Exploitation des outils statistiques pour l'intégration des données omiques en biologie végétale et animale

Emile Mardoc

UNIVERSITÉ CLERMONT AUVERGNE
ÉCOLE DOCTORALE DES SCIENCES DE LA VIE - SANTÉ - AGRONOMIE - ENVIRONNEMENT

THÈSE

Présentée par
Emile MARDOC

Pour l'obtention du grade de **Docteur d'Université**
Spécialité : **Bioinformatique**

Exploitation des outils statistiques pour l'intégration des données omiques en biologie végétale et animale

---

Date de la soutenance : 19 décembre 2023

Président du jury :

M. Said MOUZEYAR, Professeur des universités, Université Clermont-Auvergne

Rapporteurs :

M. Christophe AMBROISE, Professeur des universités, Université d'Évry-Val-d'Essone
Mme Judith BURSTIN, Directrice de recherche, INRAE Dijon
Mme Andrea RAU, Directrice de recherche, INRAE Jouy-en-Josas

Examinateurs :

Mme Marie-Laure MARTIN, Directrice de recherche, INRAE Paris-Saclay
M. Engelbert MEPHU NGUIFO, Professeur des universités, Université Clermont-Auvergne

Directeur de thèse :

M. Jérôme SALSE, Directeur de recherche, INRAE Clermont-Ferrand

Co-Encadrante de thèse :

Mme Muriel BONNET, Directrice de recherche, INRAE Clermont-Ferrand

Invité :

M. Sébastien DÉJEAN, Ingénieur de recherche, Université Paul Sabatier Toulouse 3

Résumé

Les avancées de ces dernières années dans les technologies de production de données biologiques sont à l'origine de la massification de données dites omiques, telles que les données génomiques (ADN), transcriptomiques (ARNm), protéomiques (protéines), métabolomiques (métabolites), *etc.* Ces données offrent la possibilité de décrire, en théorie, les processus biologiques les plus complexes mis en oeuvre par tous les systèmes biologiques en interaction avec leur environnement. L'enjeu méthodologique est alors de pouvoir intégrer, c'est-à-dire analyser simultanément, ces données de nature et provenance diverses pour répondre à différents questionnements scientifiques. Dans ce contexte, l'objectif de cette thèse est de proposer une approche méthodologique pour intégrer des données omiques produites dans différents contextes et l'appliquer à différents questionnements biologiques concrets chez les plantes et animaux.

Un tutoriel en 6 étapes a été développé pour préparer et mener l'intégration des données omiques, à destination des biologistes non-experts de l'intégration multi-omiques. Le tutoriel détaille alors les étapes à effectuer avant d'intégrer les données omiques, ces étapes correspondant à 1- l'acquisition des données et leur structuration sous forme matricielle, 2- la définition de la question biologique et de la stratégie intégrative associée, 3- le choix de l'outil intégratif adapté à la question et aux données, 4- le pré-traitement des données, 5- l'analyse préliminaire par jeu de données, 6- l'intégration multi-omiques. Concernant spécifiquement l'étape d'intégration (6), parmi 13 outils sélectionnés et présentés dans le manuscrit, nous avons exploité l'outil `mixOmics` et développé la fonction cimDiablo_v2 pour intégrer les données par réduction de dimension.

Ces développements méthodologiques ont été proposés dans l'optique de s'adapter à différents contextes biologiques, à savoir pour répondre à différentes questions biologiques classées en 3 stratégies intégratives (description, sélection, prédiction), en intégrant différents types de données omiques (génomiques, transcriptomiques, protéomiques, *etc.*) à différents "niveaux" (par espèces, individus, tissus, gènes, conditions expérimentales, *etc.*). Ces développements ont alors été testés sur plusieurs jeux de données biologiques, comme preuve de concept : premièrement, sur des données plantes (peuplier et céréales), afin d'identifier les profils d'interactions entre la méthylation de l'ADN et l'expression des gènes pour différentes populations géographiques d'individus (peuplier) et de stades de développement du grain (céréales), puis, deuxièmement, sur des données animales (bovin), afin d'identifier les signatures moléculaires de la composition tissulaire ou chimique des carcasses bovines en sélectionnant les protéines fortement liées aux phénotypes de la composition corporelle.

Chez les plantes, nous avons durant cette thèse 1- hiérarchisé les facteurs majeurs de variabilité des données omiques, 2- regroupé les gènes selon leur profil de méthylation et d'expression, et 3- identifié des gènes dits *master regulators-drivers*, à savoir des gènes régulés par méthylation / expression pour l'ensemble des populations (chez le peuplier) ou stades de développement du grain (chez les céréales), puis étudié leurs fonctions biologiques. Chez les animaux (bovins), nous avons proposé une liste de protéines candidates de 7 phénotypes liés à la composition corporelle, et donc à l'efficience de conversion des rations en poids de muscle, qui pourra être utilisée pour de futures études prédictives.

Mots-clés : *Données omiques*, *Intégration de données*, *Tutoriel*, *mixOmics*, *cimDiablo_v2*, *Peuplier*, *Bovin*.

Abstract

The recent advancements in biological data production technologies have led to the proliferation of *omics* data, such as genomic (DNA), transcriptomic (mRNA), proteomic (proteins), metabolomic (metabolites) data, *etc.* These data theoretically provide the opportunity to describe the most complex biological processes implemented by all biological systems interacting with their environment. Thus, the methodological challenge is to integrate, *i.e.* simultaneously analyze, these data from diverse nature and source to address various scientific questionnings. In this context, the objective of this thesis is to propose a methodological approach to integrate omics data produced in different contexts and apply it to various concrete biological questions in plants and animals.

A six-step tutorial has been developed to prepare and conduct the integration of omics data, decidated to biologists non-expert in multi-omics integration. The tutorial outlines the steps to be followed before integrating omics data, these steps corresponding to 1- data acquisition and structuring in matrix form, 2- defining the biological question and the associated integrative strategy, 3- choosing the integrative tool suitable for the selected question and the data, 4- data pre-processing, 5- preliminary data analysis, 6- multi-omics integration. Regarding the integration step (6), among 13 selected tools presented in the manuscript, we made use of the `mixOmics` tool and developed the cimDiablo_v2 function to integrate data through dimension reduction.

These methodological developments were proposed with the aim of being adaptable to various biological contexts, *i.e.* to address different biological questions classified into 3 integrative strategies (description, selection, prediction), by integrating different types of omics data (genomic, transcriptomic, proteomic, etc.) and at various levels (by species, individuals, tissues, genes, experimental conditions, *etc.*). These developments were then tested on several biological datasets as a proof of concept : firstly, on plant data (poplar and cereals) to identify interaction profiles between DNA methylation and gene expression for different geographical populations of individuals (poplar) and developmental stages of grain (cereals), and secondly, on animal data (bovine) to identify molecular signatures of tissue or chemical composition of bovine carcasses by selecting proteins strongly associated with body composition phenotypes.

In plants, we 1- ranked the major factors of omics data variability, 2- grouped genes based on their methylation and expression profiles, and 3- identified genes called "master regulators-drivers", *i.e.* genes regulated though methylation / expression interplay for all considered populations (in poplar) or stages of grain development (in cereals), and studied their biological functions. In animals (bovine), we proposed a list of candidate proteins for 7 phenotypes related to body composition, and thus, to feed conversion efficiency, which can be used for future predictive studies.

Key-words : *Omics data*, *Data integration*, *Tutorial*, *mixOmics*, *cimDiablo_v2*, *Poplar*, *Bovine*.

Remerciements

Comme l'exige la tradition, il est de mon devoir d'exprimer ici ma chaleureuse reconnaissance envers toutes les personnes m'ayant accompagné durant cette thèse, et d'exprimer aux yeux de tous et en quelques mots trois années de côtoiements, de partages et d'échanges. Ces remerciements seront donc brefs, puisqu'ils ne sauraient d'aucune manière rendre justice à toutes les interactions qui sont, peut-être même au-delà de mes travaux de thèse, ma véritable fierté. Voyez donc ces quelques paragraphes comme un pâle reflet de la gratitude et l'affection que je vous porte et que je vous ai, je l'espère, bien mieux témoignées tout au long de mon séjour au milieu des volcans d'Auvergne.

Tout d'abord, merci à **M. Christophe AMBROISE**, **Mme Judith BURSTIN** et **Mme Andrea RAU**, rapporteurs de mon jury de thèse, ainsi qu'à **Mme Marie-Laure MARTIN**, **M. Engelbert MEPHU NGUIFO** et **M. Said MOUZEYAR**, examinateurs du jury, d'avoir accepté mon invitation et d'apporter leur regard critique sur mes travaux.

Je tiens aussi à remercier tout particulièrement mon superviseur de thèse **Jerôme SALSE**, ma co-encadrante **Muriel BONNET**, ainsi que **Alyssa IMBERT** et **Mamadou Dia SOW**, pour tout ce que vous m'avez apporté, bien sûr d'un point de vue scientifique, mais aussi relationnel et humain, et pour m'avoir soutenu jusqu'au bout de la thèse.

Un grand merci à **Sébastien DÉJEAN**, **Samir DOU** et **Marie TAILLANDIER**, qui avec Alyssa m'ont suffisamment parlé de mathématiques pour que j'évite de devenir fou dans un monde de biologistes.

Merci aussi à mes deux équipes d'accueil, **PaleoEvo** et l'**équipe Biomarqueurs**, pour m'avoir bien intégré malgré le covid au début de thèse et des aller-retours assez désordonnés entre les deux équipes en fin de thèse. Merci à mes deux unités d'accueil, l'**UMR GDEC** et l'**UMR Herbivores**, une fois de plus pour m'avoir chaleureusement intégré. Et pour généraliser encore plus, merci aux sites de **Crouel** et de **Theix** pour leur accueil, et particulièrement aux cuisiniers de deux sites pour mes repas spéciaux. Merci aussi à l'**INRAE** ainsi qu'à l'**ED SVSAE** de l'UCA de m'avoir accepté comme doctorant, et à l'**ISITE CAP20-25** d'avoir financé ma thèse.

Merci aux organisatrices des JJC BAP 2023, **Larissa ADAMIK**, **Meriem BANOUH**, **Domitille COQ–ETCHEGARAY**, **Caroline FREY**, **Christelle GINOT** et **Clea SIGURET**, qui ont été suffisamment coriaces pour subir les très nombreuses réunions sans trop se plaindre (sauf une mais je dirais pas qui).

Merci à tous les **petits zigotos** qui ont transformés les pauses déjeuners en cour de récré, dirigés par notre chère **Happiness Manager**. Merci aussi à tous mes **camarades** de rando, de vélo, de badminton, de soirées, de savon (promis c'est pas aussi bizarre que ça en a l'air), *etc.* Merci à ma **famille** pour leur soutien, et notamment à ce petit fayot de **frère** que je ne pouvais pas ne pas citer à cause de ses propres remerciements.

Enfin, merci à **tous ceux qui liront cette thèse**, courage, il ne vous reste "qu'un peu plus de 300 pages".

Table des matières

Table des figures

Liste des tableaux

Liste des termes et abréviations

ACP Analyse en Composantes Principales

ADN Acide DésoxyriboNucléique

AE Auto-Encodeur

ANOVA *Analysis Of VAriance*

ARNm Acide RiboNucléique messager

BS-seq *Bisulfite sequencing*

CAH Classification Ascendante Hiérarchique

CG Cytosine-Guanine

CHG Cytosine-H-Guanine, avec H = Adénine, Cytosine ou Thymine

CHH Cytosine-H-H, avec H = Adénine, Cytosine ou Thymine

CNN *Convolutional Neural Networks*

DDA *Data-Dependent Acquisition*

DIA *Data-Independent Acquisition*

DESeq *Differential Expression analysis for Sequence count data*

FAIR *Findable, Accessible, Interoperable, and Reusable*

FDR *False Discovery Rate*

GLM *Generalized Linear Model*

KPCA *Kernel Principal Component Analysis*

LDA *Linear Discriminant Analysis*

MB-PLS *Multi-Blocks PLS*

MB-PLSDA *Multi-Blocks PLS Discriminant Analysis*

miRNA micro-Acide RiboNucléique

MLR *Multiple Linear Regression*

MS *Mass Spectrometry*

NGS *Next-Generation Sequencing*

PGD Plan de Gestion des Données

PLS *Projection to Latent Structure*

PLS-DA	*Projection to Latent Structure Discriminant Analysis*
rbd	*read by density*
RLE	*Relative Log Expression*
RNA-seq	séquençage de l'ARN
RPKM	*Reads Per Kilobase Million*
SNPs	*Single Nucleotide Polymorphisms*
sPLS	*sparse PLS*
sPLSDA	*sparse PLS Discriminant Analysis*
SWATH-MS	*Sequential Windows Acquisition of All Theoretical Spectra-Mass Spectrometry*
TMM	*Trimmed Mean of M values*
TPM	*Transcripts Per Kilobase Million*
WGBS	*Whole Genome Bisulfite Sequencing*

Contributions de la thèse

Articles en premier auteur :

— Mardoc et al. (2024) : *Genomic data integration tutorial, a plant case study*

— Mardoc et al. (en préparation) : *Integrating liver, muscle and adipocyte tissue proteomes to identify drivers of body and growth composition in bovine species*

Articles en co-auteur :

— Sow et al. (2023) : *Epigenetic Variation in Tree Evolution : a case study in black poplar (Populus nigra)*

— Bellec et al. (2023) : *Tracing 100 million years of grass genome evolutionary plasticity*

Communications orales :

— *Multi-omics integration : denoising cereal data using cimDiablo_v2 to answer several biological questions*, Journées scientifiques de l'École Doctorale (JEDs), mai 2022, Clermont-Ferrand

— *Omics data integration using a new workflow and tool : case study on Poplar*, Journées Jeunes Chercheurs du Département Biologie et Amélioration des Plantes (JJC BAP 2023), avril 2023, Clermont-Ferrand

Enseignement :

— *Biological Data Integration, M2 Plant/Plasticity*, 4h30 de cours (avec Sébastien Déjean et Jérôme Salse), 4h30 de TP (avec Samir Dou), octobre-novembre 2022, Université Clermont-Auvergne (UCA), Clermont-Ferrand

— *Statistiques 1, N2 Sciences de la vie*, 13h30 de TD, février-avril 2023, Université Clermont-Auvergne (UCA), Clermont-Ferrand

Première partie

Introduction

Chapitre 1

Contexte biologique : les données omiques

Dans ce premier chapitre, nous introduisons pour débuter la notion de données "omiques", avec une vision générale sur les données incluses dans cette classification et dénomination, leurs caractéristiques et les principaux questionnements biologiques qui peuvent y être associés. Dans un deuxième temps, nous nous focalisons sur les types de données omiques principalement étudiés durant la thèse, à savoir les données méthylomiques, transcriptomiques, protéomiques et phénotypiques.

1.1 État de l'art des données omiques

1.1.1 Définition et principales caractéristiques

Les avancées technologiques récentes dans le domaine de la biologie et de la médecine, notamment l'apparition des technologies de séquençage nouvelle-génération (ou *Next-Generation Sequencing* (NGS)) (Voelkerding et al., 2009; Slatko et al., 2018), sont à l'origine de la production massive de données génomiques à moindre coût tout en améliorant la précision et la fiabilité de ces données (Schneider and Orchard, 2011). Les données dîtes "omiques" (ou *omics*) font référence aux différents types de données biologiques, de terminaison "-omique", produites par des technologies haut-débit et caractérisant le système biologique aux différentes échelles moléculaires. Parmi elles se trouvent notamment données génomiques (Acide DésoxyriboNucléique (ADN)), épigénomiques (méthylation de l'ADN, modification des histones, *etc.*), transcriptomiques (Acide RiboNucléique messager (ARNm)), protéomiques (protéines) et métabolomiques (métabolites), comme représenté en Figure 1.1. La Figure 1.2 présente l'évolution du nombre de publications sur ces principaux types de données omiques sur les deux dernières décennies. Cependant, il existe une multitude d'autres données omiques, comme par exemple les données lipidomiques, radiomiques, ionomiques, interactomiques, *etc.* Le domaine des données omiques étant encore en pleine évolution, l'appartenance de certains types de données biologiques au sein du groupe des données dites omiques ne fait pas encore consensus, par exemple avec le cas des données phénotypiques qui sont ou non acceptées en tant que données omiques selon les études. Dans la suite de ce manuscrit, nous considérerons les données phénotypiques comme des données omiques, ou tout du moins comme une variable pouvant être intégrées aux

autres données classiquement considérées comme omiques (expression, méthylation, protéines, *etc.*).

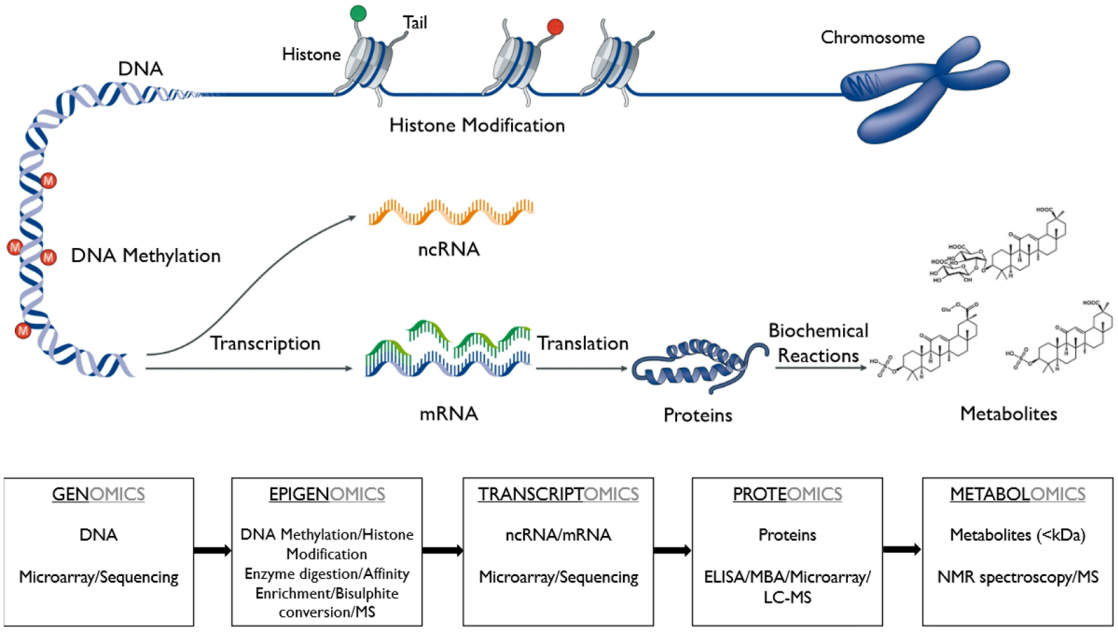


FIGURE 1.1 Principales données omiques et technologies associées
Figure et légende de Manríquez (2020). *For each of the omics data, the bottom flow chart specifies the corresponding biological structure being characterised and the different platforms employed for their quantification.Image taken and modified from Joosten et al. (2018). ncRNA : non-coding RNA, mRNA : messenger RNA, MS : Mass Spectrometry, ELISA : Enzyme-Linked Immunosorbent Assays, MBA : Multiplexed Bead Assay, LC-MS : Liquid Chromatography-Mass Spectrometry, NMR : Nuclear Magnetic Resonance.*

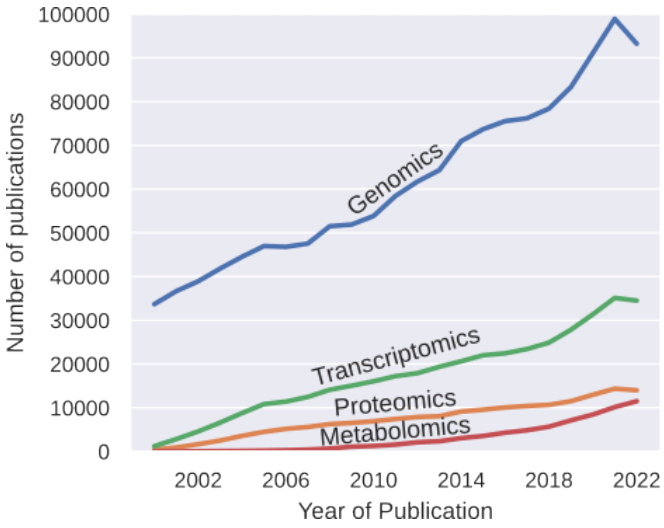


FIGURE 1.2 Évolution du nombre de publications sur les données omiques
Figure inspirée de Noor et al. (2019) et utilisant le code `https://gitlab.com/elad.noor/cosb18-multiomics-integration/` pour chercher dans le moteur de recherche PubMed le nombre de publications associées aux principaux types de données omiques.

Les données omiques représentent donc un ensemble de données moléculaires diverses, caractérisées par plusieurs particularités :

— **Données hétérogènes** : les données omiques englobent une multitude de données biologiques qui sont par essence de différentes natures à la fois de par leur provenance et leur signification biologique, mais aussi par la manière ou méthodologie dont ces données sont produites (ex : colorimétrie, séquences, spectres, imagerie, *etc.*) et normalisées (ex : comptage, pourcentage). Pour analyser ces données, il est donc nécessaire d'utiliser des outils spécifiques prenant en compte cette hétérogénéité.
— **Volume des données** : ces données sont souvent produites en très grands volumes, avec parfois des milliers voire millions de nucléotides, gènes, protéines, *etc.* soulevant des défis technologiques pour parvenir à les produire, les stocker et, comme il sera abordé ici, les analyser.
— **Déséquilibre dans les effectifs** : il est aussi très courant d'obtenir des données d'effectifs déséquilibrés. Il peut s'agir d'un déséquilibre au sein d'un jeu de données omiques, puisque ces données sont généralement produites pour un grand nombre de variables (ex : nucléotides, gènes, protéines, *etc.*) mais peu d'observations (ex : patients, animaux, plantes, *etc.*), ou bien d'un déséquilibre d'effectifs entre les variables de différents jeux de données, avec par exemple des milliers de protéines face à une dizaine de variables phénotypiques caractérisant quelques individus. Les deux cas nécessitent le développement d'outils spécifiques à l'analyse de ces données.
— **Données manquantes** : les valeurs manquantes peuvent être, d'une part, des données non produites pour une observation pour un jeu de données, par exemple si les données sont produites sur des patients pendant plusieurs semaines et qu'un patient quitte l'expérience avant le terme de l'expérimentation, et d'autre part une valeur manquante pour un couple observation/variable, par exemple si la méthode de production des données est peu précise et que certaines valeurs sont sous le seuil de détection (Flores et al., 2023).

1.1.2 Les questions biologiques

Les données omiques sont produites dans différents contextes, mais toujours dans l'optique de répondre à une ou plusieurs questions biologiques préalablement définies. Nous avons fait le choix pour la suite du manuscrit de regrouper ces questions biologiques très variées en trois catégories illustrées en Figure 1.3. Dans cette figure, les données omiques sont illustrées sous forme matricielle. Les colonnes correspondent d'une part à différentes variables omiques pour deux types de données omiques, et d'autre part à différents traits phénotypiques. Les lignes correspondent à différentes observations (des individus, des gènes, *etc.*) structurées en deux groupes. Nous proposons ainsi un cadre conceptuel de l'intégration des données omiques sous forme de matrices afin de représenter une multitude de contextes biologiques, dont l'ensemble des analyses menées durant la thèse et présentées dans la suite du manuscrit.

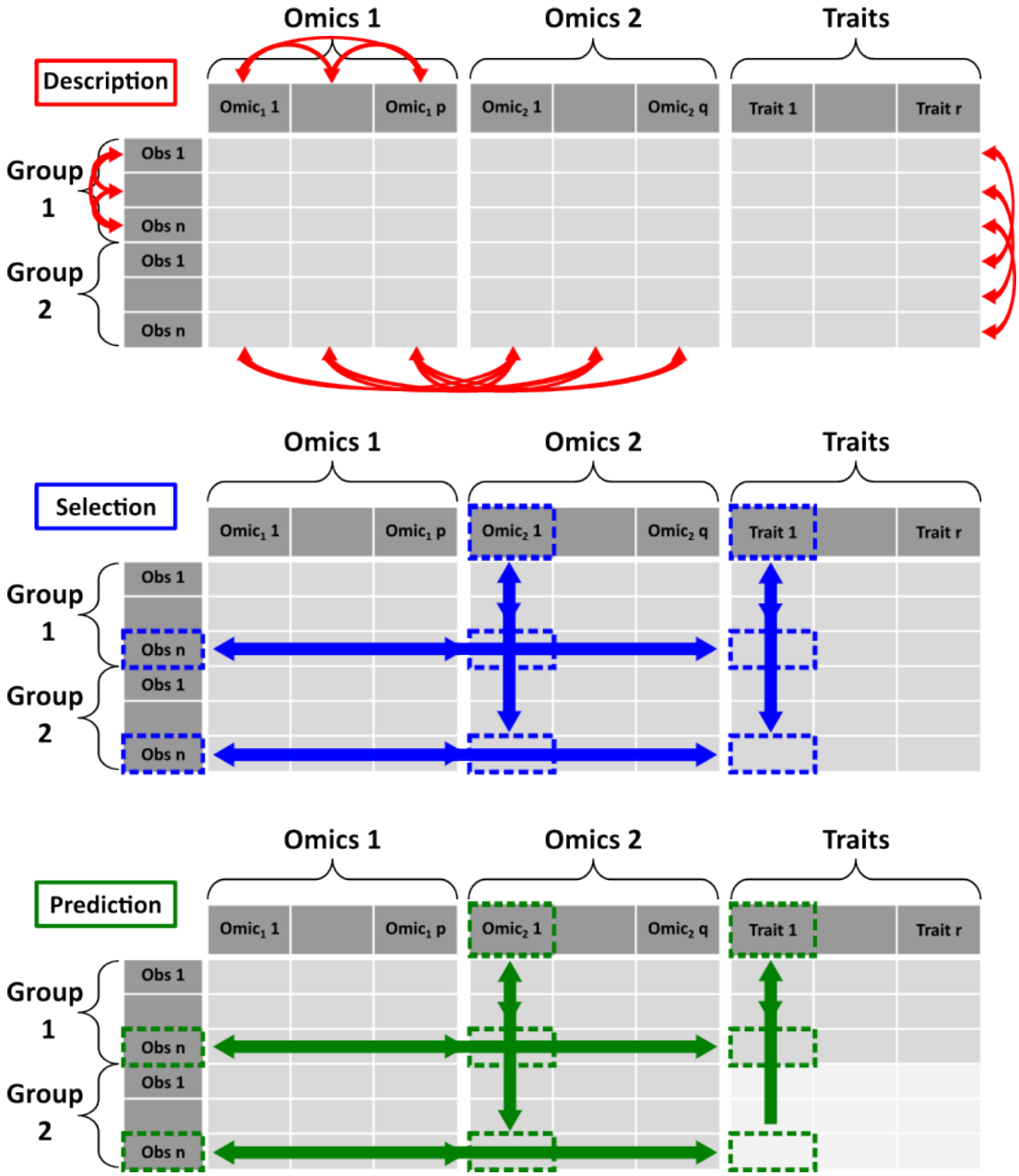


FIGURE 1.3 Schéma illustratif des trois grands types de questionnements biologiques sur un même jeu de données type

Les données sont représentées ici sous forme matricielle. Deux blocs matriciels correspondent aux données de deux types de données omiques (ex : génomiques, transcriptomiques, protéomiques, *etc.*), le troisième bloc représente des traits phénotypiques ou autres données biologiques non-omiques (ex : nombre/poids des grains pour les plantes, ou poids de carcasse pour les animaux). Chaque colonne représente une variable et chaque ligne une observation. Généralement, les observations sont des individus et les variables omiques sont alors des gènes, des protéines, *etc.* Il est aussi possible que les gènes, protéines, *etc.* soient considérés comme les observations, et les variables omiques peuvent alors être des données omiques produites à différents temps (cinétique), dans différentes conditions, *etc.*. Ici, deux groupes d'observations sont représentés et peuvent correspondre par exemple aux données produites pour les mêmes individus avant puis après traitement, ou à différentes espèces. Trois types de questions biologiques sont illustrés sur ces données. La description des relations peut s'effectuer entre variables d'un même bloc et/ou d'autres blocs, ainsi qu'entre observations d'un même groupe et/ou observations appariées dans différents groupes. La sélection peut s'effectuer sur les observations (selon leur profil atypique pour certaines variables), ou sur les variables (fortement liées entre elles pour certains individus). La prédiction peut s'effectuer sur des traits inconnus à partir de données omiques et des connaissances sur les relations omiques-traits provenant d'autres observations.

Les trois catégories de questionnements scientifiques abordés lors de l'intégration des données omiques sont pour la suite de la thèse définis autour des thèmes de (1) description, (2) sélection, (3) prédiction :

— *Description, exploration des données via leurs interactions omiques* : l'objectif est ici d'étudier les données de manière exploratoire, c'est-à-dire sans *a priori*, que ce soit pour déterminer les relations majeures entres les variables omiques et/ou entre les observations en prenant en compte la variabilité omique. Dans la Figure 1.3, ces différentes relations sont représentées avec, en colonnes, les interactions entre variables omiques issues d'un même jeu de données (ex : comparaison de l'abondance de plusieurs protéines pour différents individus) ainsi que les interactions multi-omiques (ex : impact de la méthylation sur l'expression des gènes), et en lignes les relations entre observations d'un même groupe (ex : individus similaires dans leur profil omique) ou comparaison des observations différenciées selon le groupe (ex : individus au profil omique modifié à la suite d'un traitement). Ces analyses exploratoires sont souvent suivies d'analyses complémentaires, que nous catégorisons en "sélection" et "prédiction".

— *Sélection d'entités biologiques au profil particulier* : le but est ici de sélectionner des observations/variables biologiques (ex : individus, gènes, protéines) associées à un profil omique (ex : très forte expression) et/ou phénotypique particulier, souvent sous certaines conditions (ex : en réaction à un stress ou traitement). Dans la Figure 1.3, l'$Observation\ n$ (ex : un individu avant et après traitement) a un profil particulier pour le $Trait\ 1$ (ex : individu malade avant traitement, sain après traitement) et la variable omique $Omic_2\ 1$ (ex : protéine peu sécrétée chez cet individu avant traitement, fortement sécrétée après traitement). Ainsi, si l'objectif est de sélectionner des variables aux fortes interactions et que l'interaction entre les variables $Trait\ 1$ et $Omic_2\ 1$ est aussi forte pour d'autres observations que l'$Observation\ n$, ces deux variables peuvent être sélectionnées. Si l'objectif est de sélectionner des observations pour lesquelles les profils omiques et phénotypiques sont corrélés et que les interactions entre les variables $Trait\ 1$ et $Omic_2\ 1$ sont aussi fortes pour d'autres variables, l'$Observation\ n$ pourra être sélectionnée.

— *Prédiction d'un phénotype à partir des données omiques* : enfin, il est possible de souhaiter prédire un caractère, trait, phénotype,*etc.*, inconnu à partir de données omiques acquises et des connaissances sur les interactions entre ce type de données et le caractère choisi. La Figure 1.3 illustre ceci de la manière suivante : pour des interactions omiques et phénotypiques identifiées pour certaines observations (ex : gènes ou individus) du premier groupe, cette connaissance peut être utilisée sur les observations du deuxième groupe pour lesquelles seules les données omiques ont été produites afin de prédire les phénotypes connus seulement du premier groupe. Ceci peut être utilisé par exemple pour prédire la réaction d'un individu à un traitement pour lequel la réaction pour d'autres patients est déjà connue, prédire le poids d'un animal avant l'abattage en

se servant de données d'abondance de protéines, ou encore prédire la résistance d'une céréale à un stress en se basant sur les connaissances de résistance d'autres espèces à ce stress.

1.2 Données omiques de la thèse

Au cours de la thèse, plusieurs jeux de données omiques, à la fois chez les plantes et les animaux ont été utilisés, afin de pouvoir aborder les trois catégories de questionnements scientifiques précédents et les méthodes ou approches d'intégration des données associées. Les données omiques mises en jeu consistent en des données de méthylation de l'ADN (méthylomique) et d'expression des gènes (transcriptomique) produites pour différentes espèces végétales, à savoir le peuplier *Populus nigra*, le maïs *Zea mays* et *Brachypodium distachyon*, ainsi que des données d'abondance de protéines (protéomique) et phénotypes produites chez une espèce animale, le bovin *Bos taurus*. Nous introduisons dans un premier temps ces types de données omiques séparément ainsi que leur utilisation actuelle dans le domaine végétal (méthylation, expression) et animal (protéines), puis, dans un second temps, nous introduisons l'état de l'art des connaissances sur les interactions méthylation-expression et protéines-phénotypes pour ces différentes espèces.

1.2.1 Les 3 types de données omiques utilisés durant la thèse

Données épigénétiques (méthylation de l'ADN)

La méthylation de l'ADN consiste en la fixation d'un groupement méthyle (composé d'un atome de carbone et de trois atomes d'hydrogène) à une base nucléotidique de l'ADN (adénine ou cytosine) (Moore et al., 2013). La méthylation fait partie des processus épigénétiques, c'est-à-dire des mécanismes modifiant l'expression des gènes sans en changer la séquence, et ce de manière réversible (déméthylation de l'ADN), transmissible (lors des divisions cellulaires) et adaptative (Samantara et al., 2021). Parmi les trois mécanismes de régulation épigénétique, la modification des histones, les petits ARN (micro-Acide RiboNucléique (miRNA)) et la méthylation, seul le méthylome sera décrit dans le manuscrit car exploité dans le cadre de ces travaux. Il existe différents types de méthylation, notamment la méthylation en position 5' de la cytosine (représentée en Figure 1.4) qui, pour les espèces végétales, peut prendre différentes formes : Cytosine-Guanine (CG), Cytosine-H-Guanine, avec H = Adénine, Cytosine ou Thymine (CHG), et Cytosine-H-H, avec H = Adénine, Cytosine ou Thymine (CHH). De nombreuses méthodes sont utilisées pour produire les données de méthylation, autant sur l'entièreté du génome d'un individu que sur des zones spécifiques de son génome (Singer, 2019; Li and Tollefsbol, 2021; Mattei et al., 2022; Agius et al., 2023). L'une des méthodes employées actuellement pour produire ces données de méthylome est le séquençage après conversion bisulfite (*Whole Genome Bisulfite Sequencing (WGBS)*) (Baubec and Akalin, 2016), aussi

appelé *Bisulfite sequencing* (BS-seq) (Cokus et al., 2008), utilisé pour obtenir le niveau de méthylation de chaque cytosine à l'échelle du génome entier. Ce niveau de méthylation est alors, pour chaque cytosine, la valeur moyenne du nombre de fois où elle est méthylée dans le groupe de cellules considéré. De manière générale, la méthylation de l'ADN est présentée sous forme de densité ($density = \frac{methylated\_cytosines}{methylated\_cytosines+unmethylated\_cytosines}$) ou pourcentage ($\% = density \times 100$) en considérant toutes les cytosines d'une région génique (ou génomique au sens large) choisie. Cependant, d'autres normalisations ont été proposées comme le *read by density (rbd)* ($rbd = density \times methylated\_cytosines$) présenté et utilisé dans Bellec et al. (2023) et Sow et al. (2023). La Figure 1.5 représente l'impact de ces normalisations pour différentes configurations de méthylation. Ainsi, une limite du pourcentage de méthylation, souvent considéré dans la littérature, est que l'effectif de cytosines dans la région génique/génomique considérée n'est pas pris en compte, avec un même pourcentage pour les cas 1 et 2 alors que bien plus de cytosines sont méthylées dans le cas 2. L'intérêt de la normalisation *rbd* est alors de considérer simultanément la proportion de cytosines méthylées et le nombre de ces cytosines, avec la valeur de *rbd* la plus grande lorsque beaucoup de cytosines sont présentes et toutes méthylées (cas 2), puis lorsque peu de cytosines sont présentes mais toutes méthylées (cas 1), et enfin lorsque relativement peu des cytosines présentes sont méthylées (cas 3). Nous avons donc choisi cette normalisation pour les différentes analyses de données méthylomiques de la thèse.

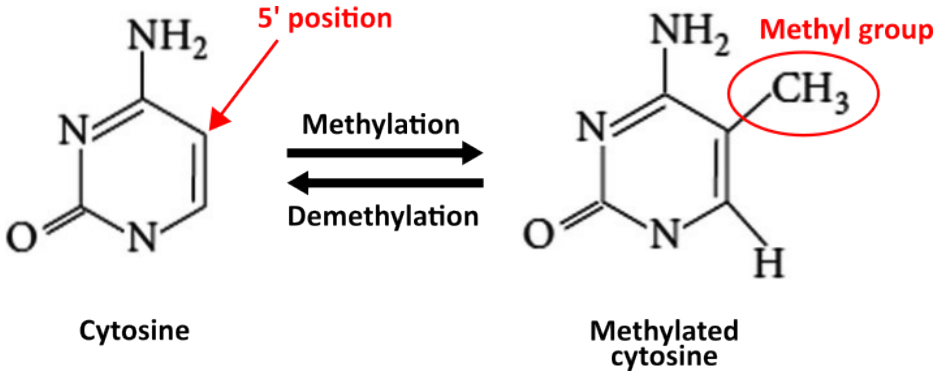
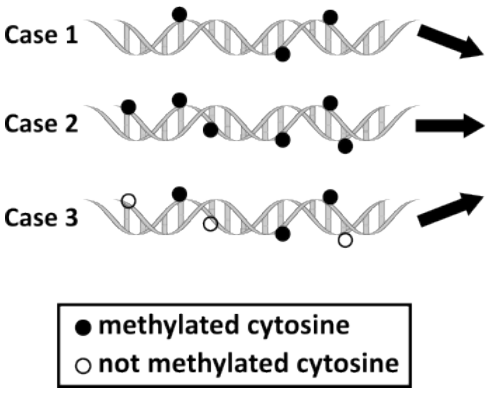


FIGURE 1.4 Méthylation d'une cytosine en position 5'
Représentation de l'attachement (méthylation) et le détachement (déméthylation) d'un groupe méthyle à la position 5' de la cytosine.



meth C	not meth C	density	%	rbd
300	0	1	100	300
600	0	1	100	600
300	300	0.5	50	150

FIGURE 1.5 Normalisations pourcentage et *rbd* pour différentes configurations de méthylation
Trois cas représentent trois régions de l'ADN différemment méthylées : 1- peu de cytosines toutes méthylées, 2- de nombreuses cytosines toutes méthylées, 3- de nombreuses cytosines dont la moitié seulement sont méthylées. Le tableau présente le nombre de cytosines méthylées (*meth C*) et non méthylées (*not meth C*), ainsi que les valeurs de densité, pourcentage et *rdb* de cytosines méthylées dans chaque cas

De nombreuses revues résument les différentes études menées sur la méthylation dans le domaine végétal, que ce soit pour tenter de comprendre les mécanismes de la méthylation (Zhang et al., 2018) et déméthylation de l'ADN (Parrilla-Doblas et al., 2019) ainsi que leur rôle dans le développement de la plante (Varriale, 2017), étudier les différences de profils de méthylation selon les individus (Takuno et al., 2016), identifier leurs réponses à différents stress (Akhter et al., 2021; Liu et al., 2022) ou encore prédire leur rôle potentiel en amélioration variétale (Shaikh et al., 2022). Plusieurs études de la méthylation de l'ADN ont été menées ces dernières années sur différentes espèces végétales, par exemple sur la croissance du grain de riz *via* la remobilisation de l'azote lors de sa croissance (Fan et al., 2022), l'impact de la déméthylation sur la maturation des fruits de tomates (Lang et al., 2017), le développement des grains de soja (An et al., 2017), l'évolution des chromosomes sexuels de l'asperge (Li et al., 2021c), mais aussi sur des réponses à différents stress biotiques ou abiotiques, comme l'infection de *Brassica rapa* à l'albugo (Tirnaz et al., 2022), la réponse au stress thermique chez le concombre (Lai et al., 2017) ou encore la réponse de la fraise à différents stress (température, hormones, sel) (López et al., 2022).

Le méthylome à l'échelle du génome a aussi été étudié sur les différentes espèces de la thèse (peuplier et céréales). Chez le peuplier, des études ont été par exemple menées sur la résistance à la sécheresse (Zhou et al., 2023), à la carence en nutriments des sols (Schönberger et al., 2016; Su et al., 2018) ou à différents pathogènes (Xiao et al., 2021), ainsi que l'identification de méthylation spécifique du déterminisme sexe (Bräutigam et al., 2017) ou dans un objectif de développement d'outils d'interactions méthylation-phénotype (Champigny et al., 2020). Chez les céréales, le méthylome a été utilisé chez *Brachypodium*, espèce modèle des graminées, pour étudier l'impact de la méthylation sur certaines variations phénotypiques (Eichten et al., 2020) et ses différences selon l'environnement (Skalska et al., 2020), chez le maïs pour la meilleure

compréhension du fonctionnement de son génome (Li et al., 2015; Xu et al., 2020; Noshay et al., 2021) et l'impact des éléments transposables (Noshay et al., 2019).

Données transcriptomiques

Les données transcriptomiques correspondent aux données issues de la transcription de l'ADN, notamment l'expression des gènes sous forme d'abondance d'ARNm. Les principales données d'ARNm actuellement produites sont des données de séquençage de l'ARN (RNA-seq) (Wang et al., 2009) qui ont progressivement remplacé les puces à ADN (*microarrays*) (Zhao et al., 2014; Mantione et al., 2014). Le séquençage des ARN produit en effet des "lectures" (*reads*) correspondant à des fragments d'ARN qui sont ensuite alignées sur un génome de référence, le comptage du nombre d'alignement représentant alors le niveau d'expression du gène. Plusieurs normalisations existent avec pour but de palier entre autres aux biais techniques liés au processus de séquençage et surtout rendre les échantillons comparables (Abbas-Aghababazadeh et al., 2018; Evans et al., 2018; Zhao et al., 2021). En transcriptomique, il existe en général deux approches de normalisation représentées en Figure 1.6 : entre échantillons (ex : entre différentes variétés) et au sein d'un même échantillon (ex : entre des groupes de gènes). Au sein d'un même échantillon (tissu/organe/individu), l'idée est de corriger les sources de variabilité pouvant être liées principalement à la taille du gène. En général dans les génomes, les gènes ont des tailles différentes et par conséquent le comptage des lectures RNA-seq est proportionnel à la taille du gène. Différentes normalisations sont proposées pour la prise en compte de la taille du gène dont les plus connus sont *Reads Per Kilobase Million (RPKM)* (Mortazavi et al., 2008) et *Transcripts Per Kilobase Million (TPM)* (Li and Dewey, 2011). Entre échantillons, la normalisation a pour but de comparer un même gène entre deux conditions biologiques distinctes, la principale source de variabilité étant la taille des librairies (couvertures ou profondeur de séquençage) entre les échantillons après le séquençage. Il existe également différentes méthodes de normalisation dont les plus utilisés sont *Relative Log Expression (RLE)* (aussi appelée *Differential Expression analysis for Sequence count data* (DESeq)) (Anders and Huber, 2010) et *Trimmed Mean of M values (TMM)* (Robinson and Oshlack, 2010). Ainsi, le type de normalisation à appliquer dépendra de la question biologique posée et des données à disposition et à normaliser en conséquence.

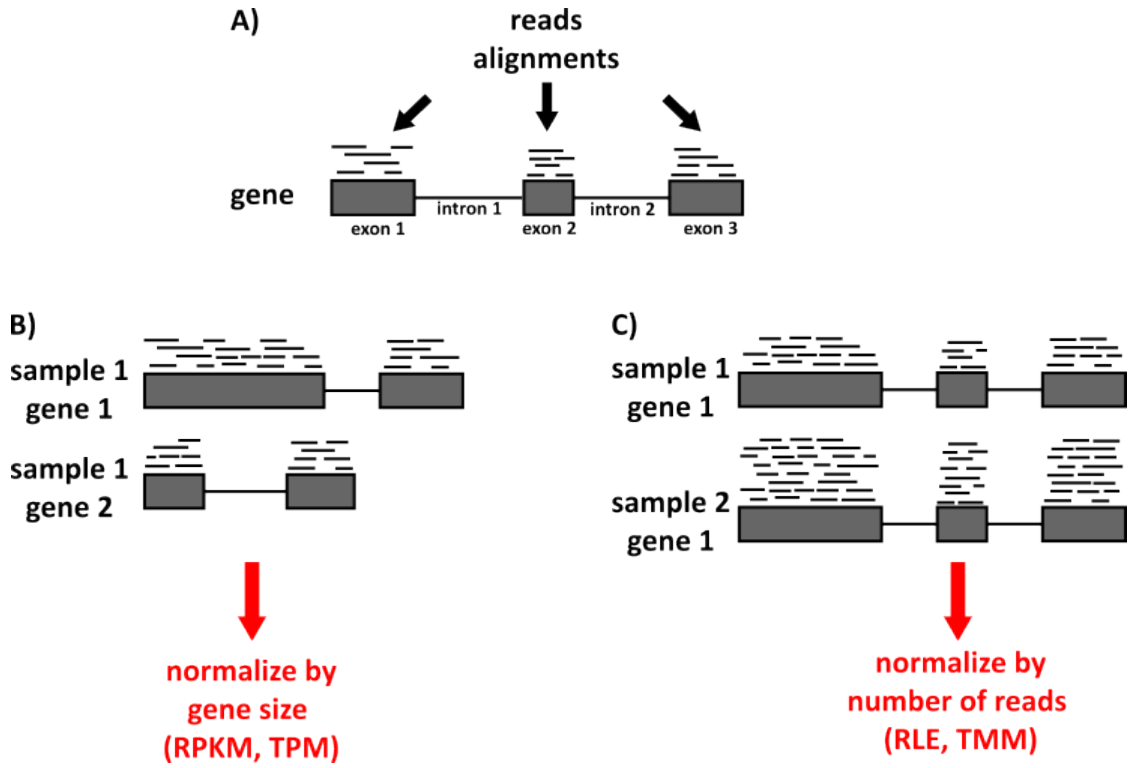


FIGURE 1.6 Normalisations par gènes ou par échantillons pour différentes configurations d'expression génomique

A- exemple d'alignement des lectures (*reads*) sur les exons d'un gène composé d'introns et exons. B- comparaison de l'expression (nombre de lectures alignées) de deux gènes de tailles différentes provenant d'un même échantillon, et normalisations proposées pour corriger le biais induit par la taille du gène. C- comparaison de l'expression d'un même gène provenant de deux échantillons différents, et normalisations proposées pour corriger le biais induit par la provenance de l'échantillon.

En biologie végétale, les analyses transcriptomiques sont utilisées sur différentes espèces pour répondre à différentes questions (Satya et al., 2022), notamment pour étudier la régulation des gènes en réponse à différents stress (Imadi et al., 2015; Wang et al., 2020b), par exemple la résistance au froid chez le riz (Shen et al., 2014), à la sécheresse chez l'eucalpyptus (Teshome et al., 2023) ou le pin (Fox et al., 2018), à la "mémoire" inter-générationnelle des plantes suite à un stress abiotique (Alves De Freitas Guedes et al., 2019), *etc.* Des données transcriptomiques ont aussi été comparées pour étudier la formation des organes et la reproduction pour différentes plantes (Julca et al., 2021) ou identifier des interactions moléculaires spécifiques entre la luzerne et son symbiote (Kang et al., 2020). Enfin, les progrès technologiques permettent maintenant la production et l'analyse des données transcriptomiques à l'échelle d'une cellule (*single-cell*) (Shulse et al., 2019; Rich-Griffin et al., 2020; Bobrovskikh et al., 2021; Shaw et al., 2021), par exemple sur le développement des racines d'*Arabidopsis thaliana* (Ryu et al., 2019; Zhang et al., 2019a) ou sur le rôle des mécanismes de fixation et de rejet du CO2 des feuilles de l'espèce *Bienertia sinuspersici* dans la photosynthèse (Han et al., 2023).

Concernant les espèces étudiées durant la thèse (peuplier et céréales), il existe aussi de nombreuses recherches se basant sur les données transcriptomiques. Chez le peuplier, des analyses ont par exemple été menées sur la régulation de gènes à différents stress, que ce soit à la séche-

resse (Yu et al., 2021) ou la salinité (Luo et al., 2017), sur la formation et le développement des racines adventives (Xiao et al., 2020; Luo et al., 2021; Yu et al., 2022), sur la formation du cambium vasculaire (Kim et al., 2019) ou encore l'expression génique à l'échelle cellulaire à partir d'une analyse *single-cell* des racines du peuplier (Chen et al., 2021b). Chez les céréales, des données transcriptomiques ont aussi été produites chez le maïs pour étudier sa résistance aux maladies (Yuan et al., 2020; He et al., 2021; Pingault et al., 2021), aux variations de contraintes abiotiques (Li et al., 2017) notamment de température (Li et al., 2021b; Sun et al., 2023a), ou encore pour la sélection de gènes du maïs liés à la résistance à différents herbicides (Sun et al., 2023b). Enfin, des études de l'expression des gènes de *Brachypodium* se sont concentrées sur la réponse aux pathogènes (Zhu et al., 2021) ou à certains stress abiotiques comme la faible teneur en phosphate (Zhao et al., 2018) ou la forte teneur en eau du sol (Rivera-Contreras et al., 2016), sur l'identification de gènes régulant les hormones (Kakei et al., 2015), et plus récemment sur l'analyse de données temporelles pour identifier des gènes impliqués dans l'immunité bactérienne (Ogasahara et al., 2022).

Données protéomiques

Les données protéomiques font aussi partie des données omiques les plus souvent produites et analysées, servant d'intermédiaire entre les données transcriptomiques et métabolomiques. Les méthodes de production de données les plus utilisées actuellement sont les méthodes de spectrométrie de masse (*Mass Spectrometry* (MS)) qui, à partir d'une matrice biologique dont les protéines ont été extraites et digérées en peptides, ionisent ces peptides, puis séparent les ions selon leur masse relative à leur charge, tout ceci afin de quantifier l'abondance des protéines. Au moins deux méthodes de quantification des signaux peptidiques sont utilisées. Une approche globale reposant sur l'extraction des courants d'ions à partir des données acquises en mode *Data-Dependent Acquisition* (DDA). Depuis les années 2010, une deuxième approche est basée sur l'isolement dans de larges fenêtres de masse et la fragmentation séquentielle de tous les précurseurs en mode *Data-Independent Acquisition* (DIA). L'une des plus récentes et précises méthodes de protéomique est la *Sequential Windows Acquisition of All Theoretical Spectra-Mass Spectrometry* (SWATH-MS) (Gillet et al., 2012; Ludwig et al., 2018). Cette méthode utilise le mode DIA, des instruments de type haute résolution/masse précise, et l'information contenue dans les spectres MS2 pour quantifier, au moins théoriquement, l'ensemble des peptides/protéines contenues dans un échantillon. Pour ce faire, la méthode détermine l'intensité des signaux des ions pour les différents ratios masse/charge à différents instants, puis analyse pour les ions ciblés la distribution de leur intensité au cours du temps, l'aire sous la courbe de distribution étant la valeur quantitative représentant l'abondance de cet ion. Cette méthode a comme principal intérêt de pouvoir quantifier des milliers de protéines tout en étant fiable et reproductible dans ses résultats, rendant possible la production massive puis la comparaison de données pour un grand nombre d'individus.

La méthode SWATH-MS étant encore récente, peu d'analyses ont été menées avec cette technologie dans le domaine de l'élevage. La méthode SWATH-MS et le traitement des données a par exemple mis en évidence des protéines associées au développement cérébral dans un modèle de fœtus porcin avec un retard de croissance (Valent et al., 2019), et des protéines du lait ont été comparées pour des chèvres de trois régions chinoises dans un but de sélection génétique (Wang et al., 2023). Toutefois, les principaux articles sur les animaux d'élevage concernent les bovins. Des analyses ont par exemple été menées pour évaluer les relations entre les abondances de protéines du muscle et la tendreté de la viande (Brandi et al., 2021; López-Pedrouso et al., 2021; Bons et al., 2021), ou encore le rôle des protéines de cellules mammaires dans la production de lait (Luo et al., 2018), d'autres pour étudier le rôle des protéines des ovaires dans la puberté et fertilité (Tahir et al., 2019) ou l'impact de la température sur les protéines du plasma séminal (Boe-Hansen et al., 2020). Enfin, des analyses ont été réalisées dans un cadre médical pour améliorer les traitements à la douleur et aux inflammations par la connaissance des réponses des protéines plasmatiques (Ghodasara et al., 2022). L'article Bons et al. (2021) compare SWATH-MS à deux autres méthodes de protéomiques visant la quantification absolue d'un nombre limité de protéines, afin de prédire à partir de l'abondance de protéines ciblées, la tendreté ou la teneur en lipides intramusculaires des viandes bovines.

1.2.2 Interactions multi-omiques

Au-delà des analyses conduites sur chaque type de données omiques individuellement, comme présenté dans la section précédente, l'étude de leurs interactions n'est pas propre à cette thèse, comme discuté ci-dessous.

Méthylation-expression

La méthylation de l'ADN, en tant que processus épigénétique, joue un rôle potentiellement majeur dans la régulation génomique. L'impact de la méthylation de l'ADN sur l'expression des gènes a ainsi été étudié depuis de nombreuses années, avant même l'essor de la production des données omiques (Razin and Cedar, 1991). Si l'effet de la méthylation semblait initialement porter sur la répression de l'expression des gènes (Newell-Price et al., 2000), d'autres effets ont depuis été déterminés, la méthylation pouvant alors, dans certains cas, aussi être associée par exemple à une augmentation de l'expression du gène (Ehrlich and Lacey, 2013).

Les interactions entre méthylation et expression ont été étudiées massivement ces dernières années chez diverses plantes (Niederhuth and Schmitz, 2017). En se restreignant aux analyses utilisant des données transcriptomiques RNA-seq et méthylomiques BS-seq, nous pouvons par exemple citer les corrélations positives (CHH promoteur) et négatives (CG et CHG corps du gène) entre méthylation et expression sous différents stress abiotiques (froid, chaleur, UV, salinité) chez le bambou (Ding et al., 2022), aux interactions méthylation-expression dans la

réponse développementale liée à de faibles teneurs en phosphate chez *Arabidposis thaliana* (Yong-Villalobos et al., 2015) ou la tomate (Tian et al., 2021) mais aussi au lien méthylation-expression en réponse à la sécheresse chez le riz (Wang et al., 2020a), ou encore chez le haricot (Zhao et al., 2022) où la méthylation du corps du gène est négativement corrélée à l'expression des gènes. De la même manière, la résistance aux larves de pyrales a été étudiée chez le soja (Zeng et al., 2021) montrant une relation négative entre la méthylation de l'ADN et l'expression des gènes en réponse à l'infection. Des données méthylomiques et transcriptomiques ont aussi été intégrées pour investiguer différents phénomènes tels que la polyploïdie et l'hétérosis chez le riz (Rao et al., 2023), la régulation de la photosynthèse chez l'ananas (Shi et al., 2021), la comparaison de caractéristiques de méthylation et leur impact sur l'expression des gènes pour différentes espèces (Seymour and Gaut, 2020), pour différents tissus de la canne à sucre (Xue et al., 2022), selon le contexte de méthylation chez *Physcomitrella patens* (Domb et al., 2020) et entre des acacias et leur progéniture (Zhang et al., 2021).

Sur les espèces étudiées dans la suite de cette thèse, plusieurs analyses faisant intervenir à la fois des données méthylomiques et transcriptomiques ont déjà été publiées. Chez le peuplier, les liens entre la méthylation de l'ADN et l'expression génique ont été étudiés en réponse à différents processus biologiques comme la réponse à la sécheresse (Liang et al., 2019), l'impact sur la croissance des tiges (Zhang et al., 2020), lors des transitions des phases de dormance et de formation du cambium (Chen et al., 2021a), ou encore lors des processus de greffage (Han et al., 2022). Enfin, les travaux sur les données peuplier présentées dans les chapitres suivants font suite à ceux notamment de Sow et al. (2021) sur les réponses de la méthylation de l'ADN et de l'expression génique lors d'une sécheresse, en utilisant des lignées génétiquement modifiées pour la méthylation de l'ADN. Chez les céréales, des travaux similaires ont été menés sur l'impact des éléments transposables sur la méthylation et l'expression génique (Wyler et al., 2020) ainsi que sur les interactions méthylation-expression sur les feuilles et bourgeons chez *Brachypodium* (Roessler et al., 2016), mais aussi chez le maïs dans le cadre d'étude de différences de méthylation et d'expression des gènes (Anderson et al., 2018) et l'effet de la déméthylation sur l'expression génique dans l'endosperme du grain (Xu et al., 2022).

Protéines-phénotypes

Comme annoncé précédemment, certains types de données biologiques ne font pas consensus dans leur appartenance ou non à la famille des données omiques. C'est le cas pour les données phénotypiques, pour lesquelles le terme "phénomiques" est parfois employé mais restent majoritairement considérées comme non omiques. Néanmoins, qu'elles le soient ou non, les données phénotypiques se distinguent par leur présence dans un grand nombre d'analyses omiques. Ainsi, les articles cités précédemment pour leur utilisation de la méthode SWATH-MS dans la production de données protéomiques font intervenir des phénotypes, souvent dans le but de déterminer des protéines biomarqueurs d'un ou plusieurs de ces traits.

1.3 Points clés

— Les avancées technologiques des dernières décennies sont à l'origine de la production toujours plus massive de données, produites à la fois sur plus d'individus et avec plus de données par individu. Cette croissance de données acquises ouvre de nouvelles possibilités pour répondre à des questions biologiques plus complexes par la prise en compte dans les analyses des interactions multi-omiques.

Il existe une multitude de types de données omiques, à savoir de données biologiques caractérisant le système aux différentes échelles moléculaires, dont certaines se démarquent en étant plus souvent citées, produites et analysées, notamment les données génomiques, épigénomiques, transcriptomiques, protéomiques et métabolomiques.

Cette abondance de données produites implique de nouveaux défis techniques pour analyser ces données de grandes dimensions, hétérogènes, d'effectifs déséquilibrés et pouvant contenir des valeurs manquantes.

— Afin d'appréhender la notion d'intégration multi-omiques, nous avons défini un cadre conceptuel représentant sous forme de matrices la principale configuration de données ainsi que leur intégration basée sur trois questionnements scientifiques, à savoir (1) la description des relations entre variables omiques et/ou observations, (2) la sélection d'observations au profil omique (et/ou phénotypique) particulier et (3) la prédiction de variables omiques (et/ou de phénotypes) à partir de données omiques.

Dans la suite de la thèse, afin d'aborder les méthodologies nécessaires pour appréhender les différents questionnements scientifiques de l'intégration des données omiques, les données omiques utilisées sont chez les plantes (peuplier et céréales) la méthylation de l'ADN (BS-seq) et l'expression des gènes (RNA-seq), et chez les animaux d'élevage (le bovin) les protéines de différents tissus (SWATH-MS) associées à plusieurs variables phénotypiques.

Chapitre 2

Cadre statistique : des analyses uni-variées aux multi-variées

2.1 Contexte statistique

2.1.1 Les statistiques pour l'analyse de données

La science des données (*data science*) est une science relativement récente, en plein essor depuis le début du XXI^{e} siècle et la production exponentielle de données issue de la démocratisation des outils informatiques et d'internet (Donoho, 2017). Elle fait référence à la fois à la production, au stockage et à l'analyse des données d'origines très diverses. La science des données joue maintenant un rôle essentiel dans la plupart des secteurs, autant dans le domaine de l'industrie, du transport, *etc.* que de la biologie, et fait pour cela intervenir différentes disciplines mathématiques et informatiques, notamment les statistiques pour l'analyse de données (Cao, 2018).

Il existe ainsi de nombreuses méthodes statistiques applicables aux données biologiques développées pour répondre à différentes questions scientifiques, par exemple pour sélectionner des variables d'intérêt, prédire des phénotypes inconnus, regrouper les variables biologiques les plus similaires ou avec les plus fortes interactions, *etc.* La diversité des approches statistiques couplée à la généricité de la plupart des méthodes permet aussi l'adaptation à différents types de données, que ce soient des données qualitatives et/ou quantitatives, de plus ou moins grandes dimensions, indépendantes ou corrélées, appariées ou non, *etc.*

L'une des manières de catégoriser ces différents outils est de les séparer selon le nombre de variables, voire de jeux de données à analyser. Il est alors possible d'analyser les données par variable avec l'analyse uni-variée, par couple de variables avec l'analyse bi-variée, pour un nombre supérieur de variables avec l'analyse multi-variée, et enfin pour plusieurs jeux de données omiques avec l'intégration multi-omiques.

L'analyse uni-variée, c'est-à-dire l'analyse de chaque variable (ex : un gène, une protéine) séparément des autres, est sans doute l'analyse la plus simple à mener, et est suffisante pour répondre à un certain nombre de questions. Les analyses bi-variées et multi-variées au sein d'un jeu de données omiques considèrent plusieurs variables afin de répondre à des questionnements plus complexes. Enfin, l'analyse simultanée de différents jeux de données omiques, appelée "intégration multi-omiques", approfondit encore plus les liens biologiques entre les données à

disposition, en contrepartie d'une difficulté accrue à les analyser et à interpréter correctement les résultats obtenus.

2.1.2 Vocabulaire

Les domaines des statistiques et science des données possèdent leur propre langage et une multitude de termes spécifiques rendant difficile leur compréhension par des non-initiés. Certains termes fréquemment employés, et nécessaires pour la compréhension des différentes sections du manuscrit, sont donc présentés ci-dessous :

— **Variables, observations, individus, échantillons** : une variable correspond à une caractéristique qui peut varier d'une observation à une autre. Une variable peut ainsi être par exemple un gène, une protéine, *etc.*, et une observation un patient, une plante, un animal, *etc.* Les termes "individus" et "échantillons" sont parfois utilisés à la place d' "observations". Toutefois, tous ces termes peuvent aussi être employés autrement selon l'étude, c'est pourquoi il est nécessaire de clarifier au début de chaque étude la nature des données et les termes choisis pour les définir, par exemple sous forme de métadonnées. Dans cette thèse, nous appelons "variables" les colonnes des jeux de données, et "observations" les lignes, comme proposé dans Wickham (2014). Ainsi chez les données plantes du Chapitre 5, les gènes seront considérés en tant qu'observations, et les protéines chez les animaux du Chapitre 6 seront considérées en tant que variables.

— **Valeurs, rangs** : les données consistent en des valeurs pour différentes variables et observations. La plupart des outils d'analyses de données analysent les valeurs, mais il existe aussi des outils analysant à la place les rangs des données, à savoir leur numéro d'indexation après avoir ordonné les valeurs, dans le but de ne pas biaiser les résultats par des valeurs trop particulières.

— **Qualitatif, quantitatif** : une manière de caractériser les données est d'utiliser les termes *qualitatif*, représentant des groupes, et *quantitatif*, représentant des quantités. Des exemples sont présentés en Table 2.1.

— **Indépendant, corrélé, apparié** : les données indépendantes sont comme leur nom l'indique sans lien *a priori*, c'est-à-dire que la modification de l'une n'a pas d'impact direct sur la modification de l'autre. Les données corrélées sont au contraire des données liées selon la métrique choisie. Enfin, les données appariées sont produites pour un même individu à différents instants ou dans différentes conditions.

— **Normalisation, transformation, normal, gaussien, normalité** : normaliser les données revient à leur appliquer une transformation afin que les données transformées suivent une loi normale, aussi appelée loi gaussienne, qui est une des lois de probabilités les plus utilisées en statistiques. Néanmoins, ce terme est aussi souvent employé pour référencer d'autres types de transformations des données, par exemple en centrant et réduisant les données ou en changeant d'échelle de sorte à obtenir des valeurs entre 0

et 1. Dans cette thèse, le terme de *transformation* sera privilégié à celui de *normalisation* lorsque la transformation en question ne porte pas particulièrement sur l'obtention de données gaussiennes, et le terme *gaussien* sera privilégié à celui de *normal* afin d'éviter des ambiguïtés. La "normalité" des données signifie que les données suivent une loi gaussienne.

— **Centré, réduit** : les données sont centrées lorsque leur moyenne est nulle (égale à 0), et réduites lorsque leurs variance et écart-type sont égaux à 1, afin de comparer des variables d'échelles initialement différentes.

— **Dimensions du jeu de données** : les dimensions des données représentent leur nombre de lignes et colonnes.

— **Combinaison linéaire** : une combinaison linéaire est une relation mathématique entre plusieurs termes, construite en multipliant chacun d'eux par un coefficient nommé "poids" puis en sommant tous ces résultats. Le choix des poids dépend de la méthode et de l'utilisation de la combinaison linéaire.

— **Effet batch** : effet provenant de l'expérimentation et introduisant un biais non souhaité dans les données produites, par exemple une différence non souhaitée des conditions expérimentales de production de plantes entre deux chambres de cultures.

— **Valeur aberrante (*outlier*), valeur extrême (*extremum*)** : valeur singulière, contrastée par rapport aux autres valeurs mesurées. Nous proposons dans ce manuscrit de parler de valeur "aberrante" lorsque la valeur provient par exemple d'une erreur de mesure ou de saisie et est à supprimer, et de valeur "extrême" lorsqu'elle représente un effet biologique recherché.

— **Supervisé, non-supervisé** : les algorithmes supervisés prennent en compte deux groupes de données, à savoir les explicatives (données d'entrée) et les expliquées (données de sortie). L'objectif est alors d'identifier les relations entre ces données afin de pouvoir expliquer, puis prédire les données expliquées à partir des données explicatives. Au contraire, les algorithmes non supervisés n'ont pas de données de sortie, et ont pour but de révéler l'information cachée dans les données.

— **Apprentissage automatique (*Machine Learning*)** : branche de l'analyse de données et de l'intelligence artificielle. Consiste pour un modèle ou un algorithme à apprendre à partir des données, *i.e.* à choisir ses paramètres en fonction des données. Les données sont généralement séparées en deux groupes : le jeu de données d'apprentissage (ou d'entraînement) sur lequel le modèle s'entraîne, et le jeu de données de test sur lequel le modèle est testé.

— **Phénomène de "surapprentissage" (*overfitting*)** : phénomène apparaissant lors de l'apprentissage d'un modèle sur un jeu de données d'entraînement, et qui consiste à déterminer des paramètres trop dépendants du jeu de données. Le modèle n'est alors pas généralisable à de nouvelles données.

— **Validation croisée** : méthode d'estimation de paramètres qui consiste à employer de multiples fois un modèle de paramètre optimal inconnu en faisant varier ce dernier. Pour cela, la validation croisée sépare les données en deux groupes, le jeu de données d'apprentissage et celui de test, et répète le processus plusieurs fois en répartissant différemment les données dans les groupes.

— **Variable latente, composante** : une variable latente est une variable calculée à partir des données. On oppose alors les variables latentes, obtenues indirectement, aux variables directement observables et mesurables qui composent les jeux de données. Les composantes sont des variables latentes calculées par combinaison linéaire des variables des données.

	Types de données	**Exemples**
Quantitatives : représentent des quantités	continues	proportions/pourcentages : valeurs allant de 0 à 1 ou à 100
		distances : valeurs positives ou nulles
	discrètes	comptages : 0, 1, 2, 3, *etc.*
Qualitatives : représentent des groupes	binaires	0 ou 1
		oui ou non
		traitement A ou traitement B
	groupes cardinaux (sans ordre, nominal)	traitement A ou traitement B ou pas de traitement ⇒ sans ordre particulier
	groupes ordinaux (avec ordre, ordinal)	forte réponse ou réponse moyenne ou pas de réponse à un traitement ⇒ ordonnées par score

TABLE 2.1 Données qualitatives et quantitatives
Exemples de données qualitatives et quantitatives classées en différentes catégories.

2.2 De l'uni-varié au multi-varié : analyses mono-omiques

Face à la multitude de méthodes existantes dans le domaine de l'analyse de données, et le nombre de paramètres à prendre en compte pour choisir les méthodes adaptées à la question biologique et aux données accessibles, de nombreux schémas et tableaux récapitulatifs ont été proposés pour présenter les principales méthodes à utiliser selon le contexte. La Figure 2.1 par exemple propose de choisir l'outil statistique selon le nombre de variables considérées, le fait qu'elles soient qualitatives et/ou quantitatives, le nombre de groupes dans la variable, *etc.* D'autres critères peuvent être ajoutés, comme `https://statswithcats.files.wordpress.com/2010/08/selection-methods-8-21-2010.png` proposant de sélectionner les méthodes selon l'objectif d'analyse, à savoir pour décrire, classifier, comparer, prédire ou expliquer les données.

Nous proposons de classer en Tableau 2.2 certains des principaux outils, méthodes, mesures pour l'analyse d'une, deux ou plus de deux variables. Chaque méthode est alors catégorisée dans une seule des grandes familles de méthodes, bien que certaines méthodes puissent en réalité être associées à plusieurs familles.

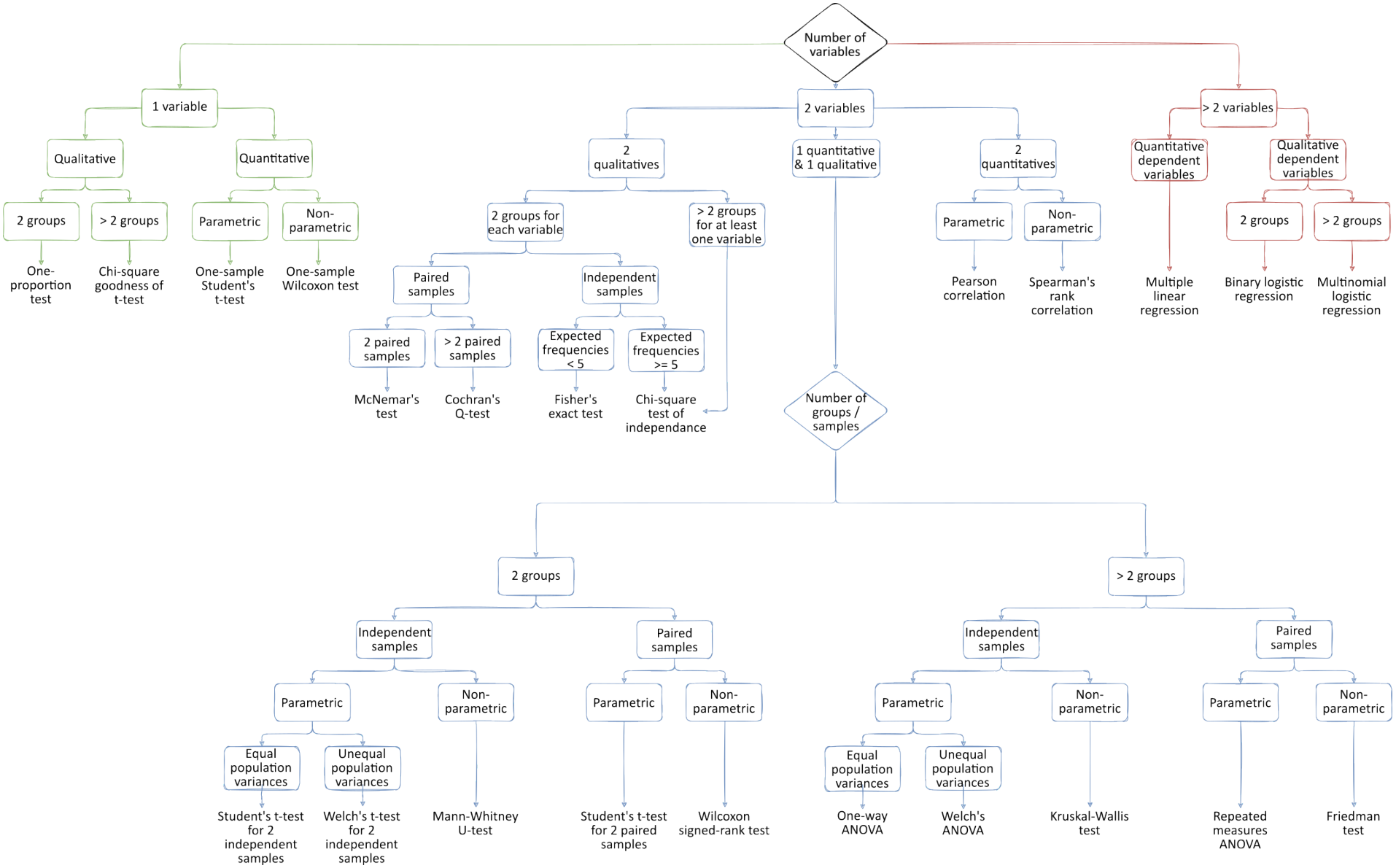


FIGURE 2.1 Organigramme pour la sélection de méthodes et tests statistiques selon les données à disposition.

Organigramme proposant de choisir l'outil statistique à utiliser selon le nombre de variables, le type qualitatif ou quantitatif des données, le nombre de groupes, la dépendance des données, la normalité supposée des données (test paramétrique), le nombre d'échantillons et leur répartition équilibrée ou non dans différentes populations. Figure modifiée à partir de `https://statsandr.com/blog/what-statistical-test-should-i-do/`, juin 2023.

Nombre de variables	Type de méthode	Objectif	Méthode / Algorithme
1 variable (Univarié)	Statistiques descriptives	Décrire les variables à partir de quelques métriques	Moyenne
			Médiane
			Variance
			Ecart-type
	Graphiques sur les distributions	Visualiser la distribution des variables	Diagramme en boîte (*Box plot*)
			Histogramme
			Diagramme en bâtons (*Bar plot*)
	Tests statistiques	Comparer les variables ou observations à une valeur théorique	Test de Student (t-test) de conformité
			Test de Wilcoxon
			Test du Khi-2 de conformité
			Test d'*ANOVA* (Analyse de la variance)
			Test de Shapiro
2 variables (Bivarié)	Corrélations	Mesurer la ressemblance entre couples de variables ou observations	Corrélation de Pearson
			Corrélation de Spearman
	Tests statistiques	Comparer les variables ou observations entre elles	Test de Student (t-test) d'homogénéité
			Test de Mann-Whitney-Wilcoxon (U-test)
			Test du Khi-2 d'homogénéité
	Graphiques en X-Y	Visualiser les observations par couple de variables	Nuage de points (*Scatter plot*)
	Modèles et régressions linéaires	Exprimer une variable à partir d'une autre	Régression linéaire simple
2 variables ou plus (Multivarié)	Factorisation matricielle	Réduire le nombre de variables tout en conservant la majeure partie de l'information	ACP (*PCA*) (Analyse en Composantes Principales)
	Classification (*clustering*)	Regrouper les variables ou observations	CAH (*HAC*) (Classification Ascendante Hiérarchique)
			K-moyennes (*K-means*)
	Modèles et régressions linéaires	Exprimer une variable comme la combinaison d'autres variables	MLG (*GLM*) (Modèle Linéaire Généralisé)
			MLR (Régression Linéaire Multiple)
			PLS (*Projection to Latent Structure*)
			Régressions Lasso / Ridge / Elastic Net
	Analyses discriminantes	Identifier les variables spécifiques à un groupe d'observations	*PLS-DA* (*Projection to Latent Structure - Discriminant Analysis*)
			ADL (*LDA*) (Analyse Discriminante Linéaire)
	Arbres de décision	Prédire la valeur d'une variable à partir d'autres variables	Forêt aléatoire (*Random forest*)

TABLE 2.2 Principales méthodes de l'analyse uni-, bi- et multi-variée de données
Description des principales méthodes d'analyse de données selon le nombre de variables considérées, le type de méthode et l'objectif général de ce type de méthodes.

Ces familles de méthodes font partie des plus fréquemment utilisées pour catégoriser les différentes méthodes d'analyse, bien qu'il soit possible de les classer autrement, par exemple en méthodes supervisées, semi-supervisées et non-supervisées (Alloghani et al., 2020). Afin d'aider les biologistes non-experts à orienter leur choix dans l'analyse des données omiques, et non dans le but d'apporter une description exhaustive des méthodes/approches à disposition, nous résumons ces familles de méthodes de la manière suivante :

— **Statistiques descriptives** : comprennent les mesures classiques pour décrire simplement les données. Les principales mesures sont la moyenne et la médiane, exprimant respectivement la somme des valeurs sur le nombre de valeurs considérées, et la va-

leur pour laquelle la moitié des valeurs considérées lui sont inférieures ou égales, et l'autre moitié supérieures ou égales. La variance et l'écart-type mesurent la dispersion des valeurs, l'écart-type étant la racine carrée de la variance. D'autres mesures peuvent aussi être utilisées comme les quantiles, les minimum et maximum, le mode, l'étendue, *etc.* Les statistiques descriptives sont généralement utilisées au début de chaque analyse puisqu'elles expliquent avec quelques métriques simples certaines des principales informations des variables et jeux de données, qui peuvent ensuite être comparés selon ces métriques. Il est aussi possible grâce à elles d'identifier par exemple des valeurs aberrantes.

— **Graphiques sur les distributions** : représentent graphiquement la distribution des données. Les diagrammes en boîtes, ou *boxplots*, représentent simultanément la médiane, les quartiles, le minimum et le maximum des données. Les histogrammes représentent empiriquement la distribution de données continues en comptant le nombre de valeurs pour différents intervalles, les diagrammes en bâtons (*barplots*) sont leurs équivalents sur données discrètes. Ces analyses sont généralement menées au début de chaque analyse pour visualiser les données et compléter les informations obtenues par les analyses statistiques descriptives utilisant des métriques. Il est alors possible de voir comment les valeurs se répartissent globalement mais aussi identifier des valeurs particulières.

— **Tests statistiques** : rejettent ou non une hypothèse faite sur les données (appelée "hypothèse nulle"). Le test de Student compare la moyenne de la variable observée à une moyenne théorique ou bien les moyennes de deux variables, sous condition de normalité (*i.e.* de suivre une loi gaussienne), alors que le test de Mann-Whitney-Wilcoxon compare les rangs des observations après avoir ordonné les variables sans condition sur leur distribution. Les tests du khi2 s'appliquent à une ou deux variables qualitatives, en uni-varié en testant par exemple l'adéquation de la répartition observée à une répartition théorique, et en bi-varié l'équation entre deux répartitions observées. Les tests d'analyse de variance (ou *Analysis Of VAriance* (ANOVA)) testent l'égalité des moyennes pour différents groupes contenus dans une variable qualitative afin de déterminer si les observations sont toutes issues de la même population ou si un groupe se distingue. Le test de Shapiro teste la normalité des données. Une multitude d'autres tests statistiques testent par exemple les variances, les valeurs extrêmes, *etc.* De plus, les tests sont souvent séparés entre d'une part les tests paramétriques, *i.e.* lorsqu'une condition sur les données est exigée comme la normalité de ces données, et les tests non-paramétriques, utilisables dans plus de contextes mais généralement moins puissants que les tests paramétriques lorsque les conditions sur les données sont satisfaites. Le choix du test dépend donc de la question posée, mais aussi des données à disposition, notamment sur le nombre de variables, le fait qu'elles soient qualitatives et/ou quantitatives, qu'elles soient appariées ou non, *etc.* Enfin, lorsque plusieurs tests sont menés, la probabilité de rejeter au

moins une fois à tord l'hypothèse nulle augmente, il est alors nécessaire d'effectuer une correction de tests multiples pour contrôler cette erreur (Noble, 2009). Ainsi, les tests statistiques sont utilisés pour comparer des variables ou observations, par exemple dans le but d'identifier des gènes différentiellement exprimés, un individu aux valeurs significativement différentes des autres individus, *etc.*, et ceci avec un seuil sur le taux d'erreur accepté afin de justifier mathématiquement du rejet ou non de l'hypothèse nulle.

— **Corrélations** : expriment la similarité dans la variabilité de deux variables ou observations. La corrélation la plus couramment utilisée est celle de Pearson, qui mesure la dépendance entre deux variables gaussiennes en calculant leur covariance divisée par leurs écart-types. La corrélation de Spearman est une mesure analogue dans le contexte non paramétrique, c'est-à-dire sans *a priori* sur les distributions, en comparant le rang des données ordonnées plutôt que leurs valeurs. Les corrélations sont une manière simple d'identifier quelles variables/observations sont les plus semblables en terme de variabilité, et sont donc utilisées directement pour chaque couple de variables ou observations, les résultats se trouvant alors généralement sous forme de matrice de corrélations, ou bien en tant que métrique utilisée par un autre algorithme.

— **Graphiques en X-Y** : représentent sur un graphique la première variable par rapport à la deuxième. Chaque point correspondant à une observation a alors pour coordonnées ses valeurs sur ces deux variables. Ces graphiques sont utilisés pour visualiser les liens entre les points pour les deux variables, par exemple pour identifier des points regroupés et donc semblables (ex : clusters de gènes, de protéines, d'individus, *etc.*), ou au contraire des points au profil particulier (valeur extrême, individu atypique, *etc.*).

— **Modèles et Régressions linéaires** : cherchent à exprimer une variable expliquée Y par une combinaison linéaire d'éléments d'une ou plusieurs variables explicatives X. Les modèles linéaires simples étant exigeants sur les types de relations entre les variables, le modèle linéaire généralisé (ou *Generalized Linear Model* (GLM)) ajoute une fonction de lien afin de généraliser le modèle à des relations plus complexes. Des tests statistiques tels que les tests de Student et ANOVA font aussi appel à des modèles linéaires, tout comme l'ensemble des régressions linéaires. Alors que la régression linéaire simple n'a qu'une variable explicative X en plus de la variable expliquée Y, la Régression Linéaire Multiple (ou *Multiple Linear Regression* (MLR)) la généralise à plusieurs variables explicatives. La régression *Projection to Latent Structure* (PLS) calcule des variables latentes analogues aux composantes de l'Analyse en Composantes Principales (ACP) et mène la régression de la ou des variables expliquées Y sur ces variables latentes au lieu des variables explicatives. Les régressions Lasso, Ridge et Elastic Net consistent en des régressions linéaires auxquelles sont ajoutées un terme de régularisation, *i.e.* une contrainte sur les coefficients du modèle, dans le but d'éviter que les coefficients calculés ne soient trop dépendants des données (phénomène de "surapprentissage", ou

overfitting). Ces analyses sont effectuées par exemple pour prédire à partir de certaines variables une variable beaucoup plus difficile à acquérir (ex : prédire certains caractères phénotypiques au terme de la croissance à partir d'abondances de protéines produites durant la croissance de l'individu) ou identifier les principales variables X explicatives de Y (ex : quelles sont les variables de méthylation de l'ADN qui sont les plus explicatives de l'expression des gènes).

— **Factorisations matricielles** : considèrent les jeux de données comme étant des matrices qu'il est possible de décomposer en un produit de matrices de dimensions inférieures, afin de réduire les dimensions tout en conservant la variabilité des jeux de données. L'ACP se concentre sur un jeu de données correspondant à une seule matrice. Elle calcule alors de manière itérative des composantes contenant une majeure partie de la variabilité initiale non encore considérée par les précédentes composantes. La concaténation de ces composantes donnent une première matrice qui, multipliée par une matrice de poids, se rapproche sensiblement de la matrice initiale. Les régression PLS et PLS multi-blocs (ou *Multi-Blocks PLS* (MB-PLS)) utilisent des factorisations matricielles sur plusieurs matrices. Ces méthodes ont pour intérêt de réduire les données tout en conservant la majeure partie de l'information, et sont donc principalement utilisées sur des données de grandes dimensions, afin de n'analyser que quelques composantes au lieu de milliers de variables.

— **Classifications (*clustering*)** : ont pour but de regrouper les variables ou observations selon leur degré de ressemblance. La Classification Ascendante Hiérarchique (CAH) regroupe progressivement les éléments ou groupes d'éléments, créant ainsi un arbre qu'il est possible de couper au seuil choisi. Il peut être utilisé pour déterminer le nombre optimal de groupes ainsi que l'appartenance des éléments dans ces groupes. L'algorithme des K-moyennes, ou *K-means*, crée K groupes et détermine de manière itérative à quel groupe doit appartenir chacun des éléments, modifiant ainsi les groupes, jusqu'à convergence du modèle. Ces méthodes sont utilisées pour identifier des groupes de gènes, protéines, individus, *etc.* aux profils omiques similaires, et/ou séparer les principaux profils en différents groupes.

— **Analyses discriminantes** : cherchent à trouver la combinaison de variables explicatives expliquant, séparant et prédisant au mieux des groupes prédéfinis. La *Projection to Latent Structure Discriminant Analysis* (PLS-DA) est alors la version discriminante de la régression PLS pour laquelle la variable expliquée Y est qualitative et non plus quantitative. L'analyse discriminante linéaire (ou *Linear Discriminant Analysis* (LDA)) cherche une combinaison linéaire de variables de X séparant au mieux les groupes de la variable expliquée Y. Ces analyses sont par exemple utilisées pour définir les gènes ou protéines liées à une typologie de pathologies cancéreuses ou à l'efficacité d'un traitement.

— **Arbres de décision** : sont construits de manière à déterminer / prédire la classe (pour

un arbre de classification) ou la valeur (pour un arbre de régression) d'une variable d'intérêt à partir des autres variables pour une observation donnée. La racine et chaque nœud de décision de l'arbre contient une règle de division sur une variable qui spécifie comment les données doivent être séparées en fonction de cette caractéristique. Le chemin déterminé par les décisions successives mène à une des feuilles de l'arbre indiquant la classe / valeur finalement prise par la variable d'intérêt. Afin d'éviter un phénomène de "surapprentissage" (ou *overfitting*), l'algorithme *Random Forest* produit plusieurs arbres en choisissant aléatoirement les variables à considérer puis, lors de la classification, conserve la valeur majoritairement déterminée par les différents arbres. Les arbres de décision peuvent être utilisés par exemple pour prédire des typologies de pathologies cancéreuses de patients à partir de variables omiques.

2.3 Points clés

— La science des données est un domaine récent utilisant entre autres les statistiques pour l'analyse des données dans différents secteurs, dont la biologie. Comme tout domaine scientifique, il possède un vocabulaire spécifique avec quelques concepts majeurs à maîtriser pour comprendre les outils et pouvoir analyser correctement les données. Une classification des concepts majeurs a été proposée afin d'introduire les méthodes qui seront utilisées dans le cadre de ce travail de thèse.

— Il existe de nombreux outils pour mener différents types d'analyses. Au sein d'un jeu de données omiques, l'analyse peut être uni-variée, bi-variée ou multi-variée. Lorsque plusieurs jeux de données omiques sont analysés simultanément, on parle d'intégration multi-omiques. Plus l'analyse prend en compte de variables et jeux de données, plus elle permet de répondre à des questionnements biologiques complexes, mais cela rend aussi plus difficile l'analyse de ces données. Les outils de l'intégration multi-omiques sont décrits dans la prochaine section du manuscrit.

Chapitre 3

Cadre statistique : l'intégration multi-omique

L'intégration de données, c'est-à-dire l'analyse simultanée de données hétérogènes, est un domaine en fort développement depuis la fin des années 2010 afin de répondre au contexte de production massive de données omiques. L'intégration multi-omiques a alors pour objectif d'apporter des réponses à des questions biologiques de plus en plus complexes par l'analyse conjointe de différents types de données omiques pour l'étude fine du fonctionnement de systèmes biologiques complexes. Ce nouveau domaine de l'analyse de données biologiques est ainsi associé à ses propres concepts, intérêts, enjeux, et bien sûr outils et méthodes spécifiques introduits ici.

3.1 Enjeux biologiques de l'intégration multi-omiques

En biologie, l'intégration multi-omiques consiste à identifier les variations omiques caractérisant les individus, populations ou espèces, pour mettre en évidence les mécanismes clés de leur développement, de leur adaptation ou de leur tolérance à des contraintes biotiques ou abiotiques en interaction avec leur écosystème. Ces analyses intégratives peuvent alors être menées, par exemple, pour conduire à de l'amélioration variétale adaptée au contexte agricole actuel et à venir, notamment au changement climatique et à la transition agro-écologique. L'intégration de différentes données omiques représente ainsi des enjeux à la fois à l'échelle des populations d'une espèce (intra-spécifique) et/ou entre espèces (inter-spécifique). À l'échelle populationnelle, l'intégration offre notamment la possibilité d'identifier des traits d'intérêt agronomique sur quelques individus, qui pourraient être adaptés à des conditions pédo-climatiques particulières. De même, à l'échelle multi-espèces, un objectif est de transférer les connaissances d'une espèce étudiée à une espèce proche (recherche translationnelle), par exemple en généralisant certains résultats obtenus par l'intégration de données omiques d'une espèce modèle comme *Arabidopsis thaliana* à d'autres espèces d'intérêt agronomique au génome plus complexe, comme le blé.

3.2 Les grands concepts de l'intégration

3.2.1 Différentes approches pour catégoriser les méthodes intégratives

Devant la multitude des questions biologiques (que nous avons catégorisées en (1) description, (2) sélection et (3) prédiction dans le Chapitre 1) et de données omiques associées (allant du génome jusqu'au phénotype), des méthodes et outils adaptés ont été développés pour intégrer des données par différentes approches.

Une des premières manières de catégoriser ces méthodes a été proposée dans Ritchie et al. (2015a), les séparant en intégration basée sur la concaténation, la transformation ou la modélisation des données, qui a inspiré les travaux de Zitnik et al. (2019) pour catégoriser les méthodes en *early*, *intermediate* ou *late integration*. L'objectif est alors d'identifier si l'intégration se fait directement sur les données avec l'analyse s'effectuant sur le résultat de la concaténation, si l'analyse s'effectue sur chaque jeu de données séparément puis ces résultats d'analyses sont intégrés, ou enfin si l'algorithme fait un entre-deux. Ces termes ont depuis été réutilisés dans Picard et al. (2021). La plus grande limite de cette approche reste néanmoins la difficulté à classer concrètement une nouvelle méthode dans l'une de ces catégories.

Une deuxième catégorisation possible est, lorsque les jeux de données sont sous forme matricielle, de parler d'intégration horizontale (N-intégration) ou verticale (P-intégration) pour indiquer que l'intégration est réalisée respectivement par lignes (mêmes lignes mais nombre de colonnes potentiellement différent entres les jeux de données) ou par colonnes (de manière analogue). La Figure 3.1 illustre ces différents types d'intégration.

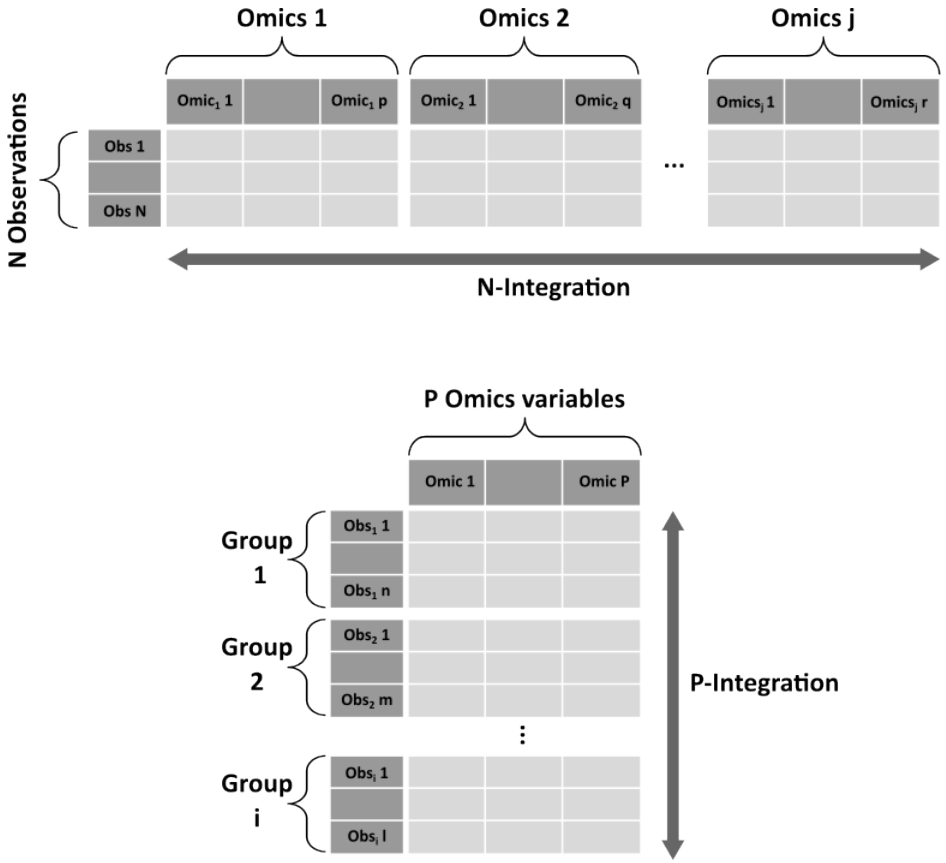


FIGURE 3.1 Illustration de l'intégration horizontale (N-intégration) et verticale (P-intégration)
Illustration des N- et P- intégrations sur des données matricielles formées d'observations en lignes et de variables omiques en colonnes. En N-intégration, plusieurs blocs correspondent à différents types de données omiques, tous avec les mêmes N observations. En P-intégration, plusieurs blocs correspondent à différents groupes d'observations, tous avec les mêmes P variables omiques.

Enfin, comme proposé par exemple dans Subramanian et al. (2020) et Athieniti and Spyrou (2023), il est possible de distinguer les méthodes intégratives en différentes familles de méthodes mathématiques ; chaque famille est caractérisée par une approche / un concept répondant à un des enjeux de l'intégration multi-omiques, notamment le fait d'intégrer des données hétérogènes et/ou de grandes dimensions. De plus, le type de questionnements biologiques auquel ces méthodes répondent peut aussi être pris en compte. Nous proposons dans la suite de cette section du manuscrit une classification des principaux outils d'intégration des données omiques en fonction des méthodologies associées et des questionnements scientifiques abordés, afin que le biologiste non mathématicien puisse sélectionner le meilleur outil ou approche en fonction de ses besoins.

3.2.2 Principales familles de méthodes

Nous proposons de classer les méthodes intégratives en méthodes de réduction de dimension, statistiques bayésiennes, basées sur la similarité, de réseaux et de réseaux de neurones artificiels. La Section 3.3 présente certains outils R d'intégration de données appartenant à une ou plusieurs de ces familles de méthodes.

— Les **méthodes de réduction de dimension** répondent à l'une des difficultés majeures lors du traitement des données omiques : gérer la grande dimension des données, plus particulièrement des données de plusieurs milliers voire millions de variables produites pour seulement quelques observations. La réduction de dimension conserve la majeure partie de l'information dans un nombre restreint de variables. Pour cela, il est possible de sélectionner les variables qui contiennent le plus d'information, ou créer des variables latentes contenant l'information extraite des variables initiales (méthodes de factorisation). Les deux approches peuvent aussi être combinées pour créer des variables latentes à partir de variables sélectionnées. Si le but est d'identifier les variables expliquant le plus la variabilité d'intérêt, une sélection de variables sera plus adaptée ; si la question est au contraire de résumer au mieux l'information en un nombre restreint de variables, le calcul de variables latentes sera plus adapté. Les concepts de réduction de dimension ainsi que certains des principaux algorithmes applicables sur un, deux ou plus de deux jeux de données omiques sont présentés par exemple dans Meng et al. (2016b) et Jia et al. (2022).
— Les **méthodes statistiques bayésiennes** supposent que les données sont issues de distributions probabilistes. Pour commencer, l'utilisateur pose une hypothèse sur les distributions *a priori* des données, puis ajuste ces distributions en observant / analysant ces données. Le principal inconvénient de ce type de méthodes est la difficulté à choisir la distribution *a priori*, un mauvais choix ayant un impact très important sur les distributions finales. Cependant, en définissant les distributions *a priori* à partir de l'expertise et des connaissances biologiques, les résultats deviennent aussi plus faciles à interpréter biologiquement que pour d'autres méthodes. Les méthodes bayésiennes sont particulièrement adaptées pour répondre à la problématique de l'hétérogénéité des données, puisqu'elles rendent possible l'analyse simultanée de données hétérogènes *via* le choix d'une distribution spécifique à chaque jeu de données. Les statistiques bayésiennes sont par exemple introduites dans Van De Schoot et al. (2021), et l'article Chu et al. (2022) présente différentes méthodes bayésiennes utilisées pour l'intégration mono- et multi-omiques dans le domaine de la cancérologie.
— Les **méthodes basées sur la similarité** entre variables ont pour principe de mesurer pour chaque couple de variables leur taux de ressemblance. La métrique de similarité peut alors être choisie pour prendre en compte l'hétérogénéité des données, afin par exemple de définir une valeur de similarité entre deux patients considérant simultanément différents types de jeux de données. Calculer la similarité entre les couples de variables n'est cependant pas suffisant en soi, et correspond seulement à une des étapes de la méthode. Ces méthodes sont alors généralement poursuivies d'une étape de clustering, et peuvent être très différentes les unes des autres selon leur manière d'exploiter les valeurs de similarité entre variables. Certains des outils basés sur la similarité sont

par exemple présentés dans Gliozzo et al. (2022) pour l'intégration de réseaux de similarité de patients, ou dans Wang and Kurgan (2019) pour la prédiction de l'impact de médicaments sur les protéines produites par les patients.

— Les **réseaux** sont des graphes composés de nœuds et d'arêtes les liant. Dans le cadre des données omiques, les nœuds représentent les individus, les gènes, les protéines, *etc.* et les arêtes quantifient les liens entre ces nœuds. Ces réseaux peuvent ainsi intégrer des données hétérogènes, en considérant par exemple simultanément chaque gène d'un jeu de données transcriptomiques et chaque protéine d'un jeu de données protéomiques comme un nœud. Les méthodes intégratives utilisant des réseaux sont souvent aussi des méthodes bayésiennes ou basées sur de la similarité (Hawe et al., 2019; Agamah et al., 2022).

— Les **réseaux de neurones artificiels**, très utilisés dans le *Deep Learning*, commencent aussi à émerger dans les analyses omiques (Aggarwal, 2018; Kang et al., 2022). Les modèles de réseaux de neurones convolutifs (ou *Convolutional Neural Networks* (CNN)) sont performants sur les analyses d'images par exemple dans le domaine médical. L'Auto-Encodeur (AE) généralise une partie des méthodes de réduction de dimension à des cas non linéaires. Ces méthodes sont néanmoins encore peu utilisées dans l'intégration multi-omiques en comparaison à d'autres domaines scientifiques, puisqu'elles sont adaptées à des données comportant à la fois énormément de variables et d'observations, ce qui est encore rarement le cas pour les données omiques produites.

Il est possible de séparer les méthodes en d'autres familles. Par exemple, Athieniti and Spyrou (2023) ajoute les méthodes basées sur les noyaux (*kernel*) et d'apprentissage profond (*deep learning*), et Subramanian et al. (2020) les méthodes multi-variées, de fusion et de corrélation.

3.2.3 L'intégration omique à différents "niveaux"

Toutes les tentatives précédentes pour catégoriser les différentes méthodes intégratives restent néanmoins limitées et ne représentent que partiellement les multiples possibilités d'intégration existantes et à venir. Notamment, avec la production croissante de données, l'enjeu n'est plus seulement d'analyser des matrices comportant toujours plus de lignes et colonnes, mais bien d'analyser une multitude de données apportant de nouveaux "niveaux" d'informations biologiques, en utilisant par exemple des données issues de plusieurs espèces, populations, individus, tissus, *etc.* pour différents types de données omiques, temporalités, conditions expérimentales, *etc.* simultanément.

Cette problématique a notamment été illustrée dans la Figure 3.2 de Rajasundaram and Selbig (2016). En Figure 3.2.a, le cas classique d'intégration est représenté, à savoir une intégration selon 3 "niveaux" (lignes, colonnes, jeux de données). En Figure 3.2.b, 3 "niveaux" d'intégration sont encore présentés, mais l'intégration est plus complexe en considérant non plus 2 mais k jeux de données. En Figure 3.2.c, un quatrième "niveau" d'intégration est ajouté,

avec maintenant deux groupes de jeux de données. Cette intégration à 4 "niveaux" peut par exemple correspondre à des données produites pour n individus, à k temporalités, pour p gènes (expression des gènes) et q protéines (abondance des protéines). Il est alors envisageable, avec les progrès en terme de production de données, de souhaiter intégrer des données avec encore plus de "niveaux" d'intégration (difficilement représentables sous forme matricielle), soulevant ainsi des défis méthodologiques et statistiques majeurs dans les années à venir.

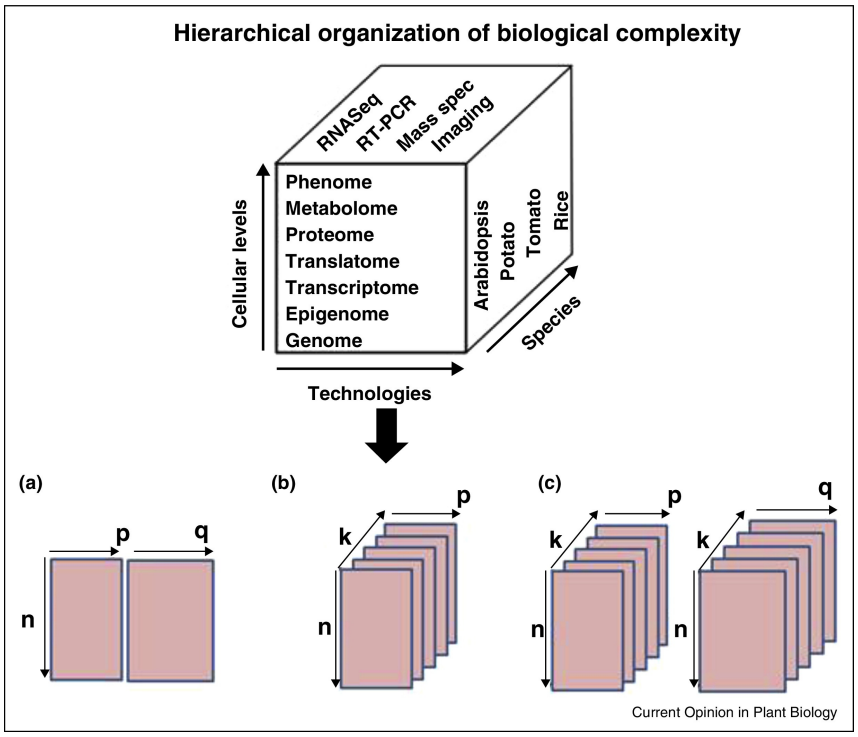


FIGURE 3.2 Intégration omique à différents "niveaux"
Figure et légende de Rajasundaram and Selbig (2016). *The complexity of biological data is multi-dimensional owing to advances in high-throughput technologies. Heterogeneity of the generated data is attributed to the measurement of different cellular levels using a wide range of techniques across various plant species. A schematic representation of possible types of data analysis problems is depicted. (a) Depicts the most common form of integrative analysis problems wherein two data sets with n observations and variables p and q are analyzed. (b) Represents the case where there are k data sets with the same n observations and the same number of variables p. This is an example of a typical descriptive type of integrative analysis. (c) Illustrates the situation where there are two sets of k data sets where k can pertain to time points or experimental conditions on n observations with variables p or q.*

3.3 Principaux outils R

L'abondante quantité d'outils développés pour intégrer les données omiques rend très difficile le choix de l'un d'eux adapté à la question biologique et aux données produites. Il existe ainsi des outils spécifiques à une question biologique précise quand d'autres sont plus génériques, des outils ayant de fortes contraintes sur les données quand d'autres tolèrent une grande diversité de types de données, certains outils en "clique-bouton" simplifient l'utilisation pour les novices mais limitent les possibilités de paramétrage pour les initiés, des outils sont en lignes, d'autres dans différents langages de programmation, *etc.* De nombreuses publications

ont donc été proposées ces dernières années pour présenter et comparer certains outils et/ou méthodes d'intégration multi-omiques.

En nous restreignant, par choix, aux outils développés dans le langage de programmation R (largement utilisé par les biologistes) et cités dans un des articles de comparaison sélectionnés et parus avant 2023, à savoir Bersanelli et al. (2016); Cai et al. (2022); Hesami et al. (2022); Lovino et al. (2021); Rappoport and Shamir (2018); Sathyanarayanan et al. (2020); Subramanian et al. (2020); Vahabi and Michailidis (2022), nous avons répertorié 39 outils R d'intégration multi-omiques (Annexe A). Il en existe cependant beaucoup d'autres, dont une grande partie se retrouvent listés notamment dans le github de Mike Love (`https://github.com/mikelove/awesome-multi-omics`).

Parmi ces 39 outils, nous nous sommes intéressés plus particulièrement à 13 d'entre eux en suivant les critères suivants :

— nous avons pour commencer sélectionné les outils cités dans plus de deux des huit articles de comparaison, en nous restreignant néanmoins uniquement aux outils d'intégration d'une grande variété de données omiques. Les outils identifiés de cette manière sont *BCC* (Lock and Dunson, 2013), *iCluster* et ses versions *iClusterPlus* et *iClusterBayes* (Shen et al., 2009; Mo et al., 2013, 2018), *JIVE* (Lock et al., 2013), *LRAcluster* (Wu et al., 2015), *MCIA* (et *MCOA*) (Bady et al., 2004), *mixOmics* (Rohart et al., 2017), *moCluster* (Meng et al., 2016a), *MOFA* (Argelaguet et al., 2018, 2020), *NEMO* (Rappoport and Shamir, 2019), *PINS* (et *PINSPlus*) (Nguyen et al., 2017, 2018) et *SNF* (Wang et al., 2014).

— nous avons ensuite ajouté à cette liste les outils *mixKernel* (Mariette and Villa-Vialaneix, 2018) et *RGCCA* (Tenenhaus and Tenenhaus, 2011), cités dans un seul de ces articles mais étant fréquemment utilisés pour l'intégration de données *via* la réduction de dimension.

Ainsi, le Tableau 3.1 détaille les caractéristiques principales de ces 13 outils finalement sélectionnés, dans le but d'aider le biologiste à choisir rapidement un outil d'intégration approprié parmi certains des outils les plus couramment cités. Ce tableau possède toutefois plusieurs limites, puisqu'il apparaît que la majorité des outils sélectionnés sont issus de méthodes non-supervisées utilisées pour décrire les relations entre observations et/ou variables omiques, mais pas pour sélectionner des observations/variables particulières ni pour prédire des phénotypes à partir de données omiques. De plus, ce tableau présente principalement des méthodes de réduction de dimension ou statistiques au détriment des autres approches. De plus, ces outils d'intégration ont été choisis en partie pour leur généricité afin d'être utilisables sur de nombreux types de données omiques, ce qui peut dans certains cas les rendre moins performants que des outils spécifiquement dédiés à des données particulières et qui considèrent les particularités de ces données.

Tool's name	Biological question	Method and tool's characteristics				Data characteristics	
	Main questions' types	Supervised/unsupervised	Methods families	Summary	Updated	Omics	Hypothesis
BCC (Bayesian Consensus Clustering)	I) Description of samples' interactions	Unsupervised	Bayesian statistics	Computes a samples' clustering for each omics dataset by using a probabilistic model, then merges clusters to get a consensus cluster across omics datasets	No	Multi-omics (quantitative)	Normal distribution Different omics on the same set of samples
iCluster (iClusterPlus / iClusterBayes)	I) Description of samples' interactions	Unsupervised	Bayesian statistics / Dimension reduction	Starts with a latent variables regression across datasets by using a probabilistic model, then uses these joint latent variables for samples' clustering	Yes	Multi-omics (quantitative and qualitative)	Linearity assumption Normal noise distribution Different omics on the same set of samples
JIVE (Joint and Individual Variation Explained)	I) Description of samples/variables' interactions	Unsupervised	Dimension reduction	Decomposes each dataset in three terms: a joint effect (across datasets), an individual effect (specific to the dataset) and a noise effect	No	Multi-omics (quantitative)	Linearity assumption
LRAcluster (Low-Rank Approximation Cluster)	I) Description of samples' interactions	Unsupervised	Dimension reduction	Probabilistically computes a common low-dimensional subspace across omics, then uses the K-means algorithm to cluster samples on this subspace	Yes	Multi-omics (quantitative and qualitative)	Linearity assumption Different omics on the same set of samples
MCIA (Multiple co-inertia analysis) (MCOA)	I) Description of samples/variables' interactions	Unsupervised	Dimension reduction	Projects each dataset on a subspace, then maximizes co-inertia between subspaces to get major information shared by datasets	Yes	Multi-omics (quantitative)	Linearity assumption Different omics on the same set of samples
mixKernel	I) Description of samples/variables' interactions II) Variables selection III) Phenotype prediction	Supervised Unsupervised	Dimension reduction	Transforms datasets with kernels, then applies usual dimension reduction methods	Yes	Multi-omics (quantitative and qualitative)	Datasets with the same rows or columns
mixOmics (with PCA, PLS, rCCA, Diablo...)	I) Description of samples/variables' interactions II) Variables selection III) Phenotype prediction	Supervised Unsupervised	Dimension reduction	Contains many matrix factorization methods for multivariate analysis and functions for data visualization. The main analysis method for one single dataset is the PCA. For two datasets or more, the main methods are the PLS and rCCA, and their extentions for discriminant analysis, variable selection ('sparse') and multi-blocks analysis.	Yes	Multi-omics (quantitative and qualitative)	Linearity assumption Datasets with the same rows or columns
moCluster (from MOGSA)	I) Description of samples' interactions	Unsupervised	Dimension reduction	Computes latent variables by using a PCA's extension, then clusters them and finally select the best subtype model	Yes	Multi-omics (quantitative)	Linearity assumption Different omics on the same set of samples
MOFA (Multi-Omics Factor Analysis) (MOFA2)	I) Description of samples' interactions III) Phenotype prediction	Unsupervised	Bayesian statistics / Dimension reduction	Factorizes datasets with a Bayesian approach to get a small number of latent factors usable for different purposes.	Yes	Multi-omics (quantitative and qualitative)	Linearity assumpion
NEMO (NEighborhood based Multi-Omics clustering)	I) Description of samples' interactions	Unsupervised	Similarity-based	Creates one similarity matrix by dataset, then merges them and finally clusters the merged matrix by Spectral clustering	No	Multi-omics (quantitative)	Euclidean distance metric
PINS (Perturbation clustering for data INtegration and disease Subtyping) (PINSPlus)	I) Description of samples' interactions	Unsupervised	Similarity-based / Network	Does several clustering to identify how often samples are clustered together. Clusterings are made on different datasets, with data perturbed by adding gaussian noise, and different clustering methods are used	Yes	Multi-omics (quantitative)	Different omics on the same set of samples
RGCCA (Regularized Generalized Canonical Correlation Analysis) (sGCCA)	I) Description of samples/variables' interactions II) Variables selection III) Phenotype prediction	Supervised Unsupervised	Dimension reduction	Computes latent variables for each dataset by maximizing correlations within and/or between datasets	Yes	Multi-omics (quantitative and qualitative)	Linearity assumption Different omics on the same set of samples
SNF (Similarity Network Fusion)	I) Description of samples' interactions	Unsupervised	Similarity-based / Network	Creates a similarity matrix then an associated network for each dataset, then iteratively fuses the networks to keep only strong correlations between samples across omics	No	Multi-omics (quantitative and qualitative)	Different omics on the same set of samples Euclidean distance metric

TABLE 3.1 Tableau des 13 outils considérés

Description des 13 outils d'intégrations multi-omiques avec leurs principales caractéristiques sur la/les questions biologiques, la méthode intégrative et les données à intégrer. Premièrement, la tableau indique à laquelle ou lesquelles des stratégies d'analyse intégrative chaque outil répond, que ce soit en terme de description, sélection ou prédiction. Deuxièmement, il est aussi indiqué si la méthode est supervisée ou non, et à quel(s) type(s) de familles de méthodes elle est associée, à savoir aux statistiques bayésiennes, méthodes de réduction de dimension, méthodes basées sur la similarité, de réseaux ou de réseaux de neurones artificiels. Une description succincte de la méthode est aussi présentée ainsi qu'une indication sur les éventuelles mises à jour de l'outil. Enfin, la tableau indique le type de données prises en compte par l'outil et ses principales hypothèses sur les données.

3.4 Le choix de mixOmics

Parmi les outils d'intégration multi-omiques listés dans le Tableau 3.1, l'un des plus populaires est le paquet R *mixOmics*. Grâce à son interface unique permettant d'exploiter de nombreuses méthodes de réduction de dimension ainsi que de nombreuses fonctions de visualisation des données, *mixOmics* est adaptable à différents jeux de données, mais aussi aux différentes stratégies d'intégration proposées dans la thèse (description, sélection, prédiction). Ceci le démarque de la plupart des outils intégratifs, et notamment ceux de la Table 3.1. Les outils *mixOmics*, *SNF*, *PINSPlus* et *iClusterBayes* ont par exemple été testés et comparés sur des données céréales durant mon stage de master précédant cette thèse, révélant la possibilité pour *mixOmics* de s'adapter aux différentes stratégies d'intégration des données tandis que les trois autres outils sont spécifiques au clustering des données. Les deux autres outils de la Table 3.1 abordant les trois stratégies intégratives sont alors *mixKernel*, inspiré de *mixOmics* et ajoutant une transformation noyau (discuté au Chapitre 7), et *RGCCA*, un paquet sur lequel se base une partie des algorithmes de *mixOmics* mais qui peut être plus complexe à prendre en main. Le paquet *mixOmics* a aussi l'avantage d'être particulièrement bien documenté avec son site (`http://mixomics.org`) reprenant ses principales fonctions et des études de cas sur différents jeux de données, ainsi que le livre Lê Cao and Welham (2021) qui détaille encore plus les fonctions et méthodes associées ainsi que leur utilisation dans différents exemples concrets. Enfin, *mixOmics* est aussi régulièrement mis à jour et a une communauté très active communiquant sur le forum dédié `https://mixomics-users.discourse.group`. Pour toutes ces raisons, nous avons décidé de nous concentrer particulièrement sur l'outil *mixOmics* durant cette thèse.

3.4.1 Le fonctionnement général de mixOmics

La Figure 3.3 illustre la structure de *mixOmics*, composée principalement de fonctions d'analyse de données et fonctions de visualisation des résultats.

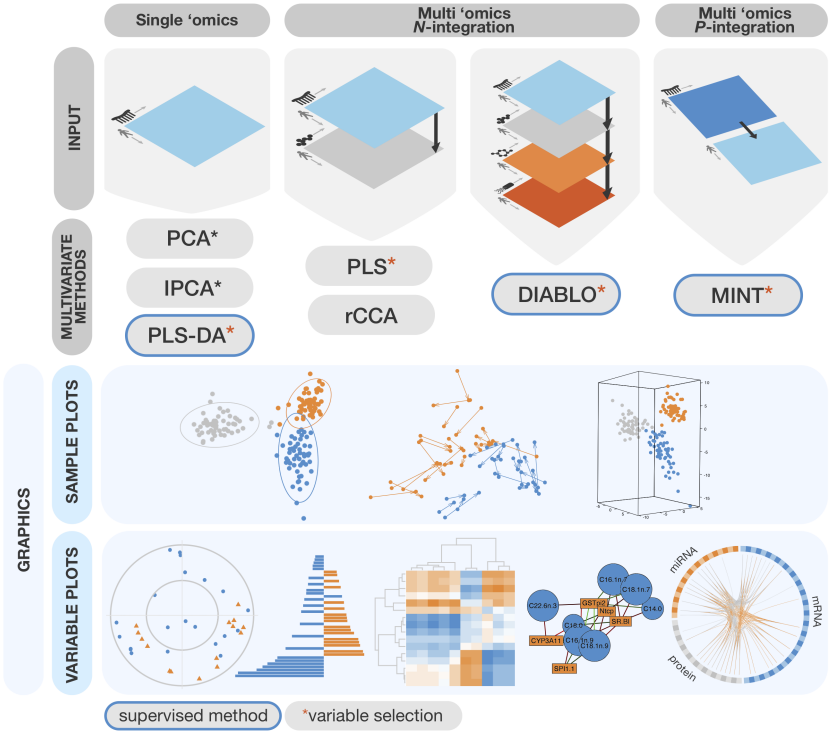


FIGURE 3.3 Structure de `mixOmics`
Figure issue de `http://mixomics.org/` synthétisant les principales fonctionnalités proposées dans `mixOmics` en terme d'analyse de données (partie supérieure) et de graphiques de visualisation des résultats obtenus (partie inférieure). Plusieurs méthodes sont alors disponibles pour l'analyse d'un, deux ou plus de deux jeux de données omiques, l'intégration s'effectuant sous forme de N- ou de P-intégration. Plusieurs fonctions graphiques sont disponibles pour afficher les lignes (appelées ici *samples*) et colonnes (*variables*) des jeux de données.

3.4.2 Les principales méthodes intégratives de mixOmics

Bien que `mixOmics` propose un assez grand nombre de méthodes intégratives, beaucoup ne sont finalement que des adaptations dans un contexte spécifique de la régression PLS. Ici sont résumées les principales méthodes de `mixOmics` et les ajouts des variantes.

— **ACP** : l'ACP (Analyse en Composantes Principales, ou *PCA*) a pour principe de créer des composantes par combinaisons linéaires des variables initiales d'un jeu de données, de manière à conserver la majeure partie de la variabilité de ces variables. Dans le cadre de l'ACP, la variabilité en question est la variance des données. Les composantes sont calculées itérativement : la première composante est calculée en maximisant sa variance, la deuxième est calculée avec comme contrainte supplémentaire d'être orthogonale à la première de sorte à ce que la variabilité de la deuxième composante ne soit pas déjà comprise dans la première. Les composantes suivantes sont elles aussi calculées de sorte à être orthogonales à toutes les composantes précédentes tout en maximisant leur variance. La Figure 3.4.A illustre, pour un jeu de données initial de n observations et p variables, sa réduction en une matrice formée de q composantes, avec q inférieur à p.

— **PLS** : la régression PLS (*Projection to Latent Structure* ou *Partial Least Squares regression*) est une méthode combinant les concepts de calcul de variables latentes et de régression linéaire. Ainsi, tout comme l'ACP, la régression PLS calcule des variables latentes (analogues aux composantes de l'ACP, et nommées ainsi dans la suite du manuscrit) par combinaison linéaire des variables dites "explicatives" (ou X) et, s'il existe plusieurs variables dites "expliquées" (ou Y), une autre combinaison linéaire est calculée sur ces variables. On parle alors de PLS1 s'il n'y a qu'une variable Y, de PLS2 s'il y en a plusieurs. Contrairement à l'ACP, la PLS ne maximise pas la variance des composantes, mais leur covariance avec la variable expliquée Y (PLS1) ou entre les deux matrices de composantes (PLS2). La Figure 3.4.B illustre la PLS2, avec d'une part les réductions du jeu de données explicatives X en composantes T et du jeu de données expliquées Y en composantes U, et d'autre part la régression des matrices de composantes explicatives T vers les expliquées U.

— **Variantes (parcimonieuse, discriminante, par blocs, canonique)** : la plupart des modèles de `mixOmics` existent sous plusieurs variantes. Les méthodes parcimonieuses (*sparse*), comme par exemple la fonction *sparse PLS* (sPLS), ajoutent un terme de "pénalité" (pénalité Lasso) lors du calcul des composantes afin de ne sélectionner qu'un nombre restreint de variables par composante. L'analyse discriminante (*DA*) avec par exemple PLS-DA est une version supervisée pour laquelle le but est de séparer au mieux les différentes classes de la variable qualitative Y par les variables quantitatives de X. Les analyses par blocs (*blocks*), comme pour la *block.pls* (MB-PLS), intègrent plusieurs jeux de données par lignes, alors que l'intégration par colonnes s'effectue avec les méthodes *mint*. La Figure 3.4.C illustre la MB-PLS pour trois blocs X afin d'expliquer le jeu de données Y. Dans le cas des régressions, le but étant d'expliquer/prédire Y à partir de X, la relation entre X et Y est asymétrique, mais il est aussi possible d'utiliser le mode *canonical* pour mener une régression inspirée de la corrélation canonique et obtenir une relation symétrique entre X et Y.

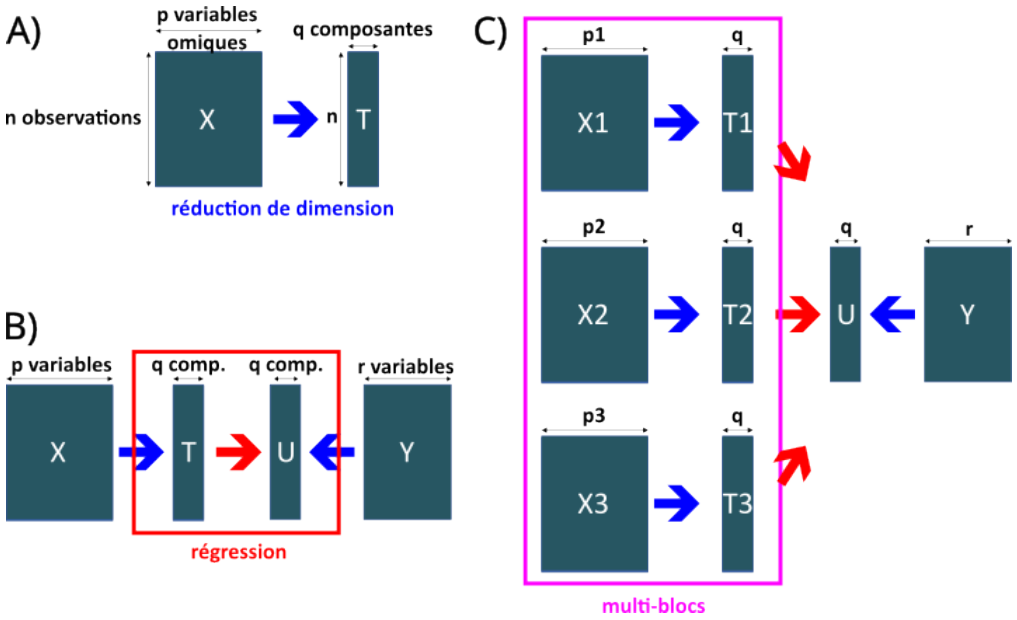


FIGURE 3.4 Illustration du fonctionnement des ACP, PLS et MB-PLS

Représentation schématique de l'ACP et des régressions PLS et MB-PLS sur des données sous forme matricielle comportant toutes le même nombre n de lignes (observations). **A-** Représentation de l'ACP : une réduction est effectuée sur un jeu de données X comportant p colonnes (variables omiques) afin d'obtenir une nouvelle matrice T comportant un nombre inférieur q de colonnes (composantes, calculées à partir des variables omiques initiales), **B-** Représentation de la régression PLS pour deux variables quantitatives (aussi appelée PLS2) : pour chaque jeu de données (X et Y), une nouvelle matrice (T et U) de q composantes est calculée, avec q inférieur aux nombres de variables p et r respectives à X et Y. Ces calculs sont effectués simultanément et dans le but de maximiser la covariance entre les composantes de T et U, correspondant à la phase de régression du jeu de données explicatives X vers le jeu de données expliquées Y, **C-** Généralisation de la régression PLS à plusieurs blocs de données (MB-PLS) : le principe est similaire à la régression PLS mais pour plusieurs blocs de données explicatives, ici trois blocs $X1$, $X2$ et $X3$. En plus de la régression de chacun de ces blocs vers le jeu de données Y, les interactions entre les composantes de $T1$, $T2$ et $T3$ sont aussi partiellement considérées selon les paramètres de la régression MB-PLS.

3.4.3 Notes sur la *block.pls*

Dans la Figure 3.3, la fonction multi-blocs mentionnée pour l'intégration de données selon les lignes (aussi appelée N-intégration) est *DIABLO*, qui correspond à la *block.splsda*, c'est-à-dire la régression PLS en version discriminante, parcimonieuse et multi-blocs. Toutefois, dans le cadre de cette thèse, la méthode principalement employée est la *block.pls* et éventuellement la *block.spls*, qui ne font pas apparaître de variable qualitative à discriminer.

De plus, la figure laisse supposer que l'ensemble des fonctions de visualisation sont applicables à l'ensemble des méthodes intégratives, ce qui n'est pas le cas. Notamment, alors que la plupart des fonctions graphiques peuvent être utilisées sur le résultat de la *block.splsda*, très peu de ces fonctions le sont sur le résultat de la *block.pls*. Ces limites seront approfondies dans les chapitres suivants.

3.5 Points clés

— L'intégration multi-omiques consiste à identifier les variations omiques caractérisant les individus, populations ou espèces, pour 1) une meilleure compréhension des mécanismes clés de leur développement, de leur adaptation ou de leur tolérance à des contraintes biotiques ou abiotiques en interaction avec leur écosystème, et 2) la prédiction de traits inconnus pour certains individus / populations / espèces à partir de leurs données omiques et des connaissances acquises sur des individus / populations / espèces proches biologiquement.

Il existe de nombreux outils et méthodes pour l'intégration de données omiques. Ici, nous décidons de les classer en méthodes de réduction de dimension, bayésiennes, basées sur la similarité, de réseaux et de réseaux de neurones artificiels.

— Nous proposons un tableau d'outils R d'intégration multi-omiques fréquemment cités et de leurs principales caractéristiques afin que les biologistes puissent choisir parmi certains des outils intégratifs les plus populaires le plus adapté à leurs questions et données.

L'un de ces outils, nommé `mixOmics`, est celui principalement utilisé dans le cadre de ce travail de thèse, notamment avec les méthodes d'ACP, de régression PLS et de régression MB-PLS analysant par réduction de dimension respectivement un, deux ou de multiples jeux de données omiques.

Chapitre 4

Objectif et stratégie de la thèse

4.1 L'objectif

Durant cette thèse, l'objectif a été de mettre en oeuvre des méthodologies d'intégration de données omiques adaptables à différents cas d'études. Trois principaux aspects de l'intégration multi-omiques ont alors été étudiés pour atteindre cet objectif (Figure 4.1) :

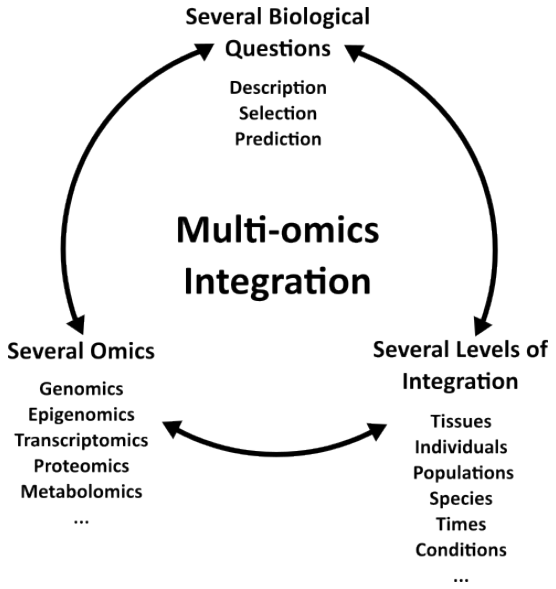


FIGURE 4.1 Représentation de la multitude de contextes pour l'intégration omique
3 aspects majeurs représentant la diversité des intégrations omiques : la question biologique, le nombre et types de données omiques, les "niveaux" d'intégration

— **La question biologique :** l'intégration de données est réalisée dans le but de répondre à une question biologique choisie. Dans cette thèse, nous catégorisons ces différentes questions en trois familles selon la stratégie choisie en terme d'intégration des données omiques pour y répondre, à savoir (1) la **description** des interactions entre les différents types de données omiques et variables biologiques associées, pour étudier par exemple l'impact de la méthylation sur l'expression génique ou l'impact de traits phénotypiques sur l'abondance des protéines, ou encore la description et catégorisation des observations (des individus ou des gènes) selon leur profil omique, par exemple l'identification

des individus/gènes au profil transcriptomique ou protéomique semblable, (2) la **sélection** de variables omiques associées à des traits phénotypiques particuliers, pour étudier par exemple les protéines à fort impact sur un trait phénotypique donné, ainsi que la sélection d'observations (des individus ou des gènes) aux profils omiques particuliers, par exemple des gènes fortement méthylés et faiblement exprimés, et (3) la **prédiction** de caractères phénotypiques inconnus à partir des connaissances des interactions omiques-phénotypiques acquises sur d'autres observations, pour prédire par exemple des traits phénotypiques à partir de données protéomiques produites sur un animal et des relations entre protéines et phénotypes préalablement déterminées sur d'autres animaux. La Figure 4.2 illustre ces différentes stratégies intégratives.

— **Les types de données omiques :** de nombreux types de données omiques sont produits, notamment des données génomiques, épigénomiques, transcriptomiques, protéomiques, métabolomiques, *etc.* Les méthodologies développées dans cette thèse doivent alors être utilisables sur ces données diverses pour s'adapter au maximum de contextes intégratifs.

— **Les "niveaux" d'intégration :** les données à intégrer peuvent avoir été produites par exemple pour différents individus ou différentes populations selon la zone géographique, différentes espèces, dans différentes conditions de développement (par exemple entre conditions de "contrôle" *vs.* "stress"), *etc.* qui sont autant de "niveaux" d'intégration (*c.f.* Section 3.2.3). Un enjeu est donc de pouvoir considérer tous ces "niveaux" lors de l'intégration des données omiques.

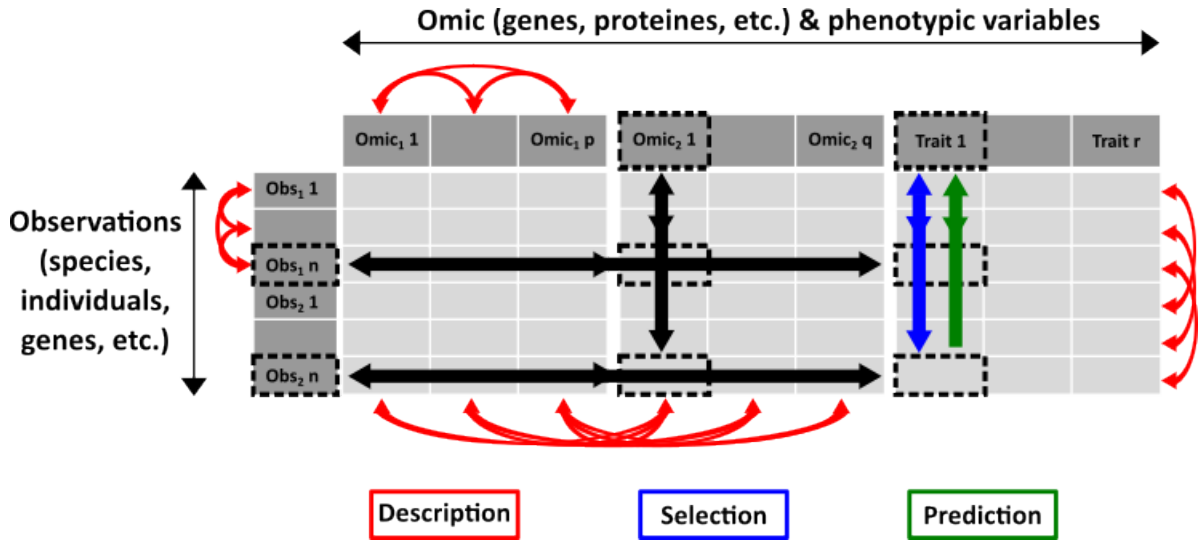


FIGURE 4.2 Résumé des trois grands types de questionnements biologiques sur un même jeu de données type

Les données sont représentées ici sous forme matricielle. Deux blocs matriciels correspondent aux données de deux types de données omiques (ex : génomiques, transcriptomiques, protéomiques, *etc.*), le troisième bloc représente des traits phénotypiques ou autres données biologiques non-omiques. Chaque colonne représente une variable (les données omiques qui, dans notre contexte, sont l'expression et la méthylation pour les analyses "plantes" et les protéines et phénotypes pour les analyses "animaux") et chaque ligne une observation (dans notre contexte, des gènes pour les analyses "plantes" et des individus pour les analyses "animaux". Généralement, les observations sont des individus et les variables omiques sont alors des gènes, des protéines, *etc.* Il est aussi possible que les gènes, protéines, *etc.* soient considérés comme les observations, et les variables omiques peuvent alors être des données omiques produites à différents instants, dans différentes conditions, *etc.*. Ici, deux groupes d'observations sont représentés et peuvent correspondre par exemple aux données produites pour les mêmes individus avant puis après traitement. Trois types de questions biologiques sont illustrés sur ces données. La description des relations peut s'effectuer entre variables d'un même bloc et/ou d'autres blocs, ainsi qu'entre observations d'un même groupe et/ou observations appariées dans différents groupes. La sélection peut s'effectuer sur les observations (selon leur profil atypique pour certaines variables), ou sur les variables (en fortes interactions entre elles pour certains individus). La prédiction peut s'effectuer sur des traits inconnus à partir de données omiques et des connaissances sur les relations omiques-traits provenant d'autres observations.

