UNIVERSITÀ DEGLI STUDI DI TRENTO

Department of Mathematics

Ph.D. in Mathematics

XXX CYCLE

The influence of the inclusion of biological knowledge in statistical methods to integrate multi-omics data

Supervisor:
Prof. Corrado Priami
Co-supervisors:
Dr. Marie Pier Scott-Boyer
Dr. Luca Marchetti

Ph.D. student:
Giulia Tini

È passato del tempo, sembra sia stato un lampo,
eravamo ragazzi e ora eccoci qua
con le crisi del caso e gli occhiali sul naso
e un'idea più realista di felicità.
"Sbagliato", Lorenzo Jovanotti Cherubini

Table of contents

List of figures

List of tables

Preface

This thesis summarizes my research in the field of multi-omics data integration as a Ph.D. student at the Department of Mathematics of the University of Trento and at The Microsoft Research - University of Trento Centre for Computational and Systems Biology (COSBI).

Multi-omics data integration is a multi-disciplinary field, which revolutionised Bioinformatics in the last decades. Its aim is analysing and interpreting information coming from multiple molecular layers, namely genomics, epigenomics, transcriptomics, proteomics, metabolomics and microbiomics, comprehensively called "omics". The accumulation of these data, generated with advanced high-throughput techniques, provides information on important biological processes. For instance, multi-omics diagnostics is nowadays considered crucial for the improvement of medical and healthcare services and for the implementation of preventive and precision medicine. The biological and medical relevance of studying multiple molecular layers together is proved by the elevate number of articles which contain the text "multi omics" in their abstract. From the PubMed repository for biomedical literature, more than 1000 papers referred to "multi omics" in the last ten years, with more than 300 of them also linked, for instance, to "cancer". However, the huge quantity of available omics data must be analysed with comprehensive systems: statistical and computational techniques are essential tools to obtain prognostic, diagnostic, and therapeutic information [248]. Additionally, mathematical approaches have been developed in the last years to simultaneously model information based on not evident/not yet studied inter-omics interactions. Those methods have proven to be powerful to combine omics data, although they do not use prior information on omics relationships.

For this reason, this thesis mainly focuses on the interplay of mathematical tools and prior knowledge about omics interactions, which are essential in the development of the multi-omics integration field and its applications. Specifically, the goal of the work described here is to show that, at least on simulated datasets, statistical multi-omics integration has better performances when both known and unknown inter-omics interactions are included in the analysis. Since prior knowledge of inter-omics interactions is not always available, we will compute it from the data with the use of multivariate statistics and networks. This thesis

deals in particular with three-omics data integration but the methodologies developed here can be easily applied to an higher number of omics.
The thesis is organized in chapters which explore increasingly complex approaches to multi-omics data integration. Specifically, I will present: a three-omics integration where inter-omics interactions are known, a study on completely unsupervised simultaneous integration and, finally, a model where known and unknown interactions are considered together. Although the biological questions addressed in the chapters of the thesis are different, it is important to consider that the biological phenomena are connected: increased knowledge about molecular mechanisms, for example, can lead to more accurate separation of patients/samples with similar medical responses. Chapters 2, 3 and 4 detail projects I contributed to during the last years and that resulted in publication or manuscripts in preparation.

Chapter 1: we provide here an overview of the main hypothesis (linear and simultaneous) of biological interactions between omics data. Additionally, this Chapter presents the state of the art statistical approaches developed to solve multi-omics integration.

Chapter 2: we test here the importance of considering known linear inter-omics relations. We combine DNA methylation, gene expression and protein levels to study adipogenesis, the process of creation and growth of adipocytes. Prior knowledge about biomolecule connections reduce the amount of data to be analysed. This leads to the observation of coordinated changes at the epigenomic, transcriptomic and proteomic levels, showing how information flows from one molecular layer to the others. Although this biological linear integration provides new insights in the adipogenic process, we consider only known omics interactions to perform it. The content of this Chapter was developed in collaboration with the Nestlé Institute of Health Science and the U.S. Food and Drug Administration, and is part of a paper under preparation:

- G. Tini, V. Varma, R. Lombardo, G. Lefebvre, P. Descombes, S. Métairon, C. Priami, J. Kaput, M.P. Scott-Boyer, "DNA methylation during human adipogenesis and the impact of fructose", preprint.