4.2 La stratégie

Pour répondre à cet objectif, nous avons proposé un tutoriel préliminaire à l'intégration de données omiques, et exploité le paquet R `mixOmics` à partir duquel nous avons développé un nouvel outil nommé cimDiablo_v2. Nous les avons ensuite testés sur différents jeux de données, produits en collaboration avec différentes équipes et partenaires, et choisis pour leur complémentarité d'expertise afin de se confronter à une grande variété de contextes biologiques et de discuter de l'efficacité d'adaptation des développements méthodologiques. De plus, les résultats biologiques obtenus ont pu être valorisés en prouvant leur complémentarité avec les résultats issus des analyses mono-omiques. Ces jeux de données, résumés dans la Table 4.1, sont les suivants :

— Données de méthylation et d'expression des gènes de peuplier produites pour différents individus de 10 populations européennes. Les objectifs biologiques sont alors de décrire les interactions entre la méthylation et l'expression des gènes à l'échelle du gé-

nome puis de sélectionner les gènes aux valeurs extrêmes pour plusieurs des variables omiques acquises, afin notamment d'identifier le rôle de la co-régulation des gènes par son expression et sa méthylation pour différentes populations de différentes origines géographiques. **Question :** ***L'intégration des données omiques permet-elle d'identifier les gènes dont la régulation par méthylation/expression est spécifique de certaines populations d'une espèce pérenne (peuplier) ou conservée entre ces populations d'origines géographiques diverses associées à différentes contraintes environnementales ?***

— Données de méthylation et d'expression des gènes de céréales produites pour différentes espèces (maïs, *Brachypodium*), pour différents stades de développement du grain. Les objectifs sont les mêmes que pour l'analyse précédente, mais avec l'objectif supplémentaire de comparer les résultats entre différentes espèces afin d'identifier les similitudes et spécificités dans les interactions omiques expression-méthylation. **Question** : ***L'intégration des données omiques permet-elle d'identifier les spécificités et similitudes de la régulation des gènes par méthylation/expression entre deux espèces (maïs et Brachypodium) issues d'un ancêtre commun datant de 65 millions d'années ?***

— Données d'abondance de protéines dans différents tissus (foie, muscle, tissu adipeux) relativement à des traits phénotypiques de la composition des carcasses chez des bovins. L'objectif est alors de sélectionner les protéines les plus explicatives de sept traits phénotypiques, afin de prédire à moyen terme sur de nouveaux individus ces traits à partir des abondances de protéines sélectionnées. **Question :** ***L'intégration des données omiques permet-elle d'identifier les signatures moléculaires de la composition tissulaire ou chimique des carcasses bovines ?***

Espèces	Individus (ou union d'individus)	Tissus (ou mélange de tissus)	Autres	Omiques	Types de questions biologiques	Projet
1 (peuplier)	10 (union des données de 2 individus pour chacune des 10 régions géographiques)	1 (mélange de cambium et xylème)	Contextes de méthylation	2 (méthylomiques, transcriptomiques)	Description Sélection	EPITREE
2 (maïs, *Brachypodium*)	1 (mélange de graines de plusieurs individus)	1 (mélange de graines)	Contextes de méthylation Stades de développement du grain	2 (méthylomiques, transcriptomiques)	Description Sélection	SeedENCODE
1 (bovin)	17 (séparés en 2 groupes alimentaires)	3 (foie, muscle, tissu adipeux)	Différence d'adiposité induite par le régime alimentaire	2 (protéomiques, phénotypes)	(Description) Sélection (Prédiction)	Cradecha

TABLE 4.1 Résumé des principales caractéristiques des données exploitées pour l'intégration multi-omiques dans le cadre de la thèse

Le tableau décrit les facteurs de variabilité des données, à savoir selon le nombre d'espèces, d'individus, de tissus (et organes) et d'autres spécificités sur les données, ainsi que les types de données omiques et phénotypiques, catégories de questions biologiques abordées et projet d'origine des différents jeux de données de la thèse.

Pour l'analyse portant sur les plantes, les travaux discutés dans les sections suivantes ont

été publiés dans les articles Sow et al. (2023) (***Epigenetic Variation in Tree Evolution : a case study in black poplar (Populus nigra)***) et Bellec et al. (2023) (***Tracing 100 million years of grass genome evolutionary plasticity***) dont je suis co-auteur, et l'article Mardoc et al. (2024) (***Genomic data integration tutorial, a plant case study***) pour lequel je suis premier auteur. Pour la partie portant sur les animaux, les travaux discutés dans les sections suivantes ne sont pas encore soumis pour publication mais sont présentés sous forme d'article à soumettre (Mardoc et al. (en préparation), ***Integrating liver, muscle and adipocyte tissue proteomes to identify drivers of body and growth composition in bovine species***).

Ma contribution à l'ensemble de ces publications est détaillée dans les Chapitres 5 et 6 présentant et discutant les résultats de l'application du tutoriel ainsi que de `mixOmics` et cimDiablo_v2 pour l'intégration multi-omiques chez les plantes (Chapitre 5) et les animaux (Chapitre 6).

Deuxième partie

Résultats

Chapitre 5

Interactions méthylation-expression chez les plantes

5.1 Introduction et contexte

L'intégration des données omiques produites par des technologies haut-débit est une manière relativement récente d'analyser les données en prenant en compte la complexité des interactions biologiques pour répondre à de multiples questions, que ce soit pour décrire les données, sélectionner des observations et/ou variables aux profils particuliers ou prédire des variables à partir des (interactions des) données omiques. Ces intégrations nécessitent alors des outils capables de répondre à une ou plusieurs de ces grandes questions pour différents jeux de données, autant en terme de type de données omiques (génomiques, transcriptomiques, protéomiques, *etc.*) que de "niveaux" d'intégration (par individus, espèces, populations, temporalités, conditions expérimentales *etc.*).

L'article Mardoc et al. (2024) présente de nouveaux développements méthodologiques pour répondre à ces besoins d'adaptation à différents contextes biologiques, à la fois par (1) un tutoriel présentant les étapes à suivre pour préparer l'intégration des données, et (2) l'exploitation du paquet R `mixOmics` et le développement de la fonction cimDiablo_v2 inspirée de ces méthodes de réduction de dimension.

Cet article propose aussi un cas concret d'utilisation de ces développements, de type "preuve de concept", sur des données omiques obtenues chez le peuplier dans le but d'étudier l'impact de la méthylation de l'ADN sur l'expression des gènes pour différentes populations européennes de peupliers (Figure 5.1), que ce soit en étudiant les interactions méthylation-expression à l'échelle du génome entier (description) ou en sélectionnant des groupes de gènes aux profils particuliers dans leurs interactions omiques pour étudier leurs fonctions et rôles biologiques (sélection). Les objectifs sont alors à la fois méthodologiques en testant le tutoriel et cimDiablo_v2 sur des données omiques complexes, et biologiques afin (1) d'affiner nos connaissances sur l'une des interactions multi-omiques clés (méthylation *vs.* expression) de la régulation génomique, et (2) de mieux comprendre l'évolution des populations de peupliers en décryptant les bases omiques de leur réponse adaptative à différents environnements géographiques et les variations environnementales associées. Les données peuplier et questions associées sont illustrées en Figure 5.2.

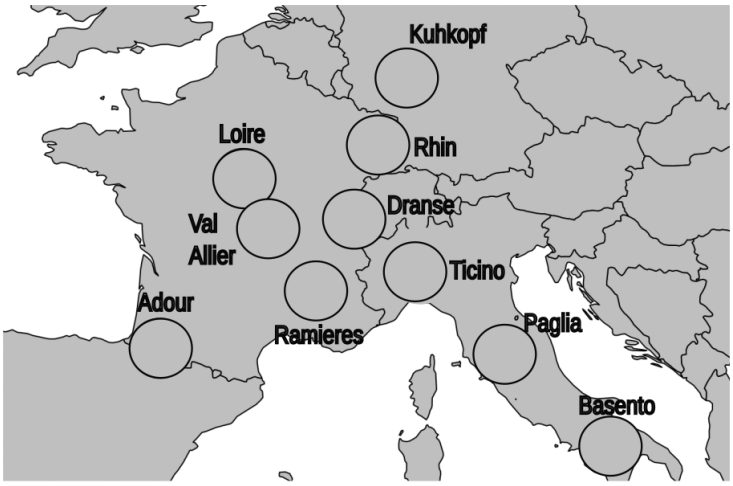


FIGURE 5.1 Illustration de l'origine géographique des 10 populations de peuplier européennes étudiées

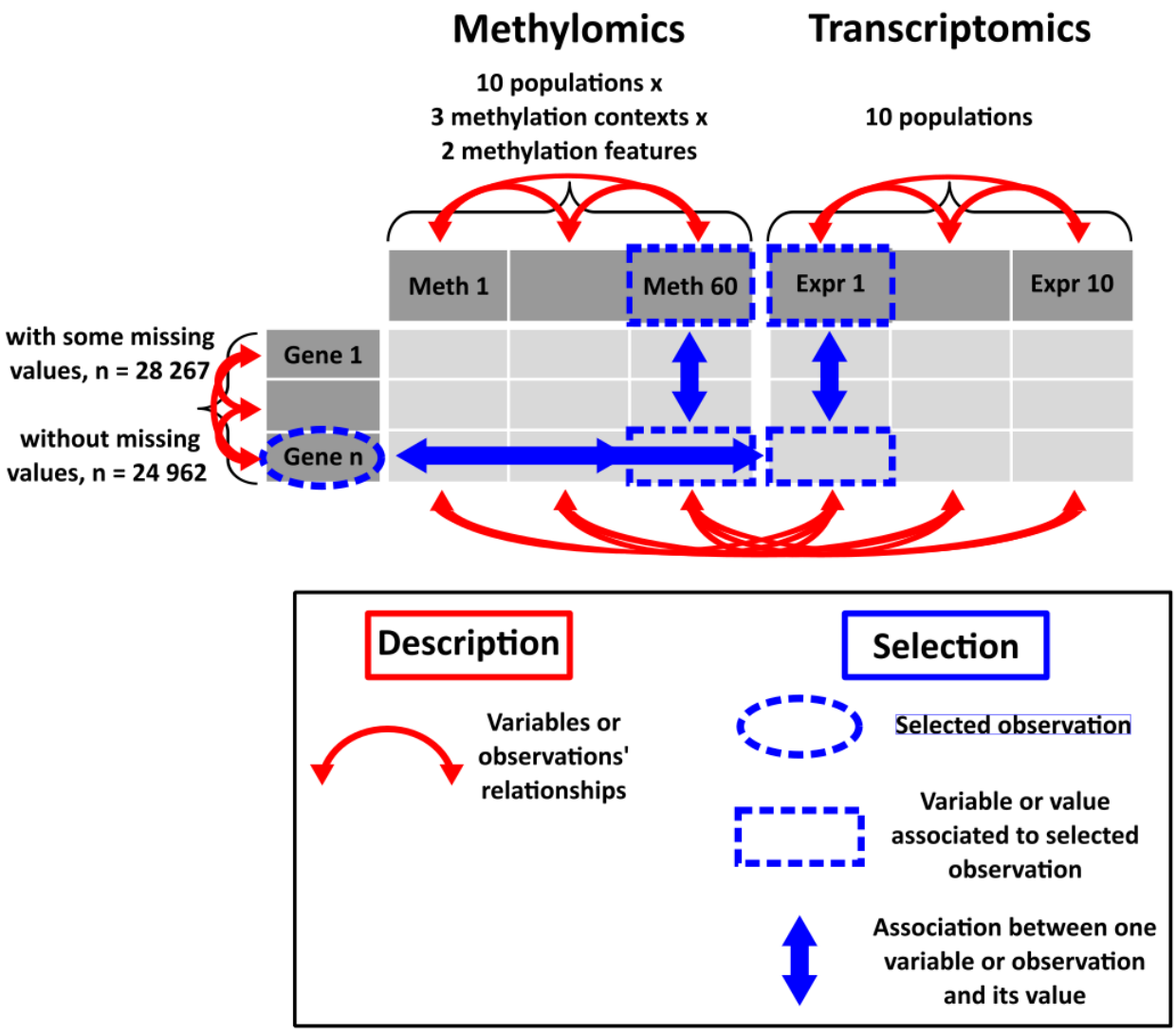


FIGURE 5.2 Représentation schématique des données peuplier utilisées et des questions d'intégration multi-omiques abordées dans le cadre de la thèse

Les données peuplier intégrées sont composées de deux blocs omiques comportant en colonnes des variables omiques et en lignes les mêmes gènes. Le bloc méthylomique est composé de 60 variables de méthylation de l'ADN (normalisation *rbd*), à savoir des variables pour 10 populations de peuplier d'Europe de l'Ouest (pour chaque gène, la valeur de méthylation est la moyenne de celle de deux individus de cette population), 3 contextes de méthylation (CG, CHG, CHH) et sur 2 régions (*features*) de l'ADN (promoteur ou corps du gène) résultant en $10 \times 3 \times 2 = 60$ variables. Pour les contextes de méthylation, C = Cytosine, G = Guanine, et H = Adénine, Cytosine ou Thymine. Le bloc transcriptomique est composé de 10 variables d'expression des gènes (normalisation TMM) pour les 10 mêmes populations (moyennage analogue à celui de la méthylation, sur les mêmes individus). Le nombre de gènes conservés après pré-traitement est de 28 267 pour les algorithmes tolérant quelques valeurs manquantes, 24 962 sinon. Les questions biologiques abordées sont la description des interactions entre les variables au sein et entre blocs dans le but de hiérarchiser les effets discriminant le plus les variables, puis la description des relations entre observations pour identifier les principaux groupes de gènes selon leurs valeurs de méthylation et expression, et enfin la sélection de gènes au profil extrême à la fois en terme de méthylation et d'expression.

5.2 Article Mardoc et al. (2024) : ***Genomic data integration tutorial, a plant case study.***

L'article Mardoc et al. (2024) (Pages 54 à 68, figures agrandies Pages 69 et 70, données supplémentaires en Annexe A Pages 199 à 205) exploite des ressources et données issues de l'article Sow et al. (2023) disponible en Annexe B (Pages 207 à 234, données supplémentaires en Annexe C Pages 235 à 246). Les approches d'analyse entre ces deux articles sont complémentaires, Sow et al. (2023) présentant les analyses uni- et bi-variées tandis que Mardoc et al. (2024) présente les analyses multi-variées conduites. Les données omiques à disposition sont issues de 20 génotypes de peuplier noir (*P. nigra*), répartis en 10 populations naturelles d'origines géographiques distinctes (Allemagne, France, Italie et Pays-bas). Pour chaque génotype, deux tissus ont été collectionnés (cambium et xylème) à raison de deux réplicas biologiques (clones) par génotype. La diversité génétique (*Single Nucleotide Polymorphisms* (SNPs)), épigénétique (méthylation de l'ADN) et l'expression des gènes (ARNm) ont été produites sur ce matériel biologique. Dans le cadre de l'article Mardoc et al. (2024), je me suis focalisé sur les données de méthylation de l'ADN et de l'expression des gènes pour étudier les interactions complexes entre méthylation et expression à l'échelle des populations, en complément des travaux de Sow et al. (2023) étudiant le rôle des variations épigénétiques sur l'évolution et l'adaptation des populations naturelles de peuplier. Ma contribution à ces deux articles est discutée dans la suite du manuscrit.

BMC Genomics

RESEARCH

Open Access

Genomic data integration tutorial, a plant case study

Emile Mardoc[1], Mamadou Dia Sow[1], Sébastien Déjean[2] and Jérôme Salse[1*]

Abstract

Background The ongoing evolution of the Next Generation Sequencing (NGS) technologies has led to the production of genomic data on a massive scale. While tools for genomic data integration and analysis are becoming increasingly available, the conceptual and analytical complexities still represent a great challenge in many biological contexts.

Results To address this issue, we describe a six-steps tutorial for the best practices in genomic data integration, consisting of (1) designing a data matrix; (2) formulating a specific biological question toward data description, selection and prediction; (3) selecting a tool adapted to the targeted questions; (4) preprocessing of the data; (5) conducting preliminary analysis, and finally (6) executing genomic data integration.

Conclusion The tutorial has been tested and demonstrated on publicly available genomic data generated from poplar (*Populus* L.), a woody plant model. We also developed a new graphical output for the unsupervised multi-block analysis, cimDiablo_v2, available at https://forgemia.inra.fr/umr-gdec/omics-integration-on-poplar, and allowing the selection of master drivers in genomic data variation and interplay.

Keywords Omics, Integration, System, Biology

Background

In recent years, the steady development of Next Generation Sequencing (NGS) and other high-throughput technologies has led to the massive production of genomic-derived data such as genome (DNA-seq), transcriptome (mRNA-seq), methylome (BS-seq), Transposase-Accessible Chromatin (ATAC-seq), *etc.* Such data allow to investigate, with an unprecedented precision and scale, the structure and evolution of genomes and their functioning in relation to phenotypes. While different types of genome-derived data (DNA variation, gene transcription, DNA-methylation, *etc.*) provide information on specific aspects of a biological system, they are ultimately interconnected and their combination likely contains information that cannot be accessed from individual data analysis. The added-value of genomic data integration (*i.e.* the combination of the data prior to the analysis, instead of analyzing each dataset separately and then combining the results) is illustrated in the literature in reducing the complexity of multiple datasets into a single dataset, considering that a combination of datasets can contain information missing in the individual datasets [1–4]. Multi-omics data integration is increasingly being used in human [5], animals [6], and microbes [7]. Data integration is also being extensively used in plant genomic research, emerging as a promising tool in green systems biology, precision plant breeding, and other biotechnological applications [8].

Genome-derived data, and omics data in general, are heterogeneous (quantitative, such as percentages or

*Correspondence:
Jérôme Salse
jerome.salse@inrae.fr
[1] UCA-INRAE UMR 1095 Genetics, Diversity and Ecophysiology of Cereals (GDEC), 5 Chemin de Beaulieu, 63000 Clermont-Ferrand, France
[2] Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, Université Paul Sabatier, Toulouse, France

counts, and qualitative, such as groups or classes) and produced in very large volumes, making data integration challenging. Analytical tools currently available for genomic data integration can be categorized on several levels. First, they differ in the statistical and mathematical framework, based on, e.g., dimension reduction, probabilistic models, or networks [9–12]. The integration procedures can be implemented at early, intermediate or late stages of data analysis; they can also be element or pathway-based, supervised or unsupervised, *etc.* [13, 14, 8]. Only a few are adaptable to large sets of biological features (species, individuals, tissues, genes, *etc.*) and genome-derived data such as genome, transcriptome, and methylome [15, 16]. Finally, the available tools can be categorized on the basis of the study objectives, e.g. referenced hereafter as description, selection and prediction [17–20]. In addition to the various approaches of data integration, it should also be recognized that data preprocessing and preliminary tests are essential for a successful implementation of data integration.

In combination with ever-increasing amounts of NGS-based genomic data available through public databases, data integration methods and approaches have the potential to transform our understanding of genome organization and gene regulation. To facilitate this progress, we propose a tutorial of best practices for genomic data integration, explained step-by-step and demonstrated on publicly available plant genome-derived data (from poplar [21]). The tutorial (Fig. 1) is structured in 6 consecutive steps that clarify the logical order of the procedures and allow to reach relevant conclusions, as illustrated using real datasets and exemplary research questions. It consists of (*i*) designing the adequate data matrix; (*ii*) formulating the targeted biological question; (*iii*) providing list of tools and methods for genomic data integration; (*iv*) data preprocessing, with considerations regarding missing values, outliers, normalization and batch effects; (*v*) conducting preliminary analysis where descriptive statistics and single omics analysis are necessary to properly understand the data structure and prevent misinterpretation; and finally (*vi*) performing genomic data integration with *mixOmics* on an illustrative case example.

Results

Genomic data matrix (Step #1)

When assessing genome-derived data from various experiments on a single individual, a group of individuals of a given species, or even on individuals from different species, one may want to gain better insight into the genomic variations and interplay of genes in the experimental design (for example, times series on a tissue exposed to a stress) used for data collection. Classically, omics data matrices consist of 'individuals' or 'samples' (biological units) arranged in lines for which available omic data ('variables') are listed in columns. However, genome-derived data, can be formatted as a matrix of genes considered here as the 'biological units', with genes arranged in lines and gene-related variables (e.g., diversity, expression, methylation, *etc.*) in columns (Fig. 1A). Such matrix can contain data for a single individual, multiple individuals of the same species (in additional columns), or even individuals from different species when comparing conserved genes. For the purpose of genomic data integration described here, we will consider genes as 'biological units' in lines and genome-derived data (expression, methylation, *etc.*) as 'variables' in columns. From such matrix, we propose a tutorial of the best practices dedicated to such genomic data integration following relevant steps of (*i*) the design of the data matrix, (*ii*) the identification of the biological questions, (*iii*) the choice of tools and methods for data integration, (*iv*) the data preprocessing, (*v*) the preliminary analysis with descriptive statistics and finally (*vi*) the genomics data integration.

In order to illustrate the use of such matrix in data integration procedures, we exploited public omics data obtained from poplar (described in [22] and [21]). The data represent transcription and cytosine methylation levels (considering the three CG, CHG and CHH contexts) for all annotated genes (considering promoter and gene-body) from ten natural poplar (*Populus nigra)* populations originating from Europe [21]. Overall, the investigated matrix consists of 70 columns (one transcriptome and six methylome columns referred to as 'variables' for each of the ten populations) and 42 950 lines (for annotated genes referred to as 'biological units') (Fig. 2A). This matrix will be used in the next steps of this tutorial.

Targeted scientific questions (Step #2)

Based on published literature, genomic data are typically integrated to answer biological questions in three general categories: (1) description of the major interplay between variables (*i.e.* genomics data) or samples (*i.e.* genes), for example how DNA methylation affects gene expression at the whole genome level?; (2) selection of 'biological units' (*i.e.* genes) considered as biomarkers for a specific genomic/phenotypic response; for example groups of genes with contrasting methylation and expression patterns, or (3) variables prediction from genomic data, for example, which combination of omic variables known in one individual or species can predict the genomic behaviors of such genes in other individual or species? considering that if we have a proven association of variables in a group of individuals, then one variable can be inferred from the other in additional individuals. Overall,

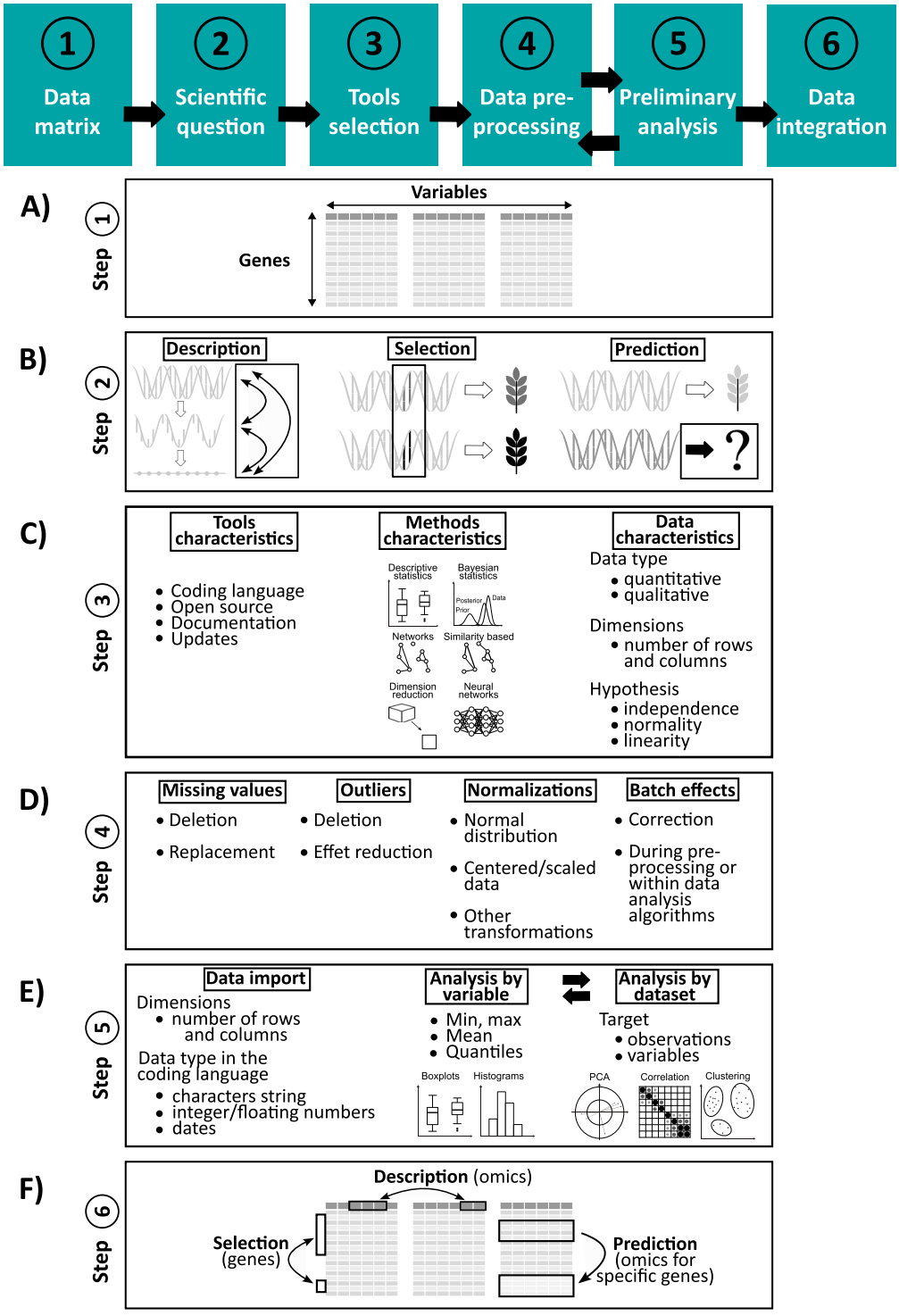


Fig. 1 Tutorial for genomic data integration. The tutorial of best practices presents the different steps to conduct multi-omics integrations: **A**, step 1 - Constructing the genomic data matrix. **B**, step 2 - Defining a clear and precise question of interest where biological questions concern describing omics interactions and interplay, selecting biomarkers specific to a trait, or predicting phenotypes from omics; **C**, step 3 - Selecting the tool, by considering tools' specificities such as their coding language, the accessibility or not to their source code, the quality of their documentation and frequency of their updates, their methods' main concepts and data requirements (see Table 1); **D**, step 4 - Preprocessing data, especially to remove or impute missing values, identify then remove outliers or reduce their impact, normalize data, correct batch effects; **E**, step 5 - Pre-analyzing data, by first importing data in the expected format with the right dimensions and types, then analyzing them by variable (univariate analysis) and dataset (multivariate analysis) to reveal major insights. **F**, step 6 - Genomic data integration for data description, selection and prediction.

specifying the targeted biological question impacts the next steps of the genomic data integration (Fig. 1B).

In the use case on poplar, we demonstrate the use of the data integration tutorial by addressing all the questions above within the description-selection-prediction continuum. At the whole genome level, the objective is to unravel the general interplay between methylome and transcriptome data, *i.e.* whether DNA methylation has an impact on gene expression. At the gene level, the objective is to identify groups of genes showing contrasted profiles for both transcriptome and methylome data and then investigate their biological functions.

Tool and method selection (Step #3)

Many tools for omics data integration have already been developed and described in published literature, and new methods keep emerging regularly. Table 1 reports 13 of the most cited tools available in R, a free software environment for statistical computing and graphics, providing lots of packages, especially in statistics and machine learning (Supplementary Table 1). Table 1 can help users to choose the most suitable tool for their particular data matrix and targeted scientific question. The selection process depends on the tools' characteristics, capabilities, methods involved, and acceptable types of data (Fig. 1C). Among these tools, *mixOmics* can address all the scientific questions in genomic data integration referenced to as data description, selection and prediction from both quantitative and qualitative data.

In the use case on poplar, we chose *mixOmics*, an open-source tool developed in the R programming language for omic data integration purposes. *mixOmics* contains several functionalities based on dimension reduction methods to explore one dataset or integrate two or more datasets, depending on the biological questions and data types available. The dimension reduction methods aim to extract the main sources of variation from datasets that are usually very large (*i.e.,* have potentially thousands of rows and/or columns). This tool is associated with an active forum (https://mixomics-users.discourse.group) and a documented website (http://mixomics.org/) that are available to help throughout the whole analysis, from the choice of the integrative and graphical functions to the interpretation of outputs. Moreover, *mixOmics* contains many functions to also analyze the data, which are generally derived from Principal Component Analysis (PCA) and Projection to Latent Structures (PLS) regression methods. These methods aim to factorize initial datasets/matrices (into: Components x Loadings + Residuals) to reduce datasets' dimensions while retaining the components with the main information from the initial omic variables. *mixOmics* also allows to display results with many graphical functions. "Sample plots" display observations (*i.e.* referred to as individuals or samples in

(See figure on next page.)

Fig. 2 Case study of genomic (expression and methylation) data integration from 10 poplar populations. **A** Genomic data matrix with 42 950 poplar genes in lines and 70 associated variables in columns (expression and methylation for 10 populations, color code in the legend at the left). Omics variables are gene expression and DNA methylation data produced for 10 populations of poplars, as presented at the bottom left legend of the figure. Methylation data were produced for 3 contexts of methylation (CG, CHG and CHH) on two gene features (gene-body or promoter). **B** Correlation matrix of the 60 methylomics and 10 transcriptomics log-transformed variables. This figure represents Spearman's correlation between each pair of omics variables. A high positive correlation between variables is represented by a deep blue point, a high negative correlation by a deep red point. No point means no correlation between variables. On the diagonal, correlations are by definition maximum and equals to one (*i.e.* correlated to themselves). The matrix' variables are arranged (see color code in the legend at the right) using a hierarchical clustering with AOE (angular order of the eigenvectors) order. **C** Loading plot of omics log-transformed, centered and scaled variables on the two first components of the PCA. Omics variable are plotted on PCA's two first principal components. For each component, the percentage of initial variance explained by this component is indicated (see color code in the legend at the left). **D** cimDiablo_v2's result on 'non-denoised' data. Left panel: Heatmap of omics integration. Each row corresponds to one gene and each column to one omics variable. Data were centered and scaled, then a cutoff was applied in [-2,2]. According to the heatmap's color code, blue corresponds to very low and red to very high methylated/expressed genes. Rows and columns' dendrograms are computed by hierarchical clusterings with the Euclidean distance and Ward method to cluster together genes and omics variables sharing similar insights. Right panel: Boxplots of k cluster groups. Using the rows dendrogram, genes were divided into four groups. For each group, the average value by population for each omics variable (methylation and gene expression) is represented. **E** cimDiablo_v2's result on 'denoised' data. Data were first centered and scaled, then 'denoised', centered and scaled a second time, before a final cutoff in [-2, 2]. **F** Comparison between 'non-denoised' and 'denoised' data for gene expression. Top panel: Boxplot of gene expression before and after the 'denoising' step. Red for 'non-denoised' data and blue for 'denoised' data. Bottom panel: MA-plot (Bland–Altman plot, where M represents the log ratio and A the mean average) of gene expression between 'denoised' and 'non-denoised' data for one of the poplar population (Adour). The x axis represents the average expression level while the y axis the log2 fold changes. Red for significant differences above |1| and black for no obvious differences. **G** Extraction of genes with extreme values (candidates) for all omics variables before and after the 'denoising' step. Left panel: Venn diagram of extracted genes before and after 'denoising'. Right panel: Heatmaps of genes with extreme values for 'non-denoised' and 'denoised' data. Gene lists are plotted using hierarchical clustering with Euclidean distance and Ward method. **H** Illustration of the Gene ontology enrichment analysis for genes with extreme values showing low expression and high methylation levels (143) after the 'denoising' step. Gene ontology enrichment has been performed using PlantGenle (https://plantgenie.org) with *Populus trichocarpa* v3.1 as background.

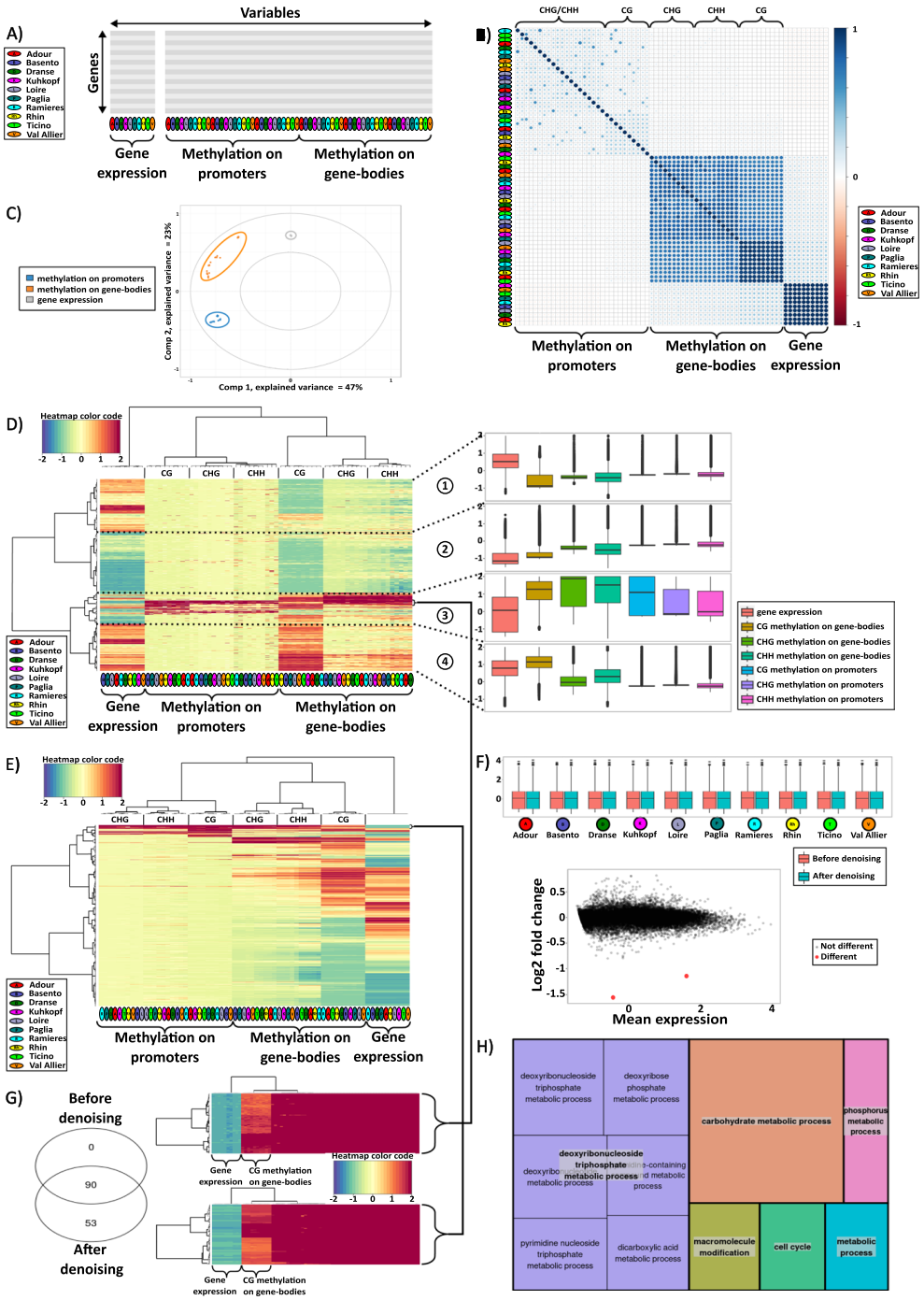


Fig. 2 (See legend on previous page.)

the current study), while "variable plots" display omics variables (*i.e.* genomic data in the current study).

Data preprocessing (Step #4)

Genomic data matrix must be preprocessed to take into account, when necessary, missing values, outliers, normalization and batch effects (Fig. 1D). Missing values can be handled by deletion (deleting each row or each column containing missing values or defining a threshold proportion of missing data over which columns-rows are deleted) or replacement (by 0, by the minimum of the values, by the average of other values, the median or a quantile and ultimately, by imputation), as summarized in [23]. Outliers, defined as an unusual value compared to the rest of the dataset (either due to error, or due to unique behavior of an investigated individual for a specific variable), can be (1) deleted from the dataset if considered as errors, (2) separated but not excluded if they represent an interesting behavior of the biological system (one analysis can focus on the outlier values, and another one on

Table 1 13 selected R tools for omics data integration

Tool's name	Scientific question	Method and tool's characteristics				Data's characteristics	
		Supervised Unsupervised	Method families	Summary	Updated	Omics	Hypothesis
BCC (Bayesian Consensus Clustering)	I) Description of samples' interactions	Unsupervised	Statistics	Computes a samples' clustering for each omics dataset by using a probabilistic model, then merges clusters to get a consensus cluster across omics datasets	No	Multi-omics (quantitative)	Normal distribution Different omics on the same set of samples
iCluster (iClusterPlus / iClusterBayes)	I) Description of samples' interactions	Unsupervised	Statistics / Dimension reduction	Starts with a latent variables regression across datasets by using a probabilistic model, then uses these joint latent variables for samples' clustering	Yes	Multi-omics (quantitative and qualitative)	Linearity assumption Normal noise distribution Different omics on the same set of samples
JIVE (Joint and Individual Variation Explained)	I) Description of samples/variables' interactions	Unsupervised	Dimension reduction	Decomposes each dataset in three terms: a joint effect (across datasets), an individual effect (specific to the dataset) and a noise effect	No	Multi-omics (quantitative)	Linearity assumption
LRAcluster (Low-Rank Approximation Cluster)	I) Description of samples' interactions	Unsupervised	Statistics / Dimension reduction	Probabilistically computes a common low-dimensional subspace across omics, then uses the K-means algorithm to cluster samples on this subspace	Yes	Multi-omics (quantitative and qualitative)	Linearity assumption Different omics on the same set of samples
MCIA (Multiple co-inertia analysis) (MCOA)	I) Description of samples/variables' interactions	Unsupervised	Dimension reduction	Projects each dataset on a subspace, then maximizes co-inertia between subspaces to get major information shared by datasets	Yes	Multi-omics (quantitative)	Linearity assumption Different omics on the same set of samples
mixKernel	I) Description of samples/variables' interactions II) Variables selection III) Phenotype prediction	Supervised Unsupervised	Dimension reduction	Transforms datasets with kernels, then applies usual dimension reduction methods	Yes	Multi-omics (quantitative and qualitative)	Datasets with the same rows or columns

Table 1 (continued)

Tool's name	Scientific question	Method and tool's characteristics				Data's characteristics	
		Supervised Unsupervised	Method families	Summary	Updated	Omics	Hypothesis
mixOmics (with PCA, PLS, rCCA, Diablo…)	I) Description of samples/variables' interactions II) Variables selection III) Phenotype prediction	Supervised Unsupervised	Dimension reduction	Contains many matrix factorization methods for multivariate analysis and functions for data visualization. The main analysis method for one single dataset is the PCA. For two datasets or more, the main methods are the PLS and rCCA, and their extentions for discriminant analysis, variable selection ('sparse') and multi-blocks analysis	Yes	Multi-omics (quantitative and qualitative)	Linearity assumption Datasets with the same rows or columns
moCluster (from MOGSA)	I) Description of samples' interactions	Unsupervised	Statistics / Dimension reduction	Computes latent variables by using a PCA's extension, then clusters them and finally select the best subtype model	Yes	Multi-omics (quantitative)	Linearity assumption Different omics on the same set of samples
MOFA (Multi-Omics Factor Analysis)(MOFA2)	I) Description of samples' interactions III) Phenotype prediction	Unsupervised	Statistics / Dimension reduction	Factorizes datasets with a Bayesian approach to get a small number of latent factors usable for different purposes	Yes	Multi-omics (quantitative and qualitative)	Linearity assumpion
NEMO (NEighborhood based Multi-Omics clustering)	I) Description of samples' interactions	Unsupervised	Similarity-based	Creates one similarity matrix by dataset, then merges them and finally clusters the merged matrix by Spectral clustering	No	Multi-omics (quantitative)	Euclidean distance metric

Table 1 (continued)

Tool's name	Scientific question	Method and tool's characteristics				Data's characteristics	
		Supervised Unsupervised	Method families	Summary	Updated	Omics	Hypothesis
PINS (Perturbation clustering for data INtegration and disease Subtyping)(PINSPlus)	I) Description of samples' interactions	Unsupervised	Similarity-based / Network	Does several clustering to identify how often samples are clustered together. Clusterings are made on different datasets, with data perturbed by adding gaussian noise, and different clustering methods are used	Yes	Multi-omics (quantitative)	Different omics on the same set of samples
RGCCA (Regularized Generalized Canonical Correlation Analysis) (sGCCA)	I) Description of samples/variables' interactions II) Variables selection III) Phenotype prediction	Supervised Unsupervised	Dimension reduction	Computes latent variables for each dataset by maximizing correlations within and/or between datasets	Yes	Multi-omics (quantitative and qualitative)	Linearity assumption Different omics on the same set of samples
SNF (Similarity Network Fusion)	I) Description of samples' interactions	Unsupervised	Similarity-based / Network	Creates a similarity matrix then an associated network for each dataset, then iteratively fuses the networks to keep only strong correlations between samples across omics	No	Multi-omics (quantitative and qualitative)	Different omics on the same set of samples Euclidean distance metric

Rows correspond to 13 selected tools and columns to the main characteristics to consider while selecting a tool for omics data integration. 'Biological question' describes which of the three main biological questions presented in the article the tool aims to answer. 'Methods and tools characteristics' details if they can be used for supervised or unsupervised analysis, at which methods' family they belong to (statistical, dimension reduction, networks, similarity-based, artificial neural networks), summarizes its functioning, and indicates if the tool is still updated with its source code's repository available in Supplementary Table 1. 'Data characteristics' describes the types of omics the tool can afford. The last column presents the main hypothesis on data, assuming they (1) follow a normal distribution (which can be tested using for instance Shapiro tests or QQ-plots), (2) share identical rows or columns (*e.g.* omics data produced on the same genes or individuals), (3) have linear interactions (*i.e.* does not consider more complex interactions such as polynomial interactions) or (4) are considered as similar according only to the Euclidean distance (*i.e.* does not consider other metrics of similarities such as correlations)

the rest of the data), or (3) transformed (normalized) in order to keep the outliers in the data matrix but reduce their effect in the analysis. If the user decides to reduce the outliers' effects, one solution is to use the logarithmic normalization of the variable. This transformation has a minor impact on low values but strongly reduces high values, while retaining the rank (relative order) of data points. However, log-transformation is not applicable to zero values, in which case errors are returned. To overcome this problem, a constant (and relatively small) value is usually added to all zeros, e.g. (+1) as we did in the case study. The appropriate increment applied to zeros prior to the log-transformation needs to be carefully considered with the biological meaning/interpretation of such transformation. In the case study, we choose +1 for log2 transformation of the transcriptomic data, because we consider that genes with TPM (Transcripts Per Million) ≤ 1 are not expressed and replacing zeros with +1 would therefore not affect the results. However, data transformations also change the scale of the variables, making it more difficult to interpret the new values. Generally, data normalization methods consist of transforming raw data, e.g. in order to obtain values that follow a Gaussian distribution to be used with parametric tests, or to center and/or scale heterogenous variables facilitating their comparison [24]. Data normality (*i.e.* a match with the Gaussian distribution also called normal distribution) can be tested for instance with Shapiro tests or QQ-plots [25]. Normalizations, as all other data transformations, should always be considered with regards to the effect on the interpretability of the results (*e.g.* does a normalization procedure that changes positive values to negative ones affect the subsequent computational steps or result interpretation?). Finally, batch effects are effects caused by a non-biological factor during any step of data production (including the preparation of biological material), undesirably biasing sub-groups of data [26]. They are generally identified during preliminary data analysis and can be corrected by functions such as *ComBat* (from R package *sva*, https://rdrr.io/bioc/sva/man/ComBat.html) and *removeBatchEffect* (from R package *limma*, https://rdrr.io/bioc/limma/man/removeBatchEffect.html).

In the illustrative case study, the data needed preprocessing due to the presence of missing values (i), and outliers (ii). In order to reduce the missingness (i), only genes with at least one expression and one methylation value were considered. Hence, from the 42 950 genes annotated in the poplar genome (v3.1), we kept 31 040 genes (72%). Secondly, we fixed the missing value cut-off to 10%, allowing only 10% of missing values for the whole dataset, further trimming the gene set to 28 267 (67% of the annotated genes). In addition, we created another dataset where no missing value was allowed for the investigated variables (genomics data), consisting of 24 962 genes (58% of the annotated genes). The outlier issue (ii) was addressed with a logarithm function, which reduced the impact of extreme values (Supplementary Fig. 1). In the case of the methylation variable (Supplementary Fig. 1A), promoter methylation remained to be strongly impacted by a few genes with extreme values, even after the log-transformation. Methylation on gene-bodies is more variable, with few values approaching zero (*i.e.* no methylation). Gene-body methylation is higher in the CG-context compared to the other sequence contexts, with a large group of highly methylated genes. Regarding transcription (Supplementary Fig. 1B), genes are divided in two groups, non-expressed *vs.* highly expressed genes.

Preliminary data analysis (Step #5)

This step consists of clarifying variable types, data dimension, and associations between rows ('biological units', *i.e.* genes) or columns ('variables', *i.e.* genomic data) of each dataset (Fig. 1E). On a single dataset level, useful methods for such preliminary analysis are (1) the Principal Component Analysis (PCA) allowing to extract the major information signal contained in the dataset, (2) the correlation matrix to highlight strongly associated pairs of variables, and (3) clustering methods such as the hierarchical clustering or the K-means algorithms to identify the most similar pairs of individuals or variables.

In our case example, the preliminary data step consisted of displaying the matrix of correlations (Fig. 2B) between each pair of variables (columns). Most variables were strongly correlated among the ten individuals (representing different populations) of poplar, indicating little variation between populations, especially for transcription and gene-body methylation in the CG context. Much weaker correlation was observed between individuals for promoter methylation; nonetheless, a cluster analysis consistently grouped all individuals in blocks of different variables. Weak correlations were observed between some omics blocks (*i.e.* between expression and gene-body methylation, or between promoter methylation and gene-body methylation). Another way to rank the genomic effects was highlighted with the PCA (Fig. 2C). Typically, a 'score plot' is used to display samples ('biological units', here genes) on the first couple of principal components, in order to identify clusters of samples and possible outliers (standard PCA plots). Here, we used a 'loading plot' to map variables on these components, in order to identify how omics variables are clustered. On the first principal component, data are divided in two groups, methylation *vs.* expression. On the second component, data are divided in two other groups, promoter methylation *vs.* gene-body methylation and expression. Overall, such preliminary data analysis revealed that the

contribution of the omic variables to the variation in the dataset follows this order, from the highest to the lowest effects, (1) the type of omics (expression *vs.* methylation), (2) the methylated compartment of genes (promoters *vs.* gene-bodies) and (3) the methylation contexts (CG *vs.* CHG and CHH), and finally (4) the different populations (individuals). Moreover, it revealed that gene-body methylation is more variable than promoter methylation, especially in the CG context. No general association between gene methylation and expression has been observed with this analysis at the whole genome level.

Genomic data integration (Step #6)

In addition to pairwise genomic data comparison [22], multiscale genomic data integration aims at characterizing hidden and potentially complex interactions between different omics data in order to provide a better comprehensive understanding of cellular and biological processes (Fig. 1F). The PLS regression function in *mixOmics*, allowing to compute a linear combination of omics variables to extract a smaller number of 'components' retaining data variability, is generally recommended for two omics datasets analyzed (see description in Step #3). Since our methylation data is more complex and was described into six variables (two gene partitions - promoters and gene-bodies - and three sequence contexts - CG, CHG, CHH), we opted for '*block.pls*', a generalization of the PLS for more than two datasets (called 'blocks'). This choice was directly made based on the insights gained from the previous preliminary analysis, particularly the different strength of association between gene expression and the various methylation subtypes. We used the *block.pls* function in its 'regression' mode, with methylation as explicative data and expression as explicated data, to focus on the impact of DNA methylation on gene expression. First, a 'design matrix' needs to be set up, considering the data being integrated. The 'design matrix' contains weights between all pairs of blocks with values multiplying each covariance between two blocks: a higher value for interactions between pairs of blocks (*i.e.* values multiplying covariance between two blocks), where a high value is assigned for interactions of high interest, and a low value is assigned for interactions of low interest [27]. In the use case on polar, the chosen weights are 0 for each block with itself, 1 for each methylome block with the transcriptomics block and 0.1 for each pair of methylome blocks. These weighting values were specifically chosen to focus on the interactions between methylome and transcriptome data, but also to take into account the interactions between the different methylome blocks/contexts or gene compartments. Such interactions are well-represented on a clustered heatmap, similar to the output of the graphical function *cimDiablo* from *mixOmics* (for multi-blocks PLS regressions). However, this function is currently applicable only for the discriminant methods *block.plsda* and *block.splsda*, and not for the non-discriminant methods *block.pls* and *block.spls*.

To overcome this limitation, here we developed in the current study *cimDiablo_v2,* a new function based on *cimDiablo* from mixOmics for non-discriminant *block.pls* and *block.spls* objects, which could take into account a 'denoising' step, and is publicly available at https://forgemia.inra.fr/umr-gdec/omics-integration-on-poplar. The 'denoising' step uses the components and loadings' matrices from *block.(s)pls*. The components are computed to keep the essential information from the initial data, mainly the variability conserved across omics variables. Therefore, the variability that is specific to one row (*e.g.* sample, individual, gene *etc.*) or one column (omics variable), referred to as 'noise', is extracted from the matrix of components and placed into a matrix of residuals, *cimDiablo_v2*. The 'denoising' step displays the data without the residuals (*i.e.* noise), corresponding to the matrix product of the components and loadings matrices, referred to as 'denoised' matrix. Most of *cimDiablo_v2* parameters are the same as for *cimDiablo*, with only a few specific to *cimDiablo_v2*. These correspond to the binary parameters to apply for data transformations: 'denoise' (to 'denoise' data as described above), 'scale2' (to center and scale data after the 'denoising' step) and 'cutoff ' (to set all values higher than 2 to 2 and all values lower than -2 to -2).

We ran *cimDiablo_v2* on the poplar data with and without the 'denoising' step (Fig. 2D, Fig. 2E, Supplementary Fig. 2). The graphical function of *cimDiablo_v2* (like *cimDiablo*) allows to display a heatmap where hierarchical clustering reveals the interactions between rows (*i.e.* genes) and between columns (*i.e.* variables) and the heatmap of rows-columns interactions. Without the 'denoising' step, the overall picture resulted in clustering of samples by omics variables (promoter methylation, gene-body methylation and gene expression), Fig. 2D. Gene-bodies appeared to be clearly more methylated than promoter regions especially in the CG context, confirming our previous results in the preliminary analysis (step #5). More marked methylation differences between populations were observed for non-CG contexts, especially for the CHH context which drives more methylation differences between populations. We then split the dendrogram into 4 groups (or clusters) of genes (rows) defining different typologies of genes in their methylation-expression regulation (Fig. 2D), delivering the following gene categories, (1) high expression level and low methylation levels for promoters and gene-body in the three methylation contexts (2) low expression level, low CG gene-body

methylation and moderate non-CG gene-body and promoter methylation levels (3) high methylation levels and moderate to low expression levels, and finally (4) high expression and CG gene-body methylation with moderate methylation in other features and contexts.

In order to focus on the general trends or 'master' genomics regulators-drivers (*i.e.* genes that are regulated though expression-methylation, or omics in general, interplay) in all populations, we used the 'denoising' step aiming at removing the variability that is specific to one row, variable or population (Fig. 2E). After the 'denoising' process, there is less variation between genomic variables per gene. At the whole genome level, this 'smoothed' effect of the 'denoising' step on the clustering heat map (Fig. 2E) can help identifying general trends or master regulators-drivers, i.e. genes with expression-methylation interplay shared between any investigated individuals, development stage, etc. At the gene level, it also allows to identify only genes with extreme omic profiles. To assess the impact of the 'denoising' step, we compare the 'non-denoised' and 'denoised' data (Supplementary Fig. 3). Interestingly, the two 'denoised' and 'non-denoised' datasets look quite similar, especially for expression data suggesting that only a few values have been changed. To precisely quantify those changes, we display MA plots to assess differences between the 'denoised' and 'non-denoised' data. Only a few genes show expression differences after the denoising step (Fig. 2F). However, for methylation data, more differences are observed after the 'denoising' step as more marked variability was initially reported between the populations, especially for promoter and non-CG gene-body methylation (Supplementary Fig. 4). Overall, we recommend the use of the 'denoising' step to assess both genome-wide and gene level interplays between genomics data in order to identify general trends as well as master regulators-drivers, or in a broader sense, to remove the variance between biological replicates in experimental setups.

In order to identify genes where the omic pattern is indicative of an association between DNA methylation and gene expression, we used the 'denoised' matrix containing no missing value (24 962 genes). Interestingly, only few genes (143) showed a contrasting pattern between expression and all methylation variables simultaneously, *i.e.* highly methylated and lowly expressed genes in all studied populations (Fig. 2G). Comparison between 'non-denoised' and 'denoised' datasets revealed that all highly methylated and lowly expressed genes identified before 'denoising' (90 genes) are found after the 'denoising' procedure, suggesting that 'denoising' does not lead to signal loss. However, 53 additional genes were specifically identified only in the 'denoised' data, suggesting that the procedure improves sensitivity of signal detection. This outcome aligns with expectations for a procedure that removes signals appearing only once (in a single omic variable or population). Gene Ontology (GO) analysis on the identified set of highly methylated and lowly expressed genes revealed enrichment in functions related to involvement in carbohydrate, cell cycle, phosphorus and metabolic processes (Fig. 2H). Among these genes, we identified *Di19* (Drought induced 19) that enhance drought tolerance in transgenic poplar plants [28]. First identified in *Arabidopsis, Di19* has been characterized as a new type of transcription factor, directly up-regulating the expression of *PR1, PR2* and *PR5* in response to drought stress [29]. The results of our integrated omic analysis may suggest that the expression of *Di19* in poplar trees could be associated with DNA methylation, suggesting a possible epigenetic regulation of this gene that can be explored in future studies to be potentially exploited in breeding schemes especially in response to drought stress.

Discussion

The constant development in sequencing methods and strategies, as well as reduction in cost, allows access in the public domain to genomic data from many plant species. How to make proper use of these data to unravel plant genome organization and regulation in different environmental contexts remains a key question for both fundamental and applied research, especially in characterizing genomic makers of crop adaptation to constraints to be exploited in breeding schemes. A tutorial of best practices when conducting genomic data integration has been proposed from a specific data matrix consisting on genes (rows) and genomic variables (columns) in order to unravel the genomic interplay of genes of several individuals of given species or individual from distinct species. The proposed tutorial applied here on the integration of genomic data can also be applied on any omics data taking into account non genome-derived data (*i.e.* proteome, metabolome, phenotype…) with individuals (*i.e.* accession) instead of genes, in lines. The tutorial has been illustrated on a case example from methylation and expression data obtained from 10 poplar populations from Europe to reveal genomic interplay (between expression and methylation) at the whole genome and gene levels. The proposed tutorial is divided in 6 steps: data matrix design, biological question, tools selection, data preprocessing, preliminary analysis and finally genomic data integration.

Regarding the targeted biological question, genomic data integration is generally conducted for the (1) description of the major interplay between variables, (2) selection of genes considered as biomarkers, (3) prediction of some variables from genomic data. Regarding the

first type of questions, one may be interested in global interplay between the omics data, for example correlations between genomic variables. When addressing the second type of questions, specific groups of genes showing similar or contrasted behaviors on one or several variables are selected. For example, users can look for genes from which genomic variation control a specific phenotype (resistance to a disease or temperature stress). Finally, regarding the third type of questions, phenotype prediction from omics data consists in transferring knowledge of omics-phenotypes interactions across individuals (plant varieties, animal species, *etc.*). Hence, it has been used in medicine to predict diseases evolution from cohorts [30], or in agronomy to predict yield (grain) production in cereals [31].

Regarding the tools available to conduct omics data integration, scientific articles already offered reviews or benchmarks of omics data integrative tools and methods to help choosing the best integrative approach. Here, we report from eight review articles [11, 13, 18–20, 32–34], 13 most cited tools in R programming (Table 1), although popular tools such as t-SNE [35] or UMAP [36], usable for non-linear dimension reduction, were not cited by these articles. The 13 tools rely on distinct methods to consider with caution depending on the biological question addressed, (1) descriptive and inferential statistics, (2) dimension reduction, (3) network and/or (4) similarity-based approaches. The descriptive statistical approaches [37] use mathematical means such as the mean, median, variance, standard deviation and graphics such as boxplots to describe the data. Statistic tests are part of the inferential statistics [38] aiming to validate or not a hypothesis on data's probabilistic distribution. Bayesian approaches [39] also belong to the inferential statistics and assume, before data analysis, that these data follow a chosen probabilistic distribution called the *prior*, then compute the *posterior* distribution by fitting the *prior* to the data. Regarding dimension reduction methods [12, 40], they aim to extract the largest part of the information contained in the data and store it in new data with lower number of dimensions. Once the omics integration tools (13 proposed) and associated methods (6 described) have been selected based on the methodological principles previously described, data need to be treated prior integration.

Regarding the data preprocessing prior to integration, four major steps have been identified, concerning missing values, outliers, normalization-transformation and the batch effects. To help to conduct data preprocessing, some algorithms have been developed for testing and comparing different methods, for instance with the R packages *missMethods* (https://rdrr.io/cran/missMethods/), *outliers* (https://rdrr.io/cran/outliers/), *bestNormalize* (https://rdrr.io/cran/bestNormalize/) and *bapred* (https://rdrr.io/cran/bapred/), overall allowing to manage the four preprocessing steps.

The preliminary analysis consists of conducting a descriptive investigation of the data in order to avoid misinterpretation of the results based on basic graphics to clarify variables' types, data dimension, basic associations between rows (samples) or columns (variables) of each dataset, *etc.* Knowing variables' ranges and distributions is very useful for data preprocessing. Indeed, users must decide whether values should be centered and/or scaled according to data ranges and distributions. Common methods are to look for minimum, median, mean and maximum values, or to derive graphs such as boxplots or histograms, when possible. Outliers are in general explicitly detected and visible by boxplots representation or in histograms when there is enough data. Before data integration, preliminary analysis of each dataset separately is strongly recommended in order to avoid mis- or over-interpretation of the results.

Other workflows from the scientific literature are usually divided in three steps: omics data, data preprocessing, simple and integrative analysis [17, 34, 41]. Compromises have been done in this study to offer a tutorial of best practices with enough information to integrate genomic datasets. Omics integrative methods for more specific purposes are presented in reviews and benchmark articles [10, 12, 16, 17, 42–47]. Moreover, some steps of the data preprocessing (redundancy, heterogeneity, *etc.*) are not presented here, but available for example in [33]. Many benchmark articles [18, 19, 32, 34] also discuss concrete effects of omics integration tools on the same datasets. From our knowledge, only [27] presents a workflow starting from the biological question of interest, that we consider should be the starting point of any workflow for conducting omic data integration. This workflow also has the advantage to be cyclic, as multi-omics integration naturally lead to new questions and then additional analysis. In complement to the previous articles, we provide here a step-by-step procedure allowing to conduct genomic data integration that we made publicly available at https://forgemia.inra.fr/umr-gdec/omics-integration-on-poplar.

Methylome and transcriptomics data from poplar have been integrated following the proposed tutorial to assess the impact of DNA methylation on gene expression, and to identify candidate genes with contrasted profiles across methylome and transcriptomics data. We developed a new function '*cimDiablo_v2*' allowing the possibility to 'denoise' data to maximize the identification of 'real' or 'strong' omics interplay where managing noise in omics data is still a challenge [48, 49], especially to remove exclusively variability with no biological meaning. Our

proposed method does not focus on removing biological meaningless information, but more precisely on removing isolated variability, *i.e.* gene variability specific to one omics variable, while keeping gene variability shared across omic variables or individuals. The case example has permitted to obtain several results with (1) the highest effect that structure the investigated omics data being the data type (transcriptomics *vs.* methylomics), then the gene feature (gene-bodies *vs.* promoters), the methylation context (CG *vs.* CHG *vs.* CHH methylation), and finally the population (10 populations from western Europe); (2) there is more variability on methylation on gene-bodies than promoters, especially the CG methylation; (3) there is no general trend between gene methylation and expression at the whole genome level; and (4) genes with contrasted expression-methylation profiles across omics variables are involved in carbohydrate, cell cycle, phosphorus and metabolic process and key functions involved in response to stresses, an important trait for the adaptation of perennial species (poplar) to different geographical environments. Overall, the use case illustrates the power of genomic data integration to identify genes driving key traits through specific genomic (expression-methylation) interplay that can be precisely identified, prior their exploitation in crop management and breeding schemes.

Conclusion

We propose a step-by-step tutorial for genomic (*i.e.* DNA-based) data integration illustrated on a case example on poplar plant consisting in (1) designing a data matrix, (2) defining a specific biological question, (3) selecting the appropriate tools, (4) performing data preprocessing, (5) conducting preliminary analysis, and (6) performing multi-omics integration. In addition, we developed *cimDiablo_v2,* a new function based on *cimDiablo* from *mixOmics* for non-discriminant *block.pls* and *block.spls* objects available at https://forgemia.inra.fr/umr-gdec/omics-integration-on-poplar and exploitable on any type of omics data.

Materials and method

Genomic data analyzed

Genome - Genomic data analyzed here have been retrieved from [21] and [22] for DNA methylation and gene expression respectively. The samples used here are initially from [21] where authors analyzed a collection of 241 genotypes of *P. nigra* populations using RNA-seq in order to assess gene expression. Recently, [22] retrieved a subset of 10 populations (*i.e.* 20 genotypes) from [21] on the same tree individuals and the same sampling time. This subset of 10 populations were analyzed using WGBS (methylome), together with the transcriptome data from [21] in order to assess the role of epigenetic regulation in driving tree species evolution and adaptation. *Methylome* - DNA methylation analysis was done in [22] on the same sample powders from [21]. Briefly, for DNA methylation, genomic DNA was extracted using a cetyl trimethyl-ammonium bromide (CTAB) protocol and whole-genome bisulfite sequencing was performed in accordance with the procedure described by [50]. Reads from sequencing were then mapped against the poplar v3.1 reference genome and methylation call realized with *BSMAP* [51] using default options, delivering 3 datasets for the 3 methylation contexts. The *Methylkit* (v1.18.0) and *genomation* (v1.32.0) R packages were used for the annotation of DNA methylation data in genomic regions (promoters and gene body). Hence, the methylome dataset consists in three methylation contexts on two genes' regions, overall producing six methylation variables by population. Methylome dataset is expressed as the number of methylated cytosines x number of methylated cytosines / number of cytosines, called rbd (read by density), [22]. *Transcriptome* - For transcriptomic (gene expression) dataset, from [21], RNA-seq was carried out with Illumina Hiseq2000 platform. Reads were mapped on the *Populus trichocarpa* v3.0 reference genome using *bowtie2* (v2.4.1) [52]. Raw counts were then normalized by Trimmed Mean of M-values (TMM) from edgeR (v3.26.4) [53] as described in [21].

Genomic data integration

Genomic data matrix - The design matrix consists of 42 950 polar genes in lines and 70 associated variables in columns. *Defining a specific biological question* - At the whole genome level, the objective is to unravel the general interplay between methylome and transcriptome data, *i.e.* if DNA methylation has an impact on gene expression. At the genes level, the objective is to identify groups of genes showing contrasted profiles for both transcriptome and methylome data and then investigate their biological functions. *Selecting appropriate tool* - We use *mixOmics* for genomic data integration associated with an active forum (https://mixomics-users.discourse.group) and a documented website (http://mixomics.org/). *Performing data preprocessing* - To deal with missing values, only genes with at least one expression and one methylation values and less than 10% of missing values for the whole dataset were considered. To reduce outliers' impact, data were log-transformed with $\log_2(1+x)$, where x represents methylation / expression values and 1 the constant number added when dealing with zero values. *Conducting preliminary analysis* - A matrix of correlations was performed on preprocessed data without missing values using Spearman correlation and AOE (Angular Order

of the Eigenvectors) criteria for variables clustering. PCA analysis was conducted on preprocessed data with missing values, firstly centered and scaled, to compute 2 components. *Performing multi-omics integration - mixOmics block.pls* regression was conducted on 6 methylation on 1 expression blocks, with a design matrix composed of 1 between expression and methylation, 0.1 between methylation blocks and 0 within each block, and finally 2 components computed by block. Data were first centered and scaled in block.pls, then several cimDiablo_v2 results were obtained depending on if data are 'denoised', centered and scaled a second time and/or cut in [-2, 2]. For cimDiablo_v2 plots, a hierarchical clustering was used with the Euclidean distance and Ward method. Master regulators-drivers were selected both on 'non-denoised' and 'denoised' data, by selecting genes with all methylation values higher than 1 and all expression values lower than -1. For the comparison of 'non-denoised' and 'denoised' data, log2 fold changes cut-off above |1| were applied using MA-plot, and a Bland-Altman plot for visual representation of genomic data.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09833-0.

Additional file 1: Supplementary Figure 1. Histograms of methylomics and transcriptomics' logged distributions from one poplar (Adour) population. **Supplementary Figure 2.** Omics data integration with cimDiablo_v2. **Supplementary Figure 3.** Boxplots of k cluster groups in each poplar population for gene expression and methylation. **Supplementary Figure 4.** Comparison between 'non-denoised' and 'denoised' data for methylation in gene-body and promoter for CG, CHG and CHH contexts.

Additional file 2: Supplementary Table 1. 39 R tools for multi-omics data integration. **Supplementary Table 2.** Gene annotation of 'no-denoised' and 'denoised' candidate genes. Annotation information have been retrieved from PlantGenIE (https://plantgenie.org/) with Populus trichocarpa v3.1 as a reference. The column 'common_before_and_after_denoising' indicates whether the gene is shared between 'denoised' and 'no-denoised' data or not (TRUE/FALSE).

Acknowledgements
Not applicable.

Authors' contributions
All authors edited and helped to improve the manuscript, read and approved the final manuscript. EM conducted all the analysis; MDS participated to the analysis; SD provided advises during the analysis conduction; JS assumed the coordination of the funded project

Funding
The current study has been funded by the ANR EpiTree project (ANR-17-CE32-0009) and ISITE CAP 2025 from Auvergne University (axis Big DATA) and INRAE BAP department.

Availability of data and materials
Genomic data integration code in R is made publicly available at https://forgemia.inra.fr/umr-gdec/omics-integration-on-poplar. Datasets produced in the current article are made available at https://entrepot.recherche.data.gouv.fr/privateurl.xhtml?token=d946bb29-4698-4bee-9c6b-c2d98558ca8a. Plants material is directly involved in the study with omics data derived from natural poplar populations described in the material and method section of the current manuscript.

Declarations

Ethics approval and consent to participate
Data described in the current manuscript follow the FAIR (Findable, Accessible, Interoperable, Reusable) principles in term of data availability, reproducibly and free access.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

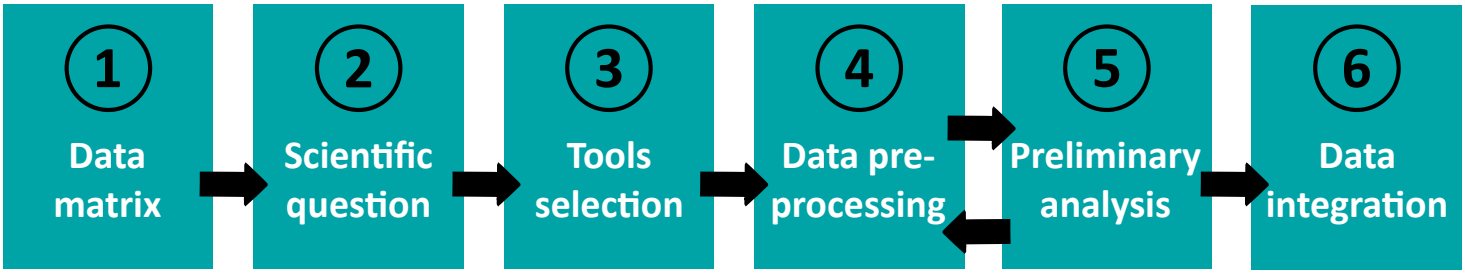
Received: 21 July 2023 Accepted: 22 November 2023
Published online: 17 January 2024

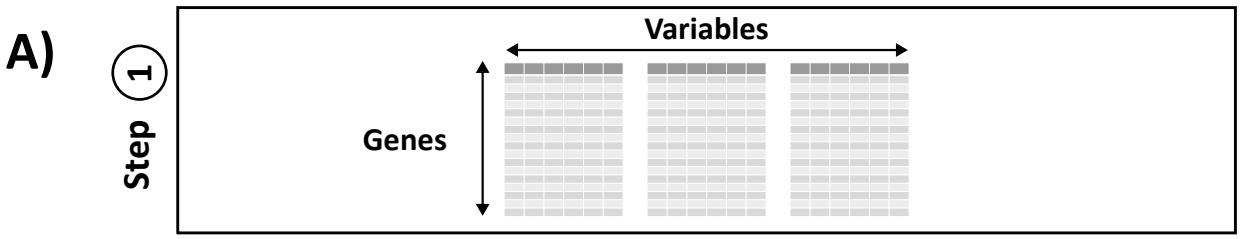
References

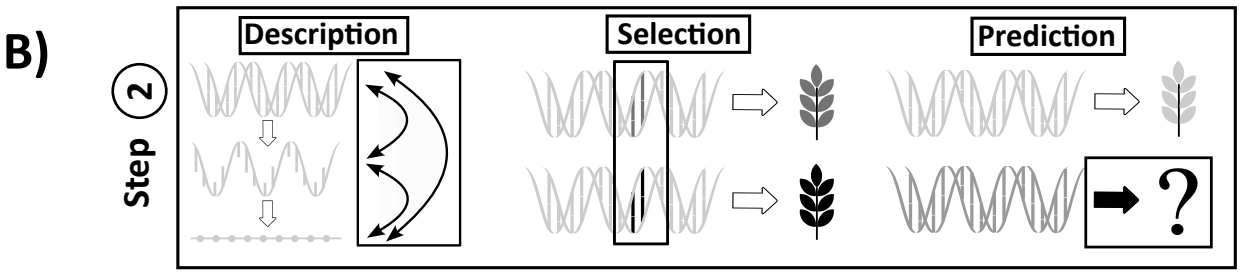
1. Tabakhi S, Suvon MNI, Ahadian P, Lu H. Multimodal learning for multi-omics: a survey. 2022.
2. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv. 2021;49:107739.
3. Krassowski M, Das V, Sahu SK, Misra BB. State of the field in multi-omics research: from computational needs to data mining and sharing. Front Genet. 2020;11:610798.
4. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front Genet. 2017;8:84.
5. Shetty SA, Smidt H, De Vos WM. Reconstructing functional networks in the human intestinal tract using synthetic microbiomes. Curr Opin Biotechnol. 2019;58:146–54.
6. Kim D-Y, Kim J-M. Multi-omics integration strategies for animal epigenetic studies — A review. Anim Biosci. 2021;34:1271–82.
7. Gutleben J, Chaib De Mares M, Van Elsas JD, Smidt H, Overmann J, Sipkema D. The multi-omics promise in context: from sequence to microbial isolate. Crit Rev Microbiol. 2018;44:212–29.
8. Jamil IN, Remali J, Azizan KA, Nor Muhammad NA, Arita M, Goh H-H, et al. Systematic Multi-Omics Integration (MOI) approach in plant systems biology. Front Plant Sci. 2020;11:944.
9. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform. 2022;23:bbab454.
10. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat Commun. 2021;12:124.
11. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. Bioinforma Biol Insights. 2020;14:117793221989905.
12. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016;17:628–41.
13. Vahabi N, Michailidis G. Unsupervised multi-omics data integration methods: a comprehensive review. Front Genet. 2022;13:854752.
14. Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–46.
15. Zhang R, Zhang C, Yu C, Dong J, Hu J. Integration of multi-omics technologies for crop improvement: status and prospects. Front Bioinforma. 2022;2:1027457.
16. Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, et al. Systems biology and multi-omics integration: viewpoints from the metabolomics research community. Metabolites. 2019;9:76.
17. Li C, Gao Z, Su B, Xu G, Lin X. Data analysis methods for defining biomarkers from omics data. Anal Bioanal Chem. 2022;414(1):235–50.

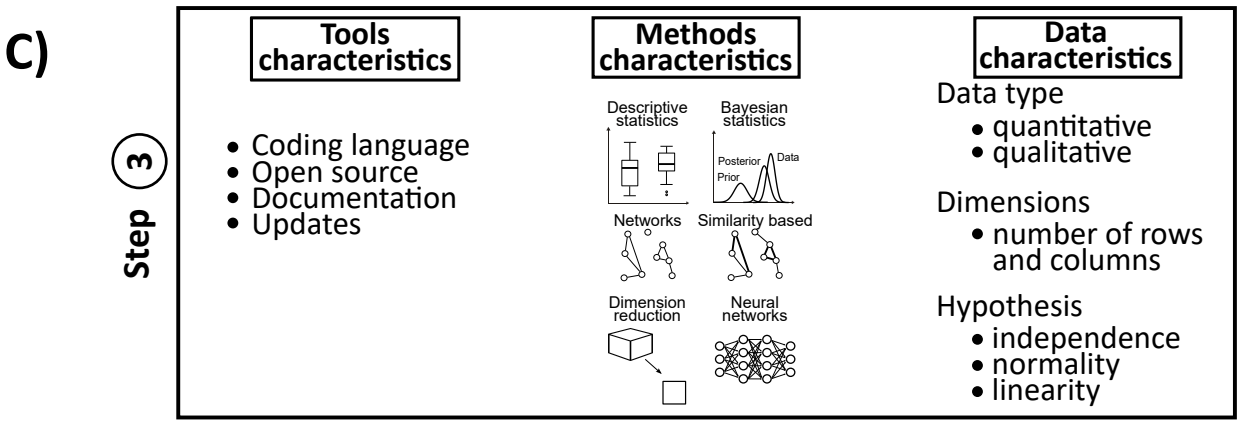
18. Lovino M, Randazzo V, Ciravegna G, Barbiero P, Ficarra E, Cirrincione G. A survey on data integration for multi-omics sample clustering. Neurocomputing. 2022;488:494–508.
19. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2018;46:10546–62.
20. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016;17:S15.
21. Chateigner A, Lesage-Descauses M-C, Rogier O, Jorge V, Leplé J-C, Brunaud V, et al. Gene expression predictions and networks in natural populations supports the omnigenic theory. BMC Genomics. 2020;21:416.
22. Sow MD, Rogier O, Lesur I, Daviaud C, Mardoc E, Sanou E, et al. Epigenetic Variation in Tree Evolution: a case study in black poplar (*Populus nigra*). preprint. Evol Biol; 2023. BIORXIV:2023.07.16.549253.
23. Song M, Greenbaum J, Luttrell J, Zhou W, Wu C, Shen H, et al. A review of integrative imputation for multi-omics datasets. Front Genet. 2020;11:570255.
24. Walach J, Hron K, Filzmoser P. Data Normalization and Scaling: Consequences for the Analysis in Omics Sciences. In: Jaumot J, Bedia C, Tauler R, editors. Comprehensive Analytical Chemistry. Data Analysis for Omics Sciences: Methods and Applications. Elsevier; 2018. p. 165–96.
25. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab. 2012;10:486–9.
26. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733–9.
27. Duruflé H, Selmani M, Ranocha P, Jamet E, Dunand C, Déjean S. A powerful framework for an integrative study with heterogeneous omics data: from univariate statistics to multi-block analysis. Brief Bioinform. 2021;22:bbaa166.
28. Wu C, Lin M, Chen F, Chen J, Liu S, Yan H, et al. Homologous drought-induced 19 proteins, PtDi19-2 and PtDi19-7, enhance drought tolerance in transgenic plants. Int J Mol Sci. 2022;23:3371.
29. Liu W-X, Zhang F-C, Zhang W-Z, Song L-F, Wu W-H, Chen Y-F. Arabidopsis Di19 functions as a transcription factor and modulates PR1, PR2, and PR5 expression in response to drought stress. Mol Plant. 2013;6:1487–502.
30. Wang Y, Huang X, Li F, Jia X, Jia N, Fu J, et al. Serum-integrated omics reveal the host response landscape for severe pediatric community-acquired pneumonia. Crit Care. 2023;27:79.
31. Zaghum MJ, Ali K, Teng S. Integrated Genetic and Omics Approaches for the Regulation of Nutritional Activities in Rice (Oryza sativa L.). Agriculture. 2022;12:1757.
32. Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience. 2022;25:103798.
33. Hesami M, Alizadeh M, Jones AMP, Torkamaneh D. Machine learning: its challenges and opportunities in plant system biology. Appl Microbiol Biotechnol. 2022;106(9–10):3507–30.
34. Sathyanarayanan A, Gupta R, Thompson EW, Nyholt DR, Bauer DC, Nagaraj SH. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. Brief Bioinform. 2020;21:1920–36.
35. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
36. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018.
37. Ibe OC. Chapter 8 – Introduction to Descriptive Statistics. In: Fundamentals of Applied Probability and Random Processes (Second Edition), Academic Press; 2014;253–74.
38. Ibe OC. Chapter 8 – Introduction to Inferential Statistics. In: Fundamentals of Applied Probability and Random Processes. (Second Edition), Academic Press. 2014:275–305.
39. Van De Schoot R, Depaoli S, King R, Kramer B, Märtens K, Tadesse MG, et al. Bayesian statistics and modelling. Nat Rev Methods Primer. 2021;1:1.
40. Jia W, Sun M, Lian J, Hou S. Feature dimensionality reduction: a review. Complex Intell Syst. 2022;8:2663–93.
41. Taverna F, Goveia J, Karakach TK, Khan S, Rohlenova K, Treps L, et al. BIOMEX: an interactive workflow for (single cell) omics data interpretation and visualization. Nucleic Acids Res. 2020;48:W385–94.
42. Eicher T, Kinnebrew G, Patt A, Spencer K, Ying K, Ma Q, et al. Metabolomics and multi-omics integration: a survey of computational methods and resources. Metabolites. 2020;10:202.
43. Ma A, McDermaid A, Xu J, Chang Y, Ma Q. Integrative methods and practical challenges for single-cell multi-omics. Trends Biotechnol. 2020;38:1007–22.
44. Rautenstrauch P, Vlot AHC, Saran S, Ohler U. Intricacies of single-cell multi-omics data integration. Trends Genet. 2022;38:128–39.
45. Stanojevic S, Li Y, Garmire LX. Computational Methods for Single-Cell Multi-Omics Integration and Alignment. ArXiv220106725 Q-Bio. 2022.
46. Wei Z, Zhang Y, Weng W, Chen J, Cai H. Survey and comparative assessments of computational multi-omics integrative methods with multiple regulatory networks identifying distinct tumor compositions across pan-cancer data sets. Brief Bioinform. 2021;22:bbaa102.
47. Wu P. Dimension reduction methods for nonlinear association analysis with applications to omics data. 2021.
48. Patruno L, Maspero D, Craighero F, Angaroni F, Antoniotti M, Graudenzi A. A review of computational strategies for denoising and imputation of single-cell transcriptomic data. Brief Bioinform. 2020;22(4):bbaa222.
49. Gupta S, Gupta A. Dealing with noise problem in machine learning datasets: a systematic review. Procedia Comput Sci. 2019;161:466–74.
50. Daviaud C, Renault V, Mauger F, Deleuze J-F, Tost J. Whole-Genome Bisulfite Sequencing Using the Ovation Ultralow Methyl-Seq Protocol. In: Tost J, editor. DNA Methylation Protocols. New York: Springer New York; 2018. 83–104.
51. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10:232.
52. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
53. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25.

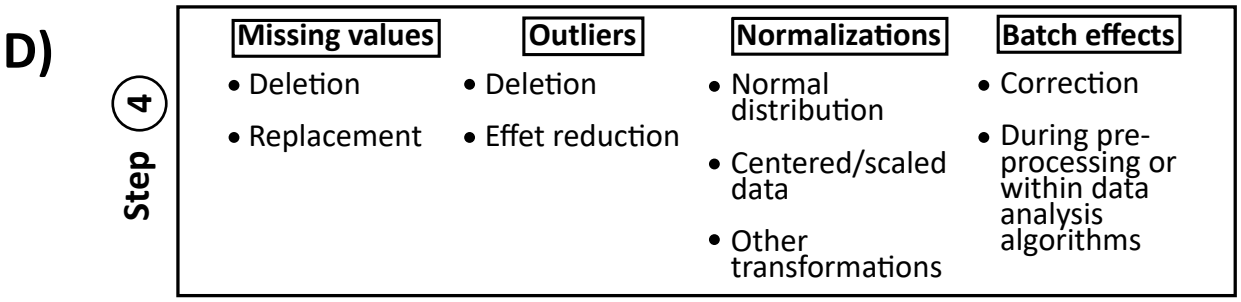
Publisher's Note

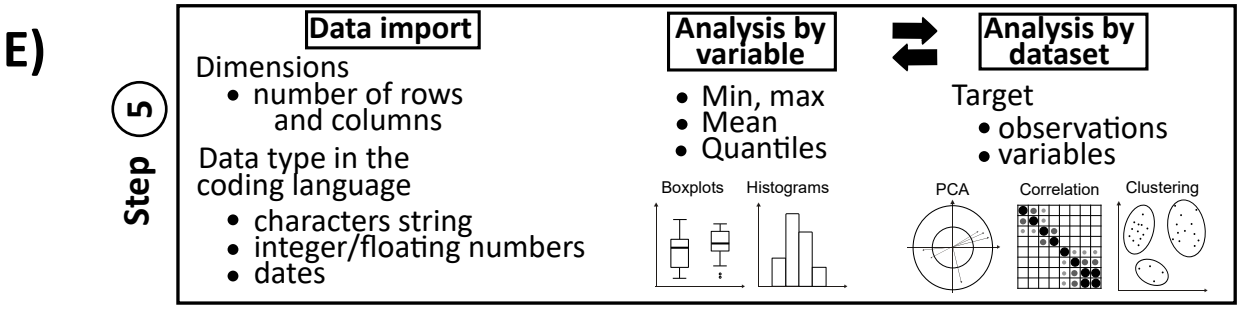

1
Data matrix
2
Scientific question
3
Tools selection
4
Data pre-processing
5
Preliminary analysis
6
Data integration


A)
Step 1
Variables
Genes


B)
Step 2
Description
Selection
Prediction
?


C)
Step 3
Tools characteristics
Coding language
Open source
Documentation
Updates
Methods characteristics
Descriptive statistics
Bayesian statistics
Posterior
Prior
Data
Networks
Similarity based
Dimension reduction
Neural networks
Data characteristics
Data type
quantitative
qualitative
Dimensions
number of rows and columns
Hypothesis
independence
normality
linearity


D)
Step 4
Missing values
Deletion
Replacement
Outliers
Deletion
Effet reduction
Normalizations
Normal distribution
Centered/scaled data
Other transformations
Batch effects
Correction
During pre-processing or within data analysis algorithms


E)
Step 5
Data import
Dimensions
number of rows and columns
Data type in the coding language
characters string
integer/floating numbers
dates
Analysis by variable
Min, max
Mean
Quantiles
Boxplots
Histograms
Analysis by dataset
Target
observations
variables
PCA
Correlation
Clustering

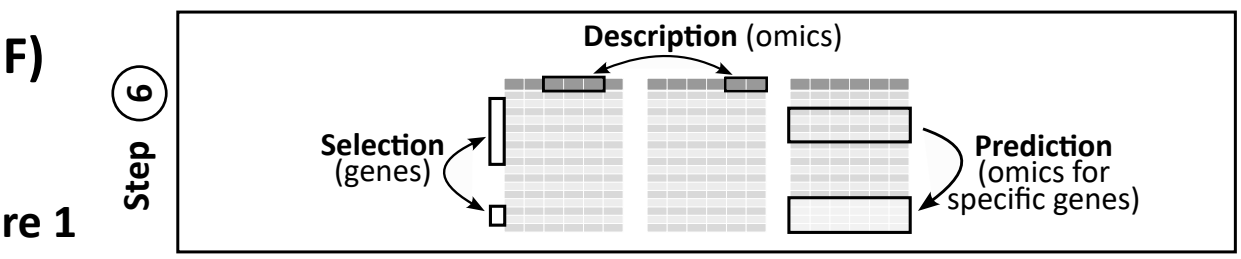

F)
Step 6
Description (omics)
Selection (genes)
Prediction (omics for specific genes)

Figure 1

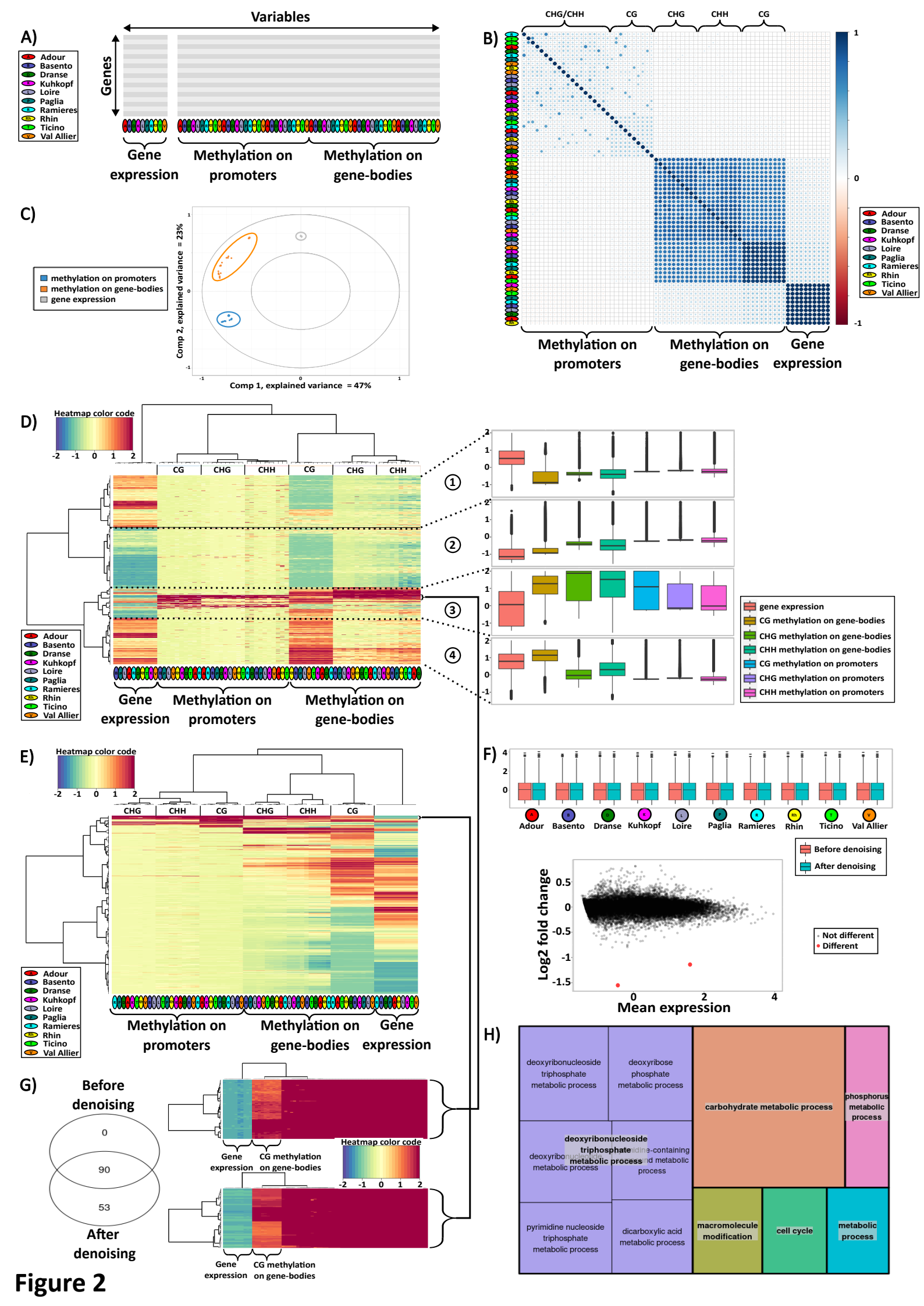

A)
Variables
Genes
Adour
Basento
Dranse
Kuhkopf
Loire
Paglia
Ramieres
Rhin
Ticino
Val Allier
Gene expression
Methylation on promoters
Methylation on gene-bodies
B)
CHG/CHH
CG
CHG
CHH
CG
1
0
-1
Methylation on promoters
Methylation on gene-bodies
Gene expression
C)
methylation on promoters
methylation on gene-bodies
gene expression
Comp 2, explained variance = 23%
Comp 1, explained variance = 47%
D)
Heatmap color code
-2 -1 0 1 2
CG CHG CHH CG CHG CHH
Gene expression
Methylation on promoters
Methylation on gene-bodies
1
2
3
4
gene expression
CG methylation on gene-bodies
CHG methylation on gene-bodies
CHH methylation on gene-bodies
CG methylation on promoters
CHG methylation on promoters
CHH methylation on promoters
E)
Heatmap color code
-2 -1 0 1 2
CHG CHH CG CHG CHH CG
Methylation on promoters
Methylation on gene-bodies
Gene expression
F)
Adour
Basento
Dranse
Kuhkopf
Loire
Paglia
Ramieres
Rhin
Ticino
Val Allier
Before denoising
After denoising
Log2 fold change
Mean expression
Not different
Different
G)
Before denoising
0
90
53
After denoising
Gene expression
CG methylation on gene-bodies
Heatmap color code
-2 -1 0 1 2
H)
deoxyribonucleoside triphosphate metabolic process
deoxyribose phosphate metabolic process
carbohydrate metabolic process
phosphorus metabolic process
deoxyribonucleoside triphosphate metabolic process
pyrimidine nucleoside triphosphate metabolic process
dicarboxylic acid metabolic process
macromolecule modification
cell cycle
metabolic process

Figure 2

5.3 Synthèse des principaux résultats

L'article Mardoc et al. (2024) présente d'une part des réponses à des questionnements méthodologiques, et d'autre part des réponses à des questionnements scientifiques en complément de celles présentées dans Sow et al. (2023).

En terme de résultats méthodologiques, nous avons proposé de nouveaux développements pour l'intégration des données omiques chez les plantes.

1. Un nouveau tutoriel en 6 étapes a ainsi été proposé pour, à partir de la question biologique choisie (et des stratégies d'intégration associées : description, sélection, prédiction) et des données à disposition, mener une analyse intégrative des données omiques pas à pas. Ces étapes sont donc *i*- la préparation des données sous forme matricielle, *ii*- la détermination de la question biologique et des stratégies d'intégration, *iii*- le choix de l'outil intégratif, *iv*- le pré-traitement des données, *v*- l'analyse préliminaire des données et *vi*- l'intégration multi-omiques. De manière générale, cette procédure a montré une bonne adaptabilité aux différentes spécificités sur la problématique et les données peuplier aboutissant à de nouvelles connaissances biologiques.
2. Une nouvelle fonction cimDiablo_v2 a aussi été implémentée puis testée sur les données peuplier pour étudier l'impact de la méthylation de l'ADN sur l'expression des gènes sur des données dont le nombre de lignes, aussi couramment appelées observations et correspondant ici à des gènes, est largement supérieur au nombre de colonnes, à savoir ici des variables omiques de méthylation et d'expression. Les nouveautés de cette fonction par rapport à celles de `mixOmics` sont, premièrement, la représentation simultanée des lignes et colonnes des données pour les résultats de la régression PLS multi-blocs non discriminante, et deuxièmement la possibilité de "débruiter" les données, c'est-à-dire de ne conserver qu'une partie de la variabilité biologique partagée entre plusieurs variables omiques. Cette intégration multi-omiques, présentée dans l'article Mardoc et al. (2024) dont je suis premier auteur, a participé à compléter l'analyse uni-et bi-variée décrite dans l'article Sow et al. (2023) dont je suis co-auteur.

Le code de l'analyse des données peuplier en suivant les étapes du tutoriel ainsi que la fonction cimDiablo_v2 sont accessibles au lien suivant :
`https://forgemia.inra.fr/umr-gdec/omics-integration-on-poplar`

En terme de résultats biologiques, l'analyse intégrative a été utilisée dans deux des trois grands questionnements identifiés durant cette thèse, plus précisément pour 1- décrire les interactions omiques (impact de la méthylation sur l'expression des gènes) et 2- sélectionner des gènes au profil omique particulier dans leurs interactions omiques (fortes valeurs de méthylation et/ou expression), à l'échelle des différentes populations de peuplier (*i.e.* peupliers d'origines géographiques différentes). Nous avons ainsi montré :

1. qu'à l'échelle du génome entier, il n'y a pas de lien systématique entre la méthylation et

l'expression des gènes. Cependant, certains gènes montrent des profils contrastés pour la méthylation et l'expression suggérant un possible contrôle de l'expression de ces gènes par le biais de la méthylation de l'ADN.

2. que l'analyse uni-et bi-variée dans Sow et al. (2023) permet d'identifier des fonctions biologiques (de gènes) contrôlées spécifiquement par la méthylation et discriminant des populations d'origines géographiques différentes.
3. que l'analyse multi-variée complémentaire de Mardoc et al. (2024) a également identifié des gènes dits *master regulators-drivers*, c'est-à-dire des gènes régulés par méthylation / expression pour l'ensemble des populations de peupliers, indépendamment de leur origine géographique.

Ces résultats, décrits dans les deux articles joints au manuscrit (Sow et al. (2023) et Mardoc et al. (2024)), sont discutés en détails dans la section suivante.

5.4 Discussion

L'interprétation des résultats de l'article Mardoc et al. (2024) est axée selon deux orientations, à savoir les résultats méthodologiques (tutoriel, `mixOmics` et cimDiablo_v2) et scientifiques (interactions entre méthylation et expression) au regard du système biologique étudié, *i.e.* des populations de peupliers de contextes géographiques (et donc de conditions pédoclimatiques) différentes.

5.4.1 Sur les résultats méthodologiques

5.4.1.1 Le tutoriel : théorie et application aux données peuplier

Le tutoriel a été proposé à destination des biologistes non-mathématiciens, pour les aider à mener correctement l'ensemble des étapes préliminaires nécessaires à une intégration multi-omiques efficace, c'est-à-dire répondant à une question biologique déterminée. La difficulté majeure a été de concevoir une procédure à la fois complète, précise, concise et adaptable à la majeure partie des projets d'intégration de données omiques, comme décrit dans la section introductive du manuscrit (Chapitre 4 Pages 43 à 47). Il est pour l'instant difficile de connaître l'impact réel de cette procédure étant donné sa très récente publication. Cependant, son utilisation sur les données peuplier offre des premiers éléments de réponse et de discussion.

Le tutoriel étant structuré en six étapes, voici quelques remarques critiques sur chacune d'elles :

— **Données sous forme matricielle** : pour commencer, nous définissons le cadre pour lequel le tutoriel a été développé, c'est-à-dire pour des données matricielles avec de nombreux gènes en lignes et quelques variables en colonnes. Toutefois, nous conseillons vivement de produire les données après avoir déterminé la question biologique et les

outils à utiliser, afin notamment de produire un nombre suffisant de données (répétitions biologiques/techniques) pour obtenir des résultats statistiquement significatifs. De plus, le tutoriel peut aussi être testé sur d'autres jeux de données, notamment pour quelques individus en lignes et de multiples variables omiques en colonnes (*c.f.* Chapitre 6).

— **Question biologique** : l'intégration multi-omiques est réalisée dans le but de répondre à une question biologique précise. Dans le tutoriel présenté, nous avons alors considéré qu'une analyse intégrative était associée à un outil d'intégration choisi pour répondre à une seule question biologique. Cependant, il arrive qu'une même analyse intégrative soit menée pour répondre à plusieurs questions en utilisant plusieurs outils, comme montré sur les analyses de la thèse. De plus, la question initialement définie peut s'adapter en cours d'analyse, par exemple si le pré-traitement des données ou les analyses préliminaires révèlent de nouvelles informations majeures sur les données.

— **Sélection de l'outil intégratif** : cette étape a pour but de choisir parmi de nombreux outils d'intégration un qui soit adapté à la fois à la question biologique, aux données à disposition, mais aussi aux préférences de l'utilisateur. Dans le cas d'outils développés dans le langage de programmation R, une table de 13 outils fréquemment cités et non exhaustive est jointe au tutoriel, aidant à choisir parmi quelques outils génériques ayant déjà fait leurs preuves. Toutefois, l'ensemble des critères présentés dans cette partie du tutoriel est suffisamment générique pour s'appliquer à d'autres outils, au-delà des 13 outils plus particulièrement décrits et discutés dans ce travail de thèse.

— **Pré-traitement des données** : l'étape de pré-traitement des données est souvent considérée comme la plus consommatrice en temps, puisqu'elle implique de nettoyer les données de leurs valeurs manquantes et aberrantes, de corriger les effets *batch* et d'appliquer d'éventuelles normalisations sur ces données. Il faut alors dans un premier temps identifier toutes ces caractéristiques, puis déterminer la meilleure stratégie pour les traiter, ce qui nécessite de multiples échanges entre les différents acteurs du projet. De plus, bien qu'aucun ordre d'exécution de ces différentes étapes de pré-traitement ne soit indiqué dans le tutoriel, nous conseillons de mener en premier la délétion des données comportant beaucoup de valeurs manquantes, puis les autres pré-traitements et enfin l'imputation des valeurs manquantes restantes.

— **Analyses préliminaires** : avant d'analyser simultanément les différents jeux de données, il est nécessaire de prendre le temps d'étudier les données séparément afin d'identifier différents types d'informations qui peuvent réapparaître dans l'analyse intégrative, voire pourraient biaiser cette dernière. Dans un premier temps, il est important de vérifier que les données sont du type souhaité, comme par exemple des chiffres, chaînes de caractères, dates, *etc.* L'importation des données est aussi l'occasion de choisir le nombre de jeux de données, généralement sous forme matricielle, et de déterminer dans ce cas le nombre de lignes et colonnes de ces derniers, même s'il est possible d'opé-

rer ces changements plus tard selon les pré-requis des méthodes d'analyses employées. Concernant les analyses préliminaires, nous conseillons d'étudier les données à la fois par variables et par jeux complets de données afin d'avoir une compréhension fine des données aux différentes échelles permises par le système expérimental dans lequel les données ont été générées. Enfin, le tutoriel présente certaines des manières les plus classiques de mener ces analyses, mais il existe d'autres approches fréquemment utilisées, par exemple les tests statistiques, régressions linéaires, nuages de points, *etc.* ainsi qu'éventuellement des méthodes plus spécifiques aux données acquises. Cette étape est centrale pour alimenter et revenir à l'étape précédente (Pré-traitement) pour adopter la stratégie adéquate concernant les données manquantes et aberrantes, les effets *batch* et le choix de la normalisation.

— **Intégration des données omiques** : enfin, l'intégration de différents jeux de données omiques sous forme matricielle est possible, à partir de l'outil intégratif sélectionné sur les données pré-traitées pour lesquelles des analyses préliminaires ont été effectuées, le but étant de répondre à la question biologique initialement spécifiée.

Dans la littérature, d'autres tutoriels ont été proposés ces dernières années pour préparer l'intégration de données omiques. C'est par exemple le cas dans Taverna et al. (2020); Sathyanarayanan et al. (2020) et Li et al. (2021a) qui présentent des procédures en trois étapes, à savoir de commencer par produire les données omiques, puis les pré-traiter et enfin mener les analyses simples et intégratives. Il semble donc que la définition de la question biologique et la sélection de l'outil intégratif ne soient pas clairement indiqués dans ces travaux et présentent un ajout notable de notre tutoriel. À notre connaissance, la seule autre procédure débutant clairement par le choix de la question biologique est celle de Duruflé et al. (2021) (Figure 5.3.A). Celle-ci comporte 12 étapes, dont certaines facultatives, afin de s'adapter au contexte. Ce tutoriel est présenté sous forme de boucle entre les étapes d'analyses et celles d'interprétation biologique des résultats et d'élaboration de nouvelles hypothèses et questions biologiques. Dans notre cas, nous avons décidé de simplifier la procédure en limitant le nombre d'étapes et en présentant un parcours linéaire afin de nous concentrer sur la réalisation d'une intégration multi-omiques (Figure 5.3.B).

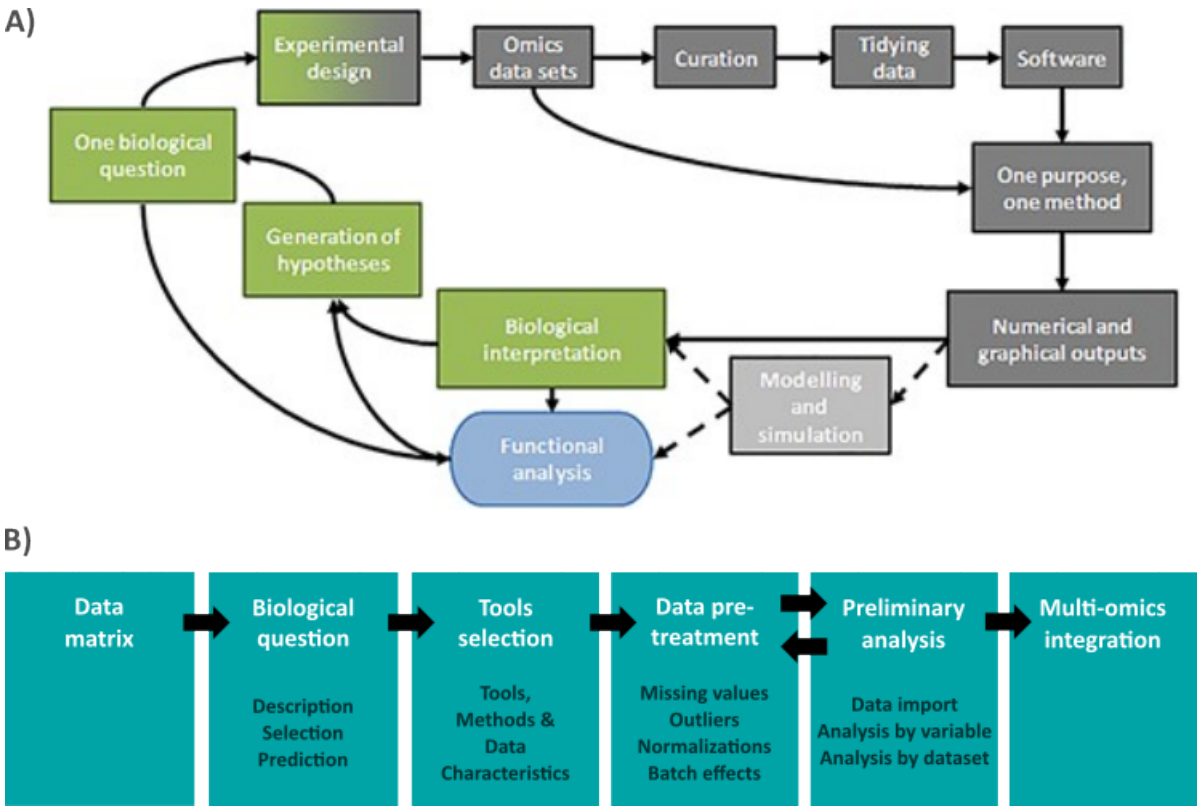


FIGURE 5.3 Comparaison de tutoriels d'intégration multi-omiques
A- Figure et légende de Duruflé et al. (2021). *Workflow for our multi-omics integrative studies. The different parts of this article are represented with grey boxes and the green boxes close the workflow with biological concepts. The workflow converges towards the functional analysis required to validate the whole study.* **B-** Synthèse du tutoriel et légende de Mardoc et al. (2024). *The tutorial presents the different steps to conduct multi-omics integrations : Step 1 - Constructing the genomics data matrix. Step 2 - Defining a clear and precise question of interest where biological questions concern describing omics interactions and interplay, selecting biomarkers specific to a trait, or predicting phenotypes from omics ; Step 3 - Selecting the tool, by considering tools' specificities such as their coding language, the accessibility or not to their source code, the quality of their documentation and frequency of their updates, their methods' main concepts and data requirements ; Step 4 - Preprocessing data, especially to remove or impute missing values, identify then remove outliers or reduce their impact, correct batch effects, normalize data ; Step 5 - Pre-analyzing data, by first importing data in the expected format with the right dimensions and types, then analyzing them by variable (univariate analysis) and dataset (multivariate analysis) to reveal major insights. Step 6 - Genomics data integration for data description, selection and prediction.*

La Figure 5.4 résume comment le tutoriel a été utilisé sur les données peuplier avec le paquet R `mixOmics` et sa fonction cimDiablo. Les détails étape par étape sont illustrés dans l'article Mardoc et al. (2024). Cependant, avant l'intégration des données omiques de peuplier, un développement d'une nouvelle version de cimDiablo, nommée cimDiablo_v2, était nécessaire pour être utilisée sur ces données multi-blocs exclusivement quantitatives.

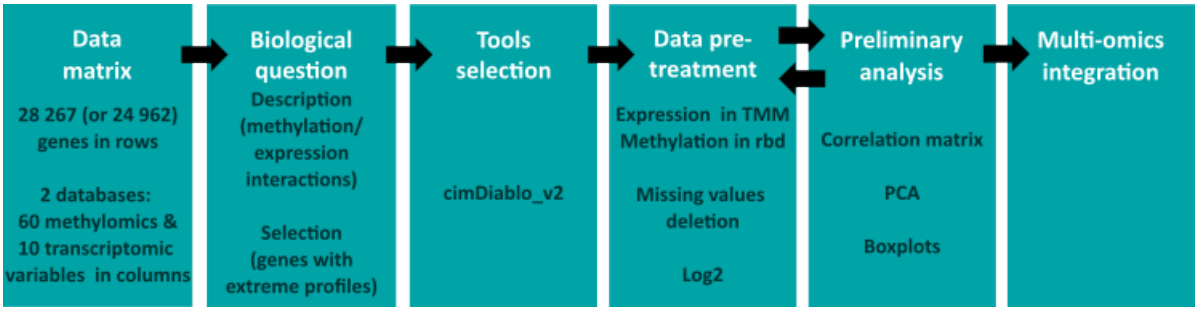


FIGURE 5.4 Application du tutoriel sur les données peuplier
Les données sont composées de 28 267 gènes en lignes (et 24 962 lorsqu'aucune valeur manquante initialement présente dans les données n'est tolérée) pour un jeu de données de 60 variables de méthylation et un de 10 variables d'expression. Les questions abordées sont la description des interactions entre la méthylation et l'expression des gènes à l'échelle du génome, et la sélection de gènes au profil de méthylation et d'expression particulier. L'outil intégratif choisi est cimDiablo_v2. Les données d'expression sont normalisées en *TMM* et celles de méthylation en *rbd*, les gènes comportant des valeurs manquantes sont supprimés pour la sélection des gènes au profil extrême, une log-transformation est appliquée à l'ensemble des données. Les principales analyses préliminaires ont été réalisées par jeu de données avec des matrices de corrélation et ACP et par variable avec des *boxplots*.

5.4.1.2 La fonction cimDiablo_v2 inspirée du paquet R mixOmics : théorie et application aux données peuplier

Les origines du développement de cimDiablo_v2 et ses deux nouveautés

Parmi les outils intégratifs développés en R, l'un des plus utilisés est le paquet `mixOmics`, proposant différentes méthodes de régression associées à des fonctions graphiques et d'optimisation des paramètres. `mixOmics` semble donc tout indiqué pour s'adapter à divers types d'intégrations, et notamment pour répondre à différentes questions biologiques autant en termes de description, de sélection et de prédiction.

Toutefois, bien que ces fonctions graphiques et d'optimisation de paramètres soient applicables à la plupart des méthodes régressives de `mixOmics`, seules quelques fonctions graphiques et aucune d'optimisation ne sont proposées pour la version multi-blocs non discriminante de la régression PLS appelée *block.pls* (MB-PLS), qui est pourtant la méthode adaptée à un grand nombre de cas dont ceux abordés durant sur cette thèse. C'est dans ce contexte que nous nous sommes intéressés à la fonction graphique *cimDiablo* développée pour la fonction *block.splsda*, affichant simultanément les lignes (observations, *i.e.* les gènes ou les individus) et colonnes (variables, *i.e.* les données omiques) des données. Cette caractéristique offre la possibilité théorique de s'adapter à tout type de questions, à savoir de décrire, sélectionner et prédire des variables (colonnes) et observations (lignes) en considérant simultanément les relations entre lignes et relations entre colonnes avec ses dendrogrammes ainsi que les relations entre lignes et colonnes avec sa *heatmap*.

J'ai développé une nouvelle version de cette fonction, appelée cimDiablo_v2, utilisable sur les résultats de la *block.pls*. La première différence notable avec *cimDiablo* concerne les données de sortie Y, étant pour la *Multi-Blocks PLS Discriminant Analysis* (MB-PLSDA)

un vecteur de données qualitatives, et une matrice de données quantitatives pour la version non discriminante MB-PLS. Ainsi, la matrice Y des MB-PLS est concaténée aux matrices X dans cimDiablo_v2, et la colonne représentant Y dans *cimDiablo* disparaît dans cimDiablo_v2. Deuxièmement, *cimDiablo* se contente de représenter les données initiales, centrées et réduites. La réduction de dimension menée par la régression n'est alors pas utilisée. Au contraire, avec cimDiablo_v2, il est possible de prendre en compte le résultat de la régression MB-PLS en activant le mode "débruitage", dans l'objectif d'identifier pour des observations (ici des gènes) les interactions omiques (ici l'expression et la méthylation) conservées entre "sources biologiques" distinctes (ici entre les différentes populations de peuplier étudiées).

Le "débruitage" de cimDiablo_v2 représente les données initiales auxquelles ont été soustraites l'information non capturée par les composantes lors de la régression, c'est-à-dire l'information résiduelle. L'hypothèse du "débruitage" de cimDiablo_v2 est alors que ce résidu correspond à du "bruit" (plus précisément une variabilité omique non recherchée), la gestion du "bruit" étant un challenge majeur lors des analyses biologiques (Patruno et al., 2020; Gupta and Gupta, 2019). Ainsi, cela ne signifie pas que le résidu/le "bruit" (ici les interactions omiques spécifiques des populations et non partagées entre elles) n'a pas d'intérêt biologique en soi (ceci est clairement étudié par les approches uni-bi-variées de Sow et al. (2023)), mais qu'il contient de la variabilité non considérée pour répondre à la question pour laquelle la MB-PLS est utilisée, cimDiablo_v2 "débruite" donc les données en supprimant ce résidu. Chez le peuplier, les données sont "débruitées" au profit de la recherche de *master regulators-drivers*, c'est-à-dire des gènes dont la régulation omique est indépendante des variabilités observées entre les populations. Le fonctionnement théorique de ce "débruitage", son impact concret, et sa validation sur les données peuplier sont développés dans la suite de cette section.

Le fonctionnement théorique du "débruitage" de cimDiablo_v2

La Figure 5.5 représente 3 notions majeures dans la compréhension du fonctionnement de `mixOmics` et de cimDiablo_v2, à savoir les notions de réduction de données, de factorisation matricielle et de "débruitage".

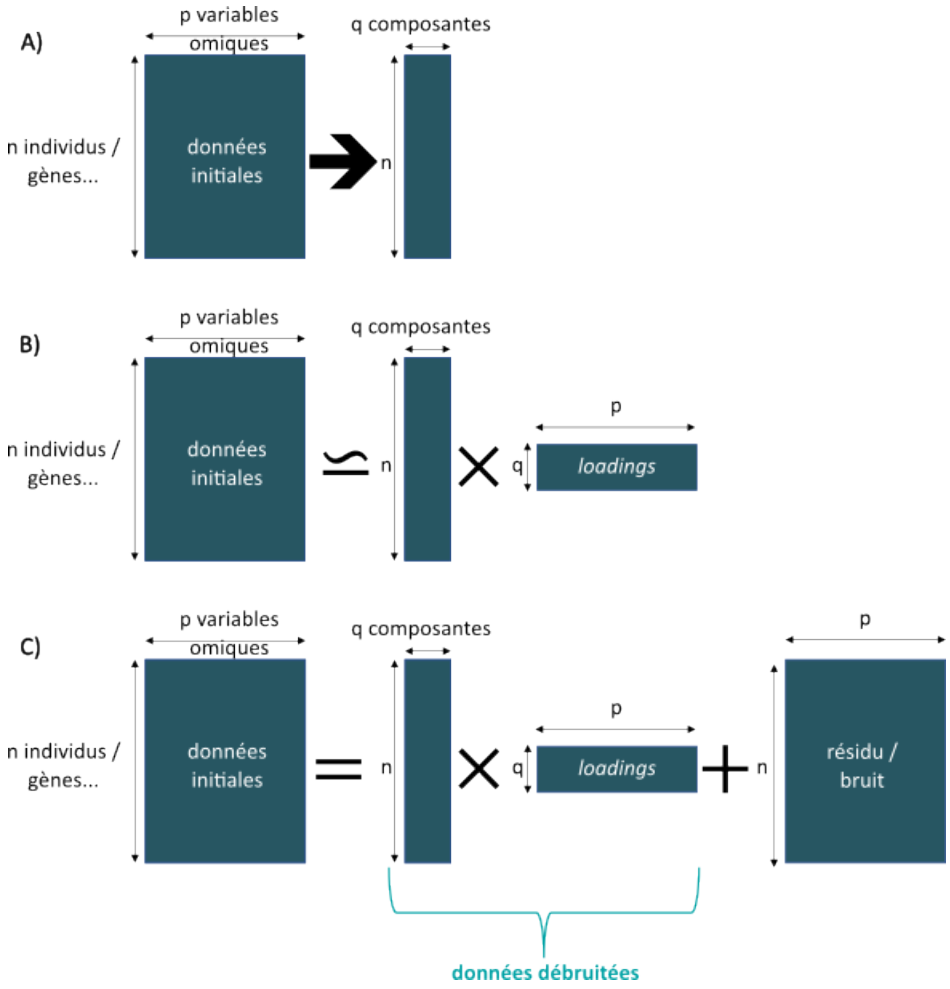


FIGURE 5.5 Illustration des principaux concepts utilisés dans mixOmics et cimDiablo_v2
Les principaux concepts sont illustrés sur un seul jeu de données théorique composé de n lignes (observations telles que des individus, des gènes, *etc.*) et p colonnes (variables omiques). **A-** la réduction de dimension représente le passage du jeu de données initial à une nouvelle matrice contenant approximativement la même information mais en un nombre restreint de colonnes (appelées composantes). L'information (ou variabilité) conservée est la variance s'il n'y a qu'un jeu de données (*cf.* ACP), la covariance ou corrélation pour deux jeux de données (PLS, CCA), ou des pseudo-corrélations dans le cas de la MB-PLS. Plus il y a de composantes et plus la variabilité du jeu de données est conservée. **B-** la factorisation matricielle est la manière pour réduire les données employée par `mixOmics` et cimDiablo_v2, l'objectif étant de factoriser les données initiales en deux matrices de dimensions inférieures, la première étant la matrice recherchée (la matrice des composantes) et la deuxième étant une matrice de poids (ou *loadings*) liant les composantes calculées aux variables initiales. **C-** lors de la factorisation matricielle, une partie de la variabilité des données initiales n'est pas conservée dans la matrice des composantes et se retrouve dans la matrice résiduelle. L'hypothèse de cimDiablo_v2 est que ce résidu correspond au "bruit" (variabilité non souhaitée masquant partiellement la variabilité d'intérêt biologique) contenu dans les données initiales, cimDiablo_v2 "débruite" alors les données en représentant graphiquement les données sans ces résidus (c'est-à-dire le produit matriciel de la matrice des composantes par celle des poids). Cette action est nommée dans cimDiablo_v2 "débruitage des données".

Le "débruitage" de cimDiablo_v2 est effectué à partir des sorties de la fonction *block.pls* (MB-PLS) de `mixOmics`, qui ne représentent qu'une faible partie de tous les vecteurs, matrices, valeurs, *etc.* utilisés durant la régression. Les deux seules sorties susceptibles d'être utilisées sont alors la matrice des composantes et la matrice des poids (*loadings*) pour chaque bloc (*i.e.* chaque jeu de données). Dans le cas où il n'y a qu'un bloc, c'est-à-dire pour l'ACP, le

produit de la matrice des composantes par celle des poids correspond aux données soustraites du résidu (Figure 5.5.C). En effet, à chaque étape, une nouvelle composante t et un nouveau vecteur des poids a sont calculés, et la nouvelle matrice X correspond à la précédente à laquelle est soustraite le produit vectoriel de la composante par les poids (Algorithme 1). Une fois le nombre de composantes souhaité atteint, la matrice restante correspond au résidu, et par reconstruction la matrice initiale peut s'exprimer ainsi :

$$X = t_1 \cdot a'_1 + t_2 \cdot a'_2 + \cdots + t_q \cdot a'_q + residual = T \cdot A' + residual$$
$$\Leftrightarrow denoised\_data = X - residual = T \cdot A'$$

avec t_i et a_i les i^{emes} vecteurs des composantes et poids, T et A la concaténation de ces vecteurs pour former la matrice des composantes et celle des poids, " $\cdot$ " le produit vectoriel ou matriciel, " $'$ " la transposée vectorielle ou matricielle et q le nombre total de composantes.

Algorithm 1: Pseudo-algorithme du fonctionnement de l'ACP de mixOmics. Issu de Lê Cao and Welham (2021).

INITIALISE the first component t using SVD(X), or a column of X
  FOR EACH dimension
    UNTIL CONVERGENCE of a
      (a) Calculate loading vector $a = X'/t't$ and norm a to 1
      (b) Update component $t = Xa/a'a$
    DEFLATE : current X matrix $\tilde{X} = X - ta'$ and update $X = \tilde{X}$

Pour le cas plus complexe des régressions MB-PLS, l'approche choisie dans cimDiablo_v2 est exactement la même pour chaque bloc, afin de reconstruire les données initiales à partir du produit matriciel des composantes par les poids. Cependant, les interactions entre composantes et poids sont plus complexes pour les régressions PLS que pour l'ACP. En effet, bien qu'il n'existe pas dans le livre référence de `mixOmics` Lê Cao and Welham (2021) de pseudo-algorithme pour la régression MB-PLS, son fonctionnement est très proche de celui de la PLS2 (Algorithme 2). Tout comme pour l'ACP, la PLS2 fonctionne de manière itérative, et à chaque itération 1 - calcule une composante et le vecteur de poids (*loadings*) associé pour chaque matrice (X pour les données explicatives, Y pour les données expliquées), 2- calcule d'autres vecteurs à partir des matrices de données, des composantes et des poids, et 3- soustrait aux matrices de données l'information contenue par ces différents vecteurs. En régression MB-PLS, les relations entre les différentes matrices de données, composantes et poids sont légèrement plus complexes mais le principe général reste le même.

Algorithm 2: Pseudo-algorithme du fonctionnement de la PLS2 de mixOmics. Issu de Lê Cao and Welham (2021).

INITIALISE the first set of components (t, u) as the first column of X and Y

FOR EACH dimension h

UNTIL CONVERGENCE of a

Calculate the vectors associated to X :

(a) Loading vector $a = X'u/u'u$ and norm a to 1

(b) Component $t = Xa/a'a$

Calculate the vectors associated to Y :

(c) Loading vector $b = Y't/t't$ and norm b to 1

(d) Component $u = Yb/b'b$

END CONVERGENCE

Calculate the regression coefficients :

$c = X't/t't$ related to the information from X

$d = Y't/t't \propto b$

$e = Y'u/u'u$ related to the information from Y

DEFLATION :

$\tilde{X} = X - tc'$

Regression mode : $\tilde{Y} = Y - td'$

Canonical mode : $\tilde{Y} = Y - ue'$

Increment to dimension $h + 1$

Ainsi, à la troisième grande étape de l'itération (**DEFLATION**), est soustrait aux données explicatives X (ex : méthylation)) le produit vectoriel $t \cdot c'$, avec t sa composante et c un vecteur relatif à X mais n'étant pas son vecteur des poids. De la même manière, en mode régressif (*i.e.* mode utilisé pour toutes les analyses de la thèse), on soustrait aux données expliquées Y (ex : expression) le produit vectoriel $t \cdot d'$, avec t la composante de X et d un vecteur proportionnel à b le vecteur des poids de Y. Cette étape est donc notablement différente de celle de l'ACP, mais aussi de celle de "débruitage" de cimDiablo_v2, pour laquelle les données sont reconstruites, avec pour chaque jeu de données, le produit vectoriel de sa composante par ses poids, soient $t \cdot a'$ pour X et $u \cdot b'$ pour Y.

En résumé, le fonctionnement théorique de la version actuelle de cimDiablo_v2 utilise la méthode régressive généralisant au cas multi-blocs l'Algorithme 2 pour le calcul des composantes et poids, mais s'approche plus de l'Algorithme 1 pour le "débruitage" des données. Il serait donc intéressant dans le futur de mieux comprendre les choix faits sur les vecteurs de la régression PLS2 puis MB-PLS en s'intéressant notamment aux travaux de Michel et Arthur Tenenhaus (Tenenhaus, 1998; Tenenhaus and Tenenhaus, 2011) pour pouvoir évaluer l'impact théorique de cette approche par rapport à celle de cimDiablo_v2. De plus, il serait aussi sou-

haitable d'implémenter une nouvelle version du "débruitage" de `mixOmics` avec ces choix de produits vectoriels, ce qui nécessiterait toutefois de préalablement modifier la fonction *block.pls* pour accéder à ces vecteurs. Cette perspective de thèse pourra être discutée avec les développeurs de `mixOmics`. Enfin, le "débruitage" présenté pouvant théoriquement être applicable aux ACP et PLS, il serait intéressant d'implémenter le "débruitage" sur ces méthodes et les tester sur différents jeux de données.

Conséquences théoriques et observées du "débruitage"

Gestion des valeurs manquantes : Une conséquence indirecte de ce "débruitage" des données concerne la gestion des valeurs manquantes. En effet, la fonction *block.pls* utilisée par cimDiablo_v2 est utilisable avec des valeurs manquantes bien que cela puisse engendrer une absence de convergence de l'algorithme notamment si la proportion de valeurs manquantes est trop abondante. Toutefois, malgré un certain nombre de valeurs manquantes dans les données initiales, les vecteurs des composantes et poids associés ne contiennent, par construction, aucune valeur manquante, et leur produit, correspondant aux données "débruitées", ne contient donc pas non plus de valeur manquante. Le "débruitage" de cimDiablo_v2 impute ainsi automatiquement les valeurs manquantes. L'utilisateur de cimDiablo_v2 doit donc décider s'il souhaite profiter de cette imputation ou s'il préfère gérer en amont les valeurs manquantes. Nous n'avons pas étudié en détail les bases méthodologiques de cette imputation durant la thèse, mais, afin d'en étudier les effets, nous avons utilisé des jeux de données avec des valeurs manquantes imputées par cimDiablo_v2 pour décrire les interactions entre méthylation et expression à l'échelle du génome, et avons utilisé les données sans valeur manquante pour sélectionner les gènes candidats discutés pour leur fonctions biologiques dans l'article Mardoc et al. (2024).

"Lissage" des données : Bien que l'objectif théorique du "débruitage" de cimDiablo_v2 soit de représenter exclusivement la variabilité biologique capturée par les composantes de la MB-PLS, son fonctionnement pratique est plus subtil. En effet, cimDiablo_v2 "lisse" les données par lignes (gènes), conservant ainsi la variabilité des gènes partagée par plusieurs colonnes (variables omiques). Ce "lissage" a un intérêt sur les données peuplier puisqu'il permet de se concentrer sur la variabilité des gènes par groupes de variables omiques, notamment par jeux de données omiques, et d'atténuer l'effet des différentes populations de peuplier, identifiant ainsi les interactions majeurs entre méthylation et expression à l'échelle de l'espèce et non des interactions spécifiques à chaque population (étudiées par les approches uni-bi-variées, Sow et al. (2023)). En d'autres termes, nous avons pu identifier les interactions omiques à l'échelle de l'espèce, c'est-à-dire partagées par tous les individus étudiés, on parle alors de *master regulators-drivers*. La Figure 5.6 illustre cet effet sur le groupe de gènes contenant les

fortes valeurs de méthylation des promoteurs, le "débruitage" accentuant les lignes de fortes valeurs de méthylation stables entre les différentes populations.

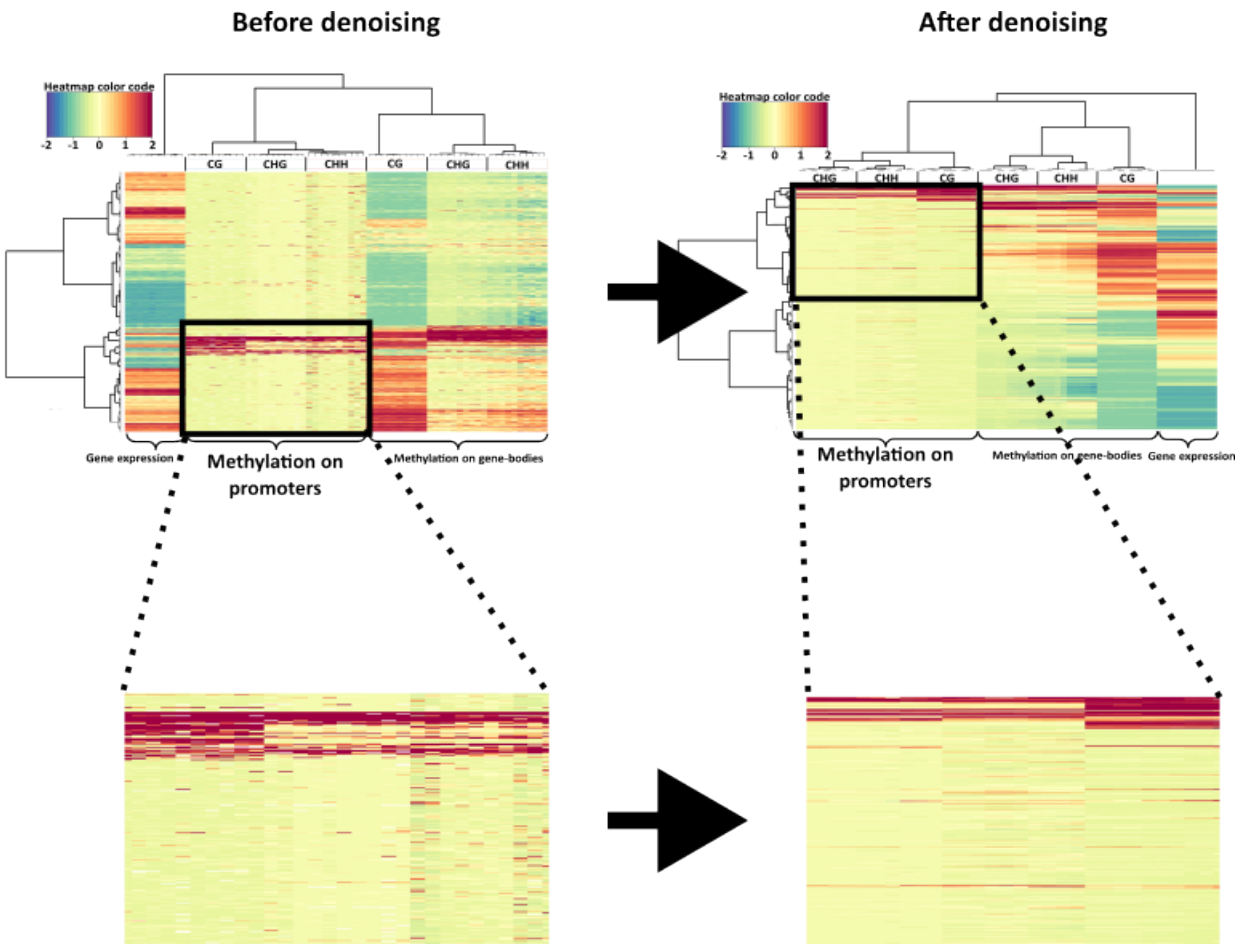


FIGURE 5.6 Effet du "débruitage" sous forme de "lissage" sur la méthylation des promoteurs des gènes de peuplier

Chaque ligne correspond à un gène, chaque colonne à une variable omique. Les lignes et colonnes sont regroupées selon les dendrogrammes issus du clustering hiérarchique avec la méthode Ward et la distance euclidienne. Les valeurs représentées selon le code couleur de la *heatmap* sont celles des données initiales, centrées et réduites, "débruitées", de nouveau centrées et réduites, et enfin coupées en [-2, 2] (toute valeur inférieure à -2 est fixée à -2, toute valeur supérieure à 2 est fixée à 2). Un zoom est effectué sur la figure avant "débruitage" (à gauche) puis après "débruitage" (à droite) sur le groupe de gènes contenant ceux avec une forte méthylation sur le promoteur.

Pour expliquer cette observation "visuelle" et en définir les bases mathématiques, il faut une fois de plus revenir au fonctionnement théorique des régressions PLS et du "débruitage" de cimDiablo_v2. Pour rappel, le "débruitage" consiste à calculer le produit des composantes par les vecteurs des poids. Or, pour chaque ligne du jeu de données initial (*i.e.* chaque gène), chaque composante contient une seule valeur égale à une somme pondérée de toutes les valeurs des variables sur cette ligne. Ainsi, au terme de la réduction de dimension, l'information par ligne est simplifiée de p variables initiales à q composantes, avec $p \gg q$. Finalement, lors du produit des composantes par les matrices de poids, les données sont reconstruites à partir de données simplifiées par lignes, il y a donc moins de variabilité par ligne après "débruitage" par rapport aux données initiales, d'où un "lissage" par lignes.

Notons que bien que ce "lissage" s'effectue par ligne, il ne "lisse" pas l'intégralité des valeurs de la ligne de la même manière, mais selon les poids portés à chaque variable et qui sont fortement liés à la similarité des variables représentée notamment par le dendrogramme des colonnes. On pourrait alors parler de "lissage" de données par lignes par groupes de variables, comme illustré en Figure 5.7.

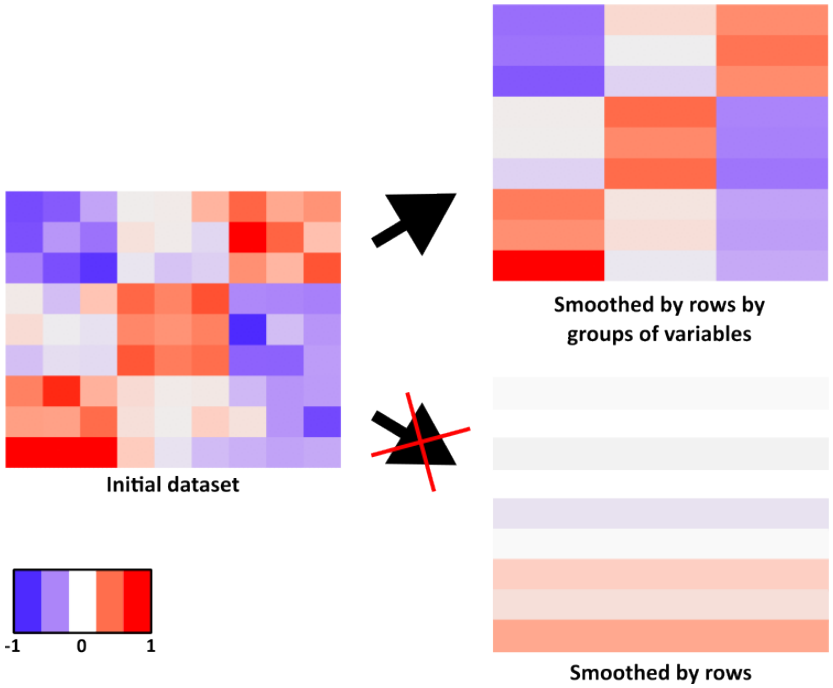


FIGURE 5.7 Schéma du "lissage" par ligne par groupe de variables.
À partir d'un jeu de données initial contenant ici 9 lignes (observations) et 9 colonnes (variables) avec différents profils de valeurs (légende en bas à gauche), le "débruitage" de cimDiablo_v2 "lisse" les valeurs par lignes lorsque des valeurs sont semblables avec celles des colonnes adjacentes (en haut à droite de la figure) et non sur toute la ligne indépendemment des similarités entre colonnes (en bas à droite de la figure).

Changement d'échelles : Une autre conséquence observable du "débruitage" de cimDiablo_v2 sur les données peuplier est le changement d'échelle principalement observé sur les données d'expression. En effet, avant de "débruiter" les données avec cimDiablo_v2, les données sont généralement centrées et réduites par la MB-PLS de `mixOmics` afin de comparer les données hétérogènes. Les variables omiques ont donc avant "débruitage" une moyenne de 0 et un écart-type de 1. Toutefois, l'étape de "débruitage" peut "dilater" les échelles des variables, que ce soit en agrandissant ou réduisant les échelles. Sur les données peuplier, c'est particulièrement le cas pour les données d'expression (Figure 5.8).

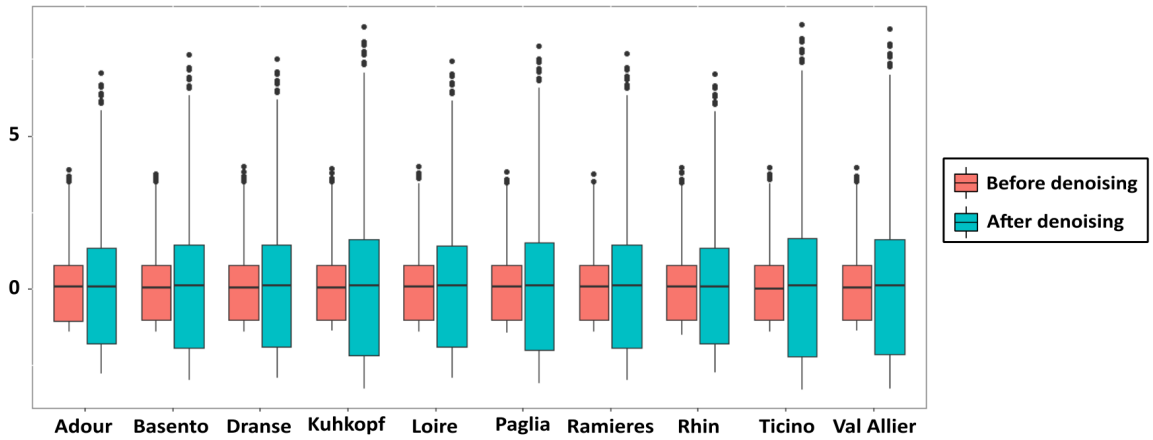


FIGURE 5.8 Distribution des variables d'expression avant contre après "débruitage"
Diagrammes en boîte des distributions des 10 variables d'expression (une par population de peuplier, axe des abscisses) avant (en rouge) contre après (en bleu) "débruitage" de cimDiablo_v2

Les causes précises de cette dilatation des échelles, qui semble principalement toucher les variables expliquées Y d'après plusieurs tests réalisés sur les données peuplier en échangeant le bloc d'expression de Y à X et l'un des blocs de méthylation de X à Y, sont encore à préciser et étudier.

En ce qui concerne les conséquences, elles peuvent être plus ou moins visibles selon les données. Pour les données peuplier, les échelles des variables d'expression étant fortement augmentées, le risque était sur les résultats graphiques de cimDiablo_v2 de ne plus voir aucune variation pour la méthylation si le *cutoff* en [-2,2] (*i.e.* valeurs supérieures à 2 fixées à 2, valeurs inférieures à -2 fixées à -2) n'était pas appliqué, ou d'avoir la majorité des valeurs d'expression fixées à 2 ou -2 et donc ne voir que peu de la variabilité des données d'expression en appliquant le *cutoff* en [-2,2]. Nous avons donc ajouté un paramètre facultatif dans cimDiablo_v2 pour centrer et réduire les données par variable après l'étape de "débruitage". Nous manquons encore de recul actuellement pour savoir si cette étape doit être systématiquement réalisée, c'est pourquoi nous conseillons pour l'instant de toujours utiliser cimDiablo_v2 avec et sans cette étape afin de déterminer celle qui a le plus d'intérêt par rapport à la question posée.

Mesure de l'effet "débruitage"

Le concept de "débruitage" de cimDiablo_v2 étant encore récent mais aussi complexe, aucune métrique de quantification de cet effet ne fait pour l'instant consensus. Deux approches ont alors été utilisées sur ces données, en comparant les valeurs des données et les listes de gènes sélectionnés avant et après "débruitage".

Approche 1 : Comparaison des valeurs : Ici, l'objectif est simplement de comparer les valeurs de chaque gène pour chaque variable omique avant et après "débruitage".

Une première idée a alors été de calculer directement la différence entre les matrices avant

et après "débruitage", qui sont représentées sous forme de *heatmaps* par cimDiablo_v2. Cependant, le "débruitage" n'impacte pas seulement les valeurs, mais aussi le clustering des lignes et des colonnes, les matrices ne sont alors pas directement comparables. La solution pour contourner ce problème est alors d'imposer un ordre fixe des lignes et colonnes. Cette approche a cependant l'inconvénient de rendre la *heatmap* plus difficile à interpréter biologiquement puisque les lignes et les colonnes ne sont plus regroupées selon leur ressemblance.

En suivant cette idée, et sachant que les données sont "débruitées" sous forme de "lissage" par lignes, il serait alors possible d'utiliser une métrique pour quantifier puis comparer la variabilité des données par lignes (ici des gènes) avant et après l'étape de "débruitage". Cependant, cette idée fait face à deux difficultés. Premièrement, quelle mesure de variabilité choisir (ex : l'écart-type de la variable, sa distance euclidienne, *etc.*) ? Deuxièmement, s'il existe pour une même ligne plusieurs groupes de valeurs, par exemple un groupe de très fortes valeurs de méthylation et un autre de très faibles valeurs d'expression, le "débruitage" s'effectue par groupe et les mesures de variabilité appliquées à l'ensemble de la ligne ne sont pas adaptées. Pour ces raisons, cette approche n'a pas été retenue.

Pour mesurer l'effet du "débruitage" des gènes, nous avons donc comparé les valeurs avant et après "débruitage" pour chaque gène et variable omique indépendamment des autres valeurs. Nous avons alors utilisé des *MA-plots* pour visualiser ces différences et déterminer les valeurs significativement différentes. La Figure 5.9 montre que globalement peu de gènes sont significativement différents (au seuil ± 1) pour les variables d'expression et de méthylation sur le corps du gène, et un peu plus pour la méthylation des promoteurs. À titre d'exemple, même pour la méthylation CHH des promoteurs qui est la plus impactée par le "débruitage", seuls 4.7% des 24 962 gènes sont significativement différents. Ainsi, malgré un changement bien visible du "débruitage" sur ces *MA-plots* comme sur les *heatmaps* de cimDiablo_v2, les données initiales sont peu altérées.

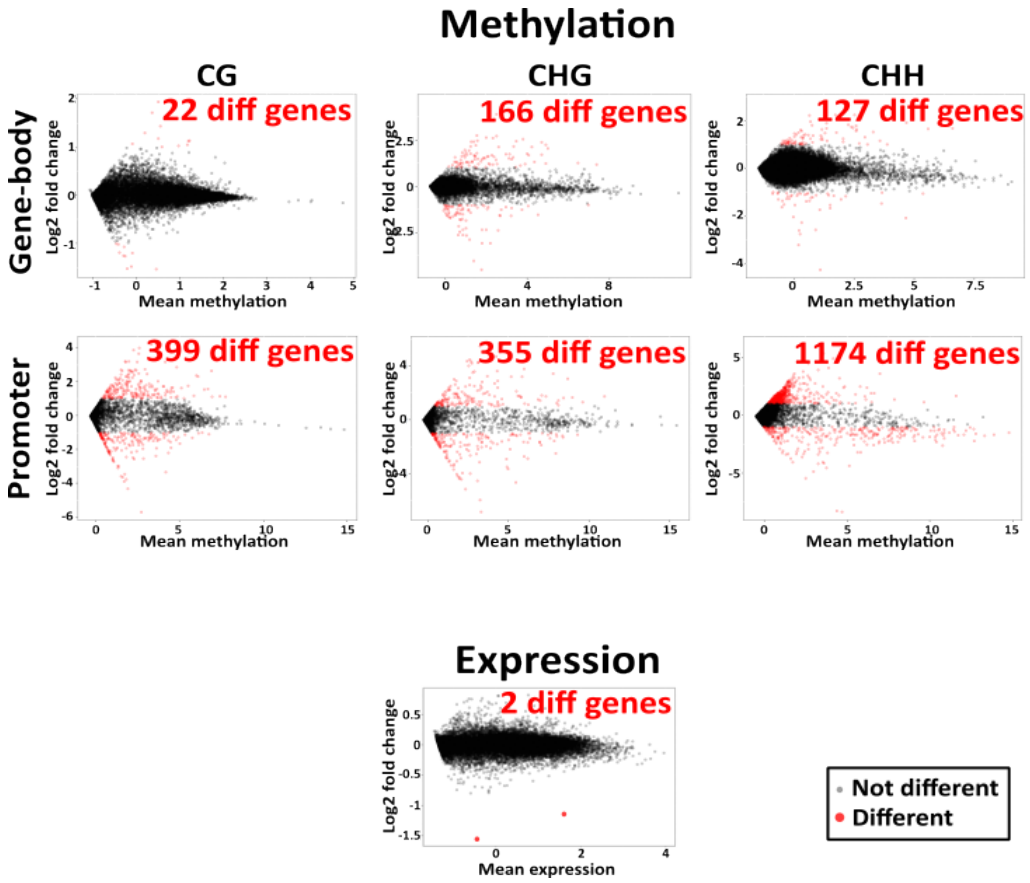


FIGURE 5.9 MA-plots de la méthylation et expression des gènes avant contre après "débruitage"

MA-plot pour chaque variable de la population de peuplier Adour. L'axe des abscisses représente la moyenne d'expression ou méthylation de chaque gène avant contre après "débruitage", et l'axe des ordonnées le log2 *fold change* (fonction logarithmique appliquée sur la valeur après "débruitage" divisée par celle avant "débruitage"). Avec un seuil de $\pm$ 1 sur l'axe des ordonnées, les gènes significativement différentiellement exprimés ou méthylés sont identifiés (illustrés en rouge).

Approche 2 : Comparaison des listes de gènes sélectionnés : Nous avons aussi sélectionné avant et après "débruitage" des gènes au profil extrême, à savoir des gènes de très forte méthylation pour toutes les variables méthylomiques et faible expression pour toutes les variables transcriptomiques, puis avons comparé les listes obtenues. 90 gènes ont ainsi été identifiés avant "débruitage", et ont tous été retrouvés après "débruitage". Toutefois, 53 gènes supplémentaires ont été identifiés après "débruitage", cette différence d'effectifs pouvant s'expliquer de deux manières. Premièrement, le "débruitage" offre la possibilité de révéler de nouveaux gènes dont la variabilité pour certaines des variables omiques empêchent l'identification avant "débruitage". Deuxièmement, l'identification de ces gènes peut aussi provenir du changement d'échelle et donc des seuils de sélection des extremums qui sont aussi des conséquences du "débruitage".

5.4.2 Sur les résultats biologiques

L'intégration des données peuplier a été réalisée dans le but à la fois de décrire les interactions entre méthylation et expression des gènes, et sélectionner des groupes de gènes aux profils multi-omiques extrêmes. Il est néanmoins nécessaire, avant cette intégration, de savoir s'il est possible de répondre à ces objectifs par des méthodes statistiques et computationnelles plus classiques. Sur ces données, il est en effet possible, avec les analyses bi-variées d'une variable de méthylation avec une variable d'expression, d'identifier des interactions omiques pour chaque couple de variables et des gènes au profil particulier pour ces variables. L'article Sow et al. (2023) a été mené en partie pour répondre à cette problématique par des méthodes uni-bi-variées.

L'analyse uni-bi-variée (Sow et al. (2023)) : Cet article étudie le rôle des variations épigénétiques dans l'évolution et l'adaptation des populations naturelles de peupliers noirs à l'échelle macro- (évènement ancien de polyploïdie) et micro- (adaptation locale à différentes zones géographiques d'origine) évolutive. Des données génétiques (SNPs), épigénétiques (méthylation de l'ADN) et transcriptomiques (expression des gènes) ont été utilisées pour répondre à ces questionnements scientifiques. Dans cet article, Sow et al. ont exploité ces données pour établir :

1. Le profil de méthylation du génome de peuplier selon la région génomique ciblée (promoteur ou corps du gène), et ce pour les trois contextes de méthylation (CG, CHG et CHH).

   Résultat obtenu : La méthylation en contexte CG du corps des gènes est plus élevée comparée aux promoteurs. Cependant pour les contextes CHG et CHH, la méthylation promoteur est similaire au corps du gène.

2. Le pan-épigénome de peuplier, c'est-à-dire la conservation ou spécificité de la méthylation entre les populations étudiées.

   Résultat obtenu : Une plus forte stabilité est observée pour la méthylation de l'ADN dans les contextes CG et CHG et plus variable entre les populations en contexte CHH.

3. Les différences d'expression et de méthylation entre les copies des gènes dupliqués.

   Résultat obtenu : Environ pour 3/4 des gènes dupliqués, il existe une différence significative en terme d'expression pour les copies d'une même paire. De manière similaire, la moitié des gènes dupliqués sont différentiellement méthylés entre les copies de la même paire de gènes.

4. La différenciation génétique et épigénétique des populations de peuplier.

   Résultat obtenu : La méthylation de l'ADN apparaît comme un possible marqueur de différenciation des populations à l'image des marqueurs génétiques (SNPs), structurant épi-génétiquement les individus en fonction de leur origine/proximité géographique.

5. L'impact de la méthylation de l'ADN sur l'expression des gènes.

   Résultat obtenu : Des groupes de gènes apparaissent faiblement méthylés et fortement exprimés (définis dans Sow et al. (2023) comme "*Hypo/Up*"), impliqués dans des fonctions basales cellulaires (ribosome) et de différenciation cellulaire (cambium), et d'autres fortement méthylés et faiblement exprimés ("*Hyper/Down*"), impliqués dans des fonctions immunitaires et de réponse au stress.

C'est dans ce cadre que je me suis intéressé à la régulation omique (méthylation-expression) à l'ensemble des populations naturelles de peupliers *via* une approche intégrative prenant en compte les interactions complexes entre les différentes variables omiques, les régions génomiques, les contextes de méthylation et les populations pour mieux appréhender la régulation génomique dans le complexe d'espèce peuplier.

L'approche multi-variée (Mardoc et al. (2024)) : L'analyse multi-variée va plus loin que les analyses uni- et bi-variées menées dans l'article Sow et al. (2023), puisqu'elle analyse simultanément les différentes variables des données peuplier, en considérant les différents types de données omiques mais aussi les différentes populations, contextes de méthylation et régions du gène sur laquelle la méthylation a lieu. cimDiablo_v2 hiérarchise alors ces effets et offre une vision plus globale de la variabilité des données. De plus, dans l'objectif d'identifier des gènes au profil extrême pour la plupart des variables, par exemple les *master regulators-drivers* à échelle de l'espèce, indépendamment des individus, cimDiablo_v2 est plus adapté que les analyses uni- et bi-variées puisque son "débruitage" atténue la variabilité propre à une variable/population pour conserver celle partagée entre plusieurs variables omiques/populations.

L'intégration multi-variée des données peuplier a ainsi été utilisée pour décrire les interactions entre variables omiques à l'échelle du génome (sans "débruitage") puis sélectionner des gènes au profil multi-omiques extrême (sans, puis avec "débruitage"). Les principaux résultats obtenus à partir de cimDiablo_v2 sont les suivants (Figure 5.10) :

1. L'intégration des différentes variables omiques a été utilisée pour hiérarchiser les facteurs de variabilité des données, montrant que l'effet majeur est le type de données omiques (méthylation, expression), puis la région génomique sur laquelle la méthylation a lieu (promoteur, corps du gène), le contexte de méthylation (CG, CHG, CHH) et enfin l'effet populationnel (10 populations européennes de peuplier).
2. Il a aussi résulté de cette intégration qu'il n'y a pas d'impact systématique de la méthylation sur l'expression des gènes à l'échelle du génome entier : des gènes sont identifiés dans les 4 typologies méthylation-expression considérées, à savoir relativement faiblement méthylés et exprimés, fortement méthylés et exprimés, faiblement méthylés et fortement exprimés, et fortement méthylés et faiblement exprimés.
3. Enfin, des gènes au profil méthylation-expression extrême pour l'ensemble des variables omiques considérées ont été identifiés et documentés.

Ces trois résultats majeurs sont discutés en détails dans la suite de cette section du manuscrit.

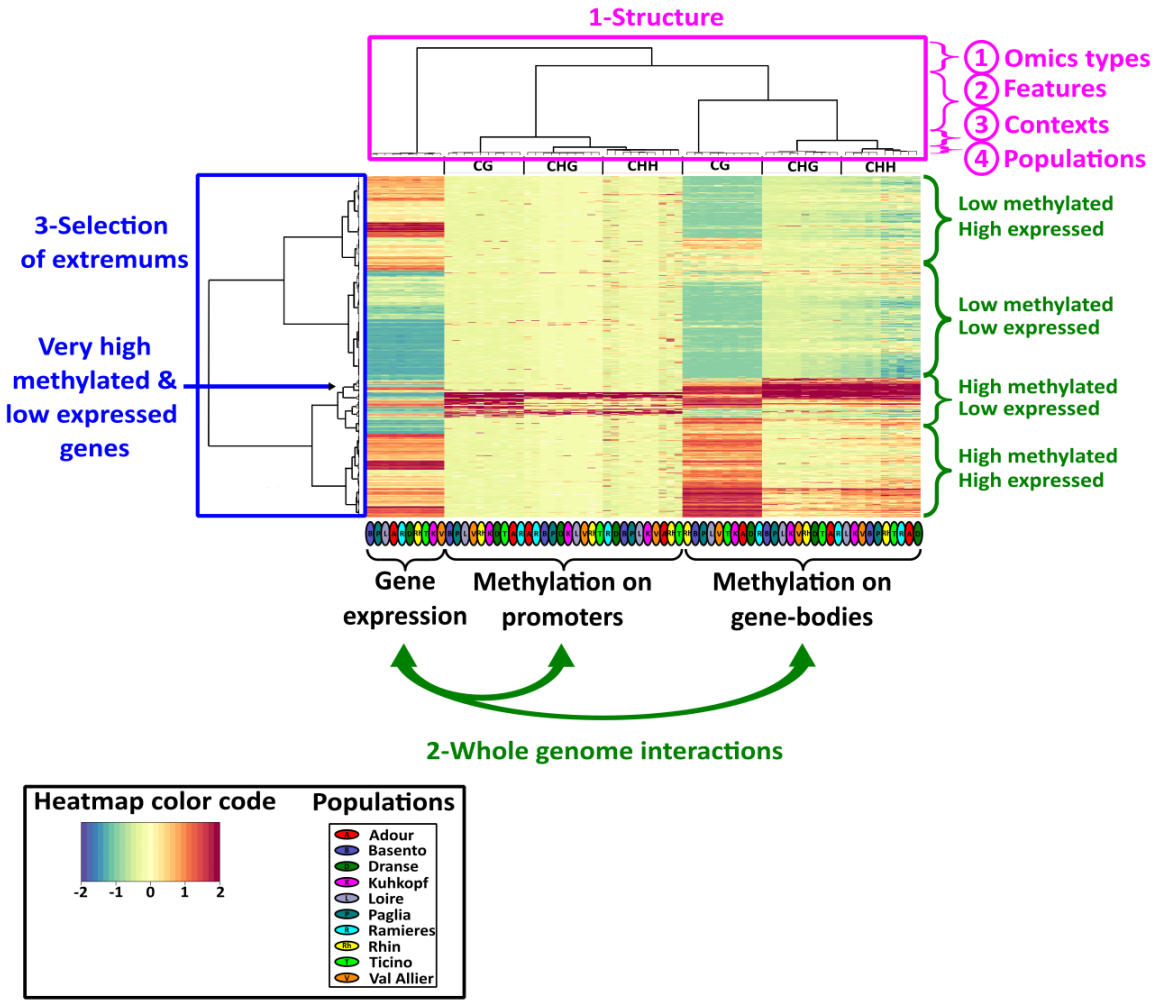


FIGURE 5.10 Principaux résultats de cimDiablo_v2 sur les données peuplier
Chaque ligne correspond à un gène, chaque colonne à une variable omique. Les lignes et colonnes sont regroupées selon les dendrogrammes issus du clustering hiérarchique avec la méthode Ward et la distance euclidienne. Les valeurs représentées selon le code couleur de la *heatmap* sont celles des données initiales, centrées et réduites, puis coupées en [-2, 2] (toute valeur inférieure à -2 est fixée à -2, toute valeur supérieure à 2 est fixée à 2). Les 10 populations de peuplier sont indiquées dans la légende en bas à gauche de la figure. Les principaux résultats sont indiqués sur la figure : 1- les effets regroupant plus ou moins fortement les variables sont hiérarchisées, 2- les interactions entre les variables sont analysées à l'échelle du génome, 3- des gènes au profil particulier sont sélectionnés.

5.4.2.1 Hiérarchisation/structuration des variables omiques

En intégrant les données de méthylation et d'expression des gènes avec l'outil cimDiablo_v2, il est possible d'utiliser l'algorithme de clustering hiérarchique afin de regrouper ensemble d'une part les gènes (en lignes) et d'autre part les variables omiques (en colonnes) aux variabilités les plus semblables. Nous avons donc utilisé le dendrogramme des colonnes pour structurer les variables omiques, et ainsi identifier que l'effet majeur séparant les variables est le type de données omiques (méthylation, expression), puis la région génique sur laquelle la

méthylation a lieu au sein du gène (promoteur, corps du gène), puis le contexte de méthylation (CG, CHG, CHH) et enfin l'effet populationnel (10 populations de peuplier). Ces résultats confirment ceux obtenus sur les mêmes données avec la matrice de corrélations ainsi que l'ACP (Figure 5.11).

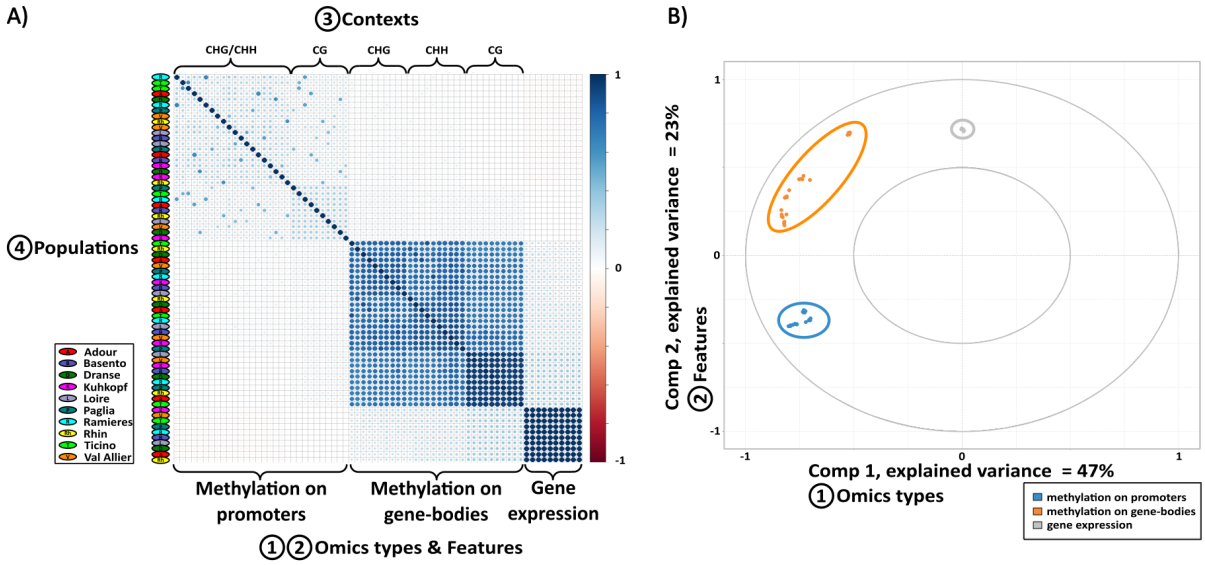


FIGURE 5.11 Matrice de corrélations et ACP des données peuplier
Analyses descriptives effectuées sur les données peuplier composées de 60 variables de méthylation et des 10 variables d'expression, toutes log-transformées. **A-** matrice de corrélations de Spearman des variables. Une forte corrélation positive entre deux variables est représentée par un point bleu foncé, une forte corrélation négative par un point rouge foncé. Une absence de point signifie une absence de corrélation. Toutes les variables ont une corrélation maximale (de 1) avec elles-mêmes, ces corrélations étant représentées sur la diagonale. La légende des 10 populations de peuplier se trouve à gauche de la figure. Les données sont réordonnées premièrement par 1) le type de données omiques et 2) la région génique (*feature*) de méthylation, puis par 3) le contexte de méthylation. **B-** représentation des différentes variables sous forme de points sur les deux premières composantes de l'ACP. Le code couleur des principaux groupes de points est légendé en bas à droite de la figure. La première composante sépare les variables selon le type de données omiques, la seconde selon la région génique de méthylation.

De manière générale, il est attendu que la différence majeure entre groupes de variables provienne du type de données omiques, mais le reste de la structuration révélée par cette analyse n'a à notre connaissance pas encore été étudiée, notamment concernant l'impact plus important de la région de méthylation (corps du gène *vs.* promoteur) par rapport au contexte de méthylation (CG, CHG, CHH). Ceci illustre la nécessité de l'intégration multi-variées des données par rapport aux analyses classiques uni-bi-variées pour rendre visible la structure générale des données et les déterminismes génomiques qui ont le plus d'effet sur la régulation des gènes et génomes.

La conséquence immédiate de ce résultat est l'effet majeur de la région génique de méthylation avec une forte disparité selon que la méthylation a lieu sur le promoteur ou le corps du gène. La variabilité induite par le contexte de méthylation est relativement mois importante, et les contextes peuvent même être confondus par exemple entre la méthylation CHG et CHH au niveau des promoteurs. Il semble ainsi qu'à l'échelle du génome, la variabilité induite par l'effet population soit très peu marquée comparée aux autres sources de variabilité considérées.

Ce résultat apporte également un argument supplémentaire de l'importance de l'intégration omique par rapport aux méthodes uni-bi-variées sur le niveau de variabilité capturée pour l'effet population comparé aux autres variables, et de l'interprétation scientifique qui en découle.

5.4.2.2 Description des interactions méthylation-expression à l'échelle du génome

À l'échelle du génome, nous n'avons pas identifié d'impact systématique de la méthylation sur l'expression des gènes. En effet, les différentes analyses des Figures 5.10 et 5.11 séparent en premier les variables de méthylation et d'expression, laissant supposer une relativement faible interaction entre la méthylation et l'expression génique. Le dendrogramme des lignes de la Figure 5.10 couplé à la *heatmap* précise et illustre ce constat. Il est ainsi possible d'identifier des groupes de gènes dans les 4 configurations ou typologies, à savoir faiblement méthylés et fortement exprimés (26.1%), et faiblement méthylés et exprimés (38.3%), fortement méthylés et faiblement exprimés (8.4%), et fortement méthylés et exprimés (27.2%) (Figure 5.12).

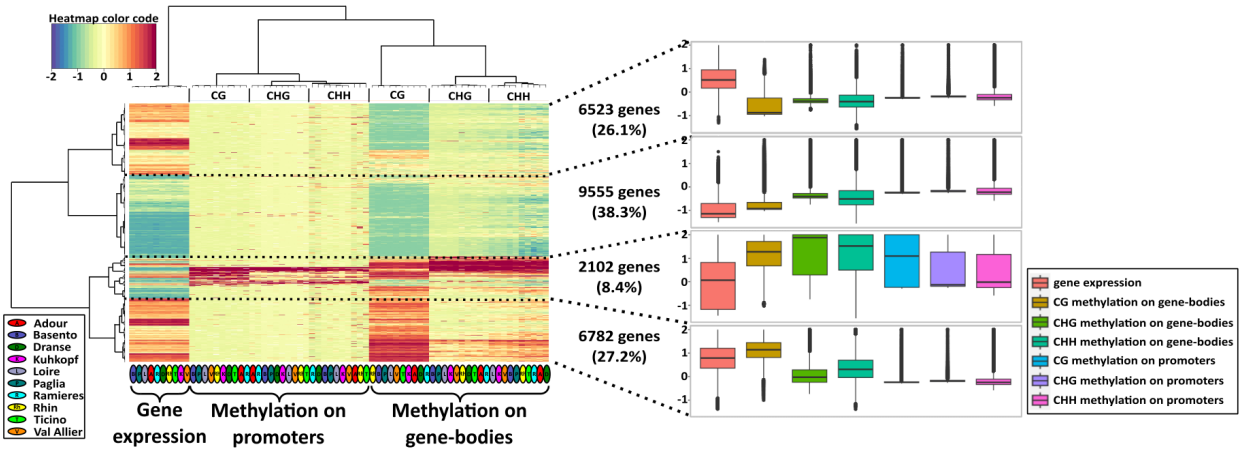


FIGURE 5.12 cimDiablo_v2 sur les données peuplier avec 4 profils de régulation méthylation-expression

Chaque ligne correspond à un gène, chaque colonne à une variable omique. Les lignes et colonnes sont regroupées selon les dendrogrammes issus du clustering hiérarchique avec la méthode Ward et la distance euclidienne. Les valeurs représentées selon le code couleur de la *heatmap* sont celles des données initiales, centrées et réduites, puis coupées en [-2, 2] (toute valeur inférieure à -2 est fixée à -2, toute valeur supérieure à 2 est fixée à 2). Les 10 populations de peuplier sont indiquées dans la légende en bas à gauche de la figure. Pour chacun des 4 groupes de gènes identifiés selon leur profil méthylation-expression, leurs distributions sur les différents blocs sont représentés sous forme de *boxplots* (légende à droite) pour chacune des 10 populations (axe des abscisses).

L'absence de lien systématique entre les variables de méthylation et d'expression est cohérente avec les conclusions de Li et al. (2012), qui ont montré que la méthylation dans les régions promotrices ne réprimait que quelques gènes fortement méthylés chez le riz. Ce résultat confirme également les résultats obtenus à partir des mêmes données brutes en analyses bi-variées dans Sow et al. (2023). Dans les analyses bi-variées, seuls quelques gènes au profil méthylation-expression particulier se démarquent, à savoir les gènes de très forte méthylation étant systématiquement faiblement exprimés et ceux de très forte expression étant généralement faiblement méthylés, pour chaque génotype considéré séparément (Figure 5.13). Dans

les analyses multi-variées, l'application de la log-transformation et du *cutoff* atténuant les valeurs extrêmes, les gènes au profil méthylation-expression extrême ne sont plus identifiables : toutes les configurations possibles de méthylation-expression sont alors identifiées dans cimDiablo_v2, et ceci en considérant simultanément les différentes populations. Il ressort donc de ces analyses complémentaires que les liens entre méthylation et expression sont complexes et ne sont distinguables qu'en fonction des contextes de méthylation et des régions génomiques ciblées pour certains gènes uniquement selon le niveau de méthylation, comme proposé dans Sow et al. (2023), dans Niederhuth and Schmitz (2017) et Bewick and Schmitz (2017).

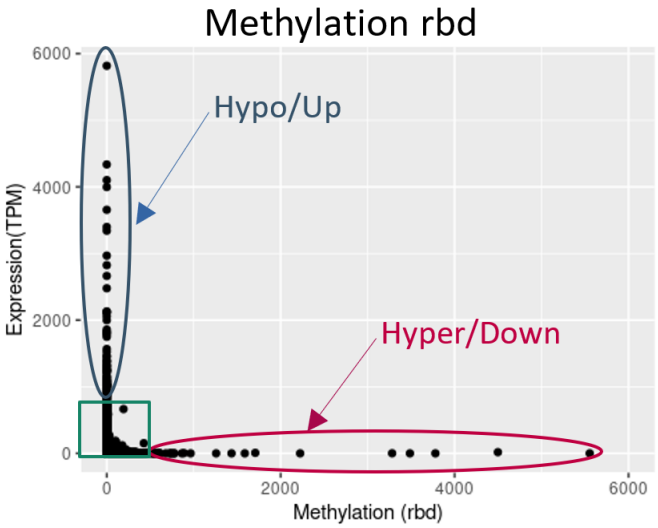


FIGURE 5.13 Nuage de points des gènes pour un couple méthylation-expression
Figure issue de Sow et al. (2023). Chaque point représente un gène. L'axe des abscisses représente leur méthylation en *rbd*, et l'axe des ordonnées leur expression en *TPM*. Trois profils se distinguent : les gènes fortement exprimés et faiblement méthylés, les fortement méthylés et faiblement exprimés et les faiblement méthylés et exprimés.

Pour préciser les interactions entre méthylation et expression sur ces données, des analyses par quantile ont été effectuées dans l'article Sow et al. (2023) (Figure 5.14). Notons que ces analyses sont effectuées avec la méthylation en pourcentage, mais avec des résultats similaires obtenus avec la normalisation de la méthylation en *rbd* (résultats non publiés). Il ressort de ces analyses que pour le promoteur en CG, une corrélation négative est observée entre méthylation et expression (Figure 5.14.A), avec les gènes fortement méthylés qui sont faiblement exprimés et vice-versa, comme documenté dans la littérature chez d'autres espèces (Kon and Yoshikawa, 2014; Nuo et al., 2016; Ma et al., 2020). Pour le corps du gène (en CG), on observe une relation positive entre méthylation et expression (Ball et al., 2009; Bewick and Schmitz, 2017) pour les faibles et moyennes valeurs de méthylation qui devient négative lorsque la méthylation est très forte (Figure 5.14.B). Il semble néanmoins qu'indépendamment du contexte et de la région de méthylation, il existe un niveau de méthylation à partir duquel l'expression du gène est fortement réprimée.

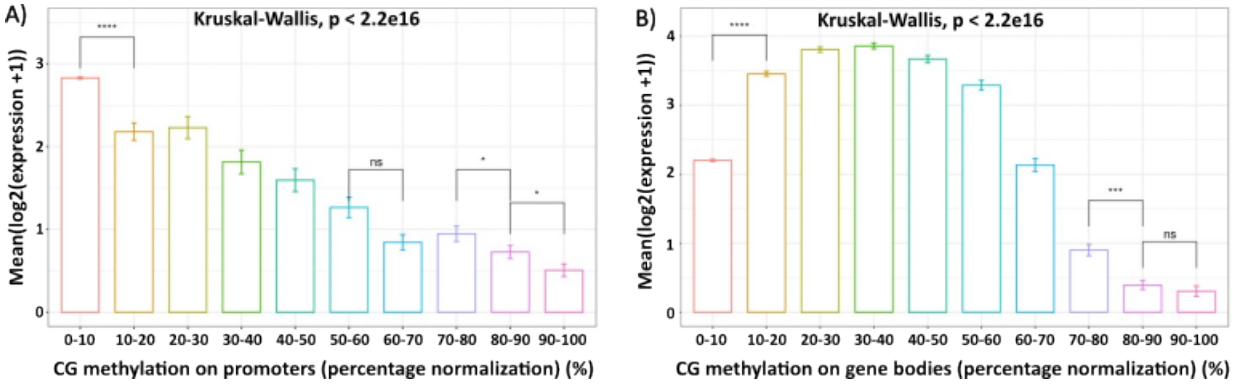


FIGURE 5.14 Relation entre expression et méthylation par quantile (méthylation en pourcentage)

Figure et légende issues de Sow et al. (2023). A, Relationship between promoter DNA methylation in CpG context and gene expression. Methylation data are splitted into 10 quantiles with each quantile capturing 10% of the methylation level. Expression data are shown as the mean expression of genes in log2. B, Relationship between gene body DNA methylation in CpG context and gene expression. Methylation data are splitted into 10 quantiles with each quantile capturing 10% of the methylation level. Expression data are shown as the mean expression of genes in log2.

Ainsi, nos travaux présentent des relations complexes entre la méthylation et l'expression des gènes, avec une corrélation négative ou positive en fonction des régions génomiques cibles et le contexte de méthylation, mais sans relation systématique entre méthylation et expression à l'échelle du génome entier, et enfin avec un effet "dose", c'est-à-dire un seuil de méthylation (quel que soit le compartiment du gène considéré) à partir duquel l'expression des gènes est réprimée.

5.4.2.3 Sélection de gènes aux profils multi-omiques extrêmes

Bien qu'aucun impact fort de la méthylation sur l'expression des gènes n'ait été déterminé à l'échelle du génome entier, mis à part l'effet "dose" précédemment décrit, un groupe de gènes très fortement méthylés pour l'ensemble des variables méthylomiques est visible sur la Figure 5.10. Ce résultat est d'autant plus intéressant que la méthylation de ces 90 gènes (Figure 5.15.A) semble stable quels que soient la région génomique cible, le contexte de méthylation et la population, montrant une fois de plus l'apport de l'intégration par rapport aux analyses uni-bi-variées pour détecter de tels profils. Pour identifier de manière plus précise ces types de gènes stables (c'est-à-dire présentant la même régulation omique indépendamment des populations), nous avons exploité la fonction de "débruitage" de cimDiablo_v2 pour éliminer la variabilité résiduelle liée à l'effet population et ainsi identifier les *master regulators-drivers*. 143 gènes ont alors été identifiés, dont les 90 initialement identifiés avant l'étape de "débruitage", ces gènes ayant des profils de méthylation et d'expression stables quelle que soit la population considérée (Figure 5.15.B). L'interprétation de ces différences d'effectifs selon le "débruitage" a été discutée en fin de section précédente (Section 5.4.1.2). L'analyse d'enrichissement en ontologie de ces gènes révèle une implication dans les processus métaboliques de

l'ARN (desoxyrubunleoside, carbohydrate) et du cycle cellulaire. Parmi ces gènes se trouve notamment *Di19* (*Drought induced 19*), initialement identifié chez *Arabidopsis thaliana* (Liu et al., 2013) puis chez le peuplier transgénique (Wu et al., 2022) comme jouant un rôle dans la tolérance à la sécheresse, une propriété très intéressante notamment chez le peuplier qui est une espèce pérenne constamment soumise aux variations climatiques auxquelles l'espèce doit s'adapter.

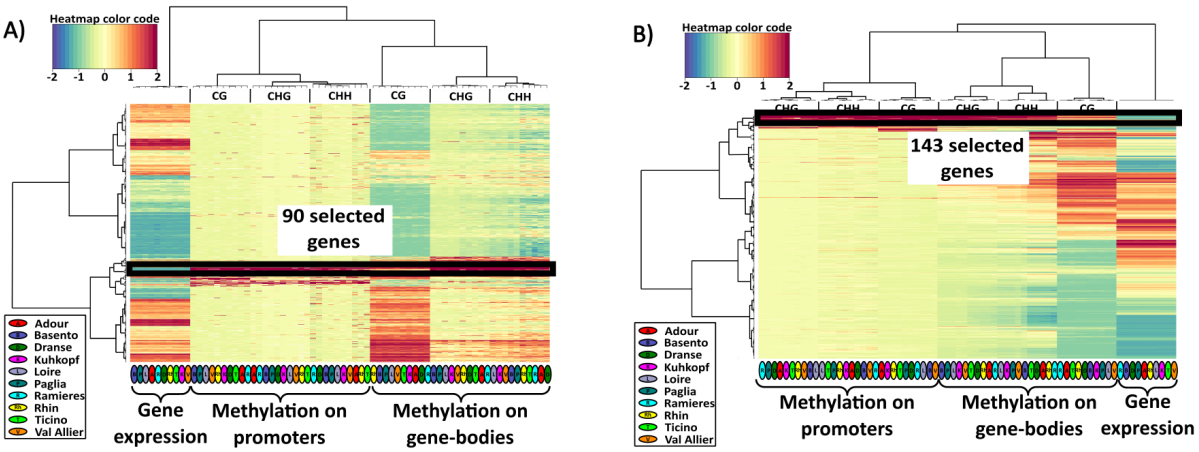


FIGURE 5.15 Gènes fortement méthylés identifiés par cimDiablo_v2 sur les données peuplier avant et après "débruitage"

cimDiablo_v2 sur les données **A-** avant "débruitage" et **B-** après "débruitage". Chaque ligne correspond à un gène, chaque colonne à une variable omique. Les lignes et colonnes sont regroupées selon les dendrogrammes issus du clustering hiérarchique avec la méthode Ward et la distance euclidienne. Les valeurs représentées selon le code couleur de la *heatmap* sont celles des données initiales, centrées et réduites, "débruitées" (en **B** seulement), de nouveau centrées et réduites (en **B** seulement), et enfin coupées en [-2, 2] (toute valeur inférieure à -2 est fixée à -2, toute valeur supérieure à 2 est fixée à 2). Les 10 populations de peuplier sont indiquées dans la légende en bas à gauche de la figure. Les groupes de gènes sélectionnés pour leur très forte méthylation sur l'ensemble des variables de méthylation sont encadrés.

5.4.3 Conclusions

Dans cet article, des résultats à la fois méthodologiques et scientifiques ont été obtenus.

Premièrement, un tutoriel ainsi qu'un outil, appelé cimDiablo_v2, implémenté à partir du paquet R `mixOmics` ont été développés dans le but de rendre possible différents types d'intégration à partir d'une variété de données omiques et pour répondre à de multiples questions biologiques autant pour la description des données, la sélection de variables/observations ou encore la prédiction de traits. Le **tutoriel** se démarque en étant, à notre connaissance, la seule procédure linéaire partant de la question biologique afin d'arriver jusqu'à l'intégration multi-omiques, et a montré son adaptabilité sur les données peuplier. La fonction **cimDiablo_v2**, basée sur la régression MB-PLS de `mixOmics` afin de réduire la dimension des données, a été développée dans l'objectif principal de "débruiter" les données. Cependant, bien que les premiers résultats sur les données semblent prometteurs, ayant abouti à de nouveaux résultats biologiques d'intérêt, plusieurs limitations ont été identifiées notamment sur la métrique à employer pour quantifier l'impact du "débruitage". L'utilisation de cimDiablo_v2 sur d'autres

jeux de données est alors nécessaire pour améliorer nos connaissances sur l'impact réel du "débruitage" des données.

Secondement, cet article présente aussi des **résultats biologiques** issus de l'intégration multi-omiques de données peuplier, complémentaires aux analyses uni-, bi- et multi-variées menées dans Sow et al. (2023). L'analyse intégrative a pu clarifier et apporter des résultats nouveaux particulièrement importants pour la compréhension plus approfondie du système biologique peuplier.

1. Un premier résultat de l'intégration de données de méthylation et d'expression des gènes a été de classer par ordre d'importance les facteurs de variabilité omique (expression, méthylation), avec l'effet majeur de la région génique (promoteur, corps du gène), puis le contexte de méthylation (CG, CHG, CHH) et enfin l'effet mineur des populations.
2. Le deuxième résultat a été de démontrer qu'à l'échelle du génome entier, il n'existe pas d'impact systématique de la méthylation sur la régulation génique. Néanmoins, au-delà d'un seuil de méthylation, qui diffère selon la variable de méthylation considérée, l'expression du gène semble systématiquement réprimée. Quatre configurations ou typologies de régulation ont été ainsi caractérisées, à savoir fortement méthylés et exprimés, fortement méthylés et faiblement exprimés, faiblement méthylés et fortement exprimés, et faiblement méthylés et exprimés.
3. Enfin, il a résulté de l'intégration des données l'identification de 143 *master regulators-drivers* dont le profil de méthylation et d'expression est contrasté (forte méthylation *vs.* faible expression) et stable quelle que soit la population considérée. Ces gènes sont principalement impliqués dans le métabolisme des ARNs et le cycle cellulaire. Parmi eux se trouve notamment le gène *Di19* (*Drought induced 19*), jouant un rôle important dans la tolérance à la sécheresse chez le peuplier et *Arabidopsis thaliana*. Ces résultats suggèrent possiblement le rôle de la co-régulation de gènes par la méthylation de l'ADN et le niveau expression pour des gènes impliqués dans des caractères d'adaptation comme la tolérance aux stress abiotiques.

5.5 Exploitation des outils et approches développées pour d'autres thématiques de recherche en cours dans l'équipe

En complément de l'article précédent qui présente l'utilisation du tutoriel et cimDiablo_v2 sur des données peuplier, d'autres travaux ont aussi été réalisés avec les mêmes outils pour répondre à des questionnements scientifiques similaires mais cette fois non plus entre populations d'une même espèce mais entre différentes espèces (chez les céréales).

5.5.1 Intégration multi-omiques chez *Brachypodium* et le maïs

Contexte : L'article Bellec et al. (2023) (en Annexe D Pages 247 à 271, données supplémentaires en Annexe E Pages 273 à 299) s'intéresse à la plasticité génomique des céréales durant les 100 millions d'années d'évolution depuis leur origine (ancêtre commun) à nos jours *via* les modifications de la structure et la régulation des gènes et génomes. Dans ce contexte, mon travail a porté précisément sur l'étude de la régulation omique pour deux des dix espèces étudiées dans cet article, à savoir le maïs et *Brachypodium*, et plus précisément sur l'impact de la méthylation de l'ADN sur l'expression des gènes *via* l'outil `mixOmics` et cimDiablo_v2.

Question biologique : Quelles ont été la nature et l'ampleur des modifications génomiques sur une période évolutive de 100 millions d'années ayant abouti, à partir d'un ancêtre commun, aux différentes espèces de céréales actuelles ?

Données : Comme pour le peuplier, il s'agit ici également d'intégrer par gènes des données méthylomiques et transcriptomiques. Néanmoins, ces données ne font pas intervenir la notion d'individus ou populations comme pour les données peuplier, mais ajoute ici chez les céréales un effet stade de développement puisque les données omiques sont produites pour 3 stades de développement du grain (9, 16 et 28 degré jours pour *Brachypodium* et 7, 15 et 35 pour le maïs, équivalents en terme physiologique), amenant finalement à 3 variables d'expression (3 stades de développement du grain) et 18 variables de méthylation (3 stades $\times$ 3 contextes $\times$ 2 régions du gène). Les deux espèces ont été analysées distinctement avec 12952 gènes pour *Brachypodium* et 41748 gènes pour le maïs ; il aurait aussi été envisageable de les analyser simultanément en conservant uniquement les gènes orthologues entre les deux espèces, ce qui aurait néanmoins grandement diminué le nombre de gènes considérés et donc l'intérêt de l'analyse portant sur les relations méthylation-expression à l'échelle du génome entier. Ces données ont été normalisées avec l'expression en TPM et la méthylation en rbd, un logarithme étant ensuite appliqué à l'ensemble des données. De plus, et contrairement aux données peuplier, ces données-ci comportent un très fort taux de valeurs manquantes dû à la faible profondeur de séquençage utilisée lors de la production des données de méthylation, nous obligeant à conserver beaucoup de gènes comportant jusqu'à 65% de valeurs manquantes (en considérant les 18 variables de méthylation et les 3 d'expression). Les résultats finaux sont donc à prendre avec beaucoup de recul dans l'étude des interactions globales entre expression et méthylation à l'échelle du génome entier. Toutefois, pour la recherche de gènes candidats, seuls les gènes sans valeur manquante ont été conservés, rendant dans ce contexte les conclusions plus robustes.

Les résultats majeurs de l'article : Bellec et al. ont étudié l'impact de la conservation des gènes, des inversions génomiques et de la duplication des génomes sur leur régulation omique (mutation, expression, méthylation) et ont montré que :

— Les gènes situés dans les régions inversées du génome sont majoritairement moins méthylés (promoteurs) et plus exprimés que les gènes situés dans des régions conservant l'orientation ancestrale.
— Les gènes conservés sont majoritairement moins méthylés (promoteurs et corps du gène) et plus exprimés que les gènes spécifiques à chaque espèce (non conservés).
— Les gènes dupliqués issus de la duplication ancestrale sont moins méthylés (corps du gène) que les gènes en singletons (simple copie).
— Plus de 50% des gènes en paires issus de la duplication ancestrale demeurent co-exprimés.
— Plus un gène est méthylé sur le corps du gène, moins il est affecté par les mutations.
— Les gènes les plus conservés en termes de séquence sont les plus exprimés.
— Les gènes les plus méthylés sur le promoteur ont tendance à être les moins exprimés.

Ces analyses ont été conduites par des approches uni- et bi-variées comparant, pour chaque contexte génomique (gène inversés *vs.* non inversés, gènes dupliqués *vs.* gènes en singletons, gènes conservés *vs.* gènes spécifiques d'espèces) séparément, et pour chaque type de données omiques indépendamment. Dans ce contexte d'étude, mon apport a consisté à étudier les interactions entre la méthylation de l'ADN et l'expression des gènes en considérant simultanément les différents contextes (CG, CHG, CHH) et différentes régions géniques (promoteur et corps du gène) de méthylation ainsi que différents stades de développement du grain. Ces interactions ont été étudiées à l'échelle du génome entier, puis des gènes aux profils omiques particuliers (gènes fortement méthylés et faiblement exprimés, et gènes fortement exprimés et faiblement méthylés) ont été identifiés chez les deux espèces afin d'étudier leur fonction biologique.

Résultats de l'intégration à l'échelle du génome : Les Figures 5.16 et 5.17 représentent respectivement les résultats de cimDiablo_v2 sur les données *Brachypodium* et *maïs*, avec les options de "débruitage" des données, de deuxième standardisation pour les centrer et réduire, et de *cutoff* en [-2,2] (*i.e.* toutes les valeurs supérieures à 2 sont fixées à 2, toutes les valeurs inférieures à -2 sont fixées à -2). Contrairement aux données peuplier, la séparation entre les différentes variables omiques est beaucoup plus confuse, puisqu'il ne semble pas y avoir, pour *Brachypodium*, de hiérarchisation claire des données selon la région génique ou le contexte de méthylation, le seul effet visible étant que le stade de développement du grain est généralement la variabilité la moins discriminante (à l'instar de l'effet populationnel chez le peuplier). Pour le maïs, il apparaît que le contexte CHH se distingue des deux autres, et que la région génique sur laquelle la méthylation a lieu distingue plus les variables que le contexte CG ou CHG. Une hypothèse pour expliquer les différences observées entre les données peuplier et céréales (maïs et *Brachypodium*) est que la hiérarchisation moins visible des données chez les céréales est due à leur plus fort taux de valeurs manquantes, particulièrement important chez *Brachypodium*. Néanmoins, pour ces deux céréales comme pour le peuplier, il est possible d'identifier visuellement un groupe de gènes de forte méthylation pour la plupart des variables et de faible expression génique, comme détaillé dans la section suivante.

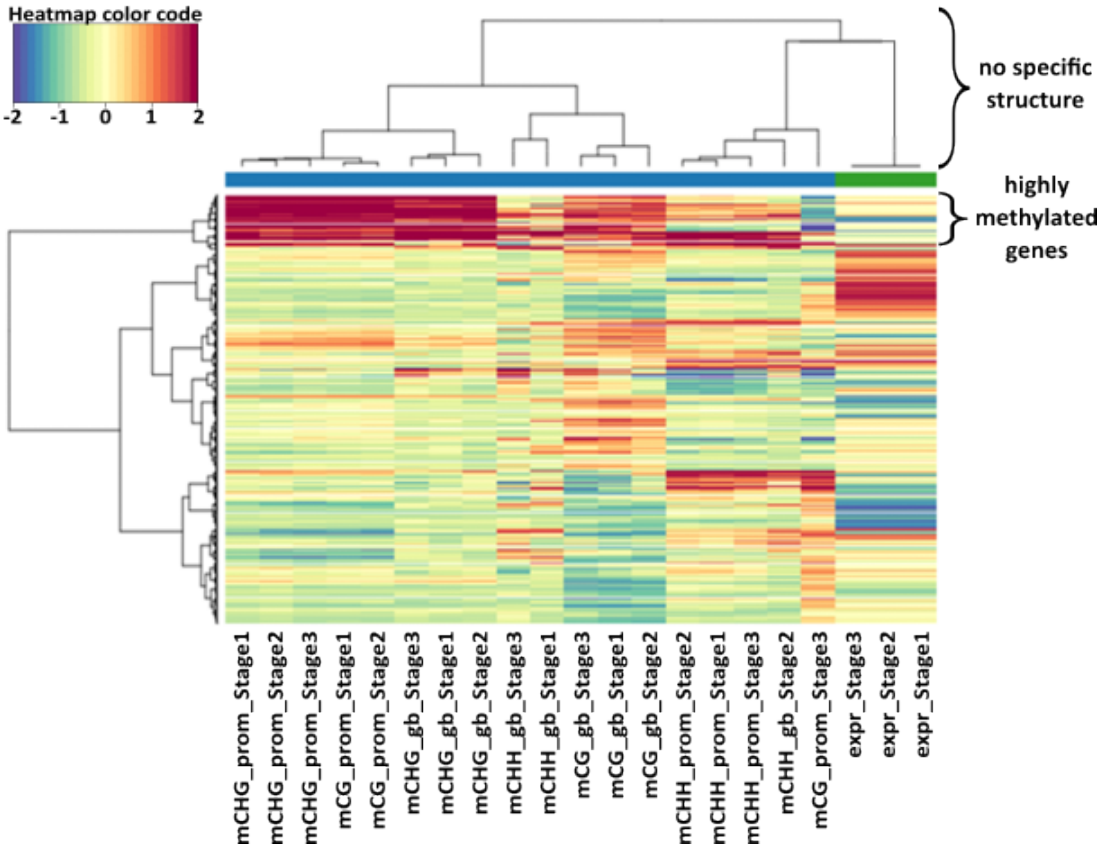


FIGURE 5.16 cimDiablo_v2 sur les données *Brachypodium*

Chaque ligne correspond à un gène, chaque colonne à une variable omique. Les lignes et colonnes sont regroupées selon les dendrogrammes issus du clustering hiérarchique avec la méthode Ward et la distance euclidienne. Les valeurs représentées selon le code couleur de la *heatmap* sont celles des données initiales, centrées et réduites, "débruitées", de nouveau centrées et réduites, et enfin coupées en [-2, 2] (toute valeur inférieure à -2 est fixée à -2, toute valeur supérieure à 2 est fixée à 2). mCG, mCHG et mCHH représentes les trois contextes de méthylation, "prom" et "gb" respectivement le promoteur et le corps du gène (*gene body*), et "Stage1", "Stage2" et "Stage3" les trois stades de développement du grain à 9, 16 et 28 jours.

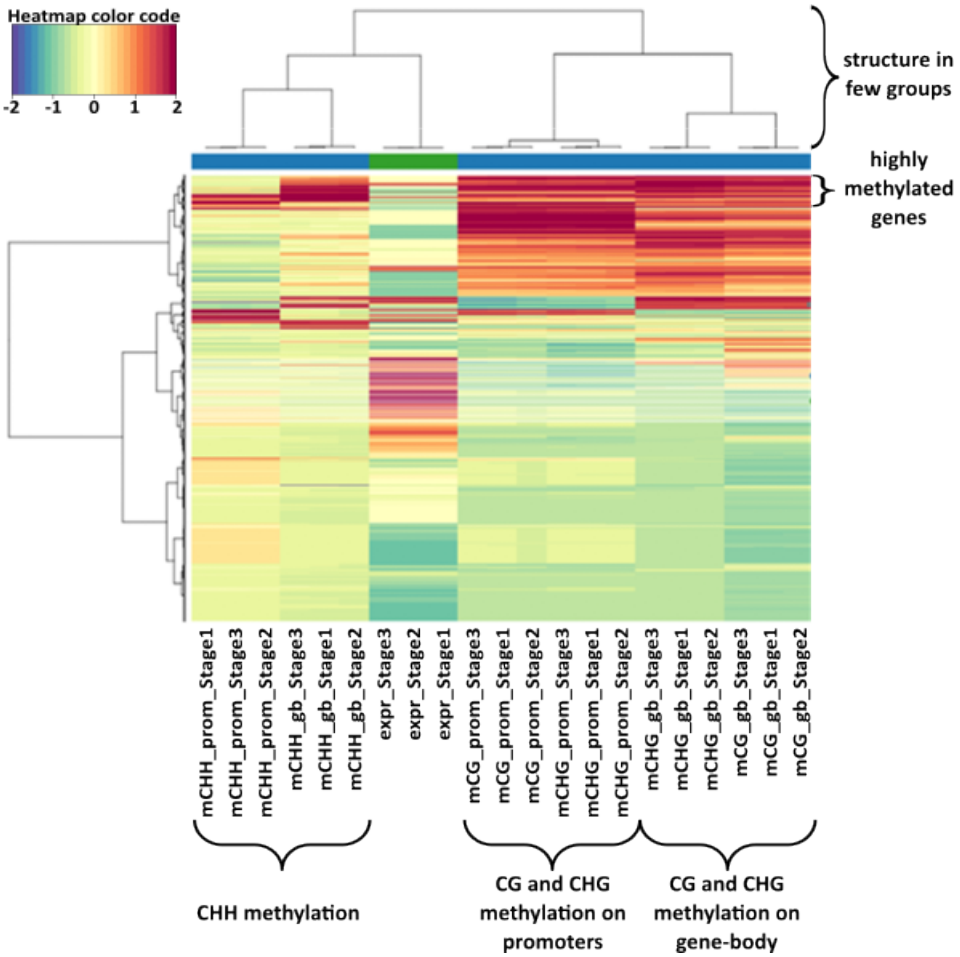


FIGURE 5.17 cimDiablo_v2 sur les données maïs

Chaque ligne correspond à un gène, chaque colonne à une variable omique. Les lignes et colonnes sont regroupées selon les dendrogrammes issus du clustering hiérarchique avec la méthode Ward et la distance euclidienne. Les valeurs représentées selon le code couleur de la *heatmap* sont celles des données initiales, centrées et réduites, "débruitées", de nouveau centrées et réduites, et enfin coupées en [-2, 2] (toute valeur inférieure à -2 est fixée à -2, toute valeur supérieure à 2 est fixée à 2). mCG, mCHG et mCHH représentes les trois contextes de méthylation, "prom" et "gb" respectivement le promoteur et le corps du gène (*gene body*), et "Stage1", "Stage2" et "Stage3" les trois stades de développement du grain à 7, 15 et 35 jours.

Résultats de l'intégration à l'échelle des gènes : Nous avons par la suite identifié les gènes au profil extrême (entre méthylation et expression) en nous inspirant des analyses bi-variées. En effet, en affichant les nuages de points des gènes avec une de leur variable de méthylation en abscisse et en ordonnée la variable d'expression au même stade de développement du grain, avec les données initiales avant application du logarithme, les résultats font ressortir trois typologies de profils omiques : quelques gènes très fortement méthylés et faiblement exprimés, quelques gènes faiblement méthylés et très fortement exprimés, et enfin la majorité des gènes aux valeurs de méthylation et d'expression relativement faibles (exemple en Figure 5.18). Ces profils initialement très marqués le sont beaucoup moins après application du logarithme, ce qui est un avantage pour le "débruitage" de cimDiablo_v2 qui fonctionne mieux sur des données pour lesquelles il n'y a pas de valeurs trop extrêmes, mais ceci rend

plus difficile leur identification. Ainsi, nous avons décidé, après "débruitage", d'appliquer la transformation inverse au logarithme, c'est-à-dire l'exponentielle, pour obtenir de nouveau des données aux trois grands types de profils de méthylation-expression. Il suffit alors de choisir des seuils à partir desquelles une valeur est considérée comme extrême pour pouvoir identifier ces gènes particuliers. L'intérêt de ce passage par la fonction cimDiablo_v2 pour finalement retourner à des analyses bi-variées réside dans le "débruitage" de cette fonction : alors que les analyses bi-variées classiques identifient les gènes au profil extrême pour au moins une interaction méthylation-expression, le "débruitage" ne conserve que celles qui sont partagées par plusieurs variables de méthylation et/ou expression. Les gènes sélectionnés par les analyses bi-variées après "débruitage" sont donc ceux partagés par plusieurs variables. Certains des gènes sélectionnés après "débruitage" le sont aussi par les analyses bi-variées classiques, et d'autres peuvent ne l'être qu'après "débruitage" s'ils n'étaient pas considérés comme suffisamment extrêmes pour l'ensemble des interactions méthylation-expression. Les effectifs de gènes aux valeurs extrêmes identifiés sont résumés en Table 5.1, et les principales fonctions biologiques associées par enrichissement GO sont liées de manière générale au processus de signalisation cellulaire (*transport, transmembrane, ligase, cation, ion, acid*) pour les gènes faiblement méthylés mais fortement exprimés (définis dans Bellec et al. (2023) comme "*Hypo/Up*"), et aux activités de transcription et d'hydrolase *(purine/ribonucleotide binding, hydrolase activity, ion binding, heat shock protein binding*) pour les gènes fortement méthylés mais faiblement exprimés ("*Hyper/Down*"). Il semblerait donc que les gènes de signalisation (fonctions importantes pour l'activité cellulaire) soient faiblement ou pas méthylés pour potentiellement assurer leur propre fonctionnement (pour un effet "protecteur" de la régulation) tandis que l'expression des gènes liée à la transcription est plus contrôlée par la méthylation de l'ADN (pour un effet "adaptatif" de la régulation).

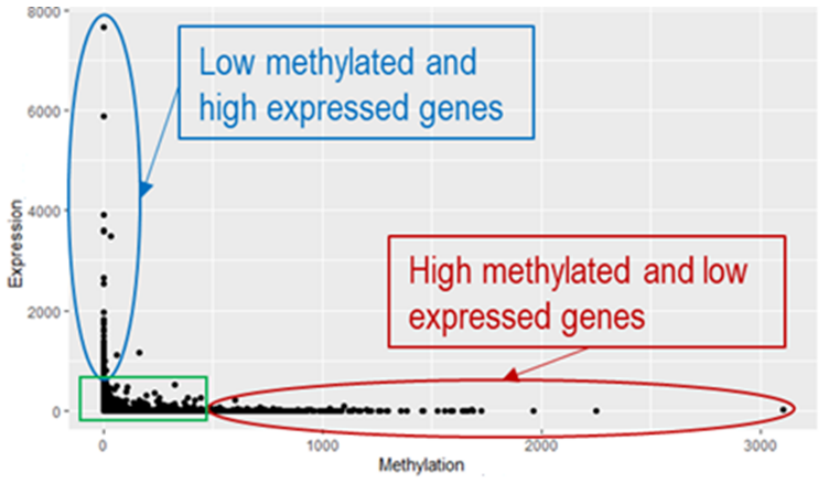


FIGURE 5.18 Nuage de points des gènes pour un couple méthylation-expression
Chaque point représente un gène. L'axe des abscisses représente leur méthylation en *rbd*, et l'axe des ordonnées leur expression en *TMM*. Trois profils se distinguent : les gènes fortement exprimés et faiblement méthylés, les fortement méthylés et faiblement exprimés et les faiblement méthylés et exprimés.

	CG		CHG		CHH	
	Hyper / Down	Hypo / Up	Hyper / Down	Hypo / Up	Hyper / Down	Hypo / Up
Brachy promoter	12	146	24	130	4	132
Brachy gene body	330	11	31	11	39	11
Maize promoter	35	7	38	8	10	8
Maize gene body	252	9	265	9	4	9

TABLE 5.1 Effectifs d'extremums identifiées pour les différentes variables omiques de maïs et *Brachypodium*
Pour le maïs et *Brachypodium*, des gènes aux profils extrêmes ont été sélectionnés pour les différents contextes (CG, CHG, CHH) et régions (promoteur, corps du gène) de méthylation. "Hyper/Down" désigne ici les gènes fortement méthylés et faiblement exprimés, "Hypo/Up" les gènes fortement exprimés et faiblement méthylés.

5.6 Points clés

— D'un point de vue méthodologique, un tutoriel pour préparer l'intégration de données omiques ainsi qu'une nouvelle fonction cimDiablo_v2 ont été développés, présentés, testés et rendus accessibles. Ils ont pour objectifs de s'adapter à différentes questions biologiques (regroupées en questions de description, sélection et prédiction), différents types de données omiques (génomiques, épigénomiques, transcriptomiques, protéomiques, métabolomiques, *etc.*) et différents "niveaux" d'intégration (par tissus, individus, populations, espèces, temporalités, conditions, *etc.*). La fonction cimDiablo_v2 a pour particularité de "débruiter" les données en se servant des résultats de la régression PLS par blocs du paquet R `mixOmics`.

Le tutoriel a ainsi pu être utilisés dans différents contextes et aboutir à des résultats concluants. La fonction cimDiablo_v2 s'est elle aussi correctement adaptée à ces différentes données pour lesquelles les lignes correspondaient aux gènes et les colonnes aux variables omiques (méthylation et expression). Certains des points présentés ici concernant le "débruitage" de cimDiablo_v2 sont complétées dans les chapitres suivants suite à leur utilisation sur les données animales.

Ces développements méthodologiques ont été utilisés dans différentes analyses, dans le but de déterminer les interactions majeures entre méthylation et expression des gènes chez le peuplier ainsi que chez les céréales (*Brachypodium*, maïs).

— D'un point de vue scientifique, des réponses ont été apportées aux différentes questions biologiques posées dans l'introduction de ce document (Chapitre 4 Pages 43 à 47) :

L'intégration des données omiques permet-elle d'identifier les gènes dont la régulation par méthylation/expression est spécifique de certaines populations d'une espèce pérenne (peuplier) ou conservée entre ces populations d'origines géographiques diverses associées à différentes contraintes environnementales ?

L'analyse intégrative de données d'expression et de méthylation de 10 populations de peuplier d'origines géographiques distinctes en Europe montre que :

1. les facteurs de variabilité omique sont, par ordre d'importance et au-delà du type de données (expression, méthylation), les régions géniques (promoteur, corps du gène), puis le contexte de méthylation (CG, CHG, CHH) et enfin les populations ayant un effet mineur constaté.
2. à l'échelle du génome entier, il n'existe pas d'impact générique de la méthylation sur la régulation génique, mis à part un effet fort de "dose", *i.e.* un seuil de méthylation (quel que soit le compartiment du gène) à partir duquel l'expression du gène est réprimée. Quatre configurations ou typologies de régulation génique ont été ainsi caractérisées, à savoir fortement méthylés et exprimés, fortement méthylés et faiblement exprimés, faiblement méthylés et fortement exprimés, et faiblement méthylés et exprimés.
3. 143 *master regulators-drivers* ont un profil de méthylation et d'expression contrasté (forte méthylation et faible expression) et stable quelle que soit la population considérée. Ces gènes sont principalement impliqués dans le métabolisme des ARNs et le cycle cellulaire. Parmi eux se trouve notamment le gène Di19 (*Drought induced 19*), jouant un rôle important dans la tolérance à la sécheresse chez le peuplier et *Arabidopsis thaliana*. Ces résultats suggèrent le rôle de la co-régulation de gènes par la méthylation de l'ADN et le niveau expression pour des gènes impliqués dans des caractères d'adaptation comme la tolérance aux stress abiotiques.

L'intégration des données omiques permet-elle d'identifier les spécificités et similitudes de la régulation des gènes par méthylation/expression au cours du développement du grain entre deux espèces (maïs et Brachypodium) issues d'un ancêtre commun datant de 65 millions d'années ?

L'analyse intégrative de données d'expression et de méthylation de deux espèces *Brachypodium* et maïs montre que :

1. les facteurs de variabilité omique, par ordre d'importance et au-delà du type de données (expression, méthylation), la région génique considérée (promoteur, corps du gène), le stade de développement du grain étant généralement la variabilité la moins discriminante.
2. trois types de profils de régulation génique ont été identifiés : quelques gènes très fortement méthylés et faiblement exprimés, quelques gènes faiblement méthylés et très fortement exprimés, et enfin et surtout la majorité des gènes aux valeurs de méthylation et d'expression relativement faibles. Les gènes aux valeurs extrêmes incluent des gènes faiblement méthylés et fortement exprimés impliqués dans des processus de signalisation cellulaire (*transport, transmembrane, ligase, cation, ion,*

acid), et des gènes fortement méthylés et faiblement exprimés impliqués dans des activités de transcription et d'hydrolase (*purine/ribonucleotide binding, hydrolase activity, ion binding, heat shock protein binding*). Ceci peut suggérer que les gènes de signalisation (fonctions importantes pour l'activité cellulaire) puissent être 'protégés' de la méthylation (et *in fine* inactivation) pour assurer leur constante expression tandis que l'expression des gènes liée à la transcription (activant ou réprimant finement des réponses physiologiques notamment d'adaptation ou de réponse à des contraintes) soit plus contrôlée par la méthylation de l'ADN.

Chapitre 6

Interactions protéines-phénotypes chez le bovin

6.1 Introduction et contexte

Dans le secteur de l'élevage et notamment de la production de viande chez le bovin, l'un des enjeux majeurs est de parvenir à un équilibre entre la proportion de masse musculaire et graisseuse sur la carcasse, de sorte à répondre aux attentes des consommateurs tout en obtenant un rendement économique maximal par animal pour l'éleveur et le transformateur. Dans ce contexte, l'intégration de diverses données omiques avec des variables phénotypiques représentatives de la composition corporelle, a pour objectifs 1- une meilleure compréhension des voies métaboliques associées à la composition corporelle, 2- la sélection d'entités biologiques telles que des gènes ou protéines associées aux phénotypes 3- à des fins prédictives basées sur la quantification de l'abondance des entités biologiques.

L'article Mardoc et al. (en préparation, ci-dessous) présente une partie des travaux réalisés dans le cadre de cette thèse sur l'intégration de données omiques de 3 organes/tissus (foie, muscle, tissu adipeux) et de 1 à 7 variable(s) phénotypique(s) de la composition chimique ou tissulaire des carcasses de 17 bovins. Pour ces bovins, une variabilité de la composition chimique ou tissulaire des carcasses a été induite par 2 stratégies nutritionnelles. L'objectif des analyses que j'ai réalisées était de sélectionner des groupes de protéines identifiées dans le tissu adipeux, le foie ou le muscle, fortement liés aux phénotypes, dans un but de découverte de marqueurs des compositions de carcasse bovine. Les données utilisées et les stratégies mises en place pour répondre à l'objectif sont illustrées en Figure 6.1.

Pour sélectionner ces protéines, plusieurs approches ont été suivies. Premièrement, plusieurs types de régressions PLS ont été menés : les PLS1 pour sélectionner les protéines par tissu les plus explicatives de la variabilité de chaque phénotype, les PLS2 pour sélectionner les protéines par tissu les plus explicatives de la variabilité de l'ensemble des phénotypes, et la MB-PLS pour sélectionner les protéines les plus explicatives de l'ensemble des phénotypes indépendamment du tissu. Deuxièmement, cimDiablo_v2 sans, puis avec "débruitage", a été utilisé dans le même objectif sélectif, cette fois pour sélectionner des protéines aux profils similaires tandis que les régressions PLS sélectionnent des protéines aux profils complémentaires. Ce chapitre présente et discute de ces différentes approches et de l'interprétation biologique des résultats.

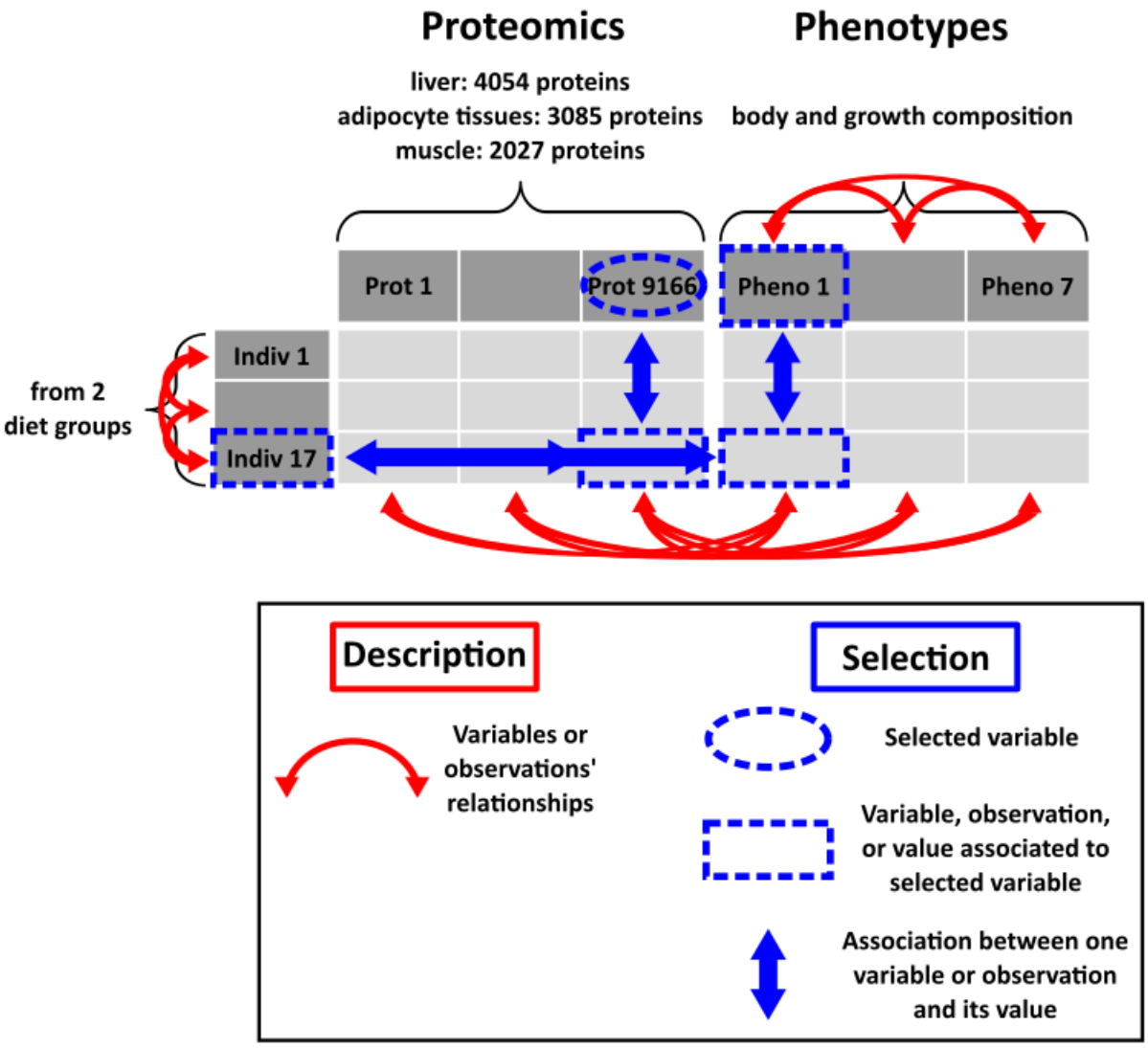


FIGURE 6.1 Représentation schématique des données bovines utilisées et des questions d'intégration multi-omiques traitées dans le cadre de la thèse

Les données bovines intégrées sont composées de deux blocs comportant en colonnes des variables et en lignes des individus. Le bloc protéomique est composé de 9166 variables, à savoir les abondances de 4054 protéines du foie, 3085 protéines du tissu adipeux et 2027 protéines du muscle. Le bloc phénotypique est composé de 7 variables de la composition chimique ou tissulaire des carcasses bovines. Les 17 individus bovins sont séparés en deux groupes selon le régime alimentaire. Le premier groupe recevait une ration composée d'ensilage de maïs supplémenté en céréales (blé et maïs). Le second groupe recevait une ration composée essentiellement d'herbe enrubannée et apportant une densité énergétique comparable à celle du groupe maïs. Les analyses réalisées avaient pour objectifs d'identifier les interactions entre les variables d'un bloc (interactions entre les différents phénotypes) et entre blocs (entre les protéines et les phénotypes), et les relations entre individus, puis de sélectionner les protéines expliquant au mieux la variabilité des phénotypes dans un objectif de recherche de marqueurs.

6.2 Article Mardoc et al. (en préparation) : *Integrating liver, muscle and adipocyte tissue proteomes to identify drivers of body and growth composition in bovine species*

L'article Mardoc et al. (en préparation) (Pages 108 à 134, données supplémentaires en Annexe F Pages 301 à 314) suivra les articles en préparation et visant à décrire l'impact des

régimes alimentaires sur les protéomes tissulaires. En effet, Alyssa Imbert a réalisé une première analyse pour étudier l'impact du régime sur le protéome de chaque tissu en combinant des analyses uni-variées et multi-variées. Les analyses uni-variées ont été effectuées à l'aide du paquet R *limma* (Ritchie et al., 2015b) en utilisant un modèle linéaire avec le régime alimentaire ajusté par la covariable centrée « âge » (Milliken and Johnson, 2001), afin de prendre en compte la différence d'âge à l'abattage induite par le régime. Une protéine est considérée comme significativement différentiellement abondante entre les deux régimes alimentaires (voir légende de la Figure 6.1) si sa p-valeur ajustée (*False Discovery Rate* (FDR), Benjamini and Hochberg (1995)) est inférieure à 10%. Une seule protéine est significativement différentiellement abondante dans le muscle, 15 dans le tissu adipeux périrénal et 221 dans le foie. En ce qui concerne les analyses multi-variées, des *sparse PLS Discriminant Analysis* (sPLSDA) ont été effectuées pour chaque tissu. Afin de sélectionner les protéines les plus impactées par le régime alimentaire, une grille de valeur a été choisie (entre 10 à 50 variables avec un pas de 5) et une validation croisée (5-fold, 20 répétitions, 2 composantes) a déterminé le nombre optimal de variables à sélectionner par composante. La première composante séparant les deux groupes, seules les protéines sélectionnées par celle-ci ont été conservées dans la suite des analyses. Ainsi, les analyses multi-variées ont sélectionné 10 protéines pour le muscle, 10 pour le foie et 15 pour le tissu adipeux.

En compléments des résultats précédents, les analyses que j'ai conduites sur ces données animales durant la thèse avaient pour objectifs d'identifier les signatures moléculaires, et parmi elles des biomarqueurs potentiels, des 7 phénotypes de la composition tissulaire ou chimique des carcasses, en combinant des méthodes régressives.

Integrating liver, muscle and adipocyte tissue proteomes to identify drivers of body and growth composition in bovine species

Mardoc E. [a,b], Imbert A. [a], Broadbent J. A. [a,d], Byrne K. A [d], Tournayre J. [a], Boby C. [a], Sepchat B. [c], Ortigues-Marty I. [a], Bonnet M. [a]

[a] INRAE, Université Clermont Auvergne, VetAgro Sup, UMR Herbivores, F-63122 Saint-Genès-Champanelle, France

[b] INRAE, Université Clermont Auvergne, UMR 1095 Genetics, Diversity and Ecophysiology of Cereals (GDEC), 5 Chemin de Beaulieu, 63000 Clermont-Ferrand, France

[c] INRAE, Systèmes d'Elevage de Ruminants de Moyenne Montagne Herbipôle (UE1414), 63122, Saint-Genès-Champanelle, France

[d] CSIRO, Agriculture and Food, St Lucia, QLD 4067 Australia

ABSTRACT

Tissular and chemical compositions of the body are important criterion of the economic value of carcasses, and of ruminant efficiency regarding the feed-to-food conversion. Tissular and chemical compositions of carcasses remain difficult to assess and may benefit from prediction based on protein abundance quantification. The aim of the present study was to determine the proteins related to the proportion of muscle, fat and bone masses within the carcass, as well as the amount of lipids, proteins and energy within the empty body weight. Additionally, a criterion of visceral adiposity, was considered as the total number of adipocytes within the perirenal adipose tissue (PRAT). These phenotypic traits were calculated in 17 Charolais bulls divergent by the proportion of adipose tissue and lipids within the carcass as the results of two feeding strategies. Muscle, PRAT and liver samples were subjected to SWATH-MS proteome measurement. Mathematical relationships between animals and proteomic datasets were determined by mono- and multi-blocks analyses. Among the 9157 proteins measured, 31 proteins were related to the one or up to five of the seven traits assayed. The muscular abundance of HIBCH involved in the catabolism of valine, the abundance of the adipose HNRNPA2B1 involved in the balance between adipose tissue and muscle masses, or the abundance of hepatic SNTB2 involved in HDL cholesterol were among the proteins the most related to both the tissular and chemical composition of the carcasses. The identified proteins hitherto poorly investigated in the bovine physiology, yet identified in monogastric models as related to the body composition, are thus interesting putative markers of carcass composition. The relationship between the abundance of these proteins and the carcass composition remains to be quantified in a higher number of bovines.

Keywords

Bovine, tissular and chemical carcass composition, proteomic, multi-block statistical analysis

Graphical abstract

Highlights

- 4054 hepatic, 3085 adipose and 2027 muscular proteins were measured by SWATH-MS proteomics
- 31 proteins were related to one of the 7 bovine traits by way of 2-step multivariate analysis
- among these proteins, some are associated to both the tissular and chemical composition of the carcasses, such as HIBCH (muscle), HNRNPA2B1 (PRAT) or SNTB2 (liver)
- CCT8 (liver), RPS24 (muscle) and DRG1 (muscle) were identified as the proteins with highest correlation to the amount of protein fixed
- OMA1 (liver), CT027 (PRAT) and COX7A2 (muscle) were identified as the proteins with highest correlation to the total number of adipocytes in PRAT

Significance

This manuscript is notable in three aspects. First, this is the first characterization of adipose, muscular and hepatic proteomes within a same animal using a sensitive and quantitative proteomic method based on SWATH-MS. Second, the multi-tissue proteome was analyzed to explain the variability of seven bovine traits related to production economics and nutrient retention — two key challenges for all stakeholders, from farmers to the beef industry. Third, we combine two multivariate analyses in order to provide short lists of proteins related to the seven traits to be tested in a larger population.

Introduction

Producing meat animals with an adequate proportion of adipose and muscular masses in the carcass is an economic challenge for the beef industry. Excess fat deposition is an energy-consuming process for cattle that is costly for farmers and non-efficient regarding feed-to-food conversion. It compromises lean accretion during growth, decreases carcass yield and thus the economic value of beef animals. Regarding farm and beef industry incomes, the economic value of carcasses is linked to the proportion of adipose and muscular masses. This significant economic challenge has motivated research on the growth of adipose tissue relatively to muscle (Bonnet et al., 2010); however, little consensus on the molecular markers of the carcass fat (Baik et al., 2017; Ceciliani et al., 2018; Bazile et al., 2019) or muscle mass (Naserkheil et al., 2022; Santiago et al., 2023; Koo et al., 2023) have been proposed. Moreover, molecular drivers were sought in one tissue from the carcass, while the relative proportion of adipose and muscular masses in the carcass is likely a complex and coordinated whole-body response to optimize nutrient partitioning between AT and muscle (Bonnet et al., 2010). Among the body organs that contribute to the coordinated whole-body response, the liver is a major

sensor of energy status, integrating short and long-term signals (Allen, 2020), and producing up to 90 % of serum protein with potential impact on the regulation of nutrient partitioning (Adams et al., 2013; McFadden, 2020). Thus, the proteome of these tissues may be indicative of these coordinated whole-body responses. Moreover, we have previously shown that protein abundances were able to predict the tenderness and marbling of bovine muscle(Bonnet et al., 2020; Picard et al., 2023). Hence, we hypothesized that an in-depth profiling and mining of proteomes from the main body tissues or organs may help to identify not only the molecular signature or the drivers of the proportion adipose and muscular masses in the bovine carcass, but also the interplay between tissues contributing to carcass or body composition. The first aim of the present study is thus to identify adipose, muscular or hepatic proteins related to seven traits of the carcass or body composition using a model of bulls divergent by the proportion of adipose tissue in the carcasses and of lipids in the empty body as a result of nutritional interventions (Sepchat et al., 2013). To select lists of proteins related to phenotypes, our second aim is to combine two multivariate statistical method. Two complementary approaches proposed by the R package *mixOmics* were used, based on the integration of omics (proteomics) and phenotypic datasets: 1- Projection to Latent Structure (PLS) regressions (Wold, 2006) were run on each tissue for each phenotype, in order to identify the combinations of tissue-specific proteins the most related to each of the seven phenotypes, 2- the multi-block version of the PLS (MB-PLS) regression inspired from generalized Canonical Correlation Analysis (Tenenhaus and Tenenhaus, 2011) was run to integrate proteomic data from all tissues and phenotypes at once, to capture the combination of proteins which, along tissues, best explains the phenotypes' variability. Selected proteins from the two approaches were then compared to determine which of the major relations between tissue-specific proteins and phenotypes are the strongest at the multi-tissue and multi-phenotypes levels, *i.e.* which tissues are the most related to each phenotype and through the contribution of which proteins. By this way, we identified, among 9157 proteins measured, 31 putative biomarkers of the chemical or tissular composition of the bovine carcasses.

Material & methods

1. Animals, Rearing, Slaughtering, and Sampling

The animal trial described herein was conducted according to relevant European guidelines (European Union procedures on animal experimentation – Directive 2010/63/EU) and with the French agreement C63 345 17 for the use of production animals in animal experimentation. Animal used in the present study were part of a larger experiment (Sepchat et al., 2013). For the present study, 17 Charolais bulls weaned at 8 months (LW of 360±33 kg) were considered. They were assigned to one of two diet treatment groups, homogenous for

the animal performances between birth and weaning. Bulls were housed in free stall on a semi-strawed area with electronic gates to measure individual intake, one pen per treatment. In order to induce a divergence in tissue (muscle/fat) growth or chemical (protein/lipid) depots in young fattening bulls, two isoenergetic diets were formulated. A wrapped hayage diet of permanent grassland supplemented with citrus pulp and wheat brewery grains (H) was compared with a corn silage diet (C) supplemented with cereals (2/3 wheat 1/3 corn) and rapeseed meal. The dietary starch concentration was respectively 26 and 385 g / kg DM for H and C. The bulls were slaughtered at the same live weight of 700 kg for an average carcass weight of 420 kg. Individual daily feed intake was recorded. Growth rates, slaughter yield and carcass composition were measured. After slaughter, hot carcasses, perirenal adipose tissues and all visceral and subcutaneous adipose tissues were weighed. Carcasses were graded according to the European beef grading system (CE 1249/2008). Within the hour after slaughter, samples of the *longissimus thoracis* (LT), perirenal adipose tissue (PRAT) and liver were excised and immediately frozen in liquid nitrogen and stored at –80°C until analysis.

2. Tissular, cellular and chemicals composition of the carcass or of the body

The bulls were characterized for their carcass composition that refers to production economics, i.e., the proportion of muscle, fat and bone within the carcass, as well as for nutrient conversion to consider the physiological needs of growing animals, i.e., the amount of lipids, proteins and energy within the empty body weight. Additionally, a criterion of visceral adiposity, the total number of adipocytes within the perirenal adipose tissue (PRAT) was calculated to identify proteins related to the hypertrophy/hyperplasia of adipocytes.

The tissular composition of the carcass was calculated according to equations established by Robelin and Geay (Robelin et al., 1975). Briefly the sixth rib was sampled and dissected, and intermuscular fat (Fat6r), muscles (Mu6r) and bones (SQ6r) were separated from each other and weighed to compute the carcass fat, carcass muscle and carcass bone weights, in kg and as followed:

Carcass fat = -21.4 + 0.2172 hot carcass weight + 56.945 Fat6r – 26.645 Mu6r + 1.074 weight of visceral and subcutaneous fats

Carcass muscle = -47.47 + 0.8357 PCC – 42.378 Fat6r + 19.363 Mu6r – 1.638 weight of visceral and subcutaneous fats

Carcass bone = + 0.14 + 75.237 SQ6rc + 11.702 Weight of the 4 cannon bones

The chemical composition of the empty body weight at slaughter (and not the carcass) was computed using the allometric coefficients proposed by Robelin et al. (Robelin et al., 1978) and weight of lipids and proteins in kg, as well as energy in MJ were calculated as followed:

Lipids: 1.1346 x total fat mass (kg)$^{0.992}$

Proteins: 0.1436 x (Empty body Weight-lipids)$^{1.0723}$

Energy: 5.48 x Proteins + 9.37 x lipids.

The total number of adipocytes in PRAT was calculated from adipocyte volume as previously described (Robelin, 1981). Briefly, at slaughter, around 20 g of PRAT was immediately placed at 37°C, fixed by osmium tetraoxyde and isolated by 8 M urea to determine adipocyte volume. The number of adipocytes per gram of lipids was then calculated [1/ (volume (cm^3) x 0.92)] as well as the number of adipocytes per gram of PRAT (number of adipocytes per gram of lipids x percentage of AT lipids). For this purpose total lipid content of AT was determined after extraction by mixing 250 mg of lyophilized tissue with chloroform:methanol, 2:1 (vol/vol) as previously described (Folch et al., 1957). The total number of adipocytes in perirenal AT was calculated (number per gram AT x AT weight).

The statistics related to percent of fat, muscle and bone in carcasses, the weight of empty body's proteins and lipids and the body's stored energy (MJ) are reported in Table 1.

3. Protein extraction and concentration.

All the reagents were purchased from Sigma-Aldrich, Saint Louis, USA (MO), except when specified. Proteins were extracted from either 80 mg of liver or muscle, and from 500 mg of adipose tissue using 1 mL of lysis buffer using a Precellys machine and Precellys tube with CK28R beads (Ozyme). Briefly, frozen tissues for each animal sample (n= 51) were mixed in a buffer containing 50 mM Tris (pH 6.8), 2% SDS, 5% glycerol, 2 mM DTT, 2.5 mM EDTA, 2.5 mM EGTA, 1x HALT Phosphatase inhibitor (Perbio 78420, Perbio Science, Villebon-sur-Yvette, France), Protease inhibitor cocktail complete MINI EDTA-free (Roche, Meylan Cedex France, 1 tablet/10 mL), 2 mM Na3VO4 and 10 mM NaF. The extracts were then boiled for 10 min at 100 °C, sonicated to reduce viscosity and centrifuged 10 min at 15,000 rpm. The supernatants were collected and stored at −80 °C until analysis. Protein concentrations were determined with a commercial protein assay (Pierce BCA reducing agent compatible kit, Thermo Scientific, Waltham, Massachusetts, United State, ref 23252) with BSA as standard.

Two hundred micrograms of protein extract was then purified, concentrated and subjected to trypsin digestion using a FASP protocol. Protein extract was applied to a 10 kDa molecular weight cut-off (MWCO) filter (Millipore, Australia) and diluted to 200 µL with UA buffer before centrifugation (20,800 x g, 15 min). The protein on the filter was washed with two 200 µL volumes of UA buffer with centrifugation (20,800 x g, 15 min). To reduce the protein on the filter, dithiothreitol (50 mM, 200 µL) was added and the solution incubated at room temperature for 50 min with shaking. The filter was washed with two 200 µL volumes of UA buffer with centrifugation (20,800 x g, 15 min). To alkylate the cysteine residues, iodoacetamide (50 mM,

100 uL) was applied to the protein on the filter with incubation for 20 min at room temperature in the dark. The filter was washed with two 200 µL volumes of UA buffer with centrifugation (20,800 x g, 15 min). The buffer was exchanged using 50 mM ammonium bicarbonate (pH 8.0) by two consecutive wash/centrifugation steps.

Sequencing grade porcine trypsin (Promega, Alexandria, Australia) was then added (4 µg in 200 µL of 50 mM ammonium bicarbonate, 2 mM CaCl2, which is a ratio of 50:1 (protein :enzyme)) to the protein on the 10 kDa filters and incubated for 16 h at 37°C in a thermomixer. The filters were transferred to fresh centrifuge tubes and the filtrate (digested peptides) was collected following centrifugation (20,800 x g, 10 min). The filters were washed with 200 µL of 100 mM ammonium bicarbonate and the filtrates were combined and lyophilized. The resultant peptides were re-suspended in 100 µL of 0.1% formic acid containing 0.1 pmol/µL of the Biognosys iRT peptides and 4 µL (equivalent to ~8 µg of total protein and 0.2 pmol of iRT) and was analyzed by LC-MS/MS.

4. LC-MS/MS SWATH analysis

a. MS Data Acquisition and Library Generation

Solubilized peptides were analyzed by LC-MS/MS by information-dependent acquisition (IDA) and to determine the optimal SWATH parameters. Therein, the peptide fractions with iRT peptides were chromatographically separated with an Ekspert nanoLC425 (Eksigent, Dublin, CA, United States) directly coupled to a TripleTOF 6600 MS (SCIEX, Redwood City, CA, United States). The peptides were desalted for 5 min on a Protecol C18P (Trajan; 3 µm, 120 Å, 10 mm × 0.3 mm) trap column at a flow rate of 10 µL/min 0.1% FA, and separated on a ChromXP C18 (3 µm, 120 Å, 150 mm × 0.3 mm) column at a flow rate of 5 µL/min at 30˚C. A linear gradient from 3-25% solvent B over 68 min was employed followed by: 5 min from 25% B to 35% B; 2 min 35% B to 80% B; 3 min at 80% B, 80-3% B, 1 min; and 8 min re-equilibration. The solvents were: (A) 5% DMSO, 0.1% FA, 94.9% water; (B) 5% DMSO, 0.1% FA, 90% acetonitrile, 4.9% water. The instrument parameters were: ion spray voltage 4500 V, curtain gas 30 psi, GS1 30 psi and GS2 30 psi, heated interface 150°C. Data were acquired in IDA mode comprising a time-of-flight (TOF)-MS survey scan followed by 30 MS/MS, each with a 40 ms accumulation time. First stage MS analysis was performed in positive ion mode, mass range *m/z* 400−1250 and 0.25 s accumulation time. Tandem mass spectra were acquired on precursor ions >150 counts/s with charge state 2−5 and dynamic exclusion for 15 s with a 100 ppm mass tolerance. Spectra were acquired over the mass range of *m/z* 100−1800 using the manufacturer's rolling

collision energy based on the size and charge of the precursor ion using Analyst TF v1.8.0 software.

b. SWATH-MS Data Acquisition

The individual digested protein extracts (4.0 µL of sample plus 0.4 pmol of iRT peptide standard, Biognosys) were chromatographically separated as described for IDA. The MS source conditions were also identical. The TOF-MS survey scan was collected over the mass range of m/z 400-1250 with a 50 ms accumulation time and the product ion mass spectra were acquired over the mass range m/z 100-1400 with 30 ms accumulation time. Variable window SWATH ranges were determined using the SWATH variable window calculator 1.0 (SCIEX) to identify 100 optimal ranges (including 1 Da overlap) spanning m/z 40000-12500 and resulting in a 3.1 s cycle time. Collision energy (CE) was determined using each window center as the input m/z for CE equations and a CE spread of 5 eV was used to allow for m/z variance across each SWATH window. Samples were acquired in a fully randomized design for each tissue, with pooled biological quality control samples interspersed periodically within batches to allow assessment of acquisition quality.

c. SWATH-MS DIA-Neural Network (DIA-NN) analysis

A comprehensive DIA-Neural Network analysis of the tissues (muscle, PRAT and liver) was performed. DIA-NN is a universal software for data-independent acquisition (DIA) proteomics data processing (https://github.com/vdemichev/diann) (Demichev et al., 2020). The mass spectrometry data was annotated by searching against a UniProt-Bos database appended to a common repository of adventitious proteins (cRAP) +iRT (U7niProt-filtered-proteome_ 9136_ Bos_ 20210520+cRAP+iRT) (75,268 proteins). Peak extraction was performed by a library-free approach, with no missed cleavages, no variable modifications, fixed modification of C by carbamidomethylation, 7-30 amino acid peptide length, with 2+ charge state, and mass ranges set to match acquisition settings. A representative log file with the complete settings used in the DIA-NN analysis are provided in supplemental material.

d. Pretreatment

Data were logged using $\log_2(1+x)$ function. All proteins containing more than 20% of missing values for at least one of the two diet groups were then removed. Finally, all other missing values were imputed by the average of the 5% lowest values by tissue. This pretreatment resulted in three proteome datasets containing respectively 4054 liver, 3085 fat and 2027 muscle proteins.

5. Statistical and Gene Ontology Analyses

Data were analyzed using R (v.4.2.1) (R Core Team, 2022), especially with Projection to Latent Structure (PLS) regressions and their multi-block version (MB-PLS) available in the R package mixOmics (v.6.20.0) (Rohart et al., 2017; Lê Cao and Welham, 2021). The main objective of these regressions is to compute linear combinations of the explicative variables (*e.g.* proteins) resulting in components maximizing their covariance with the explicated data (*e.g.* phenotypes). For PLS regressions, the explicative data consist in one single dataset and the explicated data *Y* consist in one single variable. MB-PLS generalizes these regressions for several explicative datasets and one or several explicated variables.

PLS regressions were run with the explicative dataset containing proteins variability from one tissue and the explicated variable *Y* corresponding to one of the seven phenotypes. Each combination was analyzed on centered and scaled data, leading to 3 tissues x 7 phenotypes = 21 PLS regressions. Firstly, only 2 components were computed by regression as we assumed more would be unnecessary for explaining one single phenotypic variable. To decide to finally keep 1 or 2 components, we used the *mixOmics* function *perf* to cross-validate results with 5 folds and 20 repeats, the Q2 criterion was used with a 0.0975 threshold and revealed that the second component never was necessary. For each regression, the top 10 most explicative proteins of the component were extracted.

The MB-PLS was run using the *mixOmics* function *block.spls* with three explicative datasets (one per tissue) and the seven phenotypes as explicated variables, all centered and scaled. The weights of interactions between the different blocks, stored in a design matrix, have been set to 1 for interactions between each proteomics dataset and the phenotypes, 0.1 between proteomics datasets, and 0 for a dataset with itself, in order to mainly maximize interactions between proteins and phenotypes but while partially maximizing interactions between proteins from different tissues. Firstly, the sparsity mode was used to select 30 proteins for each of 3 components and tissue, leading in 30 proteins * 3 components * 3 tissues = 270 extracted proteins. Finally, we kept only the proteins the most involved in each component by setting *loadings'* thresholds at 0.2, when *loadings* are *block.spls*' outputs containing the weights of each protein on each component. This resulted in a final selection of 70 proteins.

All the gene ontology enrichment analyses were performed using *B. Taurus* genome annotation and calculated based on hypergeometric distribution followed by false discovery rate (FDR) correction, with a

corrected p-value cut-off of 0.05. ProteInside web-service (https://umrh-bioinfo.clermont.inrae.fr/ProteINSIDE_2/) (Kaspric et al., 2015) was used to perform enrichment analysis in the Gene Ontology (GO) term Biological Processes (BP), Molecular Function (MF) and Cellular Component (CC). These analyses take into account only reviewed proteins, except if there is no reviewed proteins for an user query. In that case, all unreviewed proteins are selected. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed with ShinyGO web-service 0.77 (http://bioinformatics.sdstate.edu/go/) (Ge et al., 2020).

Results

1. Variability of the phenotypes in the 17 Charolais bulls

The proportions of AT and muscle within the hot carcass weight varied by 6 % among the bulls of the present study. The mass of lipids and proteins varied by 51 and 22 kg for an average carcass weight of 420 kg, indicating a higher variability of the lipid than protein deposition. The total number of adipocytes varied by 3.5-fold.

2. Tissue-specific relationships between proteins and a phenotype: PLS approach

For each tissue and phenotype, the 10 proteins the most explicative of the phenotype were extracted, leading to 21 lists of 10 proteins available in Supplementary Tables 1, 2 and 3. A total of 149 proteins were extracted this way, with some proteins present in several lists. Table 2 reports the mean and the range of correlations between the abundance of these proteins in each tissue for each trait. In the liver and PRAT, proteins are less correlated to Carcass_bone than to other phenotypes. The correlations between these protein abundances and phenotypes were grouped in two clusters. One included 31 proteins strongly positively correlated (average r = 0.68) with carcass fat, fixed lipids or fixed energy while negatively correlated (average r = -0.60) with carcass muscle and fixed proteins (Figure 1). These proteins were annotated or related to pathways such as oxidoreductase activity, aldehyde deshydrogenase activity, retinol binding, fatty acid degradation (Supplementary Table 4). The second cluster included 55 proteins with opposite relationships; their abundances were positively correlated (average r = 0.62) with carcass muscle and fixed proteins but negatively correlated (average r = -0.64) with fat, lipids or energy depositions (Figure 1). According to enrichment analyses, these proteins were related to pathways such as negative regulation of proteolysis, negative regulation of hydrolase activity, complement activation, or valine catabolic process (Supplementary Table 4). The proteins related to several phenotypes were summed up in the Supplemental Figure 1, resulting in 13 liver, 13 muscular and 14 adipose proteins. The total number of perirenal adipocytes as well as the

proportion of bone within the carcass were poorly correlated to the protein abundance assayed in the 3 tissues.

3. Relationships between proteins and phenotypes with all phenotypes and tissues considered simultaneously: MB-PLS approach

A total of 270 proteins was extracted by MB-PLS to explain simultaneously the seven phenotypes across tissues. Figure 2.A represents correlations between these proteins/phenotypes and the three components. Five phenotypes are highly correlated to the first component (Carcass_muscle: 0.96, Fixed_proteins: 0.72, Carcass_AT: -0.97, Fixed_lipids: -0.99 and Fixed_energy: -0.95), while the two other phenotypes are more correlated to the second (Carcass_bone: 0.82 and Total_nb_perirenal_adipocytes: 0.61) and third (Total_nb_perirenal_adipocytes: -0.78) components. The muscle carcass proportion was positively correlated to proteins fixation, and both negatively correlated to the adipocyte tissues carcass and lipids and energy fixations. The bone carcass and the total number of perirenal adipocytes are positively correlated on the second component, and negatively correlated on the third one. The number of proteins with positive and negative values on each component was well balanced, and were identified both in AT, muscle or liver.

Of these 270 proteins, we selected 70 final proteins with the highest explicative value on a component (Figure 2.B). Hence, 19 proteins were finally selected on the first component and were associated to Carcass_muscle, Fixed_proteins_ Carcass_AT, Fixed_lipids and Fixed_energy, 23 on the second component opposing Carcass_bone and Total_nb_perirenal_adipocytes to Fixed_proteins, and 28 on the third component mainly associated to Total_nb_perirenal_adipocytes. Proteins from the three tissues were used to compute each component, and the number of proteins with positive and negative values on each component seemed well-balanced, meaning respectively that phenotypes were not specifically associated to one tissue nor to positively or negatively correlated proteins. Supplementary Table 5 presents correlations between these 70 proteins and the phenotypes highly represented in the corresponding component.

4. Comparison of lists of proteins selected by PLS and multi-block PLS

The 21 lists of proteins from PLS regressions were compared to the 70 proteins from the MB-PLS regression. Supplementary Figure 2 displays the proteins from the 21 PLS lists and indicates if they are or not selected by MB-PLS. The 31 proteins at the intersection between PLS and MB-PLS results were listed in Table 3 and considered as having the most robust relationships with a phenotype. Indeed, proteins selected by both approaches were: the most explicative of a phenotype within a unique tissue according to PLS, the most related to one or more phenotypes even when tissues and phenotypes were considered simultaneously within the MB-PLS. In contrary, proteins only selected by PLS are still the most explicative proteins for these

specific tissue and phenotype, but with relatively low interactions compared to some other proteins-phenotypes interactions, while proteins selected only by MB-PLS are not the most explicative within a tissue but only when associated with proteins from other tissues to partially explain one or more phenotypes. Thus, in the liver, CCT8 was positively correlated to the muscle mass and proteins deposition and inversely correlated to the fat mass, lipids and energy depositions. The reverse was observed for SNTB2. The proteins DECR1 and PGP were positively correlated to the deposition of lipids and energy, plus fat mass for DECR1 only. The proteins OMA1 and HDHD2 were negatively and positively correlated to the number of PRAT adipocytes. No liver protein has been selected by both approaches for explaining the Carcass_bone. Within PRAT, FGB, ITIH1, HNRNPA2B1 and RIDA were related to fat and muscle masses, lipids and energy depositions; RBM14, DHX9, LOC788425 and ESAM were related to lipids or energy depositions; CT027, GMDS, MAL2 and COQ6 to the number of PRAT adipocytes, PLVAP to the proposition of bone in the carcass. Any fat proteins were related to the Fixed_proteins, suggesting a very low impact of this tissue on this phenotype. In muscle, RPS24, HIBCH, NAA15|NAA16 and DRG1 were related to the tissular and chemical carcass composition, while BCAM was exclusively related to the muscle proportion within the carcass; UBQLN1 and H4 to the muscle proportion in the carcass, ITIH5; SNRPB|SNRPN, COX7A2 and NAP1L1 to the total number of PRAT adipocytes and GPX4 to the amount of fixed lipids.

Discussion

The tissular and chemical compositions of the body are interconnected (Owens et al., 1995); however they provide complementary information with potential divergent application for end-users. Biological pathways and putative markers of the amounts of lipids, proteins and energy fixed in whole body of growing cattle are knowledge that should help to manage the physiological needs of growing animals, to quantify the nutritional value of derived meat, or to identify efficient animals regarding feed-to-food conversion. Pathways and markers related to the carcass composition and especially the proportion of fat and muscle may help to manage these traits that are related to production economics, since in Europe most of the carcass grading systems consider the fat proportion over the carcass. Here we propose pathways and putative markers of the tissular and molecular body composition thanks to a multi-tissue proteomic integrative analysis.

Such approaches were rarely used in rearing animals. For understanding mechanisms and predicting beef tenderness using protein data, Gagaoua reviewed 28 experiments showing that the classical approaches are mainly based on correlations between variables, identification of differential abundant proteins (DAP) using statistical tests or regressions (Gagaoua et al., 2021), while other analyses also use approaches inspired from the principal component analysis (PCA), especially the multi-factor analysis (MFA), for identifying relationships between muscle protein abundances and several meat and quality traits in beefs (Picard et al.,

2023), or muscle, liver and adipose transcriptomes and feed efficiency traits in pigs (Gondret et al., 2017). Moreover, networks methods were also used on more complex data (multi-tissues, multi-traits and/or multi-omics data), to construct for instance networks of gene co-expression from genome-wide association studies (GWAS), RNA-Seq, and bovine transcription factors (Cánovas et al., 2014); of weighted correlation network analysis (WGCNA), RNA-Seq, and bovine carcass traits (Silva-Vignato et al., 2019), or of differentially regulated metabolic reactions in mice liver and skeletal muscles (Egami et al., 2021). Finally, human multi-tissues data were also integrated using Bayesian model-based multi-tissue clustering algorithm developed by Erola (Erola et al., 2020). However, such very informative methods require more individual observations than those available in the present study. Thus, to select proteins highly related to one or several chosen traits, first by tissue, then with all tissues at once to catch more complex biological interactions, we used two complementary approaches. Of these, the multi-blocks generalization of the PLS regression was used to identify the best combination of proteins explaining the phenotypes. Indeed, while networks cluster similar biological entities and find the cluster the most correlated to each phenotype, the MB-PLS selects complementary entities to get a better explicative combination of proteins while avoiding the selection of proteins of similar behaviors.

Following this 2-step strategy, we identified pathways and 31 putative markers of both the tissular and chemical composition of the carcass or of the total number of adipocytes in PRAT. While in-depth hepatic, muscular and adipose proteomes were poorly related to the bone proportion within the carcass, which may be consistent with the low variability of this phenotype in the present study, they were more related to muscle mass proportion of protein deposition, as well as to the fat mass proportion and the lipid or energy deposition.

1. Molecular signature of both fat and muscle proportion in the carcass

Two muscular (Hibch, RPS24), 3 adipose (FGB, ITIH1, HNRNPA2B1) and 2 hepatic (CCT8, SNTB2) proteins were identified as inversely related to the fat or muscle masses suggesting that they may be involved in the competition or prioritization between adipose and muscle cells for the uptake and metabolism of nutrients that occurs during bovine growth (Bonnet et al., 2010). The dynamic control of nutrient partitioning and storage depends on balancing energy demand of the tissue with energy supply. In ruminants, this balance is maintained through the integration of many different signals that communicate the nutrient status of the organism to the periphery, including insulin, glucose but also proteins secreted by the tissues such as hepatokines (McFadden, 2020) adipokines or myokines (Bonnet et al., 2016). Some of the putative markers of the bovine tissular carcass composition identified in the present study, may contribute to this complex regulation of the nutrient partitioning.

Indeed, within muscle, Thioesterase 3-hydroxyisobutyryl-CoA hydrolase (Hibch), which controls the formation of 3-hydroxyisobutyrate in the valine degradation pathway was previously related to the body composition of monogastric species through a regulation of the muscular insulin sensitivity. High level of hydroxyisobutyrate, the product of Hibch activity, was related to an increase in basal skeletal muscle glucose uptake which drives glucotoxicity and impairs myocyte insulin signaling (Bishop et al., 2022). Moreover, plasma 3-hydroxyisobutyrate correlated positively to plasma and hepatic concentrations of TAG, plasma total fatty acids, plasma NEFA and insulin/glucose ratio in human (Bjune et al., 2021). Thus, the positive correlation between Hibch and muscle mass proportion may result from an insulin and glucose-related signal partly mediated by the valine catabolism pathway. Of note, even if identified only by the PLS analysis, the muscular abundance of ALDH1A1 and FHL1 were related to the carcass composition in the present study as in a previous study in the Rouge des Pres breed (Bazile et al., 2019) .

Within PRAT, the reverse relationship between the abundance of Heterogeneous nuclear ribonucleoprotein (hnRNP) A2/B1 and the carcass fat mass observed in the present study agrees with previous mice results. Indeed, an overabundance of HNRNPA2B1 in the inguinal white AT of mice decreased body weight, fat mass, and adipose tissue weights with smaller adipocyte sizes when they were subjected to diet induced AT deposition. At the opposite, a specific knockdown of HNRNPA2B1 expression in inguinal white AT of mice was reported to increase body weights and adiposity, decrease oxygen consumption, impaired cold tolerance, as well as to induce a more severe insulin resistance and hepatic steatosis compared with wild type mice (Li et al., 2022). Moreover, over expression of HNRNPA2B1 in bovine Skeletal muscle satellite cells (MuSCs) was shown to promote myogenesis (Zhang et al., 2023) by a regulatory mechanism that remains to be studied. A striking result of the present study is that 6 proteins mainly synthetized and secreted by the liver, that are thus well-known hepatokines, were quantified in the adipose tissue. All of them have been related to insulin-stimulated glucose disposal, insulin sensitivity in extrahepatic tissue or body fats. These proteins are the beta (FGB) and gamma (FGG) components of fibrinogen, two isoforms of the inter-alpha trypsin inhibitor (ITIH1 and ITIH2), two proteins related to the complement system (C3, C8_A) and one apolipoprotein (APOA2). Of these, ITIH1 and FGB only were in the short list of potential markers of carcass tissular composition identified by the 2 methods. Low abundances of FGB and FGG in the PRAT may suggest a low adipocyte inflammation and a maintained insulin-stimulated glucose disposal favoring AT growth if we assume that the biological actions of fibrinogen are the same in mice (Kang et al., 2015) and in bovine. The ITIH1 protein is known to be synthesized exclusively in hepatocytes and secreted into the bloodstream, stabilizing the extracellular matrix surrounding targeted tissues such as adipose tissue and skeletal muscle. In vivo, ITIH1 neutralization using an anti-ITIH1 antibody, increased insulin-stimulated glucose uptake in adipose tissue or skeletal muscle of wild-type mice fed a high-fat diet (Kim et al., 2019). Thus the adipose

proteins related to the tissular carcass composition are probably related to regulatory pathways sensing glucose availability and insulin sensitivity.

Within the liver, the syntrophins beta 2 (SNTB2) is a molecular adaptor protein shown to stabilize ABCA1, an essential regulator of HDL cholesterol. When challenged with a high fat diet, SNTA/B2-/- mice have similar weight gain, adiposity, serum and liver triglycerides than wild type but the glucose tolerance is impaired (Hebel et al., 2015). The Chaperonin containing TCP1 subunit 8 (CCT8) contributes to a large multi-subunit complex that mediates protein folding —especially of actin and tubulin— in eukaryotic cells but its role in animal development and growth remains to be largely explored. However, evidence suggest that CCT8 contributes to pathways that maintain the quality of cytoskeletal proteins ('proteostasis'), ensuring the appropriate function of microfilaments and microtubules including the signaling of growth factors, hormones or nutrients (Lundin et al., 2010).

Collectively the proteins that we identified associated with both the muscular and adipose masses may be involved in competition or prioritization between adipose and muscle cells for the uptake and metabolism of nutrients through the insulin sensitivity regulation.

2. A Molecular signature of both fixed lipids and fixed energy in the carcass, partly shared with the amount of fixed proteins

The hepatic abundance of SNTB2, CCT8, the adipose abundance of ITIH1, HNRNPA2B1 and the muscular abundance of RPS24 and HIBCH were not only related to the tissular composition of the carcass but also to the deposition of lipids and energy. In addition, the abundance of DECR1 and PGP in the liver; of RBM14, AKR7A3 and ESAM in PRAT; of NAA15|NAA16 and DRG1 in muscle were specifically related to lipids and energy deposition.

Within the liver, both DECR1 and PGP are mainly related to lipid metabolism pathways, which is consistent with the positive relationships between their abundances and the amount of fixed lipids and energy. The 2,4-dienoyl-CoA reductase, DECR, is one of the enzymes involved in the beta-oxidation of fatty acids. Hepatic oxidation of fatty acids mainly results in the production of acetyl-CoA which is then either completely oxidized in the TCA cycle or exported as ketone bodies or acetate; both contributing to the energy or lipid deposition at the body level. Recently, the phosphoglycolate phosphatase, PGP, was proposed to catalyze the hydrolysis of glycerol-3-phosphate (Possik et al., 2021) which forms the backbone of glycerolipids including triglycerides and phospholipids, a large part of the body lipids. Recent evidences suggest that PGP-related glucose and lipid metabolisms in hepatocytes depend on the concentration of glucose (Possik et al., 2021), probably by actively modulating the liver core of 2C-3C partition of energy substrates towards lipid

deposition. However, the physiological functions of these 2 proteins in the context of bovine growth remains to be studied.

Regarding proteins identified in PRAT, the Aldo-keto reductase family 7-member A3 AKR7A3, mainly found in liver but not yet in adipose tissue, catalyzes the reduction of aldehydes thus playing a role of in detoxification. However previous results have shown that other AKR isoforms were either positively (Ostinelli et al., 2021) or negatively (Volat et al., 2012) related to the accumulation of lipids in the adipose tissues, which suggest that AKR7A3 identified in the present study may be a novel adipose isoform favoring lipid and energy deposition in bovine. The RNA-binding protein 14, RBM14, a hnRNP-like protein as HNRNPA2B1, both being negatively associated with lipids deposition, was previously shown to positively regulate adipocyte differentiation (Firmin et al., 2017). The role of RBM14 in differentiated cells or in lipid accumulation was not yet studied, but we could hypothesize that this protein is involved in basal energy expenditure as HNRNPA2B1. The negative relationship of ESAM (endothelial cell selective adhesion molecule) with fixed energy and lipids, is in line with the relationship between this protein and the adiponectin plasma concentration (Kacso et al., 2018), an adipokine with a concentration inversely related to adiposity in cattle (Sauerwein and Häußler, 2016).

As stated above for the carcass composition, the role of the muscular RPS24 in protein folding and of HIBCH as an insulin and glucose-related signaling, is also consistent with their positive and negative relationships with lipid or energy deposition. However, the biological link between the abundances of NAA15/16, a N-terminal acetyltransferase A (NatA) complex which displays alpha (N-terminal) acetyltransferase (NAT) activity, or the developmentally-regulated GTP-binding protein 1, DRG1 (also called NEDD3 for ubiquitin E3-type ligases), and the tissular or chemical carcass compositions warrant further studies.

Of the proteins found to be related to lipid and energy deposition by 2 mathematical models, CCT8 in liver, RPS24 and DRG1 in muscle were also related to the amount of protein fixed. They are linked to protein folding and the regulation of proteolysis according to their annotation in the uniport database.

Collectively the proteins that we identified associated with both lipids and energy deposition and fat or muscle mass proportion are known to be involved in the insulin-related signaling; while those that are specifically related to lipids and energy deposition were related to lipid metabolism or adipokine signaling.

3. Molecular signatures of the number of adipocytes.

Excess deposition of nutrients in visceral adiposity decreases the feed-to-conversion ratio. Thus, identifying proteins related to visceral fat may help to control the visceral adiposity that depends on the number and

485 the volume of adipocytes within the fat tissues such as PRAT. The relative contributions of hyperplasia and
486 hypertrophy to the growth of AT vary among tissue locations and age. Hyperplasia mainly occurs during foetal
487 and/or early post-natal life, but it can also arise during later stages or in adult life in cattle such as when an
488 average diameter of approximately 90 for the adipocytes is reached (Bonnet et al., 2010). Hypertrophy begins
489 during foetal life and is the main mechanism by which AT grows after birth thanks to lipid synthesis (in
490 ruminant de novo lipogenesis occurs in AT and not in liver such as Human) and consequently lipid storage.
491 The best assessment of the visceral adiposity is the counting of the total number of adipocytes within the
492 tissue, since it considers both the number and the volume of adipocytes. In the present study we identified
493 2 hepatic (OMA1, HDHD2), 4 adipose (CT027, GMDS, MAL2, COQ6) and 4 muscular (ITIH5, SNRPB|SNRPN,
494 COX7A2, NAP1L1) proteins related to the total number of adipocytes in PRAT.

495 Of these, the hepatic zinc metallopeptidase OMA1 was related to obesity phenotypes through a
496 mitochondrial quality control protease, inactivating mitochondrial fusion and energy expenditure (Quirós et
497 al., 2013). The adipose CT027 that corresponds to an uncharacterized adipokine that was name UPF0687
498 protein / human C20orf27 homolog or Adissp (Adipose-secreted signaling protein) was recently proposed to
499 be a key regulator of white adipose tissue (WAT) thermogenesis and glucose homeostasis (Chen et al., 2022).
500 Indeed, Adissp knockout mice are defective in WAT browning, and are susceptible to high fat diet-induced
501 obesity and hyperglycaemia. The muscular cytochrome c oxidase-subunit 7A2 gene COX7A2 is a
502 mitochondrial protein found in many tissues, with key roles in steroidogenesis and oxidative metabolism.
503 Differential blood mRNA abundance of COXAA was found between lean and non-lean women, and a reverse
504 relationship was observed between the abundance of COXA7 in adipose tissue and the women adiposity
505 (Carty et al., 2014). These results suggest that low abundance of OMA1, Adissp, COXA7 may favor adipocytes
506 hypertrophy, in line with the negative relationship between OMA1, Adissp or COXA7 and the total number
507 of adipocytes in the present study. However, the underlying mechanisms remain to the explored.

508

509 Conclusion

510 Results from the present study show that the combination of PLS including one phenotype and one proteomic
511 dataset, plus a multiblock PLS including several phenotypes and proteomics datasets, was efficient to identify
512 markers and pathways related to bovine traits. Using this 2-step analysis approach, we identified 31 proteins
513 among 9157 proteins from bovine tissues or liver. These 31 proteins were not or rarely described in the
514 context of the bovine physiology or growth, probably because these proteins were not yet identified by the
515 other proteomics methods commonly used until now in bovine. In the medical field (human or rodent), the
516 31 identified proteins are very relevant with the regulatory and metabolic pathways involved in nutrient
517 partitioning or tissue growth. Thus, future work will address the relationships between the abundance of

these 31 proteins and the bovine phenotypes in a larger population in order to feed a biomarker discovery pipeline previously defined (Bonnet et al., 2020).

Authors contribution

Emile Mardoc: Software, Formal analysis, Data Curation, Visualization, Writing - original draft

Alyssa Imbert: Software, Methodology, Validation, Visualization, Project, Reviewing – original draft, Supervision of Emile Mardoc

James Broadbent: Investigation; Writing - Review & Editing; Supervision of Keren Byrne; Resources

Keren Byrne: Investigation; Writing - Review & Editing; Data Curation

Muriel Bonnet: Conceptualization, Investigation, Methodology, Validation, Visualization, Project administration, Funding acquisition, Writing - original draft, Supervision of Emile Mardoc.

Bernard Sepchat: Methodology, Validation, Writing, Review & Editing

Isabelle Ortigues-Marty: Methodology, Validation, Writing, Review & Editing

Declaration of Competing Interest

Authors declare no conflict of interests

Aknowledgement

The authors thank Region Auvergne Rhone Alpe for funding the proteomics analyses through the call “Soutien aux coopérations universitaires et scientifiques internationales 2017” (subvention number 17 009402 02). The authors acknowledge the support received from the Agence Nationale de la Recherche of the French government through the program ‘‘Investissements d’Avenir’’ (16-IDEX-0001 CAP 20-25), for the funding of the PhD grant of Emile Mardoc.

Tables

Table 1: Description of the 7 phenotypes by basic metrics. The table presents average (mean), standard deviation (SD), minimum and maximum values of the 7 phenotypes on the 17 individuals.

Phenotypic variable	Mean	SD	Minimum	Maximum
Carcass_AT, % hot carcass weight	14,82	1,912	10,2	16,92
Carcass_muscle, % hot carcass weight	71,96	1,662	70,1	76,18
Carcass_bone, % hot carcass weight	14,79	0,6654	13,7	16,17
Total_nb_perirenal_adipocytes, 10^6	6052	2786	3860	13780
Fixed_lipids, kg	56,56	14,85	26,8	78,14
Fixed_proteins, kg	66,75	6,317	58,7	80,02
Fixed_energy, MJ	3750	495	2768	4437

543

544 **Table 2: Average absolute Pearson correlations between top 10 proteins and corresponding phenotypes.**
545 For each protein selected, the absolute value of the Pearson correlation was computed between this protein
546 and the corresponding phenotype. The average of the 10 values is indicated for each phenotype and tissue,
547 with the minimum and maximum values in parentheses.

	Liver	Fat	Muscle
Carcass_AT	0,80 [0,78 - 0,86]	0,80 [0,77 - 0,85]	0,75 [0,71 - 0,90]
Carcass_muscle	0,79 [0,77 - 0,86]	0,81 [0,77 - 0,86]	0,75 [0,71 - 0,92]
Carcass_bone	0,68 [0,65 - 0,73]	0,65 [0,62 - 0,70]	0,72 [0,69 - 0,81]
Total_nb_perirenal_adipocytes	0,76 [0,72 - 0,87]	0,70 [0,64 - 0,86]	0,70 [0,63 - 0,83]
Fixed_lipids	0,81 [0,76 - 0,91]	0,83 [0,80 - 0,89]	0,73 [0,66 - 0,85]
Fixed_proteins	0,76 [0,73 - 0,81]	0,77 [0,75 - 0,82]	0,68 [0,64 - 0,76]
Fixed_energy	0,80 [0,76 - 0,86]	0,80 [0,77 - 0,86]	0,73 [0,64 - 0,90]

548

549 **Table 3: Positive and negative correlations between proteins selected by both PLS and MB-PLS regressions**
550 **and phenotypes.**

551 Tissue and corresponding gene of each protein selected by PLS and MB-PLS regressions, plus the sign of
552 correlation with associated phenotypes: red for a positive Spearman correlation, blue for a negative
553 correlation.

Tissue	Protein Name	Gene Name	Carcass_AT	Carcass_muscle	Carcass_bone	Total_nb_perirenal_adipocytes	Fixed_lipids	Fixed_proteins	Fixed_energy
Liver	TCPQ	CCT8							
Liver	F6QN89	SNTB2							
Liver	A0A3Q1LNW7.F1N5J8	DECR1							
Liver	OMA1	OMA1							
Liver	HDHD2	HDHD2							
Liver	PGP	PGP							
PRAT	A0A3Q1MG04.F1MAV0	FGB							
PRAT	ITIH1	ITIH1							
PRAT	A0A3Q1MME4	HNRNPA2B1							
PRAT	RIDA	RIDA							
PRAT	Q3ZC85	PLVAP							
PRAT	CT027	CT027							
PRAT	Q3ZBY8	GMDS							
PRAT	MAL2	MAL2							
PRAT	COQ6	COQ6							
PRAT	RBM14	RBM14							
PRAT	DHX9	DHX9							
PRAT	F1N6I4	LOC788425							
PRAT	Q2KJA9	ESAM							
Muscle	RS24	RPS24							
Muscle	HIBCH	HIBCH							
Muscle	A0A3Q1M9Y4.A0A3Q1MN02.A0A3Q1NFQ2.E1BMF6.F1N4V5	NAA15|NAA16							
Muscle	DRG1	DRG1							
Muscle	BCAM	BCAM							
Muscle	F1N1A3	UBQLN1							
Muscle	H4	H4							
Muscle	ITIH5	ITIH5							
Muscle	RSMB.RSMN	SNRPB|SNRPN							
Muscle	CX7A2	COX7A2							
Muscle	NP1L1	NAP1L1							
Muscle	GPX4	GPX4							

Figures

Figure 1: Heatmap of correlations between PLS selected proteins and the seven phenotypes.

Spearman correlations between each protein selected by PLS and the seven phenotypes, represented by a color code. Proteins and phenotypes are clustered using a hierarchical clustering with the Euclidean distance and Ward.D2 method. Two groups of proteins highly correlated to phenotypes were extracted.

Figure 2: Phenotypes and selected proteins on the first components of the multi-block PLS.

A- Dot plots of proteins and phenotypes on pairs of components computed using the *mixOmics*' function *block.spls* and displayed with the function *plotVar*. The seven phenotypes are directly displayed with their name, while the 30 most explicative proteins by tissue on each component are represented by color points according to their tissue provenance. X- and Y- axes values represent the relative implication of each variable on the corresponding component. **Left-** X- and Y- axes correspond respectively to the 1st and 2nd components. **Right-** X- and Y- axes correspond respectively to the 1st and 3rd component. **B-** *mixOmics*' *plotLoadings* of the most explicative proteins on each component, *i.e.* all proteins explaining at least 20% of the component, with weights represented by the plot scale. For each protein selected this way and identified by its Uniprot ID without the "_BOVIN" termination, its corresponding genes are indicated in parentheses and its tissue's origin by the color code.

Supplementary Figures

Supplementary Figure 1: Number of proteins selected for one or several phenotypes by PLS regressions.

Plots by tissues (liver, muscle, fat) of the number of proteins selected for one or several of the seven phenotypes, using the top 10 proteins lists from PLS regressions. For proteins selected for several phenotypes, their name and associated gene(s) are displayed.

Supplementary Figure 2: Comparison of proteins lists from PLS and MB-PLS regressions.

The top 10 proteins selected from each tissue (Liver, Fat, Muscle) on each of the 7 phenotypes using PLS regression with the *mixOmics*' function *pls* are displayed and compared to the list of proteins from the multi-block PLS regression *block.spls*. The proteins' color code indicates a positive or negative correlation between the protein and the corresponding phenotype. Each protein is identified with its Uniprot ID (without the "_BOVIN" termination), and its corresponding genes are indicated in parentheses if known.

Supplementary Tables

Supplementary Table 1: Top 10 liver proteins from PLS regression

Supplementary Table 2: Top 10 fat proteins from PLS regression

Supplementary Table 3: Top 10 muscle proteins from PLS regression

Supplementary Table 4: Functional analysis of the 2 clusters identified from the list of the 149 proteins the most explicative of each phenotype for each tissue

Supplementary Table 5: List of proteins selected by MB-PLS regression and correlations with phenotypes

Works Cited

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. Récupéré sur https://www.R-project.org/

References

Adams, A.C., Yang, C., Coskun, T., Cheng, C.C., Gimeno, R.E., Luo, Y., Kharitonenkov, A., 2013. The breadth of FGF21's metabolic actions are governed by FGFR1 in adipose tissue. Molecular Metabolism 2, 31–37. https://doi.org/10.1016/j.molmet.2012.08.007

Allen, M.S., 2020. Review: Control of feed intake by hepatic oxidation in ruminant animals: integration of homeostasis and homeorhesis. Animal 14, s55–s64. https://doi.org/10.1017/S1751731119003215

Baik, M., Kang, H.J., Park, S.J., Na, S.W., Piao, M., Kim, S.Y., Fassah, D.M., Moon, Y.S., 2017. TRIENNIAL GROWTH AND DEVELOPMENT SYMPOSIUM: Molecular mechanisms related to bovine intramuscular fat deposition in the longissimus muscle12. Journal of Animal Science 95, 2284–2303. https://doi.org/10.2527/jas.2016.1160

Bazile, J., Picard, B., Chambon, C., Valais, A., Bonnet, M., 2019. Pathways and biomarkers of marbling and carcass fat deposition in bovine revealed by a combination of gel-based and gel-free proteomic analyses. Meat Science 156, 146–155. https://doi.org/10.1016/j.meatsci.2019.05.018

Bishop, C.A., Machate, T., Henning, T., Henkel, J., Püschel, G., Weber, D., Grune, T., Klaus, S., Weitkunat, K., 2022. Detrimental effects of branched-chain amino acids in glucose tolerance can be attributed to valine induced glucotoxicity in skeletal muscle. Nutr. Diabetes 12, 20. https://doi.org/10.1038/s41387-022-00200-8

Bjune, M.S., Lindquist, C., Hallvardsdotter Stafsnes, M., Bjørndal, B., Bruheim, P., Aloysius, T.A., Nygård, O., Skorve, J., Madsen, L., Dankel, S.N., Berge, R.K., 2021. Plasma 3-hydroxyisobutyrate (3-HIB) and methylmalonic acid (MMA) are markers of hepatic mitochondrial fatty acid oxidation in male Wistar rats. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1866, 158887. https://doi.org/10.1016/j.bbalip.2021.158887

Bonnet, M., Cassar-Malek, I., Chilliard, Y., Picard, B., 2010. Ontogenesis of muscle and adipose tissues and their interactions in ruminants and other species. Animal 4, 1093–1109. https://doi.org/10.1017/S1751731110000601

Bonnet, M., Soulat, J., Bons, J., Léger, S., De Koning, L., Carapito, C., Picard, B., 2020. Quantification of biomarkers for beef meat qualities using a combination of Parallel Reaction Monitoring- and antibody-based proteomics. Food Chemistry 317, 126376. https://doi.org/10.1016/j.foodchem.2020.126376

Bonnet, M., Tournayre, J., Cassar-Malek, I., 2016. Integrated data mining of transcriptomic and proteomic datasets to predict the secretome of adipose tissue and muscle in ruminants. Mol. BioSyst. 12, 2722–2734. https://doi.org/10.1039/C6MB00224B

Cánovas, A., Reverter, A., DeAtley, K.L., Ashley, R.L., Colgrave, M.L., Fortes, M.R.S., Islas-Trejo, A., Lehnert, S., Porto-Neto, L., Rincón, G., Silver, G.A., Snelling, W.M., Medrano, J.F., Thomas, M.G., 2014. Multi-Tissue Omics Analyses Reveal Molecular Regulatory Networks for Puberty in Composite Beef Cattle. PLoS ONE 9, e102551. https://doi.org/10.1371/journal.pone.0102551

Carty, D., Akehurst, C., Savage, R., Sungatullina, L., Robinson, S., McBride, M., McClure, J., Freeman, D., Delles, C., 2014. Differential gene expression in obese pregnancy. Pregnancy Hypertension: An

International Journal of Women's Cardiovascular Health 4, 232–233. https://doi.org/10.1016/j.preghy.2014.03.011

Ceciliani, F., Lecchi, C., Urh, C., Sauerwein, H., 2018. Proteomics and metabolomics characterizing the pathophysiology of adaptive reactions to the metabolic challenges during the transition from late pregnancy to early lactation in dairy cows. Journal of Proteomics 178, 92–106. https://doi.org/10.1016/j.jprot.2017.10.010

Chen, Q., Huang, L., Pan, D., Hu, K., Li, R., Friedline, R.H., Kim, J.K., Zhu, L.J., Guertin, D.A., Wang, Y.-X., 2022. A brown fat-enriched adipokine Adissp controls adipose thermogenesis and glucose homeostasis. Nat Commun 13, 7633. https://doi.org/10.1038/s41467-022-35335-w

Demichev, V., Messner, C.B., Vernardis, S.I., Lilley, K.S., Ralser, M., 2020. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods 17, 41–44. https://doi.org/10.1038/s41592-019-0638-x

Egami, R., Kokaji, T., Hatano, A., Yugi, K., Eto, M., Morita, K., Ohno, S., Fujii, M., Hironaka, K., Uematsu, S., Terakawa, A., Bai, Y., Pan, Y., Tsuchiya, T., Ozaki, H., Inoue, H., Uda, S., Kubota, H., Suzuki, Y., Matsumoto, M., Nakayama, K.I., Hirayama, A., Soga, T., Kuroda, S., 2021. Trans-omic analysis reveals obesity-associated dysregulation of inter-organ metabolic cycles between the liver and skeletal muscle. iScience 24, 102217. https://doi.org/10.1016/j.isci.2021.102217

Erola, P., Björkegren, J.L.M., Michoel, T., 2020. Model-based clustering of multi-tissue gene expression data. Bioinformatics 36, 1807–1813. https://doi.org/10.1093/bioinformatics/btz805

Firmin, F.F., Oger, F., Gheeraert, C., Dubois-Chevalier, J., Vercoutter-Edouart, A.-S., Alzaid, F., Mazuy, C., Dehondt, H., Alexandre, J., Derudas, B., Dhalluin, Q., Ploton, M., Berthier, A., Woitrain, E., Lefebvre, T., Venteclef, N., Pattou, F., Staels, B., Eeckhoute, J., Lefebvre, P., 2017. The RBM14/CoAA-interacting, long intergenic non-coding RNA Paral1 regulates adipogenesis and coactivates the nuclear receptor PPARγ. Sci Rep 7, 14087. https://doi.org/10.1038/s41598-017-14570-y

Folch, J., Lees, M., Stanley, G.H.S., 1957. A SIMPLE METHOD FOR THE ISOLATION AND PURIFICATION OF TOTAL LIPIDES FROM ANIMAL TISSUES. Journal of Biological Chemistry 226, 497–509. https://doi.org/10.1016/S0021-9258(18)64849-5

Gagaoua, M., Terlouw, E.M.C., Mullen, A.M., Franco, D., Warner, R.D., Lorenzo, J.M., Purslow, P.P., Gerrard, D., Hopkins, D.L., Troy, D., Picard, B., 2021. Molecular signatures of beef tenderness: Underlying mechanisms based on integromics of protein biomarkers from multi-platform proteomics studies. Meat Science 172, 108311. https://doi.org/10.1016/j.meatsci.2020.108311

Ge, S.X., Jung, D., Yao, R., 2020. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics 36, 2628–2629. https://doi.org/10.1093/bioinformatics/btz931

Gondret, F., Vincent, A., Houée-Bigot, M., Siegel, A., Lagarrigue, S., Causeur, D., Gilbert, H., Louveau, I., 2017. A transcriptome multi-tissue analysis identifies biological pathways and genes associated with variations in feed efficiency of growing pigs. BMC Genomics 18, 244. https://doi.org/10.1186/s12864-017-3639-0

Hebel, T., Eisinger, K., Neumeier, M., Rein-Fischboeck, L., Pohl, R., Meier, E.M., Boettcher, A., Froehner, S.C., Adams, M.E., Liebisch, G., Krautbauer, S., Buechler, C., 2015. Lipid abnormalities in alpha/beta2-syntrophin null mice are independent from ABCA1. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1851, 527–536. https://doi.org/10.1016/j.bbalip.2015.01.012

675 Kacso, T., Bondor, C.I., Rusu, C.C., Moldovan, D., Trinescu, D., Coman, L.A., Ticala, M., Gavrilas, A.M.,
676 Potra, A.R., 2018. Adiponectin is related to markers of endothelial dysfunction and neoangiogenesis
677 in diabetic patients. Int Urol Nephrol 50, 1661–1666. https://doi.org/10.1007/s11255-018-1890-1

678 Kang, M., Vaughan, R.A., Paton, C.M., 2015. FDP-E induces adipocyte inflammation and suppresses
679 insulin-stimulated glucose disposal: effect of inflammation and obesity on fibrinogen Bβ mRNA.
680 American Journal of Physiology-Cell Physiology 309, C767–C774.
681 https://doi.org/10.1152/ajpcell.00101.2015

682 Kaspric, N., Picard, B., Reichstadt, M., Tournayre, J., Bonnet, M., 2015. ProteINSIDE to Easily
683 Investigate Proteomics Data from Ruminants: Application to Mine Proteome of Adipose and Muscle
684 Tissues in Bovine Foetuses. PLoS ONE 10, e0128086. https://doi.org/10.1371/journal.pone.0128086

685 Kim, T.H., Koo, J.H., Heo, M.J., Han, C.Y., Kim, Y.-I., Park, S.-Y., Cho, I.J., Lee, C.H., Choi, C.S., Lee, J.W.,
686 Kim, W., Cho, J.-Y., Kim, S.G., 2019. Overproduction of inter-α-trypsin inhibitor heavy chain 1 after
687 loss of $G\alpha_{13}$ in liver exacerbates systemic insulin resistance in mice. Sci. Transl. Med. 11, eaan4735.
688 https://doi.org/10.1126/scitranslmed.aan4735

689 Koo, Y., Alkhoder, H., Choi, T.-J., Liu, Z., Reents, R., 2023. Genomic evaluation of carcass traits of
690 Korean beef cattle Hanwoo using a single-step marker effect model. Journal of Animal Science 101,
691 skad104. https://doi.org/10.1093/jas/skad104

692 Lê Cao, K.-A., Welham, Z.M., 2021. Multivariate Data Integration Using R: Methods and Applications
693 with the mixOmics Package, 1st ed. Chapman and Hall/CRC, Boca Raton.
694 https://doi.org/10.1201/9781003026860

695 Li, Y., Wang, D., Ping, X., Zhang, Y., Zhang, T., Wang, L., Jin, L., Zhao, W., Guo, M., Shen, F., Meng, M.,
696 Chen, X., Zheng, Y., Wang, J., Li, D., Zhang, Q., Hu, C., Xu, L., Ma, X., 2022. Local hyperthermia therapy
697 induces browning of white fat and treats obesity. Cell 185, 949-966.e19.
698 https://doi.org/10.1016/j.cell.2022.02.004

699 Lundin, V.F., Leroux, M.R., Stirling, P.C., 2010. Quality control of cytoskeletal proteins and human
700 disease. Trends in Biochemical Sciences 35, 288–297. https://doi.org/10.1016/j.tibs.2009.12.007

701 McFadden, J.W., 2020. Review: Lipid biology in the periparturient dairy cow: contemporary
702 perspectives. Animal 14, s165–s175. https://doi.org/10.1017/S1751731119003185

703 Naserkheil, M., Manzari, Z., Dang, C.G., Lee, S.S., Park, M.N., 2022. Exploring and Identifying
704 Candidate Genes and Genomic Regions Related to Economically Important Traits in Hanwoo Cattle.
705 CIMB 44, 6075–6092. https://doi.org/10.3390/cimb44120414

706 Ostinelli, G., Vijay, J., Vohl, M.-C., Grundberg, E., Tchernof, A., 2021. AKR1C2 and AKR1C3 expression
707 in adipose tissue: Association with body fat distribution and regulatory variants. Molecular and
708 Cellular Endocrinology 527, 111220. https://doi.org/10.1016/j.mce.2021.111220

709 Owens, F.N., Gill, D.R., Secrist, D.S., Coleman, S.W., 1995. Review of some aspects of growth and
710 development of feedlot cattle. Journal of Animal Science 73, 3152.
711 https://doi.org/10.2527/1995.73103152x

712 Picard, B., Cougoul, A., Couvreur, S., Bonnet, M., 2023. Relationships between the abundance of 29
713 proteins and several meat or carcass quality traits in two bovine muscles revealed by a combination
714 of univariate and multivariate analyses. Journal of Proteomics 273, 104792.
715 https://doi.org/10.1016/j.jprot.2022.104792

Possik, E., Al-Mass, A., Peyot, M.-L., Ahmad, R., Al-Mulla, F., Madiraju, S.R.M., Prentki, M., 2021. New Mammalian Glycerol-3-Phosphate Phosphatase: Role in β-Cell, Liver and Adipocyte Metabolism. Front. Endocrinol. 12, 706607. https://doi.org/10.3389/fendo.2021.706607

Quirós, P.M., Ramsay, A.J., López-Otín, C., 2013. New roles for OMA1 metalloprotease: From mitochondrial proteostasis to metabolic homeostasis. Adipocyte 2, 7–11. https://doi.org/10.4161/adip.21999

Robelin, J., 1981. Cellularity of bovine adipose tissues: developmental changes from 15 to 65 percent mature weight. Journal of Lipid Research 22, 452–457. https://doi.org/10.1016/S0022-2275(20)34959-2

Robelin, J., Geay, Y., Barbojron, C., Jailler, R., 1978. Estimation de la composition chimique du corps entier des bovins à partir du poids des dépôts adipeux totaux. Ann. Zootech. 27, 159–167. https://doi.org/10.1051/animres:19780203

Robelin, J., Geay, Y., Jailler, R., Cuylle, G., 1975. ESTIMATION DE LA COMPOSITION DES CARCASSES DE JEUNES BOVINS A PARTIR DE LA COMPOSITION D'UN MORCEAU MONOCOSTAL PRÉLEVÉ AU NIVEAU DE LA 11e CÔTE. I. – COMPOSITION ANATOMIQUE DE LA CARCASSE. Annales de zootechnie 24, 391–402.