Chapter 3: here we move to a simultaneous perspective of multi-omics integration. We focus on unsupervised methodologies, which are able to model also unknown interactions between different data types. To understand more about this approach, we compare different unsupervised methodologies on sample classification problem. We consider both real and simulated datasets. We explore the impact of several factors that are shown to affect simultaneous integration results (method choice, presence of noise, number of integrated data, applied pre-processing). The content of this Chapter has been published in the journal *Briefings in Bioinformatics*:

- G. Tini, L. Marchetti, C. Priami, M.P. Scott-Boyer, "Multi-omics integration - a comparison of unsupervised clustering methodologies", *Briefings in Bioinformatics*, November 2017.

Chapter 4: we present here a pipeline whose aim is to combine the strengths of the simultaneous and the linear approaches, in order to solve sample classification. While the use of prior knowledge is useful to focus on known biological interactions, simultaneous methods allow the inclusion in the analysis of unknown omics relations that would be missed otherwise. The accuracy of simultaneous multi-omics data integration is improved, on simulated datasets, by the prioritization of important features. This step is performed with a multivariate linear regression method followed by the creation of a prior-knowledge network. The results obtained in this Chapter indicate that the inclusion of prior knowledge increases the power of integration methodologies.

- This work is the result of my internship at the Nestlé Institute of Health Science in Lausanne, Switzerland.

Chapter 5: in this Chapter we summarize the main results of the thesis and discuss possible future directions.

Chapter 1

Introductıon

In this Chapter we introduce the problem of multi-omics data integration. We first provide a brief overview of the different types of omics data, with a more detailed focus on DNA methylation, type of data discussed in Chapter 2. We present the different hypotheses of inter-omics interplay that will be used throughout the thesis, together with the main approaches to multi-omics data integration that are based on those hypotheses. We finally provide the state-of-the-art models that have been developed to solve the main questions and challenges of multi-omics data integration.

1.1 The "Omics Revolution"

The addition of the suffix "omics" to a molecular term connotes the comprehensive assessment of a set of molecules (like genes or proteins) [73] which contain part of the information related to the biological system under study. In their work, Hasin *et. al* well summarized the different omics data that can be assessed for the same experiment [73], namely genomics, epigenomics, transcriptomics, proteomics, metabolomics and microbiomics (see Table 1.1 for a more detailed description of those data).

In early studies, the omics collected in Table 1.1 have been investigated in isolation to look for their association with a phenotypic trait of interest. This approach, however, does not consider the interactions between different biomolecules, acknowledged by the central dogma of molecular biology. This theory, proposed by Francis Crick in 1970 [39] describes the flow of genetic information from DNA to RNA to proteins that occurs in a biological system and leads to the determination of cellular phenotypes (Figure 1.1).

Table 1.1 Omics data types that can be collected from the same experiment. For each omics the correspondent molecular levels and the description of what the omics assess are provided. Additionally, the platforms used to measure the molecular levels are added. NGS: Next Generation Sequencing; MS: Mass Spectrometry; RRBS: Reduced Representation Bisulfite Sequencing.

Omics	**Molecular level**	**Description**	**Platforms**

Genomics	Genetic variants	Identification of genetic variants associated with the phenotype	genotype arrays [221];
			NGS [217];
			exome sequencing [158]

Epigenomics	DNA methylation; histone acetylation	Genome-wide assessment of reversible modification of DNA which do not vary nucleic basis sequence	NGS [217] RRBS [136]
Transcriptomics	mRNA; miRNA; long non-coding RNA	Genome-wide qualitative (which transcripts are present) and quantitative (levels of expression) assessment of RNA levels	NGS [217], probe-based arrays [187]
Proteomics	Proteins	Quantification of peptide abundances and modifications	MS [54]
Metabolomics	Metabolites	Quantification of multiple molecules (e.g. amino acids, fatty acids, carbohydrates)	MS [54]
Microbiomics	Microbiota	Collective investigation of microorganim (bacteria, viruses) colonizing skin, mucosal surfaces, gut	NGS for metagenomics quantification [75]