Rohart, F., Gautier, B., Singh, A., Lê Cao, K.-A., 2017. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol 13, e1005752. https://doi.org/10.1371/journal.pcbi.1005752

Santiago, B., Baldassini, W., Neto, O.M., Chardulo, L.A., Torres, R., Pereira, G., Curi, R., Chiaratti, M.R., Padilha, P., Alessandroni, L., Gagaoua, M., 2023. Post-mortem muscle proteome of crossbred bulls and steers: Relationships with carcass and meat quality. Journal of Proteomics 278, 104871. https://doi.org/10.1016/j.jprot.2023.104871

Sauerwein, H., Häußler, S., 2016. Endogenous and exogenous factors influencing the concentrations of adiponectin in body fluids and tissues in the bovine. Domestic Animal Endocrinology 56, S33–S43. https://doi.org/10.1016/j.domaniend.2015.11.007

Sepchat, B., Ortigues Marty, I., Mialon, M.-M., Faure, P., Agabriel, J., 2013. Croissance et nature des dépôts de jeunes bovins charolais recevant en engraissement des rations à base d'enrubannage ou d'ensilage de maïs, in: 17. Rencontres Recherches Ruminants. Paris, France.

Silva-Vignato, B., Coutinho, L.L., Poleti, M.D., Cesar, A.S.M., Moncau, C.T., Regitano, L.C.A., Balieiro, J.C.C., 2019. Gene co-expression networks associated with carcass traits reveal new pathways for muscle and fat deposition in Nelore cattle. BMC Genomics 20, 32. https://doi.org/10.1186/s12864-018-5345-y

Tenenhaus, A., Tenenhaus, M., 2011. Regularized Generalized Canonical Correlation Analysis. Psychometrika 76, 257–284. https://doi.org/10.1007/s11336-011-9206-8

Volat, F.E., Pointud, J.-C., Pastel, E., Morio, B., Sion, B., Hamard, G., Guichardant, M., Colas, R., Lefrançois-Martinez, A.-M., Martinez, A., 2012. Depressed Levels of Prostaglandin F2α in Mice Lacking Akr1b7 Increase Basal Adiposity and Predispose to Diet-Induced Obesity. Diabetes 61, 2796–2806. https://doi.org/10.2337/db11-1297

Wold, H., 2006. Partial Least Squares, in: Kotz, S., Read, C.B., Balakrishnan, N., Vidakovic, B., Johnson, N.L. (Eds.), Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc., Hoboken, NJ, USA, p. ess1914.pub2. https://doi.org/10.1002/0471667196.ess1914.pub2

Zhang, J., Sheng, H., Zhang, L., Li, X., Guo, Y., Wang, Y., Guo, H., Ding, X., 2023. Bta-miR-206 and a Novel lncRNA-lncA2B1 Promote Myogenesis of Skeletal Muscle Satellite Cells via Common Binding Protein HNRNPA2B1. Cells 12, 1028. https://doi.org/10.3390/cells12071028

Fig 1

F1MR86.Q3T173 (FHL1)
G3MZ95 (FHL1)
TMM11 (TMEM11)
AL1A1.ALDH2 (ALDH1A1;ALDH2)
RS15A (RPS15A)
ISCA2 (ISCA2)
RS24 (RPS24)
A5PKC2 (SHBG)
HP252 ()
MYO1D (MYO1D)
E1BEL9 (COASY)
NDUB3 (NDUFB3)
F1MFZ4 (ADH4)
A0A3Q1M9Y4.A0A3Q1MN02.A0A3Q1NFQ2.E1BMF6.F1N4V5 (NAA15;NAA16)
FGF1 (FGF1)
A0A3Q1MAB7.A0A3Q1MHJ2.A0A3Q1MQX7.A0A3Q1MTL8 (DCUN1D1)
A0A3Q1LNW7.F1N5J8 (DECR1)
A0A3Q1MXL6.F1N1G7 (KIF5B)
RIDA (RIDA)
VPS29 (VPS29)
F6QN89 (SNTB2)
DRG1 (DRG1)
DHSO (SORD)
PGP (PGP)
F1N6I4 (LOC788425)
F1MVM4 (DAO)
ARPC2 (ARPC2)
GPX4 (GPX4)
PDLI2 (PDLIM2)
SNAG (NAPG)

HARS1 (HARS1)
LTOR5 (LAMTOR5)
FIBB (FGB)
A0A3Q1MA79.F1MC13 (LAMA5)
RADI (RDX)
F1N3N3 (LOC530929)
A0A3Q1LMT1.A0A3Q1LV88.A0A3Q1M3I4.A0A3Q1MQT0.A0A3Q1NLJ8.F1MIP7 (ABCC2)
VATA (ATP6V1A)
TCPZ (CCT6A)
E1BMX5 (NRP1)
F1N1S0 (SSBP1)
IF2A (EIF2S1)
DHYS (DHPS)
A0A3Q1LJ78.A4IFP8 (LPGAT1)
A0A3Q1M4V7.F1N3N3 (LOC521656;LOC530929)
A0A3Q1LK49.A5D7R6 (ITIH2)
ITIH1 (ITIH1)
F1N6Y1 (GANAB)
BHMT1 (BHMT)
E1BH06.F1MVK1 (C4A;LOC107131209)
SPA31.SPA32 (SERPINA3-1;SERPINA3-2)
AFAM (AFM)
FIBG (FGG)
HMGB2 (HMGB2)
DHX9 (DHX9)
A0A3Q1MG04.F1MAV0 (FGB)
SAT2 (SAT2)
APOA2 (APOA2)
A0A3Q1LMC1.A0A3Q1MDD4.A3KMY8 (GUSB)
DHR11 (DHRS11)
A0A3Q1M3A4.E1B805 (LOC528040)
LYSM ()
A0A3Q1LSP4.F1MX87 (C8A)
BCAM (BCAM)
F1MIL9 (GNE)
EIF3B (EIF3B)
RBM14 (RBM14)
HMGB1 (HMGB1)
Q2KJA9 (ESAM)
A0A3Q1MME4 (HNRNPA2B1)
TCPQ (CCT8)
RAP2C (RAP2C)
THRB (F2)
CAVN4 (CAVIN4)
PNPH (PNP)
A2AP (SERPINF2)
NDKB (NME2)
A0A3Q1MDG9 (LZIC)
3HIDH (HIBADH)
FETUA (AHSG)
F6Q4D3 (UBA7)
HIBCH (HIBCH)
IF6 (EIF6)
TCPB (CCT2)

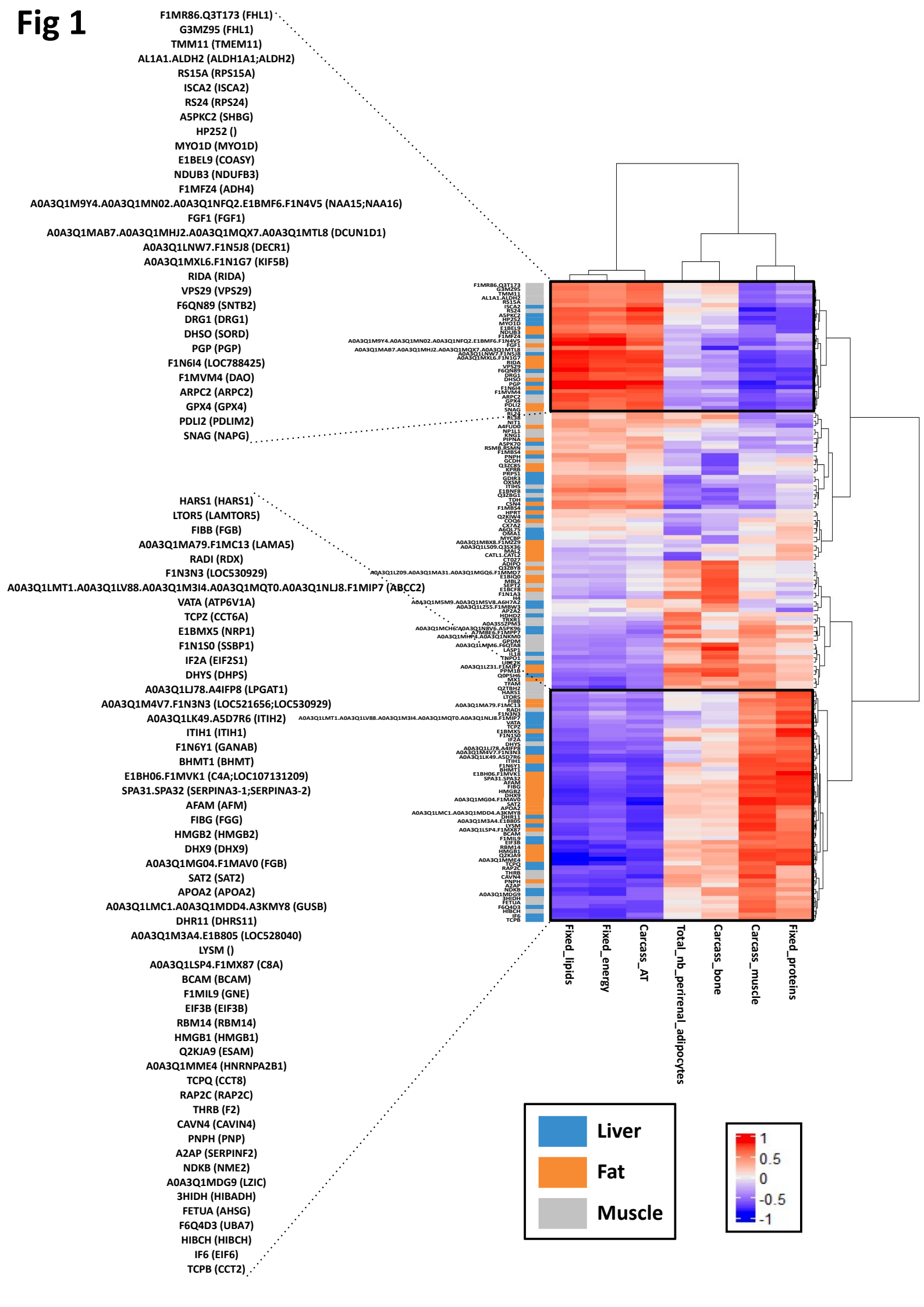
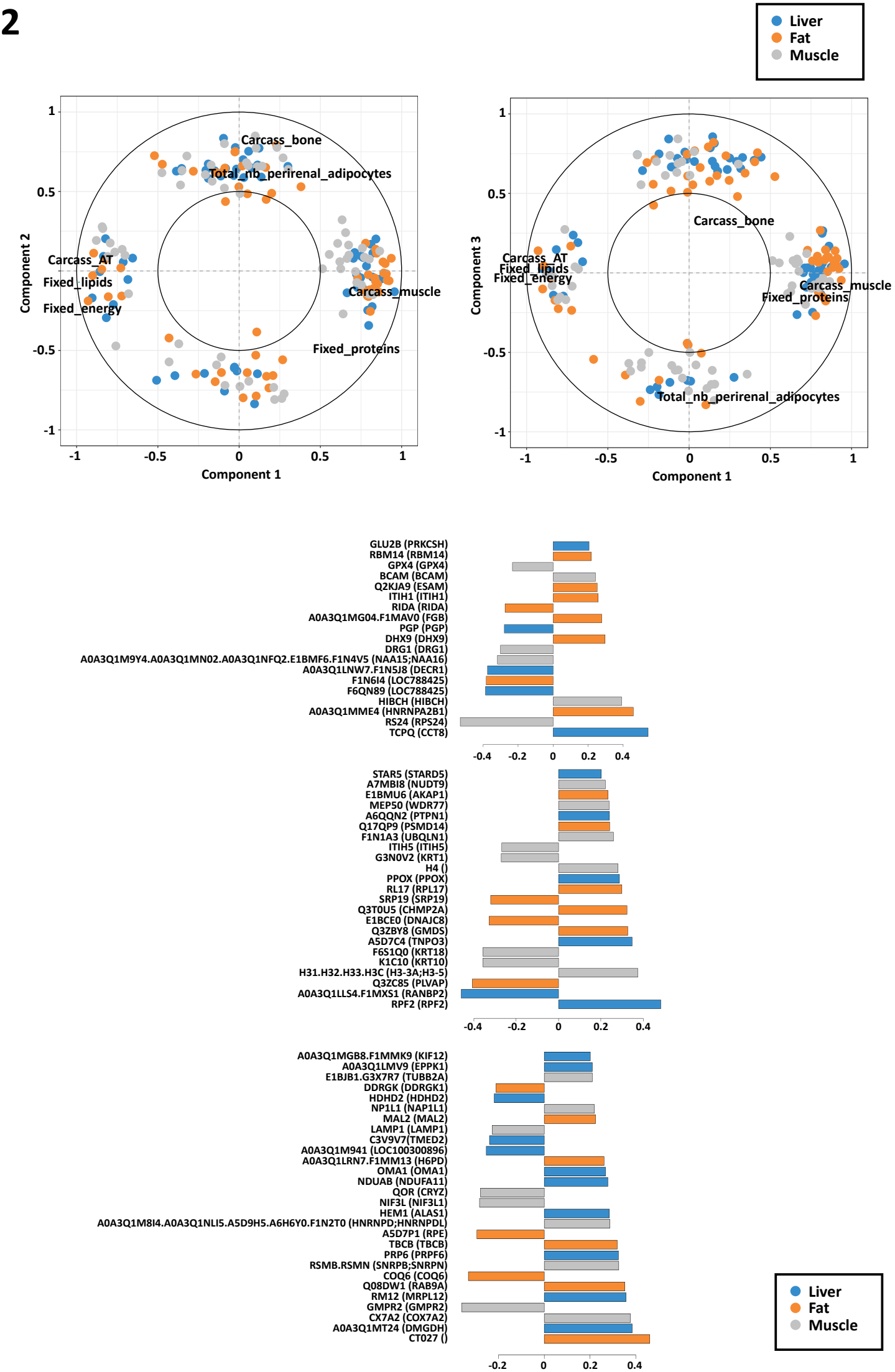


Fig 2


A)
Liver
Fat
Muscle
Carcass_bone
Total_nb_perirenal_adipocytes
Carcass_AT
Fixed_lipids
Fixed_energy
Carcass_muscle
Fixed_proteins
Component 1
Component 2
Component 3
B)
GLU2B (PRKCSH)
RBM14 (RBM14)
GPX4 (GPX4)
BCAM (BCAM)
Q2KJA9 (ESAM)
ITIH1 (ITIH1)
RIDA (RIDA)
A0A3Q1MG04.F1MAV0 (FGB)
PGP (PGP)
DHX9 (DHX9)
DRG1 (DRG1)
A0A3Q1M9Y4.A0A3Q1MN02.A0A3Q1NFQ2.E1BMF6.F1N4V5 (NAA15;NAA16)
A0A3Q1LNW7.F1N5J8 (DECR1)
F1N6I4 (LOC788425)
F6QN89 (LOC788425)
HIBCH (HIBCH)
A0A3Q1MME4 (HNRNPA2B1)
RS24 (RPS24)
TCPQ (CCT8)
STAR5 (STARD5)
A7MBI8 (NUDT9)
E1BMU6 (AKAP1)
MEP50 (WDR77)
A6QQN2 (PTPN1)
Q17QP9 (PSMD14)
F1N1A3 (UBQLN1)
ITIH5 (ITIH5)
G3N0V2 (KRT1)
H4 ()
PPOX (PPOX)
RL17 (RPL17)
SRP19 (SRP19)
Q3T0U5 (CHMP2A)
E1BCE0 (DNAJC8)
Q3ZBY8 (GMDS)
A5D7C4 (TNPO3)
F6S1Q0 (KRT18)
K1C10 (KRT10)
H31.H32.H33.H3C (H3-3A;H3-5)
Q3ZC85 (PLVAP)
A0A3Q1LLS4.F1MXS1 (RANBP2)
RPF2 (RPF2)
A0A3Q1MGB8.F1MMK9 (KIF12)
A0A3Q1LMV9 (EPPK1)
E1BJB1.G3X7R7 (TUBB2A)
DDRGK (DDRGK1)
HDHD2 (HDHD2)
NP1L1 (NAP1L1)
MAL2 (MAL2)
LAMP1 (LAMP1)
C3V9V7(TMED2)
A0A3Q1M941 (LOC100300896)
A0A3Q1LRN7.F1MM13 (H6PD)
OMA1 (OMA1)
NDUAB (NDUFA11)
QOR (CRYZ)
NIF3L (NIF3L1)
HEM1 (ALAS1)
A0A3Q1M8I4.A0A3Q1NLI5.A5D9H5.A6H6Y0.F1N2T0 (HNRNPD;HNRNPDL)
A5D7P1 (RPE)
TBCB (TBCB)
PRP6 (PRPF6)
RSMB.RSMN (SNRPB;SNRPN)
COQ6 (COQ6)
Q08DW1 (RAB9A)
RM12 (MRPL12)
GMPR2 (GMPR2)
CX7A2 (COX7A2)
A0A3Q1MT24 (DMGDH)
CT027 ()

6.3 Synthèse des principaux résultats

En terme de résultats méthodologiques, des régressions PLS entre les protéines de chaque tissu et chaque phénotype ont été menées, ainsi qu'une régression MB-PLS intégrant simultanément les trois protéomes et les sept variables phénotypiques. La mise en commun des listes de protéines des deux approches a résulté en une liste de 31 protéines quantifiées dans l'un des 3 tissus et considérées comme faisant partie des protéines les plus explicatives de la variabilité des phénotypes, à la fois au sein de leur tissu (de par la PLS), et comparativement aux autres interactions protéines-phénotypes quels que soient le tissu et le phénotype considérés (de par la MB-PLS).

En terme de résultats scientifiques, de multiples listes de protéines signant l'un des 7 phénotypes ont été produites par régressions PLS à partir de 4054 protéines du foie, 3085 protéines du tissu adipeux périrénal et de 2027 protéines musculaires. Finalement, 31 protéines des trois tissus ont été sélectionnées conjointement par PLS et MB-PLS pour expliquer la variabilité des différents phénotypes. Parmi elles, certaines sont liées à la fois à la composition tissulaire et chimique de la carcasse (SNTB2 dans le foie, HIBCH dans le muscle, HNRNPA2B1 dans le tissu adipeux), d'autres à la quantité de protéines fixées (CCT8 dans le foie, RPS24 et DRG1 dans le muscle) ou au nombre total d'adipocytes dans le tissu adipeux périrénal (OMA1 dans le foie, COX7A2 dans le muscle, CT027 dans le tissu adipeux). L'originalité de ces recherches en terme de nombre de tissus/organes et variables phénotypiques considérées, de production quantitative de données protéomiques *via* SWATH-MS et enfin d'intersection entre deux approches régressives explique que ces protéines ne soient pas ou peu identifiées actuellement chez le bovin. Cependant, les 31 protéines que nous proposons comme marqueurs potentiels de la composition chimique ou tissulaire des carcasses bovines, ont été rapportées comme liées à la composition corporelle chez l'homme ou le rongeur (HiBCH : Bjune et al. (2021) et Bishop et al. (2022), HNRNPA2B1 : Li et al. (2022), COXA7 : Carty et al. (2014), CT027 : Chen et al. (2022)) ou impliquées dans des processus biologiques contribuant à la composition corporelle, tels que la sensibilité à l'insuline (SNTB2 : Hebel et al. (2015)) ou l'utilisation tissulaire du glucose (OMA1 : Quirós et al. (2013)).

En comparant la liste des 31 protéines marqueurs potentiels des compositions de carcasse à celle des protéines dont l'abondance varie selon le groupe alimentaire (sélectionnées par la sPLSDA et/ou les analyses uni-variées), seule la protéine musculaire GPX4, sélectionnée par sPLSDA, est aussi sélectionnée par PLS1 et MB-PLS. Ainsi, 30 des 31 marqueurs potentiels sont liés aux phénotypes de composition corporelle indépendamment des régimes alimentaires étudiés. Les perspectives de ce travail sont alors 1- de valider le lien entre l'abondance de ces 30 protéines et la variabilité des phénotypes de la composition de carcasse bovine pour un nombre plus conséquent de bovins divergents par la composition corporelle, puis 2- d'étudier le pouvoir prédictif de l'abondance de ces protéines sur les phénotypes de la composition corporelle.

6.4 Discussion

Les travaux présentés dans cet article exploitent seulement une partie des analyses menées dans le cadre de cette thèse sur les données bovines, à savoir seulement l'utilisation des régressions PLS1 et MB-PLS, et pas les régressions PLS2 et cimDiablo_v2. De plus, la majeure partie du pré-traitement des données et des analyses préliminaires, deux étapes du tutoriel, n'est pas détaillée dans cet article. Ainsi, ces différents points sont présentés et discutés dans la section suivante, en complément des informations délivrées dans l'article ci-dessus.

6.4.1 Utilisation du tutoriel lors des analyses préliminaires

La Figure 6.2 résume comment le tutoriel a été utilisé sur les données bovines.

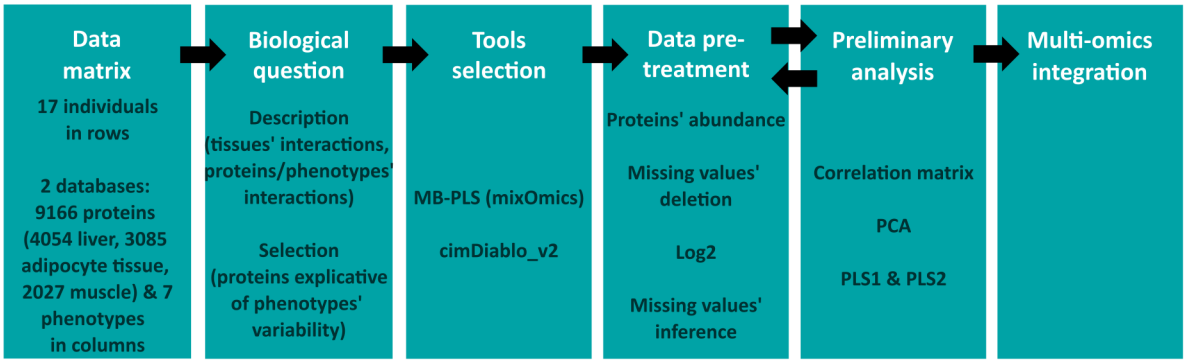


FIGURE 6.2 Application du tutoriel sur les données bovines
Les données sont composées de 17 individus en lignes pour 9166 protéines (du foie, tissu adipeux et muscle) en colonnes pour un premier jeu de données, et de 7 variables phénotypiques pour le deuxième. L'objectif étant d'identifier les signatures moléculaires de la composition corporelle, les stratégies abordées sont premièrement de décrire des interactions entre les protéines des différents tissus ainsi qu'entre les protéines et phénotypes, puis de sélectionner des protéines fortement liées aux phénotypes. Les outils intégratifs choisis sont la MB-PLS et cimDiablo_v2. Les protéines comportant plus de 20% de valeurs manquantes dans au moins l'un des groupes alimentaires sont supprimées, une log-transformation est appliquée à l'ensemble des protéines, puis les valeurs manquantes restantes sont imputées. Les principales analyses préliminaires ont été réalisées par jeu de données avec des matrices de corrélation et ACP, puis entre jeux de données avec des régressions PLS.

— **Nombre de lignes et de colonnes :** bien que le tutoriel soit initialement proposé pour des données contenant de nombreuses lignes et peu de colonnes, il a aussi été utilisé avec succès sur les données bovines structurées en peu de lignes pour de nombreuses colonnes.

— **Choix des outils d'analyse :** afin de choisir les outils d'intégrations omiques, deux critères ont été considérés. Premièrement, nous devions choisir une méthode analysant plusieurs blocs de données contenant de multiples protéines pour peu d'individus et utilisable pour sélectionner les protéines les plus explicatives de la variabilité phénotypique. Deuxièmement, nous souhaitions aussi tester cimDiablo_v2 dans ce contexte, l'outil n'ayant jamais été utilisé sur de telles données. Nous avons donc décidé d'utiliser la fonction *block.spls* de `mixOmics`, qui répond parfaitement au premier critère, mais

qui est aussi la version avec sélection de la fonction *block.pls* utilisée par cimDiablo_v2. Ainsi, une comparaison entre les deux approches était possible à l'aide de ces données. De plus, afin de compléter la comparaison de cimDiablo_v2 avec des approches similaires, nous avons aussi décidé d'utiliser lors des analyses préliminaires les régressions PLS1 et PLS2, pour sélectionner par tissu des protéines candidates et les comparer à celles de *block.spls* et cimDiablo_v2.

— **Les pré-traitements et analyses préliminaires pour caractériser la variabilité biologique induite par le régime ou l'enclos :** les 17 bovins faisaient partie d'une expérimentation incluant une quarantaine de bovins, répartis par 5 dans des enclos d'engraissement, selon leur régime alimentaire et poids vif. Les individus d'un même enclos ayant le même régime alimentaire pour des raisons pratiques (contrainte de l'éleveur fournissant un même type d'aliment pour tous les individus d'un même enclos), il n'était pas possible de corriger l'effet éventuel de l'enclos sans corriger l'effet souhaité du régime alimentaire. Afin d'identifier l'impact de ces conditions expérimentales sur les données produites, les individus ont été classifiés selon le groupe alimentaire et le numéro de l'enclos, et des analyses spécifiques ont été menées. Pour le protéome du foie exclusivement, la projection des individus sur la première composante de l'ACP a montré, en plus d'un fort effet du régime (en accord avec le grand nombre de protéines différentiellement abondantes selon le régime) (Figure 6.3.A), un effet mineur du numéro de l'enclos au sein du groupe alimentaire "*diet 1*" (Figure 6.3.B). L'effet "enclos" impactant finalement peu les données, il n'a pas été corrigé, conservant ainsi la variabilité induite par le régime alimentaire.

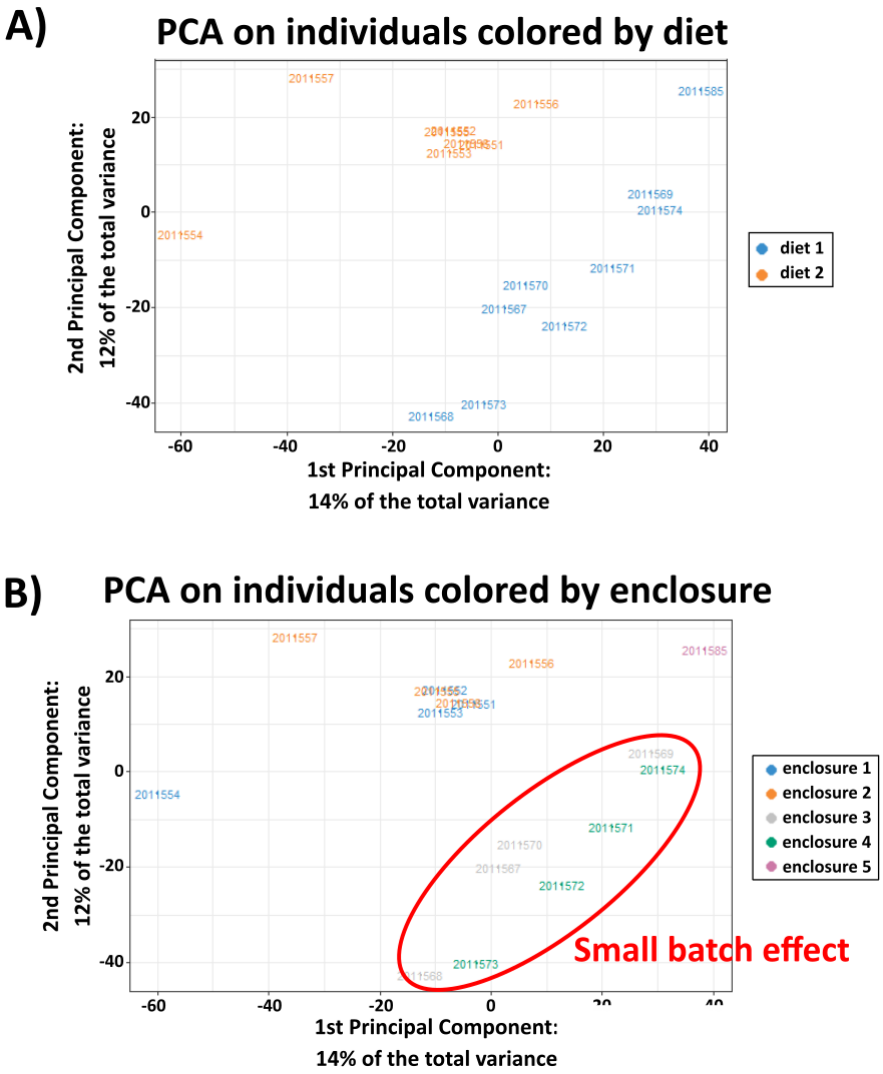


FIGURE 6.3 Individus coloriés par groupe sur l'ACP des données du foie
Représentation des 17 bovins sur les deux premières composantes de l'ACP, coloriés **A-** selon l'enclos et **B-** selon le régime alimentaire (légendes à droite des sous-figures). Un faible effet enclos est visible entre les individus des enclos 3 et 4, tous du premier régime alimentaire.

6.4.2 Approches sélectives : régressions PLS et cimDiablo_v2

Dans cet article, l'objectif étant de sélectionner des groupes de protéines fortement liées à différents phénotypes d'intérêt pour les éleveurs, plusieurs méthodes ont été testées dans le but de mener cette sélection de protéines. Les méthodes finalement conservées dans l'article sont les régressions PLS1 et MB-PLS, face à la PLS2 et cimDiablo_v2 avant et après "débruitage". Ces différentes méthodes sont discutées ici.

6.4.2.1 Sélection de protéines complémentaires avec les régressions PLS

Choix du type de régression PLS :

Les régressions PLS, que ce soit les PLS1, PLS2 ou MB-PLS, sont basées sur un concept

similaire qui est celui de calculer la combinaison linéaire des variables explicatives X maximisant une métrique telle que la covariance ou la corrélation avec la ou les variables expliquées Y. Ces régressions ont donc pour fonctionnement de sélectionner des protéines complémentaires, de sorte à expliquer au mieux Y avec le minimum de variables de X. Par construction, pour deux variables de X similaires et explicatives de la/des variables Y, seule une des deux sera sélectionnée, la deuxième n'apportant pas d'information majeure par rapport à la première pour expliquer Y.

Les différences majeures des différentes régressions PLS sont le nombre de variables et blocs de variables utilisés, ainsi que la question scientifique à laquelle chaque régression apporte une réponse (Table 6.1).

Régression PLS	**Variables explicatives X**	**Variables expliquées Y**	**Question scientifique**
PLS1	un seul bloc	une seule variable	Quels groupes de variables X expliquent au mieux la variable Y ?
PLS2	un seul bloc	plusieurs variables	Quels groupes de variables X expliquent au mieux l'ensemble des variables Y ?
MB-PLS	plusieurs blocs	une ou plusieurs variables	Quels groupes de variables X expliquent au mieux la ou l'ensemble des variables Y, avec la possibilité de privilégier ou non les variables X interagissant entre blocs ?

TABLE 6.1 Différences entre les régressions PLS selon le type de variables considéré et la question posée
Caractéristiques des régressions PLS1, PLS2 et MB-PLS en terme de variables considérées et de questions auxquelles ces régressions répondent.

Ainsi, comme présenté dans l'article, ces approches bien que très ressemblantes induisent des résultats complémentaires : alors que les PLS1 ont identifié des groupes de protéines par tissu associés à chaque variable phénotypique, la MB-PLS a identifié un ensemble de protéines des 3 tissus et n'expliquant pas un phénotype unique mais l'ensemble des phénotypes. En complément de l'article, les résultats des PLS2 listent des protéines d'un même tissu expliquant l'ensemble des phénotypes.

Nombre de composantes et de protéines sélectionnées par composante :

Le choix du nombre de composantes puis du nombre de protéines à sélectionner par composante à partir des régressions PLS est souvent difficile. Deux stratégies sont généralement utilisées : choisir les paramètres selon des arguments statistiques et algorithmiques, par exemple

en utilisant des méthodes de validation croisée et des métriques pour déterminer les paramètres optimisant les corrélations, la valeur prédictive des résultats, *etc.*, ou choisir les paramètres selon des arguments biologiques, par exemple en confrontant les résultats obtenus à ceux de la littérature. Ici, les fonctions d'optimisation des paramètres de `mixOmics` ont été testées lorsque disponibles, mais les paramètres ont principalement été choisis pour répondre à la stratégie suivante : sélectionner par différentes approches des listes de protéines candidates les plus explicatives de la variabilité des phénotypes, puis conserver uniquement les candidats les plus robustes. Dans l'article Mardoc et al. (en préparation), les régressions PLS1 et MB-PLS ont été utilisées pour déterminer les listes de protéines candidates, et seules les protéines sélectionnées par les deux approches ont été conservées. Une étude bibliographique des fonctions associées à ces protéines a ensuite été menée pour vérifier si la liste obtenue était cohérente avec les résultats de la littérature. Enfin, des travaux supplémentaires sont en préparation au sein de l'équipe pour conserver uniquement, à partir de cette liste, les protéines les plus explicatives de la variabilité des phénotypes pour un nouveau jeu de données, contenant un plus grand nombre d'individus et avec des données protéomiques produites par méthode immunologique. Les paragraphes suivants détaillent les choix menés sur le nombre de composantes et protéines par composante pour les PLS1 et MB-PLS en suivant cette stratégie.

Pour la PLS1, le nombre de composantes a été choisi en utilisant la fonction de `mixOmics` dédiée, nommée *perf*, réalisant par validation croisée plusieurs régressions pour différents nombres de composantes. La métrique $Q2$ qui en résulte (Tenenhaus, 1998) quantifie l'intérêt d'ajouter ou non une composante pour améliorer la prédiction de la variable Y. Finalement, ces résultats ont montré qu'il n'était jamais nécessaire d'ajouter de deuxième composante pour ces PLS1.

Plusieurs approches ont ensuite été testées pour sélectionner un certain nombre de protéines pour les différentes régressions PLS1. Pour commencer, tout comme précédemment, `mixOmics` propose une fonction nommée *tune* basée sur la validation croisée testant différentes valeurs et identifiant celle maximisant la prédiction. Nous avons utilisé cette fonction afin d'obtenir ce nombre optimal, qui varie de quelques protéines pour certaines régressions à plusieurs centaines lorsque le protéome est peu lié à la variable phénotypique. Cependant, nous avons finalement sélectionné le top10 des protéines explicatives dans la composante. Ce choix a été fait pour deux raisons. Premièrement, avec la fonction *tune*, nous devons assumer de considérer de la même manière quelques protéines très explicatives pour une régression et de nombreuses protéines moins explicatives pour une autre régression. En sélectionnant le top10, nous comparons toujours des protéines dont certaines sont potentiellement plus explicatives que d'autres d'un phénotype, mais en éliminant le biais sur le nombre de protéines sélectionnées. Deuxièmement, en faisant le choix de ne garder que les 10 meilleures protéines par régression, nous obtenons un faible nombre de protéines sur lequel il est possible de mener des recherches plus approfondies des fonctions biologiques associées, et de confirmer sur

une plus grande population les relations entre l'abondance des protéines et le phénotype, ce qui permet une fois de plus, à terme, d'éliminer les protéines insuffisamment explicatives de la variabilité des phénotypes.

Concernant la MB-PLS, il n'est pas encore possible d'utiliser les fonctions *perf* et *tune* de `mixOmics`. Une possibilité aurait pu être de développer une fonction locale de validation croisée pour tester les différents paramètres, mais nous avons finalement opté pour d'autres approches basées sur les fonctions graphiques de `mixOmics` disponibles pour les MB-PLS. Pour déterminer le nombre de composantes, nous avons effectué la régression avec différents nombres de composantes puis observé la position des 7 variables phénotypiques sur celles-ci. Comme montré en Figure 6.4, avec trois composantes, la première sépare et oppose la proportion de tissu adipeux, le poids de lipides et les mégajoules d'énergie à la proportion de muscles et le poids de protéines, quand les deuxième et troisième composantes séparent la proportion d'os et le nombre total d'adipocytes, tout d'abord corrélées positivement (composante 2) puis négativement (composante 3). Ainsi, les trois composantes semblent informatives et ont été conservées. Nous avons ensuite sélectionné les protéines les plus explicatives en prenant cette fois le top30 des protéines par composante et tissu auxquelles nous avons appliqué un seuil de -0.2 à 0.2 pour ne garder que les protéines les plus explicatives Figure 6.5.

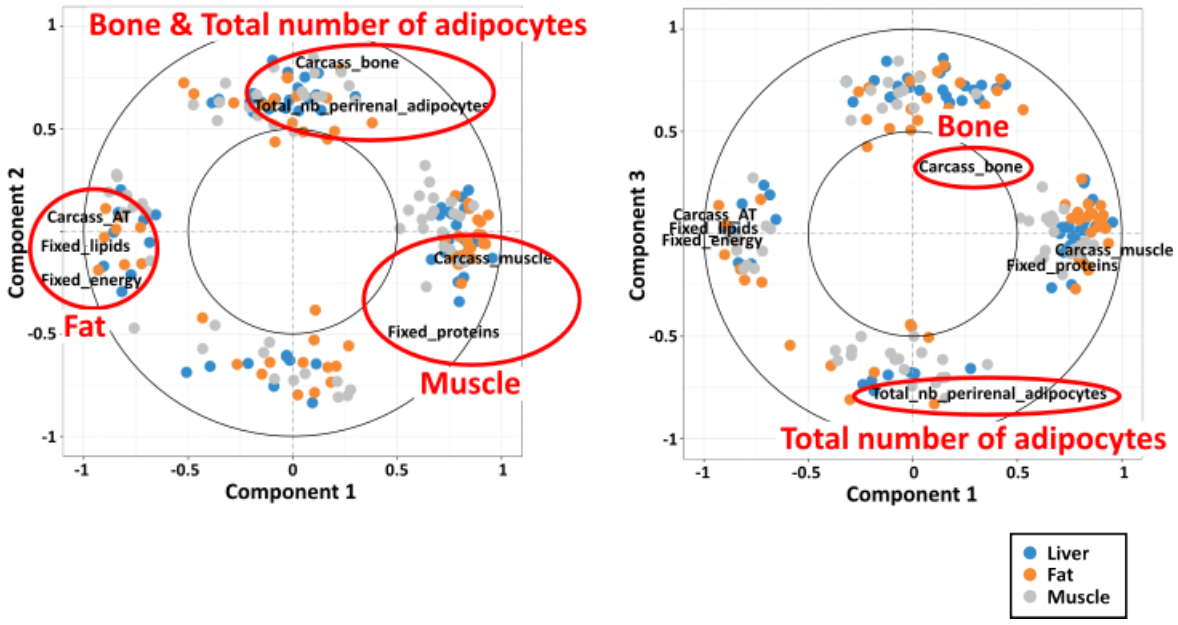


FIGURE 6.4 Variables sur les premières composantes de la MB-PLS

Représentation des protéines sélectionnées et des phénotypes sur les composantes 1 et 2 (à gauche) et 1 et 3 (à droite). Les protéines sont représentées par des points de couleur selon leur provenance (légende en bas à droite), et les phénotypes directement par leur nom. Les groupes de phénotypes au profil semblable sur les composantes sont encerclés en rouge.

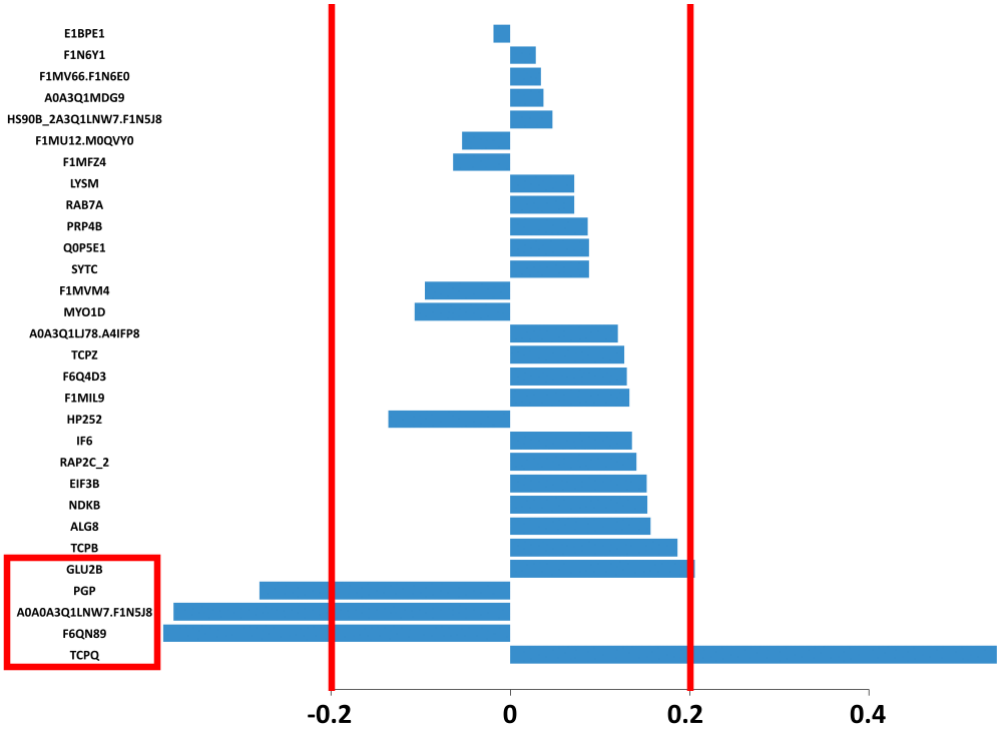


FIGURE 6.5 *Loadings* de la MB-PLS du top30 des protéines du foie sur la 1ère composante, avec sélection des protéines au seuil $\pm$ 0.2

L'axe des abscisses correspond aux *loadings* calculés par la MB-PLS. Plus cette valeur est importante en valeur absolue et plus la protéine est fortement représentée dans la composante. L'axe des ordonnées correspond au top30 des protéines, à savoir celles avec les plus fortes valeurs absolues de *loadings* sur la composante. Avec des seuils à -0.2 et 0.2 sur les *loadings*, seules les 5 protéines encadrées sont sélectionnées.

6.4.2.2 Sélection de protéines similaires avec le clustering hiérarchique de cimDiablo_v2

Contrairement aux méthodes régressives sélectionnant des protéines au profil complémentaire, la fonction cimDiablo_v2 peut être théoriquement utilisée pour sélectionner des protéines au profil similaire *via* le dendrogramme issu du clustering hiérarchique des variables omiques.

En effet, lors de l'utilisation de cimDiablo_v2 sur les données de l'article, l'intégration s'est effectuée sur les trois blocs protéomiques correspondant aux trois tissus ainsi que le bloc des données phénotypiques. Les variables de ces blocs sont ensuite affichées en colonnes du graphique et ordonnées par le clustering hiérarchique de `mixOmics`, qui associe itérativement les variables et groupes de variables les plus similaires. Il suffit donc théoriquement, dans le contexte de la recherche de protéines fortement associées aux phénotypes, de chercher dans le dendrogramme les protéines les plus proches des différents phénotypes. Plusieurs limitations sont toutefois à prendre en considération :

— **Métrique de distance ou similarité :** afin de regrouper les variables et groupes de variables les plus semblables, il est nécessaire de choisir une métrique explicitant ce degré de ressemblance ou de différence. Plusieurs options sont proposées dans `mixOmics`, mais la seule métrique de corrélation actuellement disponible est la corrélation de Pearson, ce qui peut être gênant pour des données non gaussiennes. Pour nos données pro-

téomiques, les distributions ne sont pas gaussiennes car asymétriques même après log-transformation (Figure 6.6), il se peut donc qu'un léger biais soit introduit. Les métriques de distances comme la distance euclidienne ont quant à elles la particularité de séparer les variables négativement corrélées, ce qui n'est pas souhaité dans ce contexte biologique précis où nous souhaitons identifier les protéines fortement corrélées positivement ou négativement aux phénotypes.

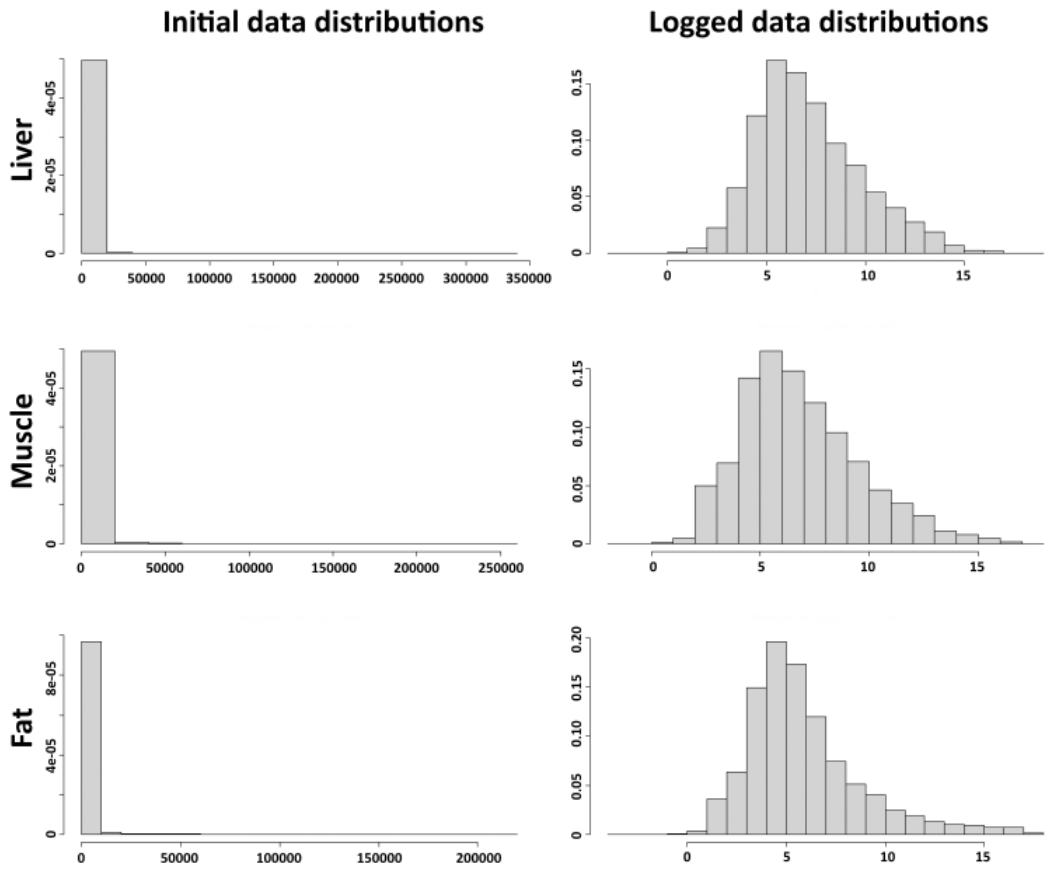


FIGURE 6.6 Distributions des données protéomiques avant et après log-transformation
Histogrammes des abondances de protéines du foie, du muscle et du tissu adipeux avant et après log-transformation. L'axe des abscisses représente les abondances ou log-abondances, l'axe des ordonnées les densités.

— **Méthode de clustering :** de la même manière, la métrique de distance ou similarité choisie ne s'appliquant que pour comparer deux variables, il est nécessaire de choisir la méthode pour comparer des groupes de variables. Plusieurs options sont proposées dans `mixOmics`, notamment les méthodes *complete linkage* et *Ward*. Cependant, seule la première implémentation de *Ward* est actuellement disponible et a été utilisée sur les données animales, quand la deuxième implémentation *Ward.D2* est généralement à privilégier.

— **Association directe ou par un intermédiaire :** toujours dans le but d'associer ensemble les variables et groupes de variables similaires, il faut aussi prendre en compte le fait que le clustering hiérarchique n'identifie pas les variables fortement liées à une

ou plusieurs variables d'intérêt, mais regroupe itérativement les variables et groupes de variables se ressemblant le plus. Avec cette méthode, si un groupe de variables est lié à deux autres groupes, ces derniers vont être considérés comme directement associés même s'ils ne le sont que par l'intermédiaire du premier groupe. Ainsi, il est possible d'identifier des protéines proches des variables phénotypiques sans qu'elles soient pourtant directement liées.

— **Nombre de nœuds :** le dendrogramme étant un arbre mathématique composé de feuilles, branches et nœuds, l'objectif est de démarrer de la feuille correspondant à la variable phénotypique d'intérêt, et de monter de nœuds en nœuds pour sélectionner les branches les plus proches composées de protéines. Il reste toutefois la question du nombre de branches et nœuds à sélectionner. En effet, plus le nombre de nœuds augmente, plus les protéines sélectionnées sont éloignées de la variable phénotypique. Bien qu'il existe de nombreuses méthodes pour déterminer le nombre optimal de nœuds à sélectionner dans un dendrogramme, elles renvoient généralement des résultats assez différents les uns des autres, c'est pourquoi il est couramment décidé de chercher la valeur consensus entre les différentes approches, ou bien de déterminer visuellement le nombre optimal de nœuds. Cependant, ces méthodes coupent généralement le dendrogramme en quelques grands clusters, ce qui n'est pas l'objectif dans ce contexte pour lequel on souhaite au contraire obtenir de très petits clusters contenant les protéines les plus proches du phénotype. Finalement, la méthode employée durant les analyses sur les données de l'article a été de fixer à 4 le nombre de nœuds considérés, et si besoin de mettre des poids de moins en moins importants en avançant dans la structure du dendrogramme afin de conserver l'information du degré de ressemblance des protéines avec la variable phénotypique.

Finalement, cimDiablo_v2 sans "débruitage" sur les données animales, avec la corrélation de Pearson et méthode *Ward*, résulte en la Figure 6.7. Les lignes correspondent aux 17 individus bovins représentés par leur identifiant. On peut alors remarquer que le régime alimentaire joue un rôle majeur dans la discrimination des groupes, avec les deux branches principales du dendrogramme des lignes (les deux groupes de composition corporelle) correspondant presque aux deux régimes, même si les hauteurs du dendrogramme montrent qu'il y a plus de différences entre chacun des individus qu'entre les deux branches principales. Les colonnes représentent les variables issues des 4 blocs, c'est-à-dire principalement les 9166 protéines des trois tissus, mais aussi les 7 variables phénotypiques. Le dendrogramme des colonnes sépare assez distinctement les protéines en 2 ou 3 grands groupes. Ici, afin de conserver la variabilité des valeurs extrêmes, nous n'avons pas utilisé le *cutoff* de cimDiablo_v2.

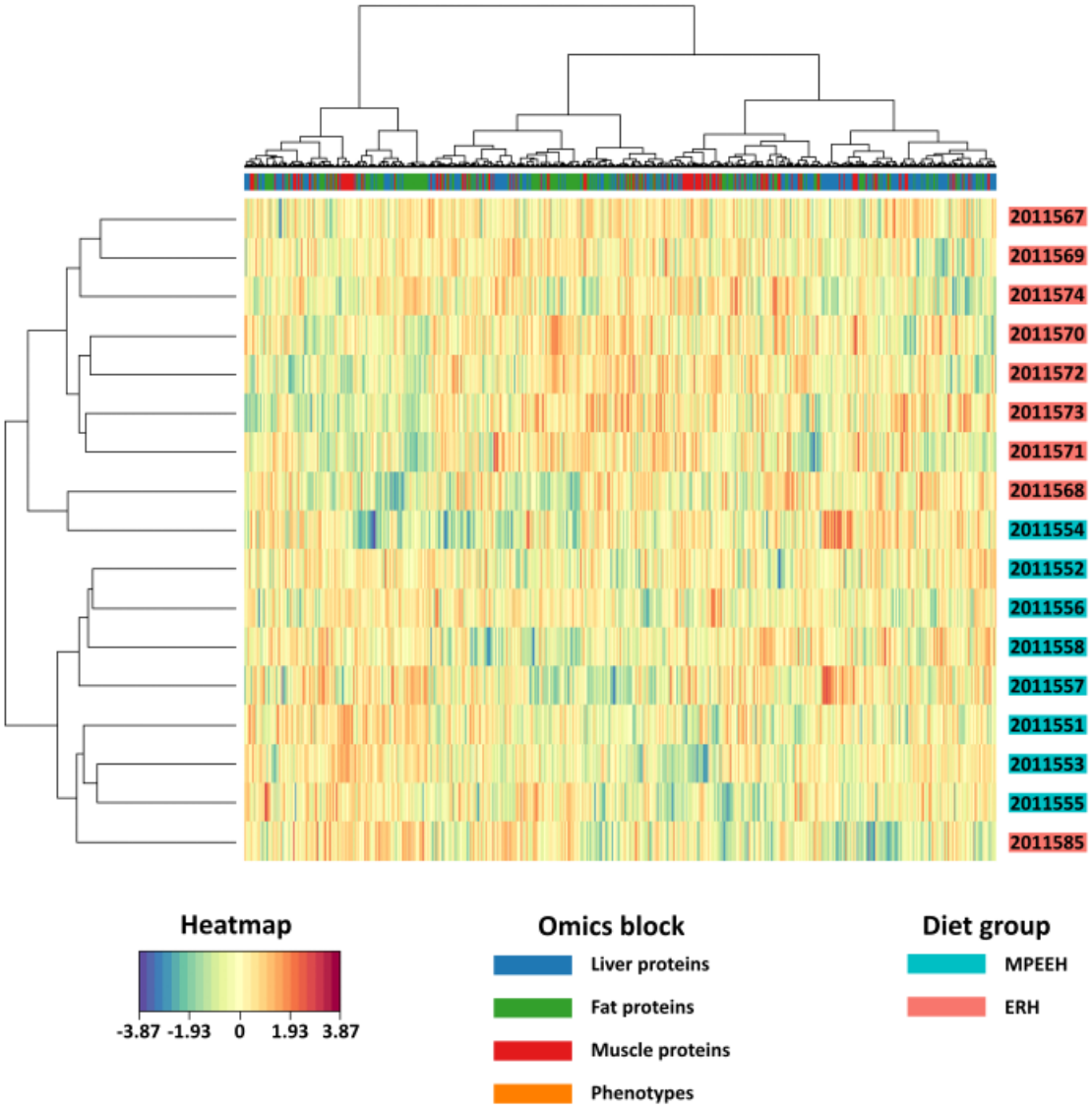


FIGURE 6.7 cimDiablo_v2 sur données animales, non "débruitées" et sans *cutoff*
Chaque ligne correspond à un individu bovin, chaque colonne à une protéine ou à un phénotype. Les lignes et colonnes sont regroupées selon les dendrogrammes issus du clustering hiérarchique avec la méthode Ward et la distance euclidienne. Les valeurs représentées selon le code couleur de la *heatmap* sont celles des données initiales centrées et réduites. Le code couleur des blocs de données ainsi que celui du régime alimentaire sont indiqués en bas de la figure : ERH pour le régime riche en herbe, MPEEH pour le régime riche en céréales.

Pour chacune des variables phénotypiques, toutes les protéines se trouvant à 4 nœuds ou moins dans le dendrogramme ont été sélectionnées. Les sous-dendrogrammes correspondant sont présentés en Figure 6.8. Les listes ainsi obtenues ont ensuite été comparées aux top10 des PLS1, mais aucune des protéines ne s'est révélée être conjointement sélectionnées dans les deux approches. Étant donné le faible effectif en protéines dans ces différentes listes, les listes de cimDiablo_v2 sans "débruitage" ont aussi été comparées à celles des PLS1 pour lesquelles la fonction *tune* a été utilisée avec la sélection d'au moins 100 protéines par régression puis conservation des protéines stables à au moins 20%. Le résultat de ces comparaisons est représenté en Figure 6.9.

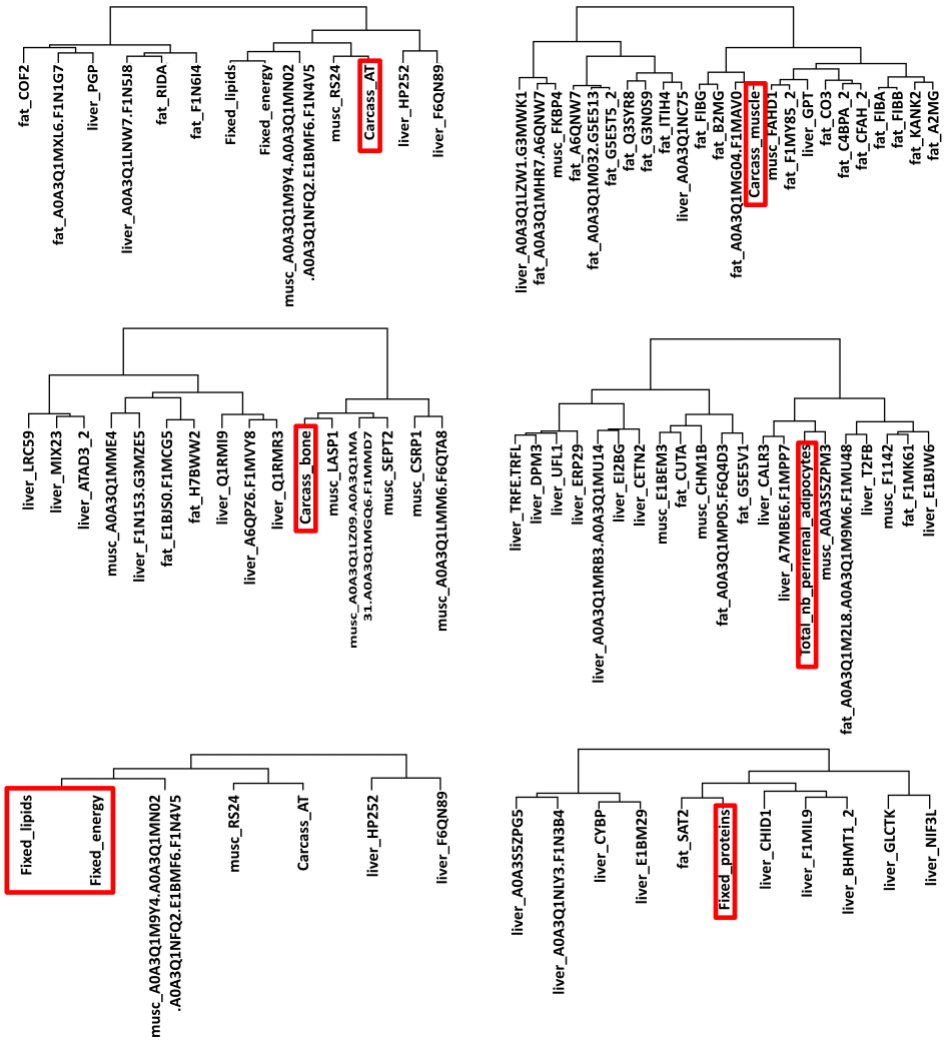


FIGURE 6.8 Zoom sur les parties du dendrogramme de cimDiablo_v2 contenant les variables phénotypiques

À partir du dendrogramme des colonnes de cimDiablo_v2, un zoom est effectué sur les branches comprenant les variables phénotypiques (à partir de chaque phénotype, le dendrogramme est remonté de 4 nœuds/intersections). Les phénotypes sont encadrés en rouge, les autres variables sont les protéines sélectionnées pour leur profil similaire au(x) phénotype(s) du même sous-dendrogramme.

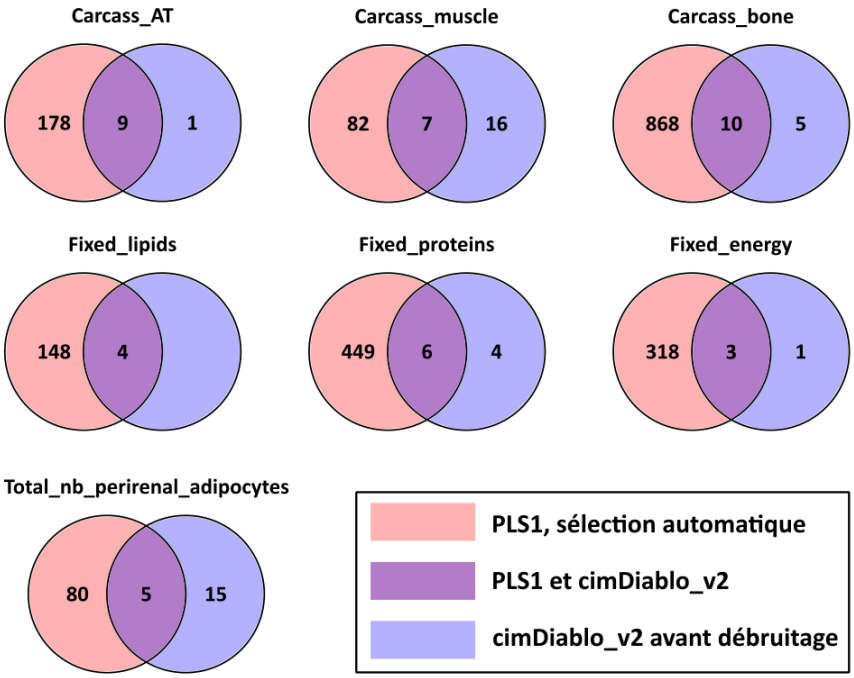


FIGURE 6.9 Diagramme de Venn des listes de protéines sélectionnées par PLS1 et par cimDiablo_v2 avant "débruitage"

Pour chacune des 7 variables phénotypiques, le diagramme de Venn indique le nombre de protéines sélectionnées par les deux approches et celles sélectionnées par seulement une des deux approches (*cf.* code couleur dans la légende à bas à droite).

Comme nous pouvons le voir, même avec plus de protéines sélectionnées par PLS, il reste de nombreuses protéines exclusivement sélectionnées par cimDiablo_v2.

Ceci peut s'expliquer par la différence d'objectifs entre les deux approches. Les régressions PLS cherchent le minimum de protéines dont la mise en commun est la plus explicative de la variable phénotypique. Les protéines ainsi sélectionnées ont des profils complémentaires. Au contraire, le clustering hiérarchique utilisé par cimDiablo_v2 regroupe les variables (protéines et phénotypes) au profil similaire selon une métrique choisie. Ainsi, les protéines exclusivement sélectionnées par PLS apportent des informations complémentaires à celles communes entre les deux approches, quand les protéines exclusives à cimDiablo_v2 ont un profil similaire à au moins une des protéines communes aux deux approches.

Dans l'article, nous avons fait le choix de ne présenter que les résultats des régressions PLS et pas ceux de cimDiablo_v2 sans "débruitage" pour deux raisons. Premièrement, comme indiqué, le paramétrage de cimDiablo_v2 n'est pas encore abouti pour ce type d'analyses, notamment sur le choix du nombre de nœuds et la sélection de variables directement ou indirectement corrélées. Deuxièmement, le but étant d'identifier un groupe de protéines le plus explicatif de la variabilité des phénotypes, il nous paraît plus pertinent de nous concentrer sur des protéines au profil complémentaire (régression PLS) que des protéines au profil similaire (cimDiablo_v2). Néanmoins, un des objectifs de la thèse étant de tester le "débruitage" de cimDiablo_v2 dans différents contextes, nous l'avons donc testé sur ces données animales (en complément des

données plantes du chapitre précédent) et avons comparé les listes de protéines sélectionnées par cimDiablo_v2 sans et avec "débruitage", comme discuté dans la section suivante.

6.4.2.3 Sélection de protéines avec le "débruitage" de cimDiablo_v2

Le "débruitage" de cimDiablo_v2 conserve théoriquement les variations majeures du jeu de données, ce qui peut mener à l'identification de protéines liées aux variables phénotypiques sans être biaisé par une éventuelle valeur extrême pour un individu spécifique. Cependant, nous montrons ici que le "débruitage" de cimDiablo_v2 n'est finalement pas applicable pour sélectionner les protéines explicatives des phénotypes sur ces données, et plus généralement pour sélectionner les colonnes d'un jeu de données.

Choix des paramètres :

Tout d'abord, afin de "débruiter" des données, il est nécessaire de choisir les paramètres de la régression MB-PLS, principalement la matrice de design et le nombre de composantes.

La matrice de design représente les poids des interactions entre les différents blocs de données. Ici, nous avons choisi de privilégier les interactions entre les blocs protéomiques et le bloc phénotypique (poids = 1), de prendre en compte plus faiblement les interactions dans les différents blocs protéomiques (poids = 0.1) et de ne pas prendre en compte les interactions d'un bloc avec lui-même (poids = 0).

Le choix du nombre de composantes est plus difficile car il n'existe pas dans `mixOmics` de fonction pour choisir ce nombre optimal pour la régression MB-PLS utilisée par cimDiablo_v2 (*i.e. block.pls*), et nous manquons encore de suffisamment de recul sur le "débruitage" de cimDiablo_v2 pour proposer une méthode efficace pour choisir ce nombre. De manière générale, puisque chaque composante contient de l'information non redondante avec celle des composantes préalablement calculées, il faut trouver un compromis pour ne pas choisir trop ou trop peu de composantes. En effet, en sélectionnant trop peu de composantes, une partie de la variabilité biologique ne sera pas contenue dans les données "débruitées". Au contraire, avec trop de composantes, une partie du "bruit" risque d'être conservée dans les dernières composantes et apparaître dans les données "débruitées". La Figure 6.10 illustre l'impact du nombre de composantes (avec 5 et 2 composantes) sur 1- le clustering des colonnes, 2- le clustering des lignes et 3- la variabilité de la *heatmap*. Pour la suite de cette analyse, nous avons arbitrairement choisi d'effectuer le "débruitage" avec 5 composantes, l'intérêt portant moins sur la méthode pour choisir le nombre de composantes que sur son impact concret sur le "débruitage" et les listes de protéines finalement sélectionnées.

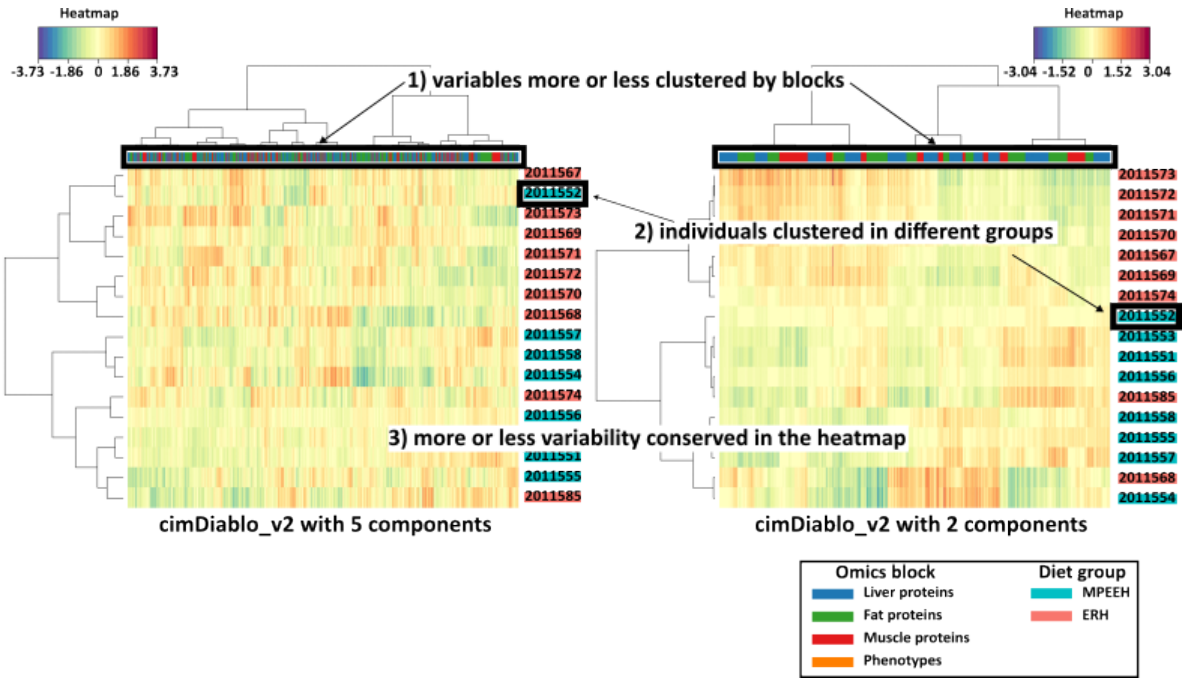


FIGURE 6.10 cimDiablo_v2 sur les données animales avec 5 puis 2 composantes
Chaque ligne correspond à un individu bovin, chaque colonne à une protéine ou à un phénotype. Les lignes et colonnes sont regroupées selon les dendrogrammes issus du clustering hiérarchique avec la méthode Ward et la distance euclidienne. Les valeurs représentées selon le code couleur de la *heatmap* sont celles des données initiales centrées et réduites, "débruitées" en utilisant 5 composantes (à gauche) ou 2 composantes (à droite), et de nouveau centrées et réduites. Le code couleur des blocs de données ainsi que celui du régime alimentaire (ERH pour le régime riche en herbe, MPEEH pour le régime riche en céréales) sont indiqués en bas de la figure. Les différences majeures entre les deux résultats de cimDiablo_v2 sont indiquées sur la figure : 1- la différence de clustering des colonnes, 2- la différence de clustering des lignes, 3- la différence de variabilité conservée et représentée dans la *heatmap*.

Sélection de protéines et comparaison avec les données non "débruitées" :

Afin de sélectionner des protéines liées aux phénotypes, la méthode est la même pour cimDiablo_v2 avec "débruitage" que sans, à savoir remonter de 4 nœuds dans le dendrogramme à partir des variables phénotypiques pour identifier les protéines qui leurs sont les plus proches. De nouvelles listes ont donc été extraites, puis comparées aux listes issues de cimDiablo_v2 avant "débruitage". Les diagrammes de Venn correspondants sont présentés en Figure 6.11.

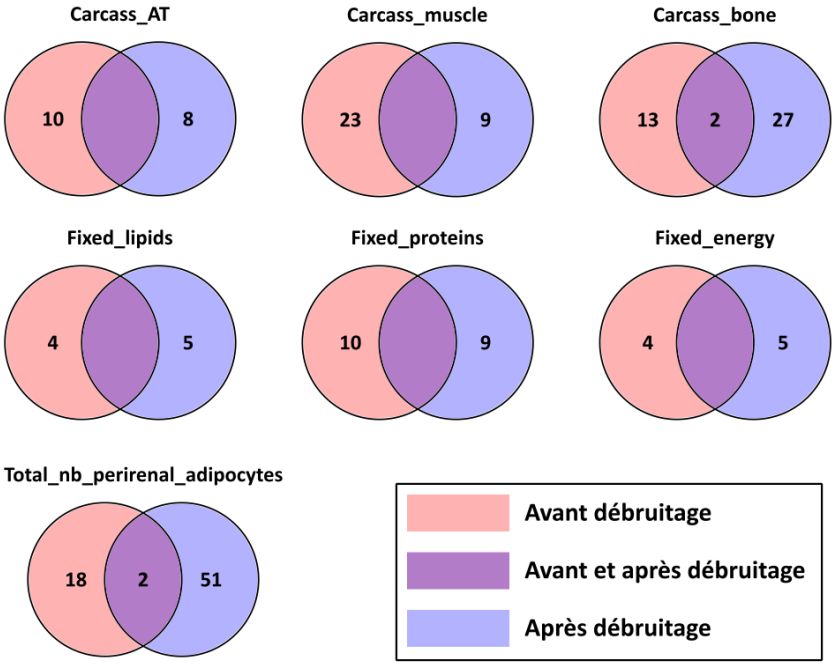


FIGURE 6.11 Diagramme de Venn des listes de protéines sélectionnées par cimDiablo_v2 avant et après "débruitage"

Pour chacune des 7 variables phénotypiques, le diagramme de Venn indique le nombre de protéines sélectionnées par les deux approches et celles sélectionnées par seulement une des deux approches (*cf.* code couleur dans la légende à bas à droite).

Les résultats indiquent que seules 2 protéines liées à la proportion d'os et 2 au nombre total d'adipocytes périrénaux sont communes, et aucune pour les autres phénotypes. Pour tenter de comprendre les différences entre ces listes de protéines, nous avons observé les corrélations entre les protéines sélectionnées avant et après "débruitage" et leurs phénotypes associés. La Figure 6.12 illustre ces corrélations (en valeurs absolues). Avant "débruitage", les corrélations sont globalement fortes, signifiant que les protéines sélectionnées sont bien celles au profil similaire avec le phénotype. Les quelques protéines faiblement corrélées au phénotype sont alors des protéines fortement corrélées à d'autres protéines du groupe, elles-mêmes fortement corrélées au phénotype. Après "débruitage", les corrélations sont globalement plus faibles et avec une seule corrélation supérieure à 0.7. Les tests de Student et de Wilcoxon entre ces deux distributions indiquent, avec des p-valeurs respectives de 1.234e-08 et 1.375e-08, une différence significative des corrélations avant contre après "débruitage". Il semble donc qu'après "débruitage", cimDiablo_v2 n'identifie pas ou peu de protéines corrélées aux phénotypes, c'est-à-dire que ce "débruitage" n'est pas du tout adapté pour sélectionner les protéines associées aux phénotypes dans ce contexte.

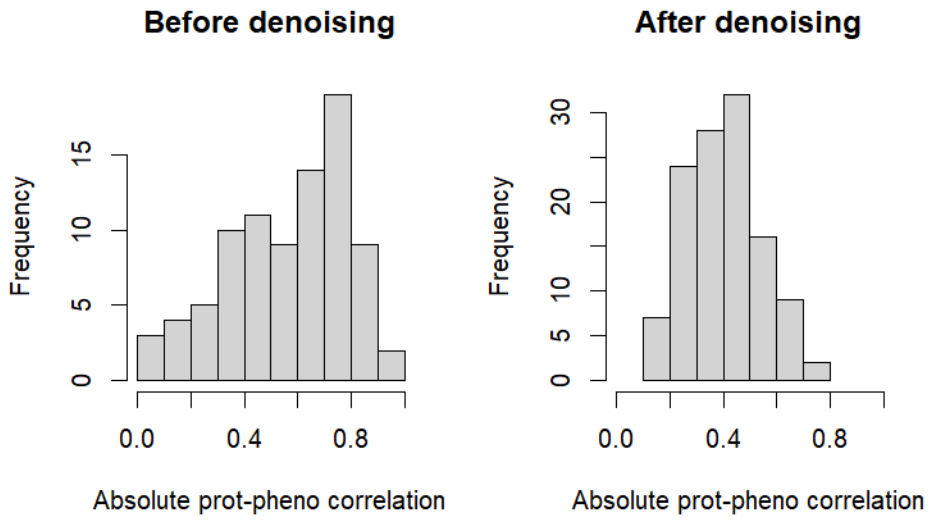


FIGURE 6.12 Distributions des corrélations entre les phénotypes et les protéines sélectionnées avant et après "débruitage"

Pour chaque protéine sélectionnée *via* le dendrogramme de cimDiablo_v2 avant et après "débruitage", sa corrélation de Pearson (en valeur absolue) avec le phénotype correspondant est calculée. Les histogrammes représentent la distribution de ses valeurs absolues de corrélation avant (à gauche) et après (à droite) "débruitage".

L'hypothèse principale expliquant ceci est la manière dont cimDiablo_v2 "débruite" les données, c'est-à-dire le "lissage" des données selon les lignes. Ainsi, le "débruitage" modifie les données de sorte à accentuer les grands profils de variabilité par lignes (ici des individus bovins) au détriment de la variabilité des colonnes (protéines et phénotypes). Le "lissage" des lignes a donc sur les données animales l'effet inverse de l'effet souhaité, puisque nous souhaitons au contraire accentuer les profils de variabilité des colonnes pour identifier les protéines aux profils similaires avec les phénotypes, comme illustré sur la Figure 6.13. Une idée primaire serait alors de transposer les données afin que les individus correspondent aux colonnes et les protéines et phénotypes aux lignes, puis appliquer le même procédé que sur les données peuplier. Cependant, cela changerait toute l'analyse, avec en premier lieu l'impossibilité d'utiliser la fonction *block.pls* pour intégrer les données, cet algorithme étant implémenté pour considérer les différents types de données omiques/phénotypiques en différents blocs en colonnes et non en lignes. Il faudrait alors étudier la possibilité d'adapter la fonction cimDiablo_v2 à la fonction intégrative *mint.pls*, ou étudier plus en détail la question d'un "débruitage" par colonnes à partir de cimDiablo_v2 avec *block.pls*, ce qui n'a pas été réalisé pour l'instant.

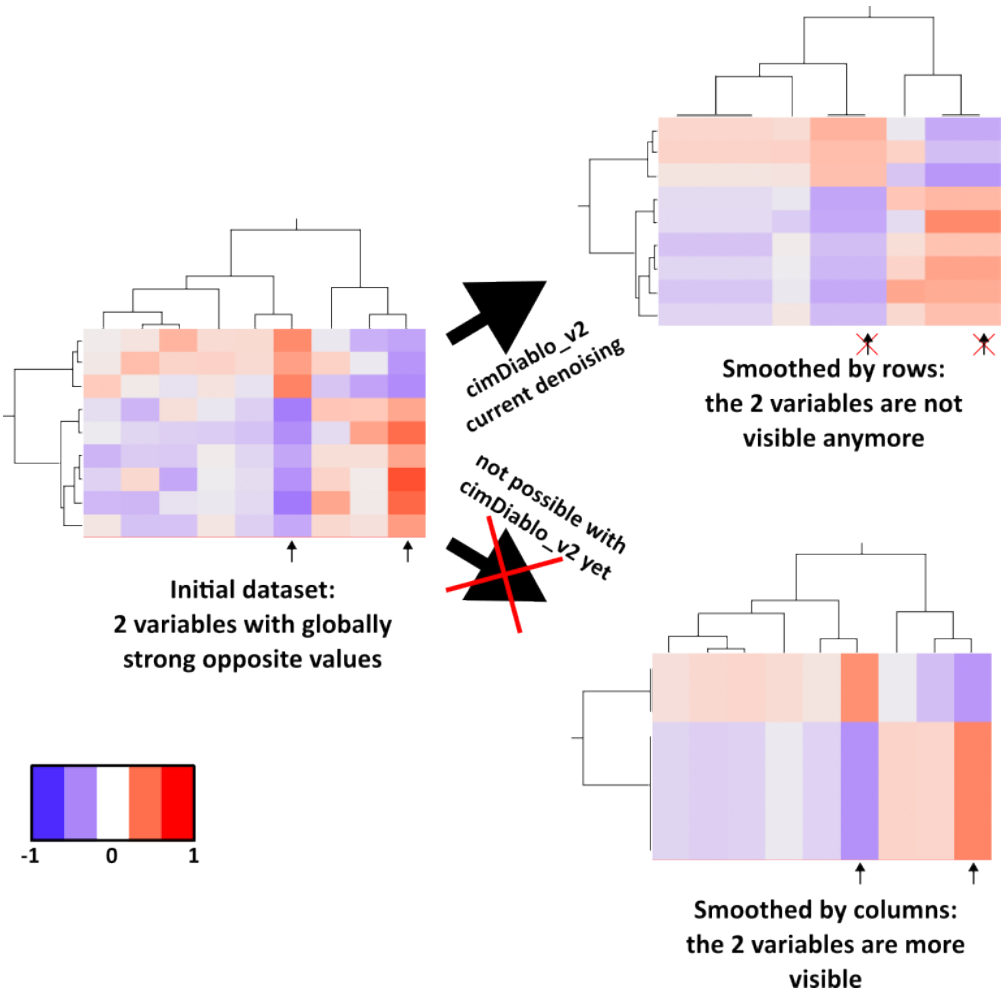


FIGURE 6.13 Schéma de la sélection de variables après un "lissage" par lignes ou colonnes.
À partir d'un jeu de données initial contenant ici 9 lignes (observations) et 9 colonnes (variables) avec différents profils de valeurs (légende en bas à gauche) et notamment deux variables aux valeurs plus fortes (positives ou négatives) relativement à celles des autres variables, le but est de pouvoir les identifier après "débruitage". Avec le "lissage" par lignes de cimDiablo_v2 (en haut à droite de la figure), ces variables ne sont plus identifiables. Avec un "lissage" par colonnes par groupes de lignes (en bas à droite de la figure), ces deux variables sont clairement identifiables.

6.5 Points clés

— D'un point de vue méthodologique, l'intégration des données animales a été l'occasion de tester sur un nouveau jeu de données le tutoriel et la fonction cimDiablo_v2, mais aussi de comparer différentes approches de régression PLS.

Le tutoriel a montré des résultats satisfaisants dans la préparation à l'intégration des données protéomiques et phénotypiques. Toutefois, ce tutoriel ayant été conçu pour être utilisable dans divers projets, son utilisation doit être associée à une expertise biologique sur les données utilisées et questionnements scientifiques associés, comme ça a été le cas dans ce projet avec de multiples interactions avec les co-auteurs de l'article.

Plusieurs types de régression PLS, à savoir les PLS1, PLS2 et MB-PLS ont été utilisées sur ces données et comparées. Il résulte de cette comparaison deux éléments

majeurs : ces méthodes se différencient premièrement par le nombre de variables et de blocs de variables considérés, et deuxièmement par l'interprétation biologique des listes obtenues par ces différentes approches. Ainsi, pour cet article, nous avons fait le choix de conserver les résultats de la PLS1 et de la MB-PLS et de mettre en avant la complémentarité de ces approches dans le but de sélectionner un faible nombre de protéines, à la fois les plus explicatives de variables phénotypiques au sein d'un tissu (PLS1), et avec des interactions fortes avec ce/ces phénotypes comparativement aux autres interactions protéines-phénotypes en considérant simultanément les protéines des trois tissus et les sept phénotypes (MB-PLS).

Nous avons aussi montré des limites d'utilisation de la fonction cimDiablo_v2 pour sélectionner des protéines candidates. Sans "débruitage", il est possible d'utiliser cimDiablo_v2 pour identifier les protéines les plus corrélées à chaque phénotype. Cependant, avec "débruitage", il n'est plus possible de mener ce type de sélection. En effet, le "débruitage" implique un "lissage" des données par lignes modifiant la variabilité des colonnes, empêchant alors le clustering des protéines et phénotypes selon leur corrélation.

— Ces analyses ont résulté en une liste de 31 protéines très fortement associées aux phénotypes, et que nous proposons comme marqueurs potentiels de la composition chimique et tissulaire des carcasses, à vérifier sur d'autres bovins. La pertinence de cette liste est attestée par les relations déjà établies entre certaines des protéines et la composition corporelle chez l'homme ou le rongeur. Parmi ces protéines, nous avons notamment identifié SNTB2 (foie), HIBCH (muscle) et HNRNPA2B1 (tissu adipeux) liées à la composition tissulaire et chimique de la carcasse, CCT8 (foie), RPS24 (muscle) et DRG1 (muscle) liées à la quantité de protéines fixées, ainsi que OMA1 (foie), COX7A2 (muscle), et CT027 (tissu adipeux) associées au nombre total d'adipocytes dans le tissu adipeux périrénal.

Troisième partie

Conclusion et perspectives

Chapitre 7

Conclusions et perspectives

Durant cette thèse, l'objectif a été de proposer des développements méthodologiques dans le but d'intégrer des données omiques diverses et à différents "niveaux" d'intégration pour répondre à différentes questions biologiques (Chapitre 4). Les principaux points qui ressortent de l'utilisation du tutoriel, de `mixOmics` et de cimDiablo_v2 sur les données plantes et animales de la thèse, autant d'un point de vue scientifique (résultats biologiques) que méthodologique (concepts, apports et limites du tutoriel, `mixOmics` de cimDiablo_v2) sont résumés ici, et des perspectives sont proposées.

7.1 En terme de biologie

7.1.1 Conclusion des principaux résultats scientifiques

Nous résumons dans cette section les résultats biologiques obtenus en reprenant les questionnements scientifiques présentés initialement dans la section introductive de ce document (Chapitre 4 Pages 43 à 47).

7.1.1.1 Chez les plantes

— ***L'intégration des données omiques permet-elle d'identifier les gènes dont la régulation par méthylation/expression est spécifique de certaines populations d'une espèce pérenne (peuplier) ou conservée entre ces populations d'origines géographiques diverses associées à différentes contraintes environnementales ?***

Les travaux menés dans le cadre de la thèse sur 10 populations de peuplier d'origines géographiques distinctes en Europe montrent que :

1. les facteurs majeurs qui impactent les variations omiques observées sont le type de données (expression, méthylation) et les régions géniques (promoteur, corps du gène), puis les contextes de méthylation (CG, CHG et CHH). Ces résultats suggèrent donc que pour la méthylation, il existe une différence majeure entre méthylation des promoteurs et corps des gènes et ce pour les différents contextes de méthylation. Sur le corps du gène, l'article Feng et al. (2010) montre en effet que la méthylation est présente dans divers organismes des règnes animal et végétal, et suggère que ce

phénomène pourrait avoir une origine ancienne antérieure au dernier ancêtre commun des plantes et animaux, ce qui pourrait expliquer son impact majeur conservé entre espèces ou contextes de méthylation. Par ailleurs, chez les animaux, seule la méthylation CG est observée (He et al., 2011), ce qui pourrait en partie expliquer les niveaux de méthylation plus élevés observés en contexte CG dans nos analyses chez les plantes (Sow et al., 2023).

2. à l'échelle du génome entier, il n'existe pas d'impact systématique de la méthylation sur la régulation génique (Nuo et al., 2016; Bewick and Schmitz, 2017; Ma et al., 2020). Néanmoins, au-delà d'un seuil de méthylation, qui diffère selon la variable de méthylation considérée, l'expression du gène semble systématiquement réprimée.
3. 143 *master regulators-drivers* présentant un profil de méthylation et d'expression fortement contrasté (forte méthylation et faible expression) et stable quelle que soit la population considérée sont impliqués dans le métabolisme des ARNs et le cycle cellulaire, incluant le gène Di19 (*Drought induced 19*), jouant un rôle dans la tolérance à la sécheresse (Wu et al., 2022).

— ***L'intégration des données omiques permet-elle d'identifier les spécificités et similitudes de la régulation des gènes par méthylation/expression au cours du développement du grain entre deux espèces (maïs et Brachypodium) issues d'un ancêtre commun datant de 65 millions d'années ?***

Les travaux menés dans le cadre de la thèse sur deux espèces de céréales montrent que :

1. le facteur majeur de la variation omique est à nouveau le type de données (expression, méthylation) puis la région génique considérée (promoteur, corps du gène), le stade de développement du grain étant généralement la variabilité la moins discriminante (au même titre que l'effet populationnel chez le peuplier).
2. trois types de profils ont été identifiés : quelques gènes très fortement méthylés et faiblement exprimés, quelques gènes faiblement méthylés et très fortement exprimés, et enfin la majorité des gènes aux valeurs de méthylation et d'expression relativement faibles. Les gènes faiblement méthylés et fortement exprimés sont impliqués dans des processus de signalisation cellulaire (*transport, transmembrane, ligase, cation, ion, acid*) et les gènes fortement méthylés et faiblement exprimés sont impliqués dans des activités de transcription et d'hydrolase.

L'intégration des données omiques chez les plantes (peuplier et céréales) démontre que le contexte puis la région génomique sur laquelle s'applique la méthylation a le plus fort impact sur l'expression des génomes/gènes au-delà de toute autre composante d'un système expérimental (les différents génotypes, stades, populations et contraintes considérés). L'intégration des données omiques a alors permis d'identifier les gènes dont la régulation par expression/méthylation est spécifique de ces composantes du système biologique, ainsi que les gènes

dont la régulation omique par expression/méthylation est stable, *i.e.* maintenue entre toutes les composantes du système expérimental. Les fonctions biologiques ainsi identifiées peuvent ouvrir de nouvelles pistes sur à la fois les déterminismes génomiques de la réponse spécifique des plantes à leur environnement, et à l'inverse sur les fonctions biologiques génériques, *i.e.* dont la régulation doit être maintenue quelles que soient les conditions environnementales, afin, à terme, de comprendre le rôle fonctionnel majeur de tels gènes qui apparaissent centraux à la biologie de la plante.

7.1.1.2 Chez les animaux

— ***L'intégration des données omiques permet-elle d'identifier les signatures moléculaires de la composition tissulaire ou chimique des carcasses bovines ?***

Les travaux menés dans le cadre de la thèse montrent que :

1. Chez le bovin, nous avons identifié 31 protéines du foie, du muscle ou du tissu adipeux dont la variabilité est la plus fortement associée à celles des variables de la composition des carcasses, à la fois lorsque les protéines d'un même tissu (PLS1) et des 3 tissus (MB-PLS) sont considérées.
2. Parmi ces protéines, nous avons notamment identifié SNTB2 (foie), HIBCH (muscle) et HNRNPA2B1 (tissu adipeux) liées à la composition tissulaire et chimique de la carcasse, CCT8 (foie), RPS24 (muscle) et DRG1 (muscle) liées à la quantité de protéines fixées, ainsi que OMA1 (foie), COX7A2 (muscle), et CT027 (tissu adipeux) associées au nombre total d'adipocytes dans le tissu adipeux périrénal.
3. Ces protéines, bien que peu étudiées chez le bovin, sont connues chez l'homme et le rongeur comme étant soit associées à la composition corporelle (Bjune et al., 2021) et en particulier à l'adiposité (Carty et al., 2014; Li et al., 2022; Chen et al., 2022), soit comme régulatrices de voies métaboliques impliquées dans la composition corporelle, par exemple la sensibilité à l'insuline (Hebel et al., 2015) ou le métabolisme du glucose (Quirós et al., 2013; Bishop et al., 2022). Cette liste de 31 protéines constitue donc l'étape de découverte de biomarqueurs selon le processus classiquement utilisé en médecine humaine (Rifai et al., 2006; Surinova et al., 2011) et récemment adapté à l'espèce bovine (Bonnet, 2018; Bonnet et al., 2020). Aussi, les perspectives de cette étude sont de vérifier un différentiel d'abondance des protéines identifiées sur de nouveaux groupes de bovins divergents par la composition corporelle, afin de "qualifier" ces biomarqueurs. Enfin, les relations entre l'abondance de ces protéines et les valeurs de composition corporelle seront mesurées sur une population représentative de ce caractère biologique. À l'issue de ce processus, un outil de phénotypage basé sur la quantification de protéines pourra être envisagé afin de remplacer les méthodes actuelles, longues, coûteuses et destructives pour estimer la composition des carcasses, et donc peu compatibles avec

la sélection génétique par exemple.

7.1.2 Perspectives pour l'intégration omique d'autres types de données biologiques

Au-delà des différents jeux de données étudiées durant la thèse, il existe d'autres types de données omiques qui ouvrent de nouvelles perspectives dans l'intégration de données omiques : les données temporelles et cellulaires. Nous discutons ici ces différents types de données et les potentiels cas d'utilisation de cimDiablo_v2 dans ce contexte.

7.1.2.1 Données temporelles avec timeOmics

Les données temporelles sont des données produites sur les mêmes observations à différents instants, par exemple sur plusieurs heures, jours, semaines ou mois, qui nécessitent généralement des adaptations méthodologiques. Ces données sont alors intégrées pour répondre à différentes questions, par exemple : Comment un stress hydrique modifie-t-il l'expression du génome lors du développement des plantes ? Comment la croissance pondérale module-t-elle le protéome des tissus adipeux et musculaire chez le bovin ? Comment les patients réagissent-ils à un traitement ?, *etc.* Le paquet R `timeOmics` (Bodein et al., 2022) a été développé pour intégrer des données temporelles tout en conservant la structure générale de `mixOmics`. Ainsi, lorsque les blocs omiques de `mixOmics` sont des matrices comportant classiquement les individus en lignes et variables omiques en colonnes, `timeOmics` considère en plus ces matrices à différents instants. Ce paquet réalise alors les étapes de pré-traitement, modélisation, clustering et enfin validation. La Figure 7.1 illustre ce fonctionnement.

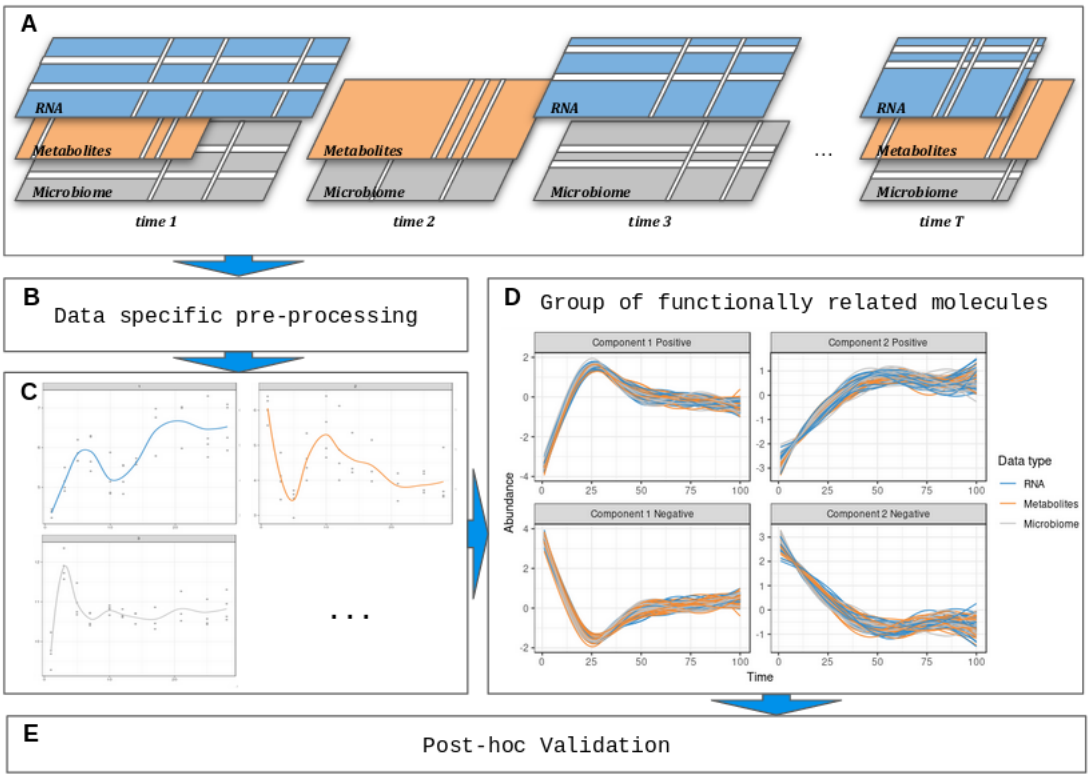


FIGURE 7.1 Fonctionnement de timeOmics

Figure et légende issues de `https://github.com/abodein/timeOmics`, octobre 2023. ***A-*** *timeOmics is a generic data-driven framework to integrate multi-Omics longitudinal data measured on the same biological samples and select key temporal features with strong associations within the same sample group. The main steps of timeOmics are :* ***B-*** *a pre-processing step. Normalize and filter low-expressed features, except those not varying over time.* ***C-*** *a modelling step. Capture inter-individual variability in biological/technical replicates and accommodate heterogeneous experimental designs.* ***D-*** *a clustering step. Group features with the same expression profile over time. Feature selection step can also be used to identify a signature per cluster.* ***E-*** *a post-hoc validation step. Ensure clustering quality.*

Plusieurs possibilités sont envisageables pour utiliser cimDiablo_v2 sur des données produites à différentes temporalités. Premièrement, il est possible de placer dans les mêmes blocs les variables qui diffèrent par leur temporalité, ou au contraire de créer un bloc par temporalité. La deuxième option, inspirée de `timeOmics` et notamment de la Figure 7.1.C, consiste à modéliser chaque variable en fonction du temps en considérant les valeurs des différentes observations. Ainsi, le modèle détermine la valeur de chaque variable à chaque instant de la cinétique. cimDiablo_v2 pourrait ainsi être utilisé sur ces données composées des données temporelles en lignes et variables en colonnes. Cette deuxième option ne serait cependant pas toujours adaptée, notamment si l'intérêt de l'intégration concerne spécifiquement la sélection d'individus/gènes au profil particulier, les observations étant prises en compte dans la modélisation mais n'apparaissant plus sur les données intégrées par cimDiablo_v2.

7.1.2.2 Données *single-cell*

Les données *single-cell* (parfois traduites en "cellule unique") sont de plus en plus produites pour une compréhension plus fine de la complexité biologique en produisant les données omiques pour chaque cellule plutôt que de les moyenner sur un ensemble de cellules d'un tissu ou organe. Il devient alors possible de décrire plus précisément les interactions ou spécificités cellulaires (ex : les adipocytes et fibres musculaires qui constituent le muscle ont-ils des spécificités métaboliques selon la composition des carcasses ?). Ainsi, plusieurs méthodes ont été proposées ces dernières années pour produire des données *single-cell* de plusieurs types de données omiques simultanément (Kashima et al., 2020; Athaya et al., 2023). Les différentes familles de méthodes présentées dans ce manuscrit sont exploitées dans les analyses *single-cell*, à savoir les méthodes de réduction de dimension, bayésiennes, basées sur la similarité, en réseaux ou réseaux de neurones artificiels (Adossa et al., 2021; Stanojevic et al., 2022; Athaya et al., 2023).

Les données ainsi obtenues sont généralement représentées sous différentes matrices comportant toutes les mêmes gènes en lignes et les mêmes cellules en colonnes, avec une matrice par type de données omiques. Ainsi, comme illustré en Figure 7.2, l'intégration de ces données devrait idéalement être réalisée simultanément sur les gènes et les cellules pour profiter pleinement de la structure des données. Néanmoins, en les intégrant seulement par gènes pour utiliser cimDiablo_v2, la réduction de dimension s'effectuerait sur les cellules afin d'en extraire les informations majeures, ce qui pourrait donner des résultats intéressants bien qu'il soit difficile pour l'instant d'imaginer l'impact du "débruitage" sur de telles données. Enfin, dans le cas où les données sont produites pour des milliers de cellules, l'intégration de telles données serait aussi l'occasion de tester cimDiablo_v2 sur des données contenant à la fois de nombreuses lignes et colonnes, ce qui n'a pas été réalisé dans le cadre de cette thèse. Pour toute autre structuration de données *single-cell* que celle de la Figure 7.2, il faudrait étudier au cas par cas son éventuelle compatibilité avec cimDiablo_v2.

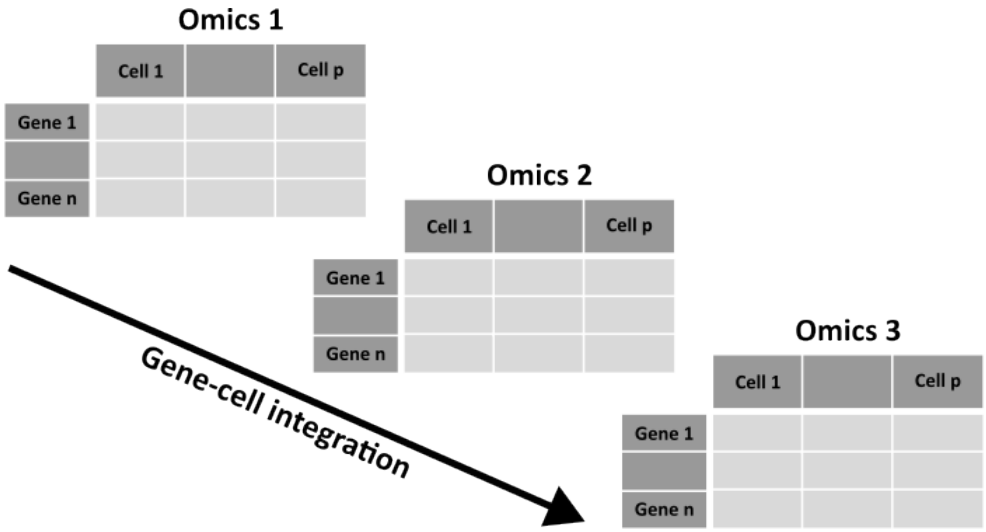


FIGURE 7.2 Schéma illustratif de l'intégration de données *single-cell*
Pour des données *single-cell* composées de différents blocs omiques avec les mêmes gènes en lignes et cellules en colonnes, l'intégration s'effectue non pas par lignes (gènes) ou par colonnes (cellules), mais par les deux à la fois.

7.1.2.3 Ajout d'une ou plusieurs variables qualitatives

Toutes les analyses présentées dans ce manuscrit portent sur des données quantitatives, bien que de nombreuses analyses intégratives fassent aussi intervenir une ou plusieurs variables qualitatives, que ce soit pour discriminer les observations par rapport à cette variable (ex : quels gènes sont différentiellement exprimés entre variétés de blé, ou en réponse à une contrainte par exemple thermique ou hydrique ? quelles protéines sont différentiellement abondantes dans les tissus de bovins en fonction de leur alimentation Herbe *vs.* Maïs ?) ou au contraire pour analyser les données tout en considérant le biais représenté par cette variable (ex : quels sont les gènes les plus fortement exprimés, autant pour les plantes stressées que les plantes témoins ?). Ces deux cas sont considérés dans `mixOmics`, respectivement par les versions discriminantes et les versions *mint* des différents types de régressions PLS.

Les fonctions non discriminante (*block.pls*) et discriminante (*block.plsda*) de la MB-PLS de `mixOmics` étant très similaires dans leur implémentation, il est possible d'adapter cimDiablo_v2 aux objets de *block.plsda* presque sans changement. Cette version discriminante de cimDiablo_v2, qui peut aussi être considérée comme la version avec "débruitage" de la fonction *cimDiablo* de `mixOmics`, a été testée brièvement sur les données céréales. La Figure 7.3 illustre cette utilisation sur les données méthylomiques et transcriptomiques du maïs associées à une variable qualitative indiquant si les gènes sont en copie unique (singleton) ou en paire (dupliqué) dans le génome du maïs. L'objectif d'une telle analyse résidait dans l'étude de l'évolution des gènes après des évènements de duplication du génome, processus récurrent au cours de l'évolution des génomes de plantes, sur la régulation (méthylation/expression) des gènes. En centrant cette analyse chez le maïs, ayant subi une duplication totale de son génome il y a 5 millions d'années, est-ce que les gènes présents en paires aujourd'hui ou retournés à l'état de

singletons ont des profils de régulation différents ? Dans cette figure, la variable qualitative est ainsi divisée en de nombreux groupes de quelques gènes en singletons ou paires, se traduisant par une absence de lien entre le nombre de copies du gène dans le génome et la régulation de ces gènes (méthylation et expression). Ce résultat semble s'opposer aux conclusions de l'article Bellec et al. (2023) qui montre chez le maïs que les gènes en simple copie seraient en moyenne significativement plus exprimés et méthylés que les gènes en double copie. Toutefois, une différence majeure entre ces approches est que cimDiablo_v2 présente l'entièreté des valeurs quand les tests statistiques utilisés dans l'article Bellec et al. (2023) comparent seulement leurs moyennes. Finalement, bien que nous ayons montré que cimDiablo_v2 pouvait être utilisé sur les objets *block.plsda* après quelques modifications mineures du code, il est nécessaire d'étudier l'impact du "débruitage" de cimDiablo_v2 en analyse discriminante avant de rendre accessible cette fonctionnalité. Néanmoins, ce "débruitage" s'effectuant sous forme de "lissage" par lignes tout comme la discrimination des observations (les gènes ou les individus), et les fonctions *block.pls* et *block.plsda* étant très similaires, il n'y a pour l'instant pas de raison de penser que cimDiablo_v2 en version discriminante ne fonctionnerait pas.

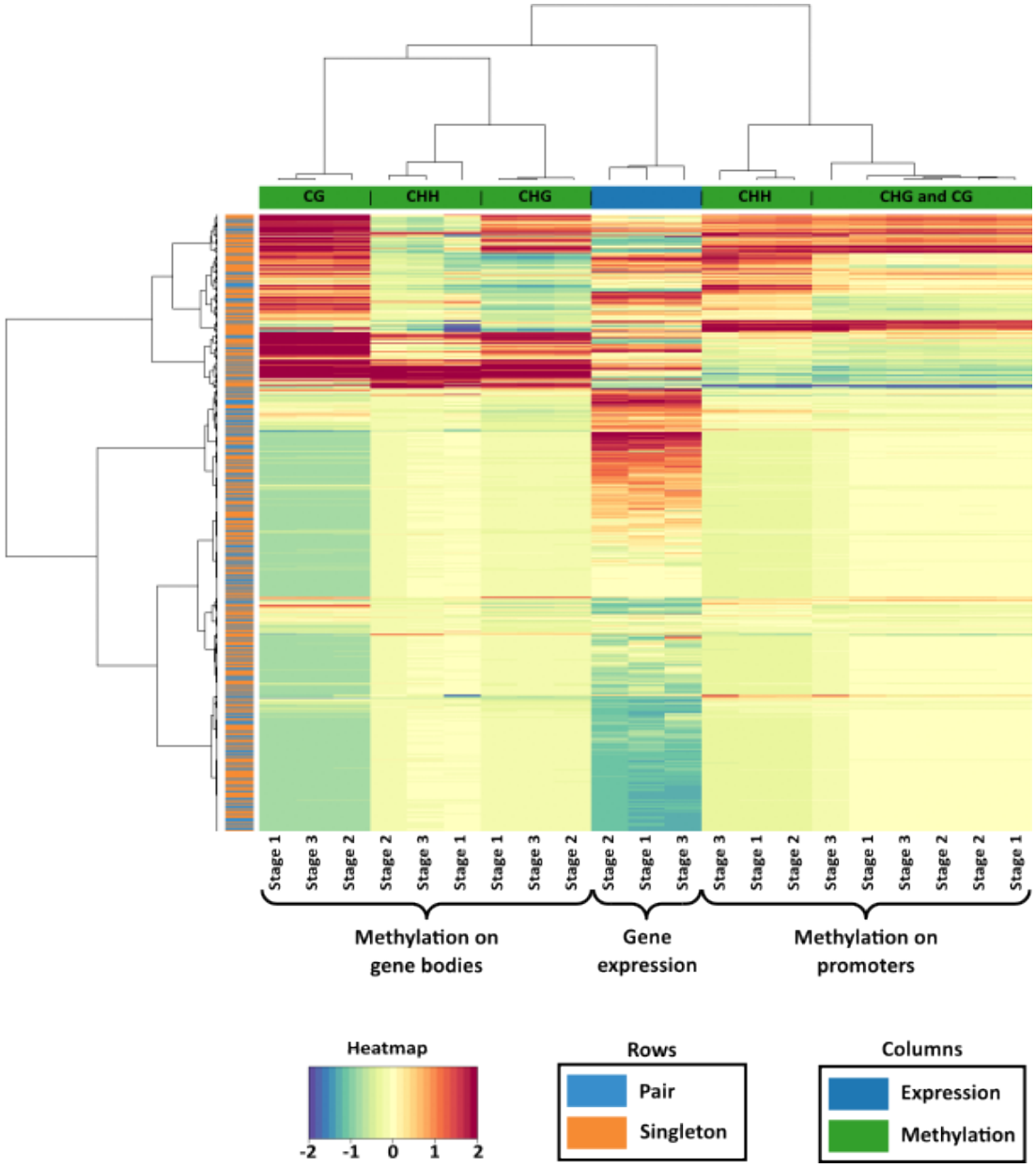


FIGURE 7.3 cimDiablo_v2 en version discriminante sur les données maïs

Chaque ligne correspond à un gène, chaque colonne à une variable omique. Les lignes et colonnes sont regroupées selon les dendrogrammes issus du clustering hiérarchique avec la méthode Ward et la distance euclidienne. Les valeurs représentées selon le code couleur de la *heatmap* sont celles des données initiales, centrées et réduites, "débruitées", de nouveau centrées et réduites, et enfin coupées en [-2, 2] (toute valeur inférieure à -2 est fixée à -2, toute valeur supérieure à 2 est fixée à 2). "Stage 1", "Stage 2" et "Stage 3" sont les stades de développement du grain à 7, 15 et 35 jours. Les deux blocs de variables omiques sont légendés en bas à droite de la figure. Pour chaque gène, la légende en bas de la figure indique s'il se trouve en un (singleton) ou deux (pair) copies dans le génome.

Concernant la version *mint*, aucune tentative n'a été réalisée durant la thèse pour la tester avec cimDiablo_v2. Cependant, bien que le "lissage" par lignes de cimDiablo_v2 ne semble théoriquement pas poser de problème à cet algorithme, et que les *mint.pls* et *mint.block.pls* possèdent aussi en valeurs de sorties les composantes et les poids utilisés par cimDiablo_v2 pour "débruiter" les données, il n'existe pas encore dans `mixOmics` de fonction graphique comme

cim ou *cimDiablo* pour représenter les résultats sous forme de *heatmap* avec dendrogrammes. Un développement supplémentaire serait donc nécessaire avant de tenter un "débruitage" avec l'approche *mint*.

7.2 En terme de méthodologie

7.2.1 Conclusion des principaux résultats méthodologiques

Dans le cadre de cette thèse, les multiples manières d'intégrer les données ont été abordées sous trois prismes : selon les types de données biologiques (génomiques, épigénomiques, transcriptomiques, protéomiques, métabolomiques, phénotypiques, *etc.*), les "niveaux" d'intégration de ces données (par gènes, protéines, tissus, individus, populations d'individus, espèces, stades développement, conditions expérimentales, *etc.*) et les questions biologiques (résumées en stratégies de description, de sélection et de prédiction).

Deux développements méthodologiques ont alors été proposés : un tutoriel en 6 étapes présentant les principaux points à aborder à partir de la question biologique et des données acquises pour mener à une intégration concluante des données, et une exploitation de l'outil `mixOmics` avec une nouvelle fonction nommée cimDiablo_v2 pour intégrer différentes données et répondre à différentes question tout en "débruitant" les données pour ne conserver qu'une partie de la variabilité des données.

7.2.1.1 Le tutoriel

Le tutoriel est composé de 6 étapes : 1- l'acquisition des données et leur structuration sous forme matricielle, 2- la définition de la question biologique, 3- le choix de l'outil intégratif selon les données et la question biologique, 4- le pré-traitement des données, 5- les analyses préliminaires, et 6- l'intégration multi-omiques.

Ce tutoriel a été utilisé durant la thèse sur les données peuplier et les données bovines. Le Tableau 7.1 résume les principaux points à chaque étape du tutoriel pour chaque analyse, révélant les différences majeures entre ces approches.

	Data matrix	Biological question	Tools selection	Data pre-treatment	Preliminary analysis	Multi-omics integration
Poplar	28 267 (or 24 962) genes in rows 2 databases: 60 methylomics & 10 transcriptomic variables in columns	Description (methylation/ expression interactions) Selection (genes with extreme profiles)	cimDiablo_v2	Expression in TMM Methylation in rbd Missing values deletion Log2	Correlation matrix PCA Boxplots	No whole genome systematic methylation/expression interactions 143 candidate genes selected with high methylation and low expression across poplar poluations, with functions enriched for involvement in carbohydrate, cell cycle, phosphorus and metabolic process
Bovine	17 individuals in rows 2 databases: 9166 proteins (4054 liver, 3085 adipocyte tissue, 2027 muscle) & 7 phenotypes in columns	Description (tissues interactions, proteins/phenotypes interactions) Selection (proteins explicative of phenotypes' variability)	MB-PLS (mixOmics) cimDiablo_v2	Proteins abundance Missing values deletion Log2 Missing values inference	Correlation matrix PCA PLS1 & PLS2	31 candidate proteins selected within and across tissues, relevant with the regulatory and metabolic pathways involved in nutrient partitioning or tissue growth

TABLE 7.1 Résumé des principales étapes du tutoriel sur les données peuplier et bovines de la thèse

Les principales différences entre les deux analyses à chaque étape du tutoriel sont les suivantes. Les types de données omiques et phénotypiques ainsi qu'une structuration matricielle avec peu ou beaucoup de lignes comparé au nombre de colonnes. Les questions biologiques précises, bien que regrouper dans les mêmes catégories de description et sélection. L'utilisation ou non de la MB-PLS. Les normalisations spécifiques au type de données omiques et la gestion des valeurs manquantes. L'utilisation ou non de régressions PLS. Les résultats différents selon tous les points précédents.

Finalement, il ressort de cette double utilisation du tutoriel les points suivants :

— Le tutoriel, bien qu'initialement développé sur des jeux de données comportant des gènes en lignes et variables omiques en colonne (Chapitre 5), est aussi parfaitement utilisable sur d'autres jeux de données, notamment des individus en lignes et données omiques en colonnes (Chapitre 6).

— En pratique, il est possible de considérer simultanément plusieurs questions biologiques et d'utiliser plusieurs outils d'intégration des données, bien que le tutoriel propose plutôt de ne considérer qu'une question et un outil intégratif à la fois afin que les utilisateurs puissent dans un premier temps appréhender l'intégration des données omiques par des cas simples, et ainsi éviter de se perdre dans une procédure et des analyses trop complexes.

— Le tutoriel est présenté de manière linéaire, avec un ordre d'étapes à suivre. Cependant, en pratique, il est courant de revenir à une étape antérieure pour préciser certains points. C'est tout particulièrement le cas avec les étapes de pré-traitement des données et d'analyses préliminaires, qui sont en pratique menées conjointement.

— Finalement, le tutoriel a montré son adaptabilité dans différents contextes, amenant de nouveaux résultats biologiques dans les deux analyses (plantes et animaux).

— Le tutoriel a été rendu public dans une volonté de prendre en compte les principes *Findable, Accessible, Interoperable, and Reusable* (FAIR) (Wilkinson et al., 2016).

7.2.1.2 mixOmics et son développement cimDiablo_v2

mixOmics

Parmi les 13 outils présentés dans le cadre de ce manuscrit, `mixOmics` a l'avantage d'être l'un des plus adaptables aux différents types de données omiques mais aussi et surtout aux questions biologiques et stratégies intégratives associées (de description, sélection et prédiction). Ceci provient, d'une part, de la multitude de méthodes de réduction de dimension implémentées et, d'autre part, des nombreuses fonctions de visualisation des données associées à ces méthodes. `mixOmics` se démarque aussi des autres outils intégratifs développés en R par son interface unie et ses multiples ressources (site web et livre dédiés, forum actif), expliquant sa popularité dans le domaine de l'intégration multi-omiques.

cimDiablo_v2

La fonction cimDiablo_v2, inspirée des fonctions de `mixOmics`, apporte deux nouveautés par rapport aux autres fonctions de ce paquet. Premièrement, cimDiablo_v2 est la première fonction représentant graphiquement à la fois les lignes et colonnes des données intégrées par la fonction de régression PLS multi-blocs *block.pls* de `mixOmics`. Deuxièmement, au-delà de son utilisation sur les données brutes normalisées, cette fonction possède une nouvelle option dite de "débruitage" afin de conserver la variabilité des lignes partagée pour plusieurs colonnes (variables) tout en atténuant la variabilité biologique propre à une unique colonne. Le "débruitage" de cimDiablo_v2 représente ainsi les données initiales auxquelles ont été soustraites la variabilité non conservée par la MB-PLS (*i.e.* le résidu de la MB-PLS), qui peut être d'origine biologique ou non.

Cette fonction a été testée sur différentes données biologiques, révélant ses points forts mais aussi ses limites selon le contexte d'étude.

Selon la question biologique et la stratégie intégrative associée : La fonction cimDiablo_v2 a été utilisée pour répondre à des questions biologiques appelant trois grand types d'objectifs. Chez les plantes, les données ont été intégrées pour décrire les interactions omiques puis sélectionner des gènes au profil multi-omiques particulier, et chez les animaux, pour sélectionner des protéines fortement associées aux phénotypes dans un but de prédiction phénotypique. Malgré des similarités entre les objectifs, la sélection portait, chez les plantes, sur les gènes (en lignes) au profil extrême pour les différentes variables omiques (en colonnes), tandis que la sélection portait, chez le bovin, sur les protéines (en colonnes) au profil de variabilité chez 17 bovins similaire à celui des variables phénotypiques (aussi en colonnes). La double différence concerne donc le fait de sélectionner des lignes (plantes) ou des colonnes (bovin), et le fait de sélectionner d'une part selon des valeurs extrêmes (plantes) et d'autre part selon des varia-

bilités corrélées (bovin). Nous avons alors pu montrer que le "débruitage" de cimDiablo_v2, sous forme de "lissage" des lignes, était parfaitement adapté dans la sélection des lignes (gènes chez les plantes), mais pas des colonnes (protéines chez les bovins). Pour répondre au questionnement scientifique lié à l'analyse des données animales, nous avons ainsi opté pour des approches PLS.

Selon les types de données omiques : Chez les plantes, les données intégrées durant cette thèse ont été des données de méthylation de l'ADN ainsi que des données transcriptomiques, contrairement aux données protéomiques et phénotypiques intégrées chez les animaux d'élevage. Malgré les différences biologiques fondamentales entre ces données, il n'y a pas eu de difficulté majeure à appliquer cimDiablo_v2 dans l'une ou l'autre de ces analyses. Néanmoins, notons que cimDiablo_v2 n'est utilisable pour l'instant que sur des données continues, et qu'il pourrait alors être intéressant de tester le "débruitage" de cimDiablo_v2 avec la fonction *cimDiablo* qui considère en plus des variables quantitatives une variable qualitative.

Selon les "niveaux" d'intégration : Pour utiliser cimDiablo_v2, les données doivent être réparties en différentes matrices appelées "blocs" (*blocks*) possédant le même nombre de lignes. Ainsi, parmi les multiples "niveaux" d'intégration possibles, l'utilisateur doit ré-arranger les données pour qu'elles correspondent aux trois "niveaux" de cimDiablo_v2 que sont les lignes, les colonnes, et les blocs de colonnes. Pour les données bovines, nous avons considéré les 17 individus en lignes, l'ensemble des phénotypes et les tissus protéomiques en blocs, les variables phénotypiques et protéines en colonnes. Pour les données peuplier, le ré-arrangement a été plus complexe, avec des données par gènes, par individus de différentes zones géographiques, par méthylation pour différents contextes (CG, CHG, CHH) et régions géniques (promoteur, corps du gène), et expression. Les données des individus ont alors été moyennées par zone géographique, et chaque combinaison d'un des trois contextes de méthylation avec l'une des deux régions géniques étudiées a fait l'objet d'un bloc dit "méthylomique" afin de restreindre le nombre de "niveaux" d'intégration.

7.2.2 Perspectives méthodologiques

7.2.2.1 Sur le tutoriel

Le tutoriel étant en cours de publication, il n'a pas encore été utilisé par d'autres personnes, que ce soient d'autres statisticiens mais aussi et surtout des biologistes non mathématiciens. Il sera donc intéressant d'avoir leur retour dans les prochains mois afin d'éventuellement améliorer le tutoriel.

7.2.2.2 Sur mixOmics et cimDiablo_v2

Durant cette thèse, plusieurs pistes d'amélioration à `mixOmics` ont été identifiées, principalement concernant la fonction *block.pls*. En effet, il semble que le développement des fonctionnalités de `mixOmics` se concentrent plus sur la version discriminante (MB-PLSDA) que le version non-discriminante (MB-PLS) de la régression PLS multi-blocs. Ainsi, peu des fonctions graphiques et aucune d'optimisation de paramètres de ce paquet R ne sont pour l'instant utilisables pour la *block.pls*, ce qui a conduit au développement de cimDiablo_v2. Il serait donc intéressant, de la même manière, d'adapter les autres fonctions utilisables en MB-PLSDA à la MB-PLS.

Certaines pistes de réflexion et d'amélioration concernant cimDiablo_v2 ont aussi été abordées dans les chapitres précédents. Ainsi, nous avons notamment mentionné la possibilité de développer le "débruitage" de cimDiablo_v2 sur l'ACP et la PLS pour étudier plus en détail et dans des cas plus simples le fonctionnement et l'impact du "débruitage" sur notamment le "lissage" par lignes, le changement d'échelles, la proportion de valeurs significativement différentes, et la séparation entre la variabilité biologique d'intérêt et la variabilité résiduelle. Nous pouvons aussi imaginer dans le même objectif de tester le "débruitage" sur des données simulées.

Nous proposons ici d'autres pistes d'amélioration, autant sur des développements statistiques que sur l'intégration d'autres types de données biologiques.

Transformation à noyau avec mixKernel

Afin d'utiliser une méthode linéaire comme la régression PLS sur des données interagissant de manière non linéaire, une possibilité est de transformer préalablement ces données avec une méthode à noyau (*kernel*). Le principe de cette méthode est de projeter ces données sur un nouvel espace mathématique de plus grandes dimensions dans lequel les relations sont linéaires, comme illustré en Figure 7.4. L'ACP à noyau (ou *Kernel Principal Component Analysis* (KPCA)) est illustrée sous forme matricielle en Figure 7.5.

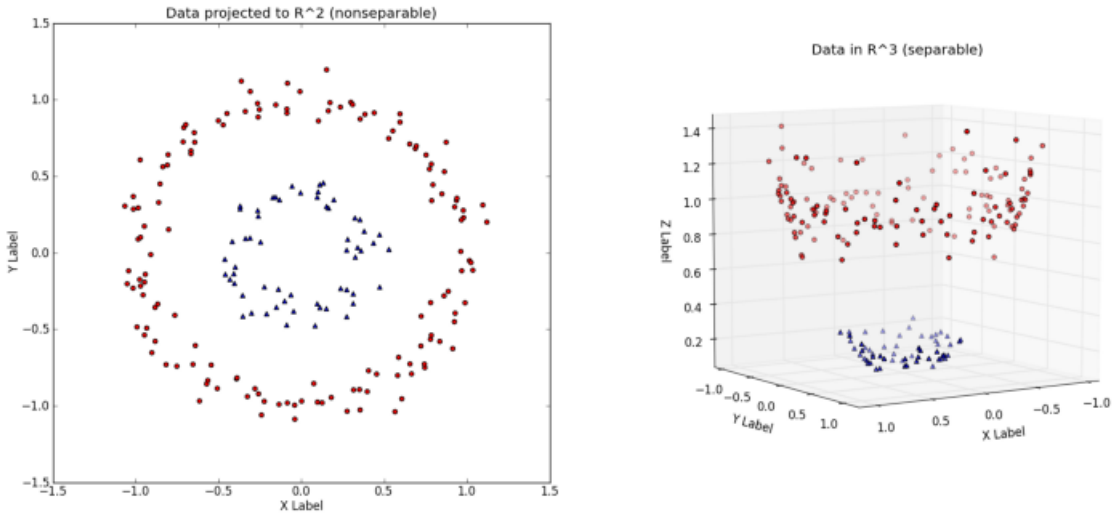


FIGURE 7.4 Exemple d'utilisation de l'"astuce du noyau" (*Kernel trick*)
Figure issue de `https://towardsdatascience.com/understanding-the-kernel-trick-e0bc6112ef78`, juin 2023. Deux groupes de données représentés par deux cercles ne peuvent être séparés par une simple droite (relation linéaire dans un espace de deux dimensions) dans le graphique de gauche, mais peuvent l'être par un plan (relation linéaire dans un espace de trois dimensions) dans le graphique de droite en ajoutant une variable Z calculée à partir des deux axes initiaux X et Y (avec $Z = X^2 + Y^2$).

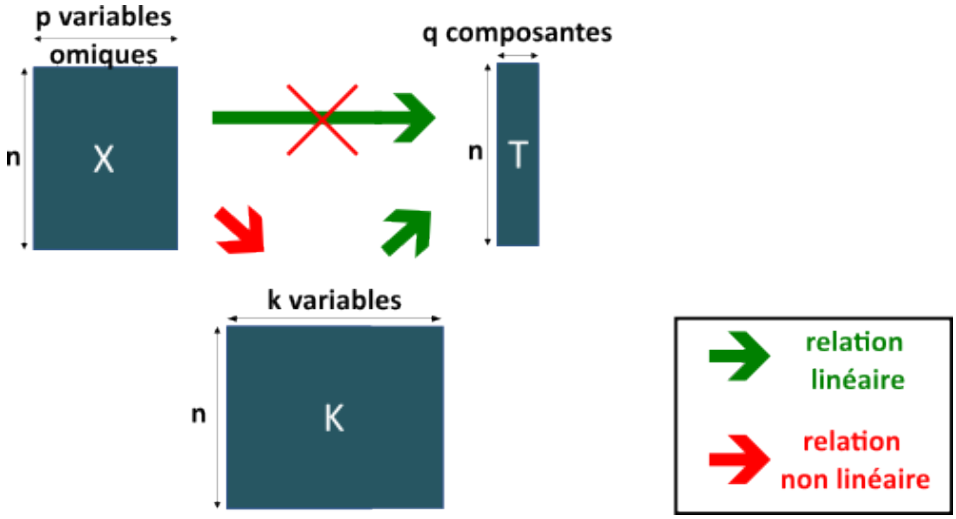


FIGURE 7.5 Illustration matricielle de l'ACP à noyau
Pour un jeu de données X sous forme matricielle, l'ACP crée une nouvelle matrice T contenant approximativement la même variabilité mais avec moins de colonnes, ces nouvelles colonnes appelées composantes étant des combinaisons linéaires des colonnes initiales appelées variables. Si les relations entre variables sont non linéaires, les composantes calculées ne contiendront pas la variabilité des données initiales. La transformation noyau crée alors un jeu de données intermédiaire K de plus grande dimension dans lequel les nouvelles variables interagissent de manière linéaire.

Le paquet `mixKernel` (Mariette and Villa-Vialaneix, 2018), compatible avec `mixOmics`, se concentre sur ces méthodes à noyaux en implémentant notamment l'ACP à noyau afin de contourner l'hypothèse de linéarité de l'ACP. Cependant, il ne semble pas exister actuellement dans `mixKernel` de méthode reprenant celles de régressions PLS avec ajout d'un noyau bien qu'elle soit implémentée par ailleurs, par exemple ici : `https://rdrr.io/cran/pls`

`/man/kernelpls.fit.html`. Un développement serait donc nécessaire pour adapter les régressions PLS au format de *mixOmics*. De plus, la transformation par noyau s'effectuant avant la réduction de dimension et par jeu de données, la généralisation de la régression PLS à noyau à la régression MB-PLS à noyau ne semble théoriquement pas poser de problème (Figure 7.6). Toutefois, en pratique, une des difficultés des méthodes à noyaux étant de choisir la transformation adaptée aux données, cette complexité augmente avec le nombre de blocs à transformer.

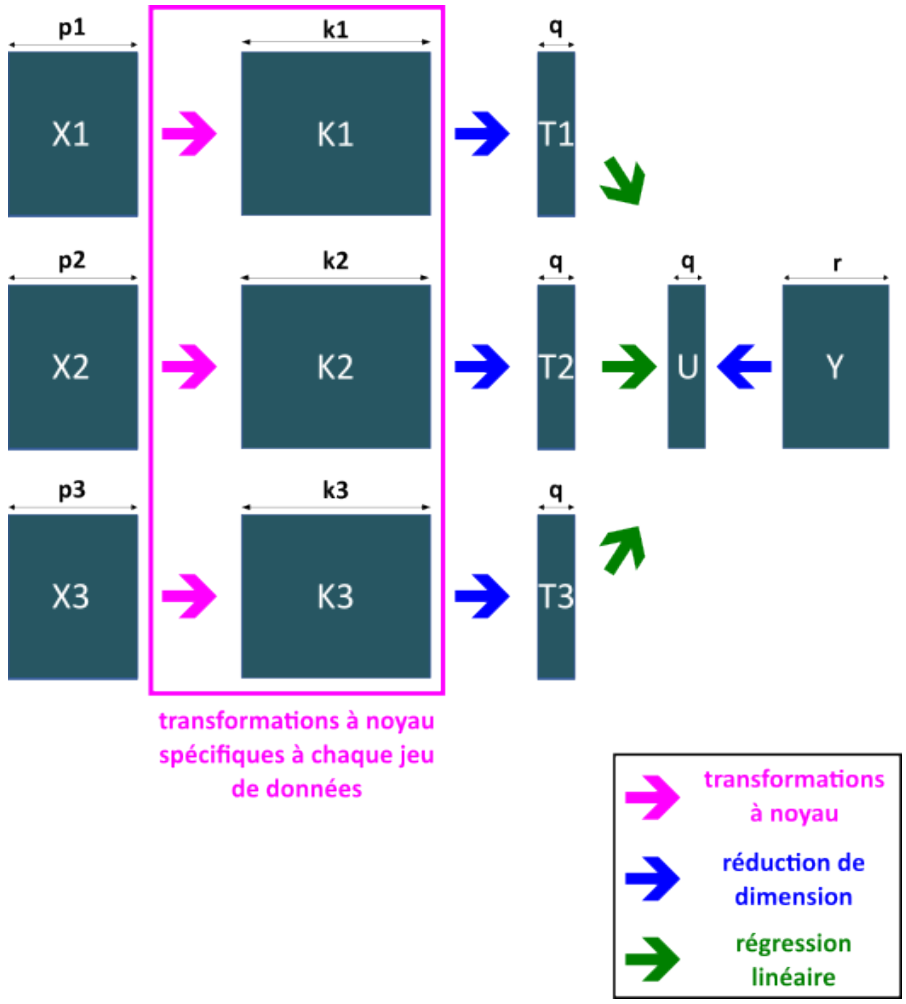


FIGURE 7.6 Illustration théorique du fonctionnement de la MB-PLS à noyau
Représentation schématique théorique de la MB-PLS à noyau sur des données sous forme matricielle comportant toutes le même nombre de lignes (observations) mais des nombres potentiellement différents de colonnes (variables). Le principe est similaire à celui de la régression MB-PLS, à savoir la réduction des blocs de données explicatives (ici trois blocs $X1$, $X2$ et $X3$) et du bloc de données expliquées (Y) en de nouveaux blocs ($T1$, $T2$, $T3$ et U) de sorte à maximiser les covariances entre ces nouveaux blocs. Une transformation noyau est alors applicable pour chaque jeu de données X avant de calculer les matrices T.

Évolution des régressions linéaires en approches non linéaires avec les auto-encodeurs

Les auto-encodeurs appartiennent à la famille des réseaux de neurones artificiels, et ont été proposés pour la première fois dans Kramer (1991). Ils sont formés de trois parties : l'encodeur,

la partie encodée (couche latente), et le décodeur. La Figure 7.7 représente cette architecture symétrique.

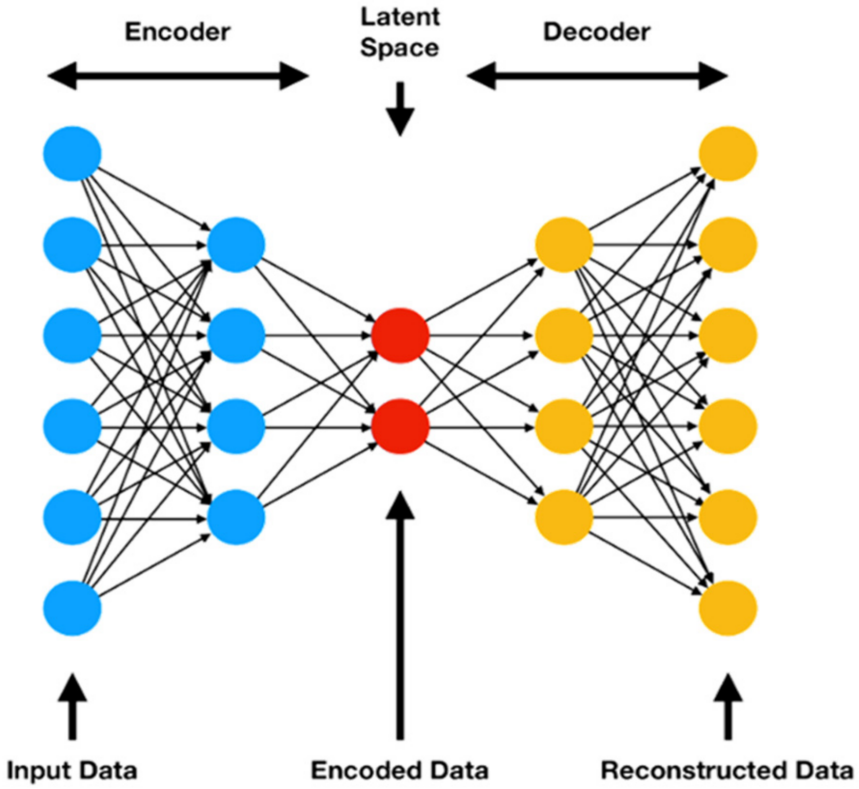


FIGURE 7.7 Architecture simplifiée d'un auto-encodeur
Figure issue de Tien et al. (2021). L'auto-encodeur est composé de trois parties : l'encodeur, la partie encodée (couche latente), et le décodeur. La première couche de l'encodeur correspond aux données initiales, avec par exemple un neurone par protéine, par gène, *etc.* Ces données sont encodées par une une ou plusieurs couches contenant de moins en moins de neurones jusqu'à obtenir une couche de données encodées de taille inférieure mais contenant pratiquement la même information. Cette couche est alors décodée à son tour par des couches contenant de plus en plus de neurones jusqu'à obtenir une nouvelle couche de taille identique à la couche initiale.

Durant la phase d'entraînement de l'algorithme ayant pour but de paramétrer le réseau, l'objectif de l'algorithme est, à partir de données initiales en première couche, de reconstruire ces données avec le plus de précision possible en couche de sortie. En ceci, l'auto-encodeur se rapproche fortement des thématiques étudiées durant cette thèse, à savoir la réduction et le "débruitage" des données. En effet, en tentant de reconstruire en sortie les données initiales à partir de couches comportant moins de neurones, la majeure partie de l'information initiale nécessaire à la reconstruction doit être synthétisée dans la couche encodée, à l'instar des composantes de l'ACP ou des régressions PLS. De plus, l'information initiale non contenue dans la couche encodée correspond au "bruit" et n'apparaît plus dans la couche de sortie, les données reconstruites sont alors des données "débruitées". Les auto-encodeurs ont donc des utilisations multiples, pour à la fois compresser l'information des données et les "débruiter". Toutefois, si les auto-encodeurs sont bien une généralisation à un cas non linéaire de l'ACP, ce n'est pas le cas pour les régressions PLS. En effet, dans la régression PLS, la réduction de chaque jeu

de données ne se fait pas indépendamment mais en lien avec les réductions des autres jeux de données, ce qui n'est pas proposé dans l'auto-encodeur classique. À notre connaissance, seul Zhang et al. (2019) a proposé une généralisation de l'auto-encodeur au cas de la PLS, et aucune structure n'a encore été proposée pour la MB-PLS.

L'avantage majeur des auto-encodeurs par rapport aux méthodes de réduction de dimension employées durant la thèse est de considérer des relations non linéaires entre les variables. En effet, bien que chaque neurone soit calculé à partir d'une combinaison linéaire des neurones de la couche précédente, une fonction particulière appelée "fonction d'activation" est ensuite appliquée à ce résultat, avec pour conséquence d'obtenir une relation non linéaire avec les nœuds de la couche précédente. De plus, avec la première couche correspondant aux variables omiques et une fonction d'activation linéaire, la deuxième couche sera bien formée de combinaisons linéaires des données initiales. Cependant, même dans ce cas, la troisième couche sera une combinaison linéaire de la deuxième, mais pas des données omiques initiales. D'un point de vue biologique, les auto-encodeurs offrent donc la possibilité de considérer des relations omiques plus complexes qu'avec des ACP.

Néanmoins, les principales limites des réseaux de neurones artificiels s'appliquent aussi aux auto-encodeurs. Premièrement, il n'existe pas ou peu de méthodes pour choisir les hyperparamètres (nombre de couches, de neurones par couche, la fonction d'activation, *etc.*) et deuxièmement, l'entraînement des auto-encodeurs nécessite une grande quantité de données pour ajuster les paramètres, notamment les poids des arêtes liant les neurones. Ceci explique pourquoi les auto-encodeurs sont pour l'instant principalement utilisés dans le domaine médical et particulièrement en cancérologie pour lequel ces jeux de données sont disponibles. Des analyses ont par exemple été menées dans ce domaine pour classifier (Franco et al., 2021; Madhumita and Paul, 2022; Zhang et al., 2019b) puis prédire (Lupat et al., 2023) des typologies de pathologies cancéreuses, regrouper des patients (Lemsara et al., 2020), prédire des variables cliniques (Tan et al., 2020; Ma and Zhang, 2019), *etc.* Toutefois, une approche possible avec les réseaux de neurones artificiels est de récupérer un réseau pré-entraîné sur d'autres données et ajuster ses paramètres à de nouvelles données moins abondantes sans modifier les hyperparamètres. Ainsi, il est envisageable que les auto-encodeurs soient fortement utilisés en biologie végétale et animale dans les prochaines années si une structure d'auto-encodeurs fait ses preuves sur un jeu de données et est adaptable à différents contextes biologiques.

Dans le cadre de cette thèse, les auto-encodeurs sont donc intéressants pour généraliser les approches de réduction de dimension à un cas non linéaire en remplaçant les composantes de l'ACP par les données encodées de l'auto-encodeur.

7.3 Conclusions et perspectives générales

7.3.1 Conclusions et perspectives générales de la thèse

L'intégration des données omiques est un domaine vaste qui peut être abordé de multiples manières. Dans cette thèse, nous avons identifié comme enjeux méthodologiques majeurs de l'intégration multi-omiques 1-le besoin d'intégrer différents types de données omiques, 2- à différents "niveaux" d'intégration et 3- pour répondre à différents types de questions biologiques. Nous avons ainsi développé un tutoriel, exploité l'outil `mixOmics` et développé la fonction cimDiablo_v2 dans l'objectif qu'ils puissent s'adapter au plus de contextes d'intégration possibles. Ainsi, nous avons décidé de tester ces développements sur des données omiques produites dans différents projets pour répondre à différents questionnements. Nous avons ainsi pu répondre aux questionnements scientifiques posés dans le cadre l'intégration de données omiques chez les plantes (peuplier et céréales) et les animaux (bovins).

Certaines limitations majeures dans l'intégration de données omiques n'ont pas ou peu été abordées durant cette thèse. Premièrement, nous avons uniquement considéré les données déjà mises sous forme matricielle, et composées de données quantitatives produites pour quelques centaines ou milliers de gènes/protéines et quelques dizaines de variables transcriptomiques/ méthylomiques ou d'individus. Ainsi, nous ne nous sommes pas confrontés à la problématique des données de très grandes dimensions (*Big Data*), ni à toutes les problématiques de production et transformation des données pour obtenir les données matricielles. De la même manière, les questions de la gestion des valeurs manquantes, de l'hétérogénéité et des différences d'effectifs entre les données n'ont été que brièvement abordées durant cette thèse.

En outre, les analyses de cette thèse se sont concentrées sur le paquet `mixOmics` et la réduction de dimension. Il serait donc intéressant d'exploiter d'autres méthodes, que ce soient d'autres méthodes de réduction de dimension comme les auto-encodeurs, ou des méthodes très différentes comme celles de réseaux, seulement brièvement décrites dans le cadre de ce travail la thèse.

7.3.2 Conclusions et perspectives générales de l'intégration multi-omiques

Plus généralement, dans le domaine de l'intégration de données omiques, la multiplication des publications méthodologiques et scientifiques des dernières années montrent un réel intérêt de l'approche intégrative pour obtenir de nouvelles connaissances biologiques. Néanmoins, beaucoup de travail dans le domaine reste encore à faire.

Premièrement, nous soulignons l'importance de suivre les instructions du Plan de Gestion des Données (PGD) pour chaque projet de recherche, afin de répondre aux attentes scientifiques en terme de Science Ouverte et d'Éthique de la recherche, ainsi qu'utiliser des données suivant le principe FAIR. Dans le domaine de l'intégration de données omiques, cela passe notamment par la considération des limites actuelles des méthodes intégratives dès le choix de la question

scientifique et du protocole expérimental, afin de produire des données compatibles avec les outils d'intégration disponibles. Pour cela, il est primordial d'engager des discussions régulières entre tous les agents participant à la conception du projet, à la production, au traitement et à l'analyse des données. Produire les données et méta-données selon le principe FAIR est aussi un enjeu majeur dans la réutilisation des données dans d'autres contextes intégratifs.

Deuxièmement, un autre enjeu majeur de l'intégration de données concerne la gestion des multiples problématiques liées aux données omiques, par exemple la gestion des valeurs manquantes au sein d'un jeu de données ou entre plusieurs jeux de données, des données temporelles, des données *single-cell*, des données de très grands volumes, des données très hétérogènes, *etc.* Le développement de méthodologies adaptées à ces différentes spécificités des données sera donc à poursuivre dans les prochaines années.

Enfin, avec la production toujours plus massive et à moindre coût des données omiques, la biologie entre progressivement dans le domaine de la *Big Data*, amenant des questions en termes de stockage, de transfert, d'accessibilité et d'analyse des données. Des changements logistiques majeurs, notamment la centralisation du stockage des données et des calculs dans des structures dédiées (*Data centers*), ont déjà débuté et seront à poursuivre. Les enjeux actuels sont alors de former l'ensemble de la communauté biologique à cette nouvelle approche de travail centralisé, ainsi que de développer de nouvelles méthodes pour intégrer ces données omiques de plusieurs giga-, téra- voire péta-octets de données.

Bibliographie

Abbas-Aghababazadeh, F., Li, Q., and Fridley, B. L. (2018). Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PLOS ONE*, 13(10) :e0206312.

Adossa, N., Khan, S., Rytkönen, K. T., and Elo, L. L. (2021). Computational strategies for single-cell multi-omics integration. *Computational and Structural Biotechnology Journal*, 19 :2588–2596.

Agamah, F. E., Bayjanov, J. R., Niehues, A., Njoku, K. F., Skelton, M., Mazandu, G. K., Ederveen, T. H. A., Mulder, N., Chimusa, E. R., and 'T Hoen, P. A. C. (2022). Computational approaches for network-based integrative multi-omics analysis. *Frontiers in Molecular Biosciences*, 9 :967205.

Aggarwal, C. C. (2018). An Introduction to Neural Networks. In *Neural Networks and Deep Learning*, pages 1–52. Springer International Publishing, Cham.

Agius, D. R., Kapazoglou, A., Avramidou, E., Baranek, M., Carneros, E., Caro, E., Castiglione, S., Cicatelli, A., Radanovic, A., Ebejer, J.-P., Gackowski, D., Guarino, F., Gulyás, A., Hidvégi, N., Hoenicka, H., Inácio, V., Johannes, F., Karalija, E., Lieberman-Lazarovich, M., Martinelli, F., Maury, S., Mladenov, V., Morais-Cecílio, L., Pecinka, A., Tani, E., Testillano, P. S., Todorov, D., Valledor, L., and Vassileva, V. (2023). Exploring the crop epigenome : a comparison of DNA methylation profiling techniques. *Frontiers in Plant Science*, 14 :1181039.

Akhter, Z., Bi, Z., Ali, K., Sun, C., Fiaz, S., Haider, F. U., and Bai, J. (2021). In Response to Abiotic Stress, DNA Methylation Confers EpiGenetic Changes in Plants. *Plants*, 10(6) :1096.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., and Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In Berry, M. W., Mohamed, A., and Yap, B. W., editors, *Supervised and Unsupervised Learning for Data Science*, pages 3–21. Springer International Publishing, Cham. Series Title : Unsupervised and Semi-Supervised Learning.

Alves De Freitas Guedes, F., Menezes-Silva, P. E., DaMatta, F. M., and Alves-Ferreira, M. (2019). Using transcriptomics to assess plant stress memory. *Theoretical and Experimental Plant Physiology*, 31(1) :47–58.

An, Y.-q. C., Goettel, W., Han, Q., Bartels, A., Liu, Z., and Xiao, W. (2017). Dynamic Changes of Genome-Wide DNA Methylation during Soybean Seed Development. *Scientific Reports*, 7(1) :12263.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10) :R106.

Anderson, S. N., Zynda, G. J., Song, J., Han, Z., Vaughn, M. W., Li, Q., and Springer, N. M. (2018). Subtle Perturbations of the Maize Methylome Reveal Genes and Transposons Silenced by Chromomethylase or RNA-Directed DNA Methylation Pathways. *G3 Genes|Genomes|Genetics*, 8(6) :1921–1932.

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., and Stegle, O. (2020). MOFA+ : a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1) :111.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6).

Athaya, T., Ripan, R. C., Li, X., and Hu, H. (2023). Multimodal deep learning approaches for single-cell multi-omics data integration. *Briefings in Bioinformatics*, page bbad313.

Athieniti, E. and Spyrou, G. M. (2023). A guide to multi-omics data collection and integration for translational medicine. *Computational and Structural Biotechnology Journal*, 21 :134–149.

Bady, P., Dolédec, S., Dumont, B., and Fruget, J.-F. (2004). Multiple co-inertia analysis : a tool for assessing synchrony in the temporal variability of aquatic communities. *Comptes Rendus Biologies*, 327(1) :29–36.

Ball, M. P., Li, J. B., Gao, Y., Lee, J.-H., LeProust, E. M., Park, I.-H., Xie, B., Daley, G. Q., and Church, G. M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnology*, 27(4) :361–368.

Baubec, T. and Akalin, A. (2016). Genome-Wide Analysis of DNA Methylation Patterns by High-Throughput Sequencing. In Aransay, A. M. and Lavín Trueba, J. L., editors, *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing*, pages 197–221. Springer International Publishing, Cham.

Bellec, A., Sow, M. D., Pont, C., Civan, P., Mardoc, E., Duchemin, W., Armisen, D., Huneau, C., Thévenin, J., Vernoud, V., Depège-Fargeix, N., Maunas, L., Escale, B., Dubreucq, B., Rogowsky, P., Bergès, H., and Salse, J. (2023). Tracing 100 million years of grass genome evolutionary plasticity. *The Plant Journal*, page tpj.16185.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society : Series B (Methodological)*, 57(1) :289–300.

Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., and Milanesi, L. (2016). Methods for the integration of multi-omics data : mathematical aspects. *BMC Bioinformatics*, 17(S2) :S15.

Bewick, A. J. and Schmitz, R. J. (2017). Gene body DNA methylation in plants. *Current Opinion in Plant Biology*, 36 :103–110.

Bishop, C. A., Machate, T., Henning, T., Henkel, J., Püschel, G., Weber, D., Grune, T., Klaus, S., and Weitkunat, K. (2022). Detrimental effects of branched-chain amino acids in glucose tolerance can be attributed to valine induced glucotoxicity in skeletal muscle. *Nutrition and Diabetes*, 12(1) :20.

Bjune, M. S., Lindquist, C., Hallvardsdotter Stafsnes, M., Bjørndal, B., Bruheim, P., Aloysius, T. A., Nygård, O., Skorve, J., Madsen, L., Dankel, S. N., and Berge, R. K. (2021). Plasma 3-hydroxyisobutyrate (3-HIB) and methylmalonic acid (MMA) are markers of hepatic mitochondrial fatty acid oxidation in male Wistar rats. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1866(4) :158887.

Bobrovskikh, A., Doroshkov, A., Mazzoleni, S., Cartenì, F., Giannino, F., and Zubairova, U. (2021). A Sight on Single-Cell Transcriptomics in Plants Through the Prism of Cell-Based Computational Modeling Approaches : Benefits and Challenges for Data Analysis. *Frontiers in Genetics*, 12 :652974.

Bodein, A., Scott-Boyer, M.-P., Perin, O., Lê Cao, K.-A., and Droit, A. (2022). timeOmics : an R package for longitudinal multi-omics data integration. *Bioinformatics*, 38(2) :577–579.

Boe-Hansen, G. B., Rêgo, J. P. A., Satake, N., Venus, B., Sadowski, P., Nouwens, A., Li, Y., and McGowan, M. (2020). Effects of increased scrotal temperature on semen quality and seminal plasma proteins in Brahman bulls. *Molecular Reproduction and Development*, 87(5) :574–597.

Bonnet, M. (2018). Proteomics of adipose tissue : from the molecular drivers of adipogenesis to the molecular phenotyping of ruminants. *International Journal of Health*, Animal Science and Food Safety :N. 1s (2018). Publisher : International Journal of Health, Animal Science and Food Safety.

Bonnet, M., Soulat, J., Bons, J., Léger, S., De Koning, L., Carapito, C., and Picard, B. (2020). Quantification of biomarkers for beef meat qualities using a combination of Parallel Reaction Monitoring- and antibody-based proteomics. *Food Chemistry*, 317 :126376.

Bons, J., Husson, G., Chion, M., Bonnet, M., Maumy-Bertrand, M., Delalande, F., Cianférani, S., Bertrand, F., Picard, B., and Carapito, C. (2021). Combining label-free and label-based accurate quantifications with SWATH-MS : Comparison with SRM and PRM for the evaluation of bovine muscle type effects. *PROTEOMICS*, 21(10) :2000214.

Brandi, J., Robotti, E., Manfredi, M., Barberis, E., Marengo, E., Novelli, E., and Cecconi, D. (2021). Kohonen Artificial Neural Network and Multivariate Analysis in the Identification of Proteome Changes during Early and Long Aging of Bovine *Longissimus dorsi* Muscle Using SWATH Mass Spectrometry. *Journal of Agricultural and Food Chemistry*, 69(38) :11512–11522.

Bräutigam, K., Soolanayakanahally, R., Champigny, M., Mansfield, S., Douglas, C., Campbell, M. M., and Cronk, Q. (2017). Sexual epigenetics : gender-specific methylation of a gene in the sex determining region of Populus balsamifera. *Scientific Reports*, 7(1) :45388.

Cai, Z., Poulos, R. C., Liu, J., and Zhong, Q. (2022). Machine learning for multi-omics data integration in cancer. *iScience*, 25(2) :103798.

Cao, L. (2018). Data Science : A Comprehensive Overview. *ACM Computing Surveys*, 50(3) :1–42.

Carty, D., Akehurst, C., Savage, R., Sungatullina, L., Robinson, S., McBride, M., McClure, J., Freeman, D., and Delles, C. (2014). Differential gene expression in obese pregnancy. *Pregnancy Hypertension : An International Journal of Women's Cardiovascular Health*, 4(3) :232–233.

Champigny, M. J., Unda, F., Skyba, O., Soolanayakanahally, R. Y., Mansfield, S. D., and Campbell, M. M. (2020). Learning from methylomes : epigenomic correlates of *Populus balsamifera* traits based on deep learning models of natural DNA methylation. *Plant Biotechnology Journal*, 18(6) :1361–1375.

Chen, B., Xu, H., Guo, Y., Grünhofer, P., Schreiber, L., Lin, J., and Li, R. (2021a). Transcriptomic and epigenomic remodeling occurs during vascular cambium periodicity in Populus tomentosa. *Horticulture Research*, 8(1) :102.

Chen, Q., Huang, L., Pan, D., Hu, K., Li, R., Friedline, R. H., Kim, J. K., Zhu, L. J., Guertin, D. A., and Wang, Y.-X. (2022). A brown fat-enriched adipokine Adissp controls adipose thermogenesis and glucose homeostasis. *Nature Communications*, 13(1) :7633.

Chen, Y., Tong, S., Jiang, Y., Ai, F., Feng, Y., Zhang, J., Gong, J., Qin, J., Zhang, Y., Zhu, Y., Liu, J., and Ma, T. (2021b). Transcriptional landscape of highly lignified poplar stems at single-cell resolution. *Genome Biology*, 22(1) :319.

Chu, J., Sun, N., Hu, W., Chen, X., Yi, N., and Shen, Y. (2022). The Application of Bayesian Methods in Cancer Prognosis and Prediction. *Cancer Genomics - Proteomics*, 19(1) :1–11.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184) :215–219.

Ding, Y., Zou, L.-H., Wu, J., Ramakrishnan, M., Gao, Y., Zhao, L., and Zhou, M. (2022). The pattern of DNA methylation alteration, and its association with the expression changes of non-coding RNAs and mRNAs in Moso bamboo under abiotic stress. *Plant Science*, 325 :111451.

Domb, K., Katz, A., Harris, K. D., Yaari, R., Kaisler, E., Nguyen, V. H., Hong, U. V. T., Griess, O., Heskiau, K. G., Ohad, N., and Zemach, A. (2020). DNA methylation mutants in *Physcomitrella patens* elucidate individual roles of CG and non-CG methylation in genome regulation. *Proceedings of the National Academy of Sciences*, 117(52) :33700–33710.

Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 26(4) :745–766.

Duruflé, H., Selmani, M., Ranocha, P., Jamet, E., Dunand, C., and Déjean, S. (2021). A powerful framework for an integrative study with heterogeneous omics data : from univariate statistics to multi-block analysis. *Briefings in Bioinformatics*, 22(3) :bbaa166.

Ehrlich, M. and Lacey, M. (2013). DNA methylation and differentiation : silencing, upregulation and modulation of gene expression. *Epigenomics*, 5(5) :553–568.

Eichten, S. R., Srivastava, A., Reddiex, A. J., Ganguly, D. R., Heussler, A., Streich, J. C., Wilson, P. B., and Borevitz, J. O. (2020). Extending the Genotype in *Brachypodium* by Including DNA Methylation Reveals a Joint Contribution with Genetics on Adaptive Traits. *G3 Genes|Genomes|Genetics*, 10(5) :1629–1637.

Evans, C., Hardin, J., and Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, 19(5) :776–792.

Fan, X., Peng, L., and Zhang, Y. (2022). Plant DNA Methylation Responds to Nutrient Stress. *Genes*, 13(6) :992.

Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., Hetzel, J., Jain, J., Strauss, S. H., Halpern, M. E., Ukomadu, C., Sadler, K. C., Pradhan, S., Pellegrini, M., and Jacobsen, S. E. (2010). Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, 107(19) :8689–8694.

Flores, J. E., Claborne, D. M., Weller, Z. D., Webb-Robertson, B.-J. M., Waters, K. M., and Bramer, L. M. (2023). Missing data in multi-omics integration : Recent advances through artificial intelligence. *Frontiers in Artificial Intelligence*, 6 :1098308.

Fox, H., Doron-Faigenboim, A., Kelly, G., Bourstein, R., Attia, Z., Zhou, J., Moshe, Y., Moshelion, M., and David-Schwartz, R. (2018). Transcriptome analysis of Pinus halepensis under drought stress and during recovery. *Tree Physiology*, 38(3) :423–441.

Franco, E. F., Rana, P., Cruz, A., Calderón, V. V., Azevedo, V., Ramos, R. T. J., and Ghosh, P. (2021). Performance Comparison of Deep Learning Autoencoders for Cancer Subtype Detection Using Multi-Omics Data. *Cancers*, 13(9) :2013.

Ghodasara, P., Satake, N., Sadowski, P., Kopp, S., and Mills, P. C. (2022). Investigation of cattle plasma proteome in response to pain and inflammation using next generation proteomics technique, SWATH-MS. *Molecular Omics*, 18(2) :133–142.

Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition : A New Concept for Consistent and Accurate Proteome Analysis. *Molecular and Cellular Proteomics*, 11(6) :O111.016717.

Gliozzo, J., Mesiti, M., Notaro, M., Petrini, A., Patak, A., Puertas-Gallardo, A., Paccanaro, A., Valentini, G., and Casiraghi, E. (2022). Heterogeneous data integration methods for patient similarity networks. *Briefings in Bioinformatics*, 23(4) :bbac207.

Gupta, S. and Gupta, A. (2019). Dealing with Noise Problem in Machine Learning Data-sets : A Systematic Review. *Procedia Computer Science*, 161 :466–474.

Han, Q., Song, H., Yang, C., Zhang, S., Korpelainen, H., and Li, C. (2022). Integrated DNA methylation, transcriptome and physiological analyses reveal new insights into superiority of poplars formed by interspecific grafting. *Tree Physiology*, 42(7) :1481–1500.

Han, S.-Y., Kim, W.-Y., Kim, J. S., and Hwang, I. (2023). Comparative transcriptomics reveals the role of altered energy metabolism in the establishment of single-cell C4 photosynthesis in Bienertia sinuspersici. *Frontiers in Plant Science*, 14 :1202521.

Hawe, J. S., Theis, F. J., and Heinig, M. (2019). Inferring Interaction Networks From Multi-Omics Data. *Frontiers in Genetics*, 10 :535.

He, W., Zhu, Y., Leng, Y., Yang, L., Zhang, B., Yang, J., Zhang, X., Lan, H., Tang, H., Chen, J., Gao, S., Tan, J., Kang, J., Deng, L., Li, Y., He, Y., Rong, T., and Cao, M. (2021). Transcriptomic Analysis Reveals Candidate Genes Responding Maize Gray Leaf Spot Caused by Cercospora zeina. *Plants*, 10(11) :2257.

He, X.-J., Chen, T., and Zhu, J.-K. (2011). Regulation and function of DNA methylation in plants and animals. *Cell Research*, 21(3) :442–465.

Hebel, T., Eisinger, K., Neumeier, M., Rein-Fischboeck, L., Pohl, R., Meier, E. M., Boettcher, A., Froehner, S. C., Adams, M. E., Liebisch, G., Krautbauer, S., and Buechler, C. (2015). Lipid abnormalities in alpha/beta2-syntrophin null mice are independent from ABCA1. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1851(5) :527–536.

Hesami, M., Alizadeh, M., Jones, A. M. P., and Torkamaneh, D. (2022). Machine learning : its challenges and opportunities in plant system biology. *Applied Microbiology and Biotechnology*.

Imadi, S. R., Kazi, A. G., Ahanger, M. A., Gucel, S., and Ahmad, P. (2015). Plant transcriptomics and responses to environmental stress : an overview. *Journal of Genetics*, 94(3) :525–537.

Jia, W., Sun, M., Lian, J., and Hou, S. (2022). Feature dimensionality reduction : a review. *Complex and Intelligent Systems*, 8(3) :2663–2693.

Joosten, S. C., Smits, K. M., Aarts, M. J., Melotte, V., Koch, A., Tjan-Heijnen, V. C., and Van Engeland, M. (2018). Epigenetics in renal cell cancer : mechanisms and clinical applications. *Nature Reviews Urology*, 15(7) :430–451.

Julca, I., Ferrari, C., Flores-Tornero, M., Proost, S., Lindner, A.-C., Hackenberg, D., Steinbachová, L., Michaelidis, C., Gomes Pereira, S., Misra, C. S., Kawashima, T., Borg, M., Berger, F., Goldberg, J., Johnson, M., Honys, D., Twell, D., Sprunck, S., Dresselhaus, T., Becker, J. D., and Mutwil, M. (2021). Comparative transcriptomic analysis reveals conserved programmes underpinning organogenesis and reproduction in land plants. *Nature Plants*, 7(8) :1143–1159.

Kakei, Y., Mochida, K., Sakurai, T., Yoshida, T., Shinozaki, K., and Shimada, Y. (2015). Transcriptome analysis of hormone-induced gene expression in Brachypodium distachyon. *Scientific Reports*, 5(1) :14476.

Kang, M., Ko, E., and Mersha, T. B. (2022). A roadmap for multi-omics data integration using deep learning. *Briefings in Bioinformatics*, 23(1) :bbab454.

Kang, W., Jiang, Z., Chen, Y., Wu, F., Liu, C., Wang, H., Shi, S., and Zhang, X.-X. (2020). Plant transcriptome analysis reveals specific molecular interactions between alfalfa and its rhizobial symbionts below the species level. *BMC Plant Biology*, 20(1) :293.

Kashima, Y., Sakamoto, Y., Kaneko, K., Seki, M., Suzuki, Y., and Suzuki, A. (2020). Single-cell sequencing techniques from individual to multiomics analyses. *Experimental and Molecular Medicine*, 52(9) :1419–1427.

Kim, M.-H., Cho, J.-S., Jeon, H.-W., Sangsawang, K., Shim, D., Choi, Y.-I., Park, E.-J., Lee, H., and Ko, J.-H. (2019). Wood Transcriptome Profiling Identifies Critical Pathway Genes of

Secondary Wall Biosynthesis and Novel Regulators for Vascular Cambium Development in Populus. *Genes*, 10(9) :690.

Kon, T. and Yoshikawa, N. (2014). Induction and maintenance of DNA methylation in plant promoter sequences by apple latent spherical virus-induced transcriptional gene silencing. *Frontiers in Microbiology*, 5.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2) :233–243.

Lai, Y.-S., Zhang, X., Zhang, W., Shen, D., Wang, H., Xia, Y., Qiu, Y., Song, J., Wang, C., and Li, X. (2017). The association of changes in DNA methylation with temperature-dependent sex determination in cucumber. *Journal of Experimental Botany*, 68(11) :2899–2912.

Lang, Z., Wang, Y., Tang, K., Tang, D., Datsenka, T., Cheng, J., Zhang, Y., Handa, A. K., and Zhu, J.-K. (2017). Critical roles of DNA demethylation in the activation of ripening-induced genes and inhibition of ripening-repressed genes in tomato fruit. *Proceedings of the National Academy of Sciences*, 114(22).

Lemsara, A., Ouadfel, S., and Fröhlich, H. (2020). PathME : pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. *BMC Bioinformatics*, 21(1) :146.

Li, B. and Dewey, C. N. (2011). RSEM : accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1) :323.

Li, C., Gao, Z., Su, B., Xu, G., and Lin, X. (2021a). Data analysis methods for defining biomarkers from omics data. *Analytical and Bioanalytical Chemistry*.

Li, H., Yue, H., Xie, J., Bu, J., Li, L., Xin, X., Zhao, Y., Zhang, H., Yang, L., Wang, J., and Jiang, X. (2021b). Transcriptomic profiling of the high-vigour maize (Zea mays L.) hybrid variety response to cold and drought stresses during seed germination. *Scientific Reports*, 11(1) :19345.

Li, P., Cao, W., Fang, H., Xu, S., Yin, S., Zhang, Y., Lin, D., Wang, J., Chen, Y., Xu, C., and Yang, Z. (2017). Transcriptomic Profiling of the Maize (Zea mays L.) Leaf Response to Abiotic Stresses at the Seedling Stage. *Frontiers in Plant Science*, 8.

Li, Q., Song, J., West, P. T., Zynda, G., Eichten, S. R., Vaughn, M. W., and Springer, N. M. (2015). Examining the Causes and Consequences of Context-Specific Differential DNA Methylation in Maize. *Plant Physiology*, 168(4) :1262–1274.

Li, S. and Tollefsbol, T. O. (2021). DNA methylation methods : Global DNA methylation and methylomic analyses. *Methods*, 187 :28–43.

Li, S.-F., Lv, C.-C., Lan, L.-N., Jiang, K.-L., Zhang, Y.-L., Li, N., Deng, C.-L., and Gao, W.-J. (2021c). DNA methylation is involved in sexual differentiation and sex chromosome evolution in the dioecious plant garden asparagus. *Horticulture Research*, 8 :198.

Li, X., Zhu, J., Hu, F., Ge, S., Ye, M., Xiang, H., Zhang, G., Zheng, X., Zhang, H., Zhang, S., Li, Q., Luo, R., Yu, C., Yu, J., Sun, J., Zou, X., Cao, X., Xie, X., Wang, J., and Wang, W. (2012). Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics*, 13(1) :300.

Li, Y., Wang, D., Ping, X., Zhang, Y., Zhang, T., Wang, L., Jin, L., Zhao, W., Guo, M., Shen, F., Meng, M., Chen, X., Zheng, Y., Wang, J., Li, D., Zhang, Q., Hu, C., Xu, L., and Ma, X. (2022). Local hyperthermia therapy induces browning of white fat and treats obesity. *Cell*, 185(6) :949–966.e19.

Liang, L., Chang, Y., Lu, J., Wu, X., Liu, Q., Zhang, W., Su, X., and Zhang, B. (2019). Global Methylomic and Transcriptomic Analyses Reveal the Broad Participation of DNA Methylation in Daily Gene Expression Regulation of Populus trichocarpa. *Frontiers in Plant Science*, 10 :243.

Liu, W.-X., Zhang, F.-C., Zhang, W.-Z., Song, L.-F., Wu, W.-H., and Chen, Y.-F. (2013). Arabidopsis Di19 Functions as a Transcription Factor and Modulates PR1, PR2, and PR5 Expression in Response to Drought Stress. *Molecular Plant*, 6(5) :1487–1502.

Liu, Y., Wang, J., Liu, B., and Xu, Z. (2022). Dynamic regulation of DNA methylation and histone modifications in response to abiotic stresses in plants. *Journal of Integrative Plant Biology*, 64(12) :2252–2274.

Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, 29(20) :2610–2616.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1).

Lovino, M., Randazzo, V., Ciravegna, G., Barbiero, P., Ficarra, E., and Cirrincione, G. (2021). A survey on data integration for multi-omics sample clustering. *Neurocomputing*, page S0925231221018063.

Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., and Aebersold, R. (2018). Data-independent acquisition-based SWATH-MS for quantitative proteomics : a tutorial. *Molecular Systems Biology*, 14(8) :e8126.

Luo, C., Zhao, S., Dai, W., Zheng, N., and Wang, J. (2018). Proteomic Analysis of Lysosomal Membrane Proteins in Bovine Mammary Epithelial Cells Illuminates Potential Novel Lysosome Functions in Lactation. *Journal of Agricultural and Food Chemistry*, 66(49) :13041–13049.

Luo, J., Nvsvrot, T., and Wang, N. (2021). Comparative transcriptomic analysis uncovers conserved pathways involved in adventitious root formation in poplar. *Physiology and Molecular Biology of Plants*, 27(9) :1903–1918.

Luo, W., Zhang, C., Zhang, J., Jiang, D., Guo, W., and Wan, D. (2017). Transcriptome analysis of four poplars exposed to continuous salinity stress. *Biochemical Systematics and Ecology*, 70 :311–319.

Lupat, R., Perera, R., Loi, S., and Li, J. (2023). Moanna : Multi-Omics Autoencoder-Based Neural Network Algorithm for Predicting Breast Cancer Subtypes. *IEEE Access*, 11 :10912–10924.

Lê Cao, K.-A. and Welham, Z. M. (2021). *Multivariate Data Integration Using R : Methods and Applications with the mixOmics Package*. Chapman and Hall/CRC, Boca Raton, 1 edition.

López, M.-E., Roquis, D., Becker, C., Denoyes, B., and Bucher, E. (2022). DNA methylation dynamics during stress response in woodland strawberry (*Fragaria vesca*). *Horticulture Research*, 9 :uhac174.

López-Pedrouso, M., Lorenzo, J. M., Di Stasio, L., Brugiapaglia, A., and Franco, D. (2021). Quantitative proteomic analysis of beef tenderness of Piemontese young bulls by SWATH-MS. *Food Chemistry*, 356 :129711.

Ma, M., Chen, X., Yin, Y., Fan, R., Li, B., Zhan, Y., and Zeng, F. (2020). DNA Methylation Silences Exogenous Gene Expression in Transgenic Birch Progeny. *Frontiers in Plant Science*, 11 :523748.

Ma, T. and Zhang, A. (2019). Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). *BMC Genomics*, 20(S11) :944.

Madhumita and Paul, S. (2022). Capturing the latent space of an Autoencoder for multi-omics integration and cancer subtyping. *Computers in Biology and Medicine*, 148 :105832.

Manríquez, J. G. (2020). *Univariate and Multivariate Statistical Approaches for the Analyses of Omics Data : Sample Classification and Two-block Integration*. PhD thesis, Imperial College London.

Mantione, K. J., Kream, R. M., Kuzelova, H., Ptacek, R., Raboch, J., Samuel, J. M., and Stefano, G. B. (2014). Comparing Bioinformatic Gene Expression Profiling Methods : Microarray and RNA-Seq. *Medical Science Monitor Basic Research*, 20 :138–142.

Mardoc, E., Sow, M. D., Déjean, S., and Salse, J. (2024). Genomic data integration tutorial, a plant case study. *BMC Genomics*, 25(1) :66.

Mariette, J. and Villa-Vialaneix, N. (2018). Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, 34(6) :1009–1015.

Mattei, A. L., Bailly, N., and Meissner, A. (2022). DNA methylation : a historical perspective. *Trends in Genetics*, 38(7) :676–707.

Meng, C., Helm, D., Frejno, M., and Kuster, B. (2016a). moCluster : Identifying Joint Patterns Across Multiple Omics Data Sets. *Journal of Proteome Research*, 15(3) :755–765.

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C. (2016b). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4) :628–641.

Milliken, G. A. and Johnson, D. E. (2001). *Analysis of Messy Data, Volume III : Analysis of Covariance*. Chapman and Hall/CRC, 0 edition.

Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K. S., and Hilsenbeck, S. G. (2018). A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, 19(1) :71–86.

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., and Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11) :4245–4250.

Moore, L. D., Le, T., and Fan, G. (2013). DNA Methylation and Its Basic Function. *Neuropsychopharmacology*, 38(1) :23–38.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7) :621–628.

Newell-Price, J., Clark, A. J., and King, P. (2000). DNA Methylation and Silencing of Gene Expression. *Trends in Endocrinology and Metabolism*, 11(4) :142–148.

Nguyen, H., Shrestha, S., and Nguyen, T. (2018). PINSPlus : Clustering Algorithm for Data Integration and Disease Subtyping. *ResearchGate*.

Nguyen, T., Tagett, R., Diaz, D., and Draghici, S. (2017). A novel approach for data integration and disease subtyping. *Genome Research*, 27(12) :2025–2039.

Niederhuth, C. E. and Schmitz, R. J. (2017). Putting DNA methylation in context : from genomes to gene expression in plants. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1860(1) :149–156.

Noble, W. S. (2009). How does multiple testing correction work ? *Nature Biotechnology*, 27(12) :1135–1137.

Noor, E., Cherkaoui, S., and Sauer, U. (2019). Biological insights through omics data integration. *Current Opinion in Systems Biology*, 15 :39–47.

Noshay, J. M., Anderson, S. N., Zhou, P., Ji, L., Ricci, W., Lu, Z., Stitzer, M. C., Crisp, P. A., Hirsch, C. N., Zhang, X., Schmitz, R. J., and Springer, N. M. (2019). Monitoring the interplay between transposable element families and DNA methylation in maize. *PLOS Genetics*, 15(9) :e1008291.

Noshay, J. M., Liang, Z., Zhou, P., Crisp, P. A., Marand, A. P., Hirsch, C. N., Schmitz, R. J., and Springer, N. M. (2021). Stability of DNA methylation and chromatin accessibility in structurally diverse maize genomes. *G3 Genes|Genomes|Genetics*, 11(8) :jkab190.

Nuo, M., Yuan, J., Yang, W., Gao, X., He, N., Liang, H., Cang, M., and Liu, D. (2016). Promoter methylation and histone modifications affect the expression of the exogenous DsRed gene in transgenic goats. *Genetics and Molecular Research*, 15(3).

Ogasahara, T., Kouzai, Y., Watanabe, M., Takahashi, A., Takahagi, K., Kim, J.-S., Matsui, H., Yamamoto, M., Toyoda, K., Ichinose, Y., Mochida, K., and Noutoshi, Y. (2022). Time-series transcriptome of Brachypodium distachyon during bacterial flagellin-induced pattern-triggered immunity. *Frontiers in Plant Science*, 13 :1004184.

Parrilla-Doblas, J. T., Roldán-Arjona, T., Ariza, R. R., and Córdoba-Cañero, D. (2019). Active DNA Demethylation in Plants. *International Journal of Molecular Sciences*, 20(19) :4683.

Patruno, L., Maspero, D., Craighero, F., Angaroni, F., Antoniotti, M., and Graudenzi, A. (2020). A review of computational strategies for denoising and imputation of single-cell transcriptomic data. *Briefings in Bioinformatics*, page bbaa222.

Picard, M., Scott-Boyer, M.-P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Computational and Structural Biotechnology Journal*, 19 :3735–3746.

Pingault, L., Varsani, S., Palmer, N., Ray, S., Williams, W. P., Luthe, D. S., Ali, J. G., Sarath, G., and Louis, J. (2021). Transcriptomic and volatile signatures associated with maize defense against corn leaf aphid. *BMC Plant Biology*, 21(1) :138.

Quirós, P. M., Ramsay, A. J., and López-Otín, C. (2013). New roles for OMA1 metalloprotease : From mitochondrial proteostasis to metabolic homeostasis. *Adipocyte*, 2(1) :7–11.

Rajasundaram, D. and Selbig, J. (2016). More effort — more results : recent advances in integrative 'omics' data analysis. *Current Opinion in Plant Biology*, 30 :57–61.

Rao, X., Ren, J., Wang, W., Chen, R., Xie, Q., Xu, Y., Li, D., Song, Z., He, Y., Cai, D., Yang, P., Lyu, S., Li, L., Liu, W., and Zhang, X. (2023). Comparative DNA-methylome and transcriptome analysis reveals heterosis- and polyploidy-associated epigenetic changes in rice. *The Crop Journal*, 11(2) :427–437.

Rappoport, N. and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms : review and cancer benchmark. *Nucleic Acids Research*, 46(20) :10546–10562.

Rappoport, N. and Shamir, R. (2019). NEMO : cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, 35(18) :3348–3356.

Razin, A. and Cedar, H. (1991). DNA methylation and gene expression. *Microbiological Reviews*, 55(3) :451–458.

Rich-Griffin, C., Stechemesser, A., Finch, J., Lucas, E., Ott, S., and Schäfer, P. (2020). Single-Cell Transcriptomics : A High-Resolution Avenue for Plant Functional Genomics. *Trends in Plant Science*, 25(2) :186–197.

Rifai, N., Gillette, M. A., and Carr, S. A. (2006). Protein biomarker discovery and validation : the long and uncertain path to clinical utility. *Nature Biotechnology*, 24(8) :971–983.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015a). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2) :85–97.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015b). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7) :e47–e47.

Rivera-Contreras, I. K., Zamora-Hernández, T., Huerta-Heredia, A. A., Capataz-Tafur, J., Barrera-Figueroa, B. E., Juntawong, P., and Peña-Castro, J. M. (2016). Transcriptomic analysis of submergence-tolerant and sensitive Brachypodium distachyon ecotypes reveals oxidative stress as a major tolerance factor. *Scientific Reports*, 6(1) :27686.

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3) :R25.

Roessler, K., Takuno, S., and Gaut, B. S. (2016). CG Methylation Covaries with Differential Gene Expression between Leaf and Floral Bud Tissues of Brachypodium distachyon. *PLOS ONE*, 11(3) :e0150002.

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017). mixOmics : An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11) :e1005752.

Ryu, K. H., Huang, L., Kang, H. M., and Schiefelbein, J. (2019). Single-Cell RNA Sequencing Resolves Molecular Relationships Among Individual Plant Cells. *Plant Physiology*, 179(4) :1444–1456.

Samantara, K., Shiv, A., De Sousa, L. L., Sandhu, K. S., Priyadarshini, P., and Mohapatra, S. R. (2021). A comprehensive review on epigenetic mechanisms and application of epigenetic modifications for crop improvement. *Environmental and Experimental Botany*, 188 :104479.

Sathyanarayanan, A., Gupta, R., Thompson, E. W., Nyholt, D. R., Bauer, D. C., and Nagaraj, S. H. (2020). A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. *Briefings in Bioinformatics*, 21(6) :1920–1936.

Satya, P., Bhattacharjee, S., Sarkar, D., Roy, S., Sharma, L., and Mandal, N. A. (2022). Transcriptomics in Plant. In Singh, R. L., Mondal, S., Parihar, A., and Singh, P. K., editors, *Plant Genomics for Sustainable Agriculture*, pages 99–127. Springer Nature Singapore, Singapore.

Schneider, M. V. and Orchard, S. (2011). Omics Technologies, Data and Bioinformatics Principles. In Mayer, B., editor, *Bioinformatics for Omics Data*, volume 719, pages 3–30. Humana Press, Totowa, NJ. Series Title : Methods in Molecular Biology.

Schönberger, B., Chen, X., Mager, S., and Ludewig, U. (2016). Site-Dependent Differences in DNA Methylation and Their Impact on Plant Establishment and Phosphorus Nutrition in Populus trichocarpa. *PLOS ONE*, 11(12) :e0168623.

Seymour, D. K. and Gaut, B. S. (2020). Phylogenetic Shifts in Gene Body Methylation Correlate with Gene Expression and Reflect Trait Conservation. *Molecular Biology and Evolution*, 37(1) :31–43.

Shaikh, A. A., Chachar, S., Chachar, M., Ahmed, N., Guan, C., and Zhang, P. (2022). Recent Advances in DNA Methylation and Their Potential Breeding Applications in Plants. *Horticulturae*, 8(7) :562.

Shaw, R., Tian, X., and Xu, J. (2021). Single-Cell Transcriptome Analysis in Plants : Advances and Challenges. *Molecular Plant*, 14(1) :115–126.

Shen, C., Li, D., He, R., Fang, Z., Xia, Y., Gao, J., Shen, H., and Cao, M. (2014). Comparative transcriptome analysis of RNA-seq data for cold-tolerant and cold-sensitive rice genotypes under cold stress. *Journal of Plant Biology*, 57(6) :337–348.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22) :2906–2912.

Shi, Y., Zhang, X., Chang, X., Yan, M., Zhao, H., Qin, Y., and Wang, H. (2021). Integrated analysis of DNA methylome and transcriptome reveals epigenetic regulation of CAM photosynthesis in pineapple. *BMC Plant Biology*, 21(1) :19.

Shulse, C. N., Cole, B. J., Ciobanu, D., Lin, J., Yoshinaga, Y., Gouran, M., Turco, G. M., Zhu, Y., O'Malley, R. C., Brady, S. M., and Dickel, D. E. (2019). High-Throughput Single-Cell Transcriptome Profiling of Plant Cell Types. *Cell Reports*, 27(7) :2241–2247.e4.

Singer, B. D. (2019). A Practical Guide to the Measurement and Analysis of DNA Methylation. *American Journal of Respiratory Cell and Molecular Biology*, 61(4) :417–428.

Skalska, A., Stritt, C., Wyler, M., Williams, H. W., Vickers, M., Han, J., Tuna, M., Savas Tuna, G., Susek, K., Swain, M., Wóycicki, R. K., Chaudhary, S., Corke, F., Doonan, J. H., Roulin, A. C., Hasterok, R., and Mur, L. A. J. (2020). Genetic and Methylome Variation in Turkish Brachypodium Distachyon Accessions Differentiate Two Geographically Distinct Subpopulations. *International Journal of Molecular Sciences*, 21(18) :6700.

Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, 122(1).

Sow, M. D., Le Gac, A., Fichot, R., Lanciano, S., Delaunay, A., Le Jan, I., Lesage-Descauses, M., Citerne, S., Caius, J., Brunaud, V., Soubigou-Taconnat, L., Cochard, H., Segura, V., Chaparro, C., Grunau, C., Daviaud, C., Tost, J., Brignolas, F., Strauss, S. H., Mirouze, M., and Maury, S. (2021). RNAi suppression of DNA methylation affects the drought stress response and genome integrity in transgenic poplar. *New Phytologist*, 232(1) :80–97.

Sow, M. D., Rogier, O., Lesur, I., Daviaud, C., Mardoc, E., Sanou, E., Duvaux, L., Civan, P., Delaunay, A., Descauses, M.-C. L., Benoit, V., Le-Jan, I., Buret, C., Besse, C., Durufle, H., Fichot, R., Le-Provost, G., Guichoux, E., Boury, C., Garnier, A., Senhaji-Rachik, A., Jorge, V., Ambroise, C., Tost, J., Plomion, C., Segura, V., Maury, S., and Salse, J. (2023). Epigenetic Variation in Tree Evolution : a case study in black poplar (*Populus nigra*). preprint, Evolutionary Biology.

Stanojevic, S., Li, Y., and Garmire, L. X. (2022). Computational Methods for Single-Cell Multi-Omics Integration and Alignment. *arXiv :2201.06725 [q-bio]*. arXiv : 2201.06725.

Su, Y., Bai, X., Yang, W., Wang, W., Chen, Z., Ma, J., and Ma, T. (2018). Single-base-resolution methylomes of Populus euphratica reveal the association between DNA methylation and salt stress. *Tree Genetics and Genomes*, 14(6) :86.

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, 14 :117793221989905.

Sun, G., Yu, H., Wang, P., Lopez-Guerrero, M., Mural, R. V., Mizero, O. N., Grzybowski, M., Song, B., van Dijk, K., Schachtman, D. P., Zhang, C., and Schnable, J. C. (2023a). A role for heritable transcriptomic variation in maize adaptation to temperate environments. *Genome Biology*, 24(1) :55.

Sun, L., Yang, M., Su, W., Xu, H., Xue, F., Lu, C., and Wu, R. (2023b). Transcriptomic analysis of maize uncovers putative genes involved in metabolic detoxification under four safeners treatment. *Pesticide Biochemistry and Physiology*, 194 :105465.

Surinova, S., Schiess, R., Hüttenhain, R., Cerciello, F., Wollscheid, B., and Aebersold, R. (2011). On the Development of Plasma Protein Biomarkers. *Journal of Proteome Research*, 10(1) :5–16.

Tahir, M. S., Nguyen, L. T., Schulz, B. L., Boe-Hansen, G. A., Thomas, M. G., Moore, S. S., Lau, L. Y., and Fortes, M. R. S. (2019). Proteomics Recapitulates Ovarian Proteins Relevant to Puberty and Fertility in Brahman Heifers (Bos indicus L.). *Genes*, 10(11) :923.

Takuno, S., Ran, J.-H., and Gaut, B. S. (2016). Evolutionary patterns of genic DNA methylation vary across land plants. *Nature Plants*, 2(2) :15222.

Tan, K., Huang, W., Hu, J., and Dong, S. (2020). A multi-omics supervised autoencoder for pan-cancer clinical outcome endpoints prediction. *BMC Medical Informatics and Decision Making*, 20(S3) :129.

Taverna, F., Goveia, J., Karakach, T. K., Khan, S., Rohlenova, K., Treps, L., Subramanian, A., Schoonjans, L., Dewerchin, M., Eelen, G., and Carmeliet, P. (2020). BIOMEX : an interactive workflow for (single cell) omics data interpretation and visualization. *Nucleic Acids Research*, 48(W1) :W385–W394.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized Generalized Canonical Correlation Analysis. *Psychometrika*, 76(2) :257–284.

Tenenhaus, M. (1998). *La régression PLS : théorie et pratique*. Editions Technip, Paris.

Teshome, D. T., Zharare, G. E., Ployet, R., and Naidoo, S. (2023). Transcriptional reprogramming during recovery from drought stress in *Eucalyptus grandis*. *Tree Physiology*, 43(6) :979–994.

Tian, P., Lin, Z., Lin, D., Dong, S., Huang, J., and Huang, T. (2021). The pattern of DNA methylation alteration, and its association with the changes of gene expression and alternative splicing during phosphate starvation in tomato. *The Plant Journal*, 108(3) :841–858.

Tien, C.-W., Huang, T.-Y., Chen, P.-C., and Wang, J.-H. (2021). Using Autoencoders for Anomaly Detection and Transfer Learning in IoT. *Computers*, 10(7) :88.

Tirnaz, S., Miyaji, N., Takuno, S., Bayer, P. E., Shimizu, M., Akter, M. A., Edwards, D., Batley, J., and Fujimoto, R. (2022). Whole-Genome DNA Methylation Analysis in Brassica rapa subsp. perviridis in Response to Albugo candida Infection. *Frontiers in Plant Science*, 13 :849358.

Vahabi, N. and Michailidis, G. (2022). Unsupervised Multi-Omics Data Integration Methods : A Comprehensive Review. *Frontiers in Genetics*, 13 :854752.

Valent, D., Yeste, N., Hernández-Castellano, L. E., Arroyo, L., Wu, W., García-Contreras, C., Vázquez-Gómez, M., González-Bulnes, A., Bendixen, E., and Bassols, A. (2019). SWATH-MS quantitative proteomic investigation of intrauterine growth restriction in a porcine model reveals sex differences in hippocampus development. *Journal of Proteomics*, 204 :103391.

Van De Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., and Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1) :1.

Varriale, A. (2017). DNA Methylation in Plants and Its Implications in Development, Hybrid Vigour, and Evolution. In Rajewsky, N., Jurga, S., and Barciszewski, J., editors, *Plant Epigenetics*, pages 263–280. Springer International Publishing, Cham. Series Title : RNA Technologies.

Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-Generation Sequencing : From Basic Research to Diagnostics. *Clinical Chemistry*, 55(4) :641–658.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3) :333–337.

Wang, C. and Kurgan, L. (2019). Review and comparative assessment of similarity-based methods for prediction of drug–protein interactions in the druggable human proteome. *Briefings in Bioinformatics*, 20(6) :2066–2087.

Wang, C., Zhao, R., Zhao, Z., Liu, N., Cheng, J., and Guo, M. (2023). Proteomic characterization and comparison of milk fat globule membrane proteins of Saanen goat milk from 3 habitats in China using SWATH-MS technique. *Journal of Dairy Science*, 106(4) :2289–2302.

Wang, G., Li, H., Meng, S., Yang, J., Ye, N., and Zhang, J. (2020a). Analysis of Global Methylome and Gene Expression during Carbon Reserve Mobilization in Stems under Soil Drying. *Plant Physiology*, 183(4) :1809–1824.

Wang, X., Li, N., Li, W., Gao, X., Cha, M., Qin, L., and Liu, L. (2020b). Advances in Transcriptomics in the Response to Stress in Plants. *Global Medical Genetics*, 07(02) :030–034.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq : a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1) :57–63.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 'T Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., Van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1) :160018.

Wu, C., Lin, M., Chen, F., Chen, J., Liu, S., Yan, H., and Xiang, Y. (2022). Homologous Drought-Induced 19 Proteins, PtDi19-2 and PtDi19-7, Enhance Drought Tolerance in Transgenic Plants. *International Journal of Molecular Sciences*, 23(6) :3371.

Wu, D., Wang, D., Zhang, M. Q., and Gu, J. (2015). Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation : application to cancer molecular classification. *BMC Genomics*, 16(1) :1022.

Wyler, M., Stritt, C., Walser, J.-C., Baroux, C., and Roulin, A. C. (2020). Impact of Transposable Elements on Methylation and Gene Expression across Natural Accessions of *Brachypodium distachyon*. *Genome Biology and Evolution*, 12(11) :1994–2001.

Xiao, D., Zhou, K., Yang, X., Yang, Y., Ma, Y., and Wang, Y. (2021). Crosstalk of DNA Methylation Triggered by Pathogen in Poplars With Different Resistances. *Frontiers in Microbiology*, 12 :750089.

Xiao, Z., Zhang, Y., Liu, M., Zhan, C., Yang, X., Nvsvrot, T., Yan, Z., and Wang, N. (2020). Coexpression analysis of a large-scale transcriptome identified a calmodulin-like protein regulating the development of adventitious roots in poplar. *Tree Physiology*, 40(10) :1405–1419.

Xu, G., Lyu, J., Li, Q., Liu, H., Wang, D., Zhang, M., Springer, N. M., Ross-Ibarra, J., and Yang, J. (2020). Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. *Nature Communications*, 11(1) :5539.

Xu, Q., Wu, L., Luo, Z., Zhang, M., Lai, J., Li, L., Springer, N. M., and Li, Q. (2022). DNA demethylation affects imprinted gene expression in maize endosperm. *Genome Biology*, 23(1) :77.

Xue, Y., Zou, C., Zhang, C., Yu, H., Chen, B., and Wang, H. (2022). Dynamic DNA methylation changes reveal tissue-specific gene expression in sugarcane. *Frontiers in Plant Science*, 13 :1036764.

Yong-Villalobos, L., González-Morales, S. I., Wrobel, K., Gutiérrez-Alanis, D., Cervantes-Peréz, S. A., Hayano-Kanashiro, C., Oropeza-Aburto, A., Cruz-Ramírez, A., Martínez, O., and Herrera-Estrella, L. (2015). Methylome analysis reveals an important role for epigenetic changes in the regulation of the *Arabidopsis* response to phosphate starvation. *Proceedings of the National Academy of Sciences*, 112(52).

Yu, D., Janz, D., Zienkiewicz, K., Herrfurth, C., Feussner, I., Chen, S., and Polle, A. (2021). Wood Formation under Severe Drought Invokes Adjustment of the Hormonal and Transcriptional Landscape in Poplar. *International Journal of Molecular Sciences*, 22(18) :9899.

Yu, Y., Meng, N., Chen, S., Zhang, H., Liu, Z., Wang, Y., Jing, Y., Wang, Y., and Chen, S. (2022). Transcriptomic profiles of poplar (Populus simonii × P. nigra) cuttings during adventitious root formation. *Frontiers in Genetics*, 13 :968544.

Yuan, G., He, X., Li, H., Xiang, K., Liu, L., Zou, C., Lin, H., Wu, J., Zhang, Z., and Pan, G. (2020). Transcriptomic responses in resistant and susceptible maize infected with Fusarium graminearum. *The Crop Journal*, 8(1) :153–163.

Zeng, W.-Y., Tan, Y.-R., Long, S.-F., Sun, Z.-D., Lai, Z.-G., Yang, S.-Z., Chen, H.-Z., and Qing, X.-Y. (2021). Methylome and transcriptome analyses of soybean response to bean pyralid larvae. *BMC Genomics*, 22(1) :836.

Zhang, Peng, Zhou, Ji, and Wang (2019). An Improved Autoencoder and Partial Least Squares Regression-Based Extreme Learning Machine Model for Pump Turbine Characteristics. *Applied Sciences*, 9(19) :3987.

Zhang, H., Lang, Z., and Zhu, J.-K. (2018). Dynamics and function of DNA methylation in plants. *Nature Reviews Molecular Cell Biology*, 19(8) :489–506.

Zhang, T.-Q., Xu, Z.-G., Shang, G.-D., and Wang, J.-W. (2019a). A Single-Cell RNA Sequencing Profiles the Developmental Landscape of Arabidopsis Root. *Molecular Plant*, 12(5) :648–660.

Zhang, X., Zhang, J., Sun, K., Yang, X., Dai, C., and Guo, Y. (2019b). Integrated Multi-omics Analysis Using Variational Autoencoders : Application to Pan-cancer Classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 765–769, San Diego, CA, USA. IEEE.

Zhang, Y., Liu, C., Cheng, H., Tian, S., Liu, Y., Wang, S., Zhang, H., Saqib, M., Wei, H., and Wei, Z. (2020). DNA methylation and its effects on gene expression during primary to secondary growth in poplar stems. *BMC Genomics*, 21(1) :498.

Zhang, Z., Liu, J., Sun, Y., Wang, S., Xing, X., Feng, X., Pérez-Pérez, J. M., and Li, Y. (2021). Genome-wide high-resolution mapping of DNA methylation reveals epigenetic variation in the offspring of sexual and asexual propagation in Robinia pseudoacacia. *Plant Cell Reports*, 40(12) :2435–2447.

Zhao, P., Ma, B., Cai, C., and Xu, J. (2022). Transcriptome and methylome changes in two contrasting mungbean genotypes in response to drought stress. *BMC Genomics*, 23(1) :80.

Zhao, P., Wang, L., and Yin, H. (2018). Transcriptional responses to phosphate starvation in Brachypodium distachyon roots. *Plant Physiology and Biochemistry*, 122 :113–120.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE*, 9(1) :e78644.

Zhao, Y., Li, M.-C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H., and McShane, L. M. (2021). TPM, FPKM, or Normalized Counts ? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine*, 19(1) :269.

Zhou, J., Xiao, L., Huang, R., Song, F., Li, L., Li, P., Fang, Y., Lu, W., Lv, C., Quan, M., Zhang, D., and Du, Q. (2023). Local diversity of drought resistance and resilience in *Populus tomentosa* correlates with the variation of DNA methylation. *Plant, Cell and Environment*, 46(2) :479–497.

Zhu, G., Gao, C., Wu, C., Li, M., Xu, J.-R., Liu, H., and Wang, Q. (2021). Comparative transcriptome analysis reveals distinct gene expression profiles in Brachypodium distachyon infected by two fungal pathogens. *BMC Plant Biology*, 21(1) :304.

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine : Principles, practice, and opportunities. *Information Fusion*, 50 :71–91.

Appendices

Annexe A

Mardoc et al. (2024), Données supplémentaires concernant les analyses peuplier

Les Jeux de Données supplémentaires sont disponibles au lien suivant :
`https://entrepot.recherche.data.gouv.fr/privateurl.xhtml?token=d946bb29-4698-4bee-9c6b-c2d98558ca8a`

Le code est disponible au lien suivant :
`https://forgemia.inra.fr/umr-gdec/omics-integration-on-poplar`

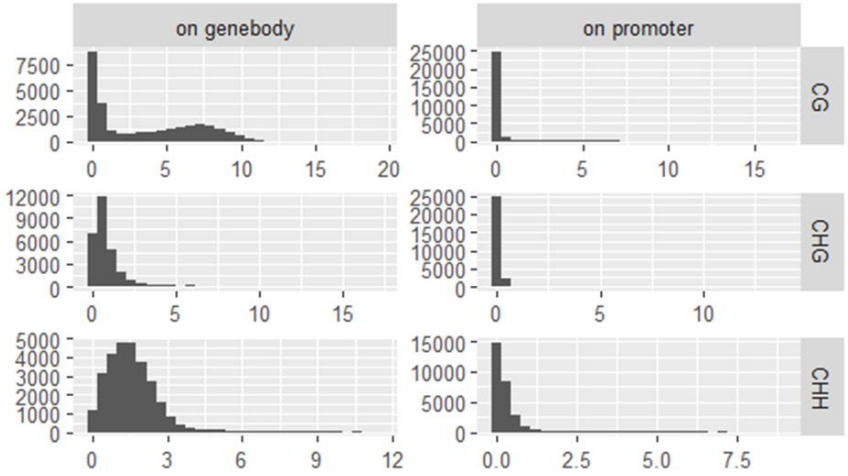
Supplementary Table 1: 39 R tools for multi-omics data integration.
Table containing R tools with their sources and review articles citing them. The 8 review articles are Sathyanarayanan et al., 2020 (https://doi.org/10.1093/bib/bbz121); Lovino et al., 2021 (https://doi.org/10.1016/j.neucom.2021.11.094); Cai et al., 2022 (https://doi.org/10.1016/j.isci.2022.103798); Hesami et al., 2022 (https://doi.org/10.1007/s00253-022-11963-6); Bersanelli et al., 2016 (https://doi.org/10.1186/s12859-015-0857-9); Rappoport and Shamir, 2018 (https://doi.org/10.1093/nar/gky889); Subramanian et al., 2020 (https://doi.org/10.1177/1177932219899051); Vahabi and Michailidis, 2022 (https://doi.org/10.3389/fgene.2022.854752)

tool's name	tool's complete name	DOI	source code	Sathyanarayanan	Lovino	Cai	Hesami	Bersanelli	Rappoport	Subramanian	Vahabi
ab-SNF	association-signal annotation boosted SNF	https://doi.org/10.1093/bioinformatics/btz124	https://github.com/pfruan/abSNF/								X
ANF	Affinity Network Fusion	https://doi.org/10.1109/BIBM.2017.8217682	https://rdrr.io/bioc/ANF/								X
BCC	Bayesian Consensus Clustering	https://doi.org/10.1093/bioinformatics/btt425	https://github.com/ttriche/bayesCC	X			X		X	X	X
Clusternomics		https://doi.org/10.1371/journal.pcbi.1005781	https://github.com/evelinag/clusternomics						X		X
COCA	Cluster-Of-Cluster Assignment	https://doi.org/10.1016/j.cell.2014.06.049	https://github.com/acabassi/coca						X		X
CNAmet		https://doi.org/10.1093/bioinformatics/btr019	https://csbi.ltdk.helsinki.fi/CNAmet/	X				X		X	X
HNMDRP	Heterogeneous Network-based Method for Drug Response Prediction	https://doi.org/10.1038/s41598-018-21622-4	https://github.com/USTC-HIlab/HNMDRP			X					
iCluster		https://doi.org/10.1093/bioinformatics/btp543	https://rdrr.io/cran/iCluster/								
(iClusterPlus)	integrative Clustering	https://doi.org/10.1073/pnas.1208949110	https://rdrr.io/bioc/iClusterPlus/	X	X	X	X	X	X	X	X
(iClusterBayes)		https://doi.org/10.1093/biostatistics/kxx017	https://rdrr.io/bioc/iClusterPlus/man/iClusterBayes.html								
iDRW	integrative Directed Random Walk	https://doi.org/10.1186/s12920-018-0389-z	https://github.com/sykim122/iDRW								X
iGC	integrated analysis of GE and CNA	https://doi.org/10.1186/s12859-016-1438-2	https://rdrr.io/bioc/iGC/	X							
INF	Integrative Network Fusion	https://doi.org/10.3389/fonc.2020.01065	https://gitlab.fbk.eu/ MPBA/INF								X
integrOmics		https://doi.org/10.1093/bioinformatics/btp515	Not available anymore					X			
intNMF	Integrative Non-negative Matrix Factorization	https://doi.org/10.1371/journal.pone.0176278	https://rdrr.io/cran/IntNMF/		X						
IS-Kmeans	Integrative Sparse k-means	https://doi.org/10.1214/17-AOAS1033	https://github.com/Caleb-Huo/IS-Kmeans								X
JIVE	Joint and Individual Variation Explained	https://doi.org/10.1214/12-AOAS597	https://rdrr.io/cran/r.jive/		X	X	X		X	X	X
LRAcluster	Low-Rank Approximation cluster	https://doi.org/10.1186/s12864-015-2223-8	https://rdrr.io/github/xlucpu/MOVICS/man/LRAcluster.html		X				X	X	X
MCIA	Multiple co-inertia analysis	http://www.numdam.org/item/RSA_1996__44_2_35_0/	https://rdrr.io/bioc/omicade4/man/mcia.html		X			X	X	X	X
(MCOA)		https://doi.org/10.1016/j.crvi.2003.10.007	https://rdrr.io/cran/multiblock/man/mcoa.html								
MethylMix		https://doi.org/10.1093/bioinformatics/btv020	https://rdrr.io/bioc/MethylMix/	X							
mixKernel		https://doi.org/10.1093/bioinformatics/btx682	https://rdrr.io/cran/mixKernel/								X
mixOmics		https://doi.org/10.1371/journal.pcbi.1005752	https://rdrr.io/bioc/mixOmics/	X		X			X	X	
moCluster (from MOGSA)		https://doi.org/10.1021/acs.jproteome.5b00824	https://rdrr.io/bioc/mogsa/			X	X			X	X
MOFA	Multi-Omics Factor Analysis	https://doi.org/10.15252/msb.20178124	https://github.com/bioFAM/MOFA								
(MOFA2)		https://doi.org/10.1186/s13059-020-02015-1	https://github.com/bioFAM/MOFA2		X	X	X			X	X
MOGSA (containing moCluster)	Multi-Omics Gene Set Analysis	https://doi.org/10.1074/mcp.TIR118.001251	https://rdrr.io/bioc/mogsa/								X
MSFA	Multi-Study Factor Analysis	https://doi.org/10.1111/biom.12974	https://rdrr.io/github/rdevito/MSFA/		X						
NEMO	NEighborhood based Multi-Omics clustering	https://doi.org/10.1093/bioinformatics/btz058	https://github.com/Shamir-Lab/NEMO		X	X	X			X	X
netClass (containing stSVM)		https://doi.org/10.1093/bioinformatics/btu025	https://rdrr.io/cran/netClass/					X			
NuChart		https://doi.org/10.1371/journal.pone.0075146	ftp://fileserver.itb.cnr.it/nuchart					X			
PINS	Perturbation clustering for data INtegration and disease Subtyping	https://doi.org/10.1101/gr.215129.116	(use PINSPlus instead)								
(PINSPlus)		https://doi.org/10.13140/RG.2.2.22969.06246	https://rdrr.io/cran/PINSPlus/		X				X	X	X
PLRS	Piecewise Linear Regression Splines	https://doi.org/10.1093/bioinformatics/btt082	https://rdrr.io/bioc/plrs/	X							
PMA	Penalized multivariate analysis	https://doi.org/10.1093/biostatistics/kxp008	https://rdrr.io/cran/PMA/							X	X
RGCCA	Regularized Generalized Canonical Correlation Analysis	https://doi.org/10.1007/s11336-011-9206-8	https://rdrr.io/cran/RGCCA/		X	X					
(sGCCA)		https://doi.org/10.1093/biostatistics/kxu001	https://rdrr.io/cran/RGCCA/man/sgcca.html								
RWRF	Random Walk with Restart for multi-dimensional data Fusion	https://doi.org/10.1186/s12859-021-04029-3	https://github.com/Sepstar/RWRF/								X
RWR-M	Random Walk with Restart for Multiplex networks	https://doi.org/10.1093/bioinformatics/bty637	https://github.com/alberto-valdeolivas/RWR-MH								X
SCFA	Subtyping via Consensus Factor Analysis	https://doi.org/10.3389/fonc.2020.01052	https://github.com/duct317/SCFA								X
SNF	Similarity Network Fusion	https://doi.org/10.1038/nmeth.2810	http://compbio.cs.toronto.edu/SNF/SNF/Software.html	X	X	X	X	X	X	X	X
Subtype-GAN	Subtype using Generative Adversarial Network	https://doi.org/10.1093/bioinformatics/btab109	https://github.com/haiyang1986/Subtype-GAN								X
tensorBSS	Tensorial Blind Source Separation	https://doi.org/10.1186/s13059-018-1455-8	https://rdrr.io/cran/tensorBSS/		X						
TransNet	Transkingdom Network	https://doi.org/10.1007/978-1-4939-8728-3_15	https://github.com/richrr/TransNetDemo								X
WGCNA	Weighted Gene Coexpression Network Analysis	https://doi.org/10.1186/1471-2105-9-559	https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/								X

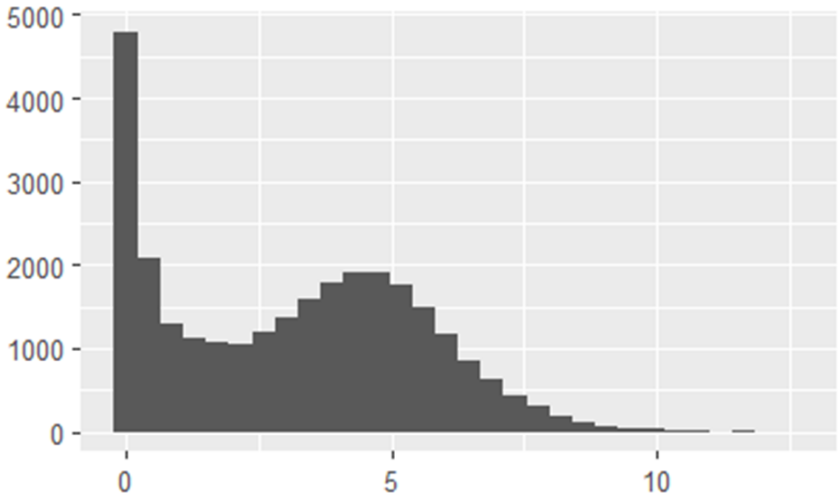
Supplementary Figure 1: Histograms of methylomics and transcriptomics' logged distributions from Adour population

*Histograms represent **A**- methylome and **B**- gene expression distributions for the specific 'Adour' population, with methylation by contexts (CG, CHG or CHH methylation) and gene features (on gene-body or promoter). On the X-axis lies methylation normalized in rbd and logged, gene expression normalized in TMM and logged. On the Y-axis lies counts of genes.*

A)

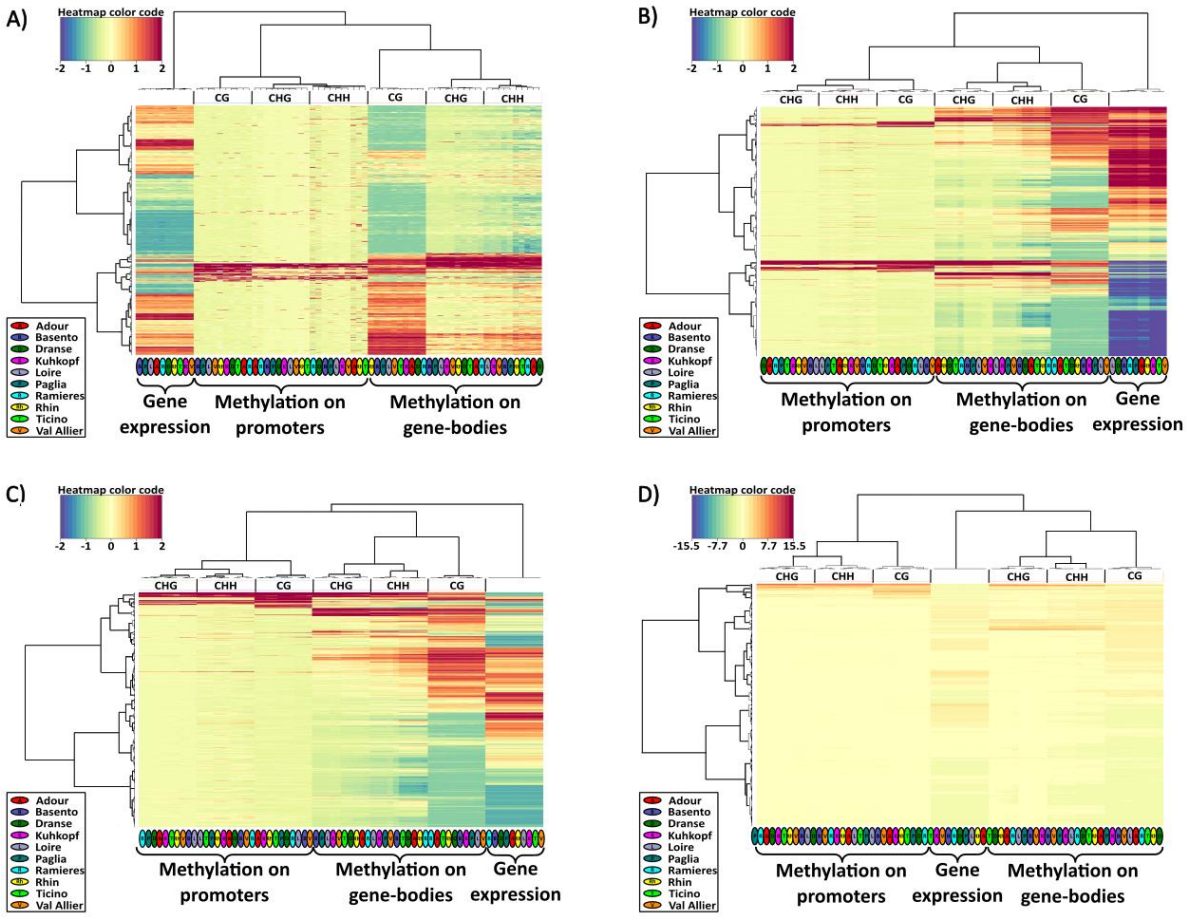


B)



Supplementary Figure 2: cimDiablo_v2 results for different parameters

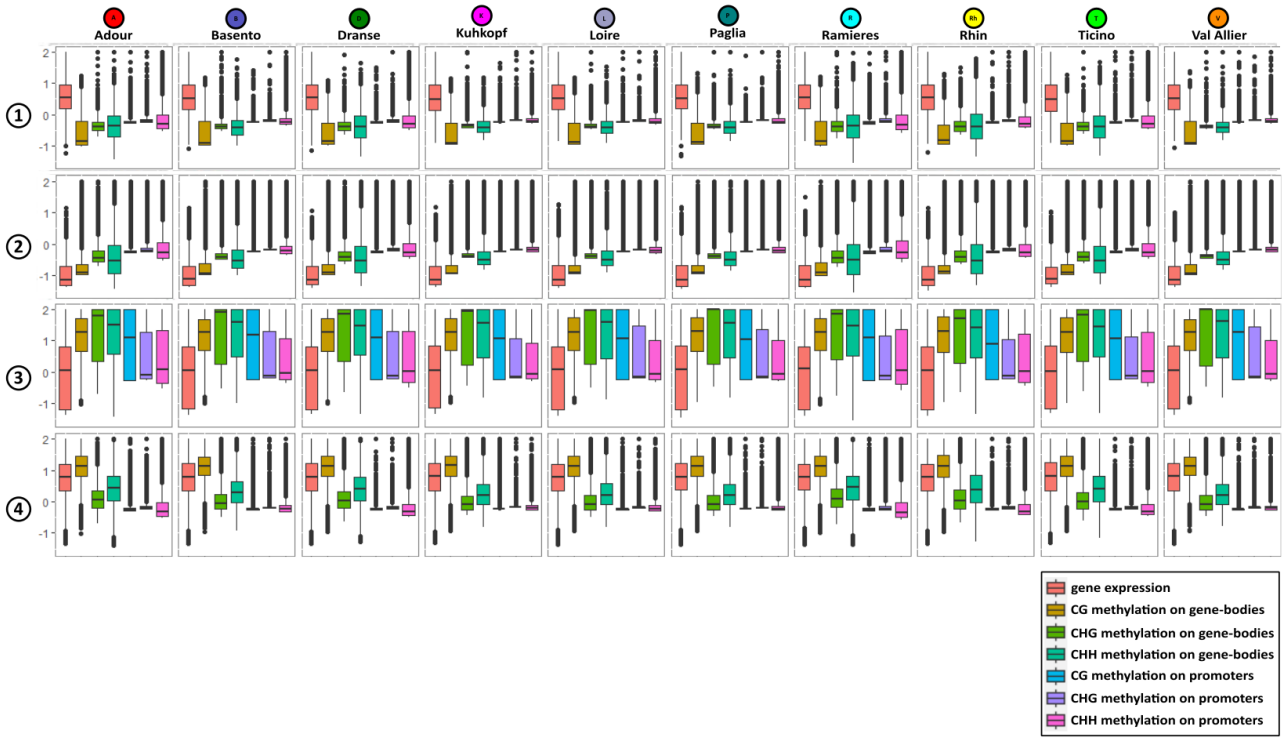
Each row corresponds to one gene and each column to one omic variable. Omics variables are gene expression and DNA methylation data produced for 10 populations of poplars, as presented in the down left legend of each sub-figure. Methylomics data were produced for 3 contexts of methylation (CG, CHH, CHG) on two gene features (gene-body or promoter). Heatmaps' values are computed following one or several of these steps: 1- center and scale a first time initial data (done by default with block.(s)pls), 2- denoise data as presented in Mardoc et al. (2024), 3- center and scale a second time data, 4- apply a cutoff in [-2, 2]. According to the heatmap's color code, blue corresponds to very low and red to very high methylated/expressed genes. Rows and columns' dendrograms are computed by hierarchical clusterings with the euclidean distance and Ward method to cluster together genes and omics variables sharing similar insights. ***A-*** *represents centered and scaled data, without any denoising nor second centering/scaling steps, and data cut in [-2, 2].* **B-** *represents data centered and scaled before being denoised then cut in [-2, 2].* ***C-*** *represents data centered and scaled before being denoised, then centered and scaled a second time, and finally cut in [-2, 2].* ***D-*** *represents data centered and scaled before being denoised, then centered and scaled a second time, but without any cutoff.*



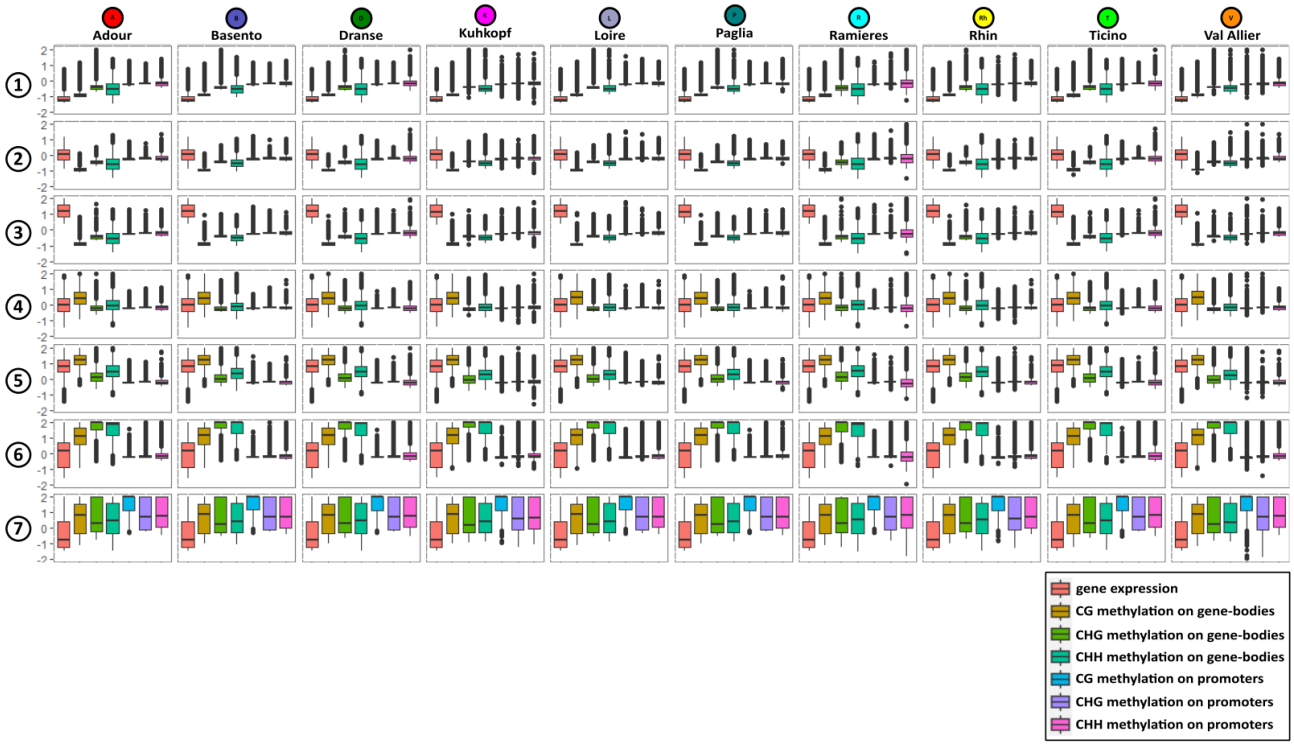
Supplementary Figure 3: Boxplots of k cluster groups in each population for gene expression and DNA methylation data.

A- Analysis of the non-denoised data. According to the row dendrogram, genes were divided into four clusters. For each cluster gene expression and methylation levels (for gene-body and promoter in the three methylation contexts) are represented as boxplots for the 10 studied populations. **B-** Boxplots of omics variations after the denoising step. According to the row dendrogram, the optimal number of k clusters was set to seven. The level of gene expression or DNA methylation is shown for each cluster in each population.

A)



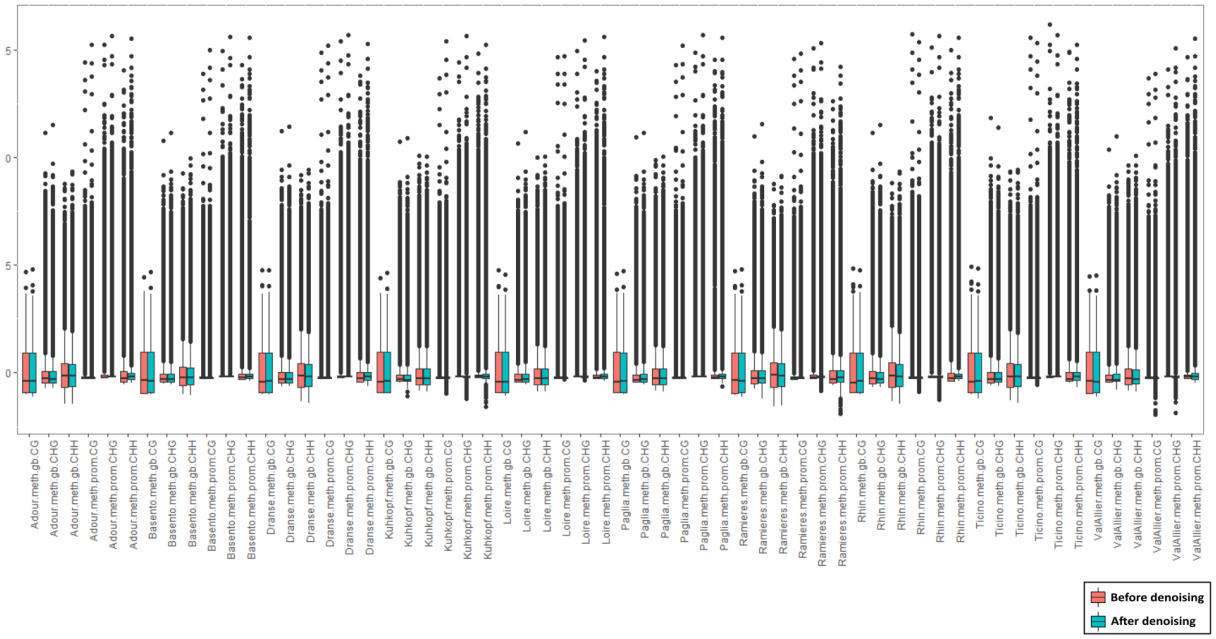
B)


Adour
Basento
Dranse
Kuhkopf
Loire
Paglia
Ramieres
Rhin
Ticino
Val Allier
1
2
3
4
5
6
7
gene expression
CG methylation on gene-bodies
CHG methylation on gene-bodies
CHH methylation on gene-bodies
CG methylation on promoters
CHG methylation on promoters
CHH methylation on promoters

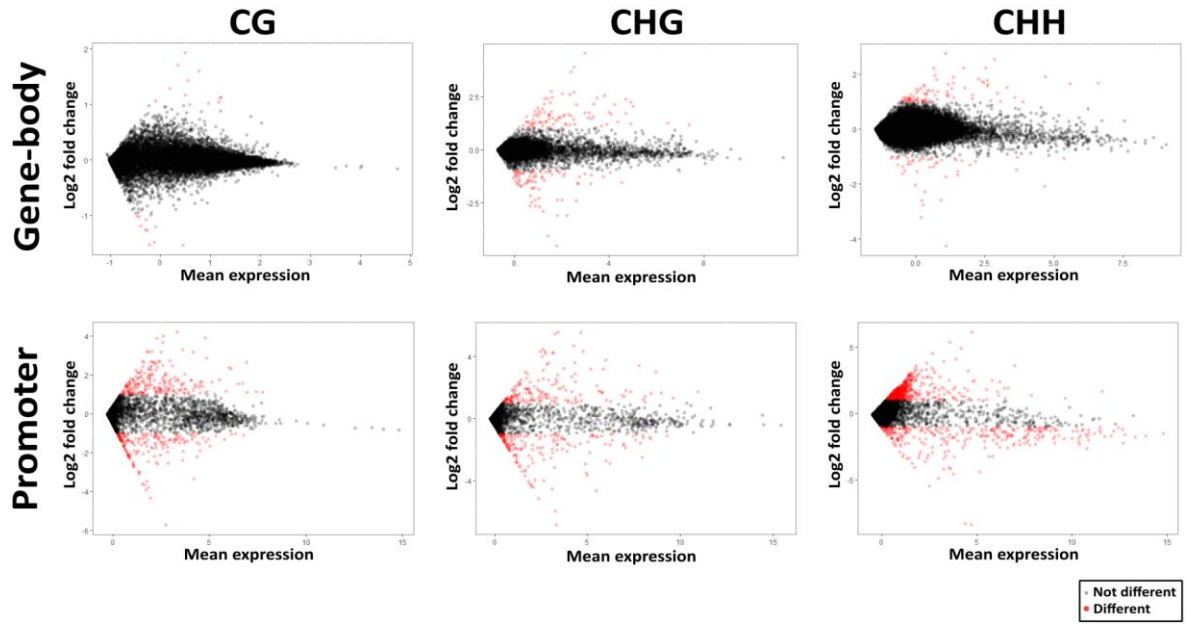
Supplementary Figure 4: Comparison between non-denoised and denoised data for methylation in gene-body and promoter for CG, CHG and CHH contexts.

A- Boxplots of methylation levels between non-denoised (red) and denoised (blue) data in the ten studied populations. gb for Gene-body methylation; prom for promoter methylation. **B-** MA-plot of Adour population for DNA methylation between non-denoised and denoised data. The x axis represents the average expression level while the y axis the log2 fold changes. Red for significant differences above |1| and blacks for no obvious differences.

A)



B)



Annexe B

Sow et al. (2023)

Epigenetic Variation in Tree Evolution: a case study in black poplar (*Populus nigra*)

Mamadou Dia Sow[1,7], Odile Rogier[2], Isabelle Lesur[3,4], Christian Daviaud[5], Emile Mardoc[1], Edmond Sanou[6], Ludovic Duvaux[3], Peter Civan[1], Alain Delaunay[7], Marie-Claude Lesage-Descauses[2], Vanina Benoit[2], Isabelle Le-Jan[7], Corinne Buret[2], Celine Besse[5], Harold Durufle[2], Régis Fichot[7], Grégoire Le-Provost[3], Erwan Guichoux[3], Christophe Boury[3], Abel Garnier[5], Abdeljalil Senhaji-Rachik[2,5], Véronique Jorge[2], Christophe Ambroise[6], Jorg Tost[5], Christophe Plomion[3], Vincent Segura[2,8]*, Stéphane Maury[7]*, Jérôme Salse[1]*

[1] INRAE/UCA UMR GDEC 1095. 5 Chemin de Beaulieu, 63100 Clermont Ferrand, France.
[2] INRAE, ONF, BioForA, F-45075 Orléans, France.
[3] INRAE, Univ. Bordeaux, BIOGECO, F-33610 Cestas, France.
[4] HelixVenture, 33700 Mérignac, France.
[5] Centre National de Recherche en Génomique Humaine, CEA-Institut de Biologie François Jacob, Université Paris-Saclay, 91000 Evry, France.
[6] LaMME, 23 Bd. de France, 91037 Évry Cedex, France.
[7] LBLGC, INRAE, Université d'Orleans, EA 1207 USC 1328, Orleans 45067, France
[8] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro Montpellier, F-34398 Montpellier, France.

*Authors for correspondence: vincent.segura@inrae.fr, stephane.maury@univ-orleans.fr and jerome.sale@inrae.fr

SUMMARY

How perennial organisms adapt to environments is a key question in biology. To address this question, we investigated ten natural black poplar (*Populus nigra*) populations from Western Europe, a keystone forest tree of riparian ecosystems. We assessed the role of (epi)genetic regulation in driving tree species evolution and adaptation over several millions of years (macro-evolution) up to a few generations (micro-evolution). At the macro-evolution scale, polar experienced differential structural (gene loss) and regulation (expression and methylation) reprogramming between sister genomic compartments inherited from polyploidization events. More interestingly, at the micro-evolution scale, both genetic and epigenetic variations differentiate populations from different geographic origins, targeting specifically genes involved in disease resistance, immune response, hormonal and stress response that can be considered as key functions of local adaptation of long lifespan species. Moreover, genes involved in cambium formation, an important functional trait for forest trees, as well as basal functions for cell survival are constitutively expressed though methylation control. These results highlight DNA methylation as a marker of population differentiation, evolutionary adaptation to diverse ecological environments and ultimately opening the need to take epigenetic marks into account in breeding strategies, especially for woody plants.

Keywords: epigenetic variation, DNA methylation, natural population differentiation, macro-evolution, micro-evolution, gene duplication, gene expression, cambium.

INTRODUCTION

Understanding the forces driving species evolution is central in biology. Evolution can be investigated at either short timescale, over a few hundreds of generations (so called micro-evolution) typically between different accessions of a species, or over long timescales of several millions of years between different species (macro-evolution). Genetic variations have long been considered as the major marker of these two facets of evolution, with specific approaches to study the extent of population dynamics and adaptive evolution at the micro-evolutionary scale (Messer et al. 2016) or lineage diversification and speciation from founder ancestral genomes at the macro-evolutionary scale (Murat et al. 2012, Pont and Salse, 2017). New discoveries over the past two decades suggest that epigenetic variation, precisely DNA methylation, also play a role in micro- and macro-evolutionary changes (Jablonka 2017).

At a macro-evolutionary timescale, flowering plants diversified from an ancestor suggested to be made of 15 protochromosomes with 22,899 protogenes dating back to 214 million years ago during the late Triassic era (Murat et al. 2017). Polyploidization (also known as whole genome duplication, WGD) has been proposed as a key mechanism through which new genetic material is generated during evolution promoting morphological and phenotypic biodiversity and contributing to the evolutionary success of the modern angiosperm species (Jaillon et al. 2007; Schnable et al. 2011; Pont et al. 2013). In that context, regulation of duplicated genes though DNA methylation may therefore represent a central phenomenon in the acquisition of novel functions and ultimately phenotypes in the newly formed polyploids compared to their diploid progenitors (El Baidouri et al. 2018; Wang et al. 2017; Chen et al. 2015; Davis et al. 2015; Jablonka 2017).

At a micro-evolutionary timescale, natural selection contributed to shape heritable epigenetic (epialleles) and genetic (alleles) variations in natural populations (Furrow and Feldman, 2013, Li et al. 2014). For instance, in the model plant *Arabidopsis thaliana*, inheritance of epialleles has been associated with population structure and heritable phenotypic variation including adaptive traits (Cortijo et al. 2014). DNA methylation represents one of the most stable epigenetic marks, corresponding to an addition of a methyl group in 5' of cytosines. In plants, methylation occurs in all cytosine contexts, *i.e.*, in CpG, CHG and CHH where H could represent A, T or C. Our understanding of the role of DNA methylation in plants has been mostly obtained from the model plant *Arabidopsis thaliana*, where DNA methylation in the coding regions of genes drives medium to high levels of expression and targets mainly housekeeping gene functions, while the methylation in the promoter region of genes is generally associated with gene silencing (Zhang et al. 2006; Maunakea et al. 2010; Bewick and Schmitz 2017). Despite a growing number of evidences unravelling the relationship between DNA methylation and gene expression, further efforts are needed to better understand methylation control over gene expression regarding different genomic contexts, and the stability of such epigenetic changes through time. Moreover, the contribution of DNA methylation in the evolution and adaptation of long lifespan species, such as trees, remains an open question (Amaral et al. 2020).

In contrast to annual plants, trees are long-living organisms that are exposed to environmental challenges over their entire lifespan (Allen et al. 2010; Anderegg et al. 2016). They have developed various stress sensory mechanisms and physiological strategies to cope with fluctuating environmental changes. The difference between annual (herbaceous) and perennial species in genome evolution is exemplified by the expansion of disease-resistance (R) genes in long-lived species as a key evolutionary process to face a wide range of abiotic and biotic threats over their lifespans (Noir et al. 2001; Ribas et al. 2011; Plomion et al. 2018; Khan and Korban, 2022). Moreover, because of relaxed purifying selection, R-genes in trees were found to display higher levels of genetic diversity (Plomion et al. 2018). It has also been shown that the rate of epigenetic mutation is much higher than genetic variations and then represents a source for short-term adaptation especially for long-living species (van der Graaf et al. 2015).

Poplar is considered as a model forest tree species. Poplars are characterized by a high genetic diversity, fast juvenile growth, vegetative propagation capacity, and the genome of several species has been sequenced (Tuskan et al. 2006; Mader et al. 2016). Over the past decade, poplar has been widely used to investigate the role of DNA methylation in phenotypic plasticity and adaptation to environmental changes (Brautigam et al. 2013; Lafon-Placette et al. 2013, 2018; Zhu et al. 2013; Plomion et al. 2016; Conde et al. 2017; Le Gac et al. 2018, 2019; Sow et al. 2018a, 2018b, 2021; Vigneaud et al. 2023). However, several open questions remain to be addressed, such as (i) does epigenetics play a role in genome reprogramming in response to polyploidization events? (ii) are populations epigenetically differentiated and to what extent does epigenetic differentiation may differ to genetic differentiation? (iii) Can local adaptation be tracked at the epigenetic level?

RESULTS

Genomic pattern of DNA methylation in black poplar

Poplar methylome on a genome-wide scale was investigated from whole genome bisulfite sequencing (WGBS) carried out on ten natural populations, covering the geographical distribution of *Populus nigra* in Western Europe (Figure 1a). DNA methylation in CpG context (averaging 32.1% to 35.4%) appeared more frequent compared to CHG (from 20.0% to 21.9%) and CHH (4.4% to 6.5%) contexts (Figure S1, Table S2). Methylated cytosines or hereafter called SMPs (for Single Methylation Polymorphisms, Figure 1b) were investigated within genes (promoters, exons and introns), intergenic regions and transposable elements (TEs) (Figure 1c and d). More than half (52%) of the SMPs in CpG context fell into the intergenic regions, 7% in promoters, 28% in exons and 13% in introns (defining 48% within genes). In the CHG context, the number of SMPs in intergenic regions decreased to 37%, while we observed an enrichment in exons (35%) and introns (21%) and only 5% in promoter regions. Similarly, in the CHH context 5% of the SMPs were located in promoter regions, 48% in genes (27% in exons and 21% in introns) and the remaining 47% in intergenic regions (Figure 1d). TEs showed a higher level of methylation in comparison to genes in all the three contexts (Figure 1c and d). While promoter and genic regions displayed similar levels of methylation in non-CpG contexts (*i.e.* CHG and CHH), gene bodies were more methylated than promoter regions in the CpG context (Figure 1c and 1d). Overall, 93% of poplar genes (40,091 out of 42,950) and 83% of annotated TEs families (6,131 out of 7,386) were covered by bisulfite-converted sequence reads in the investigated populations (Figure S2).

DNA methylation variation at the macro-evolutionary scale

It has been reported that whole genome duplications (polyploidizarions) drive (epi)genome reprogramming (Bellec et al. 2023). Within the angiosperms, poplar evolved from an Ancestral Eudicot Karyotype with 7 prochromosomes (AEK7) that went through whole genome triplication (*γ* WGT, ≈120 Mya ago) leading to AEK21 with 21 prochromosomes at the basis of any modern Eudicot species (Figure 2a), Murat et al. 2017. Within the Salicaceae botanical family, the modern poplar genome (19 chromosomes) derived from a n=12 ancestor (inherited from the AEK21 with 6 chromosomal fissions and 15 fusions) that has been duplicated (≈60 Mya WGD, to reach a n = 24 intermediate) followed by four chromosomal fissions and nine fusions (delivering a chromosomal equation of 19 = [21 + 6 - 15] x 2 + 4 - 9; Murat et al. 2015). Comparison of gene order between the inferred Eudicot ancestors (AEK7 and AEK21) and the modern poplar genome allowed to investigate the impact of ancient (ancestral γ WGT, ≈120 million years ago) and more recent (poplar specific WGD, ≈60 million years ago) polyploidization events on genome regulation (Figure 2). Polyploids are derived from two parental sub-genomes that were merged in the same nucleus and evolved through inversions, translocations, chromosomal fusion, and fission processes, leading to possible sub-genome differentiation, referred to as subgenome dominance (Bellec et al. 2023). The subgenome dominance is manifested by the differential retention of the ancestral gene content leading to least-fractionated regions (LF) and most-fractionated regions (MF) compartments in the modern poplar genome. Thus, we developed a synteny-based approach to detect genome compartments that underwent different fractionation after the ancestral (γ) WGT and poplar-specific WGD events (Figure 2). Overall, 6,442 genes were identified in LF compartments, and 3,356 genes in MF compartments (Figure 2a, b). GO enrichment for molecular functions showed that LF related genes were enriched for binding process (protein binding, mRNA binding, etc.), signaling (metal ion) and transferase activity, whereas genes in MF fractions were mostly enriched in binding and transcription factor activity (Figure S3). While duplicated genes may return into single copy (defining LF and MF compartments), some of the duplicates remain conserved as pairs in the modern genome following polyploidization events. A total of 2,020 duplicated genes inherited from the poplar-specific WGD were identified and found to be functionally enriched in binding, transcription factor activity and signaling processes (Figure S3).

To investigate the impact of polyploidization in shaping (epi)genome regulation over a macro-evolutionary scale, we compared expression and methylation profiles between genes in LF or MF fractions. No difference was observed for methylation in gene bodies (exons + introns) in CpG context between LF- and MF-located genes. However, in non-CpG contexts, MF-genes

displayed higher methylation levels in comparison to genes located in LF compartments. The methylation level in gene promoters showed a distinct pattern. While no clear bias was observed for CpG and CHG contexts, MF-located genes appeared more methylated in their promoters than genes located in LF fractions (in seven out of the ten investigated populations) in CHH context (Figure 2b). Conversely, the genes located in the LF fraction appeared to be more expressed than genes in the MF fraction, displaying an antagonist pattern with DNA methylation (Figure 2b). Overall, regarding the ancestral (γ) triplication shared within the Eudicots, the LF compartment (where ancestral genes are preferentially retained) showed higher gene expression and lower methylation for promoters in CHH context and in gene bodies in CHG and CHH contexts, compared to genes located in MF compartments (where ancestral genes have been preferentially lost). To better understand the effect of polyploidization on gene regulation, we further investigated methylation and expression profiles between duplicated genes inherited from the recent poplar specific WGD. Interestingly, about ¾ of duplicated genes showed gene expression differences between the two copies of a pair. One of the copies was more expressed than the duplicated counterpart, and most of these over-expressed copies are located in the LF fractions of the genome. For the remaining ¼ of the duplicates, the two copies displayed the same expression level (Figure 2c). Similarly, about half of the duplicated genes showed methylation differences between LF and MF copies, while half of them expressed the same methylation level (Figure 2c). Furthermore, 72% of the differentially methylated gene pairs exhibited also expression differences (Figure S4). Overall, at the macro-evolution time scale, since the ancestral polyploidization events dating back to ≈60 and ≈120 million years ago, polar experience intense structural (gene loss) and regulation (expression and methylation) reprogramming. Genes from the MF subgenome were found to be more methylated (especially in non-CpG contexts) and less expressed compared to genes located in the LF fraction.

DNA methylation variation at the micro-evolutionary scale

We then assessed the methylation landscape at a more recent evolutionary scale across ten populations, using a methylation threshold above the mean methylation in each context separately (*i.e.* a genomic feature was considered methylated when its methylation ratio was above the mean methylation observed from the ten populations, Figure 3a). We then reported between 20,433 and 21,321 genes/TEs methylated in the CpG context in the ten populations, between 9,324 and 9,785 in CHG and between 10,697 and 14,667 in CHH. Exploiting the concept of pan-genome (representing the entire set of genes within a species), consisting of a core genome (containing genes shared between all individuals of the species) and the 'dispensable' genome (containing genes specific to individuals of the species), the pan-methylome consisted of an entire set of 25,485, 13,715 and 29,651 methylated genes and TEs in CpG, CHG and CHH contexts respectively, across the ten populations. The core methylome (*i.e.,* genes/TEs methylated in the ten populations) also varied between the three methylation contexts. The core methylome represented 66.6%, 52.5% and 25.4% of the pan-methylome in the CpG, CHG and CHH context, respectively, suggesting more stable CpG and CHG methylation between populations compared to CHH methylation (Figure 3a).

Furthermore, we investigated population relatedness solely based on available SMPs in all the investigated populations, through a phylo-epigenomic approach. Strikingly, the methylation clustering (strictly excluding SNPs) in CpG and CHG contexts recovered the geographical partitioning of the 10 natural populations, while such structure signal was partially lost in the CHH context (Figure 3b, Figure S5). Namely, CpG and CHG methylation clusterings fit the geographical origin of the populations defining three sub-groups. In addition to the phylo-epigenomic clustering we classically assessed the genetic structure (from SNP data) of the ten *P. nigra* populations. The hierarchical ascendant clustering on the genomic relationship matrix revealed three different sub-groups. The first sub-group was composed of two Italian populations (Basento and Paglia). The second sub-group consisted of four populations, two originating from France (Dranse and Rhine), one from Italy (Ticino) and one from Germany (Kuhkopf). The third sub-group was made of four French populations (Ramieres, Adour, ValAllier and Loire), Figure 3b. The genetic structure of the considered populations was

consistent with their geographical origins and relatedness across west-east (sub-groups 1 and 3) and north-south (subgroups 1 and 2) axes (Figure 1a). Interestingly, the genetic structure was similar to the methylation clustering (CpG and CHG) suggesting that both genetics (DNA polymorphism) and epigenetics (DNA methylation) may act as markers of population differentiation.

Role of DNA methylation in local adaptation

To unravel the biological functions driven by genes most differentiated at the epigenetic levels between the geographical origins, we focused on SMPs located in genes (including promoters) that best support the phylo-epigenomics and genetic structure of the ten studied populations (Figure S6a). Overall, 69,189, 18,523 and 11,282 SMPs in CpG, CHG and CHH contexts respectively, show significant association with the genetic structure of the populations (Figure 4a, Figure S6b). Based on a pcadapt analysis, we identified 2,241 (in CpG), 940 (in CHG) and 389 (in CHH) genes whose methylation profiles contribute the most to the differentiation of the natural populations (Figure 4b, Figure S6c). To obtain a more robust set of candidate genes supporting geographical differentiation, we considered an intersection of the previous pcadapt-derived genes, and those differentially methylated across the populations, resulting in 271 non-redundant genes (Figure 4c). Interestingly, many of these genes were enriched in functions related to disease resistance (34 TIR-NBS-LRR class, *i.e*. R genes), immune response (8 genes), hormonal and stress response (3 genes, MeSA, MeJA and DRY2). Methylation level of the disease resistance genes showed wide diversity among *P. nigra* populations (Figure 4d, Figure S7). For instance, R-genes in Basento, Paglia and Ticino populations from the southern range (Italy) were weakly- or un-methylated, which contrasted with the northern populations such as Loire and ValAllier (France).

Interplay between DNA methylation and gene expression

To unveil the possible link between DNA methylation and gene expression at the population level, we compared methylation in % (standard normalization) or in rbd (Bellec et al. 2023) to gene expression (in TPM) (Figure S8a). When assessing the link between DNA methylation (in %, CpG) and gene expression at the whole genome level, no clear pattern could be established (Figure S8b, spearman correlation r = -0.095, *p-value* < 2.2e-16). However, were able to clearly distinguish three groups of genes combining distinct patterns of expression and methylation. Highly expressed genes in populations displaying low or no methylation (hereafter called Hypo/Up, for hypomethylated and up-regulated genes), highly methylated and lowly expressed genes (hereafter called Hyper/Down, for hypermethylated and down-expressed) and genes that are both methylated and expressed (Figure S8b). Given this result obtained with methylation expressed rbd (methylated reads are weighted with the ratio between methylated and un-methylated reads allowing to focus only on reads supporting the methylation status), we developed a different approach to study the relationship between gene expression and DNA methylation (in %, the standard methylation normalization) by creating methylation ratio quantiles with 10% increment (Figure 5a and b). Comparison between gene expression and promoter methylation percentage (in quantile) clearly showed that the expression of most genes is negatively correlated to increase in methylation level in each methylation context (Figure 5a, Figure S8c). Using the same approach, we investigated the relationship between methylation in gene bodies and gene expression. The expression of genes increased gradually with the methylation level in the CpG context before a strong decrease when methylation reached ~60% (Figure 5b). However, gene body methylation in CHG and CHH contexts displayed a similar pattern observed for promoter methylation (Figure S8d). Overall, it appears that DNA methylation at the population level affects gene expression in a discontinuous manner, having no discernible effect at low-level changes but probably leading to silencing when a certain threshold is reached, either in the promoter or the gene body, pointing out a dosage effect of methylation on gene expression.

Which physiological traits can be driven through expression-methylation interplay? To address this question, we then focused on genes showing the most contrasted methylation and gene expression profile among populations, *i.e.* Hyper/Down and Hypo/Up genes (Figure 5c). Such

genes highlight that DNA methylation may control the tissue-specific expression (from cambium and xylem in the current experiment) of targeted genes. Functional annotation of Hypo/Up genes (*i.e.* weakly- or un-methylated while highly expressed) showed enrichment in functions related to ribosome process (75 genes of housekeeping functions), cell wall and lignin metabolic process (53 genes involved in cambium differentiation), etc. that are essential functions for cell growth and differentiation (Figure 5d). Namely, *XCP2* (XYLEM CYSTEINE PEPTIDASE 2), *WOX4* (a WUSCHEL-related homeobox gene family member) and *PXY* (PHLOEM INTERCALATED WITH XYLEM) involved in xylem and phloem differentiation, were found to be constantly expressed and not methylated. Similarly, GO annotation of Hyper/Down (*i.e.* highly methylated while lowly- or not-expressed) genes revealed enrichment in immune response and glycerolipid biosynthetic process (Figure 5e) with two *MYB* transcription factors but also an important key gene for cambium formation. Namely, *TED6* (tracheary element differentiation-related) involved in differentiation of xylem vessel elements by promoting secondary cell wall formation is apparently constantly silenced by high DNA methylation level. Our data suggest that fine-tuned interplay between omics (expression/methylation) data may control the physiological development and differentiation of the considered tissues (cambium and xylem).

DISCUSSION

Contrary to model species (*A. thaliana*) or crops (tomato and rice) where most of epigenetic studies have been conducted in herbaceous and annual plants so far, poplars are woody long-lifespan trees that are repeatedly exposed to environmental challenges over decades. Therefore, trees have developed various mechanisms enabling them to adapt and to survive. Several previous studies in annual plants found correlations between methylation patterns and habitat or climate in different plant species (Kawakatsu et al. 2016; Xu et al. 2020; Galanti et al. 2022), but there is lack of evidence of such phenomenon, if existing, for perennial species. In this study, we present the analysis of natural populations of black poplars (*Populus nigra*) sampled from four distinct geographic origins (France, Italy, Germany and Netherlands) and grown in a common garden in France. We focus on cambial tissues, an important functional trait for forest trees responsible of wood formation, to assess the evolutionary and functional impact of epigenetic variations at macro- (past polyploidization events) and micro- (geographical origins) scales.

DNA methylation as a stable epigenetic mark in tree populations

How DNA methylation is acting on genomes of different tree populations? DNA methylation has been studied in many plant species, showing wide diversity in terms of methylation level and pattern (Bartels et al. 2018; Niederhuth et al. 2016). Methylation diversity is observed for all the three methylation contexts (CpG, CHG and CHH) even though CpG methylation (the predominant type) is more stable across species (Niederhuth et al. 2016). Here, we assessed DNA methylation at the population level in black poplars. Methylation level was quite stable for CpG and CHG contexts across populations, and more variable in the CHH context. About half of methylated cytosines were located in genomic features, including promoter regions. Transposable elements appeared to be more methylated in all three methylation contexts compared to genic features, a general trend observed in many species (Cao et al. 2021; Zhang et al. 2018; Zhang et al. 2006). The methylation of TEs was much higher in CpG and CHG contexts, a feature detected in almost all studied plant epigenomes (Bräutigam and Cronk 2018; Lang et al. 2018; Zemach et al. 2010). The distribution of DNA methylation across the investigated genomic features were also significantly different between the methylation contexts. While in the CpG context, genic regions are more methylated than promoters, they display similar methylation levels in the non-CpG contexts. The increased CpG methylation level in genic regions could be related to DNA methylation-driven silencing of intron-located repetitive elements (Cao et al. 2021). Moreover, the pan-methylome consisted of 25,485, 13,715 and 29,651 methylated genes and TEs in CpG, CHG and CHH contexts respectively across the studied *P. nigra* populations when applying a methylation threshold above the mean value observed in the populations in each context. Specifically, 66,6%, 52.4% and 25.4% of

poplar genes/TEs were constantly methylated (core methylome) in CpG, CHG and CHH contexts, respectively. Although DNA methylomes are available for many plant species, integrative analysis of the DNA methylome profiles in a pan-methylome manner (between individuals-tissues of the same species) has been little studied in plants. In humans, the comparison of DNA methylation patterns in different cancer types has provided novel insights into alterations that contribute to cancer development (Shimizu et al. 2021; Liu et al. 2018). Here, we show that 16,961, 7,207 and 7,543 genes and TEs are methylated in all 10 black poplar trees, in CpG, CHG and CHH context, respectively. Only the CHH methylation context expresses wide diversity between the different populations and highlights the conservative nature of DNA methylation in *P. nigra* populations (core methylome greater than 50% of the pan methylome in CpG and CHG). Overall, these results provide a comprehensive analysis of genome-wide DNA methylation across natural populations of a keystone species of the riparian forest ecosystems, highlighting the conservative nature of methylation patterns across the populations especially in CpG and CHG contexts.

DNA methylation and gene expression reprogramming following ancestral polyploidization events

Do epigenetic modifications provide a signal of long-term evolution following polyploidization events? Polyploidization (or WGD) has occurred frequently during plant genome evolution and represents an evolutionary driving force which provides extra genetic material to be specialized for phenotypic diversification, adaptation, and survival. Such an event is followed by a diploïdization (or extensive fractionation) process reverting the polyploids to diploid status through gene number reduction (Van de Peer et al. 2021; Soltis et al. 2015; Wendel, 2015; Murat et al. 2012). Ancestral gene losses during the diploidization process (accompanied by genomic structural rearrangements) may lead in some cases to subgenome dominance in the form of biased compartmentalization, with dominant (retaining more ancestral genes) and sensitive (retaining fewer ancestral genes) subgenomes, so called least-fractionated (LF) and most-fractionated (MF) regions (Alger and Edger, 2020; Cheng et al. 2016). We conducted a comparative analysis between Eudicot ancestors (AEK7 and AEK21) and the modern poplar genome to detect fractionation bias in ancestral gene retention after polyploidization events. Instead of focusing on the specific *Salicaceae* WGD which was inferred to have occurred around 60 million years ago (Dai et al. 2014) and where no subgenome dominance has been reported (Liu et al. 2017), we investigated the evolutionary impact of the γ triplication event shared by all the rosids (~120 mya) in terms of gene loss, DNA methylation and gene expression. Hence, 15% of the annotated poplar genes could be attributed to the dominant-LF compartment and 8% to the sensitive-MF compartment. Functional annotation of genes in the dominant-LF fraction showed enrichment in binding and signaling process, whereas MF-genes are more involved in transcription factor activity as previously reported (Blanc and Wolfe, 2004; Freeling, 2009; Hao et al. 2021). The LF compartment showed a higher gene expression and lower methylation levels for promoters in the CHH context and for genic regions in CHG and CHH contexts, compared to genes located in the MF compartment. This has also been reported in cotton, maize and *Brasicaceae* where the dominant-LF subgenome appeared to be more expressed and less methylated (Renny-Byfield et al. 2017; Woodhoouse et al. 2014; Schnable et al. 2011). Besides subgenome dominance, we further investigated the regulation plasticity of duplicated genes (that remain conserved as pairs after the diploidization process) derived from the specific *Salicaceae* WGD. It has been proposed that duplicated genes, initially having identical sequences and functions, tend to diverge in regulatory and coding regions, which may change their expression pattern or lead to the acquisition of new functions (Blanc and Wolfe, 2004; Xu et al. 2012). In poplar, about ¾ of the duplicated genes showed gene expression differences and are preferentially located in the dominant-LF fraction of the genome. The number of duplicated pairs with expression differences vary widely between species (Zhao et al. 2017; Schnable et al. 2011; Throude et al. 2009; Yim et al. 2009). Similarly, about half of duplicated genes in poplar showed methylation differences. Keller and Yi (2014) highlighted that for a majority of duplicate gene pairs in humans, a specific duplicate partner is consistently hypo- or hypermethylated across highly divergent tissues. Analysis of duplicated gene pairs in

cassava showed that gene body methylation and gene expression have co-evolved within a short evolutionary timescale (Wang et al. 2015; Wang et al. 2017). Here, we found that 72% of differentially expressed genes showed methylation differences. Overall, after past polyploidization events, poplar showed divergent evolutionary patterns reflected by subgenome dominance, gene expression and epigenetic differences.

DNA methylation as a marker of population differentiation alike genetic markers

To what extent do epigenetic marks differentiate populations from different geographic origins? Besides genetic variation, natural plant populations usually also harbor epigenetic variation. This idea arose by the fact that DNA methylation variation in natural plant populations is often non-random and geographically structured (Dubin et al. 2015; Garino et al. 2015; Gugger et al. 2016; Kawakatsu et al. 2016; Galanti et al. 2022). Using a phylo-epigenomic approach, we clearly established that DNA methylation, in the same manner as genetic markers, differentiate population structure according to their geographic origins. This result suggests that DNA methylation, at least for CpG and CHG contexts, could be used for genotype and/or population differentiation just like SNPs. Thus, epigenetic mechanisms could represent an additional layer of heritable phenotypic variation since these modifications can be inherited through generations and possibly explain a part of the missing heritability of Mendelian traits (Maher, 2008; Becker et al. 2011; Heckwolf et al. 2020; Noshay and Springer, 2021; Sammarco et al. 2022). Within natural populations of oak in Southern California, Platt et al. (2015) found patterns of genetic and epigenetic (mainly CpG methylation) differentiation indicating that local adaptation is operating on large portions of the oak genome. Methylation in the CpG context is more frequent but also more abundant at the gene level, suggesting that it may have more adaptive 'power' compared to non-CpG methylation. This hypothesis is supported by Gugger et al. (2016) who identified variations in CpG methylation linked to climatic variations at or near genes, suggesting a direct relationship between DNA methylation in CpG and local adaptation. However, Platt et al. (2015) did not establish a link between methylation in CHG context and local adaptation and suggested that CHG methyl-polymorphisms are not playing a significant role. In the current study we clearly established that CHG methylation clustering fits the geographic distribution of the natural populations and is similar to their genetic clustering. Nonetheless, CHG methylation levels vary more widely across species compared to CpG methylation (~10.0% in *A. thaliana* against ~26.8% in *P. trichocarpa*, Barlets et al. 2018) and may therefore have different impacts depending on the species. In our populations, the overall methylation level in CpG and CHG contexts was quite high, ranging between 32.1% to 35.4% for CpG and between 20.0% to 21.9% for CHG, depending on the considered population. These results may suggest that the geographic structure of the population origins is genetically and more interestingly epigenetically marked and that DNA methylation needs to be considered as a key marker of species micro-evolution (over millennia of evolution) in complementing the Mendelian principle of genetic inheritance as proposed by Jablonka (2017). However it is not clear whether such results also illustrate epigenetic drift or to what extent epigenetic variations follow a demographic structure determined by genetic markers (Lamka et al. 2022). Galanti et al. (2022) showed that in *Thlaspi arvense*, natural epigenetic variations are significantly associated with both genetic variation and environment of population's origin, and that the relative importance of the two factors strongly depends on methylation contexts, with environmental variations being higher in non-CG contexts. Hence, the CHG population structure reported here may support the hypothesis of a role of DNA methylation in population differentiation. While some publications mention genetic control over epigenetic variations (Becker et al. 2011; Dubin et al. 2015; Sow et al. 2018b, Alvarez et al. 2021), epigenetics could also accelerate mutational dynamics (Ossowski et al. 2010; van der Graaf et al. 2015 Johannes, 2019; Zhou et al. 2020) and therefore could mimick the geographical structuring of both phylo-genetic and phylo-methylomics structures.

DNA methylation reshapes gene expression dynamics among populations

While DNA methylation is quite often associated with regulation of gene expression, it is more likely that instead of being a simple "on-off switch", DNA methylation has a nuanced impact on

the expression of genes according to targeted genomic features and contexts (Zhang et al. 2006; Niederhuth and Schmitz, 2017; Bewick and Schmitz, 2017). However, even when DNA methylation was evaluated per genomic features (*i.e.* promoter and genic regions), we observed poor correlation between DNA methylation and gene expression on the whole genome level among populations. This result is in contrast with common knowledge that DNA methylation in promoters is associated with gene silencing (Kon and Yoshikawa, 2014; Nuo et al. 2016; Bewick and Schmitz, 2017; Ma et al. 2020), but consistent with the findings of Li et al. (2012), who showed that methylation in promoter regions repressed only a few heavily methylated genes in rice. In order to scrutinize in more depth, the relationship between DNA methylation and expression, we categorized genes according to quantiles of their methylation level. We first assessed the link between promoter methylation and gene expression and revealed a significant negative correlation with gene expression when promoters are highly methylated. The same pattern was reported during transgene inactivation in transgenic plants (Weinhold et al. 2013; Nuo et al. 2016). For methylation in genic regions, that has been positively correlated with gene expression (Ball et al. 2009; Bewick and Schmitz, 2017), we confirm, at the population level, the expected relationship in the CpG context, but only at low-to-moderate methylation levels. The positive correlation of genic methylation and expression is disrupted on the genome-wide scale by the fact that highly (methylation ratio above 60%) methylated gene bodies showed a strong negative correlation with gene expression. Overall, the link between DNA methylation and gene expression is more nuanced and complex than initially thought, with specificity on a gene-by-gene basis. Genome-wide patterns can only be disentangled when genomic features (i.e., promoter or genic regions) and sequence contexts are evaluated separately.

DNA methylation regulation of key functional trait among populations

As long-lived species, trees are characterized by the expansion of disease-resistance (R) genes as a key process to face a wide range of biotic threats over their lifespans (Plomion et al. 2018). R genes are shown to enable plants to detect specific pathogen-associated molecules and initiate signal transduction to activate defenses (Hammond-Kosack and Jones, 1997). NBS-LRR (nucleotide binding site-leucine rich repeat) genes, a class of immune receptor, are well known to play fundamental roles in disease resistance (Dangl and Jones, 2001; Kong et al. 2020). The expression levels of plant NBS-LRR genes may be regulated by different mechanisms, including DNA methylation (Kong et al. 2020). In our study, we found that genes with different methylation profiles between natural populations of distinct geographical origins were enriched in functions related to disease resistance (TIR-NBS-LRR genes). While R-genes were weakly- or un-methylated in populations originating from Italy (Basento, Paglia and Ticino), they appeared more methylated in French populations such as Loire and ValAllier. This suggests that DNA methylation could represent a key mechanism to control the expression of R-genes in trees. Indeed, previous report in common bean has shown that most of R-genes are methylated, reminiscent of the DNA methylation pattern of surrounding repeated sequences (Richards et al. 2018a). As R-genes-triggered immunity can be associated with a reduction in growth and yield, so-called ‘fitness costs’, plants use an elaborate interplay of different mechanisms to control R-gene transcript levels to avoid the associated cost of resistance in the absence of a pathogen (Richards et al. 2018b). Hence, it appears that DNA de/methylation is required for the proper expression/silencing of defense related genes (Zeng et al. 2021). We also found that genes with different methylation profiles between populations are enriched in functions related to hormonal signaling. Such a relationship between DNA methylation and phytohormone related genes has been already reported in poplar (Lafon-Placette et al. 2018; Maury et al. 2019; Zhang et al. 2020; Sow et al. 2021). Similarly, *DRY2* (DROUGHT HYPERSENSITIVE 2, response to water deprivation, response to ethylene), another important gene involved in shoot growth, was found to be differentially methylated between the studied populations, suggesting variation in response to drought stress among the poplar gene pools (Wang et al. 2020).

We then investigated genes showing strong negative correlation between DNA methylation and gene expression, namely the Hypo/Up (weakly- or un-methylated and highly expressed)

and the Hyper/Down (highly methylated while lowly- or not-expressed) genes, within the considered tissues, cambium and xylem. For Hypo/Up genes, we observed enrichment in housekeeping (ribosome activity) and cambium activity (e.g. in cell wall biosynthesis) related genes. Since housekeeping genes are required for the maintenance of basal functions for cell survival (Joshi et al. 2022), it is assumed that they are constitutively expressed and therefore depleted or lowly methylated to maintain their proper expression. Similarly, several genes essential for cambium differentiation were found weakly- or un-methylated and highly expressed. This is the case for (i) *XCP2*, a XYLEM CYSTEINE PEPTIDASE involved in programmed cell death of rays tylosis essential for heartwood formation (Avci et al. 2008; Nakaba et al. 2015; Zheng et al. 2015), (ii) *PXY*, a receptor-like kinase essential for maintaining polarity during plant vascular-tissue development (Fisher and Turner, 2007) and (iii) *WOX4*, a WUSCHEL-related homeobox gene family playing a crucial role in the regulation of cambium cell proliferation (Fisher et al. 2019). The vascular stem-cell tissue known as cambium generates phloem cells on one side and xylem cells on the other. While xylem is required for water transport, phloem is required primarily for the transport of photoassimilates (Sieburth, 2007; Etchells et al. 2015). Xylem and phloem form the vascular tissues of trees, an important functional trait for forest trees. It has been shown that in *pxy* mutants, the spatial organization of vascular development is lost and the xylem and phloem are partially interspersed (Fisher and Turner, 2007). Similarly, ectopic overexpression of *PXY* gene in hybrid poplar resulted in vascular tissue abnormalities and poor plant growth (Etchells et al. 2015). Moreover, *wox4* mutants exhibited reduced cell division activity in the cambial tissue (Fisher et al. 2019). Interestingly the two candidate genes, *PXY* and *WOX4* are known to act in the same signaling pathway, *WOX4* being downstream of the *PXY* receptor kinase to regulate stem cell proliferation (Etchells et al. 2013; Fisher et al. 2019; Hu et al. 2022). Recently, Dai et al (2023) using more than 20 sets of poplar transgenic lines have shown that *WOX4* system may coordinate genetic and epigenetic (histone marks) regulation to maintain normal vascular cambium development for wood formation.

For Hyper/Down genes, we observed an enrichment in innate immune response, suggesting a control of immune response genes by DNA methylation. In addition, *TED6* (Tracheary Element Differentiation-Related6) involved in xylem vessel differentiation (secondary cell wall) was found highly methylated and silenced. Transient RNAi of *Arabidopsis TED6* and 7 resulted in aberrant secondary cell wall formation of *Arabidopsis* root vessel elements (Endo et al. 2009). Moreover, it has been shown that a *rdd* mutant in *Arabidopsis* (defective DNA demethylation) with affected expression of many genes including *TED6* is impaired in tracheary element differentiation (Lin et al. 2020). Homology searches have identified *TED6/7*-like proteins only in the angiosperm lineage suggesting that the development of *TED6/7* proteins could have coincided with the emergence of the angiosperm lineage, and that they may have made key contributions to the evolution of water-conducting cells from tracheids to vessels (Rejab et al. 2015).

Overall, the current study shows that DNA methylation leaves genomic footprints recognizable at both macro- and micro-evolutionary scales and has a nuanced relationship with gene expression in poplar. In this study, genes with key functions for trees (disease resistance and wood formation) have been identified based on their DNA methylation patterns (on genome-wide or inter-population scales), suggesting a role of DNA methylation in their regulation. Our data showed that DNA methylation in poplar populations display natural variations, and may regulate fitness traits (disease resistance and wood formation). These results also highlight the need to take epigenetic markers into account in breeding strategies (Kakoulidou et al. 2021) together with genetic markers for both wood production and quality in the context of climate change that requires adaptation to biotic and abiotic constrains.

EXPERIMENTAL PROCEDURES

Sample collection and population structure

The initial experimental design consisted of 1,160 black poplar genotypes sampled from 14 river catchments of 4 European countries, Germany, France, Italy and Netherlands (Guet et al. 2015; Gebreselassie et al. 2017). The genetic diversity within this black poplar collection

was previously characterized using 5,600 SNPs from a 12k Infinium array (Faivre-Rampant et al. 2016) and led to the definition of a subset of 241 genotypes representative of the genetic diversity while avoiding the widespread introgression from the cultivar *P. nigra* cv. Italica. These 241 genotypes originated from 10 river catchments (Chateigner et al. 2020). Population structure analysis was carried out on this restricted set of 241 genotypes with the same set of 5,600 SNPs and the model-based ancestry estimation in the ADMIXTURE program (Alexander et al. 2009) highlighted six genetic clusters which minimized the cross-validation error (Figure 1a).

Genomic DNA extraction

For the present study, twenty genotypes (two genotypes per river catchment) were selected to be representative of the diversity of the black poplar collection. We sampled cambium and xylem on two biological replicates (clones) of each genotype, located in two blocks of a large common garden experiment at INRAE Orléans, France. The resulting 80 samples were further used to extract genomic DNA. DNA was extracPG_34_pe.bamted using a cetyl trimethy-lammonium bromide (CTAB) buffer according to Doyle & Doyle (1987). Extracted gDNA was then quantified using a Nanodrop spectrometer (Thermo Fisher Scientific, Waltham, MA, USA). An equimolar pool of the gDNA samples from the two individual clones and the two tissues (xylem and cambium, for wood formation) of each genotype was then performed. The gDNA pool for each of the 20 genotypes was sent to the CEA laboratory in Evry for both WGS (whole genome sequencing) and WGBS (whole genome bisulfite sequencing).

Whole Genome Sequencing (WGS), read alignment and variant calling

Whole genome sequencing was performed by the 'Centre National de Recherche en Génomique Humaine (CNRGH), Institut de Biologie François Jacob, CEA, Evry, France'. After a complete quality control, genomic DNA (1 µg) has been used to prepare a library for whole genome sequencing, using the Illumina TruSeq DNA PCR-Free Library Preparation Kit, according to the manufacturer's instructions. After normalization and quality control, libraries have been sequenced on a HiSeqX5 platform (Illumina Inc., CA, USA), as paired-end 150 bp reads. One lane of HiSeqX5 flow cell was used for each sample, to reach an average sequencing depth of 30x for each sample. Sequence quality parameters were assessed throughout the sequencing run and standard bioinformatics analysis of sequencing data based on the Illumina pipeline was used to generate FASTQ files for each sample. We followed the bioinformatics pipeline described in Rogier et al. 2018 with small modifications. After a read quality control with *FastQC* v0.11.7 (Andrews, 2010), sequences were trimmed using the *Trimmomatic* tool v0.38 (Bolger et al. 2014). The adapter sequences were removed, the 9th first bases and the low-quality bases were trimmed based on a PHRED score below 20 and finally the reads with a length of less than 35 nucleotides were discarded. The mapping of the reads was performed using *BWA mem* v0.7.17 (Li, 2013) on the *Populus trichocarpa* v3.1 reference genome (Tuskan et al. 2006). Then, the *Picard Toolkit* v2.18.2 allowed the removal of the duplicated reads. Three caller programs were used to identify the variants: (i) *GATK* v4.0.11.1 (McKenna et al. 2010) using the *HaplotypeCaller* tool in single-sample calling mode followed by joint genotyping of the 20 samples; (ii) *FreeBayes* v1.2.0-2 (Garrison et al. 2012) in multi-sample calling mode; and (iii) *SAMtools* v1.8 (Danecek et al. 2021) using the *mpileup* tool in multi-sample calling mode followed by the *BCFtools* v1.8 (Li, 2011) call command. Finally, we considered only biallelic intra-nigra SNPs with quality threshold ≥ 30. Raw data were filtered with *VCFtools* v0.1.15 (Danecek et al. 2011) and only SNPs identified by at least 2 callers were selected to obtain the final set of SNPs.

Whole Genome Bisulfite Sequencing (WGBS), read alignment and methylation calling and annotation

Whole genome bisulfite sequencing was performed using the Ovation Ultralow Methyl-seq kit (Tecan Genomics/Nugen, San Carlos, CA, USA, http://www.nugen.com/products/ovation-ultralow-methyl-seq-library-systems) following the published procedure (Daviaud et al. 2018). The workflow of the library preparation protocol follows a standard library preparation protocol,

in which methylated adaptors are ligated to the fragmented DNA prior to bisulfite conversion. 200 ng of genomic DNA was fragmented to a size of approximately 200 base pairs (bp). Purified and methylated adaptors compatible with sequencing on an Illumina HiSeq instrument were ligated. The resulting DNA library was purified, and bisulfite converted. A qPCR assay determined the optimal number of PCR amplification cycles (between 10 to 15 cycles) required to obtain a high diversity library with minimal duplicate reads prior to the actual library amplification. The 20 samples to be sequenced were combined at equimolar quantities in two pools and the sequencing was performed in paired end mode (2 x 150 pb) on ten lanes of two Illumina HiSeq4000 flow cells to reach a minimal theoretical coverage of 30X for each sample. Fastq files from each sample were concatenated following sequencing.
The bioinformatic pipeline for DNA methylation analysis was executed on a Galaxy instance of the IHPE (Interactions Hôtes Pathogènes Environnements) platform (http://galaxy.univ-perp.fr/, Perpignan, France; Dugé de Bernonville et al. 2022) using *Populus trichocarpa* (v3.1) as a reference genome. Reads were first trimmed with *Trim Galore* (Galaxy v0.4.3.1) prior to mapping with *BSMAP* (Galaxy v1.0.0, Xi and Li, 2009) using default settings. Average sequencing depth after the mapping step ranged from 7X to 26X depending on genotypes (Table S1). Methylation calling was then achieved with *BSMAP methylation caller* (Galaxy v1.0.0) for the detection of methylated cytosines hereafter called SMPs for single methylated polymorphisms in the three methylation contexts (CpG, CHG and CHH). Bisulfite non-conversion rate ranked from 0.4 to 1.1% (Table S1). The *Methylkit* and *genomation* R packages (v1.18.0) were used for the analysis and annotation of DNA methylation data. SMPs were annotated for gene promoters (400 bp centered on the transcription start site (TSS) which allow to target the proximal promoter containing primary regulatory elements), genic regions (exons and introns), intergenic regions and transposable elements (TEs). With methylation expressed in percentage (%), we also used the rbd (read by density) normalization approach (Bellec et al. 2023) to study the link between DNA methylation and gene expression, with rbd = mCs x ratio (0-1) and where ratio = mCs / (mCs + Cs). MCs correspond to the number of methylated reads and Cs the number of un-methylated reads. Gene Ontology (GO) annotation was inferred from the *Arabidopsis* TAIR10 gene annotations using the best blastN hit (BlastN V3.0). GO terms Enrichment was performed using the *metascape* software with default parameters (Zhou et al. 2019).

RNA-seq data recovery

The full set of RNA-seq data has already been published by Chateigner et al. 2020. We retrieved the same 20 genotypes analyzed here for WGS and WGBS and performed TPM normalization (Transcript Per Million, edgeR v3.26.4) for the comparison of different set of genes allowing to address gene size differences.

Construction of phylogenetic and methyl-phylogenomic trees

Phylogenomic trees based on DNA methylation profiles were constructed with the *methylkit* software. Bed files corresponding to the methylation data for each of the genotypes were merged to form a methylation matrix grouping all genotypes by the context of methylation, tolerating up to 30% of missing data. For the CpG context only (symmetric case), the reads supporting the two strands were merged to improve coverage. Methylated matrices for each context were then filtered with SNPs data from the WGS to discard methylation calls due to genetic C/T polymorphisms. We set a minimum coverage of 7X for all genotypes and discarded the STR-010 genotype from the Rhin population as it did not reach this minimum value. The genotypes were clustered based on the similarity of their methylation profiles for each methylation context separately using ward hierarchical clustering and pearson's correlation distance implemented in the *methylkit* software. Similarly, the phylogenetic tree was built using only SNPs without any missing value and with a minor allele frequency (MAF) above 5%. The genomic relationship matrix (GRM) was then estimated following VanRaden (2008) and the population structure evaluated by performing a hierarchical ascendant clustering using the Ward method on the GRM, converting relationships into dissimilarities.

Detection of epigenomic signatures of local adaptation and/or drift

We took advantage of the available *pcadapt* tool (v4.3.3) in order to detect epigenomic markers involved in biological adaptation (Luu et al. 2017). We focused only on SMPs located in genic features (exons and introns) including promoter regions (200 bp around the TSS). Epigenetic markers (within genes) with particular patterns (i.e. outliers) were identified with the pool option of *pcadapt*. We divided the methylation matrices by 100 so that DNA methylation data are similar to frequencies. For all omics data we ran the analysis with the default options of *pcadapt* (MAF at 0.05). The number of principal components (PCs, K) captured varied between the three methylation contexts and was fixed according to the scree plots considering Catell's rule (Cattel R.B, 1966). Thus, only principal components above the point of inflection were taken. For methylated CpGs, three components were retained while four components were selected for non-CpG contexts (CHG and CHH). We set the minimum cut-off for outlier detection at a p-value below 0.05 using Benjamini-Hochberg correction (FDR). Epigenetic markers that have disproportionately high contribution to the structure defining PCs were considered as putative markers of local adaptation. Using a different approach in order to characterize which epigenetic markers mimic the phylogenetic tree, we ran differential analysis between populations. First, we applied a filter on the standard deviation per position to select the cytosines which vary the most between the samples. Using the R package *MethylKit*, we fit a weighted fractional Logistic regression model to explain the ratio of methylated cytosines by the population structure. Weights are defined as the read coverage i.e., the sum of methylated and non-methylated cytosines after bisulfite conversion. The Chi-square Test is then used to assess the significance of the association between the cytosines and the genetic population structure. P-values are corrected by the Bonferroni approach and the significance threshold is fixed to 0.01. The comparison between the epigenetic trees built on top-markers and the phylogenetic tree are given by the tanglegram plots combining the two dendrograms for each context.

Inference of duplicated genes and genomic fractions following polyploidization

The poplar genome was compared (with blastP) to the inferred Ancestral Eudicot Karyotypes (AEK) consisting of post-γ AEK with 21 proto-chromosomes and 9,022 ordered protogenes and a pre-γ AEK with 7 proto-chromosomes and 6,284 ordered protogenes, with γ being the shared triplication at the basis of rosids. The complete method for ancestral karyotype reconstruction is published in Murat et al. 2017. Filtering out one-to-two gene relationships between respectively AEK and poplar allowed the identification of duplicated genes inherited from the poplar-specific duplication within the *Salicaceae* family (Murat et al. 2015, Bellec et al. 2023). For each ancestral region, we detected compartments of the genome that underwent different fractionation in ancestral gene retention following polyploidization, defining the LF (Least Fractionated) and MF (Most Fractionated) compartments in the poplar genome, as proposed in Bellec et al. 2023. Differences between the evolutionary structural features (i.e., LF- vs. MF- genes) were assessed using 4 statistical tests: 2 distribution tests (Kolmogorov-Smirnov and Anderson-Darling tests), Kruskal-Wallis (median comparison) and T-test (mean comparison). Significant results were considered when at least two statistical tests passed the p-value cut-off (0.05). Differentially expressed genes (DEGs) and differentially methylated genes (DMGs) between the duplicated genes were investigated with the R package edgeR (v3.38.4, Chen et al. 2017) using TPM (to account gene size differences) and rbd (Bellec et al. 2023) data respectively. Differences were assessed using the likelihood ratio test and p-values adjusted by Benjamini-Hochberg method to control the false discovery rate.

FIGURES

Figure 1: Methylation landscape in poplar. **A**, Geographical distribution of the 241 genotypes (black dot) from natural populations of *Populus nigra* representative of 6 genetic clusters (one by color) from a model-based ancestry estimation in ADMIXTURE program. **B**, Schematic representation of the SMPs (Single methylation polymorphisms). Numbers represent methylation ratio (mCs/total C) in a specific genomic position; Pop for Population; Ind for Individual. **C**, Circos plot of the distribution of DNA methylation in CpG, CHG and CHH contexts

along genes and TEs for the 19 chromosomes of poplar. Methylation data are plotted for 1Mb windows with a minimum of methylation at 25% and 10X of coverage. CpG methylation density in blue; CHG methylation density in red; CHH methylation density in green; Gene density in black; TEs density in red. **D**, Annotation and methylation level (in percentage) of the SMPs (Single Methylated Polymorphism) in genomic features, i.e., in promoters (black), exons (red), introns (green) and intergenic or TEs (blue) for the three methylation contexts, CpG, CHG and CHH. Methylation analysis are shown for the Loire population.

Figure 2: Genome evolution following ancestral polyploïdizations. **A**, Poplar genome from rosid ancestors and structural variations. Rosid ancestors with 7 reconstructed proto-chromosomes experienced a whole genome triplication (WGT) event to give rise to AEK21 with 21 proto-chromosomes. Salicales order then went through a second round of whole genome duplication (WGD) events followed by structural rearrangements ending with 19 chromosomes in poplar. Ancestral gene loss leads to different compartmentalization, LF (least-fractionnated, blue) and MF (most-fractionated, red). **B**, Evolutionary trajectory between LF (blue) and MF (red) genes in terms of methylation (promoter and gene body for CpG, CHG and CHH) and gene expression. Upper arrows indicated higher expression or methylation levels; Down arrows indicated lower expression or methylation levels; Horizontal arrows indicated no bias in expression or methylation levels. **C**, Differentially expressed genes (DEGs) and differentially methylated genes (DMGs) in all populations between duplicated gene pairs from the Salicales specific WGD. Blue for upregulated / hypermethylated LF-genes; Red for downregulated / hypomethylated genes in MF; Grey for no DEGs or DMGs between the duplicated genes in LF and MF. LogFC for log fold change. DEGs and DMGs are considered when FDR < 0.05 and logFC > 1.

Figure 3: Phylo-epi/genomics in *Populus nigra*. **A**, Pan methylome of the ten natural populations in CpG, CHG and CHH contexts. The intersection size represents the number of common features between populations and the set size (horizontal bars, in right), represents the number of methylated features (genes and TEs) in each population. **B**, Phylo-epigenomic and phylogenetic tree reconstruction for DNA methylation (CpG) and genetic (SNPs) markers. The epigenetic trees were realized with genome-wide SMPs filtered with SNPs data, coverage (>7X) and tolerating 30% of missing data. The genetic tree was done with all the genome-wide SNPs without any missing value and with a minor allele frequency above 5%.

Figure 4: Identification of genomic markers of local adaptation and/or drift. **A**, Comparison between the phylo-epigenomic tree in CpG built on differentially SMPs between the three sub-groups and the phylogenetic tree. **B**, Manhattan plots of obtained epigenetic pcadapt gene markers in CpG context with adjusted p-values (FDR, blue line). Markers above the blue line (FDR = 0.05) are considered as putative markers of local adaptation. **C**, WordCloud of identified putative genes involved in local adaptation. **D**, CpG Methylation diversity of disease resistance (R) genes in the 20 black poplar trees. The color code represents the population groups.

Figure 5: Regulation of gene expression by DNA methylation. **A**, Relationship between promoter DNA methylation in CpG context and gene expression. Methylation data are splitted into 10 quantiles with each quantile capturing 10% of the methylation level. Expression data are shown as the mean expression of genes in log2. **B**, Relationship between gene body DNA methylation in CpG context and gene expression. Methylation data are splitted into 10 quantiles with each quantile capturing 10% of the methylation level. Expression data are shown as the mean expression of genes in log2. **C**, Identification of outlier genes between promoter DNA methylation and gene expression (i.e. Hypo/Up and Hyper/Down). Methylation (x axis) is represented by quantiles with each quantile capturing 10% of the methylation level and gene expression (y axis) by logarithm. Red for CpG context, green for CHG context and blue for CHH context. Hypo/Up for hypomethylated and overexpressed genes and Hyper/Down for hypermethylated and downregulated genes. **D**, GO annotation using biological process terms

for Hypo/Up genes. **E**, Functional annotations for Hyper/Down genes using biological process terms.

AUTHOR CONTRIBUTIONS

MDS and JS wrote the manuscript. All authors edited and helped to improve the manuscript; VB, CBuret, MCLD, ILJ and AD performed the sampling; DNA and RNA extraction were realized by VB, CBuret, MCLD, AD and MDS; Genomic data production was realized by JT, CD, CBesse; Bioinformatic and statistical analysis were done by JT, AG, OR, ASR, ILK, SM, EM, CA, VS, ES and MDS; SM and JS assumed the coordination; All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

We are grateful to the to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) and the Mesocentre Clermont Auvergne bioinformatics platform (https://doi.org/10.18145/aubi) for providing computing and storage resources and to the GBFOR, INRAE, 2018, Forest Genetics and Biomass Facility (https://doi.org/10.15454/1.5572308287502317E12) for the experimental design setup and samples collection. We thank COST action (European Cooperation in Science and Technology) EPIgenetic mechanisms of Crop Adaptation To Climate cHange (EPICATCH; grant number CA19125) for active discussion.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The raw data for WGS, WGBS and RNAseq are stored in the NCBI website under the following accession numbers PRJNA818172 BioProject (WGS), PRJNA828400 BioProject (WGBS), GSE128482 (RNA-seq). The processed SNPs, methylation matrices, LF/MF, duplicated genes, pcadapt, hypo/up and Hyper/Down candidate genes can be found at: https://entrepot.recherche.data.gouv.fr/privateurl.xhtml?token=0d8bacbb-71aa-4352-9da4-cc6ed16a6190.

FUNDING

The current publication benefitted of fundings from the ANR (EPITREE ANR-17-CE32-0009-01), the ‘Région Auvergne-Rhône-Alpes’ and FEDER 'Fonds Européen de Développement Régional' (#23000816 project SRESRI 2015).

SUPPORTING INFORMATION

Figure S1. Global DNA methylation percentage distribution between the 20 *P. nigra* genotypes in CpG, CHG and CHH contexts.
Figure S2. DNA Methylation coverage on genes and TEs in the CpG, CHG and CHH contexts.
Figure S3. Gene ontology annotation using biological process terms of LF, MF and duplicated genes.
Figure S4. Venn diagram of differentially expressed genes (DEGs) and differentially methylated genes (DMGs) between the duplicated genes.
Figure S5. Phylo-epigenomic tree reconstruction for DNA methylation in CHG and CHH contexts.
Figure S6. Methylation markers involved in local adaptation.
Figure S7. DNA methylation dynamics of disease resistance (R) genes in the 20 black poplar trees.
Figure S8. Comparison between gene expression and DNA methylation.
Table S1. Mapping and methylation statistics for 20 black *P. nigra* genotypes.

REFERENCES

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655–1664.

Alger, E.I. and Edger, P.P. (2020). One subgenome to rule them all: underlying mechanisms of subgenome dominance. Current Opinion in Plant Biology 54: 108–113.

Allen, C.D. et al. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. Forest Ecology and Management 259: 660–684.

Alvarez, M., Bleich, A., and Donohue, K. (2021). Genetic differences in the temporal and environmental stability of transgenerational environmental effects. Evolution 75: 2773–2790.

Amaral J, Ribeyre Z, Vigneaud J, Sow MD, Fichot R, Messier C, Pinto G, Nolet P, Maury S (2020) Advances and Promises of Epigenetics for Forest Trees. *Forests* 11(9): 976

Anderegg, W.R.L., Klein, T., Bartlett, M., Sack, L., Pellegrini, A.F.A., Choat, B., and Jansen, S. (2016). Meta-analysis reveals that hydraulic traits explain cross-species patterns of drought-induced tree mortality across the globe. Proceedings of the National Academy of Sciences 113: 5024–5029.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Avci, U., Earl Petzold, H., Ismail, I.O., Beers, E.P., and Haigler, C.H. (2008). Cysteine proteases XCP1 and XCP2 aid micro-autolysis within the intact central vacuole during xylogenesis in Arabidopsis roots. The Plant Journal 56: 303–315.

Ball, M.P., Li, J.B., Gao, Y., Lee, J.-H., LeProust, E.M., Park, I.-H., Xie, B., Daley, G.Q., and Church, G.M. (2009). Targeted and genome-scale strategies reveal gene body methylation signatures in human cells. Nat Biotechnol 27: 361–368.

Bartels, A., Han, Q., Nair, P., Stacey, L., Gaynier, H., Mosley, M., Huang, Q., Pearson, J., Hsieh, T.-F., An, Y.-Q., and Xiao, W. (2018). Dynamic DNA Methylation in Plant Growth and Development. IJMS 19: 2144.

Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K., and Weigel, D. (2011). Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature 480: 245–249.

Bellec, A., Sow, M.D., Pont, C., Civan, P., Mardoc, E., Duchemin, W., Armisen, D., Huneau, C., Thévenin, J., Vernoud, V., Depège-Fargeix, N., Maunas, L., Escale, B., Dubreucq, B., Rogowsky, P., Bergès, H., Salse, J (2023) Tracing 100 million years of grass genome evolutionary plasticity. Plant J. doi: 10.1111/tpj.16185

Bewick, A.J. and Schmitz, R.J. (2017). Gene body DNA methylation in plants. Curr Opin Plant Biol 36: 103–110.

Blanc, G. and Wolfe, K.H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16: 1679–1691.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Bräutigam, K. et al. (2013). Epigenetic regulation of adaptive responses of forest tree species to the environment. Ecol Evol 3: 399–415.

Bräutigam, K. and Cronk, Q. (2018). DNA Methylation and the Evolution of Developmental Complexity in Plants. Front. Plant Sci. 9: 1447.

Cao, Q., Feng, Y., Dai, X., Huang, L., Li, J., Tao, P., Crabbe, M.J.C., Zhang, T., and Qiao, Q. (2021). Dynamic Changes of DNA Methylation During Wild Strawberry (Fragaria nilgerrensis) Tissue Culture. Front. Plant Sci. 12: 765383.

Cattell, R.B. (1966). The Scree Test For The Number Of Factors. Multivariate Behavioral Research 1: 245–276.

Chateigner, A., Lesage-Descauses, M.-C., Rogier, O., Jorge, V., Leplé, J.-C., Brunaud, V., Roux, C.P.-L., Soubigou-Taconnat, L., Martin-Magniette, M.-L., Sanchez, L., and Segura, V. (2020). Gene expression predictions and networks in natural populations supports the omnigenic theory. BMC Genomics 21: 416.

Chen, X., Ge, X., Wang, J., Tan, C., King, G.J., and Liu, K. (2015). Genome-wide DNA methylation profiling by modified reduced representation bisulfite sequencing in Brassica

rapa suggests that epigenetic modifications play a key role in polyploid genome evolution. Front Plant Sci 6: 836.

Chen, Y., Pal, B., Visvader, J.E., and Smyth, G.K. (2017). Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR. F1000Res 6: 2055.

Cheng, F., Sun, C., Wu, J., Schnable, J., Woodhouse, M.R., Liang, J., Cai, C., Freeling, M., and Wang, X. (2016). Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa. New Phytol 211: 288–299.

Conde, D., Le Gac, A.-L., Perales, M., Dervinis, C., Kirst, M., Maury, S., González-Melendi, P., and Allona, I. (2017). Chilling-responsive DEMETER-LIKE DNA demethylase mediates in poplar bud break. Plant Cell Environ 40: 2236–2249.

Cortijo, S. et al. (2014). Mapping the epigenetic basis of complex traits. Science 343: 1145–1148.

Dai, X. et al. (2014). The willow genome and divergent evolution from poplar after the common genome duplication. Cell Res 24: 1274–1277.

Dai, X., Zhai, R., Lin, J. et al. Cell-type-specific PtrWOX4a and PtrVCS2 form a regulatory nexus with a histone modification system for stem cambium development in Populus trichocarpa. Nat. Plants 9, 96–111 (2023).

Danecek, P. et al. (2011). The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience 10: giab008.

Dangl, J.L. and Jones, J.D.G. (2001). Plant pathogens and integrated defence responses to infection. Nature 411: 826–833.

Daviaud, C., Renault, V., Mauger, F., Deleuze, J.-F., and Tost, J. (2018). Whole-Genome Bisulfite Sequencing Using the Ovation® Ultralow Methyl-Seq Protocol. In DNA Methylation Protocols, J. Tost, ed, Methods in Molecular Biology. (Springer: New York, NY), pp. 83–104.

Davis, A.P., Benninghoff, A.D., Thomas, A.J., Sessions, B.R., and White, K.L. (2015). DNA methylation of the LIN28 pseudogene family. BMC Genomics 16: 287.

Doyle, J.J. and Doyle, J.L. eds A rapid DNA isolation procedure for small quantities of fresh leaf tissue. PHYTOCHEMICAL BULLETIN.

Dubin, M.J. et al. (2015). DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife 4: e05255.

Dugé de Bernonville, T., Daviaud, C., Chaparro, C., Tost, J., and Maury, S. (2022). From Methylome to Integrative Analysis of Tissue Specificity. In Catharanthus roseus: Methods and Protocols, V. Courdavault and S. Besseau, eds, Methods in Molecular Biology. (Springer US: New York, NY), pp. 223–240.

El Baidouri, M., Kim, K.D., Abernathy, B., Li, Y.-H., Qiu, L.-J., and Jackson, S.A. (2018). Genic C-Methylation in Soybean Is Associated with Gene Paralogs Relocated to Transposable Element-Rich Pericentromeres. Mol Plant 11: 485–495.

Endo, S., Pesquet, E., Yamaguchi, M., Tashiro, G., Sato, M., Toyooka, K., Nishikubo, N., Udagawa-Motose, M., Kubo, M., Fukuda, H., and Demura, T. (2009). Identifying New Components Participating in the Secondary Cell Wall Formation of Vessel Elements in Zinnia and Arabidopsis. The Plant Cell 21: 1155–1165.

Etchells, J.P., Provost, C.M., Mishra, L., and Turner, S.R. (2013). WOX4 and WOX14 act downstream of the PXY receptor kinase to regulate plant vascular proliferation independently of any role in vascular organisation. Development 140: 2224–2234.

Etchells, J.P., Smit, M.E., Gaudinier, A., Williams, C.J., and Brady, S.M. (2016). A brief history of the TDIF-PXY signalling module: balancing meristem identity and differentiation during vascular development. New Phytologist 209: 474–484.

Faivre-Rampant, P. et al. (2016). New resources for genetic studies in Populus nigra: genome-wide SNP discovery and development of a 12k Infinium array. Molecular Ecology Resources 16: 1023–1036.

Fischer, U., Kucukoglu, M., Helariutta, Y., and Bhalerao, R.P. (2019). The Dynamics of Cambial Stem Cell Activity. Annual Review of Plant Biology 70: 293–319.
Fisher, K. and Turner, S. (2007). PXY, a Receptor-like Kinase Essential for Maintaining Polarity during Plant Vascular-Tissue Development. Current Biology 17: 1061–1066.
Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol 60: 433–453.
Furrow, R.E. and Feldman, M.W. (2014). Genetic Variation and the Evolution of Epigenetic Regulation. Evolution 68: 673–683.
Guarino, F., Cicatelli, A., Brundu, G., Heinze, B., Castiglione, S. (2015). Epigenetic Diversity of Clonal White Poplar (*Populus alba* L.) Populations: Could Methylation Support the Success of Vegetative Reproduction Strategy? PLOS ONE 10(7): e0131480.
Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing.
Gebreselassie, M.N. et al. (2017). Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from Populus nigra (L.) natural populations. Industrial Crops and Products 107: 159–171.
Guet, J., Fabbrini, F., Fichot, R., Sabatti, M., Bastien, C., and Brignolas, F. (2015). Genetic variation for leaf morphology, leaf structure and leaf carbon isotope discrimination in European populations of black poplar (Populus nigra L.). Tree Physiology 35: 850–863.
Gugger, P.F., Fitz-Gibbon, S., PellEgrini, M., and Sork, V.L. (2016). Species-wide patterns of DNA methylation variation in Quercus lobata and their association with climate gradients. Mol Ecol 25: 1665–1680.
Hammond-Kosack, K.E. and Jones, J.D.G. (1997). Plant Disease Resistance Genes. Annual Review of Plant Physiology and Plant Molecular Biology 48: 575–607.
Hao, Y. et al. (2021). The contributions from the progenitor genomes of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible. Genome Res 31: 799–810.
Heckwolf, M.J., Meyer, B.S., Häsler, R., Höppner, M.P., Eizaguirre, C., and Reusch, T.B.H. (2020). Two different epigenetic information channels in wild three-spined sticklebacks are involved in salinity adaptation. Sci. Adv. 6: eaaz1138.
Holliday, R. and Grigg, G.W. (1993). DNA methylation and mutation. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 285: 61–67.
Hu, J., Hu, X., Yang, Y., He, C., Hu, J., and Wang, X. (2022). Strigolactone signaling regulates cambial activity through repression of WOX4 by transcription factor BES1. Plant Physiology 188: 255–267.
Jablonka, E. (2017). The evolutionary implications of epigenetic inheritance. Interface Focus. 7: 20160135.
Jaillon, O. et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463–467.
Johannes, F. (2019). DNA methylation makes mutational history. Nat. Plants 5: 772–773.
Joshi, C.J., Ke, W., Drangowska-Way, A., O'Rourke, E.J., and Lewis, N.E. (2022). What are housekeeping genes? PLOS Computational Biology 18: e1010295.
Kakoulidou, I. et al. (2021). Epigenetics for Crop Improvement in Times of Global Change. Biology 10: 766.
Keller, T.E. and Yi, S.V. (2014). DNA methylation and evolution of duplicate genes. Proc Natl Acad Sci U S A 111: 5932–5937.
Khan, A. and Korban, S.S. (2022). Breeding and genetics of disease resistance in temperate fruit trees: challenges and new opportunities. Theor Appl Genet.
Kon, T. and Yoshikawa, N. (2014). Induction and maintenance of DNA methylation in plant promoter sequences by apple latent spherical virus-induced transcriptional gene silencing. Front. Microbiol. 5.
Kong, W., Xia, X., Wang, Q., Liu, L.-W., Zhang, S., Ding, L., Liu, A., and La, H. (2020). Impact of DNA Demethylases on the DNA Methylation and Transcription of Arabidopsis NLR Genes. Frontiers in Genetics 11.
Lafon-Placette, C. et al. (2018). Changes in the epigenome and transcriptome of the poplar

shoot apical meristem in response to water availability affect preferentially hormone pathways. J Exp Bot 69: 537–551.

Lafon-Placette, C., Faivre-Rampant, P., Delaunay, A., Street, N., Brignolas, F., and Maury, S. (2013). Methylome of DNase I sensitive chromatin in Populus trichocarpa shoot apical meristematic cells: a simplified approach revealing characteristics of gene body DNA methylation in open chromatin state. New Phytol 197: 416–430.

Lamka, G.F., Harder, A.M., Sundaram, M., Schwartz, T.S., Christie, M.R., DeWoody, J.A., and Willoughby, J.R. (2022). Epigenetics in Ecology, Evolution, and Conservation. Front. Ecol. Evol. 10: 871791.

Lang, D. et al. (2018). The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. Plant J 93: 515–533.

Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359.

Le, T.-N., Schumann, U., Smith, N.A., Tiwari, S., Au, P.C.K., Zhu, Q.-H., Taylor, J.M., Kazan, K., Llewellyn, D.J., Zhang, R., Dennis, E.S., and Wang, M.-B. (2014). DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in Arabidopsis. Genome Biology 15: 458.

Le Gac, Al. et al. (2018). Winter-dormant shoot apical meristem in poplar trees shows environmental epigenetic memory. Journal of experimental botany 69.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, Q., Eichten, S.R., Hermanson, P.J., and Springer, N.M. (2014). Inheritance Patterns and Stability of DNA Methylation Variation in Maize Near-Isogenic Lines. Genetics 196: 667–676.

Li, X. et al. (2012). Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. BMC Genomics 13: 300.

Lin, W., Sun, L., Huang, R.-Z., Liang, W., Liu, X., He, H., Fukuda, H., He, X.-Q., and Qian, W. (2020). Active DNA demethylation regulates tracheary element differentiation in Arabidopsis. Science Advances 6: eaaz2963.

Liu, Y., Huang, R., Liu, Y., Song, W., Wang, Y., Yang, Y., Dong, S., and Yang, X. (2018). Insights from multidimensional analyses of the pan-cancer DNA methylome heterogeneity and the uncanonical CpG-gene associations. Int J Cancer 143: 2814–2827.

Liu, Y., Wang, J., Ge, W., Wang, Z., Li, Y., Yang, N., Sun, S., Zhang, L., and Wang, X. (2017). Two Highly Similar Poplar Paleo-subgenomes Suggest an Autotetraploid Ancestor of Salicaceae Plants. Front Plant Sci 8: 571.

Luu, K., Bazin, E., and Blum, M.G.B. (2017). pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour 17: 67–77.

Ma, M., Chen, X., Yin, Y., Fan, R., Li, B., Zhan, Y., and Zeng, F. (2020). DNA Methylation Silences Exogenous Gene Expression in Transgenic Birch Progeny. Front. Plant Sci. 11: 523748.

Mader, M., Paslier, M.-C.L., Bounon, R., Bérard, A., Rampant, P.F., Fladung, M., Leplé, J.-C., and Kersten, B. (2016). Whole-genome draft assembly of x clone INRA 717-1B4. Silvae Genetica 65: 74–79.

Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature 456: 18–21.

Maunakea, A.K. et al. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466: 253–257.

Maury, S., Sow, M.D., Le Gac, A.-L., Genitoni, J., Lafon-Placette, C., and Mozgova, I. (2019). Phytohormone and Chromatin Crosstalk: The Missing Link For Developmental Plasticity? Frontiers in Plant Science 10.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing

data. Genome Res 20: 1297–1303.
Messer, P.W., Ellner, S.P., and Hairston, N.G. (2016). Can Population Genetics Adapt to Rapid Evolution? Trends in Genetics 32: 408–418.
Murat, F., Armero, A., Pont, C., Klopp, C., and Salse, J. (2017). Reconstructing the genome of the most recent common ancestor of flowering plants. Nat Genet 49: 490–496.
Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., Crollius, H.R., and Salse, J. (2015). Understanding Brassicaceae evolution through ancestral genome reconstruction. Genome Biol 16: 262.
Murat, F., Peer, Y.V. de, and Salse, J. (2012). Decoding Plant and Animal Genome Plasticity from Differential Paleo-Evolutionary Patterns and Processes. Genome Biology and Evolution 4: 917–928.
Nakaba, S., Takata, N., Yoshida, M., and Funada, R. (2015). Continuous expression of genes for xylem cysteine peptidases in long-lived ray parenchyma cells in Populus. Plant Biotechnology 32: 21–29.
Niederhuth, C.E. et al. (2016). Widespread natural variation of DNA methylation within angiosperms. Genome Biol 17: 194.
Niederhuth, C.E. and Schmitz, R.J. (2017). Putting DNA methylation in context: from genomes to gene expression in plants. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1860: 149–156.
Noir, S., Combes, M.C., Anthony, F., and Lashermes, P. (2001). Origin, diversity and evolution of NBS-type disease-resistance gene homologues in coffee trees (Coffea L.). Mol Genet Genomics 265: 654–662.
Noshay, J.M. and Springer, N.M. (2021). Stories that can't be told by SNPs; DNA methylation variation in plant populations. Current Opinion in Plant Biology 61: 101989.
Nuo, M.T., Yuan, J.L., Yang, W.L., Gao, X.Y., He, N., Liang, H., Cang, M., and Liu, D.J. (2016). Promoter methylation and histone modifications affect the expression of the exogenous DsRed gene in transgenic goats. Genet. Mol. Res. 15.
Olinski, R., Slupphaug, G., Foksinski, M., and Krokan, H.E. (2021). Genomic Uracil and Aberrant Profile of Demethylation Intermediates in Epigenetics and Hematologic Malignancies. IJMS 22: 4212.
Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327: 92–94.
Platt, A., Gugger, P.F., Pellegrini, M., and Sork, V.L. (2015). Genome-wide signature of local adaptation linked to variable CpG methylation in oak populations. Mol Ecol 24: 3823–3830.
Plomion, C. et al. (2016). Forest tree genomics: 10 achievements from the past 10 years and future prospects. Annals of Forest Science 73: 77–103.
Plomion, C. et al. (2018). Oak genome reveals facets of long lifespan. Nat Plants 4: 440–452.
Pont, C. et al. (2013). Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. The Plant Journal 76: 1030–1044.
Pont, C. and Salse, J. (2017). Wheat paleohistory created asymmetrical genomic evolution. Current Opinion in Plant Biology 36: 29–37.
Rejab, N.A., Nakano, Y., Yoneda, A., Ohtani, M., and Demura, T. (2015). Possible contribution of TED6 and TED7, secondary cell wall-related membrane proteins, to evolution of tracheary element in angiosperm lineage. Plant Biotechnology 32: 343–347.
Renny-Byfield, S., Rodgers-Melnick, E., and Ross-Ibarra, J. (2017). Gene Fractionation and Function in the Ancient Subgenomes of Maize. Mol Biol Evol 34: 1825–1832.
Ribas, A.F., Cenci, A., Combes, M.-C., Etienne, H., and Lashermes, P. (2011). Organization and molecular evolution of a disease-resistance gene cluster in coffee trees. BMC Genomics 12: 240.
Robinson, M.D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11: R25.
Rogier, O., Chateigner, A., Amanzougarene, S., Lesage-Descauses, M.-C., Balzergue, S.,

Brunaud, V., Caius, J., Soubigou-Taconnat, L., Jorge, V., and Segura, V. (2018). Accuracy of RNAseq based SNP discovery and genotyping in Populusnigra. BMC Genomics 19: 909.

Sammarco, I., Münzbergová, Z., and Latzel, V. (2022). DNA Methylation Can Mediate Local Adaptation and Response to Climate Change in the Clonal Plant Fragaria vesca: Evidence From a European-Scale Reciprocal Transplant Experiment. Front. Plant Sci. 13: 827166.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proceedings of the National Academy of Sciences 108: 4069–4074.

Schumann, U., Lee, J.M., Smith, N.A., Zhong, C., Zhu, J.-K., Dennis, E.S., Millar, A.A., and Wang, M.-B. (2019). DEMETER plays a role in DNA demethylation and disease response in somatic tissues of Arabidopsis. Epigenetics 14: 1074–1087.

Shimizu, D. et al. (2022). Pan-cancer methylome analysis for cancer diagnosis and classification of cancer cell of origin. Cancer Gene Ther 29: 428–436.

Sieburth, L.E. (2007). Plant Development: PXY and Polar Cell Division in the Procambium. Current Biology 17: R594–R596.

Soltis, P.S., Marchant, D.B., Van de Peer, Y., and Soltis, D.E. (2015). Polyploidy and genome evolution in plants. Current Opinion in Genetics & Development 35: 119–125.

Sow, M.D. et al. (2018a). Chapter Twelve - Epigenetics in Forest Trees: State of the Art and Potential Implications for Breeding and Management in a Context of Climate Change. In Advances in Botanical Research, M. Mirouze, E. Bucher, and P. Gallusci, eds, Plant Epigenetics Coming of Age for Breeding Applications. (Academic Press), pp. 387–453.

Sow, M.D. et al. (2021). RNAi suppression of DNA methylation affects the drought stress response and genome integrity in transgenic poplar. New Phytologist 232: 80–97.

Sow, M.D., Segura, V., Chamaillard, S., Jorge, V., Delaunay, A., Lafon-Placette, C., Fichot, R., Faivre-Rampant, P., Villar, M., Brignolas, F., and Maury, S. (2018b). Narrow-sense heritability and PST estimates of DNA methylation in three Populus nigra L. populations under contrasting water availability. Tree Genetics & Genomes 14: 78.

Throude, M. et al. (2009). Structure and expression analysis of rice paleo duplications. Nucleic Acids Res 37: 1248–1259.

Tuskan, G.A. et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313: 1596–1604.

van der Graaf, A., Wardenaar, R., Neumann, D.A., Taudt, A., Shaw, R.G., Jansen, R.C., Schmitz, R.J., Colomé-Tatché, M., and Johannes, F. (2015). Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proceedings of the National Academy of Sciences 112: 6676–6681.

Van de Peer, Y., Ashman, T.-L., Soltis, P.S., and Soltis, D.E. (2021). Polyploidy: an evolutionary and ecological force in stressful times. The Plant Cell 33: 11–26.

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. Journal of Dairy Science 91: 4414–4423.

Vigneaud, J., Kohler, A., Sow, M.D., Delaunay, A., Fauchery, L., Guinet, F., Daviaud, C., Barry, K.W., Keymanesh, K., Johnson, J., Singan, V., Grigoriev, I., Fichot, R., Conde, D., Perales, M., Tost, J., Martin, F.M., Allona, I., Strauss, S.H., Veneault-Fourrey, C. and Maury, S. (2023), DNA hypomethylation of the host tree impairs interaction with mutualistic ectomycorrhizal fungus. New Phytol, 238: 2561-2577.

Wang, H., Beyene, G., Zhai, J., Feng, S., Fahlgren, N., Taylor, N.J., Bart, R., Carrington, J.C., Jacobsen, S.E., and Ausin, I. (2015). CG gene body DNA methylation changes and evolution of duplicated genes in cassava. Proc Natl Acad Sci U S A 112: 13729–13734.

Wang, L., Ko, E.E., Tran, J., and Qiao, H. (2020). TREE1-EIN3–mediated transcriptional repression inhibits shoot growth in response to ethylene. Proceedings of the National Academy of Sciences 117: 29178–29189.

Wang, X., Zhang, Z., Fu, T., Hu, L., Xu, C., Gong, L., Wendel, J.F., and Liu, B. (2017). gene body CG methylation and divergent expression of duplicate genes in rice. Sci Rep 7: 2675.

Weinhold, A., Kallenbach, M., and Baldwin, I.T. (2013). Progressive 35S promoter methylation increases rapidly during vegetative development in transgenic Nicotiana attenuata plants. BMC Plant Biol 13: 99.

Wendel, J.F. (2015). The wondrous cycles of polyploidy in plants. Am J Bot 102: 1753–1756.

Woodhouse, M.R., Cheng, F., Pires, J.C., Lisch, D., Freeling, M., and Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. Proc Natl Acad Sci U S A 111: 5283–5288.

Xu, G., Guo, C., Shan, H., and Kong, H. (2012). Divergence of duplicate genes in exon–intron structure. Proc. Natl. Acad. Sci. U.S.A. 109: 1187–1192.

Ye, Z.-H. and Zhong, R. (2015). Molecular control of wood formation in trees. Journal of Experimental Botany 66: 4119–4131.

Yim, W.C., Lee, B.-M., and Jang, C.S. (2009). Expression diversity and evolutionary dynamics of rice duplicate genes. Mol Genet Genomics 281: 495–495.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328: 916–919.

Zeng, W., Huang, H., Lin, X., Zhu, C., Kosami, K.-I., Huang, C., Zhang, H., Duan, C.-G., Zhu, J.-K., and Miki, D. (2021). Roles of DEMETER in regulating DNA methylation in vegetative tissues and pathogen resistance. J Integr Plant Biol 63: 691–706.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.-L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E., and Ecker, J.R. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126: 1189–1201.

Zhang, Y., Liu, C., Cheng, H., Tian, S., Liu, Y., Wang, S., Zhang, H., Saqib, M., Wei, H., and Wei, Z. (2020). DNA methylation and its effects on gene expression during primary to secondary growth in poplar stems. BMC Genomics 21: 498.

Zhao, M., Zhang, B., Lisch, D., and Ma, J. (2017). Patterns and Consequences of Subgenome Differentiation Provide Insights into the Nature of Paleopolyploidy in Plants. Plant Cell 29: 2974–2994.

Zhou, Y., He, F., Pu, W., Gu, X., Wang, J., and Su, Z. (2020). The Impact of DNA Methylation Dynamics on the Mutation Rate During Human Germline Development. G3 Genes|Genomes|Genetics 10: 3337–3346.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 10: 1523.

Zhu, R., Shevchenko, O., Ma, C., Maury, S., Freitag, M., and Strauss, S.H. (2013). Poplars with a PtDDM1-RNAi transgene have reduced DNA methylation and show aberrant post-dormancy morphology. Planta 237: 1483–1493.

Figure 1

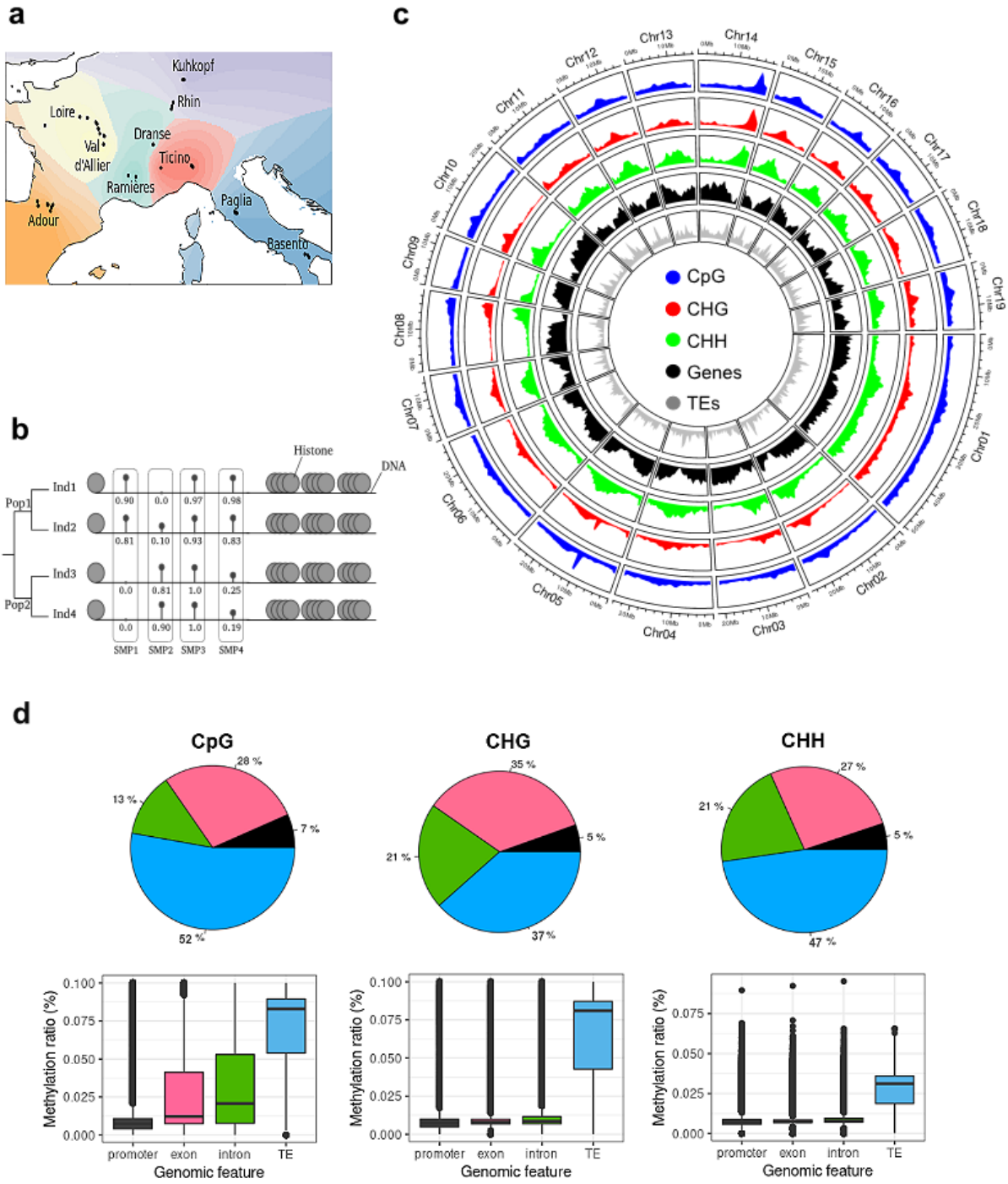


Figure 2

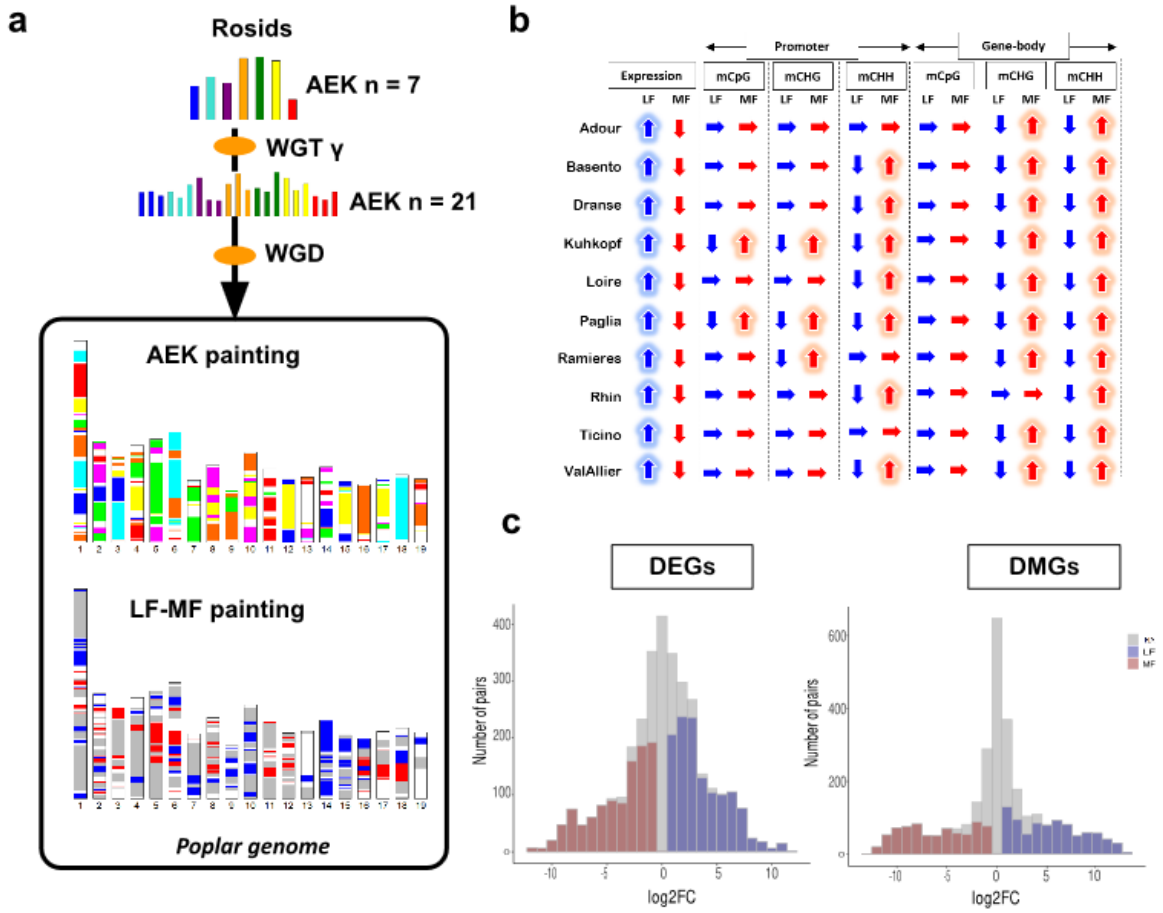


Figure 3

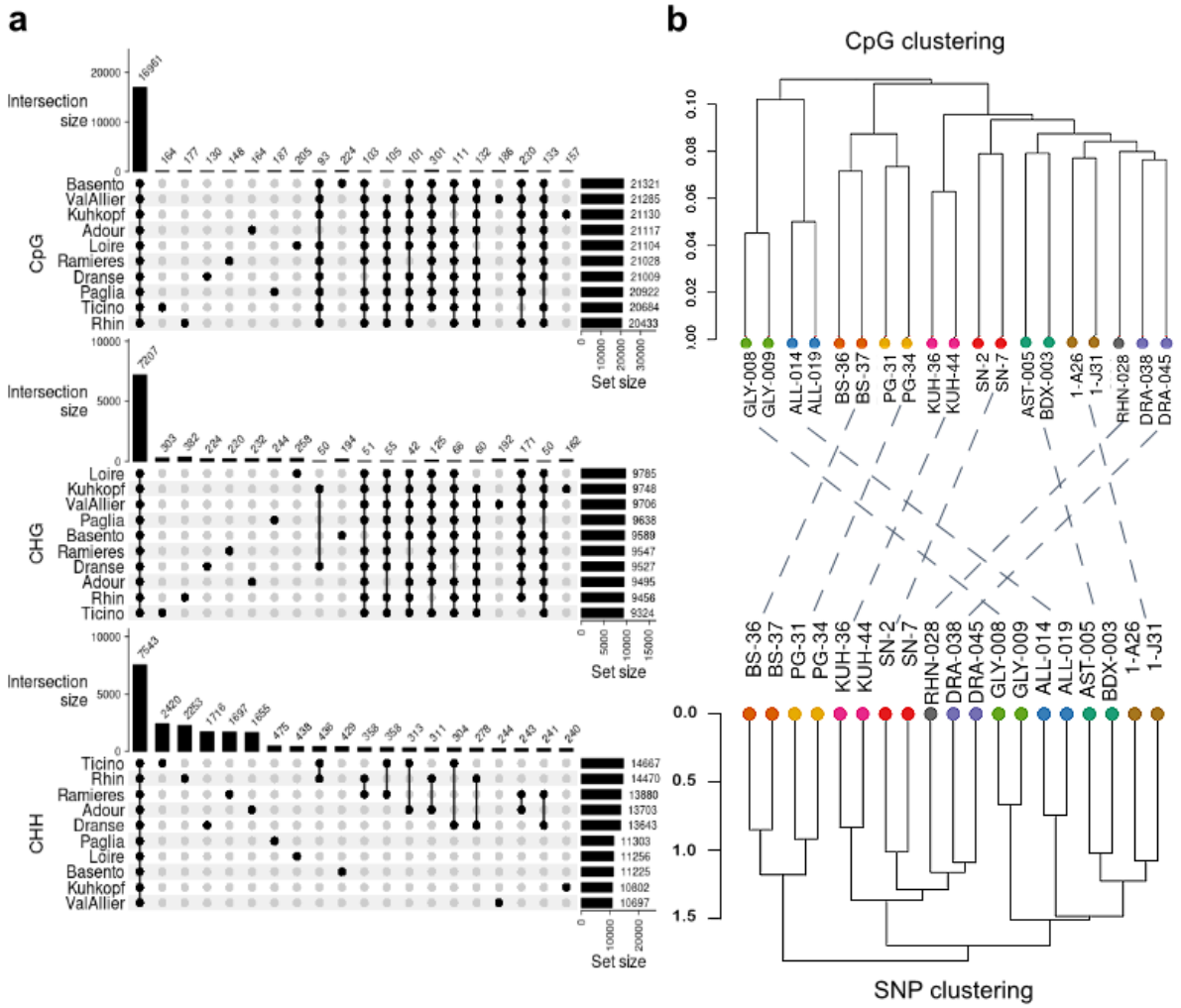


Figure 4

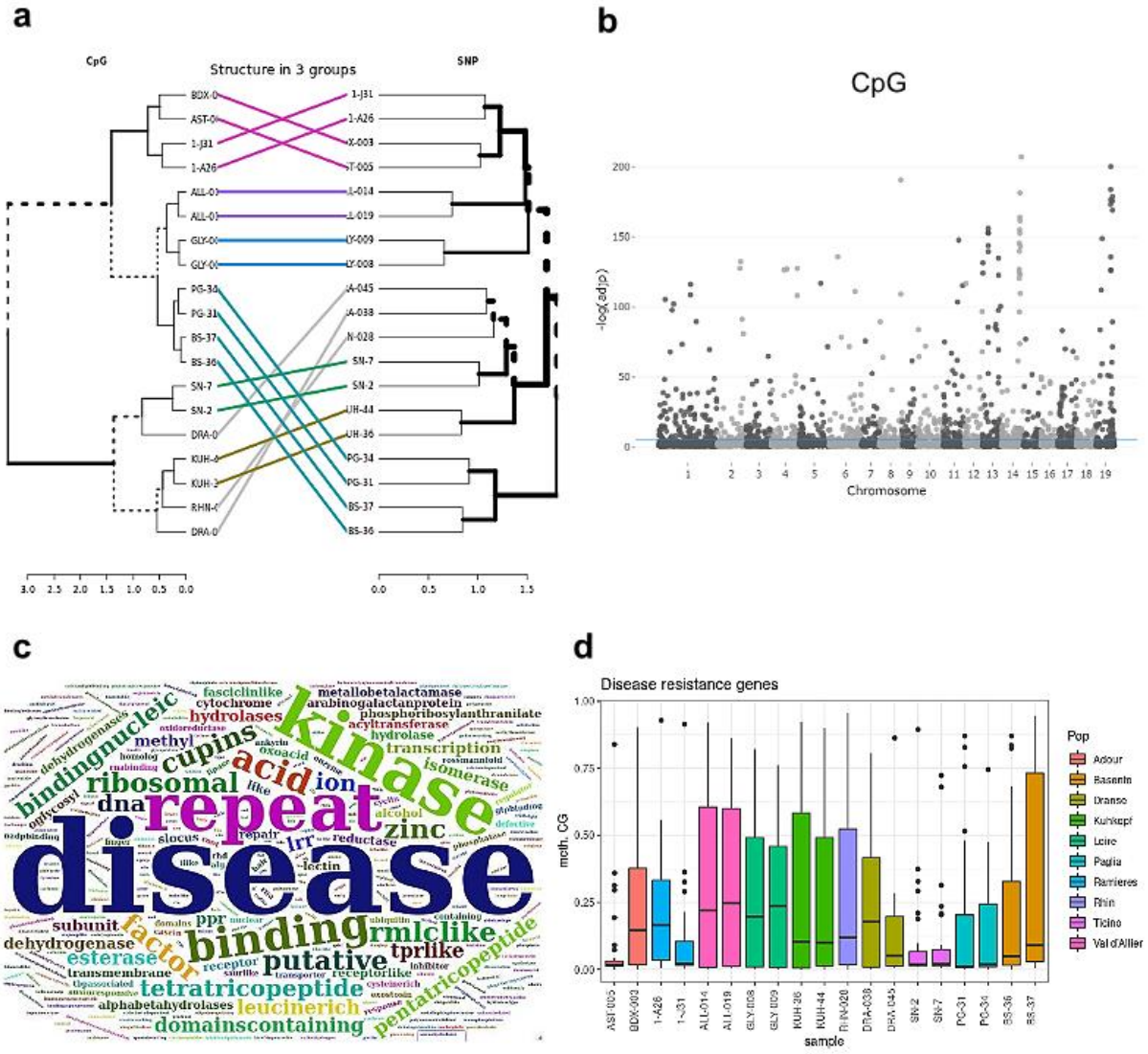
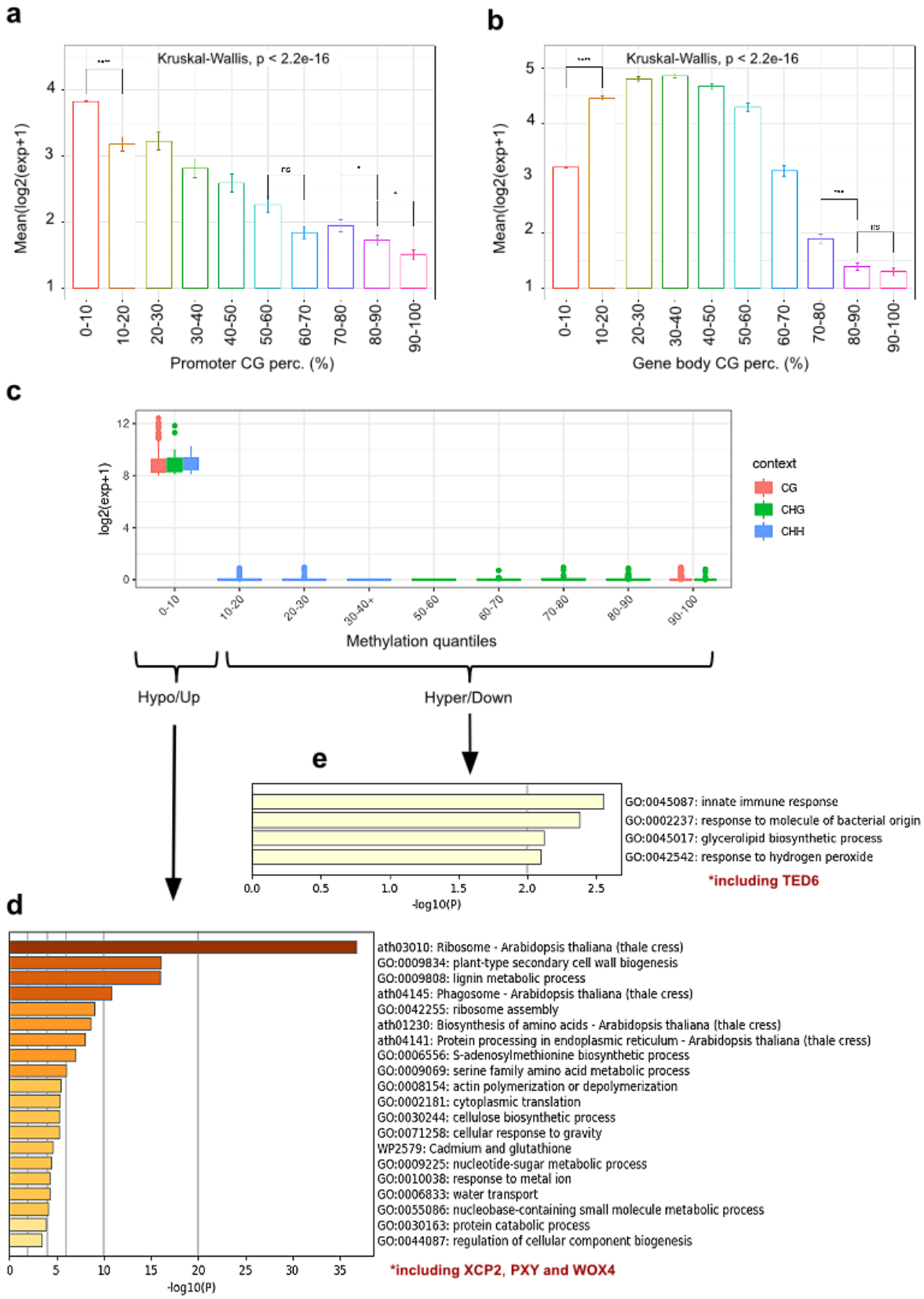


Figure 5



Annexe C

Sow et al. (2023), Données supplémentaires

Les Jeux de Données supplémentaires sont disponibles au lien suivant :
https://entrepot.recherche.data.gouv.fr/privateurl.xhtml?token=0d8bacbb-71aa-4352-9da4-cc6ed16a6190

##Supplementary Table S1: Mapping and methylation statistics for 20 black *P. nigra* genotypes of ten natural populations. Reads are mapped using BSMAP mapper Galaxy v1.0.0) using default options. Total reads and methylation percentage are shown for each methylation context (CpG, CHG and CHH).

Population	Genotype	Barcode	Aligned pairs (%)	Unique pairs (%)	Aligned pairs reads	Unique pairs reads	Coverage (X)	Non-conversion rate (%)	Total positions	CpG positions	CHG positions	CHH positions	Mean CpG methylation (%)	Mean CHG methylation (%)	Mean CHH methylation (%)
ValAllier	ALL-19	B00FLQO	32.0	25.6	48459339	38736049	21.52	0.5	110051871	9365382	15471155	85215333	34.4	21.0	4.8
	ALL-14	B00FLQP	34.8	28.1	59641958	48031616	26.48	0.4	111093134	9463906	15597276	86031951	34.9	21.3	5.1
Kuhkopf	KUH-36	B00FLQQ	35.6	28.3	58227706	46209807	25.85	0.5	110589215	9420230	15539692	85629292	34.7	21.0	5.1
	KUH-44	B00FLQR	32.9	26.5	56654421	45677245	25.15	0.4	110994738	9456256	15589122	85949359	34.1	20.7	4.4
Basento	BS-36	B00FLQS	36.7	28.6	57957234	46504399	25.73	0.4	110654162	9404047	15532685	85717429	34.8	21.0	4.8
	BS-37	B00FLQT	36.4	29.2	59126821	47501943	26.25	0.6	110678400	9420680	15552762	85704957	35.3	21.5	5.5
Loire	GLY-08	B00FLQU	30.4	24.0	48492560	38360749	21.53	0.5	110435643	9410385	15528524	85496733	34.1	21.3	5.1
	GLY-09	B00FLQV	26.0	20.6	41685669	32955624	18.51	0.5	110029626	9368122	15470208	85191295	33.7	20.7	4.7
Paglia	PG-31	B00FLQW	31.7	25.5	50724833	40852308	22.52	0.4	110148113	9363473	15475161	85309478	34.0	20.6	4.8
	PG-34	B00FLQX	16.4	12.8	27422970	21476693	12.18	0.6	108447931	9207881	15257684	83982365	32.1	20.0	5.0
Ramieres	1-A26	B00J70X	38.8	30.6	59055168	46634747	26.22	1.1	110590914	9412081	15527082	85651750	35.4	21.9	5.6
	1-J31	B00J70Y	37.5	30.0	43852288	34992247	19.47	1.0	108558265	9207207	15269020	84082037	35.2	21.6	6.5
Adour	AST-05	B00J70Z	37.4	29.8	52480388	41820542	23.30	1.0	109489505	9292260	15380077	84817167	35.1	21.6	5.5
	BDX-03	B00J710	37.2	29.8	46630447	37372882	20.70	1.0	108832863	9227188	15291025	84314649	35.0	21.6	5.6
Dranse	DRA-38	B00J711	37.6	29.9	49543346	39361060	22.00	1.0	109124626	9261507	15337309	84525809	35.3	21.7	5.3
	DRA-45	B00J712	38.2	30.4	44344812	35367413	19.69	1.0	108476077	9198690	15254777	84022609	35.1	21.6	5.2
Rhin	RHN-28	B00J713	37.4	29.9	45716058	36604476	20.30	1.0	108874616	9240478	15307318	84326819	34.9	21.6	6.4
	STR-10	B00J714	11.9	8.7	14871752	10902349	6.60	1.0	removed	removed	removed	removed	removed	removed	removed
Ticino	SN-2	B00J715	37.5	29.6	41028320	32432888	18.22	1.0	106996920	9041444	15051597	82903878	34.6	20.9	4.4
	SN-7	B00J716	37.9	30.0	45613804	36088689	20.25	1.0	108024554	9155154	15195875	83673524	35.1	21.3	5.3

##Supplementary Table S2: Evolutionary trajectory between LF and MF genes in terms of methylation (promoter and gene body for CpG, CHG and CHH) and gene expression. Numbers represent p-values from the comparison between LF and MF genes. KS = Two-sample Kolmogorov-Smirnov test; AD = Anderson-Darling k-sample test; t-test = Welch Two Sample t-test; K-W = Kruskal-Wallis rank sum test

		Adour	Basento	Dranse	Kuhkopf	Loire	Paglia	Ramieres	Rhin	Ticino	ValAllier
Expression	**KS**	0.02531	0.04252	0.05447	0.01901	0.04011	0.02819	0.02494	0.04673	0.102	0.09425
	AD	0.001112	0.003499	0.002350	0.0007120	0.003618	0.0009231	0.004782	0.003247	0.009218	0.005196
	t-test	0.03427	0.03697	0.02666	0.05132	0.0262	0.02472	0.0685	0.02699	0.02173	0.01938
	K-W	0.002287	0.00463	0.00374	0.003454	0.006513	0.002562	0.006493	0.007288	0.01763	0.00889
Promoter CpG	**KS**	0.3942	0.2912	0.5678	0.04656	0.1821	0.2849	0.4418	0.4791	0.3641	0.3957
	AD	0.256	0.2220	0.3466	0.01629	0.09652	0.03920	0.2087	0.2790	0.2129	0.2543
	t-test	0.002437	0.001141	0.009647	0.0009331	0.001058	0.001787	0.002484	0.005326	0.02506	0.008024
	K-W	0.4422	0.3869	0.5967	0.02824	0.1508	0.06207	0.992	0.5144	0.1922	0.321
Promoter CHG	**KS**	0.249	0.1167	0.3898	0.05055	0.1629	0.04987	0.1379	0.4188	0.6378	0.2784
	AD	0.1953	0.06016	0.1194	0.01218	0.1064	0.06129	0.02415	0.5470	0.3017	0.08375
	t-test	0.003113	0.0003352	0.008752	6.164e-05	0.0007002	0.007217	0.001414	0.01985	0.08483	0.0006791
	K-W	0.761	0.659	0.1241	0.07999	0.7448	0.2918	0.03231	0.5547	0.5311	0.7391
Promoter CHH	**KS**	0.1523	0.1012	0.007568	0.004983	0.02874	0.04301	0.1259	0.2347	0.2269	0.1061
	AD	0.05994	0.03324	0.002010	0.0001683	0.02210	0.009265	0.06930	0.02492	0.1339	0.01483
	t-test	0.003179	0.0001089	0.0008604	0.000102	0.0001496	0.000824	0.0006013	0.002556	0.007603	0.001287
	K-W	0.2289	0.2538	0.00504	0.0004315	0.1816	0.03664	0.7324	0.03268	0.3505	0.03368
Gene-body CpG	**KS**	0.08439	0.06094	0.1366	0.3604	0.469	0.3028	0.3584	0.5598	0.4104	0.2159
	AD	0.03074	0.08075	0.05414	0.1136	0.2321	0.1606	0.1472	0.3140	0.1752	0.1516
	t-test	0.3032	0.3627	0.2697	0.1727	0.2778	0.1661	0.2372	0.3834	0.2425	0.4317
	K-W	0.2363	0.3893	0.3145	0.6954	0.6324	0.7217	0.7976	0.6524	0.9927	0.4451
Gene-body CHG	**KS**	0.004947	0.02702	0.07927	0.01245	0.006047	0.02987	0.05466	0.1399	0.01982	0.06332
	AD	0.006918	0.04044	0.02629	0.01473	0.009409	0.04237	0.01450	0.05874	0.01174	0.01474
	t-test	0.001805	0.007094	0.001943	0.0005014	0.007667	0.009549	0.009163	0.00878	0.003603	0.02393
	K-W	0.7914	0.4523	0.05	0.1089	0.1987	0.1451	0.02957	0.6333	0.06411	0.02697
Gene-body CHH	**KS**	0.006431	0.04405	0.01009	0.02856	0.009882	0.01306	0.01997	0.03527	0.06381	0.01583
	AD	0.001489	0.01231	0.004460	0.009262	0.004013	0.003156	0.01504	0.01400	0.01629	0.005221
	t-test	0.0004202	0.0007833	9.547e-05	7.41e-05	0.0001241	0.000127	0.0002788	0.002037	0.0003728	0.000521
	K-W	0.3702	0.2658	0.1068	0.05442	0.09332	0.01805	0.1851	0.2384	0.05517	0.02443

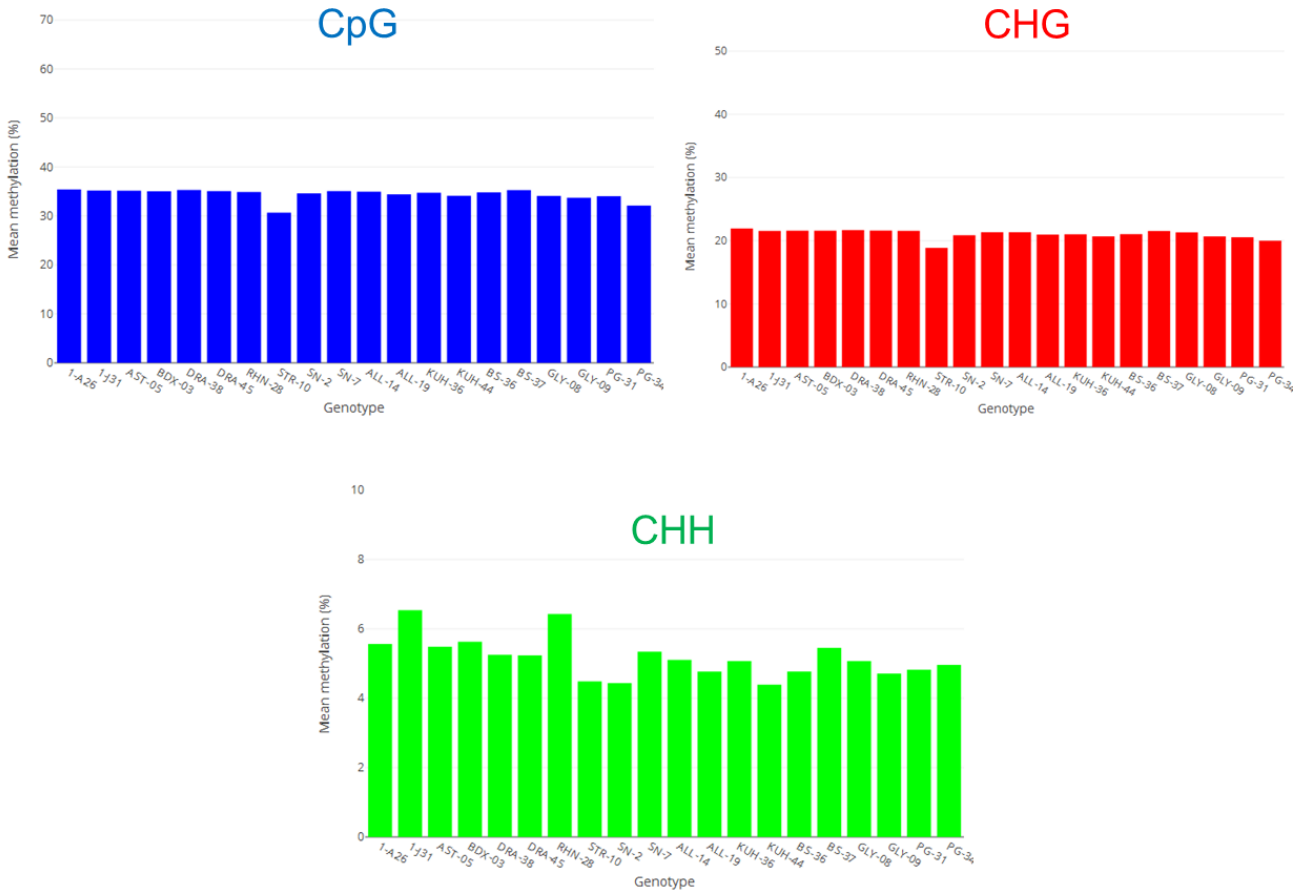


Figure S1 Global DNA methylation percentage distribution between the 20 *P. nigra* genotypes in CpG, CHG and CHH contexts.

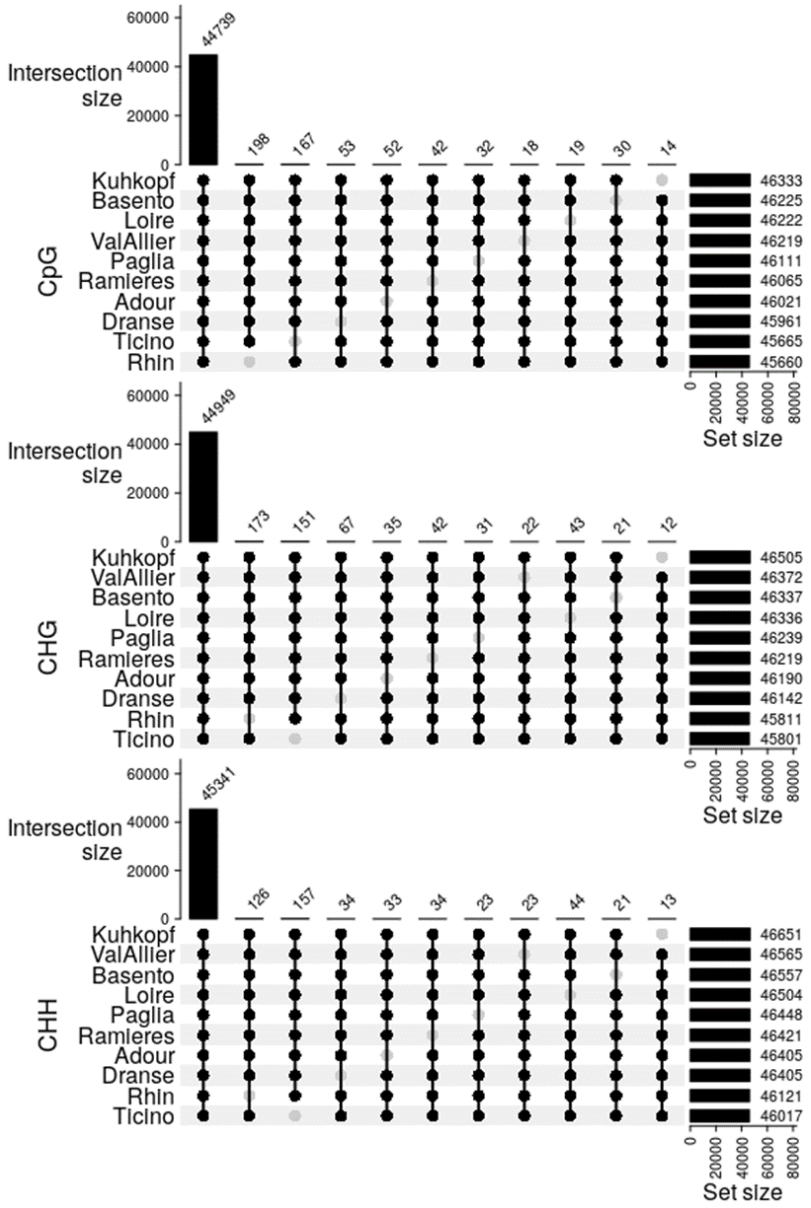


Figure S2 DNA Methylation coverage on genes and TEs in the CpG, CHG and CHH contexts. The intersection size represents the number of common features (genes and TEs) between populations and the set size (horizontal bars, in right) represents the number of methylated features (genes and TEs) in each population.

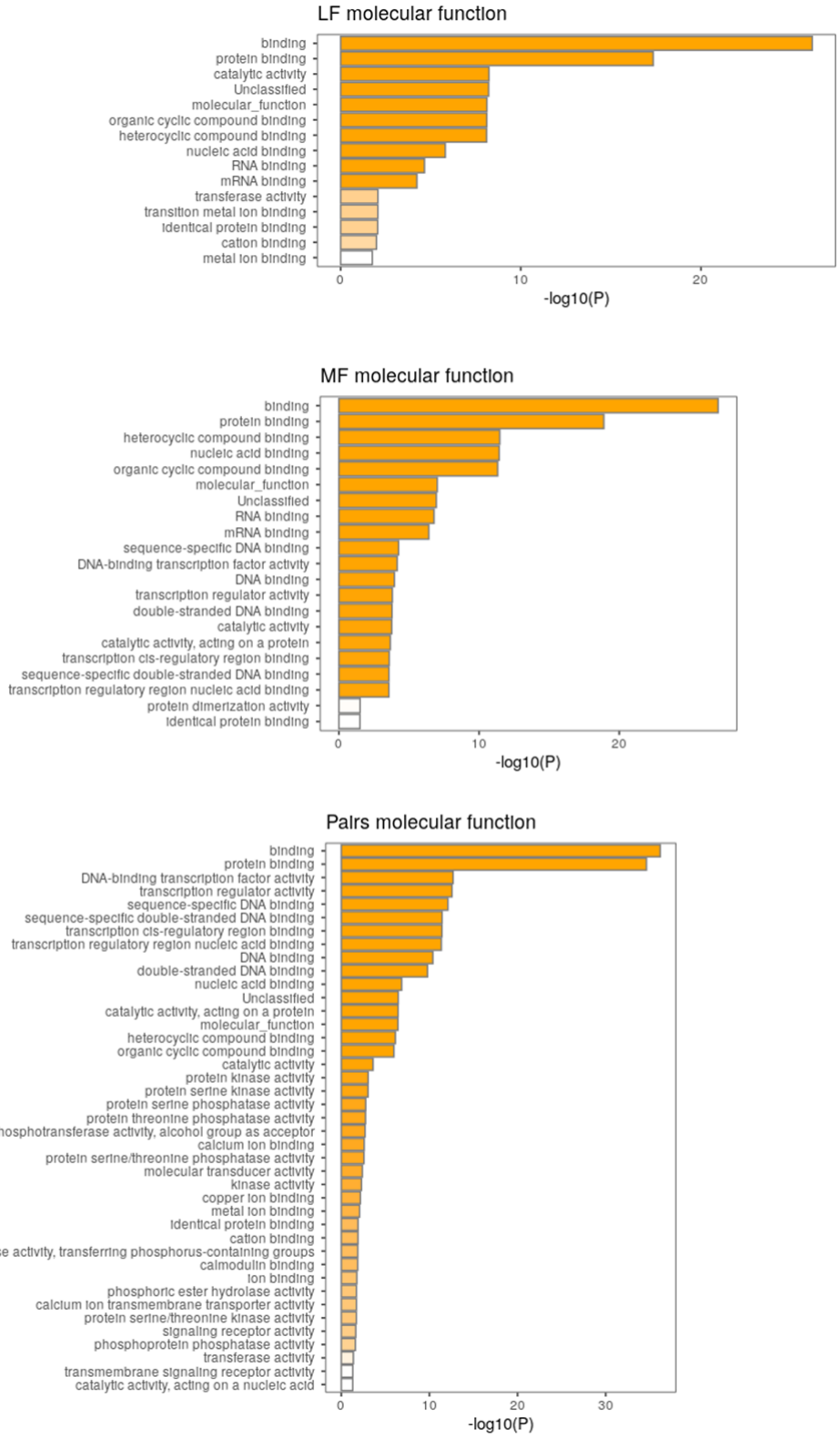


Figure S3 Gene ontology annotation using biological process terms of LF, MF and duplicated genes. Functional annotation of genes retained in the dominant-LF, sensitive-MF fraction of the genome and duplicated pairs. As metascape does not support more than 3000 gene IDs, enrichment analyses were realized in Gene Ontology database (http://geneontology.org) using the corresponding ATG terms.

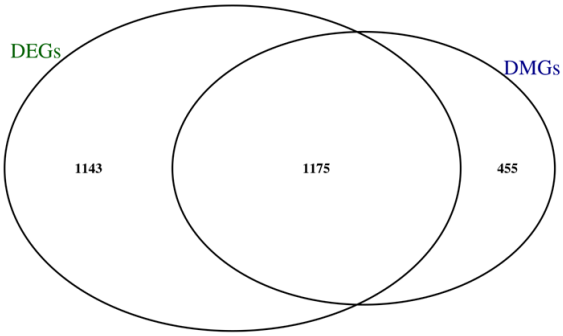


Figure S4 Venn diagram of differentially expressed genes (DEGs) and differentially methylated genes (DMGs) between the duplicated genes from the Salicales specific WGD event.

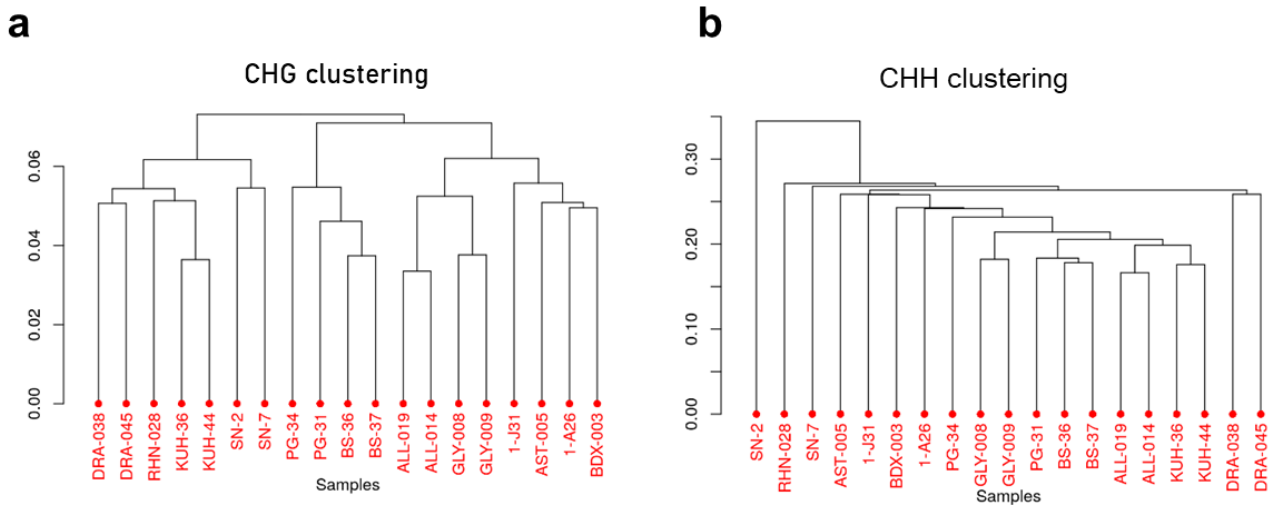


Figure S5 Phylo-epigenomic trees reconstruction for DNA methylation. A, Methyl-CHG tree was realized with genome-wide CHG SMPs filtered with SNPs data, coverage (>7X) and tolerating 30% of missing data. B, The CHH methyl tree was realized with genome-wide CHH SMPs filtered with SNPs data, coverage (>7X) and tolerating 30% of missing data.

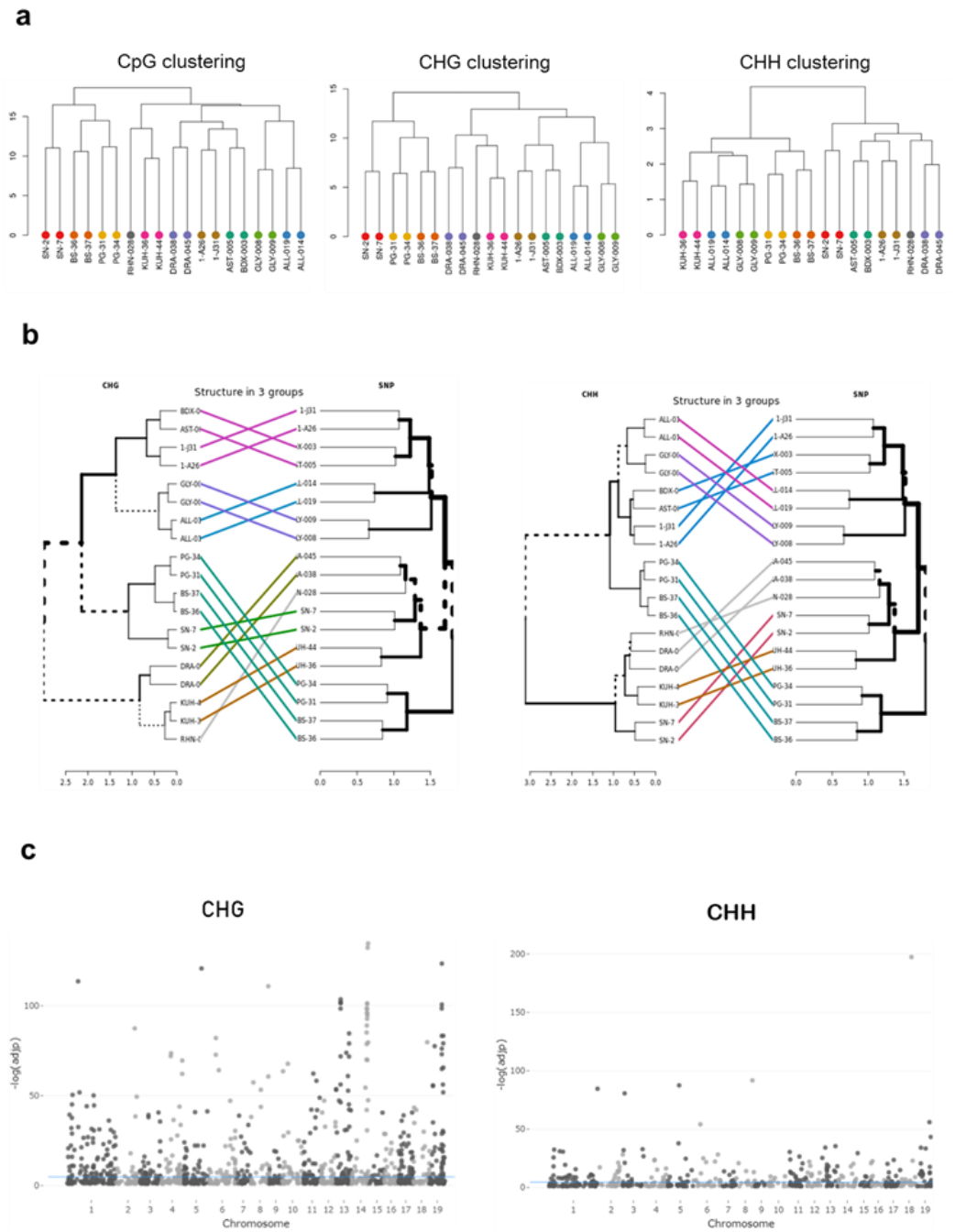


Figure S6 Methylation markers involved in local adaptation. A, Phylo-epigenomic tree in CpG, CHG and CHH contexts using only SMPs located in gene features including the promoter regions. The genotypes were clustered based on the similarity of their methylation profiles for each methylation context separately using ward hierarchical clustering and pearson correlation distance methods implemented in *methylkit* software. B, Comparison between the phylo-epigenomic tree in CHG and CHH contexts built on differentially SMPs between the three sub-groups and the phylogenetic tree (SNP). The genotypes were clustered based on the similarity of their methylation profiles for each methylation context separately using ward hierarchical clustering and pearson correlation distance methods implemented in *methylkit* software. B, Manhattan plots of obtained epigenetic pcadapt gene markers in CHG and CHH contexts with adjusted p-values (FDR, blue line). Markers above the blue line (FDR = 0.05) are considered as putative markers of local adaptation.

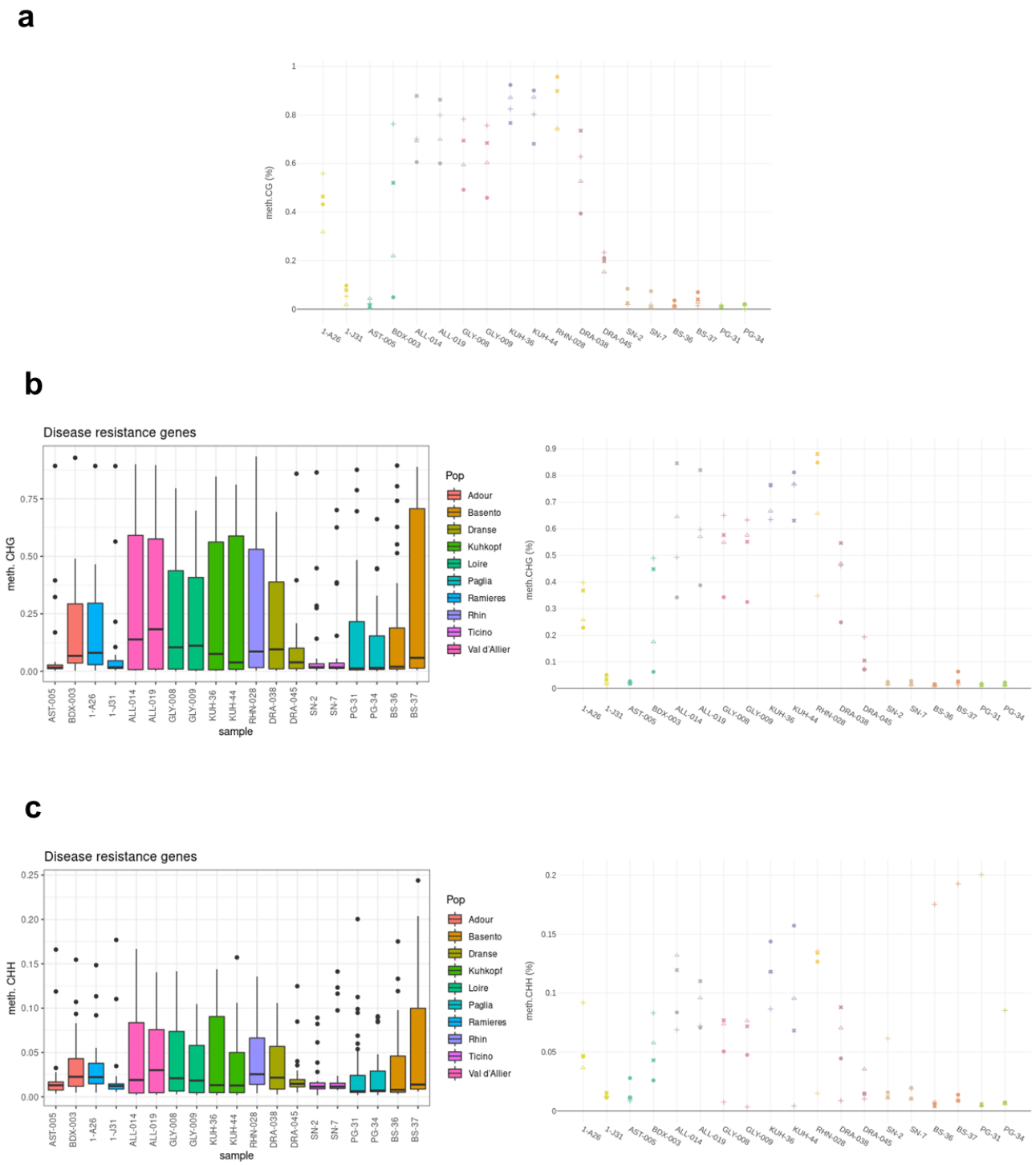
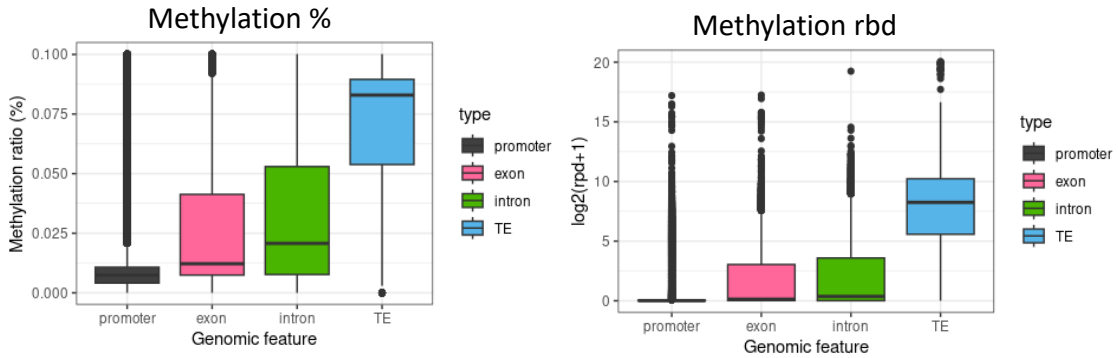
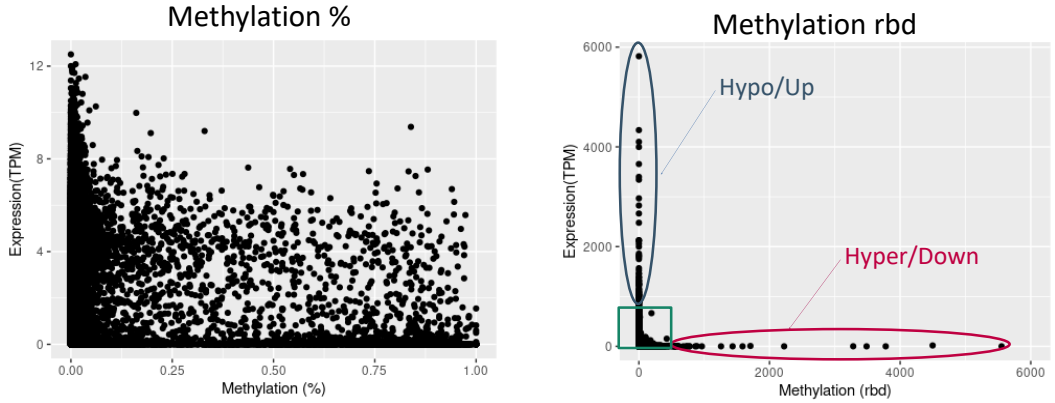


Figure S7 DNA methylation dynamics of disease resistance (R) genes in the 20 black poplar trees. A, CpG Methylation dynamics of the most variable R genes between the ten populations. B, CHG methylation dynamics of the R genes between the ten populations (left) and the most variable R genes regarding the sub-groups (right). C, CHH methylation dynamics of the R genes between the ten populations (left) and the most variable R genes regarding the sub-groups (right). The populations are represented by different color codes.

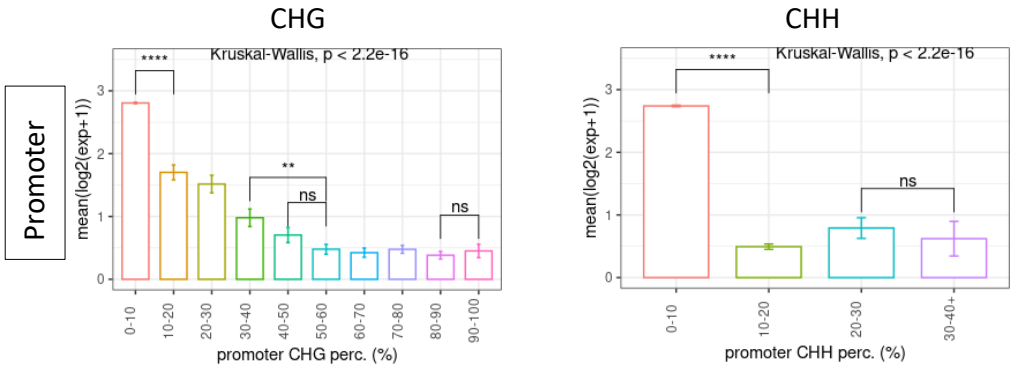
a


Methylation %
Methylation ratio (%)
0.100
0.075
0.050
0.025
0.000
promoter
exon
intron
TE
Genomic feature
type
promoter
exon
intron
TE
Methylation rbd
log2(rpd+1)
20
15
10
5
0
promoter
exon
intron
TE
Genomic feature
type
promoter
exon
intron
TE

b


Methylation %
Expression(TPM)
12
8
4
0
0.00
0.25
0.50
0.75
1.00
Methylation (%)
Methylation rbd
Hypo/Up
Hyper/Down
Expression(TPM)
6000
4000
2000
0
0
2000
4000
6000
Methylation (rbd)

c


CHG
CHH
Promoter
Kruskal-Wallis, p < 2.2e-16
****
**
ns
mean(log2(exp+1))
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
promoter CHG perc. (%)
promoter CHH perc. (%)

d

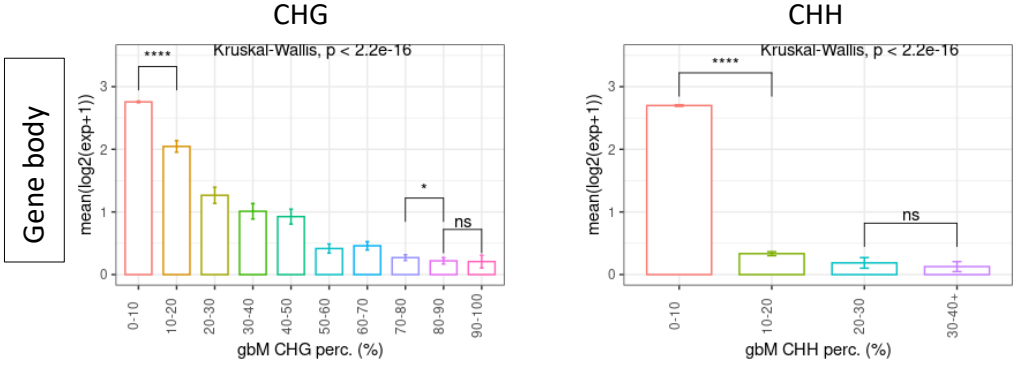

CHG
CHH
Gene body
Kruskal-Wallis, p < 2.2e-16
****
*
ns
mean(log2(exp+1))
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
gbM CHG perc. (%)
gbM CHH perc. (%)

Figure S8 Comparison between gene expression and DNA methylation. A, Comparison between methylation normalization approaches in percentage (%) and rbd (read per density) for promoters (black), exons (red), introns (green) and intergenic regions or TEs (blue). B, Comparison between gene expression and DNA methylation using percentage (left) and rbd (right) normalization approaches. Hypo/Up for hypomethylated and upregulated genes and Hyper/Down for hypermethylated and downregulated genes. C, Comparison between gene expression and promoter DNA methylation using percentage normalization approach splitted into 10 quantiles with each quantile capturing 10% of the methylation in CHG and CHH contexts. D, Comparison between gene expression and gene-body DNA methylation using percentage normalization approach splitted into 10 quantiles with each quantile capturing 10% of the methylation in CHG and CHH contexts.

Annexe D

Bellec et al. (2023)

the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2023) **114,** 1243–1266 doi: 10.1111/tpj.16185

Tracing 100 million years of grass genome evolutionary plasticity

Arnaud Bellec[1,†], Mamadou Dia Sow[2,†], Caroline Pont[2], Peter Civan[2], Emile Mardoc[2], Wandrille Duchemin[2], David Armisen[2], Cécile Huneau[2], Johanne Thévenin[3], Vanessa Vernoud[4], Nathalie Depège-Fargeix[4], Laurent Maunas[5], Brigitte Escale[5,6], Bertrand Dubreucq[3], Peter Rogowsky[4], Hélène Bergès[1] and Jerome Salse[2,*]

[1]*INRAE/CNRGV US 1258, 24 Chemin de Borde Rouge, 31320, Auzeville-Tolosane, France,*
[2]*UCA, INRAE, GDEC, 5 Chemin de Beaulieu, 63000, Clermont-Ferrand, France,*
[3]*INRAE/AgroParisTech-UMR 1318. Bat 2. Centre INRA de Versailles, route de Saint Cyr, 78026, Versailles CEDEX, France,*
[4]*INRAE/CNRS/ENS/Univ. Lyon-UMR 879, 46 allée d'Italie, 69364, Lyon Cedex 07, France,*
[5]*Arvalis–Institut du végétal, 21 chemin de Pau, 64121 Montardon, France, and*
[6]*Direction de l'agriculture de Polynésie française, Route de l'Hippodrome, 98713 Papeete, France*

Received 13 November 2022; revised 29 January 2023; accepted 24 February 2023; published online 14 March 2023.
*For correspondence (e-mail jerome.salse@inrae.fr).
†These authors contributed equally to this work.
[Correction added on 28 April 2023, after first online publication: Figures 1 and 5 have been revised in this version.]

SUMMARY

Grasses derive from a family of monocotyledonous plants that includes crops of major economic importance such as wheat, rice, sorghum and barley, sharing a common ancestor some 100 million years ago. The genomic attributes of plant adaptation remain obscure and the consequences of recurrent whole genome duplications (WGD) or polyploidization events, a major force in plant evolution, remain largely speculative. We conducted a comparative analysis of omics data from ten grass species to unveil structural (inversions, fusions, fissions, duplications, substitutions) and regulatory (expression and methylation) basis of genome plasticity, as possible attributes of plant long lasting evolution and adaptation. The present study demonstrates that diverged polyploid lineages sharing a common WGD event often present the same patterns of structural changes and evolutionary dynamics, but these patterns are difficult to generalize across independent WGD events as a result of non-WGD factors such as selection and domestication of crops. Polyploidy is unequivocally linked to the evolutionary success of grasses during the past 100 million years, although it remains difficult to attribute this success to particular genomic consequences of polyploidization, suggesting that polyploids harness the potential of genome duplication, at least partially, in lineage-specific ways. Overall, the present study clearly demonstrates that post-polyploidization reprogramming is more complex than traditionally reported in investigating single species and calls for a critical and comprehensive comparison across independently polyploidized lineages.

Keywords: grasses, paleogenomics, evolution, ancestor.

INTRODUCTION

Since Charles Darwin's seminal theory in 1859 (Darwin, 1859), the understanding of species evolution has relied on unveiling the forces promoting biodiversity through speciation and diversification, ultimately leading to morphological and phenotypic innovations. In this respect, Susumu Ohno proposed polyploidy – also referred to as whole genome duplication (WGD) – as a key contributor to such innovation (Ohno, 1970). Polyploidy is a condition of having multiple (> 2) sets of chromosomes that coexist in one nucleus and can be stably inherited by progenies. Polyploids are formed in two ways: autopolyploidization corresponding to genome doubling involving the same parental species and allopolyploidization corresponding to genome doubling involving two parental species (Leitch & Bennett, 1997). A further distinction is related to the timing of this event. Paleopolyploids are species that experienced polyploidization in their ancient past and their genomes have been subsequently re-diploidized through structural and functional reorganization. Neopolyploids are species that experienced polyploidization more recently and still possess distinguishable sets of parental chromosomes.

Polyploid species are widespread across the animal kingdom, including frog (Schmid et al., 2015), fish (Liu

et al., 2017a), insects (Li et al., 2018) and mammals (Acharya & Ghosh, 2016). In plants, polyploidy is ubiquitous in angiosperms (Van de Peer et al., 2017) and widely found in lower plants, such as gymnosperms (Li et al., 2015), ferns (Hidalgo et al., 2017) and diatoms (Parks et al., 2018). It has been proposed that polyploidy is implicated in the generation of phenotypic diversity (Landis et al., 2018; Leebens-Mack et al., 2019), species diversification (Levin & Soltis, 2018), crop domestication (Salman-Minkov et al., 2016) and adaptation (Vanneste et al., 2014), all of which is attributed to the enhanced genomic plasticity generated by WGD (Leitch & Leitch, 2008). Thus, numerous studies have suggested that polyploidy contributed to the evolutionary success of extant angiosperm species, including the model plant *Arabidopsis* (Thomas et al., 2006) and major crops such as sorghum (Paterson et al., 2009), maize (Schnable et al., 2011), *Brassicaceae* (Murat et al., 2015), wheat (Pont et al., 2013), cotton (Renny-Byfield et al., 2015) and soybean (Schmutz et al., 2010).

However, several fundamental questions about polyploidization remain open. What are the structural and functional features of the new balance between genomic plasticity, stability and evolvability? To what extent does polyploidization enhance molecular evolution, epigenetic changes and alterations in gene expression in duplicated genomic regions and genes? What mechanisms establish these changes? Addressing these questions requires a reconstruction of the gene content and genome organization of extinct ancestors predating the speciation and WGD events. The reconstruction of the ancestral genomes allows precise identification of major karyotype changes and associates them with evolutionary consequences (Pont et al., 2019b). In the present study, we make a wider use of this strategy, expanding it beyond a single species and WGD, toward a series of speciation and polyploidization events covering several species of the grass family over the last 100 million years (my). The objective is to provide generalized description of the structural and functional consequences accompanying polyploidization events during plant evolution.

RESULTS

Consequences of polyploidization events on karyotype remodeling

Grasses (*Poaceae* family) went through a whole genome duplication event around 100 Myr ago (referenced as ρ), before the divergence of the *Ehrhartoideae* (including rice), *Pooideae* (including *Brachypodium*, as well as the *Triticeae*) and the *Panicoideae* (including sorghum, setaria and maize) subfamilies. After the ancestral shared ρ tetraploidization, additional lineage-specific polyploidization events occurred in wheat (tetraploidization and hexaploidization, 0.36 and 0.01 Myr ago, respectively) and maize (tetraploidization, 5 Myr ago), Murat et al., 2017. Grasses can therefore be considered as the ideal angiosperm model system to investigate the fate of both paleo- and neo-polyploidy events on genome organization and regulation. Comparative genomics in grasses was conducted through the reconstruction of ancestral genomes. The ancestral genome is a 'median' or 'intermediate' genome consisting of the most parsimonious gene order, based on the extant genes conserved between the modern species investigated (Murat et al., 2017). A pivot species was chosen for each of the investigated subfamilies on the basis of the smallest numbers of historical polyploidization events (i.e. rice, *Brachypodium* and sorghum). Ancestral grass karyotype (AGK) were inferred by integrating paralogies and orthologies (following a synteny-based approach) from the eight investigated genomes (rice, *Brachypodium*, barley, tetraploid and hexaploid wheat, setaria, maize and sorghum), defining independent contiguous ancestral regions (CARs, also referred to as protochromosomes). This resulted in a pre-ρ AGK of seven protochromosomes (hereafter AGK7) with 10 286 ordered protogenes and a post-ρ AGK of 12 protochromosomes (hereafter AGK12) with 16 560 ordered protogenes (Figure 1a, Figure S1), refining and enriching (above 10%) the ancestral gene content of grass ancestors reported previously (Pont et al., 2019b). Protogenes are enriched in Gene Ontology (GO) (http://geneontology.org) terms related to basic cellular functions such as transferase, transporter and signaling activity ($P < 0.05$) (Figure S2a). The comparative genomics data produced here and associated inferred ancestors are available at https://urgi.versailles.inra.fr/synteny/ (Data S1).

Comparison of the gene order between the inferred ancestors (AGK7 and AGK12) and the eight modern genomes allowed the identification of losses of gene collinearity, defining synteny breakpoints (SBPs) (Figures S1 and S3, Table S1). SBPs are derived from ancestral chromosome fusions, fissions and translocations in the course of evolution. We confirmed that AGK7 was duplicated (ρ paleotetraploidization) to reach the AGK12 intermediate through fissions (2) and translocations (2), overall defining six SBPs. Remarkably, the synteny of AGK12 and rice genomes is fully conserved, whereas four (two fusions), nine (three fusions and one translocation) and 38 (18 fusions) SBPs since AGK12 have been identified in sorghum, setaria and maize, respectively. By comparing barley to diploid, tetraploid and hexaploid wheat genomes, we reconstructed the ancestral *Triticeae* karyotype (ATK7) made of seven protochromosomes covered by 12 718 conserved genes. Relative to AGK12, ATK underwent five fusions and a complex interplay of translocations shaping the modern *Triticeae* chromosomes 4 and 5, leading to 10 SBPs. The wheat A subgenome (in tetraploids and hexaploid) shows an additional SBP corresponding to a 4A/7B translocation, leading to a total of 11 SBPs. Following the karyotype evolution at the basis of the modern grasses

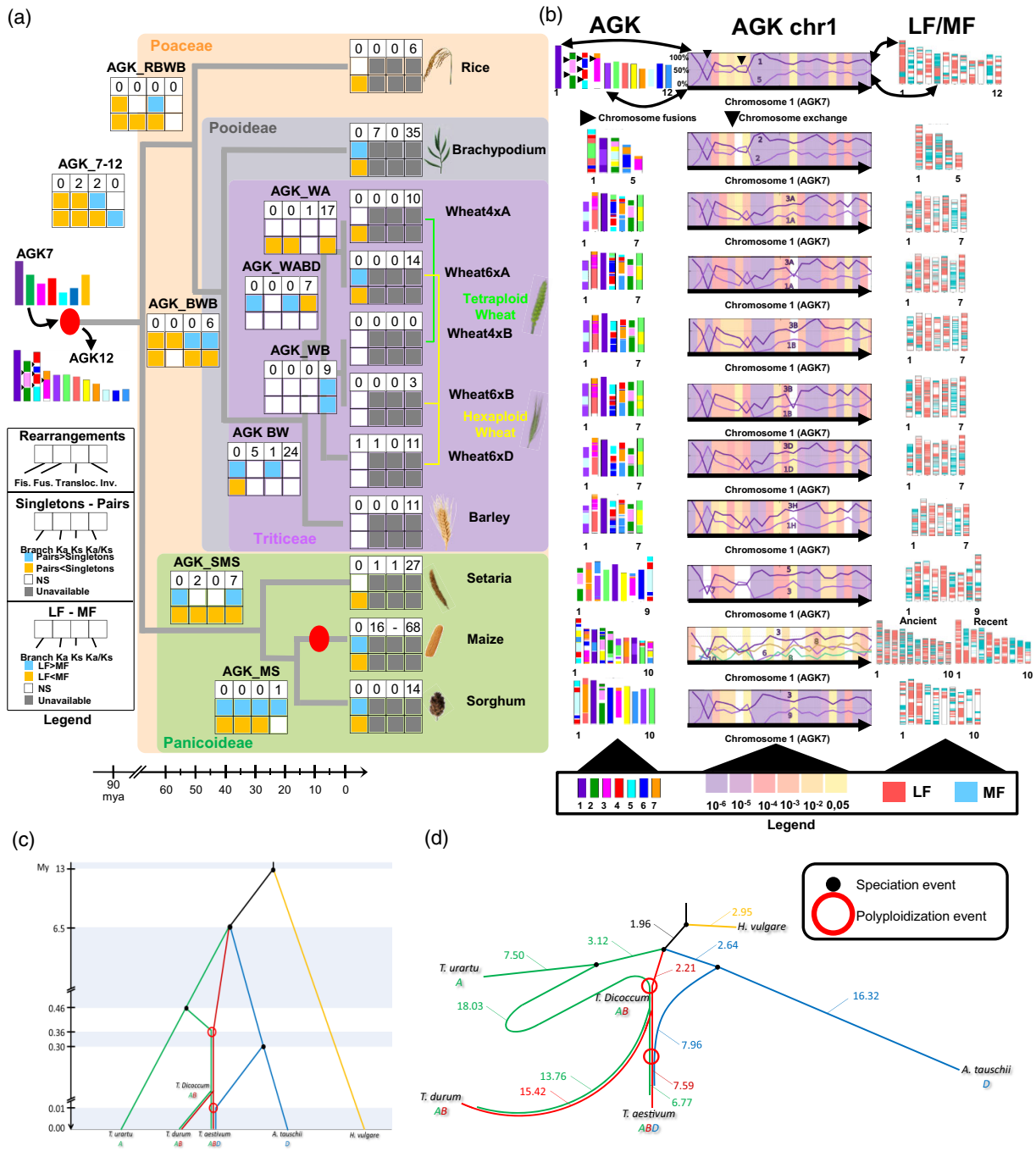


Figure 1. Evolution of grass genome structures. (a) A phylogenetic tree of grass evolution according to the time scale (bottom) in million years (Myr). Common ancestors (at the left) of extant species (at the right) reported at each tree node are abbreviated according to the initial letters of the descending species, with, for example, BWB ancestor corresponding to the speciation time point of *Brachypodium*-wheat-barley. For each ancestor, the number of chromosomal fissions (fis.), fusions (fus.), translocation (trans.) and inversions (inv.) are shown in the top rows of the minitables, together with differences in branch length, K_a, K_s, K_a/K_s between singletons and pairs (middle row in mini-tables), and LF and MF genomic fractions (bottom row in mini-tables) (see legend at bottom left). The ancestral WGD and maize-specific WGD are indicated by red dots. (b) Left: Extant genomes are illustrated with a color code for the seven ancestral chromosomes (AGK7 color code). Center: Graphs show the retention of ancestral genes for each pair of the duplicated blocks (ρ WGD) corresponding to the ancestral chromosome 1. Significant differences of gene retention between the duplicated blocks are indicated by a color code (see legend at bottom). For maize, both the ρ WGD and the recent maize-specific WGD are considered (four duplicated blocks). Right: Modern genomes are illustrated with the least fractionated (LF) and most fractionatedm (MF) genomics compartments (red for LF genes and blue for MF genes). (c) Phylogenetic tree of *Triticeae* scaled according to estimated time [in million years (Myr)] of speciation (black dots) and polyploidization (red circles) events. (d) Variations of substitution rates in the *Triticeae* lineage. The branch lengths are proportional to the substitution rates expressed in number of substitutions per site per billion years, highlighting differences in sequence dynamics between species and subgenomes.

(from AGK12) and at the basis of the *Triticeae* (from ATK7), the numbers of SBPs leading to the modern genome structures greatly differ between lineages, without any shuffling event or SBP shared among the eight investigated species.

In addition to the large chromosomal rearrangements, we investigated inversions during grass evolution (Figure 1a, Table S2, Figure S4a). Overall, the fraction of the genome covered by inversions ranges from 0.2% in rice to 22.8% in the tetraploid wheat subgenome B (Figure S4b). Inversion events have particularly impacted *Triticeae* genomes, covering 17.4% of the barley genome, 19.4% of the hexaploid wheat subgenome D and 20.6–22.8% of the wheat subgenomes A and B in the tetraploid and hexaploid contexts. Wheat subgenome A, both in the tetraploid (64 inversions since AGK12) and hexaploid (68 inversions) contexts, appears to be more dynamic (based on the number of inversions) than the B (46 and 49 inversions in tetraploid and hexaploid wheat, respectively) and D (48 inversions) subgenomes. Noteworthy, most of the subgenome A extra-inversions occurred after the tetraploidization event (Figure S4c). The maize genome went through an increase in inversions (68) after its divergence with sorghum (14 inversions) and the polyploidization event dating back to 5 Myr ago, clearly indicating the role of polyploidizations in promoting genome shuffling. Inversions appear to be associated to high gene density genomic regions compared to 1000 simulations performed by positioning the inversions randomly on the chromosomes of the investigated species (Figure S4d, Table S3) and tend to be located in gene rich high-recombination (HR) genomic regions compared to gene poor low-recombination (LR) regions (Figure S4e). We have also observed signs of TE involvement in genomic rearrangements, particularly on the chromosome 6A of hexaploid (*cv*. Chinese Spring) and tetraploid wheat (wild emmer Zavitan and durum *cv*. Svevo) showing the highest rate of inversions (Figure S5a–c). We found insertions of an long terminal repeat retrotransposon Jorge (*Copia* superfamily) at the immediate boundary of the inversions detected between Chinese Spring and Zavitan genomes and, similarly, an insertion of Laura (*Gypsy* superfamily) bordering an inversion detected between Chinese Spring and Svevo genomes.

Impact of polyploidization events on sequence divergence

To study gene divergence following polyploidization, we generated 15 810 gene trees from a total of 304 549 modern genes representing 68% of the genes annotated in the eight investigated species. Out of the 15 810 gene trees, 2599 and 3851 correspond to AGK7 (pre-ρ) and AGK12 (post-ρ) genes respectively, conserved (i.e. present with at least one homolog) in all investigated species. Phylogenetic tree reconciliation lead to the identification of 22 334 ancestral genes at the time of the maize WGD, 30 567 ancestral genes at the speciation of the AGK12 ancestor (complementing the catalog of 16 560 protogenes from the synteny-based approach described in the previous section) and 15 435 ancestral genes at each node of the speciation of the investigated species (Figure S6a, Table S4). Each node of the phylogenetic tree consists in most recent common ancestors (MRCA) referenced with an abbreviated name according to the initial letters of the descending species (e.g. with BWB ancestor corresponding to the speciation time point of *Brachypodium*-wheat-barley). Gene phylogenies allowed us to investigate substitution rate changes in conserved genes inherited from a particular speciation or duplication event, expressed as substitutions per site and per billion years. Both the ancient WGD (ρ) and the recent maize WGD are associated with an increased rate of substitutions (4.11 and 9.79, respectively), as well as speciation events (5.04 and 4.30 at the basis of the *Pooideae* and *Panicoideae*, respectively) compared to all the closest relatives (< 3.23). In the case of the ancient (ρ) WGD event, the substitution rate has undergone a net slowdown in both lineages resulting from the speciation of AGK12, with 2.15 for the *Pooideae* and 2.04 for the *Panicoideae*. In both subfamilies, the internodes leading to speciation (BWB ancestor, BW ancestor and MS ancestor) appear to evolve faster (with rates of 5.04, 3.23 and 4.3 respectively) compared to post-speciation tree branches (with 2.74 for rice, 2.17 for *Brachypodium*, 3.07 for barley, 2.84 for wheat, 2.31 for *Setaria* and 1.93 for sorghum), Figure S6b and Table S5.

To gain better insights into the consequences of recent polyploidizations on substitution rates, we focused on the *Triticeae* lineage complex represented by seven ATK protochromosomes covered by 12 718 protogenes, in selecting a set of 3905 single-copy orthologs shared by barley, diploid (*Triticum urartu* and *Aegilops tauschii*), tetraploid and hexaploid wheats (Figure 1c). Considering a shorter time scale in *Triticeae* evolution (13 Myr) (Figure S6), the count of substitutions is lower than those previously described in grasses (Table S5). Nevertheless, such *Triticeae* specific gene set provides a better understanding of the recent divergence between wheat subgenomes, contrasting the evolutionay speed of the different lineages and subgenomes in response to polyploidization (Figure 1d, Figure S6c, Table S5). Barley and the wheat lineages prior to the speciation and polyploidization events display low substitution rates, ranging from 1.96 to 2.95 (substitutions per site per billion years). The highest substitution rate of 18.03 was observed for the A lineage leading to polyploid wheats, relating to the period after the divergence from *T. uratu* (0.46 Myr ago) and before the hexaploidization event, marked by the divergence of the A subgenome in the tetraploid and hexaploid contexts (0.01 Myr ago). This long internode spans the tetraploidization event, which cannot be placed exactly. When focusing on the terminal branches, there are two apparently discontinuous categories of

substitution rates. Low substitution rates (6.77–7.96) were obtained for the A, B and D subgenomes in the hexaploid context and the diploid A lineage (*T. urartu*). High substitution rates (13.79–16.32) were obtained for the A and B subgenomes in the tetraploid context, and the diploid D lineage (*Ae. tauschii*). Although the observed branch lengths indicate that substitution rates could have been accelerated in the tetraploids or alternatively decelerated in the hexaploids, no general pattern of substitution rate change in relation to polyploidizations can be concluded. Associating substitution rates with polyploidization events is further complicated by the extreme patterns observed at the diploid level (7.5 in *T. urartu*; 16.32 in *Ae. tauschii*) and the uncertain effects of artificial selection (purifying selection, as well as domestication) that accompanied the domestication of tetraploid and hexaploid wheats.

Impact of polyploidization events on gene content

Subgenome dominance (i.e. a preferential retention of genes in one of the two homoeologous genomic regions) is another presumed consequence of polyploidization. To investigate this biased fractionation of the duplicated regions, we developed a window-based statistical approach defining the least fractionated (LF), most fractionated (MF) and unbiased compartments of the investigated genomes (Figure 1b, Figure S7a). Among the 447 744 genes annotated in the eight investigated species, a total of 158 124 (35% of annotated genes) and 74 311 (17% of annotated genes) were classified as LF or MF, respectively. The LF compartment contains genes enriched in functions related to signaling processes (transducer, transferase, transporter, kinase, etc.), whereas the MF compartment is enriched in transcription activity (binding, transcription, etc.) (Figure S2b). A similar approach has been applied for the recent duplication in maize, resulting in the annotation of 29 148 (46% of annotated genes) and 17 455 (27%) as LF- and MF-located genes, respectively (Figure S7b). Overall, we deliver a statistically-based assessment of 187 272 LF- and 91 766 MF-located extant genes in respect to the ancestral shared ρ and maize-specific WGDs (Data S1 and Table S4). For each of the pre-duplication chromosome, the two post-polyploidization (i.e. homoeologous) blocks are identified as either LF, MF or unbiased. Fraction changes (LF becoming MF or vice versa) on a duplicated block deriving from a polyploidization event may then reflect homoeologous chromosome exchanges during evolution. Based on this assumption, it appears that each of the post-WGD chromosome pairs, derived from AGK7 following the ancestral shared ρ events, experienced homoeologous chromosome exchanges particularly in the (peri-) centromeric region (Figure 1b, Figure S7).

Complementary to our analysis of genes located in LF and MF compartments, we investigated evolutionary trajectories of genes that are retained in a single copy (singletons) or two copies (pairs) after WGD (Table S4). From the gene-based phylogenetic approach described above (defining 15 810 gene trees), the singletons were defined as genes with an ancestral copy in AGK7 retained in a single copy in the eight extant species, thus defining 156 560 singletons (35% of annotated genes) in the extant genomes. The same strategy was applied in defining 15 157 singletons (24% of annotated genes) following the maize-specific WGD from AGK12. Conversely, ancestral genes in AGK7 retained in two copies in the extant species defined 113 837 genes as pairs (25% of annotated genes), whereas 14 293 genes (23% of annotated genes) were identified as pairs in maize when considering the maize-specific WGD from AGK12. Complementing the previous catalog of singletons and pairs based on phylogeny, singletons and pairs were also inferred using the synteny-based approach, delivering 16 560 ordered protogenes from AGK12 (post-ρ ancestor) and 10 286 ordered protogenes from AGK7 (pre-ρ ancestor). This led to the identification of 19 032 genes as pairs and 75 662 as singletons. By definition, the LF compartment is expected to contain more singletons relative to the MF compartment, whereas genes in the MF-compartment are more likely to be in the paired category (each gene loss in the MF compartment produces a singleton in the LF compartment and scarcity of gene losses in the LF compartment leads to a lack of singletons in the MF compartment). In accordance with this expectation, both inferences of singletons and pairs (phylogeny- or synteny-based approaches) identified sets of GO terms similar to the ones obtained for LF and MF compartments, respectively. Singletons are enriched in functions related to signaling processes, with the most significant GO terms for transferase, transporter, hydrolase and kinase (Figure S2c). Gene pairs are enriched in gene functions related to transcription activity with top significance for binding and transcription (false discovery rate < 0.05 using rice as reference) (Figure S2c).

Regarding the sequence divergence of genes (Figure 1a, Table S6), no difference in non-synonymous substitution rates (K_a) has been observed between singletons and pairs in six out of nine investigated MRCAs. However, regarding synonymous substitutions (K_s) rate, genes in pairs displayed a significantly higher rate compared to singletons for 6 out of the 9 investigated MRCAs. These differences were partially reflected in the ratios of non-synonymous to synonymous substitutions (K_a/K_s), with higher values observed in pairs for four out of nine investigated MRCAs, except for the wheat ABD ancestor exhibiting a higher K_a/K_s ratio for singletons. Differences in K_a, K_s and K_a/K_s between singletons and pairs are not significant for the remaining ancestors. To complement the investigation of grass gene sequence evolution, we examined branch lengths on phylogenetic trees (Table S6). In four out six of the more recent MRCAs [SMS ancestor (27 Myr),

MS ancestor (16 Myr), BW ancestor (13 Myr) and WABD ancestor (6.5 Myr)], as well as four out of 11 extant species, pairs have evolved faster than singletons. This is reversed in the older MRCAs [AGK12 (60 Myr), RBWB ancestor (46 Myr) and the BWB ancestor (35 Myr)] where singletons have evolved faster than pairs. Taken together, these observations support the notion that sequence divergence following WGD depends on whether the gene is conserved in multiple copies (pairs) or not (singletons). When observed significance for 58% (21 among 36) of the K_a-K_s-K_a/K_s values obtained for the nine inferred ancestors, pairs evolve faster than singletons in 71% (15 among the 21) of the observed significant differences.

At the post-polyploidization block level, comparisons between genes located in LF and MF fractions (Figure 1a, Table S6) revealed that LF genes exhibit significantly lower molecular evolution rates than MF genes in five out of nine MRCAs, considering both synonymous and non-synonymous substitution rates (LF/MF ratios ranging from 1.017 to 1.072 and from 1.026 to 1.189 for the synonymous and non-synonymous substitutions, respectively). Moreover, ancestors where the differences between the LF and MF compartments were not significant are the more recent ancestors (i.e. more distant from the original ρ WGD event), suggesting that the polyploidization effect on gene sequence is no longer active after such a long period of time or that allopolyploidization (wheat hexaploidization for example) has less effect than autopolyploidization (ancestral ρ WGD). Ratios of non-synonymous to synonymous substitutions (K_a/K_s) were found to have a more complex pattern. The K_a/K_s ratio is lower for LF genes than for MF genes in three out of nine ancestors (BWB, SMS and WA ancestors) and higher in AGK12 and WB ancestors. Overall, in complement to K_a and K_s dynamics, the analysis of branch lengths displayed a pattern coherent with previous results, but with a finer temporal inference (Figure 1a, Table S6). The genes located in the MF fraction displayed significantly longer branches compared to the LF-located genes in seven out of the nine MRCAs and seven out of the 11 extant species. These results indicate that sequence divergence of genes after polyploidization is affected by whether the gene belongs to the LF or MF fraction of the genome, with LF-located genes never evolving faster with statistical significance. However, this phenomenon of differential evolution rates appears to have occurred soon after the polyploidization event, with waning effect over longer periods of time.

Impact of polyploidization events on genome regulation (expression, methylation)

The evidence presented above (Figure 1b) documents the extent of biased gene retention after WGD (i.e. the subgenome dominance in grasses). Perhaps the most intriguing question pertains to the drivers and mechanisms of this phenomenon. Recently, evidence of a possible involvement of repetitive elements reactivated via modification of small RNAs and epigenetic marks has been presented (Wang et al., 2022). To test this hypothesis, we generated RNA-sequencing (RNA-seq) data and whole genome bisulfite sequencing datasets (BS-seq), comparable between wheat, *Brachypodium* and maize, and collected publicly available single nucleotide polymorphism (SNP) datasets (Data S2, Table S7). We investigated the relationship between the different gene status (genes located in inverted or collinear blocks, LF or MF blocks, conserved versus species-specific genes, pairs versus singletons; Figure 1a) and their evolutionary dynamics (substitution rates and nucleotide diversity) and regulation (gene expression, DNA methylation) patterns. Conclusions are raised below when omic variations in respect to the ancestral shared ρ WGD are consistent between at least two investigated species (between maize, wheat and *Brachypodium*) without a conflict (i.e. with no significant opposite signal) in the third species (Figure 2a–d).

We first assessed the possible role of genomic inversions in modifying gene regulation, following the idea that local changes in gene order (synteny decay) can disrupt *cis-trans*-regulatory interactions. Genes in inverted regions indeed exhibit higher expression and lower promoter methylation levels, in addition to higher synonymous substitutions rate (Figure 2a, Figure S8). We also observed lower expression and higher methylation levels (at promoter and gene body), as well as higher synonymous substitution rates in species-specific genes compared to the conserved ones (Figure 2b). Among the conserved genes, ancient pairs exhibit lower gene-body methylation level and elevated levels of molecular evolution (K_a and K_s) compared to singletons (Figure 2c). The recent maize duplication displayed distinct patterns, with only the gene-body methylation mimicking the omics variation observed for the ancestral ρ duplication. When comparing duplicated blocks, genes located in the MF fraction are more expressed in all species (except *Brachypodium*) and exhibit higher K_a-K_s rates (for all species except the recent maize WGD). When focusing on the most extremely fractionated LF and MF blocks at the whole genome level, defining fast evolving (FE, i.e. MF regions showing the highest loss of ancestral genes) and slow evolving (SE, i.e. regions showing the highest retention of ancestral genes) regions, the same trends were observed for gene expression, K_a, K_s and SNPs (Figure S8b). However, in maize, FE regions inherited from the ancestral ρ duplication were more methylated than SE, whereas no methylation bias was observed between FE and SE regions inherited from the maize-specific duplication (except for CG promoter), in contrast to that observed between the LF-MF compartments from the recent maize duplication.

To better understand the effect of recent polyploidizations, we investigated divergence between wheat

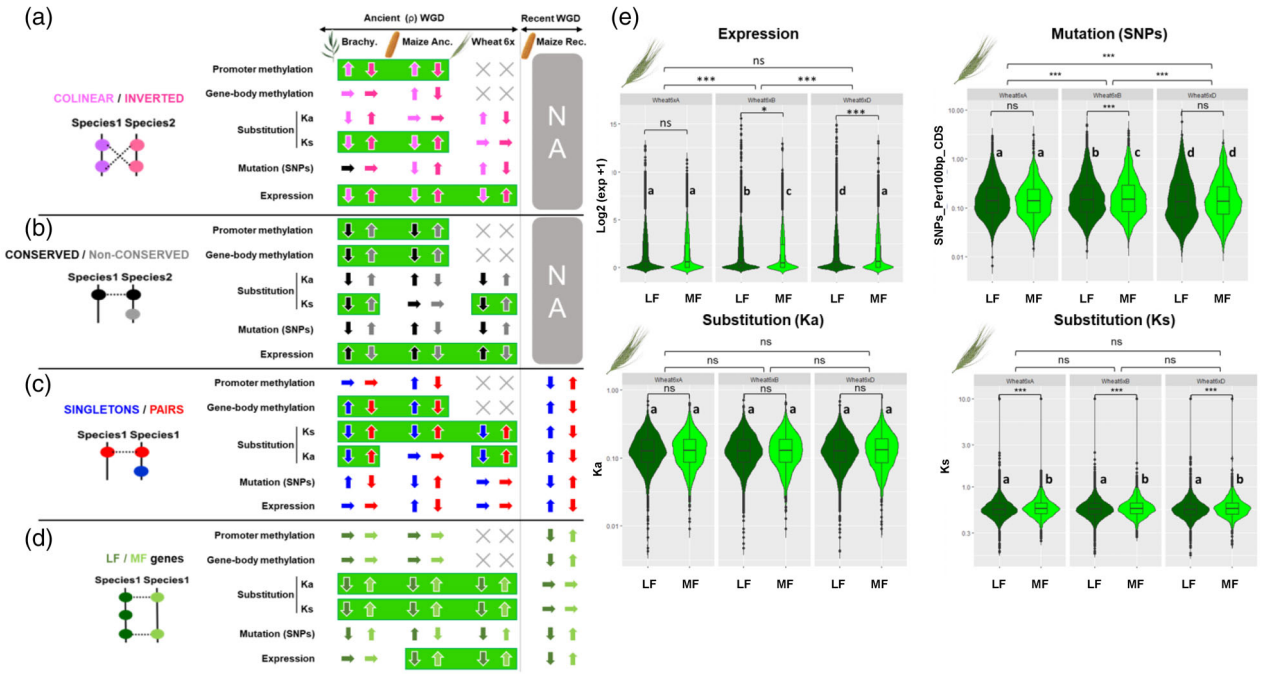


Figure 2. Regulation dynamics during grass evolution. (a) Omics differences such as methylation (promoter and gene-body), substitution (K_a and K_s), polymorphisms (SNPs) and expression are illustrated with arrows (↑ for significant increase; ↓ for significant decrease; → for no significant difference). The comparisons were performed between: (a) collinear versus inverted, (b) conserved versus non-conserved, (c) singleton versus pair and (d) LF- versus MF-located, genes. For each species and for each of the omics data considered, the first arrow (left) corresponds to collinear-conserved-singleton-LF genes and the second arrow (right) corresponds to inverted-non-conserved-pair-MF genes (on a, b, c and d panels, respectively). Omic variations in respect to the ancestral shared ρ WGD that are consistent between at least two species without a conflict (i.e. with no significant signal) in the third species are highlighted within a green background. (e) Omics variation (expression, mutations, substitution) between wheat post-polyploidization compartments (LF versus MF) inherited from the ancestral shared ρ WGD in A-B-D subgenomes regarding the wheat lineage-specific allopolyploidization.

subgenomes (A, B and D). Subgenomes display divergence in expression (A > D > B) and diversity (B > A > D) with no difference in substitution rates (K_a and K_s) (Figure S9). When ancient blocks (LF and MF) inherited from the ancestral (ρ) duplication are considered for each of the subgenomes (A, B and D), the hexaploid wheat genome can be divided into six genomic compartments (LF and MF for each A, B and D). The LF-D regions accumulate less polymorphisms (SNPs) than in A and B (LF-B > LF-A > LF-D), being the most stable compartment of the wheat genome. The same patterns of SNP differences were observed for MF genes with MF-B > MF-A > MF-D. Expression data show a more complex profile, where the B subgenome is the least expressed in both LF and MF compartments, the A subgenome appeared to be the most expressed in the LF compartment, and the A and D subgenomes are the most expressed in the MF compartments (Figure 2b, Table S8). The same analysis was performed at the duplicated genes level (singletons versus pairs). No significant differences were observed for expression, synonymous and non-synonymous substitutions between the A, B and D subgenomes. However, the D subgenome accumulated less polymorphisms for both singletons and pairs compared to the A and B subgenomes (Figure S10). This possibly illustrates that additional processes, such as domestication and selection, have shaped wheat genomes in complement to polyploidization events, either recent (between A, B and D) or ancient (between LF and MF) ones.

Impact of polyploidization events on conserved and duplicated genes

We then investigated regulation (expression and methylation) differences between pairs of conserved (between species) or duplicated (within species) genes (Figure 3). To do so, we first defined profiles for each gene by concatenating their expression or methylation status (0 for non-expressed or non-methylated, 1 otherwise) for three comparable developmental stages (stages 1, 2 and 3) of the grain development in maize, *Brachypodium* and wheat. From a repertoire of 1997 genes showing a perfect 1–2–3 gene-to-gene copy relationship in *Brachypodium*, maize and wheat, respectively (referred to as ohnologs), we found that half of the ancestral genes conserved the same expression profile in all three species (43% are expressed in all species and 6% are not expressed in all three species, Figure 3a). When looking at duplicated genes in the same species, we observed that 70% and 75% of the duplicates share the same expression profile in maize and hexaploid wheat,

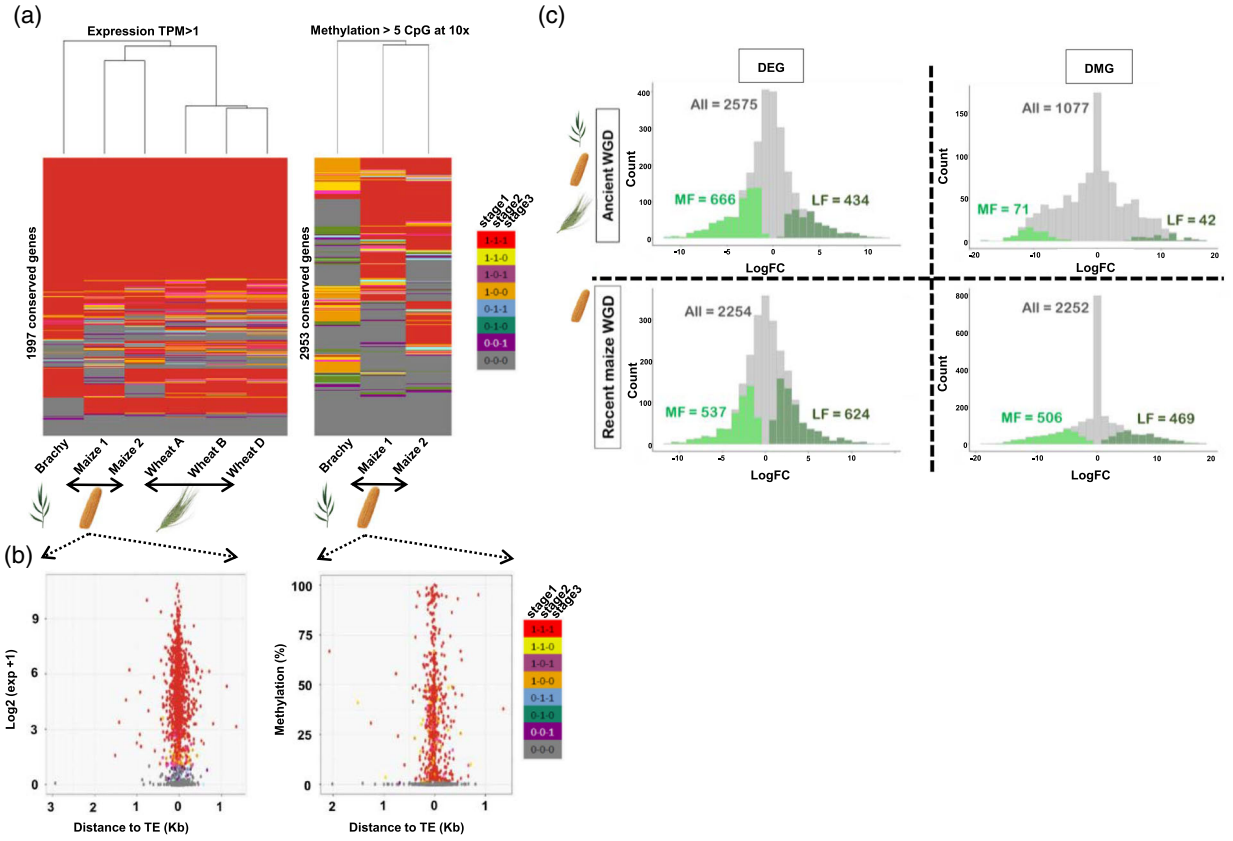


Figure 3. Regulation of conserved and duplicated genes. (a) Omics variation between 1997 genes showing a perfect 1–2–3 gene-to-gene relationship between *Brachypodium*, maize and wheat (for expression at the left) and 2953 genes showing a perfect 1–2 gene-to-gene relationship between *Brachypodium* and maize (for methylation at the right) using a color code illustrating expression/methylation profiles during the three developmental stages with 0 corresponding to no detected expression/methylation and 1 corresponding to detected expression/methylation (with the profile 1–1–1 (red) indicating genes expressed/methylated in the three developmental stages considered). (b) Average distance of the closest TE for duplicated pairs showing differences in expression and methylation in maize. The *x*-axis represents the distance to the nearest transposable element (TE) from the transcription start site (TSS) of the gene, and the *y*-axis indicates the gene expression (logarithmic scale, left) or DNA methylation (%, right) levels, respectively. (c) Differences in expression (with DEGs for differentially expressed genes, left) and methylation (with DMGs for differentially methylated genes, right) between gene pairs in *Brachypodium*, maize and wheat deriving from the ancestral shared ρ WGD (top) and for the maize-specific duplication (bottom). Pairs that are differentially expressed (FDR < 0.05) and differentially methylated (FDR < 0.05) in LF compared to MF are shown in dark green (upregulated LF genes) and light green (downregulated MF genes).

respectively, whereas 13% (249) and 5% (85) of the duplicates showed 'On/Off' expression pattern (i.e. one copy of the pair expressed, the other copy being silenced). Between the three species, a set of ancestral ohnologous genes with On/Off pattern have been identified, corresponding to 10, 9 and 1% genes conserved between *Brachypodium*/maize, *Brachypodium*/wheat and maize/wheat, respectively (Figure 3a, Table S9).

Similarly, a repertoire of 2953 genes showing a perfect 1–2 gene-to-gene relationship (i.e. ohnologs) between *Brachypodium* and maize was used to assess DNA methylation differences among conserved genes (methylation cut-off = 5 CpG at 10 × coverage) (Figure 3a, Table S9). Methylome analysis shows stronger differences between the three developmental stages in *Brachypodium*, where 46% of conserved genes show distinct pattern between stages (in contrast to 34% in maize), probably as a result of major DNA methylation variations between the two species during early embryonic stages of the grain development. Comparison between *Brachypodium* and maize ohnologs indicates higher methylation level in maize genes (22% with 1–1–1 and 23% with 0–0–0 profile between maize duplicates) than in *Brachypodium* (5% with 1–1–1 and 49% with 0–0–0 profile). We found 16% of the ancestral genes retaining the same methylation patterns between the two species with 14% corresponding to 0–0–0 profile (unmethylated genes in the three stages in both species). In maize, this percentage increases to 45%, with 22% of duplicates corresponding to 1–1–1 and 23% to 0–0–0 expression profiles. We then tested whether these differences in gene expression and DNA methylation between the ohnologs and duplicates in maize could be related to the presence of transposable elements (TEs) in proximity of genes, potentially disrupting the ancestral expression or methylation pattern. No significant correlation has been observed between the presence of TEs in the vicinity of the

gene copy showing differences in expression or methylation levels between conserved or duplicated gene pairs (Figure 3b, Figure S11).

We then investigated the expression fate of duplicated genes after WGD in *Brachypodium*, maize and wheat. After assigning the two copies of duplicates to their corresponding LF or MF compartments, we analyzed their expression and methylation differences [differentially expressed genes (DEGs) and differentially methylated genes (DMGs)] (Figure 3c). Most previous studies focus on expression differences between duplicated genes, with one copy expressed while the other one remaining silenced. However, only about half of the duplicated genes we investigated (for both ancient ρ and recent maize WGDs) exhibit expression differences, with the remaining gene pairs having the same expression pattern. When expression divergence is observed between pairs, genes from the MF compartment are overexpressed for ancient WGD (666 versus 434 genes in the LF compartment, chi-squared, $P = 2.65 \times 10^{-12}$), whereas genes in the MF compartment are significantly less expressed in the case of the recent maize WGD (537 versus 624 genes in the LF compartment, chi-squared, $P = 1.07 \times 10^{-2}$). DNA methylation also displays a distinct pattern between pairs depending on the WGD events investigated (ancient versus recent). The vast majority (90%) of ancient duplicates do not exhibit any DNA methylation differences, likely because of the age of the ρ WGD (100 Myr), which is expected to confound changes established shortly after WGD. Nevertheless, significantly higher number of observed DMGs are hypomethylated in the MF compartment (71 in MF and 42 in LF, chi-squared $P = 6.37 \times 10^{-3}$). The recent duplication event in Maize (5 Myr ago) offers more resolution, with about half of duplicated genes (43%) exhibiting DNA methylation differences. Similar numbers of these DMGs are hypomethylated in MF and LF (506 and 469, respectively; chi-squared $P = 0.236$). Finally, gene copies of pairs located in the MF compartment exhibit higher synonymous and non-synonymous substitution rates for both the ancient and recent maize-specific WGDs (Figure S12).

Interplay between genomic regulations on key traits

To investigate a possible relevance of the observed omics variations onto the genetic basis of key agronomic phenotypes or traits, we concentrate on the interplay between omics variables (i.e. synonymous and non-synonymous substitution rates, SNP density, gene expression and DNA methylation) for maize genes and their associated functions in biological processes (Figure 4). At the whole genome level in maize, we detected a weak negative correlation ($r^2 = -0.27$, $P < 2.2 \times 10^{-16}$) between promoter methylation (in CG, CHG and CHH) and gene expression, whereas a weak positive correlation ($r^2 = 0.2$, $P < 2.2 \times 10^{-16}$) has been recovered between gene expression and gene-body methylation in the CG context (Figure S13a). However, when splitting omics variable into quantiles, clear omics interplay appears. Highly methylated genes (methylation > 50%) in the promotor are less expressed, whereas hypomethylated genes (first quantiles) correspond to the most expressed genes (Figure 4a). Gene body methylation and expression level seem positively correlated (for the first quantiles) and negatively correlated for highly methylated genes (last quantiles, methylation > 58%). A negative correlation was observed between K_a and gene-body methylation as well as expression with lower expression levels as the K_a rate increases (Figure 4a), and with no clear pattern between K_s or SNP density on gene expression (Figure S13b). Interestingly, highly methylated and silenced genes belong to non-conserved (species-specific) genes, whereas conserved genes do not show a clear interplay between gene expression and DNA methylation (Figure S13c).

To study in more detail the link between DNA methylation and gene expression, gene-to-gene analysis was conducted using the multivariate approach implemented in mixOmics in maize and *Brachypodium* (Rohart et al., 2017). Overall, samples were clustered by developmental stages and the transcriptomic (gene expression) and epigenetic variables (in CG, CHG and CHH contexts) (Figure S14). Gene-to-gene analysis opened up possibilities to identify sets of genes with particular profiles, such as hypomethylated/upregulated (Hypo/Up) and hypermethylated/downregulated (Hyper/Down) genes, with expression expressed as TPM (with 'upregulated' referring to highly expressed genes and 'downregulated' to low expressed genes) and methylation expressed as read by density (rbd) (Figure 4b, Figure S15). In total, 441 and 50 genes were classified as Hypo/Up, respectively, in *Brachypodium* and maize and 440 and 604 genes were classified as Hyper/Down, respectively, in *Brachypodium* and maize. Using *Brachypodium* for GO analysis, Hypo/Up (441) genes (in CG, CHG and CHH) are enriched in GO terms such as transporter, transmembrane, ligase, cation, ion, acid, etc., relating to signaling processes, whereas Hyper/Down (397) genes (in CG and CHG excluding CHH context because of the low number of genes impacted) are enriched in functions related to purine/ribonucleotide binding, hydrolase activity, ion binding, transporter activity, heat shock protein binding, etc. (Figure S2d,e).

It has been proposed that polyploidization may promote the emergence of phenotypes leading to species diversification, adaptation and domestication (Qi et al., 2021). To assess the impact of gene or block duplication divergence (in expression and methylation) on key traits, we focused on the recent maize WGD event. Although expression divergence in most maize duplicates is not associated with methylation differences, 11% of the duplicated genes (out of 3193) display a clear interplay between

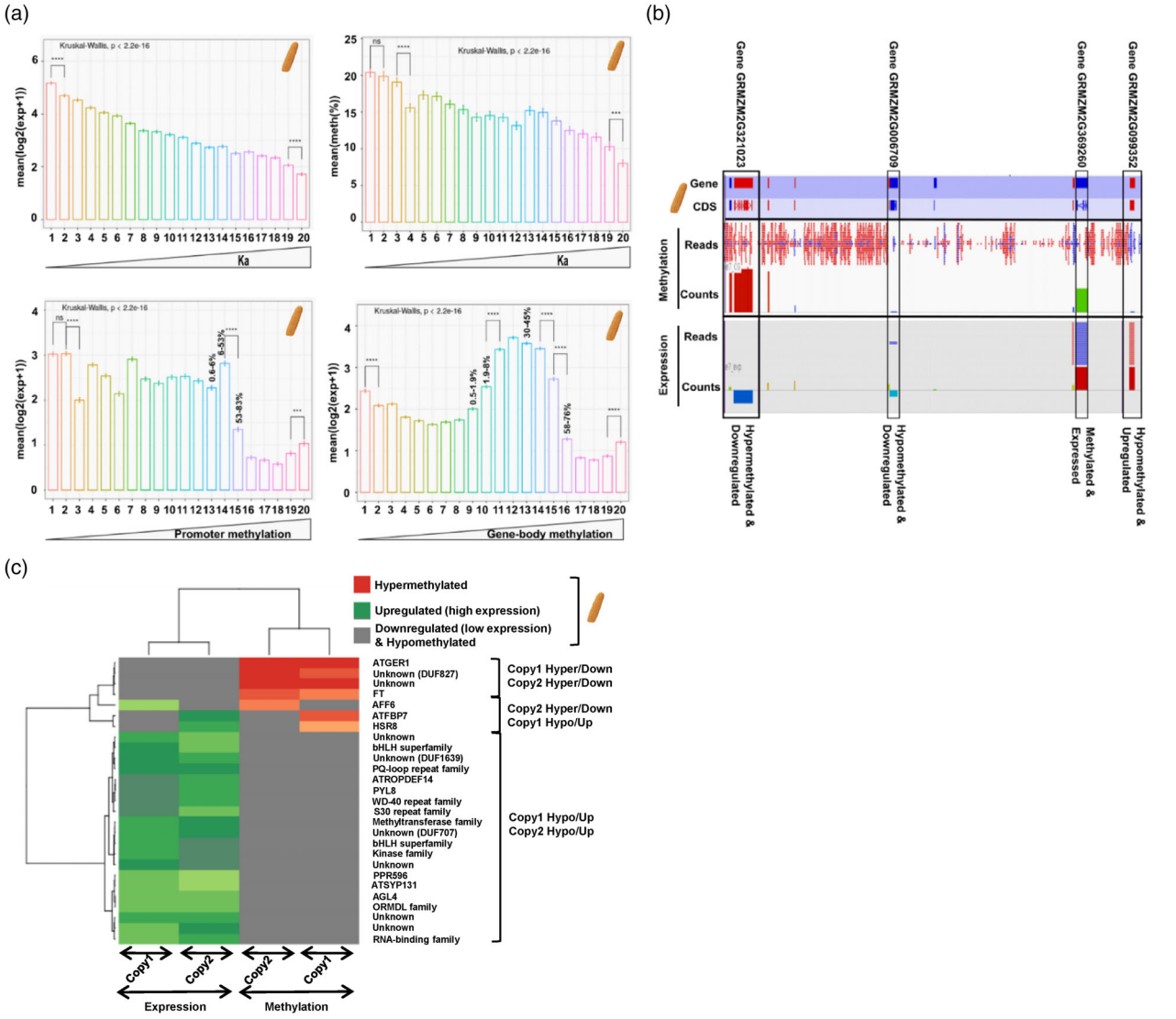


Figure 4. Multiomics regulation interplay in maize. (a) Interplay between substitution rates (K_a expressed in quantiles at the top) and gene expression (top left) or gene-body methylation (top right). Bottom: Link between promoter and gene-body DNA methylation level (in % splitted in quantiles) and gene expression [expressed as log2 (TPM) for promoter (bottom left) and gene-body (bottom right)]. (b) Zoom on a genomic region illustrating the link between gene expression and DNA methylation in maize for hypomethylated and upregulated (Hypo/Up) and hypermethylated and downregulated (Hyper/Down) genes. The coverage (top track, i.e. reads) and quantified (bottom track, i.e. counts) data are shown for methylation and expression of genes (red for high expression/methylation levels and blue for low expression/methylation levels). (c) Methylation and expression profile of duplicated trait-driving genes in maize with contrasted signature in methylation and gene expression. Gray indicates expressed/methylated genes; green indicates expressed genes; and red indicates methylated genes. Hyper/Down indicates hypermethylated and downregulated (low expression) genes and Hypo/Up indicates hypomethylated and upregulated genes (high expression). Gene names and functions are mentionned on the right as detailed in the main text.

gene expression and methylation for at least one copy of the gene pair. In particular, 8% of the duplicated genes had a hypomethylated and upregulated copy (referenced as Hypo/Up), whereas only 3% correspond to copies with the opposite pattern (hypermethylated and downregulated, i.e. referenced as Hyper/Down) (Figure 4c, Table S10). We then focused on pairs where both copies were contrasted in terms of methylation and expression profiles (Figure 4c). Among them, 20 displayed Hypo/Up status (i.e. both copies of a pair are hypomethylated and upregulated). These included bHLH, ROPGEF14, SEP2/AGL4, ORMDL, PPR596, SYP131, PYL8, CCoAOMT1 (*S*-adenosyl-L-methionine-dependent methyltransferases) involved mainly in pollen/floral development, seed germination/growth and stress response. We found four gene pairs where both copies are highly methylated and not expressed (Hyper/Down),

namely GER1/GLP1 (involved in fruit development), FT (flowering time) and two unknown genes. Finally, we found three key gene pairs where the two copies show a contrasted (complementary) omics regulation pattern, with one copy hypermethylated and downregulated, and the other copy hypomethylated and upregulated (Figure 4c). These genes correspond to ADF6 (defense response), FBP7 (seed development) and HSR8 (cell wall and glucose response).

DISCUSSION

Grasses as a key species complex to investigate polyploidization events

Polyploidization has been proposed as a driving force of plant genome evolution. However, the genomic reprogramming of duplicated compartments and genes inherited from polyploidization and underpinning such evolutionary success remains largely obscure. Polyploidization events are followed by a diploidization (or fractionation) process that consists of gene number reduction (Murat et al., 2014; Schnable et al., 2012; Woodhouse et al., 2010). The continuum of structural and regulatory modifications between the parental genome merger and the eventual return to a diploid status needs further investigation to decipher the biological functions gained from WGDs that may explain the role of recurrent polyploidizations during angiosperm evolution (Fox et al., 2020).

To investigate the evolutionary trajectories of duplicated genomes, we conducted a comparative omics analysis between eight modern grasses deriving from a 100 Myr old ancestors (AGK7-12) of 7–12 protochromosomes with 10–16 K protogenes enriched in basal functions such as transport activity. Despite widespread and recurrent polyploidizations, most grass species have markedly few chromosomes. For example, maize is expected to have $n = 28$ chromosomes ($7 \times 2 \times 2$) based on polyploidizations since AGK7 alone, but in fact has only $n = 10$. The many examples of species with far fewer chromosomes than predicted by past polyploidization events suggest a strong selective pressure that favors fewer chromosomes, although the advantages of a low chromosome number are not clear (Escudero & Wendel, 2020; Pont et al., 2019b). Extant grass genomes have derived from the inferred ancestor (AGK7 and AGK12) through distinct chromosomal fusions and fissions, as well as inversion events that tend to be located at the telomeres and more frequent in the polyploid context, without any reuse of synteny breakpoint, in contrast to that proposed in animals (Maria Maggiolini et al., 2020). Synteny breakpoints marking a recurrent inversion in chimpanzee and gorilla chromosomes homologous to the human chromosome 16 are near a conserved 23-kb inverted repeat composed of satellites, LINE and Alu elements. It is considered that this repeat mediated the inversion by bringing the chromosomal arms into close proximity, hence facilitating intrachromosomal recombination (Goidts et al., 2005). As in mammals, our analysis of the sequences surrounding inversions (on the wheat chromosome 6) supports the hypothesis of the involvement of sequence repeats in the emergence of inversions. For genes embedded in inversion, we observed a higher non-synonymous substitution rate (in *Brachypodium* and maize), higher expression and a lower methylation level (at promoter in CG-CHH-CHG contexts), suggesting that inversions may play a role in genome plasticity during evolution. In the mimetic butterfly, a set of genes coadapted for mimicry was assembled by an inversion (Joron et al., 2011). Dvorak et al. (2018) reported faster rates of genomic changes (inversions, translocations and duplications) in the *Triticeae* genomes compared to rice and sorghum, as well as faster substitution rates in the *Pooideae* branches compared to *Panicoideae* and *Oryzoideae*, with the highest rates in the A, B and *Ae. tauschii* genomes and the lowest rates in rice and sorghum. Such differences in rearrangement rates (particularly inversions) could be attributed to differences in lifespan, as suggested in mammals where the quickest rate of chromosome breaks was observed in mouse that has a short generation time (Murphy et al., 2005). We propose here that such genomic rearrangements are also favored by polyploidization events, with the merging of quasi-identical genomes in the nucleus facilitating rapid differentiation of the two parental subgenomes necessary for correct homologous (and not homoeologous) chromosome paring during meiosis.

Fate of duplicated genes and blocks following polyploidization

Deletion of duplicated genes does not occur at random but rather through a so-called subgenome dominance process, where most of the ancestral genes are preferentially retained on only one of the duplicated blocks. On the whole chromosome or genome level, this diploidization phenomenon then leads to a dominant subgenome enriched in singletons and characterized by gene retention (D or LF for 'least fractionated') and a sensitive subgenome enriched in duplicates and characterized by preferential losses (S or MF for 'most fractionated') (Salse, 2016). This phenomenon has not been observed in auto-polyploid plants, such as potato (Stupar et al., 2007), poplar (Liu et al., 2017b) and pear (Li et al., 2019).

In the present study, 25% and 23% of annotated genes were retained in pairs in the extant genomes, following the ancestral (ρ, 90 Myr ago) and maize-specific (5 Myr ago) WGDs, respectively. Pairs (or 'diploidization-resistant' genes), as well as MF-located genes, are enriched in functional annotation terms such as transcription factor and transcription regulator, in contrast to singletons (and LF-

located genes) enriched in signalization processes, in agreement with the findings of Freeling (2009). In *Brassicaceae*, LF single-copy genes are reportedly enriched in biological processes such as DNA repair and RNA interference, with the relevant genes preferentially deleted from MF1 and MF2 subgenomes (Hao, Mabry, et al., 2021). The gene balance hypothesis tries to explain such reported differential retention of certain functional groups of genes in pairs by emphasizing the protein network level. Transcription factors and transcription regulators operating in interaction within multiple protein-nucleic acid complexes are prone to being retained as pairs (or MF genes) after polyploidization (Conant et al., 2014; Freeling et al., 2015). By contrast, genes encoding proteins that function independently or are less associated with others in networks (proteins involved in DNA repair and metabolism, as well as RNA binding and interference), tend to retain only one of those copies (singletons of LF gnes) and loose the redundant ones (Freeling, 2009). Although the LF and MF blocks are conserved at orthologous position between the investigated species, indicating that the genome fractionation has been established before the grass speciation, LF–MF fraction exchanges between pairs of ancestral chromosomes can trace homoeologous chromosome exchanges during post-WGD evolution. Homoeologous exchange is common during early polyploid formation, and large chromosome segments from one subgenome can replace another as observed in *Brassica* (Stein et al., 2017), strawberry (Folta & Barbey, 2019), cotton (Page et al., 2016) and millet (Shi et al., 2019). In our present analysis of molecular evolution of duplicated genes and singletons, no clear distinction in non-synonymous substitutions (K_a) between pairs and singletons was determined, but pairs clearly displayed a significantly higher number of synonymous substitutions (K_s) compared to singletons. At the block level, LF-located genes showed significantly less synonymous substitutions than MF-located genes and non-synonymous substitutions ratios in favor of an excess of substitutions in MF genes. Phylogenetic tree branch lengths confirmed this general pattern, where pairs have evolved faster than singletons in more recent MRCAs (< 27 Myr ago) and modern species. Overall, our results suggest that evolutionary speed is affected by both whether a gene belongs to the LF/MF fraction of the genome, and whether the gene is conserved in multiple copies (pairs) or not (singletons) after the WGD event. Pairs display a significantly higher number of substitutions and longer branch length compared to singletons, and LF-located genes evolve slower than the MF-located genes, jointly making the copies of gene pairs that are located in the LF compartments, the most plastic fraction of the genome during grass evolution.

To what extent can this observed bias in molecular evolution between the LF and MF compartments be generalized across plant species? *Arabidopsis thaliana*, which experienced two polyploidization events (α and β, 24 and 40 Myr ago), retained between 23 and 29% of duplicated genes following the subgenome dominance process (Blanc et al., 2003; Thomas et al., 2006). The dominant subgenome (LF) has retained 70% of the genes since the paleoploidization event, whereas 46 and 36% of the genes have been retained in MF1 and MF2, respectively. In *Brassicaceae*, an ancestral triplication (WGT) dating back to 15 Myr ago was followed by subgenome dominance and a reported loss of 60% of the ancestral triplicates, with 50, 65 and 70% lost in the LF, MF1 and MF2 compartments, respectively (Cheng et al., 2012; Liu et al., 2014; Parkin et al., 2014). Among the genes that derive from the WGT, 13.42% of polyploidy-derived genes accumulate more transposable elements and non-synonymous mutations than other genes during evolution. Although no biased fractionation between subgenomes was detected in the allotetraploid Ethiopian cereal teff, a general subgenome dominance in the expression atlas across tissues was reported, with the B subgenome having overall higher homoeolog expression levels, perhaps as a result of its smaller size and fewer transposable elements (VanBuren et al., 2020). In maize, subgenome dominance following the lineage-specific duplication dating back to 5 Myr ago, documented in the current analysis, is in line with previous results based on 28.1% of genes retained in pairs, reporting that LF blocks contain more expressed genes, less transposable elements and accumulate fewer substitutions (Renny-Byfield et al., 2017; Schnable et al., 2011; Zhao et al., 2017). It has been reported that duplicated genes retained in rice are enriched in SNPs encoding less amino acid changes, suggesting that such 'advantageous' material/genes inherited from WGDs are highly stable (*i.e.*, slow rate of evolution) over the time (Chapman et al., 2006).

Omics reprogramming following polyploidization

Duplicated genes and blocks have been proposed as a reservoir of alternative variants of gene regulation in terms of expression and methylation. In the current analysis, we established that half of the duplicated genes (after the ancient and recent WGDs) display gene expression and DNA methylation differences, and that pairs display less methylation, less expression, more SNPs (in maize for both ancient and recent polyploidization events), lower gene-body methylation level (CG-CHH-CHG contexts) and higher rates of synonymous and non-synonymous substitutions (except for the recent duplication in Maize) compared to singletons. At the post-polyploidization block level, MF genes show less expression with no methylation bias compared to LF genes. Such divergence in expression and methylation between duplicated blocks and genes has been proposed as a source of subfunctionalization (partitioning of ancestral functions between the duplicated

genes) and neofunctionalization (gain of a non-ancestral function in one duplicate), both being key forces of evolutionary plasticity in plants (Zou et al., 2009). To what extent can such differences in regulation between duplicated blocks and genes be considered as a general post-polyploidization phenomenon in plants?

In *Arabidopsis thaliana*, 74% of genes are conserved in pairs and 3% display difference in expression (Coate et al., 2020; Duarte et al., 2006) with singletons more expressed than pairs (Wang et al., 2012). In cotton, LF-located genes are reportedly more expressed than the MF-located genes, and are associated with more TEs and small interfering RNAs that may play a role in reducing or modulating the expression of flanking genes (Renny-Byfield et al., 2015). The expression of pairs in three tissues (petals, grains and leaves) were also investigated and it was established that, among 2000 gene pairs deriving from an ancestral duplication, more than 99% display expression difference between at least one of the tissues and 93% between the three tissues. Among 1971 pairs active during the grain development (at 10, 20, 30 and 40 days post-anthesis), 84% display expression differences (Renny-Byfield et al., 2014). In maize, 65% of pairs show expression differences on average across eight tissues (Schnable et al., 2011), with expression dominance for 60% of the pairs in favor of one subgenome and 40% for the other, depending on the tissue considered (Zhao et al., 2017). However, no expression difference between maize subgenomes (Li et al., 2016) and no methylation difference (Renny-Byfield et al., 2017) have also been reported. Nonetheless, the dominant (LF) subgenome has consistently been observed to have lower methylation upstream of genes, in particular lower CHH methylation around 500 bp upstream of the TSS (Renny-Byfield et al., 2017; Zhao et al., 2017). Overall, the LF fraction in maize has been also demonstrated to have more tandem gene duplications (Edger et al., 2019), higher gene expression (Schnable et al., 2011) and lower DNA methylation (Woodhouse et al., 2014). In soybean, that experience a lineage-specific duplication 5–13 Myr ago, 80% of duplicated genes are conserved, with 40% of them showing expression difference but no bias toward any of the subgenomes (Zhao et al., 2017). The duplicated genes from the two subgenomes showed no differences in the distance to the nearest TE or in methylation levels (Zhao et al., 2017). Although only a few genes in the soybean genome have a TE nearby (< 1 kb), highly expressed soybean genes tend to be further away from TEs (Zhao et al., 2017). In *Brassicaceae*, LF-located genes are more expressed, less methylated and accumulate less SNPs than MF-located genes (Cheng et al., 2016). Parkin et al. (2014) reported that among the triplets inherited from the ancestral triplication in *Brassica*, 83% are differentially expressed in the leaf, with LF-located genes being more expressed but showing no difference in DNA methylation. *Brassica napus*, derived from a hybridization between *Brassica rapa* (An) and *Brassica olerace* (*Cn*) some 12 500 years ago, consists of MF (An) and LF (Cn) subgenomes, with 58% of pairs showing no expression difference in leaves and roots, and no subgenome dominance in expression bias for the remaining pairs associated with expression differences (Chalhoub et al., 2014). Li et al. (2020) investigated four tissues to conclude that between 70 and 85% of the retained pairs have a bias in expression toward the An subgenome depending on the tissue considered. Globally, CG, CHG and CHH methylation of genes and long terminal repeat retrotransposons are higher in the dominant BnC (LF) subgenome compared to the BnA subgenome, similarly to the results from natural *B. napus* (Chalhoub et al., 2014). Zhang et al. (2021) confirmed the previous analysis in showing that singletons are less expressed and more methylated than pairs, and that singletons in Cn (LF) are more methylated than singleton in An (MF). In pear, that experienced a polyploidization some 30 Myr ago, duplicated blocks have been reported with no difference in gene loss, substitutions, expression and methylation (Li et al., 2019). Singletons in pear are more expressed with transcripts detected in a wider range of tissues, more methylated (CG in promoters and CHG in promoters and gene bodies) and have more substitutions (K_a), as well as K_a/K_s, compared to pairs. On the other hand, 54% of retained pairs show expression differences. In *Nelumbo nucifera*, that experienced a paleotetrapolyploidization 60 Myr ago, MF-located genes are more methylated and more expressed than LF-located genes, while singletons are more expressed in a wider range of tissues, more methylated (gene body), and have less substitutions compared to pairs (Shi et al., 2020). In *Mimulus peregrinus* resynthesized and natural polyploids, the dominant subgenome had lower TE density and tended to have equal or lower expression compared to the non-dominant subgenome (Edger et al., 2017). Extensive homoeolog expression bias is also observed in hexaploid wheat (Ramírez-González et al., 2018) and tetraploid *Tragopogon mirus* (Buggs et al., 2010) and may be a common feature of recent polyploid grasses. In switchgrass (Lovell et al., 2021), the K subgenome (LF) has higher gene density, more upregulated genes and lower substitution rates compared to the N subgenome, pointing to a stronger evolutionary constraint of the K subgenome and suggesting that the potential for adaptive evolution may be differentially partitioned between subgenomes. *In the Cucurbita* genus (Sun et al., 2017), with an allotetraploidization that happened between 3–26 Myr ago, the two subgenomes have retained similar numbers of genes, and no subgenome being globally dominant in gene expression. Integrating studies from 20 angiosperm species, DeSmet et al. (2013) reported that singletons are more methylated and more expressed than pairs. Wang et al. (2017b) investigated CG

methylation of pairs in rice and concluded that genes showing methylation divergence also show expression differences. Investigations of the expressional dynamics of grass duplicates deriving from a 90 Myr ago paleotetraploidization event suggest that 57.4% (Yim et al., 2009) and up to 85% (Throude et al., 2009) of rice paleoduplicates have diverged in expression. In rice, retained ancient gene duplicates associated with high expression tend to have higher CG body methylation (Wang et al., 2013), suggesting a direct role of epigenetic regulation in structural and expressional maintenance of duplicates, preventing pseudogenization, silencing and deletion, and ultimately retaining WGD-derived genes. It has also been reported that dominantly expressed genes in D/LF subgenomes have fewer 24-bp RNAs in their 1-kb flanking regions compared to their S/MF paralogs (Woodhouse et al., 2014).

Omics regulation interplay in grasses

In grasses, we have provided evidence of the link between nucleotide substitutions, DNA methylation and genes expression. We report a negative correlation between K_a and both gene expression and DNA methylation, highlighting the interplay between (non-synonymous) nucleotide substitutions and gene regulation. In mollusk (*Crassostrea* giga), Song et al. (2018) reported a negative relationship between gene expression and non-synonymous sequence diversity, which suggests that purifying selection has played an important role in shaping genetic diversity. Highly expressed genes are often subject to strong selective constraints, and tend to evolve more slowly (Liao & Zhang, 2006). Although negative correlation between gene expression and non-synonymous substitutions is more obvious, the negative correlation between gene-body methylation and non-synonymous substitutions reported in the present study is more surprising. Most other studies found positive association of synonymous substitutions with gene-body methylation (Glastad et al., 2016; Lian et al., 2020). The negative link between gene-body methylation and non-synonymous substitutions in grasses found here may suggest that DNA methylation plays a protector role by preventing non-synonymous mutations, despite the potential for increased mutation rates in methylated CpG dinucleotides (Chuang et al., 2012; Monroe et al., 2022). Indeed, methylated genes have been suggested to be more functionally important than unmethylated genes, and then tend to evolve slowly (Sarda et al., 2012; Takuno & Gaut, 2012). Another explanation could be that methylated genes that are functionally important such as the housekeeping genes, may be under purifying selection, and therefore have less non-synonymous mutations.

To go further, we also examined the link between DNA methylation and gene expression. DNA methylation is a heritable epigenetic modification that has been shown to impact gene expression in plants and animals (Zhang et al., 2008). In *Arabidopsis*, DNA methylation of coding regions is associated with genes that are expressed at medium to high levels and enriched for housekeeping functions, whereas methylation of promoter regions is generally associated with gene repression or silencing (Bewick & Schmitz, 2017). Although no whole-genome correlation between DNA methylation and gene expression was detected in the present study, we show here, for a subset of genes, a negative link between promoter DNA methylation and gene expression, supporting the idea that DNA methylation plays a key role in gene silencing but only for the highly methylated ones (Suzuki & Bird, 2008). We report a clear methylation dosage effect on gene expression with a strong inverse correlation between promoter methylation and gene expression, and with a major reduction in gene expression (up to silencing) when promoter methylation reaches above 50%. In respect to the DNA methylation of gene bodies, which has been reported to be positively correlated with gene expression, we demonstrate here that this relationship depends on the methylation level, with high gene-body methylation (> 58%) associated with transcriptional repression. Finally, we have highlighted that these interplays mark species-specific genes (and not conserved genes), which might suggest species specificity in gene expression regulation mediated by DNA methylation modifications.

Emergence of key traits through polyploidization

What are the consequences of polyploidization in biological and phenotypic innovation? Reports highlight the role of polyploidization on species diversification and crop domestication, as well as in the establishment of important agronomic traits including stress resistance (Hannweg et al., 2016; Hias et al., 2018; Renny-Byfield & Wendel, 2014). In maize, the dominant subgenome contributes more to trait heritability than the non-dominant subgenome (Renny-Byfield et al., 2017) and, in strawberry, the dominant subgenome largely controls several biological pathways related to agriculturally valuable traits such as fruit flavor, color and aroma (Edger et al., 2019). A recent study showed that subgenome-specific selection of defense response genes contributes to the environmental adaptation of allopolyploid *Brassica napus* (Lu et al., 2019). In *Brassicaceae*, large-scale resequencing revealed that parallel selection of homoeologous genes derived from polyploidization is associated with morphotype diversification (leafy head) in *B. rapa* and *B. oleracea* (Cheng et al., 2016). In *Brassicaceae*, GO terms related to the phosphate biology process and root development were enriched in over-retained genes following the ancestral WGT. In the cotton genome, a search for genes responsible for long white fibers revealed 620 homoeologous pairs that have been subjected to domestication selection either in the A or D subgenome, whereas only 34 homoeologous pairs

exhibit selection signals in both subgenomes, indicating that the coexisting subgenomes have been under asymmetrical domestication selection (Wang et al., 2017a). By comparing the distributions of positively selected genes in cotton as well as fiber-related quantitative trait loci, Zhang et al. (2015) concluded that the A subgenome was selected for fiber improvement genes and the D subgenome was selected for stress tolerance genes. Polyploidy has been also proposed to give rise to aromatic profiles of strawberry fruits (Ulrich & Olbricht, 2013). In lupin, a dinitrogen-fixing legume that evolved from a WGT event, the number of phosphorus-use efficiency genes has increased through WGT, tandem and dispersed duplications, with WGT-derived genes enriched in GO terms related to the phosphate biology process and root development (Xu et al., 2020). In switchgrass, 75.9% of biomass SNP-heritability was attributable to the N subgenome, and only 24.1% to the K subgenome across 10 common garden experiments (Lovell et al., 2021). In the present study, we were able to identify key genes of agronomic importance with particular omics patterns when assessing methylation and expression signature of specific and recently duplicated genes in maize. Most of the duplicated genes with an identical methylation and expression signature for both copies were hypomethylated and expressed, and involved in seed development and growth. This is the case for: (i) ORMDL family gene, where reduction in gene expression resulted in sterile phenotype with abnormal pollen morphology and staining in rice (Chueasiri et al., 2014); (ii) AGL4/SEP2, where reduction in SEP activity led to the loss of normal ovule development (Favaro et al., 2003) and can subsequently influence the architecture of the inflorescence (Ma et al., 2022); (iii) SYP131, where triple mutant syp124 syp125 syp131 resulted in gametophytic defect (Slane et al., 2017); (iv) bHLH, a transcription factor involved in plant growth and development (Hao, Zong, et al., 2021; Zhou et al., 2020); (v) PYL8, where overexpressing plants exhibit hypersensitive phenotype to ABA in seed germination, seedling growth and establishment (Lim et al., 2013); and (vi) ROPGEF14, where ropgef1, ropgef9, ropgef12 and ropgef14 quadruple mutant showed reduced pollen tube elongation (Chang et al., 2013); etc. We show that all these genes are duplicated in maize and expressed during grain development with no or low methylation impact for both pairs. For three maize gene pairs, one copy was hypomethylated and overexpressed, whereas the other copy of the pair was hypermethylated and underexpressed. This was the case for: (i) ADF6 where downregulation in cotton rendered the plant tolerant to *V. dahlia* infection (Sun et al., 2021); (ii) FBP7, where downregulation resulted in maternally controlled defects in seed development in Petunia (Cheng et al., 2000; Colombo et al., 1997); and (iii) HSR8, involved in sugar responsive growth and development (Li et al., 2007) or stress tolerance (Zhao et al., 2019). Finally, four maize gene pairs were hypermethylated and silenced for both copies. Two of them are unknown genes and the remaining two correspond to FT and GER1/GLP1 genes. FT modulates flowering transition and inflorescence architecture (Wickland & Hanzawa, 2015), whereas GER1 is involved in stress response (Ilyas et al., 2016) and is expressed during fruit development and ripening in plum (El-Sharkawy et al., 2010). These observations open up possibilities for exploitation of DNA methylation in breeding, where manipulation of the methylation pattern of duplicated genes (e.g. by gene editing techniques) could bring about changes in gene expression that drive agronomically important traits, such as yield and resistance to biotic and abiotic stresses.

Model of polyploidization-driven genomic plasticity

In Figure 5(a), we propose a model of possible association between WGD and the potential consequences on genome reprogramming based on (i) significant differences observed between duplicated regions or genes and (ii) consistent differences observed across species and WGD events (Figure 5a). Following such rules, our model proposes that (i) inverted genes have higher substitution rates, are less methylated and more expressed (#1 in Figure 5)]; (ii) conserved genes have lower substitution rates, are less methylated and more expressed (#2 in Figure 5); (iii) pairs have more substitutions, are less methylated and less expressed (#3 in Figure 5); (iv) LF-located genes have less substitutions with no clear consensus on expression or methylation bias between LF/MF-located genes (#4 in Figure 5); and (v) more than 50% of the pairs show within-pair expression differences (#5 in Figure 5), MF-located and paired genes are enriched in transcription factor (TF) and regulator (TR) activity while LF-located and singleton genes are enriched in signaling (transport, transfer) activity (#6 in Figure 5).

Structural and functional post-polyploidy changes driven by subgenomic dominance seem to require some time to 'evolve' and 'stabilize', as suggested by the current analysis, as well as by results on resynthesized polyploids (Renny-Byfield et al., 2015). Consequently, such genomic reprogramming following polyploidization may lead to novelty at (i) the functional level (neo- and subfunctionalization, Lynch & Conery, 2000; Wang et al., 2012); (ii) the network level with the maintenance of a stoichiometric balance of gene product interactions (or connectivity) in macromolecular complexes (Birchler & Veitia, 2014); (iii) the phenotypic level with an heterosis effect with transgressive performances (Birchler et al., 2010); (iv) the allelic level in masking deleterious recessive mutations (Gu et al., 2003); (v) the adaptation level with escape from adaptative conflicts (Des Marais & Rausher, 2008); and (vi) the regulation level with novel expression and methylation patterns (Aury et al., 2006; Yang & Gaut, 2011), all potentially contributing

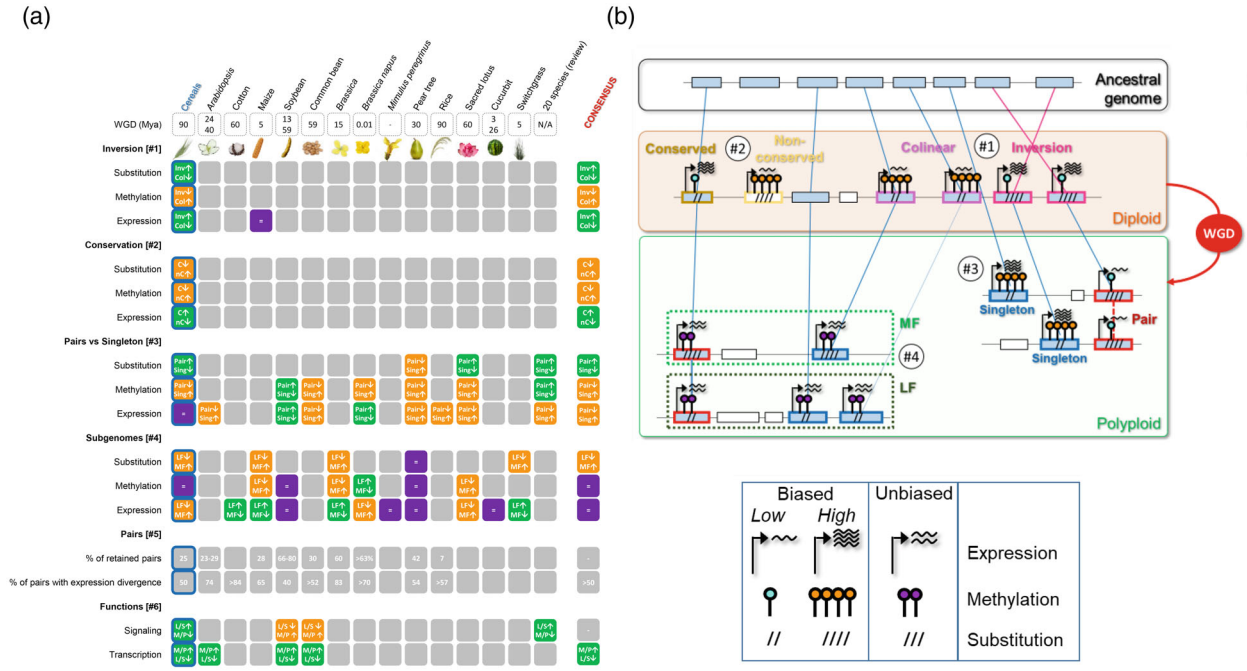


Figure 5. Model of post-polyploidization genomic reprograming. (a) Comparative analysis of the omics variation reported for inverted (referenced to as INV for inverted and COL for colinear, genes), conserved (referenced to as C for conserved and nC for non conserved, genes), duplicated (referenced to as Pair for duplicated and Sing for non-duplicated, genes), LF-/MF-located genes (referenced to as LF for least fractionated and MF for most fractionated compartments), gene pairs (with the percentage of retained pairs in the genomes and the percentage of pairs showing expression differences), and functions (referencing signaling such as transport and transfer activity and transcription such as TF and TR activity) in LF/Singleton (L/S) and MF/Pair (M/P) (lines) from the current analysis (first column) and the literature from 51 angiosperm species. When similar conclusions are made in other species compared to the present study, the color of the first column is applied to the corresponding columns. When opposite conclusions are made, different color is used. No observed differences are indicated in purple. Consensus conclusions (last column on the right) are drawn for omics trends predominantly observed in response to angiosperm paleo- and neo-polyploid events. (b) Schematic model of post-polyploidization genome reprogramming for inverted, conserved, duplicated, LF- and MF-located genes, with genes shown as colored rectangles and substitutions, expression and methylation according to the legend at the top right.

to a polyploid machinery absent from the diploid progenitors. Looking from the perspective of evolutionary outcomes backwards, it appears that post-polyploidization changes are established to strike a balance between stability and novelty. On the one hand, structural and functional partitioning of the subgenomes reinforces the differentiation of homoeologous chromosomes (subgenomes), leading to stabilization of meiotic pairing and increased fertility of the nascent polyploids by prevention of homoeologous pairing. On the other hand, these changes provide genetic redundancy 'geared' toward phenotypic plasticity and novelty that could be advantageous in many ecological contexts. However, we need to be cognisant of the fact that, by looking at the extant polyploids, we are only assessing the successful polyploid lineages and ignoring an unknown number of polyploidizations that disappeared as evolutionary dead ends, potentially creating an illusion that polyploidy is always associated with evolutionary advantage. We can speculate that striking the right balance between stability and novelty after WGD is rare, and the extant paleopolyploid lineages achieved their evolutionary success through independent trajectories, albeit showing some level of convergence. This view is consistent with the results presented here, demonstrating that diverged polyploid lineages sharing a common WGD event often present the same patterns of structural changes and evolutionary dynamics, although these patterns are difficult to generalize across independent WGD events. The lack of general patterns could partially be a result of confounding non-WGD factors (differences in TE and recombination landscapes, epigenetic regulation, GC-content, efficiency of DNA repair mechanisms, founder effects, selection constraints and other population-genetic forces) operating over millions of years after polyploidization. Nonetheless, a common mechanism of biased fractionation remains enigmatic and it appears that polyploids harness the potential of genome duplication, at least partially, in lineage-specific ways. Given that probably all grasses are derived from the ρ paleopolyploidization, it is impossible to imagine the diversification and the existence of this clade without some relation to WGD. In this sense, WGD is unequivocally linked to the evolutionary success of grasses during the past 100 Myr, although it remains difficult to attribute this success to particular genomic consequences

of polyploidization. Especially, it would be important to precisely distinguish genomic reprogramming, as directly inherited from WGD, from other processes that also impacted genome organization and regulation, such as natural selection, adaptation to natural environmental stresses and domestication. There is also the desire to distinguish genomic reorganization directly caused by WGD from other processes that can impact the genome structure and expression in non-random ways, such as natural and artificial selection. Moreover, it should always be considered that some of the differences in genome fractionation commonly attributed to WGD could be merely a reflection of differences that pre-existed in the parental genotypes of the founding polyploids.

Overall, the present study clearly demonstrates that post-polyploidization reprogramming is more complex than the traditionally reported differentiation of subgenomes and singletons-pairs. Although statistically significant differences in gene expression, DNA methylation and substitution rates can be identified in various post-WGD fractions of the extant genomes, patterns that are strictly consistent across different clades and WGDs are rare, with the clearest distinction in LF–MF fractions. The lack of generalized patterns of WGD consequences highlights our poor understanding of polyploidy-driven evolution and calls for a critical and comprehensive comparison across independently polyploidized lineages. Nonetheless, we demonstrate that a detailed characterization of omics profiles across duplicated gene copies can be useful for our understanding of trait evolution in the polyploid context, particularly in polyploid crops where it could identify targets for breeding and improvement.

EXPERIMENTAL PROCEDURES

Karyotype reconstruction, evolution and rearrangements

Genomes

The different genomes used for the present study are *Brachypodium distachyon* (phytozome v9), *Oryza sativa* (phytozome v11), *Sorghum bicolor* (phytozome v9), *Zea mays* (phytozome v11), *Setaria italica* (phytozome v8), *Triticum aestivum* (Chinese spring v1.0), *Triticum dicoccum* (Zavitan WEW v1.0) and *Hordeum vulgare* (Morex v1.0).

Ancestral karyotype reconstruction

AGKs were reconstructed according to a two-stage procedure. Although the ancestral karyotype is reconstructed in the first stage, the ancestral gene content of such karyotypes, and the gene order, are inferred in the second stage. Stage 1, in which the ancestral karyotype (i.e. the ancestral chromosome number) is determined, is based on genome alignments, using cumulative identity percentage and cumulative alignment length percentage blast parameters (Salse et al., 2009). Conserved genes (i.e. putative protogenes or pPG) are identified from these alignments and pPGs conserved in all investigated species (i.e. core protogenes or core-pPG) are then extracted. These core-pPGs are used to identify synteny blocks (SBs) using DRIMM-SYNTENY (Pham & Pevzner, 2010), removing groups of fewer than five genes. SBs are then merged using MGRA (Alekseyev & Pevzner, 2009) based on chromosome-to-chromosome orthologous relationships between the compared genomes. Stage 1 thus yields the ancestral protochromosomes (also referred to as CARs), corresponding to independent sets of blocks sharing paralogous and/or orthologous relationships in modern species. Stage 2 of the procedure is aiming at ordering protogenes on the previously defined protochromosomes. Genes (pPGs) conserved between pairs of species but not constituting core-pPGs (used in stage 1) are integrated within SBs with DRIMM-SYNTENY and then mapped onto the protochromosomes, delivering an exhaustive set of ordered protogenes. Such approach allowed the reconstruction of a pre-WGD (AGK7) ancestor by merging the post-duplication regions into a pre-duplication CAR. In this particular case, following the two stages approach described previously, the karyotyping stage (i) is performed using genes retained as pairs in the post-WGD ancestors, and the enrichment stage (ii) is performed using the remaining singletons because duplicated genes may have returned to singletons as a result of fractionation. In this particular case, singletons from both paralogous blocks are intercalated in groups (no strict order) between conserved paralogs according to their current positions in the modern genomic regions.

Synteny breakpoints characterization

Alignment of the protogenes from the post-ρ ancestor (AGK12) to the genes of extant species allowed building an atlas of synteny breakpoints (i.e. ancestral chromosome fusion–fission sites). A transition between two ancestral chromosomes on an extant chromosome defined an SBP region bounded by two extant genes, orthologs of protogenes from two distinct ancestral chromosomes. Despite ancestral chromosome fusions–fissions, we investigated inversions detected from alignments of ordered protogenes of the post-ρ AGK12 and the gene order in the extant species investigated. Inversions were detected by analyzing the relative positions of genes on chromosomes of extant species compared to the inferred AGK order. Each gene has been tagged as non-inverted (collinear) or inverted relative to AGK. An inversion consists of the inversion of the order of two consecutive genes in a modern species compared to the ancestral gene order. To focus on large inversions, we set up a minimum threshold of 10 inverted extant genes compared to AGK12 to define an inversion. Inversions were clustered into inversion families based on the value of their Jaccard index that gauges the similarity in terms of AGK gene content between inversions. Two inversions were included in a family when their Jaccard index value was 0.6 or above. Inversions have been refined by merging manually some particular inversions initially detected. When two contiguous inversions in one species are orthologous to a single inversion in a related species, they are then merged into a single one to consider the common ancestry of the inversions for both species (i.e. belonging to the same inversion family). An inversion family was considered as included in another family when 90% of its AGK gene content is present in the second one. After merging of the inversions, a new step of inversion family clustering has been performed. At the end, based on the composition of the families, we have clustered the inversions according to their evolutionary origins and placed them on the phylogenic tree of grasses. For each branch of the phylogenic tree, the rate of inversions was calculated as the number of inversions over evolutionary time in million years. For each investigated species, we performed 1000 simulations where we randomized the position of the observed inversions along the genome (thus keeping the species specific

inversion number and size distribution constant) and subsequently computed the gene density (in number of genes per Mb) for the randomized inversions.

Genome fraction inference

In each extant genome, dominance at the gene level was inferred by comparing counts of descendant genes in different extant chromosomes over non-overlapping windows of 100 ancestral genes from the pre-WGD ancestral grass karyotypes (AGK7) for the eight investigated species and AGK12 for maize and wheat. For each extant genome, we categorized the genes as LF, MF and non-significant. For each ancestral window of 100 genes, the number of ancestral genes retained in the two corresponding duplicated blocks in the extant species where compared. Significantly biased retention corresponded to $P < 0.05$, where the *P*-value is the probability of observing at least this difference between two windows based on binomial distribution (where *n* is the windows size and *P* is the mean of the compared counts divided by *n*). This procedure was repeated on the genome of Maize, tetraploid wheat and hexaploid wheat based on AGK12 gene order to detect the LF/MF genes in these species.

Cereal phylogenetic history reconstruction

Homology detection and sequence alignment

We detected homology between the cereal genes by performing an all versus all blast (Altschul et al., 1990) for which the output was processed using Silix (Miele et al., 2011), followed by Hifix (Miele et al., 2012). The amino acid sequences of the generated homologous families were aligned using MUSCLE (Edgar, 2004). GBLOCKS (Castresana, 2000) with options –b5 = h –b4 = 2 was then used to extract conserved blocks from the multiple sequence alignments. This step identified 75 families presenting either no conserved blocks or at least one sequence absent from the detected conserved blocks. These cases likely resulted from spurious homology assignment and we decided to further refine them using a walktrap algorithm (Pons & Latapy, 2006), where the inverse of the pairwise poisson distance was used to give weights to the edges between proteins. This step yielded 175 additional homologous families with conserved blocks. The complete procedure resulted in 22 129 aligned homologous gene families (containing a total of 317 660 genes from the investigated species).

Reconciled gene tree inference

Within the *Triticeae*, given the limited divergence of the species under investigation here (e.g. potentially < 10 000 years have passed since the divergence between the A subgenomes of *T. dicoccum* and *T. aestivum*), we chose to perform gene tree inference by jointly taking into account sequence and reconciliation information, which have been demonstrated to result in better quality gene trees (Scornavacca et al., 2015; Szöllosi et al., 2013). Our specific approach here most closely resembles that of TERA (Scornavacca et al., 2015), but with the added possibility of a non-binary species tree (which is needed in our case because the speciation-hybridization history between the A, B and D wheat lineage is unresolved and/or complicated by reticulation events, (Glémin et al., 2019) and whole genome duplications (two recent WGDs in the case of hexaploid wheat). For each gene, a distribution of 1000 bootstrap trees was inferred using IQTREE (Hoang et al., 2018; Nguyen et al., 2015). This distribution was then used in our reconciliation procedure to compute the reconciled gene tree that minimizes the joint reconciliation sequence score. After this, optimal branch lengths were computed for the inferred topologies using gene family codon alignment and GTR-GAMMA from RAXML (Stamatakis, 2014).

Computing K_a and K_s

Codon multiple sequence alignments were computed from the protein alignments using PAL2NAL (Suyama et al., 2006). The number of substitutions per synonymous site and per non-synonymous site (K_a and K_s, respectively) were computed using the kaks function of SEQINR package (Charif & Lobry, 2007). The K_a or K_s profiles of different sets of genes were compared using the medians of the main peak of the distributions, which we evaluated as a more reliable measure than the distribution modes. The statistical significance of these differences was tested using a bootstrap procedure. Thousand bootstraps were performed for each comparison.

Inference of substitution rates in Triticeae

To gain further insight into the evolution of wheats, we directed particular attention on *Triticeae* species. We extracted 3905 gene families in which *Triticeae* (namely here Barley, Wheat4xA, Wheat4xB, Wheat6xA, Wheat6xB, Wheat6xD, *T. urartu* and *A. tauschii*) presented exactly one ortholog each. Recently, Naser-Khdour et al. (2019) showed that model violation in phylogenetic reconstruction was both prevalent and deleterious among partition-based phylogenetic analysis. Indeed, most models of sequence evolution make the assumption that sequence evolution is stationary, reversible and homogeneous (SRH conditions). Accordingly, we tested each of the 3905 gene family DNA alignments for violating the SRH assumptions, using scripts adapted from Naser-Khdour et al. (2019). In total, 930 gene families exhibited signs of SRH conditions violation and were thus excluded from our study. The remaining 2975 gene family DNA sequences were used to construct a partitioned multiple alignment, upon which the best tree topology and branch lengths were inferred using RAXML (Stamatakis, 2014), with 100 bootstraps (the topology was constrained so that only the speciation of the wheat A, B and D lineages were left uncertain). The optimal model for each partition was inferred using IQTREE (Nguyen et al., 2015).

Comparative SNP analysis

For maize, unimputed diversity data within the third generation *Z. mays* haplotype map (Bukowski et al., 2018) were downloaded from public sources. Individuals with outlier values [> third quartile + (1.5 × interquartile range)] of missingness were removed, together with sites showing outlier values of total depth. For Brachypodium, a custom vcf file was created. Chromosome-level assemblies for 54 lines of B. dystachyon (Gordon et al., 2017) were downloaded from public sources and shredded into 300-bp fragments with a 50-bp sliding step using the GenomeTools (http://genometools.org) shredder command. The shredded fragments were converted into de-duplicated pseudo-fastq fragments with TALLY (Davis et al., 2023) and mapped onto the common reference used throughout this study using BWA MEM (Li & Durbin, 2009). After coordinate-sorting (PICARD TOOLS SortSam, http://broadinstitute.github.io/picard/) and addition of read groups (GATK AddOrReplaceReadGroups, McKenna et al., 2010), vcf files were generated separately for each chromosome using GATK4.1.8 HaplotypeCaller, restricting the SNP calling to exon space only. For hexaploid wheat, unimputed SNP data produced by Pont et al. (2019a) were used. For each species, per-gene SNP densities were calculated as the number of SNPs (> 0.05 minor allele frequency) per 100 bp of coding sequence, using the vcf files described above, relevant gene annotation files, and the

commands intersect and groupby from the BEDTOOLS suite (Quinlan & Hall, 2010).

Expression data

RNA sample preparation

Seeds from *Brachypodium* (genotype BD21-3), maize (genotype FV2) and wheat (genotype Recital) were grown and pollinated under standard conditions. Grains were sampled during the grain development at three stages (stage 1 for the cell division phase; stage 2 for the storage protein accumulation phase; and stage 3 for the dehydration phase): 9, 16 and 28 days after pollination in *Brachypodium*, at 7, 15 and 35 days after pollination in maize and at 100, 250 and 500 degree days in wheat. These developmental stages have been determined to be equivalent by morphological, biochemical and genetic criteria. Each sample was a pool of at least 10 whole seeds collected from at least two different plants for RNA extraction. Grains (approximately 1 g of tissue) were ground in liquid nitrogen and were extracted with 4.5 ml of buffer (10 mM Tris–HCl, pH 7.4, 1 mM EDTA, 0.1 M NaCl, 1% sodium dodecyl sulfate) and 3 ml of phenol–chloroform–isoamyl alcohol mixture 25:24:1. The supernatant was extracted one more time with the same phenol solution in order to eliminate proteins and starch. The nucleic acids were precipitated by addition of 0.1 volumes of 3 M sodium acetate (pH 5.2) and 2 volumes of 100% ethanol. After precipitation, RNA was rinsed one time with 70% ethanol and the pellets dissolved in RNase-free water. Purification was performed with a DNAse treatment RNase-Free DNase Set (Qiagen, Hilden, Germany), followed by the RNeasy MinElute Cleanup Kit (Qiagen). The integrity of RNA was checked with an Agilent 2100 Bioanalyzer microfluidics-based platform, using RNA 6000 Nano Chip kit and reagents (Agilent Technologies, Santa Clara, CA, USA).

Control genes used for comparative grain kinetics

Quantitative real-time PCR expression profiles have been produced with control genes known to be specifically expressed during cell division (*SUBTILISIN* gene corresponding to TaAffx.79990.1.S1 in wheat, GRMZM2G039538 in maize and Bradi1g08670-08450 in *Brachypodium*), filling (*Opaque2/SPA* gene corresponding to TaAffx.15974.1.S1 in wheat, GRMZM2G015534 in maize and Bradi1g55450 in *Brachypodium*) and dehydration (*Rab17* gene corresponding to TaAffx.58091.1.S1 in wheat, GRMZM2G079440 in maize and Bradi4g22280-22290-Bradi3g43870 in *Brachypodium*) phases of the grain development in grasses. RNA extractions were performed according to the ZR Plant RNA MiniPrep™ (Zymo Research, Irvine, CA, USA) protocol; DNase set (Qiagen) was used for RNA purification. When necessary, the residual DNA in the RNA samples was removed using the DNase set (Qiagen) and RNeasy Minelute Cleanup (Qiagen) kits. cDNA synthesis was performed with a Transcriptor first strand cDNA synthesis kit (Roche, Basel, Switzerland) in accordance with the manufacturer's instructions with 0.5 μg of RNA in 20 μl of reaction. The thermal cycling conditions were 10 min at 25°C/30 min at 55°C/5 min at 85°C. The reverse transcription reaction was diluted to a final volume of 200 μl, and 4 μl of synthesized cDNA was used as a template for a real-time PCR using a LightCycler® 480 instrument (Roche). The reactions were performed in a 10-μl volume comprising 1 × LightCycler® 480 DNA SYBR Green I Master (Roche). The quantification cycles (Cq) were analyzed using the LightCycler®480 software version 1.5.0 and normalized with reference genes using an 'Advanced Relative Quantification' profile to obtain a normalized ratio E_t^{-CqR}/E_r^{-CqR} [where CqT/CqR is the cycle number at target/reference detection threshold (crossing point) and E_t/E_r is the efficiency of target/reference amplification (10^{-1}/slope)]. The specificity of the amplification was confirmed via melting curve analysis of the final PCR products by increasing the temperature from 65°C to 95°C. PCR efficiency was calculated for each gene using a standard curve of serial dilutions and was used in the relative expression analysis. All observations were expressed as the mean ± SD.

RNA library construction and sequencing

The total RNAs were analyzed with a capillary MultiNA microchip electrophoresis system (Shimadzu, Kyoto, Japan). The total RNAs were first sheared with ultrasound (three pulses of 30 sec at 4°C). From the sheared total RNA samples, the 3′ polyA+ fragments were purified by means of oligo(dT) chromatography. An RNA adapter was then ligated to the 5′-phosphate of the 3′ fragments. First-strand cDNA synthesis was performed using an oligo(dT)-adapter primer and M-MLV H-reverse transcriptase. The resulting cDNAs were PCR-amplified using a high-fidelity DNA polymerase. For Illumina sequencing, the cDNA fractions in the size range of 250–400 bp were eluted from preparative agarose gels. Aliquots of the size fractionated cDNAs were analyzed by capillary electrophoresis. RNA-seq was performed separately for the considered samples/libraries on an illumina HiSeq (Illumina Corp., San Diego, CA, USA).

RNA-seq analysis

RNA-seq reads were cleaned using CUTADAPT (Martin, 2011) prior to mapping on the reference genomes using HISAT2 (Kim et al., 2019) with default parameters. Expressed genes were identified using TPM (Li & Dewey, 2011) threshold (> 1). Differential expression analysis was performed using EDGER (Robinson et al., 2010). Genes with less than five counts per million were dropped before library normalization using the trimmed mean of M-values method. Count distribution was assessed with generalized linear model and dispersion estimated according to McCarthy et al. (2012). A gene was considered as differentially expressed (likelihood ratio test) if its adjusted *P*-value (Benjamini–Hochberg) was below 0.05. Genes were scored according to their expression profiles with, for example, a profile 0–1–0 corresponding to an expression observed only at stage 2.

Methylation data

WGBS library construction and sequencing

Whole-genome bisulfite sequencing was performed by the Integragen laboratory (Evry, France) using the same developmental stage seeds described previously. Extracted genomic DNA was fragmented to approximately 200 bp, and methylated adapters compatible with sequencing on an Illumina HiSeq instrument were ligated. The resulting libraries were then bisulfite converted (EpiTect; Qiagen) and purified. A Real-time PCR assay was used to determine the optimal number of PCR amplification cycles required to obtain a high diversity library with minimal duplicated reads. The sequencing was realized with paired ends (2 × 100 bp) on an Illumina HiSeq 4000 platform with a minimal theoretical coverage of 15×.

DNA methylation analysis

The quality of raw reads was first checked with FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), version 0.11.7, before adapter trimming (illumina adapters), base quality (> 28) and 5′ trimming (6 bp, to reduce methylation calling bias) with TRIM GALORE, version 0.5.0 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Trimmed reads were then

mapped to corresponding reference genomes used in this study, using BISMARK, version 0.22.3 (Krueger & Andrews, 2011) under BOWTIE2, version 2.3.4.3 (Langmead et al., 2009). All DNA methylation contexts (CG, CHG and CHH) were extracted using default options of the bismark workflow. The methylKit R package (Akalin et al., 2012), version 1.18.0, was used to annotate methylation data into gene features: promoter (500 bp around the TSS) and gene-body (exons + introns) for each methylation contexts. With methylation expressed in percentage, a new normalization approach for DNA methylation was used to assess the link between gene expression and DNA methylation. We call this normalization rbd, for read by density: rbd = *mCs* × ratio (0–1) where ratio = *mCs*/(*mCs* + *Cs*). DMGs were assessed using edgeR package with rpd data (Chen et al., 2018; Robinson et al., 2010). Methylation differences were assessed using the likelihood ratio test and *P*-values adjusted by the Benjamini–Hochberg procedure to control the false discovery rate. Methylation differences were considered when adjusted *P*-values were below 0.05. Genes were scored according to their methylation profiles with, for example, a profile 0–1–0 corresponding to methylation observed only at stage 2.

Statistical analysis

Statistical analyses were performed via R software under RSTUDIO (R Core Team, 2018, https://www.R-project.org/). Differences between evolutionary structural features were assessed via Kruskal–Wallis or Wilcoxon tests, a *t*-test and/or analysis of variance (ANOVA). Tukey's post-hoc test was used when significant results were computed via ANOVA. *P* (false discovery rate) < 0.05 was considered statistically significant.

AUTHOR CONTRIBUTIONS

JT, VV, NDF, LM, BE, BD, PR and HB produced the plant samples for omics characterization and participated in the manuscript writing. CP, PC, EM, WD, DA and CH participated in the analysis of the omics data and manuscript writing. AB, MDS and JS conceived the experiment, managed the project, analyzed the data and wrote the article.

ACKNOWLEDGEMENTS

Completion of this article was supported by the 'Région Auvergne-Rhône-Alpes' and FEDER 'Fonds Européen de Développement Régional' (#23000816 project SRESRI 2015), the Institut Carnot Plant2Pro (#0001455 project SyntenyViewer 2017), The ISITE CAP2025 (#00002146 SRESRI 2015 'Pack Ambition Recherche Project' TransBlé 2018).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data described in the current manuscript are available in the Supporting information (Data S1 and S2) and on repository under the BioProject ID PRJNA949445.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Dotplot based deconvolution of modern genomes according to inferred AGKs.

Figure S2. Gene Ontology enrichment.

Figure S3. SBPs between extant cereal genomes and AGK12.

Figure S4. Inversions in grass genomes.

Figure S5. Inversions between hexaploid wheat and tetraploid wheat.

Figure S6. Phylogeny and datation for substitution rate analysis.

Figure S7. LF/MF compartments.

Figure S8. Analysis of omics data between collinear versus inverted genes, singletons versus genes in pairs and LF versus MF genes for wheat, maize and *Brachypodium*.

Figure S9. Subgenome dominance in hexaploid wheat between A-B-D subgenomes.

Figure S10. Subgenome dominance in hexaploid wheat between recent A-B-D subgenomes for each ancient LF-MF subgenomes.

Figure S11. Impact of TEs in methylation and expression differences observed between conserved genes.

Figure S12. Plasticity of duplicated genes (K_a, K_s).

Figure S13. Whole genome omics interplay.

Figure S14. Expression and methylation levels of conserved and non-conserved genes.

Figure S15. Expression and methylation outliers from mixOmics.

Data S1. Comparative genomics data between 10 grass species.

Data S2. Omics data available for wheat, maize and *Brachypodium*.

Table S1. Catalog of SBPs in cereal genomes compared to AGKs.

Table S2. Catalog of inversions (INVS) in cereal genomes compared to AGKs.

Table S3. Partitioning of genomes according to recombination rates.

Table S4. LF/MF and singleton/pair contents in cereal genomes.

Table S5. Branch length values for cereals and Triticeae phylogenetic trees.

Table S6. Substitution rate and branch length in cereals for LF/MF and singleton/pair genes.

Table S7. Global DNA methylation in *Brachypodium* and maize.

Table S8. Comparison of omic features between wheat subgenomes (A, B, D and LF and MF).

Table S9. Expression and methylation profiles of conserved genes.

Table S10. Maize duplicated genes expression and methylation regulation for key trait-driving genes.

REFERENCES

Acharya, D. & Ghosh, T.C. (2016) Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. *BMC Genomics*, **17**, 1–14.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A. *et al.* (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, **13** (10), R87.

Alekseyev, M.A. & Pevzner, P.A. (2009) Breakpoint graphs and ancestral genome reconstructions. *Genome Research*, **19**, 943–957.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M. *et al.* (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*, **444**, 171–178.

Bewick, A.J. & Schmitz, R.J. (2017) Gene body DNA methylation in plants. *Current Opinion in Plant Biology*, **36**, 103–110.

Birchler, J.A. & Veitia, R.A. (2014) The gene balance hypothesis: dosage effects in plants. *Methods in Molecular Biology*, **1112**, 25–32.

Birchler, J.A., Yao, H., Chudalayandi, S., Vaiman, D. & Veitia, R.A. (2010) Heterosis. *Plant Cell*, **22**, 2105–2112.

Blanc, G., Hokamp, K. & Wolfe, K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Research*, **13**, 137–144.

Buggs, R.J.A., Elliott, N.M., Zhang, L., Koh, J., Viccini, L.F., Soltis, D.E. *et al.* (2010) Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid Tragopogon mirus. *The New Phytologist*, **186**, 175–183.

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z. *et al.* (2018) Construction of the third-generation Zea mays haplotype map. *Gigascience*, **7**, 1–12.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540–552.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A.P., Tang, H., Wang, X. *et al.* (2014) Early allopolyploid evolution in the post-neolithic Brassica napus oilseed genome. *Science*, **345**, 950–953.

Chang, F., Gu, Y., Ma, H. & Yang, Z. (2013) AtPRK2 promotes ROP1 activation via RopGEFs in the control of polarized pollen tube growth. *Molecular Plant*, **6**, 1187–1201.

Chapman, B.A., Bowers, J.E., Feltus, F.A. & Paterson, A.H. (2006) Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 2730–2735.

Charif, D. & Lobry, J.R. (2007) SeqinR 1.0-2: a contributed Package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla, U., Porto, M., Roman, H.E. & Vendruscolo, M. (Eds.) *Structural approaches to sequence evolution. Biological and medical physics, biomedical engineering*. Berlin, Heidelberg: Springer. Available from: https://doi.org/10.1007/978-3-540-35306-5_10

Chen, Y., Pal, B., Visvader, J.E., Smyth, G.K., Andrews, S. & Macdonald, J.W. (2018) Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR, **F1000Research**, 1–40.

Cheng, F., Sun, R., Hou, X., Zheng, H., Zhang, F., Zhang, Y. *et al.* (2016) Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in Brassica rapa and Brassica oleracea. *Nature Genetics*, **48**, 1218–1224.

Cheng, X.F., Wittich, P.E., Kieft, H., Angenent, G., XuHan, X. & Van Lammeren, A.A.M. (2000) Temporal and spatial expression of MADS box genes, FBP7 and FBP11, during initiation and early development of ovules in wild type and mutant Petunia hybrida. *Plant Biology*, **2**, 693–702.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K. *et al.* (2012) Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PLoS One*, **7**(5), e36442.

Chuang, T.J., Chen, F.C. & Chen, Y.Z. (2012) Position-dependent correlations between DNA methylation and the evolutionary rates of mammalian coding exons. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 15841–15846.

Chueasiri, C., Chunthong, K., Pitnjam, K., Chakhonkaen, S., Sangarwut, N., Sangsawang, K. *et al.* (2014) Rice ORMDL controls sphingolipid homeostasis affecting fertility resulting from abnormal pollen development. *PLoS One*, **9**, e106386.

Coate, J.E., Farmer, A.D., Schiefelbein, J.W. & Doyle, J.J. (2020) Expression partitioning of duplicate genes at single cell resolution in Arabidopsis roots. *Frontiers in Genetics*, **11**, 596150.

Colombo, L., Franken, J., Van der Krol, A.R., Wittich, P.E., Dons, H.J. & Angenent, G.C. (1997) Downregulation of ovule-specific MADS box genes from petunia results in maternally controlled defects in seed development. *Plant Cell*, **9**, 703–715.

Conant, G.C., Birchler, J.A. & Pires, J.C. (2014) Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology*, **19**, 91–98.

Darwin, C.R. (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*, 1st edition. London: John Murray.

Davis, M.P.A., van Dongen, S., Abreu-goodger, C., Bartonicek, N. & Enright, A.J. (2023) Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods*, **63**, 41–49.

DeSmet, R., Adams, K.L., Vandepoele, K., van Montagu, M.C.E., Maere, S. & van de Peer, Y. (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 2898–2903.

Duarte, J.M., Cui, L., Wall, P.K., Zhang, Q., Zhang, X., Leebens-Mack, J. *et al.* (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. *Molecular Biology and Evolution*, **23**, 469–478.

Dvorak, J., Wang, L., Zhu, T., Jorgensen, C.M., Deal, K.R., Dai, X. *et al.* (2018) Structural variation and rates of genome evolution in the grass family seen through comparison of sequences of genomes greatly differing in size. *The Plant Journal*, **95**, 487–503.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 1–19.

Edger, P.P., Poorten, T.J., Van Buren, R., Hardigan, M.A., Colle, M., McKain, M.R. *et al.* (2019) Origin and evolution of the octoploid strawberry genome. *Nature Genetics*, **51**, 541–547.

Edger, P.P., Smith, R., McKain, M.R., Cooley, A.M., Vallejo-Marin, M., Yuan, Y. *et al.* (2017) Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell*, **29**, 2150–2167.

El-Sharkawy, I., Mila, I., Bouzayen, M. & Jayasankar, S. (2010) Regulation of two germin-like protein genes during plum fruit development. *Journal of Experimental Botany*, **61**, 1761–1770.

Escudero, M. & Wendel, J.F. (2020) The grand sweep of chromosomal evolution in angiosperms. *The New Phytologist*, **228**, 805–808.

Favaro, R., Pinyopich, A., Battaglia, R., Kooiker, M., Borghi, L., Ditta, G. *et al.* (2003) MADS-box protein complexes control carpel and ovule development in Arabidopsis. *Plant Cell*, **15**, 2603–2611.

Folta, K.M. & Barbey, C.R. (2019) The strawberry genome: a complicated past and promising future. *Horticulture Research*, **6**, 19–21.

Fox, D.T., Soltis, D.E., Soltis, P.S., Ashman, T.L. & van de Peer, Y. (2020) Polyploidy: a biological force from cells to ecosystems. *Trends in Cell Biology*, **30**, 688–694.

Freeling, M. (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annual Review of Plant Biology*, **60**, 433–453.

Freeling, M., Scanlon, M.J. & Fowler, J.F. (2015) Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. *Current Opinion in Genetics & Development*, **35**, 110–118.

Glastad, K.M., Gokhale, K., Liebig, J. & Goodisman, M.A.D. (2016) The caste- and sex-specific DNA methylome of the termite Zootermopsis nevadensis. *Scientific Reports*, **6**, 1–14.

Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M. *et al.* (2019) Pervasive hybridizations in the history of wheat relatives. *Science Advances*, **5**, 1–10.

Goidts, V., Szamalek, J.M., de Jong, P.J., Cooper, D.N., Chuzhanova, N., Hameister, H. *et al.* (2005) Independent intrachromosomal recombination events underlie the pericentric inversions of chimpanzee and gorilla chromosomes homologous to human chromosome 16. *Genome Research*, **15**, 1232–1242.

Gordon, S.P., Contreras-Moreira, B., Woods, D.P., Des Marais, D.L., Burgess, D., Shu, S. *et al.* (2017) Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nature Communications*, **8**, 2184.

Gu, Y.Q., Anderson, O.D., Londeorë, C.F., Kong, X., Chibbar, R.N. & Lazo, G.R. (2003) Structural organization of the barley D-hordein locus in comparison with its orthologous regions of wheat genomes. *Genome*, **46**, 1084–1097.

Hannweg, K., Steyn, W. & Bertling, I. (2016) In vitro-induced tetraploids of Plectranthus esculentus are nematode-tolerant and have enhanced nutritional value. *Euphytica*, **207**, 343–351.

Hao, Y., Zong, X., Ren, P., Qian, Y. & Fu, A. (2021) Basic Helix-Loop-Helix (bHLH) transcription factors regulate a wide range of functions in Arabidopsis. *International Journal of Molecular Sciences*, **22**, 7152.

Hao, Y., Mabry, M.E., Edger, P.P., Freeling, M., Zheng, C., Jin, L. *et al.* (2021) The contributions from the progenitor genomes of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible. *Genome Research*, **31**, 799–810.

Hias, N., Svara, A. & Keulemans, J.W. (2018) Effect of polyploidisation on the response of apple (malus × domestica Borkh.) to Venturia inaequalis infection. *European Journal of Plant Pathology*, **151**, 515–526.

Hidalgo, O., Pellicer, J., Christenhusz, M., Schneider, H., Leitch, A.R. & Leitch, I.J. (2017) Is there an upper limit to genome size? *Trends in Plant Science*, **22**, 567–573.

Hoang, D.T., Vinh, L.S., Flouri, T., Stamatakis, A. & von Haeseler, A. (2018) MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evolutionary Biology*, **18**, 1–11.

Ilyas, M., Rasheed, A. & Mahmood, T. (2016) Functional characterization of germin and germin-like protein genes in various plant species using transgenic approaches. *Biotechnology Letters*, **38**, 1405–1421.

Joron, M., Frezal, L., Jones, R.T., Chamberlain, N.L., Lee, S.F., Haag, C.R. *et al.* (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, **477**, 203–206.

Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, **37**, 907–915.

Landis, J.B., Soltis, D.E., Li, Z., Marx, H.E., Barker, M.S., Tank, D.C. *et al.* (2018) Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany*, **105**, 348–363.

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W. *et al.* (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, **574**, 679–685.

Leitch, A.R. & Leitch, I.J. (2008) Genomic plasticity and the diversity of polyploid plants. *Science*, **320**, 481–483.

Leitch, I.J. & Bennett, M.D. (1997) Polyploidy in angiosperms. *Trends in Plant Science*, **2**, 470–476.

Levin, D.A. & Soltis, D.E. (2018) ScienceDirect factors promoting polyploid persistence and diversification and limiting diploid speciation during the K – Pg interlude. *Current Opinion in Plant Biology*, **42**, 1–7.

Li, B. & Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 1–16.

Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, L., Briskine, R., Schaefer, R., Schnable, P.S., Myers, C.L., Flagel, L.E. *et al.* (2016) Co-expression network analysis of duplicate genes in maize (*Zea mays* L.) reveals no subgenome bias. *BMC Genomics*, **17**, 875.

Li, M., Wang, R., Wu, X. & Wang, J. (2020) Homoeolog expression bias and expression level dominance (ELD) in four tissues of natural allotetraploid Brassica napus. *BMC Genomics*, **21**, 1–15.

Li, Q., Qiao, X., Yin, H., Zhou, Y., Dong, H., Qi, K. *et al.* (2019) Unbiased subgenome evolution following a recent whole-genome duplication in pear (*Pyrus bretschneideri* Rehd.). *Horticulture Research*, **6**, 1–12.

Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H. *et al.* (2015) Early genome duplications in conifers and other seed plants. *Science Advances*, **1**, 1–7.

Li, Y., Smith, C., Corke, F., Zheng, L., Merali, Z., Ryden, P. *et al.* (2007) Signaling from an altered cell wall to the nucleus mediates sugar-responsive growth and development in Arabidopsis thaliana. *Plant Cell*, **19**(8), 2500–2515.

Li, Z., Tiley, G.P., Galuska, S.R., Reardon, C.R., Kidder, T.I., Rundell, R.J. *et al.* (2018) Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 4713–4718.

Lian, S., Zhou, Y., Liu, Z., Gong, A. & Cheng, L. (2020) The differential expression patterns of paralogs in response to stresses indicate expression and sequence divergences. *BMC Plant Biology*, **20**, 1–16.

Liao, B.Y. & Zhang, J. (2006) Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Molecular Biology and Evolution*, **23**, 1119–1128.

Lim, C.W., Baek, W., Han, S.W. & Lee, S.C. (2013) Arabidopsis PYl8 plays an important role for ABA signaling and drought stress responses. *Plant Pathology Journal*, **29**, 471–476.

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I.A.P. *et al.* (2014) The brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications*, **5**, 1–11.

Liu, X.L., Jiang, F.F., Wang, Z.W., Li, X.Y., Li, Z., Zhang, X.J. *et al.* (2017a) Wider geographic distribution and higher diversity of hexaploids than tetraploids in Carassius species complex reveal recurrent polyploidy effects on adaptive evolution. *Scientific Reports*, **7**, 1–10.

Liu, Y., Wang, J., Ge, W., Wang, Z., Li, Y., Yang, N. *et al.* (2017b) Two highly similar poplar paleo-subgenomes suggest an autotetraploid ancestor of salicaceae plants. *Frontiers in Plant Science*, **8**, 1–11.

Lovell, J.T., MacQueen, A.H., Mamidi, S., Bonnette, J., Jenkins, J., Napier, J.D. *et al.* (2021) Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature*, **590**, 438–444.

Lu, K., Wei, L., Li, X., Wang, Y., Wu, J., Liu, M. *et al.* (2019) Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. *Nature Communications*, **10**, 1–12.

Lynch, M. & Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Ma, Y.Q., Pu, Z.Q., Tan, X.M., Meng, Q., Zhang, K.L., Yang, L. *et al.* (2022) SEPALLATA-like genes of Isatis indigotica can affect the architecture of the inflorescences and the development of the floral organs. *PeerJ*, **10**, e13034.

Marais, D.L.D. & Rausher, M.D. (2008) Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, **454**, 762–765.

Maggiolini, F.A.M., Sanders, A.D., Shew, C.J., Sulovari, A., Mao, Y., Puig, M. *et al.* (2020) Single-cell strand sequencing of a macaque genome reveals multiple nested inversions and breakpoint reuse during primate evolution. *Genome Research*, **30**, 1680–1693.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, **17**, 10–12.

McCarthy, D.J., Chen, Y. & Smyth, G.K. (2012) Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Research*, **40**, 4288–4297.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**(9), 1297–1303.

Miele, V., Penel, S., Daubin, V., Picard, F., Kahn, D. & Duret, L. (2012) High-quality sequence clustering guided by network topology and multiple alignment likelihood. *Bioinformatics*, **28**, 1078–1085.

Miele, V., Penel, S. & Duret, L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.

Monroe, J.G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M. *et al.* (2022) Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature*, **602**, 101–105.

Murat, F., Armero, A., Pont, C., Klopp, C. & Salse, J. (2017) Reconstructing the genome of the most recent common ancestor of flowering plants. *Nature Genetics*, **49**(4), 490–496.

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H. *et al.* (2015) Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biology*, **16**, 262.

Murat, F., Zhang, R., Guizard, S., Flores, R., Armero, A., Pont, C. *et al.* (2014) Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16K protogenes. *Genome Biology and Evolution*, **6**, 12–33.

Murphy, W.J., Larkin, D.M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L. *et al.* (2005) Evolution: dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, **309**, 613–617.

Naser-Khdour, S., Quang Minh, B., Zhang, W., Stone, E.A. & Lanfear, R. (2019) The prevalence and impact of model violations in phylogenetic analysis. *Genome Biology and Evolution*, **11**, 3341–3352.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, **32**(1), 268–274.

Ohno, S. (1970) Introduction. In: Ohno, S. (Ed.) *Evolution by gene duplication*. Berlin, Heidelberg: Springer, pp. 1–2. Available from: https://doi.org/10.1007/978-3-642-86659-3_1

Page, J.T., Liechty, Z.S., Alexander, R.H., Clemons, K., Hulse-Kemp, A.M., Ashrafi, H. *et al.* (2016) DNA sequence evolution and rare homoeologous conversion in tetraploid cotton. *PLoS Genetics*, **12**, 1–22.

Parkin, I.A.P., Koh, C., Tang, H., Robinson, S.J., Kagale, S., Clarke, W.E. *et al.* (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biology*, **15**, 1–18.

Parks, M.B., Nakov, T., Ruck, E.C., Wickett, N.J. & Alverson, A.J. (2018) Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *American Journal of Botany*, **105**, 330–347.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.

Pham, S.K. & Pevzner, P.A. (2010) DRIMM-synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics*, **26**, 2509–2516.

Pons, P. & Latapy, M. (2006) Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, **10**, 191–218.

Pont, C., Leroy, T., Seidel, M., Tondelli, A., Duchemin, W., Armisen, D. *et al.* (2019a) Tracing the ancestry of modern bread wheats. *Nature Genetics*, **51**, 905–911.

Pont, C., Murat, F., Guizard, S., Flores, R., Foucrier, S., Bidet, Y. *et al.* (2013) Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. *The Plant Journal*, **76**, 1030–1044.

Pont, C., Wagner, S., Kremer, A., Orlando, L., Plomion, C. & Salse, J. (2019b) Paleogenomics: reconstruction of plant evolutionary trajectories from modern and ancient DNA. *Genome Biology*, **20**, 1–17.

Qi, X., An, H., Hall, T.E., Di, C., Blischak, P.D., McKibben, M.T.W. *et al.* (2021) Genes derived from ancient polyploidy have higher genetic diversity and are associated with domestication in Brassica rapa. *The New Phytologist*, **230**(1), 372–386.

Quinlan, A.R. & Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842.

Ramírez-González, R.H., Borrill, P., Lang, D., Harrington, S.A., Brinton, J., Venturini, L. *et al.* (2018) The transcriptional landscape of polyploid wheat. *Science*, **361**, 1–12.

Renny-Byfield, S., Gallagher, J.P., Grover, C.E., Szadkowski, E., Page, J.T., Udall, J.A. *et al.* (2014) Ancient gene duplicates in Gossypium (cotton) exhibit near-complete expression divergence. *Genome Biology and Evolution*, **6**, 559–571.

Renny-Byfield, S., Gong, L., Gallagher, J.P. & Wendel, J.F. (2015) Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. *Molecular Biology and Evolution*, **32**, 1063–1071.

Renny-Byfield, S., Rodgers-Melnick, E. & Ross-Ibarra, J. (2017) Gene fractionation and function in the ancient subgenomes of maize. *Molecular Biology and Evolution*, **34**, 1825–1832.

Renny-Byfield, S. & Wendel, J.F. (2014) Doubling down on genomes: polyploidy and crop plants. *American Journal of Botany*, **101**, 1711–1725.

Robinson, M.D., Mccarthy, D.J. & Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.A. (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology*, **13**(11), e1005752.

Salman-Minkov, A., Sabath, N. & Mayrose, I. (2016) Whole-genome duplication as a key factor in crop domestication. *Nature Plants*, **2**, 1–4.

Salse, J. (2016) Deciphering the evolutionary interplay between subgenomes following polyploidy: a paleogenomics approach in grasses. *American Journal of Botany*, **103**, 1167–1174.

Salse, J., Abrouk, M., Murat, F., Quraishi, U.M. & Feuillet, C. (2009) Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Briefings in Bioinformatics*, **10**(6), 619–630.

Sarda, S., Zeng, J., Hunt, B.G. & Yi, S.V. (2012) The evolution of invertebrate gene body methylation. *Molecular Biology and Evolution*, **29**, 1907–1916.

Schmid, M., Evans, B.J. & Bogart, J.P. (2015) Polyploidy in amphibia. *Cytogenet. Genome Research*, **145**, 315–330.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Schnable, J.C., Freeling, M. & Lyons, E. (2012) Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biology and Evolution*, **4**, 265–277.

Schnable, J.C., Springer, N.M. & Freeling, M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 4069–4074.

Scornavacca, C., Jacox, E. & Szöllosi, G.J. (2015) Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics*, **31**, 841–848.

Shi, J., Ma, X., Zhang, J., Zhou, Y., Liu, M., Huang, L. *et al.* (2019) Chromosome conformation capture resolved near complete genome assembly of broomcorn millet. *Nature Communications*, **10**, 1–9.

Shi, T., Rahmani, R.S., Gugger, P.F., Wang, M., Li, H., Zhang, Y. *et al.* (2020) Distinct expression and methylation patterns for genes with different fates following a single whole-genome duplication in flowering plants. *Molecular Biology and Evolution*, **37**, 2394–2413.

Slane, D., Reichardt, I., El Kasmi, F., Bayer, M. & Jürgens, G. (2017) Evolutionarily diverse SYP1 Qa-SNAREs jointly sustain pollen tube growth in Arabidopsis. *The Plant Journal*, **92**, 375–385.

Song, K., Li, L. & Zhang, G. (2018) Relationship among intron length, gene expression, and nucleotide diversity in the Pacific oyster Crassostrea gigas. *Marine Biotechnology*, **20**, 676–684.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stein, A., Coriton, O., Rousseau-Gueutin, M., Samans, B., Schiessl, S.V., Obermeier, C. *et al.* (2017) Mapping of homoeologous chromosome exchanges influencing quantitative trait variation in Brassica napus. *Plant Biotechnology Journal*, **15**, 1478–1489.

Stupar, R.M., Bhaskar, P.B., Yandell, B.S., Rensink, W.A., Hart, A.L., Ouyang, S. *et al.* (2007) Phenotypic and transcriptomic changes associated with potato autopolyploidization. *Genetics*, **176**, 2055–2067.

Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y. *et al.* (2017) Karyotype stability and unbiased fractionation in the paleo-allotetraploid Cucurbita genomes. *Molecular Plant*, **10**, 1293–1306.

Sun, Y., Zhong, M., Li, Y., Zhang, R., Su, L., Xia, G. *et al.* (2021) GhADF6-Mediated Actin reorganization is associated with defence against verticillium dahliae infection in cotton. *Molecular Plant Pathology*, **22**, 1656–1667.

Suyama, M., Torrents, D., Bork, P. & Delbru, M. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, **34**, 609–612.

Suzuki, M.M. & Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews. Genetics*, **9**, 465–476.

Szöllosi, G.J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. (2013) Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, **62**, 901–912.

Takuno, S. & Gaut, B.S. (2012) Body-methylated genes in arabidopsis thaliana are functionally important and evolve slowly. *Molecular Biology and Evolution*, **29**, 219–227.

Thomas, B.C., Pedersen, B. & Freeling, M. (2006) Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research*, **16**(7), 934–946.

Throude, M., Bolot, S., Bosio, M., Pont, C., Sarda, X., Quraishi, U.M. *et al.* (2009) Structure and expression analysis of rice paleo duplications. *Nucleic Acids Research*, **37**, 1248–1259.

Ulrich, D. & Olbricht, K. (2013) Diversity of volatile patterns in sixteen Fragaria vesca L. accessions in comparison to cultivars of Fragaria xananassa. *Journal of Applied Botany and Food Quality*, **86**, 37–46.

Van de Peer, Y., Mizrachi, E. & Marchal, K. (2017) The evolutionary significance of polyploidy. *Nature Reviews. Genetics*, **18**, 411–424.

VanBuren, R., Man Wai, C., Wang, X., Pardo, J., Yocca, A.E., Wang, H. *et al.* (2020) Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. *Nature Communications*, **11**, 1–11.

Vanneste, K., Baele, G., Maere, S. & van de Peer, Y. (2014) Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the cretaceous – Paleogene boundary. *Genome Research*, **32**, 1334–1347.

Wang, Z., Li, Y., Sun, P., Zhu, M., Wang, D., Lu, Z. *et al.* (2022) A high-quality *Buxus austro-yunnanensis* (Buxales) genome provides new insights into karyotype evolution in early eudicots. *BMC Biology*, **20**, 216.

Wang, M., Tu, L., Lin, M., Lin, Z., Wang, P., Yang, Q. *et al.* (2017a) Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nature Genetics*, **49**, 579–587.

Wang, X., Zhang, Z., Fu, T., Hu, L., Xu, C., Gong, L. *et al.* (2017b) Gene-body CG methylation and divergent expression of duplicate genes in rice. *Scientific Reports*, **7**, 1–11.

Wang, Y., Wang, X., Lee, T.H., Mansoor, S. & Paterson, A.H. (2013) Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship with gene expression in Oryza sativa (rice). *The New Phytologist*, **198**, 274–283.

Wang, Y., Wang, X. & Paterson, A.H. (2012) Genome and gene duplications and gene expression divergence: a view from plants. *Annals of the New York Academy of Sciences*, **1256**, 1–14.

Wickland, D.P. & Hanzawa, Y. (2015) The FLOWERING LOCUS T/TERMINAL FLOWER 1 gene family: functional evolution and molecular mechanisms. *Molecular Plant*, **8**, 983–997.

Woodhouse, M.R., Cheng, F., Pires, J.C., Lisch, D., Freeling, M. & Wang, X. (2014) Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 5283–5288.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S. *et al.* (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biology*, **8**, e1000409.

Xu, W., Zhang, Q., Yuan, W., Xu, F., Muhammad Aslam, M., Miao, R. *et al.* (2020) The genome evolution and low-phosphorus adaptation in white lupin. *Nature Communications*, **11**, 1–13.

Yang, L. & Gaut, B.S. (2011) Factors that contribute to variation in evolutionary rate among arabidopsis genes. *Molecular Biology and Evolution*, **28**, 2359–2369.

Yim, W.C., Lee, B.M. & Jang, C.S. (2009) Expression diversity and evolutionary dynamics of rice duplicate genes. *Molecular Genetics and Genomics*, **281**, 483–493.

Zhang, J., Liu, Y., Xia, E.H., Yao, Q.Y., Liu, X.D. & Gao, L.Z. (2015) Autotetraploid rice methylome analysis reveals methylation variation of transposable elements and their effects on gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, E7022–E7029.

Zhang, Q., Guan, P., Zhao, L., Ma, M., Xie, L., Li, Y. *et al.* (2021) Asymmetric epigenome maps of subgenomes reveal imbalanced transcription and distinct evolutionary trends in Brassica napus. *Molecular Plant*, **14**, 604–619.

Zhang, X., Shiu, S., Cal, A. & Borevitz, J.O. (2008) Global analysis of genetic, epigenetic and transcriptional polymorphisms in Arabidopsis thaliana using whole genome tiling arrays. *PLoS Genetics*, **4**, e1000032.

Zhao, C., Zayed, O., Zeng, F., Liu, C., Zhang, L., Zhu, P. *et al.* (2019) Arabinose biosynthesis is critical for salt stress tolerance in Arabidopsis. *The New Phytologist*, **224**, 274–290.

Zhao, M., Zhang, B., Lisch, D. & Ma, J. (2017) Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *Plant Cell*, **29**, 2974–2994.

Zhou, X., Liao, Y., Kim, S.U., Chen, Z., Nie, G., Cheng, S. *et al.* (2020) Genome-wide identification and characterization of bHLH family genes from Ginkgo biloba. *Scientific Reports*, **10**, 1–15.

Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R. & Shiu, S.H. (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiology*, **151**, 3–15.

Annexe E

Bellec et al. (2023), Données supplémentaires

Les Jeux de Données supplémentaires sont disponibles au lien suivant :

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.16185&file=tpj16185-sup-0002-SupplementaryDataSets.xlsx

Les Tables supplémentaires sont accessibles au lien suivant :

https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Ftpj.16185&file=tpj16185-sup-0003-Tables.xlsx

SUPPLEMENTARY DATA FILE

SUPPLEMENTARY DATA SETS

Supplementary Data Set 1: Comparative genomics data between 10 grass species
Supplementary Data Set 2: Omics data available for wheat, maize and *Brachypodium*

SUPPLEMENTARY TABLES

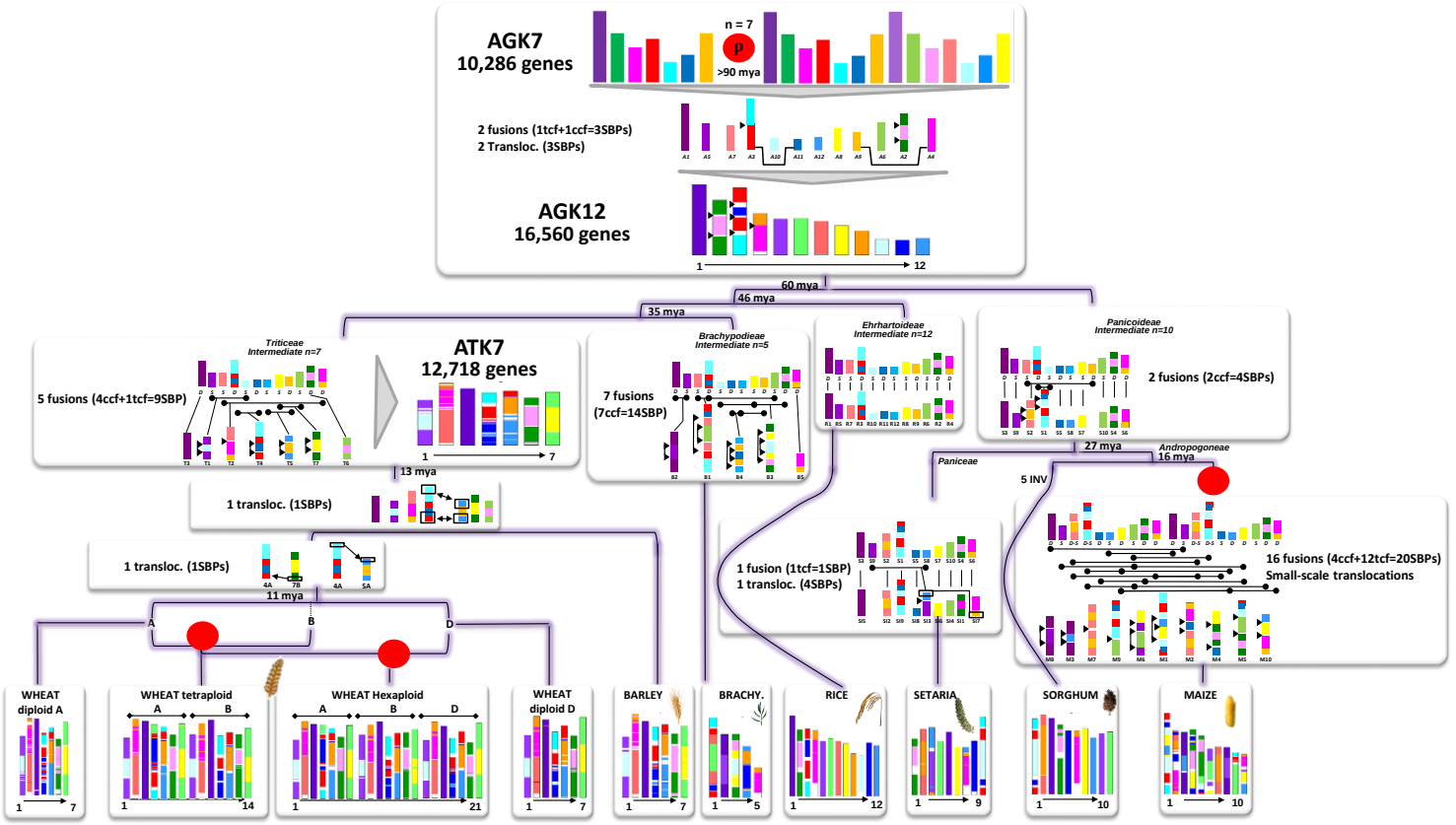
Supplementary Table 1: Catalog of SBPs in cereal genomes compared to AGKs.
Supplementary Table 2: Catalog of inversions (INVS) in cereal genomes compared to AGKs.
Supplementary Table 3: Partitioning of genomes according to recombination rates.
Supplementary Table 4: LF/MF and Singleton/Pair contents in cereal genomes.
Supplementary Table 5: Branch lengths values for cereals and Triticeae phylogenetic trees.
Supplementary Table 6: Substitution rate and branch length in cereals for LF/MF and Singleton/Pair genes.
Supplementary Table 7: Global DNA methylation in *Brachypodium* and maize.
Supplementary Table 8: Comparison of omic features between wheat subgenomes (A, B, D and LF, MF).
Supplementary Table 9: Expression and methylation profiles of conserved genes.
Supplementary Table 10: Maize duplicated genes expression and methylation regulation for key trait-driving genes.

SUPPLEMENTARY FIGURES

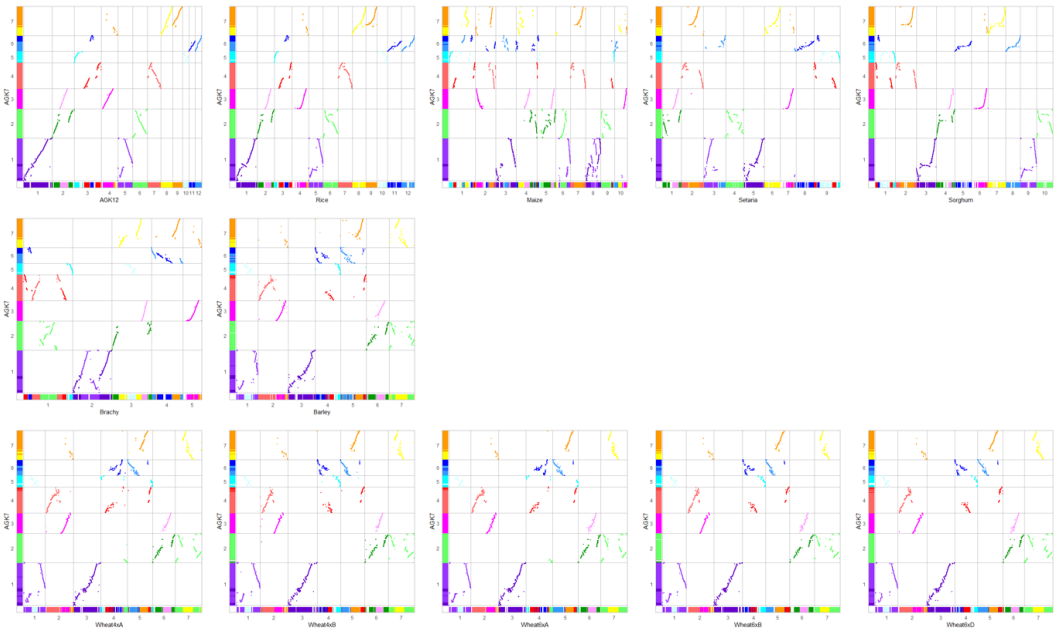
Supplementary Figure 1: Dotplot based deconvolution of modern genomes according to inferred AGKs
Supplementary Figure 2: Gene Ontology enrichment
Supplementary Figure 3: SBPs between extant cereal genomes and AGK12
Supplementary Figure 4: Inversions in grass genomes
Supplementary Figure 5: Inversions between hexaploid wheat and tetraploid wheat
Supplementary Figure 6: Phylogeny and datation for substitution rate analysis
Supplementary Figure 7: LF/MF compartments.
Supplementary Figure 8: Analysis of omics data between collinear versus inversed genes, singletons versus genes in pairs and LF- versus MF-genes for Wheat, Maize and *Brachypodium*
Supplementary Figure 9: Subgenome dominance in hexaploid wheat between A-B-D subgenomes
Supplementary Figure 10: Subgenome dominance in hexaploid wheat between recent A-B-D subgenomes for each ancient LF-MF subgenomes
Supplementary Figure 11: Impact of TEs in methylation and expression differences observed between conserved genes
Supplementary Figure 12: Plasticity of duplicated genes (Ka, ks)
Supplementary Figure 13: Whole genome omics interplay
Supplementary Figure 14: Expression and methylation levels of conserved and non-conserved genes
Supplementary Figure 15: Expression and methylation outliers from mixOmics

<u>Supplementary Figure 1</u>: Dotplot based deconvolution of modern genomes according to inferred AGKs (Ancestral Grass Karyotypes). a. Evolutionary scenario from inferred ancestors (AGK and ATK) leading to the extant grass genomes. Evolutionary events (fusions, fissions, translations) defining synteny breakpoints (SBPs) are illustrated and detailed on the tree branches. From AGK12 (structure of the modern rice genome), two ancestral chromosomes fusions (reported as ccf for centromeric-based chromosome fusions and tcf for telomeric-based chromosome fusions) contributing to 4 SBPs explain the modern sorghum genome and setaria experienced specifically an extra fusion and one complex translocation event between chromosomes 3 and 7, leading to a total of 4 and 9 SBPS in respectively setaria and sorghum. Maize experienced 16 fusions and small-scale translocations leading to 38 SBPs. *Brachypodium* experienced 14 SBPs (from 7 fusions), one common with the *Triticeae* (corresponding to a fusion between chromosomes 9 and 12 in AGK12). The modern *Triticeae* derive from a ATK7 inherited from AGK12 following 5 ancestral chromosome fusions and 2 translocations leading to 10 SBPs (and 11 for wheat subgenome A) **b**. Dotplot based deconvolution of the synteny between AGK7 (y-axis, with light and dark shades of seven colors illustrating the transition between AGK7 to AGK12) and the modern genomes (x-axis). **c**. Dotplot based deconvolution of the synteny between AGK7 (y-axis) and the modern genomes (x-axis) with blue dots for singleton genes and red dot for duplicated genes.

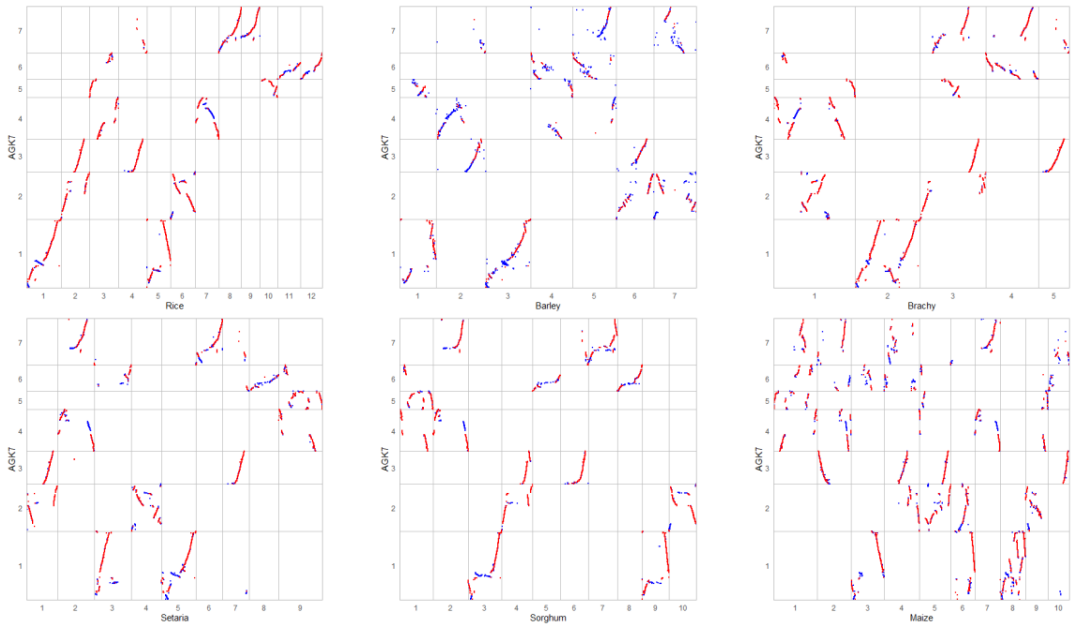
a.

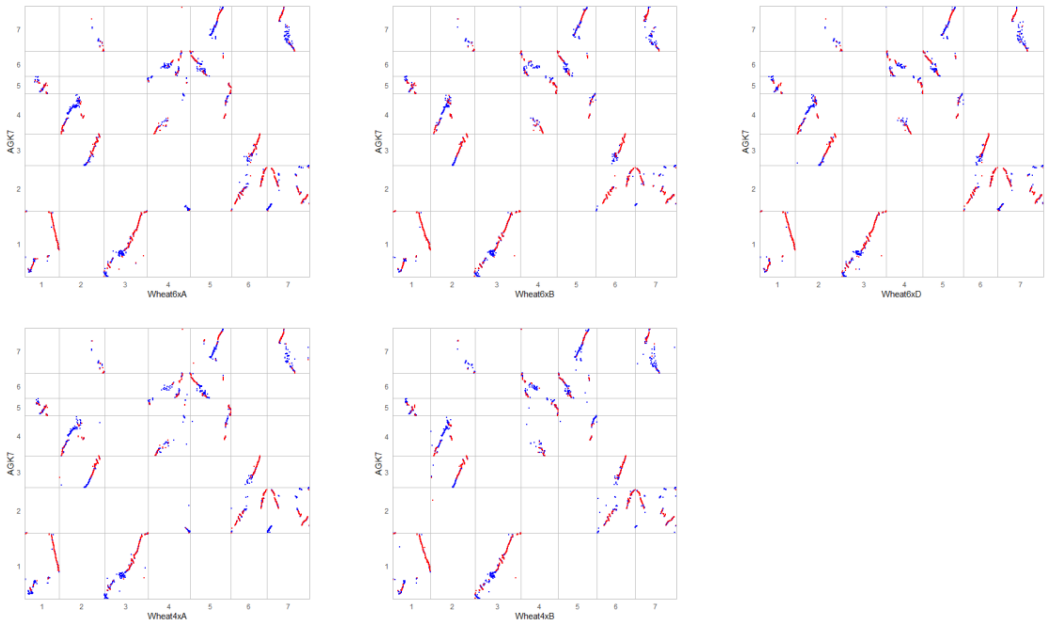


b.


AGK7
AGK12
Rice
Maize
Setaria
Sorghum
Brachy
Barley
Wheat4xA
Wheat4xB
Wheat6xA
Wheat6xB
Wheat6xD

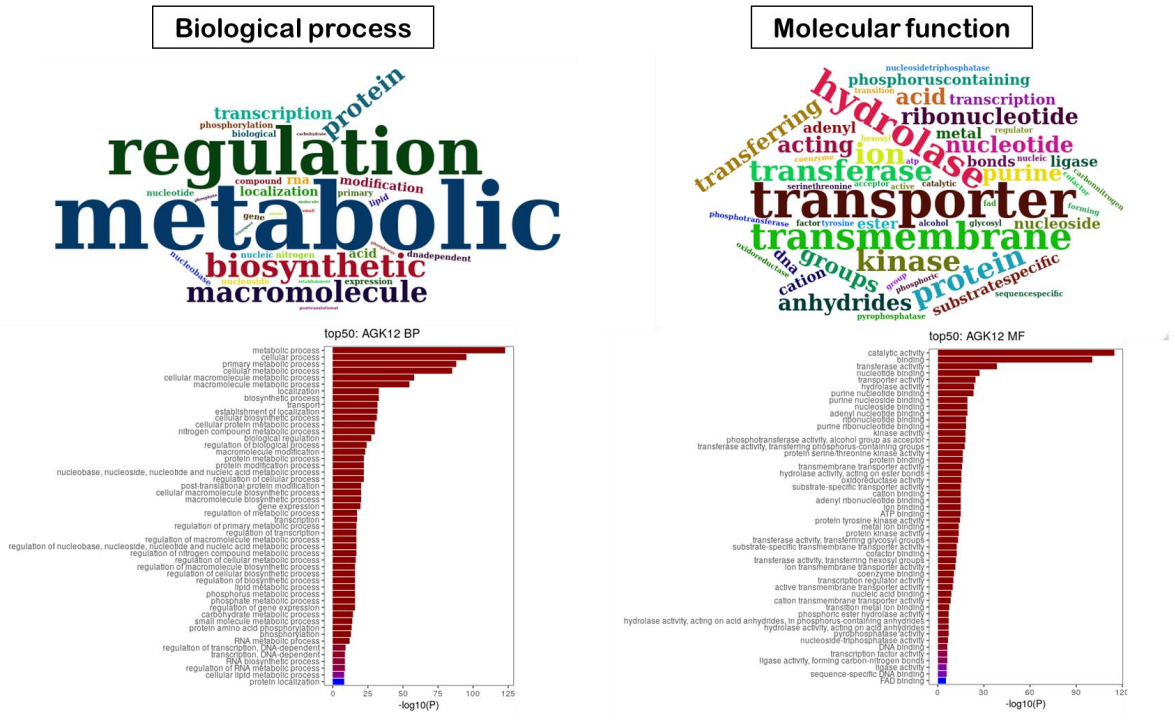
c.


AGK7
Rice
Barley
Brachy
Setaria
Sorghum
Maize

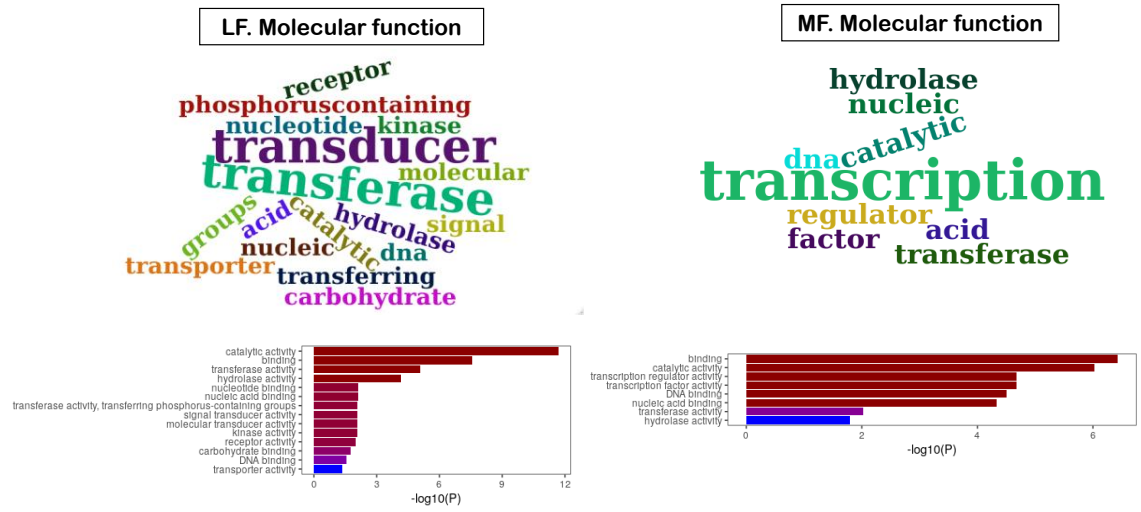

AGK7
Wheat6xA
AGK7
Wheat6xB
AGK7
Wheat6xD
AGK7
Wheat4xA
AGK7
Wheat4xB

<u>Supplementary Figure 2</u>: Gene Ontology enrichment. Biological processes (left) and/or molecular function (right) annotations highlighting GO enrichment presented by tag clouds and top 50 enriched GO terms for (**a**) Ancestral conserved genes in AGK12 (using rice GO as reference), (**b**) LF- and MF-genes (using rice GO as reference), (**c**) Singletons and duplicated genes pairs (using rice GO as reference), (**d**) Hypomethylated/Upregulated genes (in *Brachypodium*), (**e**) Hypermethylated/Downregulated genes (in *Brachypodium*).

a.

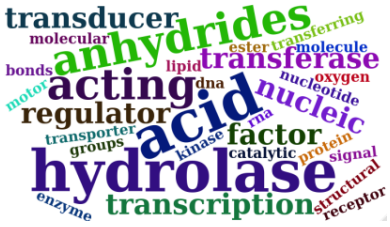


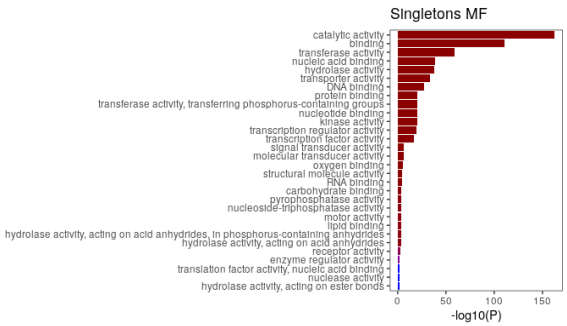
b.



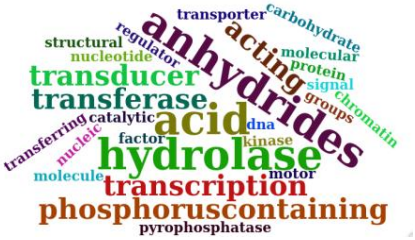
c.

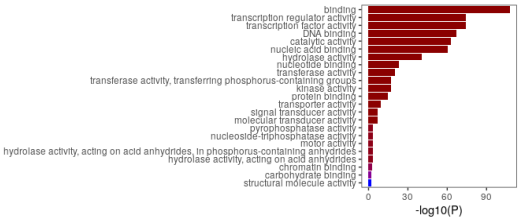
Singletons. Molecular function





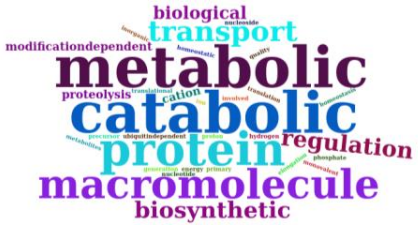
Pairs. Molecular function

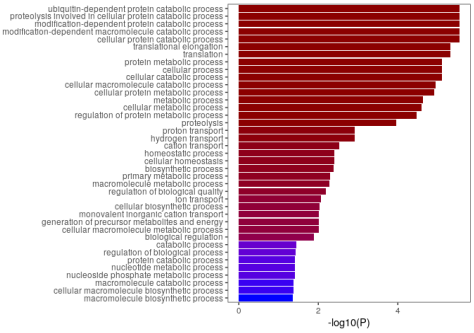




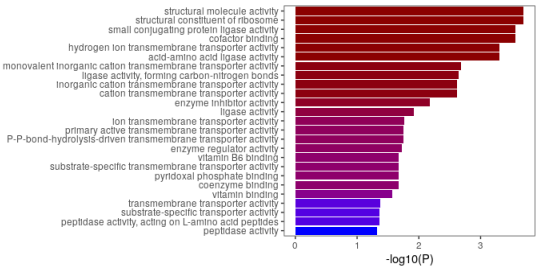
d.

Biological process



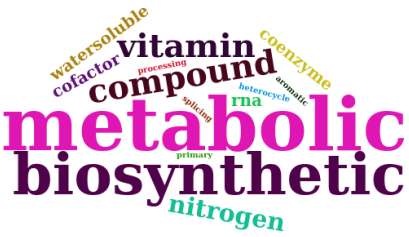


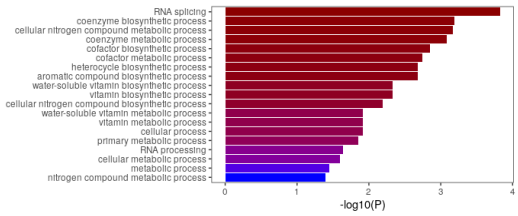
Molecular function



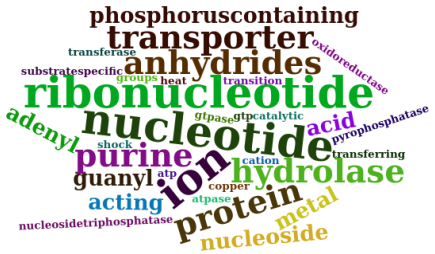
e.

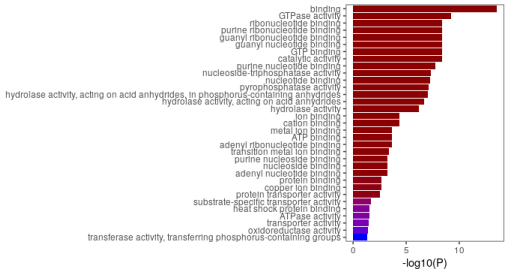
Biological process



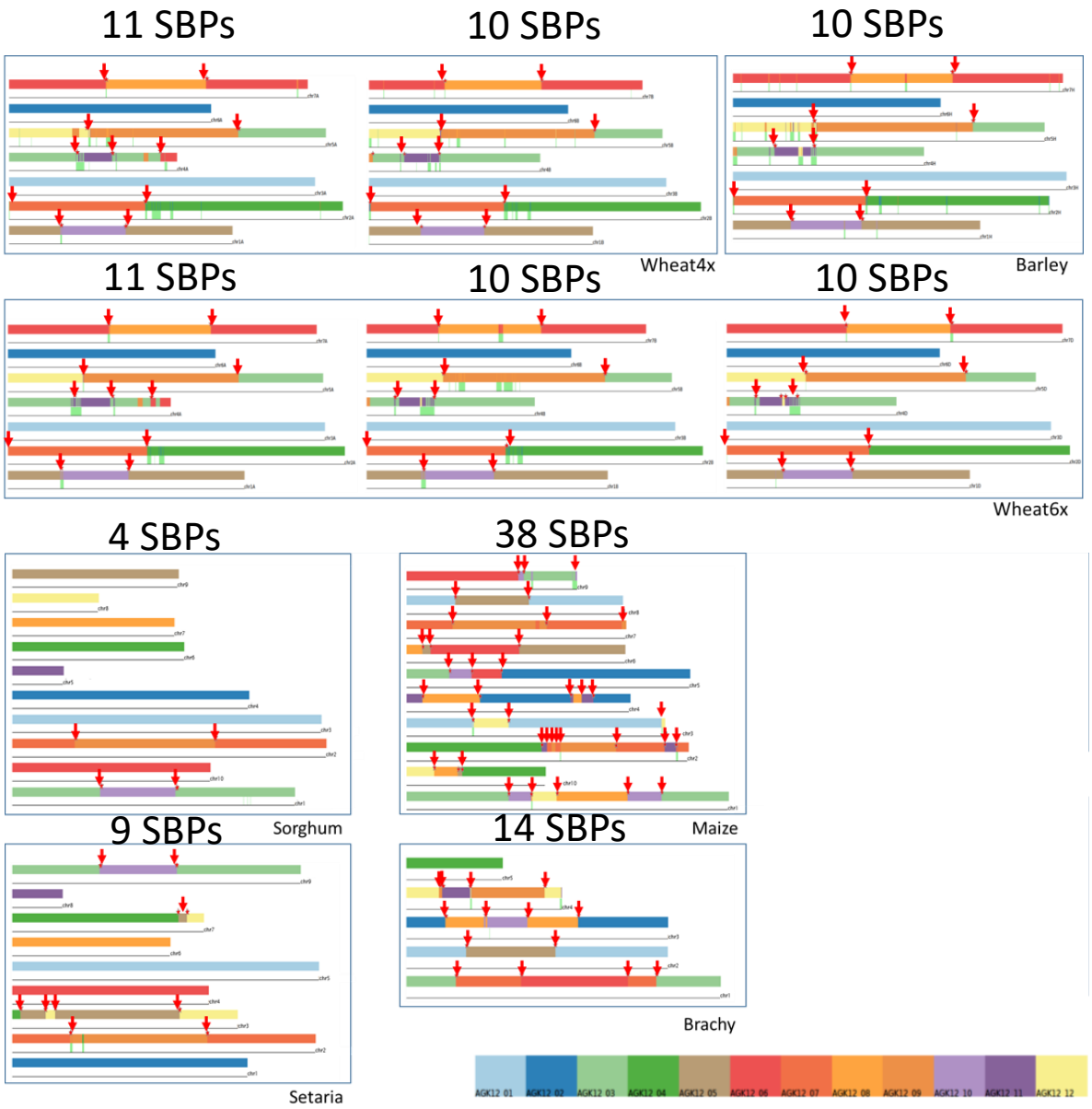


Molecular function



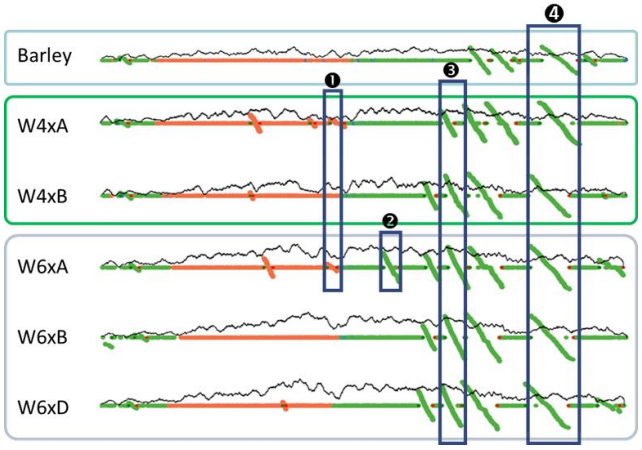


<u>Supplementary Figure 3</u>: Synteny break points (SBPs) between extant cereal genomes and AGK12. SBPs are illustrated by red arrows and extant chromosomes are colored according to the AGK12 color code (bottom).



Supplementary Figure 4: Inversions in grass genomes. a. Graphical representation of inversions within *Triticeae* chromosomal group 2 originating from AGK12 protochromosomes 4 (in green) and 6 (in red). Baseline represents extant gene orders when collinear to AGK12 protogenes order. Diagonals represents inversions defined as inverted extant gene orders in regards to AGK12 protogene order, with #1: inversions specific to subgenome A in wheat, #2: hexaploid wheat (6x) subgenome A specific inversions, #3: wheat specific inversions, #4. *Triticeae* specific inversions. **b**. Table illustrating the cumulative gene space covered by the inversions (in columns) detected in the investigated species (in lines) **c**. Upset plot of shared inversions among cereals. **d**. Gene density in inversions (red line) in regards to gene density resulting from 1000 simulations, by positioning the inversions randomly on the chromosomes. In all species, the observed gene density in the characterized inversions was higher than the values observed in simulations, ranging from an observed gene density higher than 87.3% of simulated values in *Brachypodium* to more than 97.9% for all other species of the panel. **e**. Position of inversions according to recombination rate levels within chromosomes.

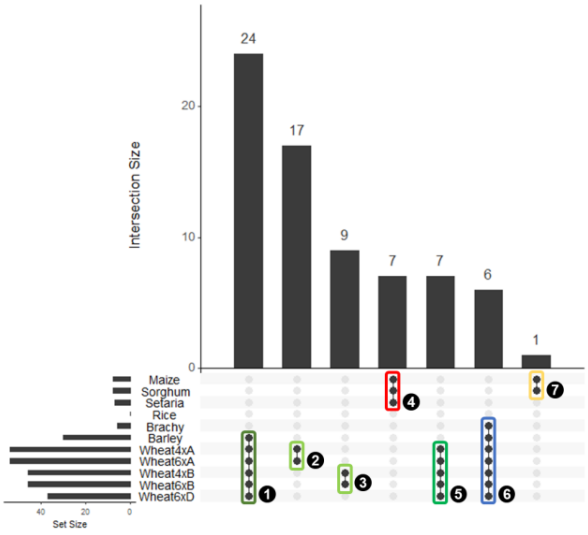
a.



b.

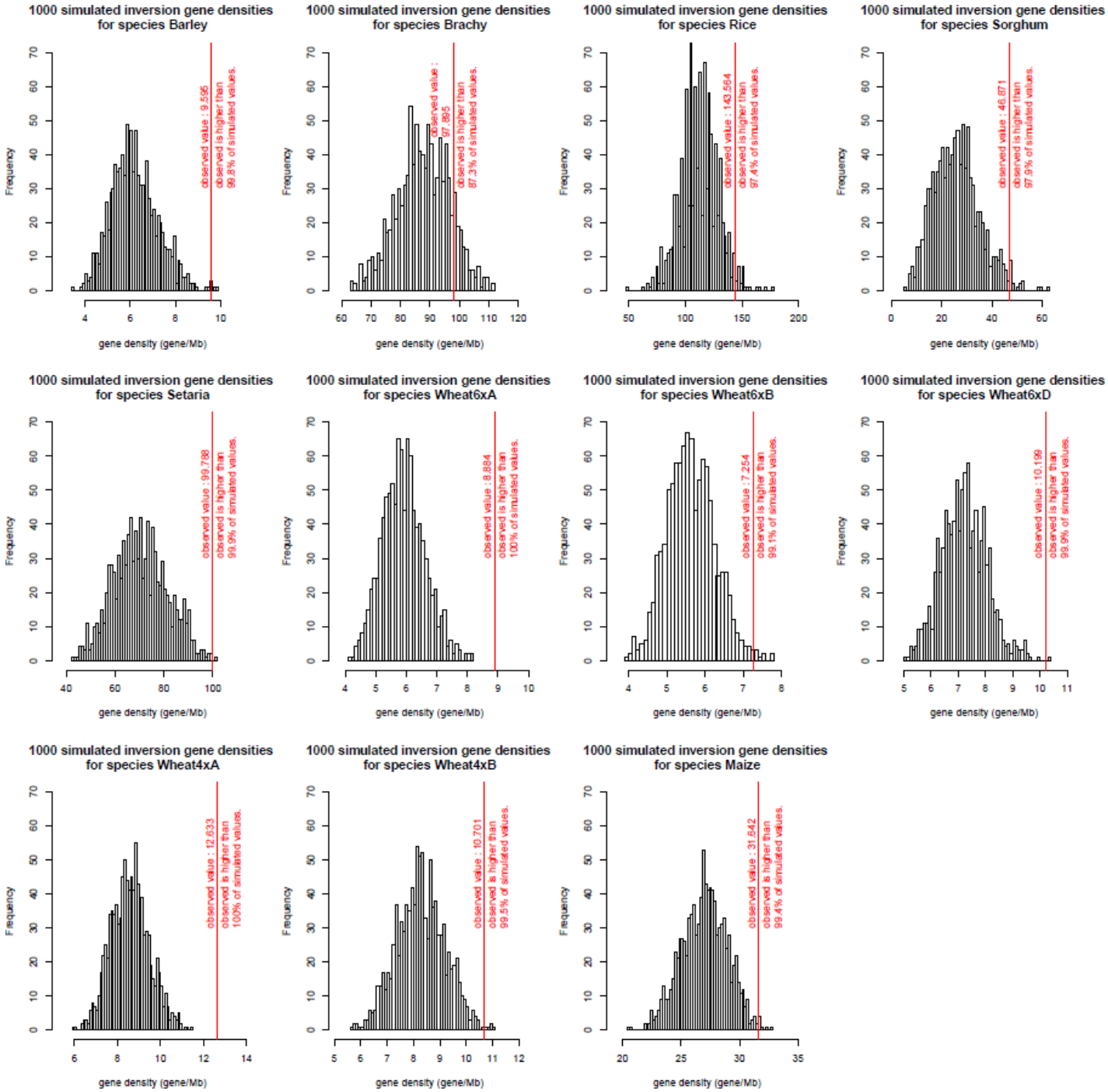
	Genome size (bp)	Cumulative inversions length (bp)	Fraction of inversions in the genome
Barley	4583636181	797115940	17,4%
Brachy	270993320	29071917	10,7%
Rice	372530218	766207	0,2%
Sorghum	659098135	82822880	12,6%
Setaria	401159754	62703147	15,6%
Wheat6xA	4934357510	1017109333	20,6%
Wheat6xB	5179939804	1145788091	22,1%
Wheat6xD	3950743238	768027870	19,4%
Wheat4xA	4899086626	1057220356	21,6%
Wheat4xB	5179230282	1182676239	22,8%
Maize	2059165538	290851024	14,1%

c.



❶: *Triticeae*-specific inversion
❷: Wheat subgenome A specific inversions
❸: Wheat subgenome B specific inversions
❹: *Panicoideae*-specific inversions
❺: Wheat specific inversions
❻: BWB specific inversions
❼: MS Specific inversions

d.

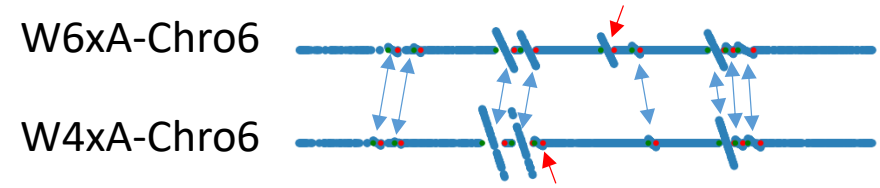

1000 simulated inversion gene densities for species Barley
observed value : 9.595
observed is higher than 99.8% of simulated values.
1000 simulated inversion gene densities for species Brachy
observed value : 97.895
observed is higher than 87.3% of simulated values.
1000 simulated inversion gene densities for species Rice
observed value : 143.564
observed is higher than 97.4% of simulated values.
1000 simulated inversion gene densities for species Sorghum
observed value : 46.871
observed is higher than 97.9% of simulated values.
1000 simulated inversion gene densities for species Setaria
observed value : 99.788
observed is higher than 99.9% of simulated values.
1000 simulated inversion gene densities for species Wheat6xA
observed value : 8.884
observed is higher than 100% of simulated values.
1000 simulated inversion gene densities for species Wheat6xB
observed value : 7.254
observed is higher than 99.1% of simulated values.
1000 simulated inversion gene densities for species Wheat6xD
observed value : 10.199
observed is higher than 99.9% of simulated values.
1000 simulated inversion gene densities for species Wheat4xA
observed value : 12.633
observed is higher than 100% of simulated values.
1000 simulated inversion gene densities for species Wheat4xB
observed value : 10.701
observed is higher than 99.5% of simulated values.
1000 simulated inversion gene densities for species Maize
observed value : 31.642
observed is higher than 99.4% of simulated values.
Frequency
gene density (gene/Mb)

e.

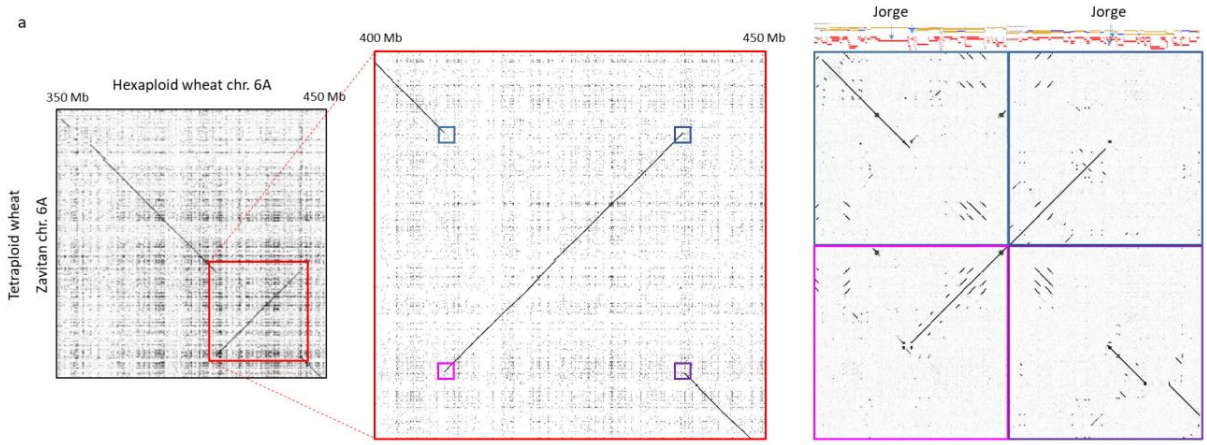
Specie	Genome Size	Cumulated LR size	Size of LR region / Total Genome Size	% of INV located in LR region
Brachy	270993320	38000000	14,0%	10,9%
Sorghum	659098135	60000000	9,1%	4,5%
Wheat6xB	5179939804	1090000000	21,0%	6,1%
Maize	2059165538	341000000	16,6%	10,5%
Barley	4583636181	1560000000	34,0%	9,8%
Setaria	401159754	70000000	17,4%	5,7%
Wheat6xD	3950743238	1040000000	26,3%	10,4%
Wheat4xA	4899086626	1780000000	36,3%	18,8%
Wheat4xB	5179230282	1180000000	22,8%	10,9%
Wheat6xA	4934357510	1720000000	34,9%	13,2%

<u>Supplementary Figure 5</u>: Inversions between hexaploid and tetraploid wheats. a. Detailed comparison of inversions on chromosome 6 of tetraploid and hexaploid wheat using the same representation as Supplementary Figure 4a. Blue arrows highlight shared inversions corresponding to common ancestral origins. Red arrows highlight species-specific inversions. **b**. Dotplot based alignment of Zavitan tetraploid wheat against Chinese spring hexaploid wheat of a genomic region of chromosome 6A (see coordinated on the figure). **c**. Dotplot based alignment Svevo tetraploid wheat against Chinese spring hexaploid wheat of a genomic region of chromosome 6A (see coordinated on the figure). In both cases both inversion boundaries correspond to same LTR retrotransposons (red rectangle).

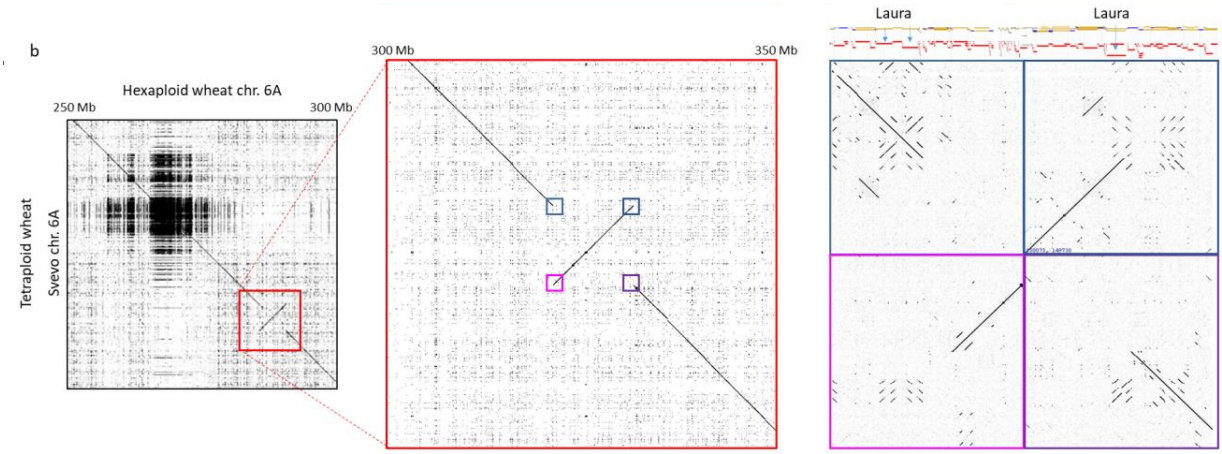
a.



b.



c.

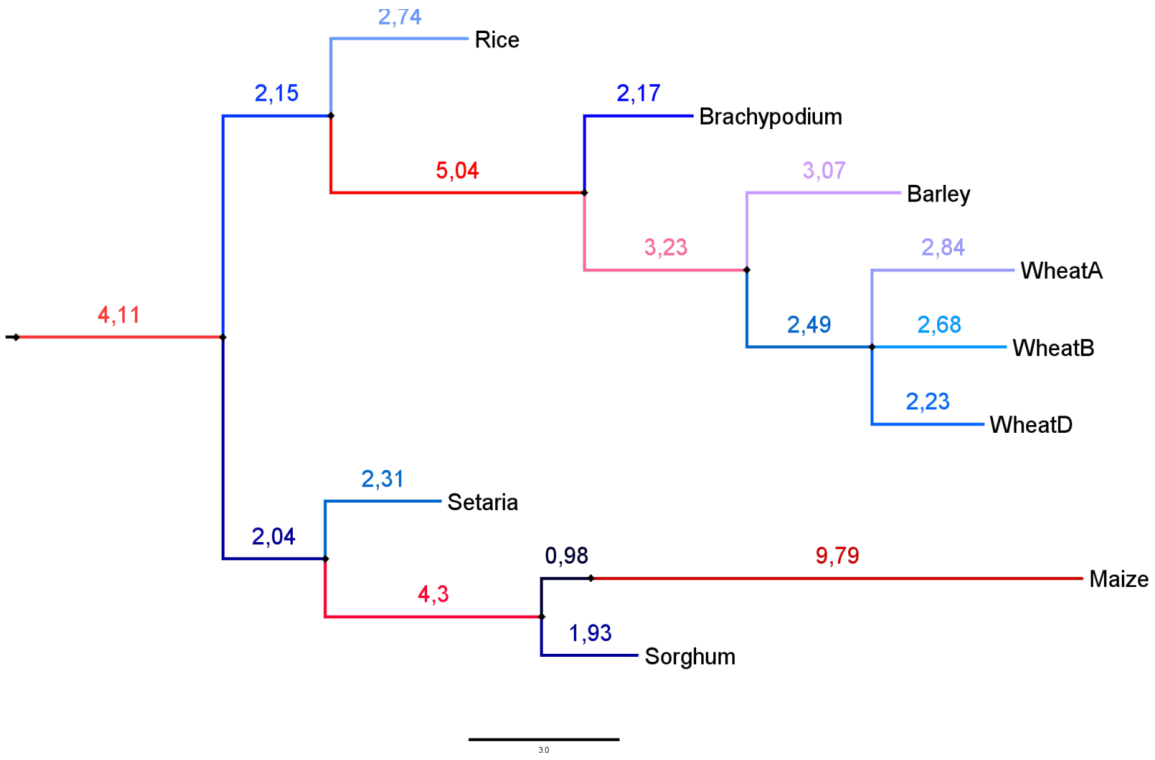


Supplementary Figure 6: Phylogeny and dating for substitution rate analysis. a. Divergence time between the different clades of cereals as estimated by Murat et al. 2017 and El Baidouri et al. 2017. **b**. Substitution rate corresponding to measured substitutions per site value normalized by branch duration in billion years among cereals. **c**. Substitution rates in *Triticeae*: the same calculation as above is applied.

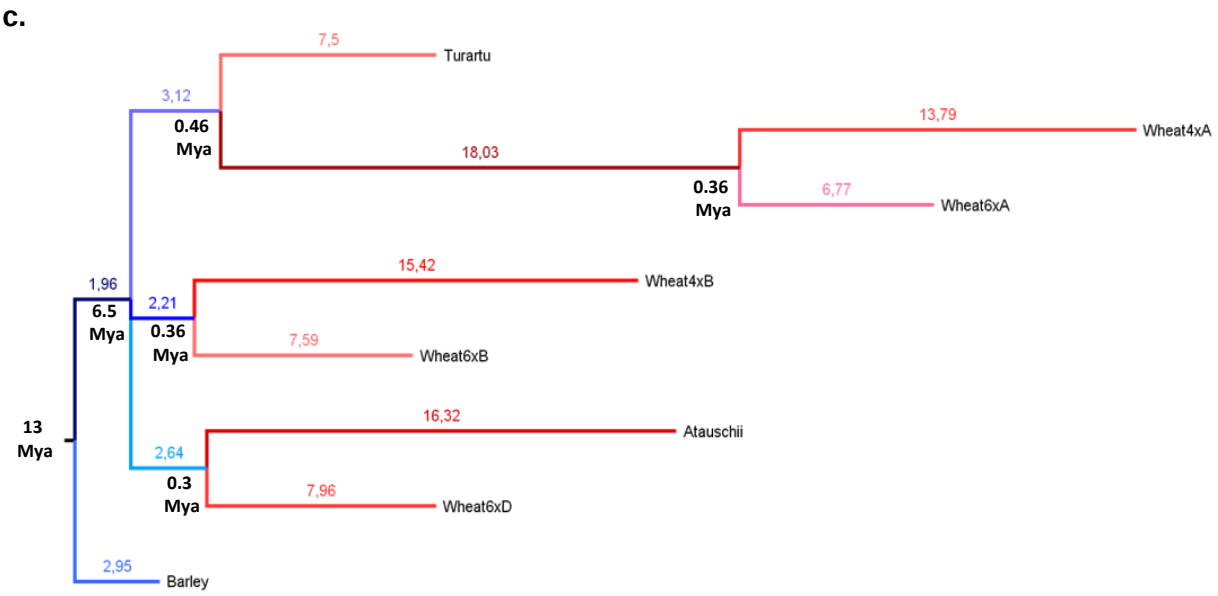
a.

Clade	Divergence time (Mya)
AGK7 WGD	90
AGK12 Speciation	60
(Rice, *Brachypodium*)	46
(Barley, *Brachypodium*)	35
(Barley, Wheat)	13
Wheats root	6.5
Tetraploid wheat	0.36
Hexaploid wheat	0.01
(Maize, Setaria)	27
(Maize, Sorghum)	16
(WheatA, *T.urartu*)	0.46
(WheatD, *A.speltoides*)	0.3

b.

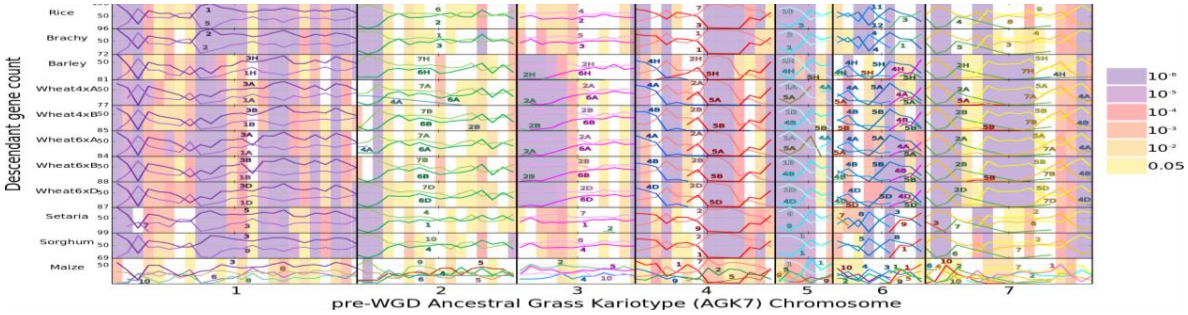


c.

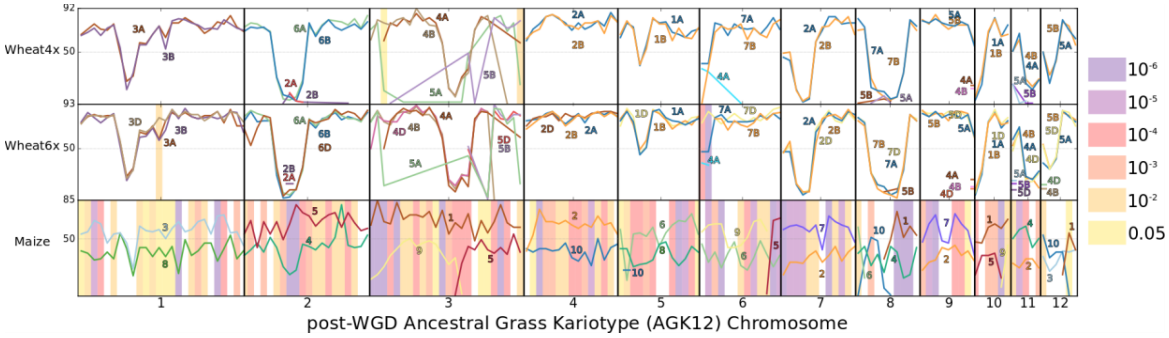

7,5
Turartu
3,12
0.46
Mya
13,79
Wheat4xA
18,03
0.36
Mya
6,77
Wheat6xA
15,42
Wheat4xB
1,96
6.5
Mya
2,21
0.36
Mya
7,59
Wheat6xB
16,32
Atauschii
13
Mya
2,64
0.3
Mya
7,96
Wheat6xD
2,95
Barley

<u>Supplementary Figure 7</u>: LF/MF compartments detection. a. LF/MF compartments inherited from the ancestral ρ WGD. For each ancestral chromosomes (AGK7, x-axis) curves show the counts of genes observed in the modern (post-ρ) duplicated regions in the extant species (y-axis) for each 100 ancestral genes windows. The background color indicates the p-value of biased fractionation statistical test. **b**. LF/MF compartments inherited from species-specific polyploidization events in tetraploid wheat, hexaploid wheat and maize from AGK12. Curves show the counts of genes observed in the post-duplication regions in the extant species for each 100 ancestral genes windows in the post-ρ ancestor (AGK12). The background color indicates the p-value of biased fractionation statistical test.

a.

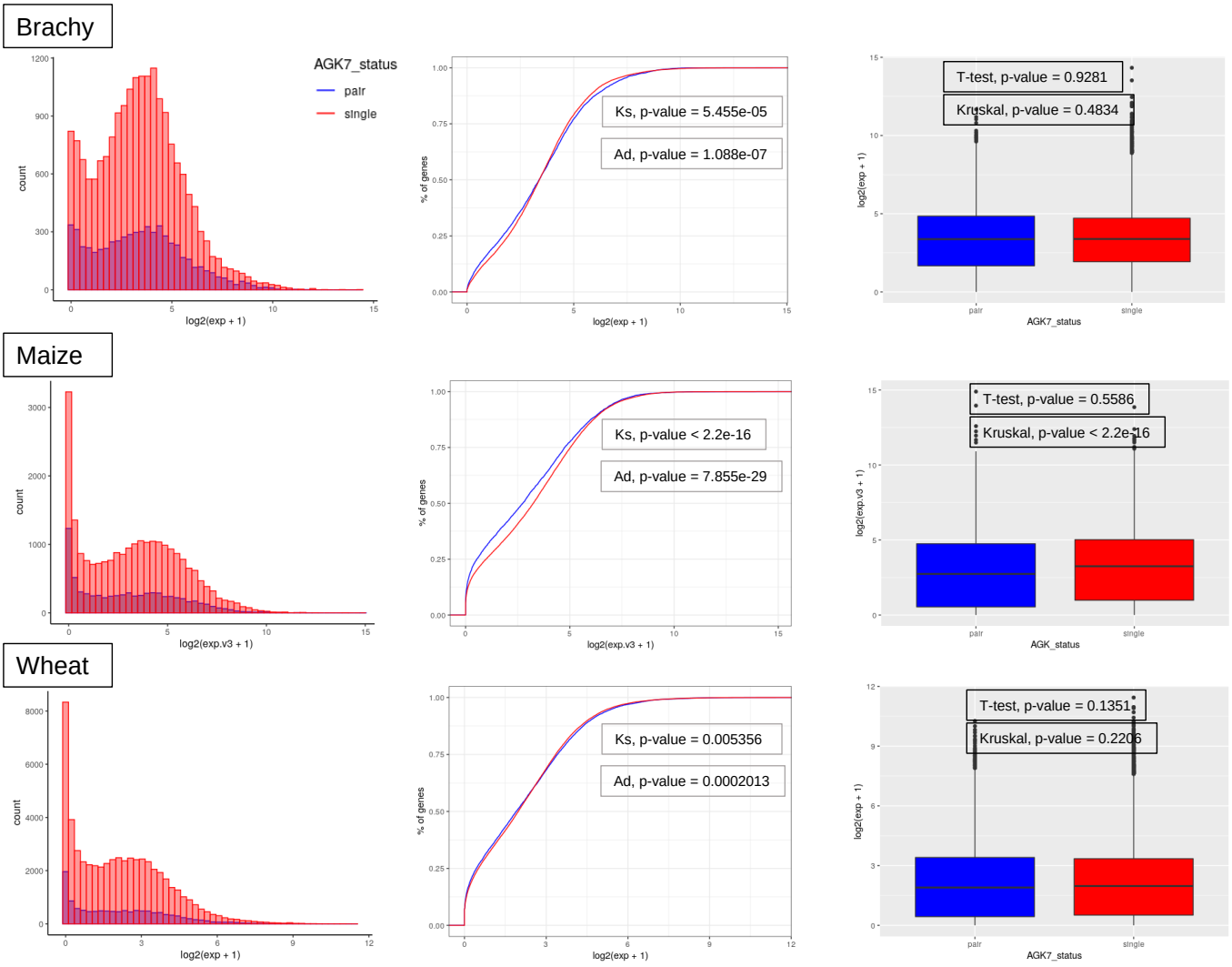


b.



<u>Supplementary Figure 8</u>: Analysis of omics data between collinear versus inversed genes, singletons versus genes in pairs and LF- versus MF-genes for Wheat, Maize and *Brachypodium*. **a**. Comparative analysis of gene expression between singletons and duplicated genes (pairs) in *Brachypodium*, maize and wheat illustrated as (left) histogram distribution plot, (center) empirical cumulative distribution and (right) boxplot. Significant differences are shown using four different statistical tests, *i.e*. Ks = Kolmogorov-Smirnov test, Ad= Anderson-Darling test, t-test= Student test, Kruskal = Kruskal-Wallis test. **b**. Data summary showing omics differences such as gene expression, DNA methylation (promoter and gene body), mutations dynamic (SNPs) and substitution rates (Ka, Ks), illustrated with arrows (↑ for increase and ↓ for decrease) between collinear and inverted, conserved and non-Conserved, singleton and pair, LF- and MF-genes. Col. = collinear; Inv. = Inverted; CS = Conserved; NCS = Non-Conserved; Sing = Singleton; pair = duplicated genes; LF = Least-Fractionated; MF = Most-Fractionated; SE = Slow evolving; FE = Fast evolving.

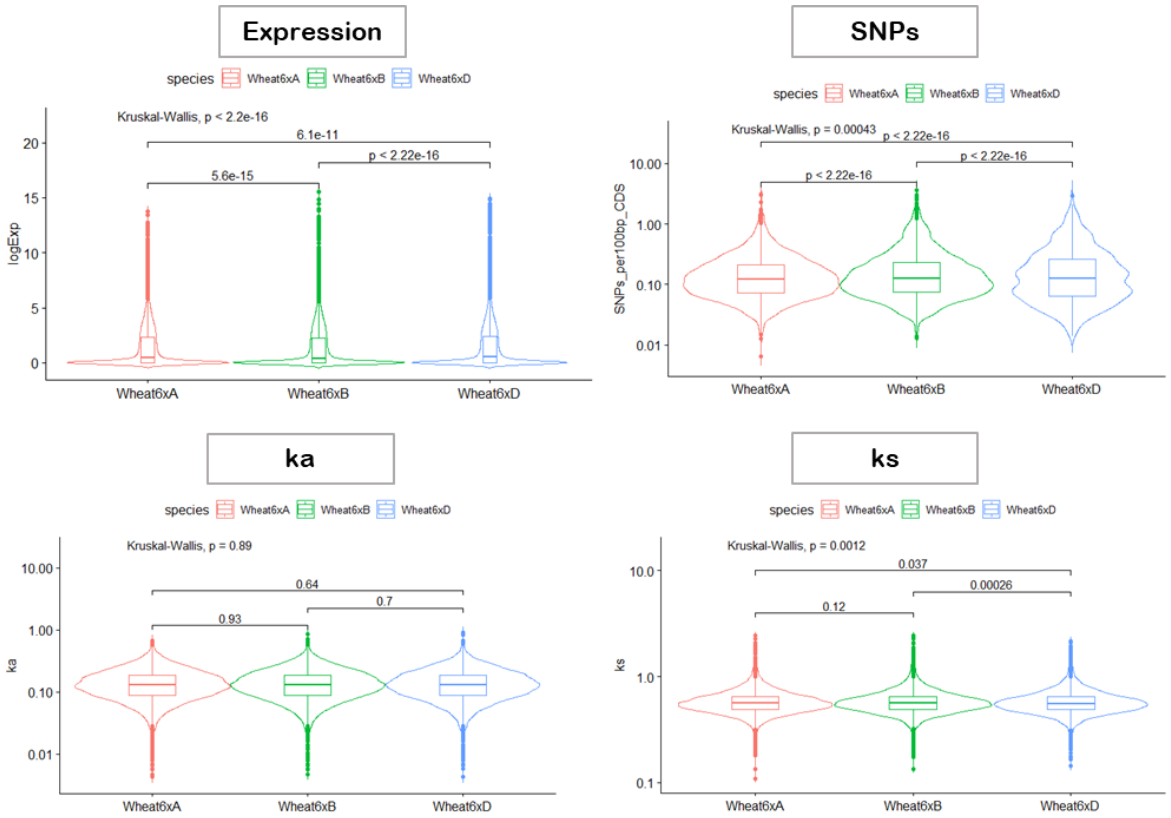
a.



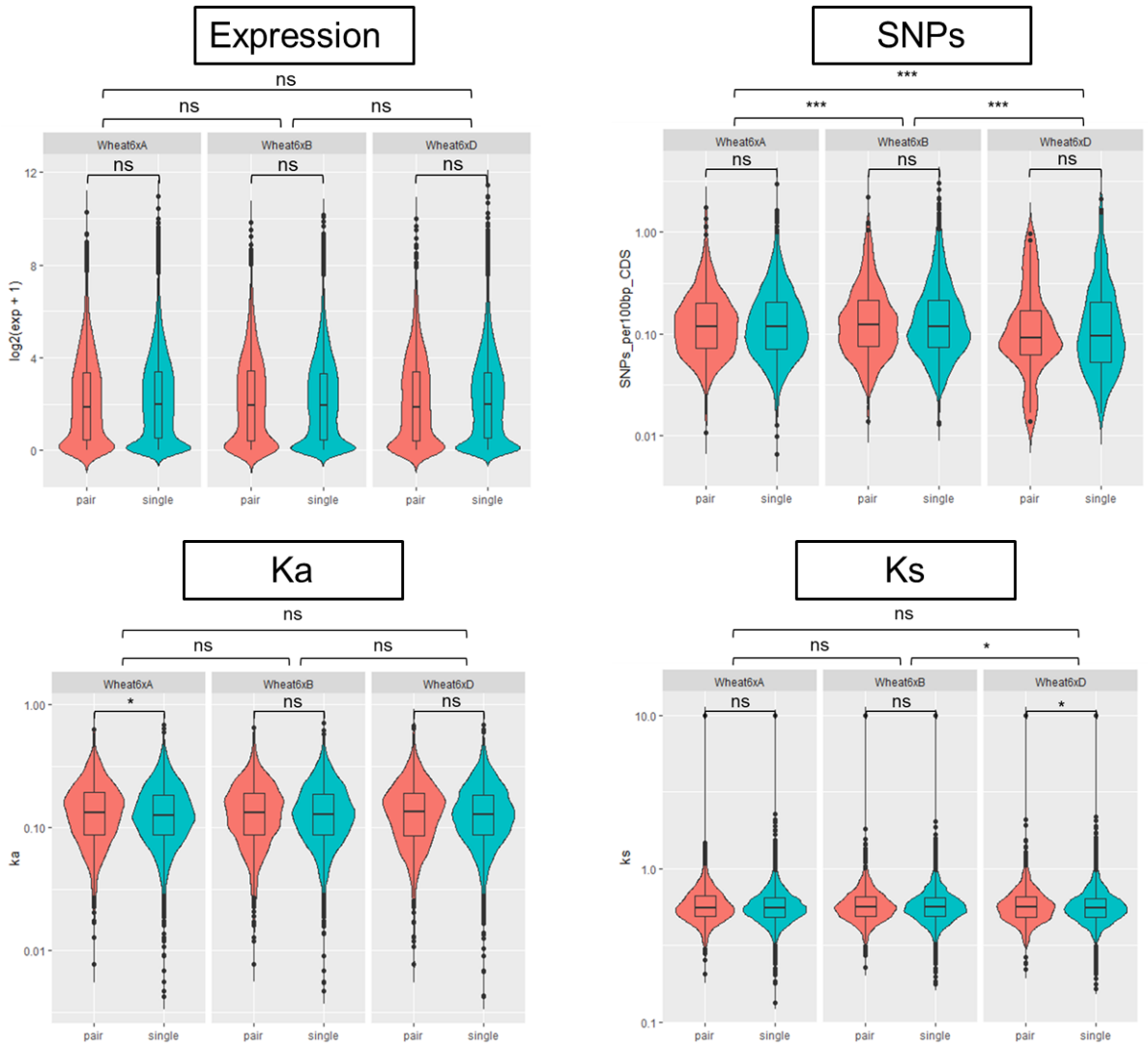
b.

		Brachy	*Brachy*	*Maize A*	*Maize A*	*Wheat6x*	*Wheat6x*	*Maize R*	*Maize R*
direction (**col**|**inv**)	Expression							X	X
	mCG promoter					X	X	X	X
	mCHG promoter					X	X	X	X
	mCHH promoter					X	X	X	X
	mCG gbM					X	X	X	X
	mCHG gbM					X	X	X	X
	mCHH gbM					X	X	X	X
	SNPs							X	X
	Ka							X	X
	Ks							X	X
Evolutionary state (**CS**|NCS)	Expression							X	X
	mCG promoter					X	X	X	X
	mCHG promoter					X	X	X	X
	mCHH promoter					X	X	X	X
	mCG gbM					X	X	X	X
	mCHG gbM					X	X	X	X
	mCHH gbM					X	X	X	X
	SNPs							X	X
	Ka							X	X
	Ks							X	X
AGK status (**sing**|**pair**)	Expression								
	mCG promoter					X	X		
	mCHG promoter					X	X		
	mCHH promoter					X	X		
	mCG gbM					X	X		
	mCHG gbM					X	X		
	mCHH gbM					X	X		
	SNPs								
	Ka								
	Ks								
Fraction (**LF**|**MF**)	Expression								
	mCG promoter					X	X		
	mCHG promoter					X	X		
	mCHH promoter					X	X		
	mCG gbM					X	X		
	mCHG gbM					X	X		
	mCHH gbM					X	X		
	SNPs								
	Ka								
	Ks								
Evolutionary speed (**SE**|**FE**)	Expression								
	mCG promoter					X	X		
	mCHG promoter					X	X		
	mCHH promoter					X	X		
	mCG gbM					X	X		
	mCHG gbM					X	X		
	mCHH gbM					X	X		
	SNPs								
	Ka								
	Ks								

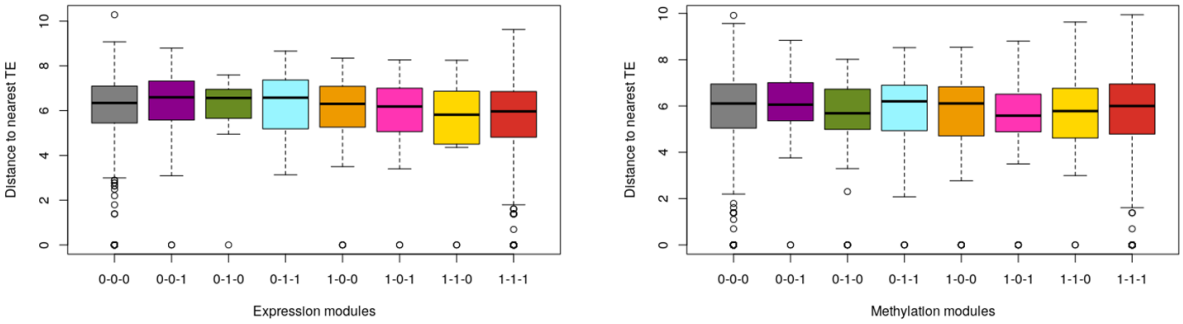
<u>Supplementary Figure 9</u>: Subgenome dominance in hexaploid wheat between A-B-D subgenomes. Comparison of gene expression (expressed as log2(exp+1)), SNP rate (expressed as SNP per 100bp in CDS), Ka and Ks values between the A, B and D subgenomes (x-axis) of hexaploid wheat. Statistical test p-value values are shown above the brackets.



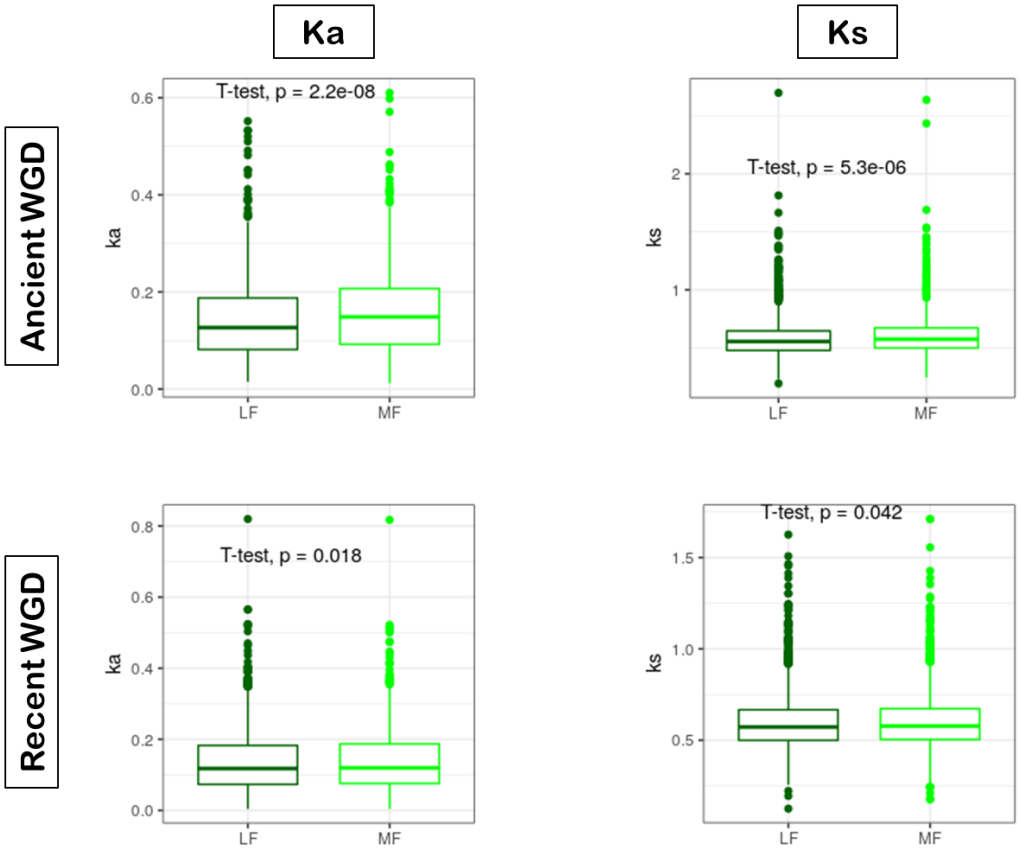
<u>Supplementary Figure 10</u>: Subgenome dominance in hexaploid wheat between recent A-B-D subgenomes for each ancient LF-MF subgenomes. Comparison of gene expression (expressed as log2(exp+1)), SNPs (expressed as SNP per 100bp in CDS), Ka and Ks values between A, B and D subgenomes of hexaploid wheat partitioned into singletons and pairs (x-axis). Values above the square brackets indicate whether the compared values show significant statistically differences. Stars indicate that the differences are significant, and their number reflects increased statistical significance. NS indicates the test is not significant. The square brackets above the boxes correspond to the comparisons between the three wheat subgenomes for a given gene category (singletons or pairs). The square brackets in the boxes correspond to comparisons between singletons and pairs within a given subgenome (either A, B or D).



<u>Supplementary Figure 11</u>: Impact of transposable elements (TEs) in methylation and expression differences observed between of conserved genes. Expression and methylation modules are represented in the x-axis and the distance to the nearest TEs (expressed in logarithm absolute value) in the y-axis. 1-1-1 corresponds to an expression observed in all the three developmental stages; 0-0-0 corresponds to genes not expressed in all the three developmental stages; Modules with both 0 and 1 represent differences in expression or methylation between the developmental stages. Zero indicates not methylated/expressed genes and 1 for methylated/expressed genes.

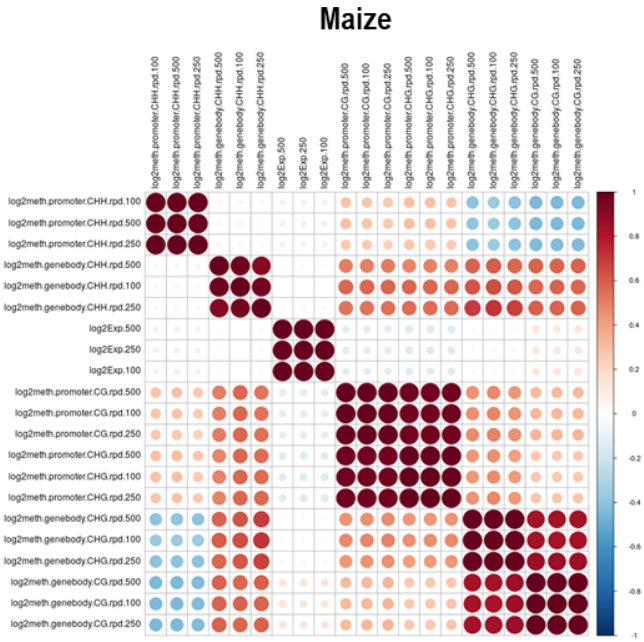


<u>Supplementary Figure 12</u>: Ka and Ks plasticity of duplicated genes. Non-synonymous (ka, left) and synonymous (ks, right) substitutions between the duplicated gene pairs in LF and MF compartments inherited from the ancestral shared (ρ) WGD (top) and the recent maize-specific WGD (bottom).

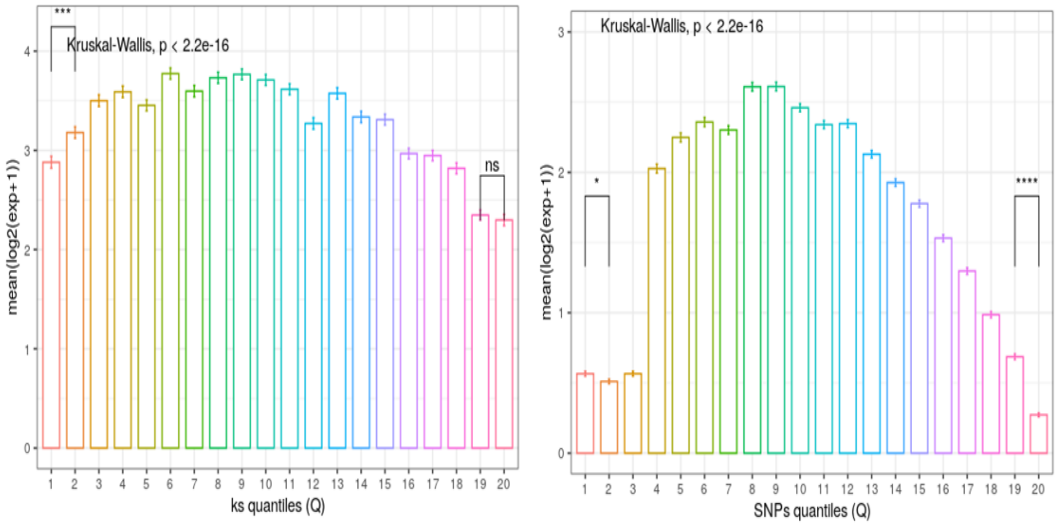


Supplementary Figure 13: Whole genome omics interplay. **a**. Whole genome correlation between genes expression and DNA methylation levels in the three contexts (CG, CHG and CHH) and for the three developmental stages (100DD, 250DD and 500DD), with red color for positive correlation and blue color for negative correlation. **b**. Interplay between gene expression level and SNP density in maize. The SNPs are represented by quantiles ranked in ascending order and gene expression by the logarithm of mean values of each quantile. **c**. Expression and methylation interplay for conserved and non-conserved genes. Methylation in gene promoter is represented in quantiles (Q1 to Q20), x-axis. Gene expression is represented as the mean value of genes in each quantile (log transformed), y-axis. Red curve for conserved genes and blue curve for non-conserved genes

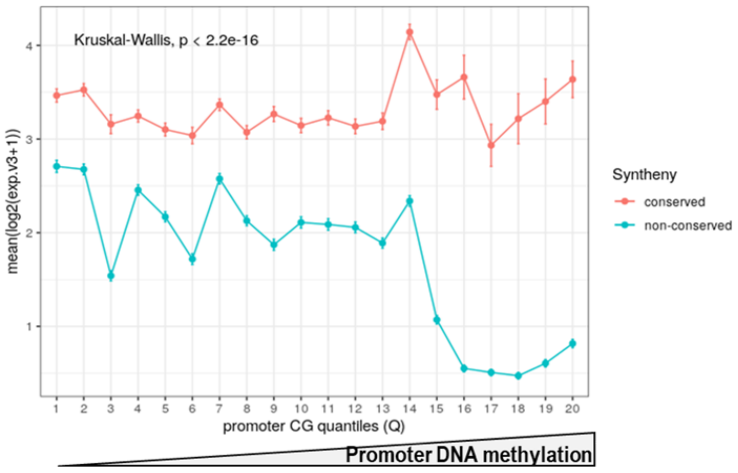
a.



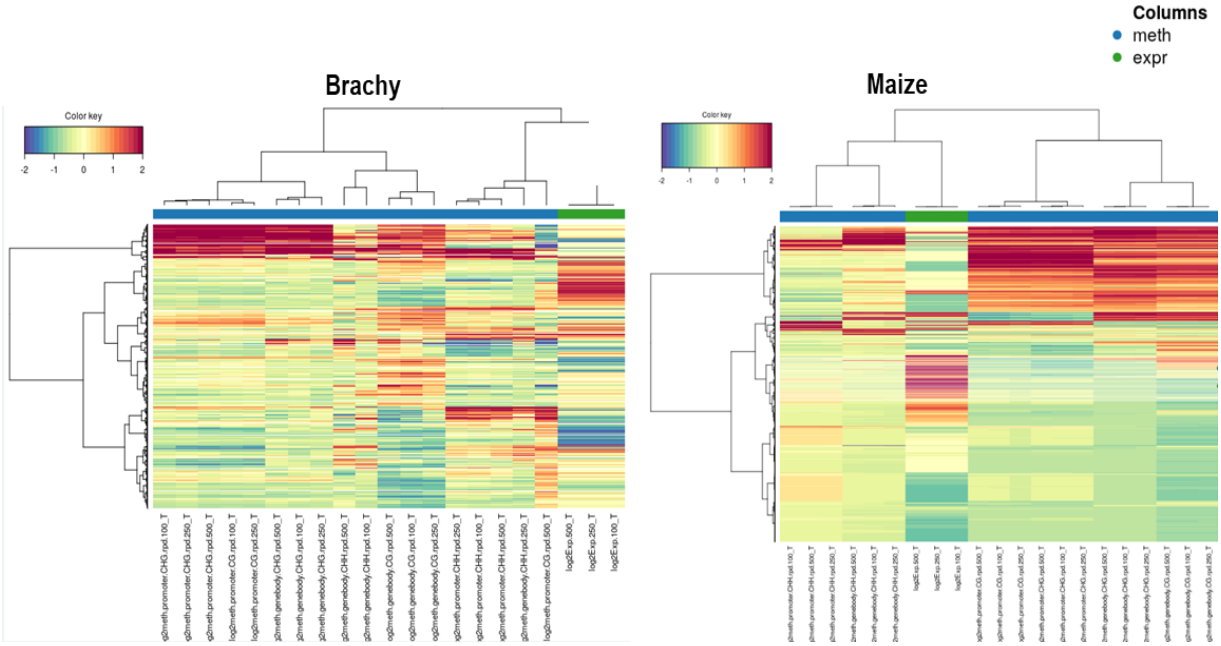
b.



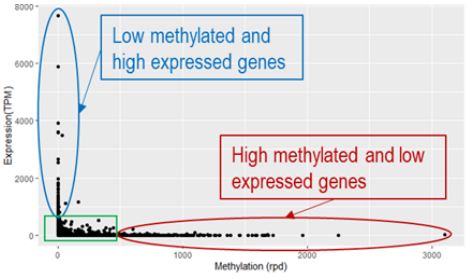
c.


Kruskal-Wallis, p < 2.2e-16
mean(log2(exp.v3+1))
promoter CG quantiles (Q)
Promoter DNA methylation
Syntheny
conserved
non-conserved

<u>Supplementary Figure 14:</u> Expression and methylation levels in Brachypodium and Maize. Mixomics data integration of (in columns) gene expression, promoter and gene-body methylation (in CG, CHG and CHH) for the three developmental stages in *Brachypodium* and maize for each of the annotated genes (in lines). Normalized expression and methylation levels of genes are shown with a color code (see legend top left)



<u>Supplementary Figure 15:</u> Expression and methylation gene outliers. ***Left panel***: Outliers detection with gene expression expressed in TPM (y-axis) and methylation expressed in rpd (read per density, x-axis). This allows the identification of hypermethyled and downregulated genes (in red) and hypomethylated and upregulated genes (in blue). ***Right panel***: Number of promoter and gene-body Hyper/Down (hypermethyled / downregulated) and Hypo/Up (hypomethylated / upregulated) genes in *Brachypodium* and maize for each methylation context (CG, CHG and CHH).



	CG		CHG		CHH	
	Hyper / Down	Hypo / Up	Hyper / Down	Hypo / Up	Hyper / Down	Hypo / Up
Brachy promoter	12	146	24	130	4	132
Brachy gene body	330	11	31	11	39	11
Maize promoter	35	7	38	8	10	8
Maize gene body	252	9	265	9	4	9

Annexe F

Mardoc et al. (en préparation), Données supplémentaires concernant les analyses bovines

Supplementary Figure 1 : Number of proteins selected for one or several phenotypes by PLS regressions.

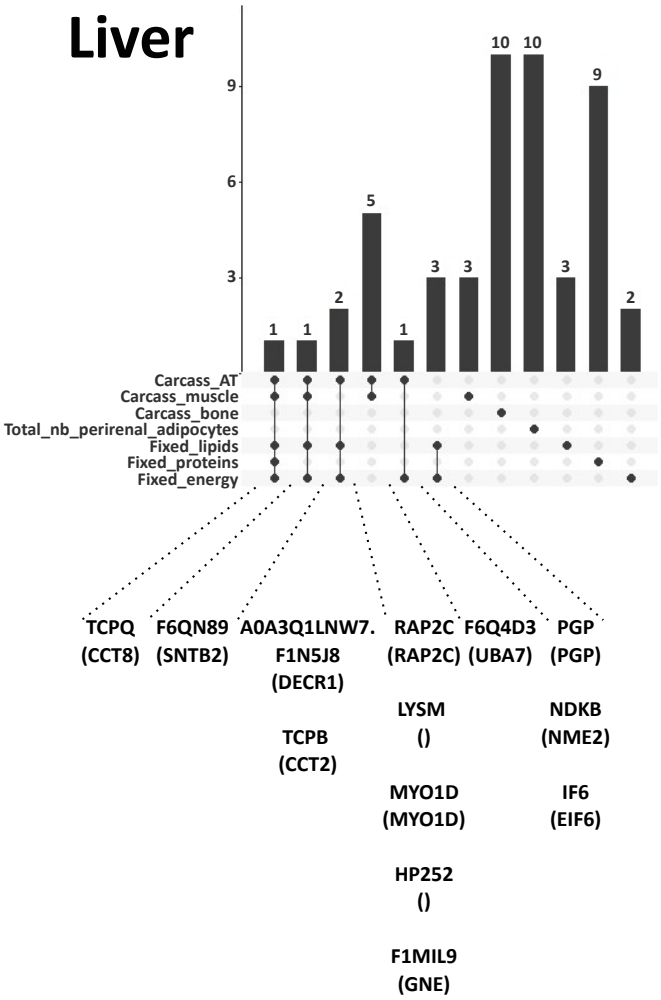
Plots by tissues (liver, muscle, fat) of the number of proteins selected for one or several of the seven phenotypes, using the top 10 proteins lists from PLS regressions. For proteins selected for several phenotypes, their name and associated gene(s) are displayed.

Supplementary Figure 2 : Comparison of proteins lists from PLS and MB-PLS regressions.

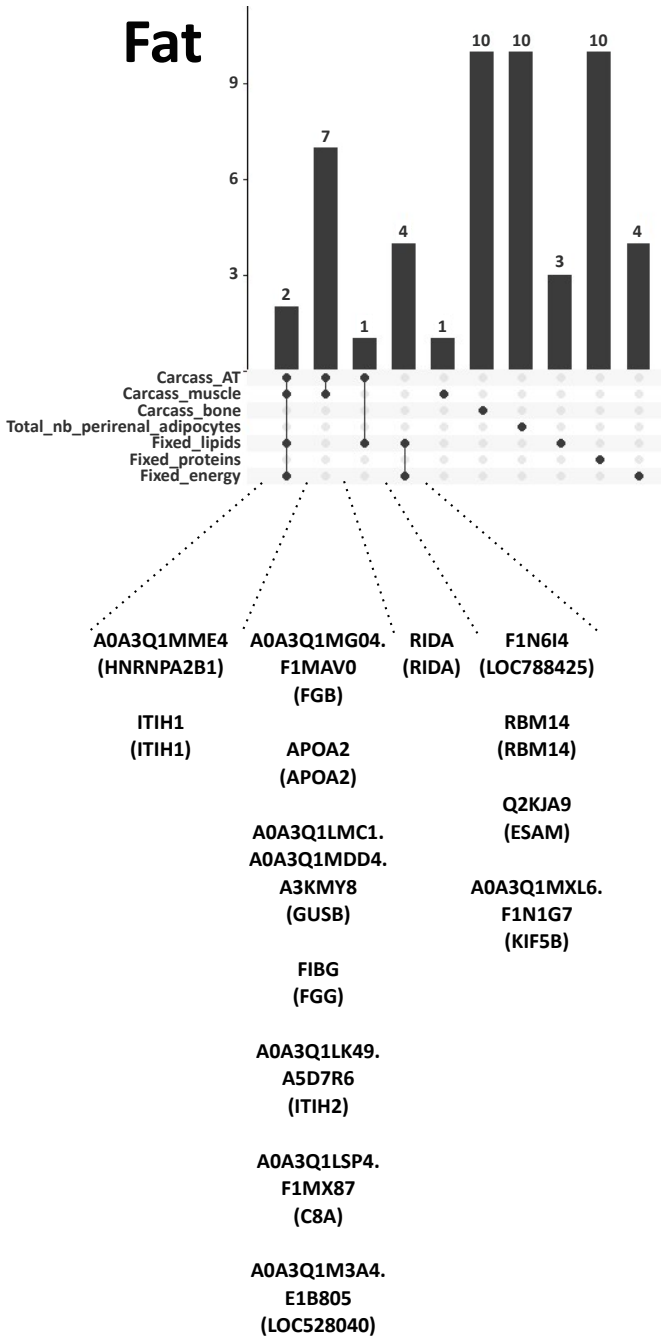
The top 10 proteins selected from each tissue (Liver, Fat, Muscle) on each of the 7 phenotypes using PLS regression with the mixOmics' function pls are displayed and compared to the list of proteins from the multi-block PLS regression block.spls. The proteins' color code indicates a positive or negative correlation between the protein and the corresponding phenotype. Each protein is identified with its Uniprot ID (without the "_BOVIN" termination), and its corresponding genes are indicated in parentheses if known.

Sup Fig 1

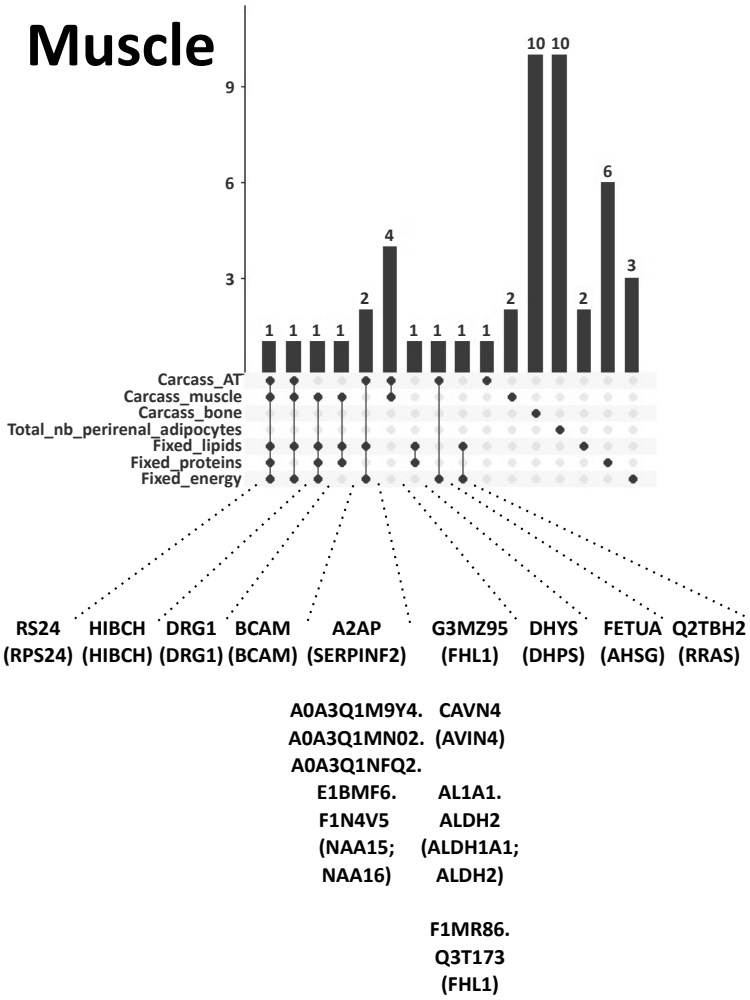
Liver



Fat



Muscle



Sup Fig 2

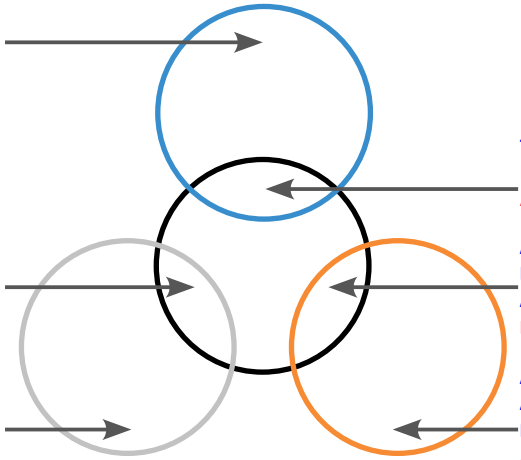
①

RAP2C (RAP2C)
MYO1D (MYO1D)
LYSM ()
TCPB (CCT2)
F6Q4D3 (UBA7)
HP252 ()
F1MIL9 (GNE)

RS24 (RPS24)
HIBCH (HIBCH)
A0A3Q1M9Y4.A0A3Q1MN02.A0A3Q1NFQ2
.E1BMF6.F1N4V5 (NAA15;NAA16)

CAVN4 (CAVIN4)
F1MR86.Q3T173 (FHL1)
G3MZ95 (FHL1)
A2AP (SERPINF2)
FETUA (AHSG)
AL1A1.ALDH2 (ALDH1A1;ALDH2)
TMM11 (TMEM11)

ISCA2 (ISCA2)



TCPQ (CCT8)
F6QN89 (SNTB2)
A0A3Q1LNW7.F1N5J8 (DECR1)

A0A3Q1MG04.F1MAV0 (FGB)
ITIH1 (ITIH1)
A0A3Q1MME4 (HNRNPA2B1)
RIDA (RIDA)

APOA2 (APOA2)
A0A3Q1LMC1.A0A3Q1MDD4.A3KMY8 (GUSB)
FIBG (FGG)
A0A3Q1LK49.A5D7R6 (ITIH2)
A0A3Q1LSP4.F1MX87 (C8A)
A0A3Q1M3A4.E1B805 (LOC528040)

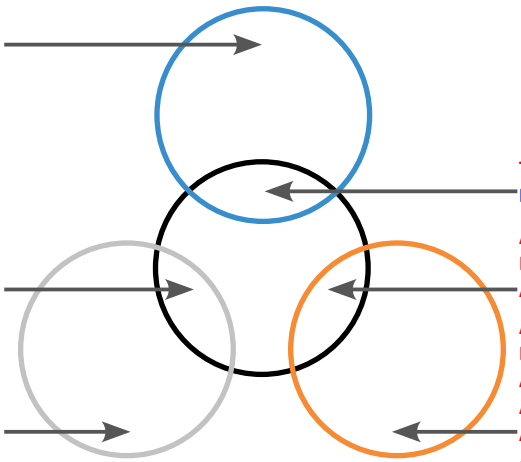
②

RAP2C (RAP2C)
LYSM ()
HP252 ()
F1MVM4 (DAO)
F1MIL9 (GNE)
DHR11 (DHRS11)
MYO1D (MYO1D)

RS24 (RPS24)
HIBCH (HIBCH)
DRG1 (DRG1)
BCAM (BCAM)

F1MR86.Q3T173 (FHL1)
G3MZ95 (FHL1)
CAVN4 (CAVIN4)
3HIDH (HIBADH)
RL24 (RPL24)
AL1A1.ALDH2 (ALDH1A1;ALDH2)

Q0P5H6 (TBL2)



TCPQ (CCT8)
F6QN89 (SNTB2)

A0A3Q1MG04.F1MAV0 (FGB)
ITIH1 (ITIH1)
A0A3Q1MME4 (HNRNPA2B1)

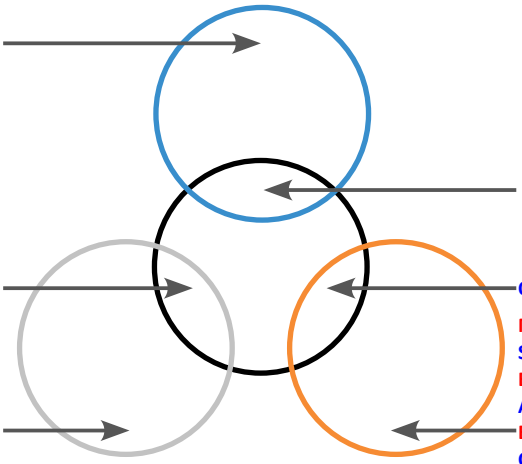
APOA2 (APOA2)
FIBG (FGG)
A0A3Q1LK49.A5D7R6 (ITIH2)
A0A3Q1LSP4.F1MX87 (C8A)
A0A3Q1M3A4.E1B805 (LOC528040)
A0A3Q1LMC1.A0A3Q1MDD4.A3KMY8 (GUSB)
FIBB (FGB)

③

TDH (TDH)
PRPS1 (PRPS1)
MYCBP (MYCBP)
A0A3Q1M5M9.A0A3Q1M5V8.A6H7A2 (ADA)
A0A3Q1LZ55.F1MBW3 (ACSL4)
E1BNF8 (FGGY)
IL18 (IL18)
PNPH (PNP)
UBE2K (UBE2K)

F1N1A3 (UBQLN1)
H4 ()

LASP1 (LASP1)
ADIPO (ADIPOQ)
A0A3Q1LZ09.A0A3Q1MA31.A0A3Q1MGQ6.F1MMD7 (ITIH4)
SEPT2 (SEPTIN2)
GCDH (GCDH)
A0A3Q1LMM6.F6QTA8 (PARVA)
A0A3Q1MAB7.A0A3Q1MHJ2.A0A3Q1MQX7
.A0A3Q1MTL8 (DCUN1D1)
TNPO1 (TNPO1)



Q3ZC85 (PLVAP)

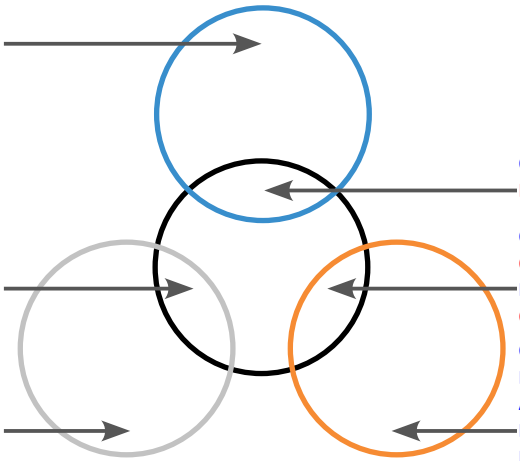
MBL2 (MBL)
SNAG (NAPG)
E1BCF8 (ZC3HAV1)
A0A3Q1MBX8.F1MZZ9 (CCDC50)
E1BIQ0 (PNPT1)
CSN4 (COPS4)
KPRB (PRPSAP2)
A0A3Q1LZ31.F1MJP7 (UAP1)
PPM1B (PPM1B)

④

F1MBS4 (SLC25A10)
A5PK70 (NME3)
OXSM (OXSM)
A6QL75 (PTP4A2)
GDIR3 (ARHGDIG)
Q2KIW4 (LCAT)
A0A3Q1MCH6.A0A3Q1N8V6.A5PK96 (ACP1)
A7MBE6.F1MPP7 (ACSM1;ACSM2A)

ITIH5 (ITIH5)
RSMB.RSMN (SNRPB;SNRPN)
CX7A2 (COX7A2)
NP1L1 (NAP1L1)

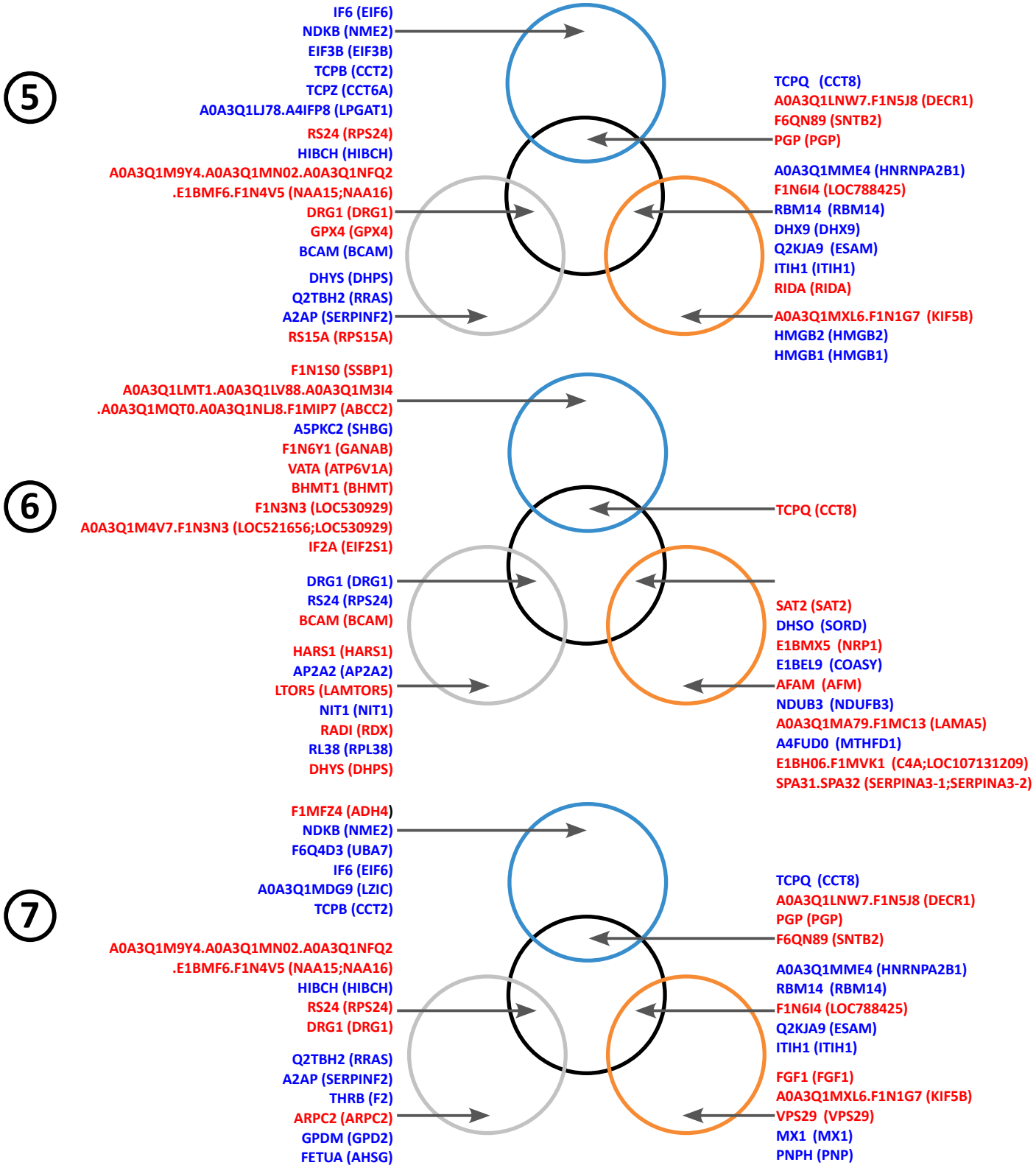
KNG1 (KNG1)
A0A3S5ZPM3 (PGD)
TFAM (TFAM)
Q3ZBG1 (RAB14)
A0A3Q1MHP4.A0A3Q1NKM0 ()
TRXR1 (TXNRD1)



OMA1 (OMA1)
HDHD2 (HDHD2)

CT027 ()
Q3ZBY8 (GMDS)
MAL2 (MAL2)
COQ6(COQ6)

CATL1.CATL2 (CTSL;CTSV)
PDLI2 (PDLIM2)
A0A3Q1LS09.Q3SX36 (TFG)
F1MBS4 (SLC25A10)
HPRT (HPRT1)
PIPNA (PITPNA)



PLS Liver

PLS Fat

PLS Muscle

MB-PLS

1 Carcass_AT

2 Carcass_muscle

3 Carcass_bone

4 Total_nb_perirenal_adipocytes

5 Fixed_lipids

6 Fixed_proteins

7 Fixed_energy

Positive correlation

Negative correlation

Supplementary Table 1: Top 10 liver proteins from PLS1 regression

Lists of the 10 most explicative proteins from liver tissue explaining each phenotype, selected by using mixomics' function pls then ordering proteins by their implication in the first computed component according to pls loadings' outputs. Proteins are defined with Uniprot ID without "_BOVIN" termination and corresponding genes are indicated if known. Pearson correlation values are given between each protein and corresponding phenotype. Finally, the column "Selected-by-MB-PLS" indicates if the selected protein using PLS regression is also selected with the multi-block PLS regression approach.

Phenotype	Uniprot	Corresponding_gene	Correlation	Selected_by_MB-PLS
Carcass_AT	TCPQ	CCT8	-0,863	TRUE
	F6QN89	SNTB2	0,819	TRUE
	RAP2C	RAP2C	-0,807	FALSE
	MYO1D	MYO1D	0,788	FALSE
	LYSM		-0,786	FALSE
	TCPB	CCT2	-0,781	FALSE
	F6Q4D3	UBA7	-0,78	FALSE
	HP252		0,78	FALSE
	F1MIL9	GNE	-0,779	FALSE
	A0A3Q1LNW7;F1N5J8	DECR1	0,778	TRUE
Carcass_muscle	TCPQ	CCT8	0,855	TRUE
	F6QN89	SNTB2	-0,813	TRUE
	ISCA2	ISCA2	-0,785	FALSE
	RAP2C	RAP2C	0,784	FALSE
	LYSM		0,783	FALSE
	HP252		-0,782	FALSE
	F1MVM4	DAO	-0,774	FALSE
	F1MIL9	GNE	0,773	FALSE
	DHR11	DHRS11	0,77	FALSE
	MYO1D	MYO1D	-0,767	FALSE
Carcass_bone	Q0P5H6	TBL2	0,729	FALSE
	TDH	TDH	-0,721	FALSE
	PRPS1	PRPS1	-0,715	FALSE
	MYCBP	MYCBP	-0,706	FALSE
	A0A3Q1M5M9;A0A3Q1M5V8;A6H7A2	ADA	0,682	FALSE
	A0A3Q1LZ55;F1MBW3	ACSL4	0,665	FALSE
	E1BNF8	FGGY	-0,665	FALSE
	IL18	IL18	0,654	FALSE
	PNPH	PNP	-0,653	FALSE
	UBE2K	UBE2K	0,648	FALSE
Total_nb_perirenal_adipocytes	OMA1	OMA1	-0,872	TRUE
	F1MBS4	SLC25A10	-0,791	FALSE
	A5PK70	NME3	-0,787	FALSE
	OXSM	OXSM	-0,767	FALSE
	A6QL75	PTP4A2	-0,759	FALSE
	GDIR3	ARHGDIG	-0,742	FALSE
	Q2KIW4	LCAT	-0,728	FALSE
	A0A3Q1MCH6;A0A3Q1N8V6;A5PK96	ACP1	0,719	FALSE
	A7MBE6;F1MPP7	ACSM1;ACSM2A	0,719	FALSE
	HDHD2	HDHD2	0,717	TRUE
Fixed_lipids	TCPQ	CCT8	-0,911	TRUE
	A0A3Q1LNW7;F1N5J8	DECR1	0,855	TRUE
	F6QN89	SNTB2	0,836	TRUE
	PGP	PGP	0,811	TRUE
	IF6	EIF6	-0,786	FALSE
	NDKB	NME2	-0,782	FALSE
	EIF3B	EIF3B	-0,775	FALSE
	TCPB	CCT2	-0,769	FALSE
	TCPZ	CCT6A	-0,767	FALSE
	A0A3Q1LJ78;A4IFP8	LPGAT1	-0,764	FALSE
Fixed_proteins	F1N1S0	SSBP1	0,805	FALSE
	A0A3Q1LMT1;A0A3Q1LV88;A0A3Q1M3I4; A0A3Q1MQT0;A0A3Q1NLJ8;F1MIP7	ABCC2	0,798	FALSE
	A5PKC2	SHBG	-0,791	FALSE
	F1N6Y1	GANAB	0,76	FALSE
	VATA	ATP6V1A	0,742	FALSE
	TCPQ	CCT8	0,74	TRUE
	BHMT1	BHMT	0,729	FALSE
	F1N3N3	LOC530929	0,728	FALSE
	A0A3Q1M4V7;F1N3N3	LOC521656;LOC530929	0,727	FALSE
	IF2A	EIF2S1	0,727	FALSE
Fixed_energy	TCPQ	CCT8	-0,856	TRUE
	A0A3Q1LNW7;F1N5J8	DECR1	0,842	TRUE
	PGP	PGP	0,825	TRUE
	F1MFZ4	ADH4	0,805	FALSE
	F6QN89	SNTB2	0,792	TRUE
	NDKB	NME2	-0,785	FALSE
	F6Q4D3	UBA7	-0,777	FALSE
	IF6	EIF6	-0,773	FALSE
	A0A3Q1MDG9	LZIC	-0,761	FALSE
	TCPB	CCT2	-0,759	FALSE

Supplementary Table 2: Top 10 fat proteins from PLS1 regression

Lists of the 10 most explicative proteins from fat tissue explaining each phenotype, selected by using mixomics' function pls then ordering proteins by their implication in the first computed component according to pls loadings' outputs. Proteins are defined with Uniprot ID without "_BOVIN" termination and corresponding genes are indicated if known. Pearson correlation values are given between each protein and corresponding phenotype. Finally, the column "Selected-by-MB-PLS" indicates if the selected protein using PLS regression is also selected with the multi-block PLS regression approach.

Phenotype	Uniprot	Corresponding_gene	Correlation	Selected_by_MB-PLS
Carcass_AT	A0A3Q1MG04;F1MAV0	FGB	-0,852	TRUE
	ITIH1	ITIH1	-0,838	TRUE
	APOA2	APOA2	-0,824	FALSE
	A0A3Q1LMC1;A0A3Q1MDD4;A3KMY8	GUSB	-0,806	FALSE
	FIBG	FGG	-0,801	FALSE
	A0A3Q1MME4	HNRNPA2B1	-0,791	TRUE
	A0A3Q1LK49;A5D7R6	ITIH2	-0,788	FALSE
	A0A3Q1LSP4;F1MX87	C8A	-0,788	FALSE
	A0A3Q1M3A4;E1B805	LOC528040	-0,775	FALSE
	RIDA	RIDA	0,774	TRUE
Carcass_muscle	A0A3Q1MG04;F1MAV0	FGB	0,856	TRUE
	ITIH1	ITIH1	0,85	TRUE
	APOA2	APOA2	0,824	FALSE
	FIBG	FGG	0,816	FALSE
	A0A3Q1LK49;A5D7R6	ITIH2	0,805	FALSE
	A0A3Q1LSP4;F1MX87	C8A	0,798	FALSE
	A0A3Q1M3A4;E1B805	LOC528040	0,79	FALSE
	A0A3Q1MME4	HNRNPA2B1	0,772	TRUE
	A0A3Q1LMC1;A0A3Q1MDD4;A3KMY8	GUSB	0,77	FALSE
	FIBB	FGB	0,767	FALSE
Carcass_bone	MBL2	MBL	0,696	FALSE
	SNAG	NAPG	-0,695	FALSE
	E1BCF8	ZC3HAV1	0,675	FALSE
	A0A3Q1MBX8;F1MZZ9	CCDC50	-0,667	FALSE
	E1BIQ0	PNPT1	0,642	FALSE
	CSN4	COPS4	-0,635	FALSE
	KPRB	PRPSAP2	-0,634	FALSE
	A0A3Q1LZ31;F1MJP7	UAP1	0,632	FALSE
	PPM1B	PPM1B	0,623	FALSE
	Q3ZC85	PLVAP	-0,62	TRUE
Total_nb_perirenal_adipocytes	CATL1;CATL2	CTSL;CTSV	-0,857	FALSE
	PDLI2	PDLIM2	-0,746	FALSE
	CT027		-0,732	TRUE
	Q3ZBY8	GMDS	0,699	TRUE
	A0A3Q1LS09;Q3SX36	TFG	-0,68	FALSE
	MAL2	MAL2	-0,678	TRUE
	F1MBS4	SLC25A10	-0,649	FALSE
	HPRT	HPRT1	-0,646	FALSE
	PIPNA	PITPNA	-0,645	FALSE
	COQ6	COQ6	0,644	TRUE
Fixed_lipids	A0A3Q1MME4	HNRNPA2B1	-0,891	TRUE
	F1N6I4	LOC788425	0,87	TRUE
	RBM14	RBM14	-0,85	TRUE
	DHX9	DHX9	-0,829	TRUE
	Q2KJA9	ESAM	-0,826	TRUE
	ITIH1	ITIH1	-0,81	TRUE
	A0A3Q1MXL6;F1N1G7	KIF5B	0,802	FALSE
	HMGB2	HMGB2	-0,798	FALSE
	RIDA	RIDA	0,796	TRUE
	HMGB1	HMGB1	-0,796	FALSE
Fixed_proteins	SAT2	SAT2	0,817	FALSE
	DHSO	SORD	-0,791	FALSE
	E1BMX5	NRP1	0,784	FALSE
	E1BEL9	COASY	-0,762	FALSE
	AFAM	AFM	0,759	FALSE
	NDUB3	NDUFB3	-0,755	FALSE
	A0A3Q1MA79;F1MC13	LAMA5	0,75	FALSE
	A4FUD0	MTHFD1	-0,749	FALSE
	E1BH06;F1MVK1	C4A;LOC107131209	0,749	FALSE
	SPA31;SPA32	SERPINA3-1;SERPINA3-2	0,749	FALSE
Fixed_energy	A0A3Q1MME4	HNRNPA2B1	-0,861	TRUE
	RBM14	RBM14	-0,836	TRUE
	F1N6I4	LOC788425	0,819	TRUE
	FGF1	FGF1	0,801	FALSE
	A0A3Q1MXL6;F1N1G7	KIF5B	0,787	FALSE
	Q2KJA9	ESAM	-0,779	TRUE
	ITIH1	ITIH1	-0,777	TRUE
	VPS29	VPS29	0,771	FALSE
	MX1	MX1	-0,769	FALSE
	PNPH	PNP	-0,767	FALSE

Supplementary Table 3: Top 10 muscle proteins from PLS1 regression

Lists of the 10 most explicative proteins from muscle tissue explaining each phenotype, selected by using mixomics' function pls then ordering proteins by their implication in the first computed component according to pls loadings' outputs. Proteins are defined with Uniprot ID without "_BOVIN" termination and corresponding genes are indicated if known. Pearson correlation values are given between each protein and corresponding phenotype. Finally, the column "Selected-by-MB-PLS" indicates if the selected protein using PLS regression is also selected with the multi-block PLS regression approach.

Phenotype	Uniprot	Corresponding_gene	Correlation	Selected_by_MB-PLS
Carcass_AT	RS24	RPS24	0,901	TRUE
	HIBCH	HIBCH	-0,781	TRUE
	CAVN4	CAVIN4	-0,751	FALSE
	F1MR86;Q3T173	FHL1	0,745	FALSE
	G3MZ95	FHL1	0,734	FALSE
	A2AP	SERPINF2	-0,718	FALSE
	FETUA	AHSG	-0,714	FALSE
	AL1A1;ALDH2	ALDH1A1;ALDH2	0,709	FALSE
	TMM11	TMEM11	0,709	FALSE
	A0A3Q1M9Y4;A0A3Q1MN02; A0A3Q1NFQ2;E1BMF6;F1N4V5	NAA15;NAA16	0,709	TRUE
Carcass_muscle	RS24	RPS24	-0,915	TRUE
	HIBCH	HIBCH	0,775	TRUE
	F1MR86;Q3T173	FHL1	-0,756	FALSE
	G3MZ95	FHL1	-0,747	FALSE
	CAVN4	CAVIN4	0,742	FALSE
	3HIDH	HIBADH	0,731	FALSE
	DRG1	DRG1	-0,718	TRUE
	RL24	RPL24	-0,707	FALSE
	BCAM	BCAM	0,707	TRUE
	AL1A1;ALDH2	ALDH1A1;ALDH2	-0,705	FALSE
Carcass_bone	LASP1	LASP1	0,812	FALSE
	ADIPO	ADIPOQ	0,74	FALSE
	A0A3Q1LZ09;A0A3Q1MA31; A0A3Q1MGQ6;F1MMD7	ITIH4	0,739	FALSE
	F1N1A3	UBQLN1	0,728	TRUE
	sept-02	SEPTIN2	0,725	FALSE
	GCDH	GCDH	-0,723	FALSE
	A0A3Q1LMM6;F6QTA8	PARVA	0,691	FALSE
	A0A3Q1MAB7;A0A3Q1MHJ2; A0A3Q1MQX7;A0A3Q1MTL8	DCUN1D1	-0,691	FALSE
	TNPO1	TNPO1	0,687	FALSE
	H4		0,685	TRUE
Total_nb_perirenal_adipocytes	ITIH5	ITIH5	-0,83	TRUE
	KNG1	KNG1	-0,763	FALSE
	A0A3S5ZPM3	PGD	0,741	FALSE
	RSMB;RSMN	SNRPB;SNRPN	-0,718	TRUE
	CX7A2	COX7A2	-0,698	TRUE
	TFAM	TFAM	0,666	FALSE
	NP1L1	NAP1L1	-0,661	TRUE
	Q3ZBG1	RAB14	-0,653	FALSE
	A0A3Q1MHP4;A0A3Q1NKM0		0,631	FALSE
	TRXR1	TXNRD1	0,627	FALSE
Fixed_lipids	RS24	RPS24	0,849	TRUE
	HIBCH	HIBCH	-0,829	TRUE
	A0A3Q1M9Y4;A0A3Q1MN02; A0A3Q1NFQ2;E1BMF6;F1N4V5	NAA15;NAA16	0,822	TRUE
	DRG1	DRG1	0,73	TRUE
	GPX4	GPX4	0,69	TRUE
	BCAM	BCAM	-0,687	TRUE
	DHYS	DHPS	-0,682	FALSE
	Q2TBH2	RRAS	-0,663	FALSE
	A2AP	SERPINF2	-0,663	FALSE
	RS15A	RPS15A	0,655	FALSE
Fixed_proteins	HARS1	HARS1	0,764	FALSE
	AP2A2	AP2A2	-0,702	FALSE
	LTOR5	LAMTOR5	0,699	FALSE
	NIT1	NIT1	-0,696	FALSE
	DRG1	DRG1	-0,691	TRUE
	RS24	RPS24	-0,671	TRUE
	RADI	RDX	0,665	FALSE
	BCAM	BCAM	0,658	TRUE
	RL38	RPL38	-0,642	FALSE
	DHYS	DHPS	0,642	FALSE
Fixed_energy	A0A3Q1M9Y4;A0A3Q1MN02; A0A3Q1NFQ2;E1BMF6;F1N4V5	NAA15;NAA16	0,902	TRUE
	HIBCH	HIBCH	-0,84	TRUE
	RS24	RPS24	0,803	TRUE
	Q2TBH2	RRAS	-0,756	FALSE
	A2AP	SERPINF2	-0,685	FALSE
	THRB	F2	-0,673	FALSE
	ARPC2	ARPC2	0,667	FALSE
	DRG1	DRG1	0,657	TRUE
	GPDM	GPD2	-0,647	FALSE
	FETUA	AHSG	-0,635	FALSE

Supplementary Table 4: Functional analysis of the 2 clusters identified from the list of the 149 proteins the most explicative of each phenotype for each tissue

Cluster 1 : 31 proteins strongly positively correlated with carcass fat, fixed lipids or fixed energy and negatively correlated with carcass muscle and fixed proteins; Cluster 2 : 55 proteins positively correlated with carcass muscle and fixed proteins and negatively correlated with fat, lipids or energy depositions.

All the enrichment analyses were performed with B. taurus genome annotation and calculated based on hypergeometric distribution followed by false discovery rate (FDR) correction, with a corrected p-value cut-off of 0.05. ProteInside web-service was used to perform enrichment analysis in the Gene Ontology (GO) term Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were perfomed with ShinyGO web-service 0.77.

	GO_BP	Name	P-value (fdr)	Gene Name
Cluster 1	GO:0017196	N-terminal peptidyl-methionine acetylation	1.587e-2	NAA15 NAA16
	GO:0110095	cellular detoxification of aldehyde	1.587e-2	ADH4 ALDH1A1
	GO:0046185	aldehyde catabolic process	1.587e-2	ADH4 ALDH2
	GO:0110096	cellular response to aldehyde	1.587e-2	ADH4 ALDH1A1
	GO:1901615	organic hydroxy compound metabolic process	1.587e-2	ADH4 DAO FGF1 PGP SORD
	GO:0018206	peptidyl-methionine modification	1.587e-2	NAA15 NAA16
	GO:0019400	alditol metabolic process	1.606e-2	PGP SORD
	GO:0006474	N-terminal protein amino acid acetylation	2.157e-2	NAA15 NAA16
	GO:0006066	alcohol metabolic process	2.679e-2	ADH4 FGF1 PGP SORD
	GO:0046164	alcohol catabolic process	2.723e-2	ADH4 SORD
	GO:0044281	small molecule metabolic process	3.135e-2	ADH4 COASY DAO DECR1 FGF1 GPX4 PGP SORD
	GO:0044282	small molecule catabolic process	3.149e-2	ADH4 DAO DECR1 SORD
	GO:0031365	N-terminal protein amino acid modification	3.293e-2	NAA15 NAA16
	GO:0051164	L-xylitol metabolic process	3.887e-2	SORD
	GO:0036088	D-serine catabolic process	3.887e-2	DAO
	GO:0046436	D-alanine metabolic process	3.887e-2	DAO
	GO:0006060	sorbitol metabolic process	3.887e-2	SORD
	GO:0046370	fructose biosynthetic process	3.887e-2	SORD
	GO:0051160	L-xylitol catabolic process	3.887e-2	SORD
	GO:0006062	sorbitol catabolic process	3.887e-2	SORD
	GO:0030392	fructosamine catabolic process	3.887e-2	ALDH1A1
	GO:0019407	hexitol catabolic process	3.887e-2	SORD
	GO:0019527	pentitol catabolic process	3.887e-2	SORD
	GO:0055130	D-alanine catabolic process	3.887e-2	DAO
	GO:1901617	organic hydroxy compound biosynthetic process	3.887e-2	DAO FGF1 PGP
	GO:1901616	organic hydroxy compound catabolic process	3.887e-2	ADH4 SORD
	GO:0030389	fructosamine metabolic process	3.887e-2	ALDH1A1
	GO:0019519	pentitol metabolic process	3.887e-2	SORD
	GO:0045861	negative regulation of proteolysis	1.763592860202999e-05	C4A F2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0051346	negative regulation of hydrolase activity	1.763592860202999e-05	APOA2 C4A ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 RDX SERPINA3-1
	GO:0042730	fibrinolysis	1.986670767991634e-05	F2 FGB FGG SERPINF2
	GO:0051248	negative regulation of protein metabolic process	0.0001494023327737758	C4A DHX9 EIF2S1 EIF6 F2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0010951	negative regulation of endopeptidase activity	0.0001718335629749467	C4A ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 SERPINA3-1
	GO:0010466	negative regulation of peptidase activity	0.0001718335629749467	C4A ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 SERPINA3-1
	GO:1901564	organonitrogen compound metabolic process	0.00020697019491331921	APOA2 ATP6V1A BHMT C4A DHPS DHX9 EIF2S1 EIF3B EIF6 F2 FGB FGG GANAB HARS1 HIBADH HIBCH ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 NME2 NRP1 PNP RAP2C RBM14 RDX SAT2 SERPINA3-1 SERPINF2 UBA7
	GO:0030195	negative regulation of blood coagulation	0.000226763745942155	F2 FGB FGG SERPINF2
	GO:1900047	negative regulation of hemostasis	0.0002299276145543636	F2 FGB FGG SERPINF2
	GO:0050790	regulation of catalytic activity	0.0002498576432382683	APOA2 C4A CCT2 DHX9 F2 HMGB2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 NRP1 RAP2C RDX SERPINA3-1 SSBP1
	GO:0050819	negative regulation of coagulation	0.0002498576432382683	F2 FGB FGG SERPINF2
	GO:0030193	regulation of blood coagulation	0.0003767459140090367 6	F2 FGB FGG SERPINF2
	GO:0061045	negative regulation of wound healing	0.0003767459140090367 6	F2 FGB FGG SERPINF2
	GO:0006958	complement activation, classical pathway	0.0003775957123306470 6	C4A C8A LOC107131209 LOC528040
	GO:1900046	regulation of hemostasis	0.0003787112270360857	F2 FGB FGG SERPINF2
	GO:0043086	negative regulation of catalytic activity	0.0003793138507119277 8	APOA2 C4A ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 RDX SERPINA3-1
	GO:0006959	humoral immune response	0.0003793138507119277 8	C4A C8A F2 FGB LOC107131209 LOC528040
	GO:0050818	regulation of coagulation	0.0008333703966419471	F2 FGB FGG SERPINF2
	GO:0030162	regulation of proteolysis	0.0009217643153715534	C4A F2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0051246	regulation of protein metabolic process	0.001010522606177786	C4A DHX9 EIF2S1 EIF3B EIF6 F2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 NRP1 RAP2C RDX SERPINA3-1 SERPINF2
	GO:1903035	negative regulation of response to wounding	0.0010673163223561858	F2 FGB FGG SERPINF2
	GO:0002455	humoral immune response mediated by circulating imm	0.0010673163223561858	C4A C8A LOC107131209 LOC528040
	GO:0031639	plasminogen activation	0.0010673163223561858	FGB FGG SERPINF2
	GO:0052548	regulation of endopeptidase activity	0.0010673163223561858	C4A ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 SERPINA3-1
	GO:0051336	regulation of hydrolase activity	0.0010673163223561858	APOA2 C4A HMGB2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 RDX SERPINA3-1
	GO:1902042	negative regulation of extrinsic apoptotic signaling path	0.0010673163223561858	FGB FGG HMGB2
	GO:0006574	valine catabolic process	0.0011281852337694297	HIBADH HIBCH
	GO:0052547	regulation of peptidase activity	0.0014847419402196824	C4A ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 SERPINA3-1
	GO:0006956	complement activation	0.0019986878915142178	C4A C8A LOC107131209 LOC528040
	GO:0080090	regulation of primary metabolic process	0.0028869622512937626	APOA2 C4A CAVIN4 CCT2 CCT6A DHX9 EIF2S1 EIF3B EIF6 F2 HMGB2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 LPGAT1 NME2 NRP1 RAP2C RBM14 RDX SERPINA3-1 SERPINF2 SSBP1
	GO:0065009	regulation of molecular function	0.0036936182836115575	APOA2 C4A CCT2 DHX9 F2 HMGB2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 NRP1 RAP2C RDX SERPINA3-1 SSBP1
	GO:0061041	regulation of wound healing	0.0037722041161398928	F2 FGB FGG SERPINF2
	GO:1902041	regulation of extrinsic apoptotic signaling pathway via c	0.004303563051391728	FGB FGG HMGB2
	GO:0048518	positive regulation of biological process	0.005720229442983612	AHSG APOA2 C4A C8A CAVIN4 CCT2 CCT6A DHPS DHX9 EIF2S1 EIF6 F2 FGB FGG HMGB2 LOC107131209 LOC528040 LPGAT1 NME2 NRP1 RAP2C RBM14 RDX SERPINF2 SSBP1
	GO:1903829	positive regulation of protein localization	0.005753030792131008 5	CCT2 CCT6A F2 FGB FGG RDX
	GO:1903034	regulation of response to wounding	0.006785254764669 8	F2 FGB FGG SERPINF2
	GO:0044092	negative regulation of molecular function	0.006785254764669 8	APOA2 C4A ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 RDX SERPINA3-1
	GO:0044281	small molecule metabolic process	0.006785254764669 8	APOA2 ATP6V1A BHMT EIF6 GNE HARS1 HIBADH HIBCH LOC530929 LPGAT1 NME2 PNP
	GO:0006573	valine metabolic process	0.007230059382683062	HIBADH HIBCH
	GO:0019724	B cell mediated immunity	0.008650752414547453	C4A C8A LOC107131209 LOC528040
	GO:0016064	immunoglobulin mediated immune response	0.008650752414547453	C4A C8A LOC107131209 LOC528040
	GO:1904851	positive regulation of establishment of protein localizat	0.008650752414547453	CCT2 CCT6A
	GO:0006957	complement activation, alternative pathway	0.010037359729095831	C8A LOC528040
	GO:0070203	regulation of establishment of protein localization to te	0.010037359729095831	CCT2 CCT6A
	GO:0009605	response to external stimulus	0.010037359729095831	AHSG C4A C8A EIF2S1 F2 FGB FGG HMGB2 LAMA5 LOC107131209 LOC528040 NRP1 RBM14 SERPINF2
	GO:0007596	blood coagulation	0.010170526105698791	F2 FGB FGG SERPINF2
	GO:1904816	positive regulation of protein localization to chromoson	0.010206436171423073	CCT2 CCT6A
	GO:0007599	hemostasis	0.010206436171423073	F2 FGB FGG SERPINF2
	GO:0071704	organic substance metabolic process	0.010206436171423073	APOA2 ATP6V1A BHMT C4A CAVIN4 CCT2 CCT6A DHPS DHRS11 DHX9 EIF2S1 EIF3B EIF6 F2 FGB FGG GANAB GNE GUSB HARS1 HIBADH HIBCH HMGB2 HNRNPA2B1 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 LOC530929 LPGAT1 NME2 NRP1 PNP RAP2C RBM14 RDX SAT2 SERPINA3-1 SERPINF2 SSBP1 UBA7
	GO:0050817	coagulation	0.010206436171423073	F2 FGB FGG SERPINF2
	GO:0019538	protein metabolic process	0.010206436171423073	APOA2 C4A DHPS DHX9 EIF2S1 EIF3B EIF6 F2 FGB FGG GANAB HARS1 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 NRP1 RAP2C RBM14 RDX SERPINA3-1 SERPINF2 UBA7
	GO:0072378	blood coagulation, fibrin clot formation	0.010206436171423073	FGB FGG
	GO:0070202	regulation of establishment of protein localization to ch	0.010206436171423073	CCT2 CCT6A
	GO:0051171	regulation of nitrogen compound metabolic process	0.010206436171423073	C4A CAVIN4 CCT2 CCT6A DHX9 EIF2S1 EIF3B EIF6 F2 HMGB2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 NME2 NRP1 RAP2C RBM14 RDX SERPINA3-1 SERPINF2 SSBP1
	GO:0032967	positive regulation of collagen biosynthetic process	0.010824432098222123	F2 SERPINF2
	GO:0044238	primary metabolic process	0.010824432098222123	APOA2 ATP6V1A BHMT C4A CAVIN4 CCT2 CCT6A DHPS DHRS11 DHX9 EIF2S1 EIF3B EIF6 F2 FGB FGG GANAB GNE GUSB HARS1 HIBADH HIBCH HMGB2 HNRNPA2B1 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 LPGAT1 NME2 NRP1 PNP RAP2C RBM14 RDX SERPINA3-1 SERPINF2 SSBP1 UBA7
	GO:0031638	zymogen activation	0.010824432098222123	FGB FGG SERPINF2

	GO:0034116	positive regulation of heterotypic cell-cell adhesion	0.010824432098222123	FGB FGG
	GO:0034109	homotypic cell-cell adhesion	0.010824432098222123	FGB FGG RDX
	GO:0045907	positive regulation of vasoconstriction	0.010824432098222123	FGB FGG
	GO:0050821	protein stabilization	0.010824432098222123	AFM APOA2 CCT2 CCT6A
	GO:0051050	positive regulation of transport	0.011747851697314 3	AHSG APOA2 DHX9 F2 FGB FGG RDX
	GO:1904814	regulation of protein localization to chromosome, telon	0.011985556738883976	CCT2 CCT6A
	GO:0006508	proteolysis	0.011985556738883976	C4A F2 FGB FGG ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 SERPINA3-1 SERPINF2 UBA7
	GO:0010714	positive regulation of collagen metabolic process	0.011985556738883976	F2 SERPINF2
	GO:0051172	negative regulation of nitrogen compound metabolic pr	0.012335369345146226	C4A DHX9 EIF2S1 EIF6 F2 HMGB2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0031589	cell-substrate adhesion	0.012963025796399706	BCAM FGB FGG LAMA5 NRP1
	GO:0072376	protein activation cascade	0.012963025796399706	FGB FGG
	GO:0008625	extrinsic apoptotic signaling pathway via death domain	0.012963025796399706	FGB FGG HMGB2
	GO:0032965	regulation of collagen biosynthetic process	0.012963025796399706	F2 SERPINF2
	GO:0019081	viral translation	0.01393725532464129	DHX9 EIF3B
	GO:0045652	regulation of megakaryocyte differentiation	0.01393725532464129	EIF6 HMGB2
	GO:0009083	branched-chain amino acid catabolic process	0.01393725532464129	HIBADH HIBCH
	GO:0030194	positive regulation of blood coagulation	0.01393725532464129	F2 SERPINF2
	GO:1900048	positive regulation of hemostasis	0.01393725532464129	F2 SERPINF2
	GO:0002443	leukocyte mediated immunity	0.014896184679101336	C4A C8A F2 LOC107131209 LOC528040
	GO:0070200	establishment of protein localization to telomere	0.015408932034394848	CCT2 CCT6A
	GO:0007160	cell-matrix adhesion	0.015408932034394848	BCAM FGB FGG NRP1
	GO:0030168	platelet activation	0.015408932034394848	F2 FGB FGG
	GO:2001237	negative regulation of extrinsic apoptotic signaling path	0.015858986246380553	FGB FGG HMGB2
	GO:0050820	positive regulation of coagulation	0.016660032625795244	F2 SERPINF2
	GO:0034114	regulation of heterotypic cell-cell adhesion	0.01783121815153944	FGB FGG
	GO:0010712	regulation of collagen metabolic process	0.01783121815153944	F2 SERPINF2
	GO:0006595	polyamine metabolic process	0.01783121815153944	DHPS SAT2
				APOA2 ATP6V1A BHMT C4A CAVIN4 CCT2 CCT6A DHPS DHX9 EIF2S1 EIF3B EIF6 F2 FGB FGG GANAB GNE HARS1 HIBADH HIBCH HMGB2 HNRNPA2B1 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 NME2 NRP1 PNP RAP2C RBM14 RDX SAT2 SERPINA3-1 SERPINF2 SSBP1
	GO:0006807	nitrogen compound metabolic process	0.01783121815153944	UBA7
	GO:0006954	inflammatory response	0.018400022819004443	AHSG C4A F2 HMGB2 LOC107131209 LOC528040
				APOA2 ATP6V1A BHMT C4A CAVIN4 CCT2 CCT6A DHPS DHRS11 DHX9 EIF2S1 EIF3B EIF6 F2 FGB FGG GANAB GNE GUSB HARS1 HIBADH HIBCH HMGB2 HNRNPA2B1 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 LOC530929 LPGAT1 NME2 NRP1 PNP RAP2C RBM14 RDX
	GO:0008152	metabolic process	0.018689764484875537	SAT2 SERPINA3-1 SERPINF2 SSBP1 UBA7
	GO:0098542	defense response to other organism	0.018689764484875537	C4A C8A F2 FGB HMGB2 LOC107131209 LOC528040 RBM14
	GO:0048522	positive regulation of cellular process	0.018689764484875537	AHSG APOA2 CAVIN4 CCT2 CCT6A DHPS DHX9 EIF2S1 EIF6 F2 FGB FGG HMGB2 LPGAT1 NME2 NRP1 RAP2C RBM14 RDX SERPINF2 SSBP1
	GO:0002252	immune effector process	0.020286605003018 4	APOA2 C4A C8A F2 LOC107131209 LOC528040
	GO:0097746	blood vessel diameter maintenance	0.020286605003018 4	FGB FGG SERPINF2
	GO:0007229	integrin-mediated signaling pathway	0.020286605003018 4	LAMA5 NME2 NRP1
	GO:0035296	regulation of tube diameter	0.020286605003018 4	FGB FGG SERPINF2
	GO:0006953	acute-phase response	0.020286605003018 4	AHSG F2
	GO:0032231	regulation of actin filament bundle assembly	0.020286605003018 4	NRP1 RDX SERPINF2
	GO:0035150	regulation of tube size	0.02079059884996597	FGB FGG SERPINF2
	GO:0031325	positive regulation of cellular metabolic process	0.021222829704794985	CAVIN4 CCT2 CCT6A DHX9 EIF6 F2 HMGB2 LPGAT1 NME2 NRP1 RAP2C RBM14 SERPINF2 SSBP1
	GO:0060255	regulation of macromolecule metabolic process	0.021222829704794985	APOA2 C4A CAVIN4 CCT2 CCT6A DHX9 EIF2S1 EIF3B EIF6 F2 HMGB2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 NME2 NRP1 RAP2C RBM14 RDX SERPINA3-1 SERPINF2 SSBP1
	GO:2000352	negative regulation of endothelial cell apoptotic proces	0.021222829704794985	FGB FGG
	GO:0009611	response to wounding	0.02149903236127843	F2 FGB FGG NRP1 SERPINF2
	GO:0060341	regulation of cellular localization	0.022474159586157144	CCT2 CCT6A DHX9 F2 FGB FGG RDX
	GO:0032964	collagen biosynthetic process	0.022474159586157144	F2 SERPINF2
	GO:0009081	branched-chain amino acid metabolic process	0.023502482831145522	HIBADH HIBCH
	GO:0002682	regulation of immune system process	0.023502482831145522	APOA2 C4A C8A DHPS EIF6 HMGB2 LOC107131209 LOC528040 RBM14
	GO:0070199	establishment of protein localization to chromosome	0.023502482831145522	CCT2 CCT6A
	GO:1904951	positive regulation of establishment of protein localizat	0.023502482831145522	CCT2 CCT6A FGB FGG
	GO:0019222	regulation of metabolic process	0.023502482831145522	APOA2 C4A CAVIN4 CCT2 CCT6A DHX9 EIF2S1 EIF3B EIF6 F2 HMGB2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 LPGAT1 NME2 NRP1 RAP2C RBM14 RDX SERPINA3-1 SERPINF2 SSBP1
	GO:0050778	positive regulation of immune response	0.023852774198222522	C4A C8A HMGB2 LOC107131209 LOC528040 RBM14
	GO:0048519	negative regulation of biological process	0.023852774198222522	AHSG APOA2 C4A DHX9 EIF2S1 EIF6 F2 FGB FGG HMGB2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 NME2 RAP2C RBM14 RDX SERPINA3-1 SERPINF2
	GO:0009891	positive regulation of biosynthetic process	0.024138139651623626	CAVIN4 CCT2 CCT6A DHX9 EIF6 F2 HMGB2 LPGAT1 NME2 RBM14 SERPINF2
	GO:2001252	positive regulation of chromosome organization	0.024180617917502464	CCT2 CCT6A SSBP1
	GO:0006413	translational initiation	0.024718983945535893	EIF2S1 EIF3B EIF6
	GO:0050878	regulation of body fluid levels	0.025335886536374034	F2 FGB FGG SERPINF2
	GO:0051054	positive regulation of DNA metabolic process	0.02673115742846895	CCT2 CCT6A DHX9 SSBP1
	GO:0071693	protein transport within extracellular region	0.02673115742846895	AFM
	GO:0051086	chaperone mediated protein folding independent of co	0.02673115742846895	CCT2
	GO:0071266	'de novo' L-methionine biosynthetic process	0.02673115742846895	BHMT
Cluster 2	GO:0110053	regulation of actin filament organization	0.02673115742846895	ESAM NRP1 RDX SERPINF2
	GO:0051173	positive regulation of nitrogen compound metabolic pr	0.02673115742846895	CAVIN4 CCT2 CCT6A DHX9 EIF6 F2 HMGB2 NME2 NRP1 RAP2C RBM14 RDX SERPINF2 SSBP1
	GO:0070198	protein localization to chromosome, telomeric region	0.02673115742846895	CCT2 CCT6A
	GO:0051240	positive regulation of multicellular organismal process	0.02673115742846895	APOA2 DHPS DHX9 F2 FGB FGG HMGB2 NME2 SERPINF2
	GO:0060695	negative regulation of cholesterol transporter activity	0.02673115742846895	APOA2
	GO:0060694	regulation of cholesterol transporter activity	0.02673115742846895	APOA2
	GO:0060627	regulation of vesicle-mediated transport	0.02673115742846895	AHSG APOA2 FGB FGG RDX
	GO:0090066	regulation of anatomical structure size	0.02673115742846895	ESAM FGB FGG RDX SERPINF2
	GO:0110114	negative regulation of lipid transporter activity	0.02673115742846895	APOA2
	GO:0033036	macromolecule localization	0.02673115742846895	ABCC2 AFM APOA2 CCT2 CCT6A DHX9 ESAM F2 FGB FGG HNRNPA2B1 LAMA5 RDX
	GO:0048584	positive regulation of response to stimulus	0.02673115742846895	C4A C8A DHX9 F2 FGB FGG HMGB2 LOC107131209 LOC528040 RBM14 SERPINF2
	GO:0031647	regulation of protein stability	0.02673115742846895	AFM APOA2 CCT2 CCT6A
	GO:0032057	negative regulation of translational initiation in respons	0.02673115742846895	EIF2S1
	GO:0032101	regulation of response to external stimulus	0.02673115742846895	AHSG F2 FGB FGG HMGB2 RBM14 SERPINF2
	GO:0032212	positive regulation of telomere maintenance via telome	0.02673115742846895	CCT2 CCT6A
	GO:0009893	positive regulation of metabolic process	0.02673115742846895	APOA2 CAVIN4 CCT2 CCT6A DHX9 EIF6 F2 HMGB2 LPGAT1 NME2 NRP1 RAP2C RBM14 RDX SERPINF2 SSBP1
	GO:0006082	organic acid metabolic process	0.02673115742846895	BHMT EIF6 HARS1 HIBADH HIBCH LOC530929 LPGAT1
	GO:0010604	positive regulation of macromolecule metabolic proces	0.02673115742846895	APOA2 CAVIN4 CCT2 CCT6A DHX9 EIF6 F2 HMGB2 NME2 NRP1 RAP2C RBM14 RDX SERPINF2 SSBP1
	GO:0032781	positive regulation of ATP-dependent activity	0.02673115742846895	DHX9 SSBP1
	GO:0032919	spermine acetylation	0.02673115742846895	SAT2
	GO:0015723	bilirubin transport	0.02673115742846895	ABCC2
	GO:0032920	putrescine acetylation	0.02673115742846895	SAT2
	GO:1905698	positive regulation of polysome binding	0.02673115742846895	DHX9
	GO:2000371	regulation of DNA topoisomerase (ATP-hydrolyzing) acti	0.02673115742846895	DHX9
	GO:0033043	regulation of organelle organization	0.02673115742846895	CCT2 CCT6A ESAM NRP1 RBM14 RDX SERPINF2 SSBP1
	GO:2000373	positive regulation of DNA topoisomerase (ATP-hydroly	0.02673115742846895	DHX9
	GO:0002253	activation of immune response	0.02673115742846895	C4A C8A LOC107131209 LOC528040 RBM14
	GO:0038190	VEGF-activated neuropilin signaling pathway	0.02673115742846895	NRP1
	GO:0039695	DNA-templated viral transcription	0.02673115742846895	DHX9
	GO:0046204	nor-spermidine metabolic process	0.02673115742846895	SAT2
	GO:0043152	induction of bacterial agglutination	0.02673115742846895	FGB
	GO:0044419	biological process involved in interspecies interaction b	0.02673115742846895	C4A C8A F2 FGB HMGB2 LOC107131209 LOC528040 NRP1 RBM14
	GO:0045740	positive regulation of DNA replication	0.02673115742846895	DHX9 SSBP1
	GO:0045834	positive regulation of lipid metabolic process	0.02673115742846895	APOA2 F2 LPGAT1
	GO:0010605	negative regulation of macromolecule metabolic proces	0.02674354510256594	APOA2 C4A DHX9 EIF2S1 EIF6 F2 HMGB2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0002449	lymphocyte mediated immunity	0.02674354510256594	C4A C8A LOC107131209 LOC528040
	GO:0030212	hyaluronan metabolic process	0.027436272344973513	ITIH1 ITIH2

GO:0002460	adaptive immune response based on somatic recombin	0.027694900936880172	C4A C8A LOC107131209 LOC528040
GO:0032728	positive regulation of interferon-beta production	0.02885154678177039	DHX9 HMGB2
GO:0002684	positive regulation of immune system process	0.02885154678177039	C4A C8A DHPS HMGB2 LOC107131209 LOC528040 RBM14
GO:0006952	defense response	0.029031387120579268	AHSG C4A C8A F2 FGB HMGB2 LOC107131209 LOC528040 RBM14
GO:2001236	regulation of extrinsic apoptotic signaling pathway	0.029929600248433854	FGB FGG HMGB2
GO:1904358	positive regulation of telomere maintenance via telome	0.029929600248433854	CCT2 CCT6A
GO:2000351	regulation of endothelial cell apoptotic process	0.029929600248433854	FGB FGG
GO:0045087	innate immune response	0.031006670652749407	C4A C8A HMGB2 LOC107131209 LOC528040 RBM14
GO:0043094	cellular metabolic compound salvage	0.03130888777376281	BHMT PNP
GO:2000249	regulation of actin cytoskeleton reorganization	0.03280867819712153	ESAM NRP1
GO:0019229	regulation of vasoconstriction	0.03280867819712153	FGB FGG
GO:0019216	regulation of lipid metabolic process	0.03298770583221967	APOA2 EIF6 F2 LPGAT1
GO:0051017	actin filament bundle assembly	0.03298770583221967	NRP1 RDX SERPINF2
GO:0003018	vascular process in circulatory system	0.03298770583221967	FGB FGG SERPINF2
GO:0017148	negative regulation of translation	0.03298770583221967	DHX9 EIF2S1 EIF6
GO:0006520	amino acid metabolic process	0.033133651232907256	BHMT HARS1 HIBADH HIBCH
GO:0042304	regulation of fatty acid biosynthetic process	0.033133651232907256	EIF6 LPGAT1
GO:0080134	regulation of response to stress	0.033133651232907256	AHSG DHX9 F2 FGB FGG HMGB2 RBM14 SERPINF2
GO:0072577	endothelial cell apoptotic process	0.033133651232907256	FGB FGG
GO:0061572	actin filament bundle organization	0.03357590693338177	NRP1 RDX SERPINF2
GO:0006403	RNA localization	0.03416423052197084	CCT2 DHX9 HNRNPA2B1
GO:0010638	positive regulation of organelle organization	0.034521662858483765	CCT2 CCT6A NRP1 SERPINF2 SSBP1
GO:1901135	carbohydrate derivative metabolic process	0.035566578750987225	ATP6V1A GANAB GNE ITIH1 ITIH2 NME2 PNP
GO:0032880	regulation of protein localization	0.03570225683802283	CCT2 CCT6A F2 FGB FGG RDX
GO:0061028	establishment of endothelial barrier	0.03570225683802283	RAP2C RDX
GO:0051130	positive regulation of cellular component organization	0.03570225683802283	AHSG APOA2 CCT2 CCT6A NRP1 SERPINF2 SSBP1
GO:0006955	immune response	0.03596455138647304	APOA2 C4A C8A F2 FGB HMGB2 LOC107131209 LOC528040 RBM14
GO:0007155	cell adhesion	0.03652252042206729	BCAM DHPS ESAM FGB FGG LAMA5 NRP1 RDX
GO:0090303	positive regulation of wound healing	0.03652252042206729	F2 SERPINF2
GO:0070527	platelet aggregation	0.03652252042206729	FGB FGG
GO:0030219	megakaryocyte differentiation	0.03652252042206729	EIF6 HMGB2
GO:0051128	regulation of cellular component organization	0.03708912909724616	AHSG APOA2 CCT2 CCT6A ESAM F2 NRP1 RBM14 RDX SERPINF2 SSBP1
GO:1900736	regulation of phospholipase C-activating G protein-coup	0.03715501907859519	F2
GO:0051496	positive regulation of stress fiber assembly	0.03715501907859519	NRP1 SERPINF2
GO:0051604	protein maturation	0.03715501907859519	DHPS FGB FGG SERPINF2
GO:0051493	regulation of cytoskeleton organization	0.03715501907859519	ESAM NRP1 RBM14 RDX SERPINF2
GO:1900738	positive regulation of phospholipase C-activating G prot	0.03715501907859519	F2
GO:0070374	positive regulation of ERK1 and ERK2 cascade	0.03715501907859519	FGB FGG SERPINF2
GO:1905776	positive regulation of DNA helicase activity	0.03715501907859519	SSBP1
GO:0032917	polyamine acetylation	0.03715501907859519	SAT2
GO:1905696	regulation of polysome binding	0.03715501907859519	DHX9
GO:0010912	positive regulation of isomerase activity	0.03715501907859519	DHX9
GO:0006427	histidyl-tRNA aminoacylation	0.03715501907859519	HARS1
GO:0060621	negative regulation of cholesterol import	0.03715501907859519	APOA2
GO:0006417	regulation of translation	0.03715501907859519	DHX9 EIF2S1 EIF3B EIF6
GO:0060620	regulation of cholesterol import	0.03715501907859519	APOA2
GO:0032102	negative regulation of response to external stimulus	0.03715501907859519	F2 FGB FGG SERPINF2
GO:0031328	positive regulation of cellular biosynthetic process	0.03715501907859519	CAVIN4 CCT2 CCT6A DHX9 EIF6 HMGB2 LPGAT1 NME2 RBM14 SERPINF2
GO:0051168	nuclear export	0.03715501907859519	DHX9 EIF6 HNRNPA2B1
GO:1905464	positive regulation of DNA duplex unwinding	0.03715501907859519	SSBP1
GO:0034113	heterotypic cell-cell adhesion	0.03715501907859519	FGB FGG
GO:1901566	organonitrogen compound biosynthetic process	0.03715501907859519	APOA2 ATP6V1A BHMT DHX9 EIF2S1 EIF3B EIF6 HARS1 NME2 PNP
GO:0008284	positive regulation of cell population proliferation	0.03715501907859519	DHPS DHX9 F2 HMGB2 NME2 SERPINF2
GO:0008612	peptidyl-lysine modification to peptidyl-hypusine	0.03715501907859519	DHPS
GO:0046340	diacylglycerol catabolic process	0.03715501907859519	APOA2
GO:0043207	response to external biotic stimulus	0.03715501907859519	C4A C8A F2 FGB HMGB2 LOC107131209 LOC528040 RBM14
GO:1903392	negative regulation of adherens junction organization	0.03715501907859519	RDX
GO:0042060	wound healing	0.03715501907859519	F2 FGB FGG SERPINF2
GO:0009892	negative regulation of metabolic process	0.03715501907859519	APOA2 C4A DHX9 EIF2S1 EIF6 F2 HMGB2 ITIH1 ITIH2 LAMTOR5 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
GO:0010911	regulation of isomerase activity	0.03715501907859519	DHX9
GO:0034249	negative regulation of amide metabolic process	0.03715501907859519	DHX9 EIF2S1 EIF6
GO:0051707	response to other organism	0.03715501907859519	C4A C8A F2 FGB HMGB2 LOC107131209 LOC528040 RBM14
GO:1904036	negative regulation of epithelial cell apoptotic process	0.03715501907859519	FGB FGG
GO:0032648	regulation of interferon-beta production	0.03727371256190624	DHX9 HMGB2
GO:1903036	positive regulation of response to wounding	0.03727371256190624	F2 SERPINF2
GO:0032608	interferon-beta production	0.03727371256190624	DHX9 HMGB2
GO:0050776	regulation of immune response	0.037951093285170906	C4A C8A HMGB2 LOC107131209 LOC528040 RBM14
GO:0098609	cell-cell adhesion	0.038068965346944976	DHPS ESAM FGB FGG LAMA5 RDX
GO:0032210	regulation of telomere maintenance via telomerase	0.03833112010905364	CCT2 CCT6A
GO:0043462	regulation of ATP-dependent activity	0.03833112010905364	DHX9 SSBP1
GO:0009607	response to biotic stimulus	0.03971016820981523	C4A C8A F2 FGB HMGB2 LOC107131209 LOC528040 RBM14
GO:0051239	regulation of multicellular organismal process	0.03982457964001827	AHSG APOA2 DHPS DHX9 EIF6 F2 FGB FGG HMGB2 NME2 NRP1 SERPINF2
GO:0032956	regulation of actin cytoskeleton organization	0.041348009676886247	ESAM NRP1 RDX SERPINF2
GO:0050766	positive regulation of phagocytosis	0.044592478118348905	AHSG APOA2
GO:0001755	neural crest cell migration	0.044592478118348905	LAMA5 NRP1
GO:0061077	chaperone-mediated protein folding	0.046053085014008274	CCT2 CCT6A
GO:0001885	endothelial cell development	0.046053085014008274	RAP2C RDX
GO:1904356	regulation of telomere maintenance via telomere lengtl	0.047407952261206164	CCT2 CCT6A
GO:0090497	mesenchymal cell migration	0.047407952261206164	LAMA5 NRP1
GO:0002376	immune system process	0.047407952261206164	APOA2 C4A C8A DHPS EIF6 F2 FGB HMGB2 LOC107131209 LOC528040 RBM14
GO:0002250	adaptive immune response	0.048314912837701986	C4A C8A LOC107131209 LOC528040
GO:0002034	maintenance of blood vessel diameter homeostasis by r	0.048314912837701986	SERPINF2
GO:0038189	neuropilin signaling pathway	0.048314912837701986	NRP1
GO:0110112	regulation of lipid transporter activity	0.048314912837701986	APOA2
GO:1904973	positive regulation of viral translation	0.048314912837701986	DHX9
GO:0042310	vasoconstriction	0.048314912837701986	FGB FGG
GO:1902903	regulation of supramolecular fiber organization	0.048314912837701986	ESAM NRP1 RDX SERPINF2
GO:0050434	positive regulation of viral transcription	0.048314912837701986	DHX9
GO:1904971	regulation of viral translation	0.048314912837701986	DHX9
GO:0032481	positive regulation of type I interferon production	0.048314912837701986	DHX9 HMGB2
GO:0006446	regulation of translational initiation	0.048314912837701986	EIF2S1 EIF3B
GO:0070945	neutrophil-mediated killing of gram-negative bacterium	0.048314912837701986	F2
GO:0090297	positive regulation of mitochondrial DNA replication	0.048314912837701986	SSBP1
GO:0003072	renal control of peripheral vascular resistance involved	0.048314912837701986	SERPINF2
GO:0090277	positive regulation of peptide hormone secretion	0.048314912837701986	FGB FGG
GO:0032233	positive regulation of actin filament bundle assembly	0.048314912837701986	NRP1 SERPINF2
GO:0002793	positive regulation of peptide secretion	0.04934209345687288	FGB FGG
GO:0006950	response to stress	0.04934209345687288	AHSG C4A C8A DHX9 EIF2S1 F2 FGB FGG HMGB2 LOC107131209 LOC528040 NRP1 RBM14 SERPINF2
GO:0032970	regulation of actin filament-based process	0.04975772570210023	ESAM NRP1 RDX SERPINF2

	GO_MF	Name	P-value (fdr)	Gene Name
	GO:0004030	aldehyde dehydrogenase [NAD(P)+] activity	0.00013582048573309 72	ADH4 ALDH1A1 ALDH2
	GO:0016491	oxidoreductase activity	0.00040983020794488 03	ADH4 ALDH1A1 ALDH2 DAO DECR1 GPX4 LOC788425 SORD
	GO:0016620	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor	0.00101136847046031 139	ADH4 ALDH1A1 ALDH2
	GO:0016903	oxidoreductase activity, acting on the aldehyde or oxo group of donors	0.00152473695651962 75	ADH4 ALDH1A1 ALDH2
	GO:0004596	peptide alpha-N-acetyltransferase activity	0.00316289023162034 4	NAA15 NAA16
	GO:0004029	aldehyde dehydrogenase (NAD+) activity	0.00316289023162034 4	ALDH1A1 ALDH2
	GO:0004033	aldo-keto reductase (NADP) activity	0.01344305911606824 2	ADH4 LOC788425
	GO:0004595	pantetheine-phosphate adenylyltransferase activity	0.01348147331274413	COASY
	GO:0046526	D-xylulose reductase activity	0.01348147331274413	SORD
	GO:0003939	L-iditol 2-dehydrogenase activity	0.01348147331274413	SORD
	GO:0016616	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	0.01348147331274413	ADH4 LOC788425 SORD
	GO:0018547	nitroglycerin reductase activity	0.01348147331274413	ALDH2
	GO:0106373	3-deoxyglucosone dehydrogenase activity	0.01348147331274413	ALDH1A1
	GO:0031320	hexitol dehydrogenase activity	0.01348147331274413	SORD
	GO:0016614	oxidoreductase activity, acting on CH-OH group of donors	0.01348147331274413	ADH4 LOC788425 SORD
	GO:0005483	soluble NSF attachment protein activity	0.01348147331274413	NAPG
	GO:0047833	D-sorbitol dehydrogenase (acceptor) activity	0.01348147331274413	SORD
	GO:0034212	peptide N-acetyltransferase activity	0.01348147331274413	NAA15 NAA16
	GO:0044877	protein-containing complex binding	0.01469166240456166 1	ARPC2 FGF1 KIF5B MYO1D NAA15 NAA16
	GO:0008967	phosphoglycolate phosphatase activity	0.01564468162161920 5	PGP
	GO:0008670	2,4-dienoyl-CoA reductase (NADPH) activity	0.01564468162161920 5	DECR1
	GO:0047066	phospholipid-hydroperoxide glutathione peroxidase activity	0.01564468162161920 5	GPX4
	GO:0004140	dephospho-CoA kinase activity	0.01564468162161920 5	COASY
	GO:0016663	oxidoreductase activity, acting on other nitrogenous compounds as donors, oxygen as acceptor	0.01564468162161920 5	ALDH2
	GO:0051903	S-(hydroxymethyl)glutathione dehydrogenase activity	0.01564468162161920 5	ADH4
	GO:0008080	N-acetyltransferase activity	0.01564468162161920 5	NAA15 NAA16
	GO:0051287	NAD binding	0.01564468162161920 5	ADH4 SORD
Cluster 1	GO:0019115	benzaldehyde dehydrogenase [NAD(P)+] activity	0.01564468162161920 5	ADH4
	GO:0035276	ethanol binding	0.01564468162161920 5	ADH4
	GO:0003884	D-amino-acid oxidase activity	0.01564468162161920 5	DAO
	GO:0043136	glycerol-3-phosphatase activity	0.01564468162161920 5	PGP
	GO:0005503	all-trans retinal binding	0.02202897913813863 7	ADH4
	GO:0003960	NADPH:quinone reductase activity	0.02202897913813863 7	ADH4
	GO:0016407	acetyltransferase activity	0.02257485754104741	NAA15 NAA16
	GO:0005198	structural molecule activity	0.02301635298333742 6	ARPC2 RPS15A RPS24 SNTB2
	GO:0000166	nucleotide binding	0.02381480630923699	ADH4 COASY DAO DECR1 DRG1 KIF5B MYO1D SORD
	GO:1901265	nucleoside phosphate binding	0.02381480630923699	ADH4 COASY DAO DECR1 DRG1 KIF5B MYO1D SORD
	GO:0043262	ADP phosphatase activity	0.02436107240322163 4	PGP
	GO:0016410	N-acyltransferase activity	0.02436107240322163 4	NAA15 NAA16
	GO:0043022	ribosome binding	0.02436107240322163 4	NAA15 NAA16
	GO:0042802	identical protein binding	0.02539232587263062	DAO DECR1 DRG1 GPX4 ISCA2 KIF5B SORD
	GO:0004024	alcohol dehydrogenase activity, zinc-dependent	0.02880615003452123 2	ADH4
	GO:0030955	potassium ion binding	0.03933433287702009 5	DRG1
	GO:0003774	cytoskeletal motor activity	0.04072920742656323	KIF5B MYO1D
	GO:0018455	alcohol dehydrogenase [NAD(P)+] activity	0.04199168523026034 6	ADH4
	GO:0004022	alcohol dehydrogenase (NAD+) activity	0.04199168523026034 6	ADH4
	GO:0003824	catalytic activity	0.04203432572363673	ADH4 ALDH1A1 ALDH2 COASY DAO DECR1 DRG1 GPX4 KIF5B LOC788425 NAA15 NAA16 PGP SORD
	GO:0003735	structural constituent of ribosome	0.04203432572363673	RPS15A RPS24
	GO:0036094	small molecule binding	0.04203432572363673	ADH4 COASY DAO DECR1 DRG1 KIF5B MYO1D SORD
	GO:0043021	ribonucleoprotein complex binding	0.04203432572363673	NAA15 NAA16
	GO:0016918	retinal binding	0.04203432572363673	ADH4
	GO:0017076	purine nucleotide binding	0.04203432572363673	ADH4 COASY DECR1 DRG1 KIF5B MYO1D SORD
	GO:0031420	alkali metal ion binding	0.04464938361200659	DRG1
	GO:0019841	retinol binding	0.04464938361200659	ADH4
	GO:0008131	primary amine oxidase activity	0.04649606407744376	DAO
	GO:0030554	adenyl nucleotide binding	0.04649606407744376	ADH4 COASY DECR1 KIF5B MYO1D SORD
	GO:0016661	oxidoreductase activity, acting on other nitrogenous compounds as donors	0.04649606407744376	ALDH2
	GO:0070402	NADPH binding	0.04981278359799661	DECR1
	GO:0004866	endopeptidase inhibitor activity	6.06229790813214 6e-06	AHSG C4A ITIH1 ITIH2 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0030414	peptidase inhibitor activity	6.06229790813214 6e-06	AHSG C4A ITIH1 ITIH2 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0061135	endopeptidase regulator activity	6.06229790813214 6e-06	AHSG C4A ITIH1 ITIH2 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0061134	peptidase regulator activity	1.38275332849228e-05	AHSG C4A ITIH1 ITIH2 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0004857	enzyme inhibitor activity	2.05311623124963 57e-05	AHSG APOA2 C4A ITIH1 ITIH2 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0140678	molecular function inhibitor activity	0.00170578439934561 81	AHSG APOA2 C4A ITIH1 ITIH2 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0140662	ATP-dependent protein folding chaperone	0.00756342778286085 7	CCT2 CCT6A CCT8
	GO:0004867	serine-type endopeptidase inhibitor activity	0.00872010097760543 4	ITIH1 ITIH2 SERPINA3-1 SERPINF2
	GO:0003743	translation initiation factor activity	0.00872010097760543 4	EIF2S1 EIF3B EIF6
	GO:0003713	transcription coactivator activity	0.00872010097760543 4	DHX9 HMGB2 NME2 RAP2C RBM14
	GO:0030234	enzyme regulator activity	0.00872010097760543 4	AHSG APOA2 C4A ITIH1 ITIH2 LOC107131209 LOC528040 NRP1 SERPINA3-1 SERPINF2
	GO:0050839	cell adhesion molecule binding	0.00872010097760543 4	BCAM ESAM FGB FGG LAMA5
	GO:0044183	protein folding chaperone	0.01324288639533491	CCT2 CCT6A CCT8
	GO:0140657	ATP-dependent activity	0.01814056877185049 3	ABCC2 ATP6V1A CCT2 CCT6A CCT8 DHX9 UBA7
	GO:0005539	glycosaminoglycan binding	0.01814056877185049 3	F2 ITIH1 ITIH2 NRP1
	GO:0051082	unfolded protein binding	0.02061307658533017 4	CCT2 CCT6A CCT8
	GO:0008135	translation factor activity, RNA binding	0.02061307658533017 4	EIF2S1 EIF3B EIF6
	GO:0043022	ribosome binding	0.02102925496175001 4	DHX9 EIF2S1 EIF6
	GO:0019782	ISG15 activating enzyme activity	0.02386256222824797 6	UBA7
	GO:0033919	glucan 1,3-alpha-glucosidase activity	0.02386256222824797 6	GANAB
	GO:0008442	3-hydroxyisobutyrate dehydrogenase activity	0.02386256222824797 6	HIBADH
	GO:0033679	3'-5' DNA/RNA helicase activity	0.02386256222824797 6	DHX9
	GO:0008431	vitamin E binding	0.02386256222824797 6	AFM
	GO:0001069	regulatory region RNA binding	0.02386256222824797 6	DHX9
	GO:0072555	17-beta-ketosteroid reductase activity	0.02386256222824797 6	DHRS11
	GO:0034038	deoxyhypusine synthase activity	0.02386256222824797 6	DHPS
	GO:0090079	translation regulator activity, nucleic acid binding	0.02386256222824797 6	EIF2S1 EIF3B EIF6
	GO:0003860	3-hydroxyisobutyryl-CoA hydrolase activity	0.02386256222824797 6	HIBCH
	GO:0005540	hyaluronic acid binding	0.02386256222824797 6	ITIH1 ITIH2
Cluster 2	GO:0015127	bilirubin transmembrane transporter activity	0.02386256222824797 6	ABCC2
	GO:0009384	N-acylmannosamine kinase activity	0.02386256222824797 6	GNE
	GO:0090600	alpha-1,3-glucosidase activity	0.02386256222824797 6	GANAB
	GO:0097367	carbohydrate derivative binding	0.02386256222824797 6	ABCC2 ATP6V1A CCT2 CCT6A CCT8 DHX9 F2 HARS1 ITIH1 ITIH2 NME2 NRP1 RAP2C
	GO:0008761	UDP-N-acetylglucosamine 2-epimerase activity	0.02386256222824797 6	GNE
	GO:0047150	betaine-homocysteine S-methyltransferase activity	0.02386256222824797 6	BHMT
	GO:0016887	ATP hydrolysis activity	0.02394141433227 3	CCT2 CCT6A CCT8 DHX9
	GO:0098772	molecular function regulator activity	0.02437157751829562	AHSG APOA2 C4A HMGB2 ITIH1 ITIH2 LOC107131209 LOC528040 NRP1 SERPINA3-1 SERPINF2
	GO:0098631	cell adhesion mediator activity	0.02506529679529604	BCAM ESAM
	GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	0.02647991324988483 2	GANAB GNE GUSB
	GO:0045182	translation regulator activity	0.02647991324988483 2	EIF2S1 EIF3B EIF6
	GO:0016712	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen	0.02804663747993814 8	LOC521656 LOC530929
	GO:0003712	transcription coregulator activity	0.02996922104347555	DHX9 HMGB2 NME2 RAP2C RBM14
	GO:0043021	ribonucleoprotein complex binding	0.03076889612032923	DHX9 EIF2S1 EIF6
	GO:1990518	single-stranded 3'-5' DNA helicase activity	0.03336153309814923	DHX9
	GO:0000253	3-keto sterol reductase activity	0.03336153309814923	DHRS11
	GO:0005055	laminin receptor activity	0.03336153309814923	BCAM
	GO:0004566	beta-glucuronidase activity	0.03336153309814923	GUSB
	GO:0004821	histidine-tRNA ligase activity	0.03336153309814923	HARS1
	GO:0045142	triplex DNA binding	0.03336153309814923	DHX9
	GO:0072582	17-beta-hydroxysteroid dehydrogenase (NADP+) activity	0.03336153309814923	DHRS11
	GO:0016798	hydrolase activity, acting on glycosyl bonds	0.03617288676369915 4	GANAB GNE GUSB
	GO:0003824	catalytic activity	0.03802234431930324	ATP6V1A BHMT CCT2 CCT6A CCT8 DHPS DHRS11 DHX9 F2 GANAB GNE GUSB HARS1 HIBADH HIBCH LOC521656 LOC530929 LPGAT1 NME2 NRP1 PNP RAP2C SAT2 UBA7
	GO:0070053	thrombospondin receptor activity	0.04543031187091215 6	F2
	GO:1905172	RISC complex binding	0.04543031187091215 6	DHX9
	GO:0070653	high-density lipoprotein particle receptor binding	0.04543031187091215 6	APOA2
	GO:0019003	GDP binding	0.04621298200123315	NME2 RAP2C
	GO:0016874	ligase activity	0.04751816797329457	ATP6V1A HARS1 UBA7

	GO_CC	Name	P-value (fdr)	Gene Name
Cluster 1	GO:0005829	cytosol	0.00034904996340886206	ADH4 ALDH1A1 DAO DCUN1D1 DECR1 DRG1 FGF1 GPX4 KIF5B NAA15 NAA16 RPS24 SORD VPS29
	GO:0031415	NatA complex	0.0010990753054557648	NAA15 NAA16
	GO:0005737	cytoplasm	0.0026519776696218326	ADH4 ALDH1A1 ARPC2 COASY DAO DCUN1D1 DECR1 DRG1 FGF1 GPX4 KIF5B LOC788425 NAA15 NAA16 NAPG NDUFB3 RPS24 SNTB2 SORD TMEM11 VPS29
	GO:0031414	N-terminal protein acetyltransferase complex	0.0026519776696218326	NAA15 NAA16
	GO:0005739	mitochondrion	0.005137137019081023	COASY DAO DECR1 GPX4 NAPG NDUFB3 SORD TMEM11
	GO:0005622	intracellular anatomical structure	0.010063615036456277	ADH4 ALDH1A1 ARPC2 COASY DAO DCUN1D1 DECR1 DRG1 FGF1 GPX4 KIF5B LOC788425 MYO1D NAA15 NAA16 NAPG NDUFB3 RPS15A RPS24 SNTB2 SORD TMEM11 VPS29
	GO:0036195	muscle cell projection membrane	0.04215428948922547	ARPC2
	GO:0031975	envelope	0.04215428948922547	DAO GPX4 NDUFB3 SORD TMEM11
	GO:0031966	mitochondrial membrane	0.04215428948922547	DAO NDUFB3 SORD TMEM11
	GO:0036194	muscle cell projection	0.04215428948922547	ARPC2
	GO:0031967	organelle envelope	0.04215428948922547	DAO GPX4 NDUFB3 SORD TMEM11
	GO:0030906	retromer, cargo-selective complex	0.04507588225886602	VPS29
	GO:0005740	mitochondrial envelope	0.04565873090053588	DAO NDUFB3 SORD TMEM11
Cluster 2	GO:0005615	extracellular space	3.843095786384986e-06	AFM AHSG APOA2 BHMT C4A C8A F2 FGB FGG HMGB2 LAMA5 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0005576	extracellular region	6.094612105018686e-06	AFM AHSG APOA2 BHMT C4A C8A F2 FGB FGG HMGB2 ITIH1 ITIH2 LAMA5 LOC107131209 LOC528040 SERPINA3-1 SERPINF2
	GO:0005577	fibrinogen complex	1.4701081106306992e-05	FGB FGG SERPINF2
	GO:0005832	chaperonin-containing T-complex	0.00015586843704312015	CCT2 CCT6A CCT8
	GO:0101031	protein folding chaperone complex	0.002096216597698602	CCT2 CCT6A CCT8
	GO:0030054	cell junction	0.005388416792474428	ABCC2 C4A EIF2S1 EIF3B ESAM FGB LAMA5 LOC107131209 NRP1 RAP2C RDX
	GO:0070161	anchoring junction	0.01048382061536991	ABCC2 C4A ESAM LOC107131209 NRP1 RAP2C RDX
	GO:0031091	platelet alpha granule	0.011065056888211246	FGB FGG

	KEGG Pathway	Enrichment FDR	Genes
Cluster 1	Metabolic pathways	0.00344757978389607	ALDH1A1 ALDH2 PGP NDUFB3 DAO SORD ADH4 COASY GPX4
	Pantothenate and CoA biosynthesis	0.00825909038127396	ALDH2 COASY
	Fatty acid degradation	0.0175395320676454	ALDH2 ADH4
	Pyruvate metabolism	0.0175395320676454	ALDH2 ADH4
	Arginine and proline metabolism	0.0191995224588003	ALDH2 DAO
	Glycolysis / Gluconeogenesis	0.0240667651914018	ALDH2 ADH4
	Retinol metabolism	0.0245616199502928	ALDH1A1 ADH4
	Endocytosis	0.0245616199502928	KIF5B VPS29 ARPC2
Cluster 2	Complement and coagulation cascades	9.0630648130245e-08	C8A FGG F2 SERPINF2 FGB C4A
	Coronavirus disease	8.27825174347648e-05	C8A FGG F2 NRP1 FGB C4A
	Metabolic pathways	0.0066187071551755	GUSB HIBADH GNE BHMT ATP6V1A SAT2 HIBCH PNP GANAB NME2
	Staphylococcus aureus infection	0.0365307502241675	FGG C4A
	Linoleic acid metabolism	0.0484192277720815	LOC521656, LOC530929
	Platelet activation	0.0484192277720815	FGG F2 FGB
	Systemic lupus erythematosus	0.0484192277720815	C8A C4A

Supplementary Table 5: List of proteins selected by MB-PLS regression and correlations with phenotypes

List of proteins from the three tissues selected by *mixomics* ' function *block.spls* to explain the best the 7 phenotypes. Sparsity mode was used to select 30 proteins by tissue for each of the 3 components. Finally, we kept only proteins implicated in at least 20% of a component according to loadings' outputs of *block.spls* . Proteins are defined with Uniprot ID without "_BOVIN" termination and corresponding genes are indicated if known. Pearson correlation values is given for proteins and phenotypes highly associated to the same component (phenotypes implicated in at least 35% of the component).

Tissue	Uniprot	Corresponding_gene	Correlation_to_phenotype						
			Carcass_AT	Carcass_muscle	Carcass_bone	Total_nb_perirenal _adipocytes	Fixed_lipids	Fixed_proteins	Fixed_energy
Liver	TCPQ	CCT8	-0,863	0,855			-0,911	0,74	-0,856
Muscle	RS24	RPS24	0,901	-0,915			0,849	-0,671	0,803
Fat	A0A3Q1MME4	HNRNPA2B1	-0,791	0,772			-0,891	0,642	-0,861
Muscle	HIBCH	HIBCH	-0,781	0,775			-0,829	0,462	-0,84
Liver	F6QN89	SNTB2	0,819	-0,813			0,836	-0,657	0,792
Fat	F1N6I4	LOC788425	0,751	-0,732			0,87	-0,702	0,819
Liver	A0A3Q1LNW7;F1N5J8	DECR1	0,778	-0,737			0,855	-0,563	0,842
Muscle	A0A3Q1M9Y4;A0A3Q1MN02; A0A3Q1NFQ2;E1BMF6;F1N4V5	NAA15;NAA16	0,709	-0,665			0,822	-0,226	0,902
Muscle	DRG1	DRG1	0,693	-0,718			0,73	-0,691	0,657
Fat	DHX9	DHX9	-0,756	0,732			-0,829	0,728	-0,763
Liver	PGP	PGP	0,752	-0,703			0,811	-0,444	0,825
Fat	A0A3Q1MG04;F1MAV0	FGB	-0,852	0,856			-0,779	0,661	-0,724
Fat	RIDA	RIDA	0,774	-0,734			0,796	-0,633	0,752
Fat	ITIH1	ITIH1	-0,838	0,85			-0,81	0,601	-0,777
Fat	Q2KJA9	ESAM	-0,753	0,751			-0,826	0,662	-0,779
Muscle	BCAM	BCAM	-0,707	0,707			-0,687	0,658	-0,617
Muscle	GPX4	GPX4	0,644	-0,625			0,69	-0,62	0,632
Fat	RBM14	RBM14	-0,736	0,718			-0,85	0,562	-0,836
Liver	GLU2B	PRKCSH	-0,759	0,735			-0,755	0,648	-0,699
Liver	RPF2	RPF2			0,624	0,416		-0,433	
Liver	A0A3Q1LLS4;F1MXS1	RANBP2			-0,612	-0,369		0,592	
Fat	Q3ZC85	PLVAP			-0,62	-0,442		0,32	
Muscle	H31;H32;H33;H3C	H3-3A;H3-5			0,657	0,352		-0,289	
Muscle	K1C10	KRT10			-0,501	-0,293		0,581	
Muscle	F6S1Q0	KRT18			-0,618	-0,122		0,59	
Liver	A5D7C4	TNPO3			0,593	0,362		-0,442	
Fat	Q3ZBY8	GMDS			0,532	0,699		-0,133	
Fat	E1BCE0	DNAJC8			-0,474	-0,487		0,469	
Fat	Q3T0U5	CHMP2A			0,46	0,484		-0,394	
Fat	SRP19	SRP19			-0,575	-0,214		0,564	
Fat	RL17	RPL17			0,578	0,351		-0,282	
Liver	PPOX	PPOX			0,577	0,35		-0,458	
Muscle	H4				0,685	0,357		-0,21	
Muscle	G3N0V2	KRT1			-0,54	-0,144		0,615	
Muscle	ITIH5	ITIH5			-0,41	-0,83		0,0366	
Muscle	F1N1A3	UBQLN1			0,728	0,236		-0,322	
Fat	Q17QP9	PSMD14			0,555	0,305		-0,333	
Liver	A6QQN2	PTPN1			0,477	0,337		-0,451	
Muscle	MEP50	WDR77			0,407	0,553		-0,341	
Fat	E1BMU6	AKAP1			0,475	0,542		-0,228	
Muscle	A7MBI8	NUDT9			0,488	0,195		-0,585	
Liver	STAR5	STARD5			0,487	0,554		-0,265	
Fat	CT027					-0,732			
Liver	A0A3Q1MT24	DMGDH				-0,602			
Muscle	CX7A2	COX7A2				-0,698			
Muscle	GMPR2	GMPR2				0,604			
Liver	RM12	MRPL12				-0,576			
Fat	Q08DW1	RAB9A				-0,345			
Fat	COQ6	COQ6				0,644			
Muscle	RSMB;RSMN	SNRPB;SNRPN				-0,718			
Liver	PRP6	PRPF6				-0,638			
Fat	TBCB	TBCB				-0,642			
Fat	A5D7P1	RPE				0,309			
Muscle	A0A3Q1M8I4;A0A3Q1NLI5; A5D9H5;A6H6Y0;F1N2T0	HNRNPD;HNRNPDL				-0,513			
Liver	HEM1	ALAS1				-0,589			
Muscle	NIF3L	NIF3L1				0,417			
Muscle	QOR	CRYZ				0,479			
Liver	NDUAB	NDUFA11				-0,607			
Liver	OMA1	OMA1				-0,872			
Fat	A0A3Q1LRN7;F1MM13	H6PD				-0,338			
Liver	A0A3Q1M941	LOC100300896				0,578			
Liver	C3V9V7	TMED2				0,577			
Muscle	LAMP1	LAMP1				0,496			
Fat	MAL2	MAL2				-0,678			
Muscle	NP1L1	NAP1L1				-0,661			
Liver	HDHD2	HDHD2				0,717			
Fat	DDRGK	DDRGK1				0,289			
Muscle	E1BJB1;G3X7R7	TUBB2A				-0,439			
Liver	A0A3Q1LMV9	EPPK1				-0,627			
Liver	A0A3Q1MGB8;F1MMK9	KIF12				-0,55			