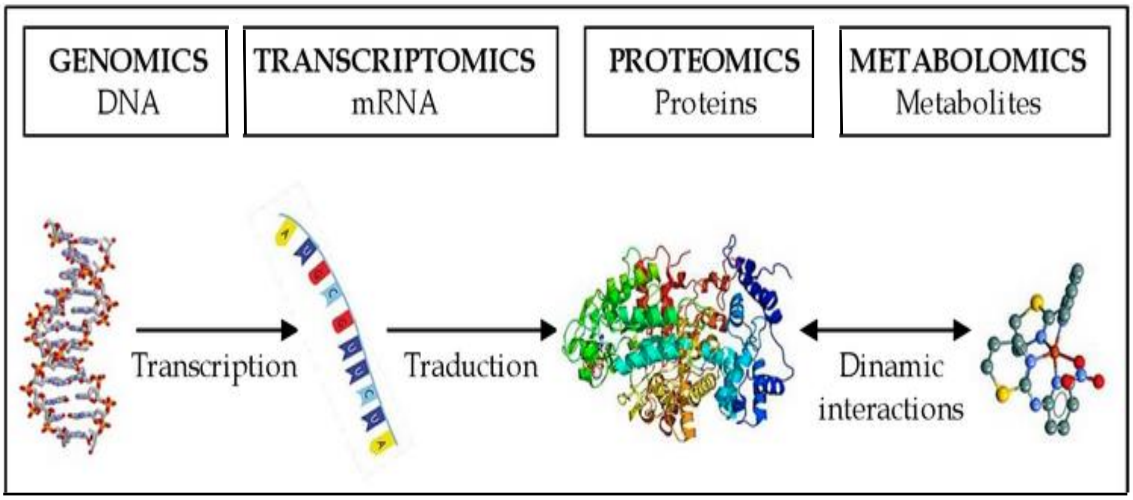


Fig. 1.1 Genetic information flow as described by the central dogma of molecular biology (from [151]). Instructions on DNA are transcribed into messenger RNA (mRNA) and then translated into proteins. Metabolites are the end-product of biological processes, obtained by dynamical interactions with proteins.

The central dogma of molecular biology can be thought as the first theoretical step towards multi-omics data integration. However, the importance of this field in Bioinformatics has increased only in the last decades, with the sequencing if entire human genome in 2001, with the Human Genome Project [82].

Improvements in high-throughput techniques (*e.g.* Next Generation Sequencing [217], Mass Spectrometry [54]) also permitted cost-efficient access to measurements of multiple molecular levels. These technological improvements, coupled with the availability of several repositories of accessible data, such as The Cancer Genome Atlas Project (TCGA) [204] or The Encyclopedia of DNA Elements Project (ENCODE) [51], lead to what can be called "Omics Revolution" [91]. However, there is now a need for adequate analytical methods namely improving multi-omics integration [156] to unleash the full potential of this ever-increasing amount of massive data.

1.2 Linear and simultaneous integration: challenging the central dogma of biology

The growing availability of data describing complex traits has challenged the central dogma of molecular biology. The alternative theories that were proposed gave new perspectives to the multi-omics integration field.

Deeper investigation of biochemical processes, allowed by the advent of recent technologies, has uncovered molecular activities which were not taken into account by Crick. Examples are: reverse transcription (copying of DNA into RNA [21, 22]), post-transcription RNA processing [77] and post-translation protein modification (*e.g.* protein methylation [109], cis- and trans-splicing [182]).

On the basis of those discoveries, in 2009, James Shapiro re-formulated the central dogma by stating, for example, that the flow of information from one molecular layer to the others is not unidirectional. Shapiro also stated that every element of the genome interacts directly or indirectly with many other genomics components [190]. Few years later, in 2012, Denis Noble proposed a theory of no privileged level of causation in biological systems, that is it cannot be assumed that an organism is completely defined by its genome alone [160]. Starting from the example of the cardiac rhythm, Noble showed that feedback cycles among molecular layers not only exist, but are necessary parts of biological processes.

Following the hypotheses of omics interplay provided by the central dogma of molecular biology and by its alternative theories, two main approaches to multi-omics data integration can be distinguished: linear or simultaneous integration [175]. In Figure 1.2 we provide a graphical representation of the two different hypotheses of interplay: the hypothesis of linear interaction is referred as "Hypothesis A", while "Hypothesis B" describes the simultaneous interaction of omics.

1.2.1 Linear multi-omics integration

The first approach to multi-omics data integration that we will discuss in this thesis, assumes linear and hierarchical interactions between omics data (based on the Hypothesis A, Figure 1.2). Following the central dogma of molecular biology (Figure 1.1), variations in the DNA lead to gene expression changes, which in turn are responsible of level of protein production and thus of different phenotypes appearance [39]. Although, as pointed out before, this view of biology has proved to be oversimplified, the linear integration ensures to obtain biological meaningful results, since it is based on already known biological processes.

Multivariate regression and variable selection methods are common tools used to mathemati-

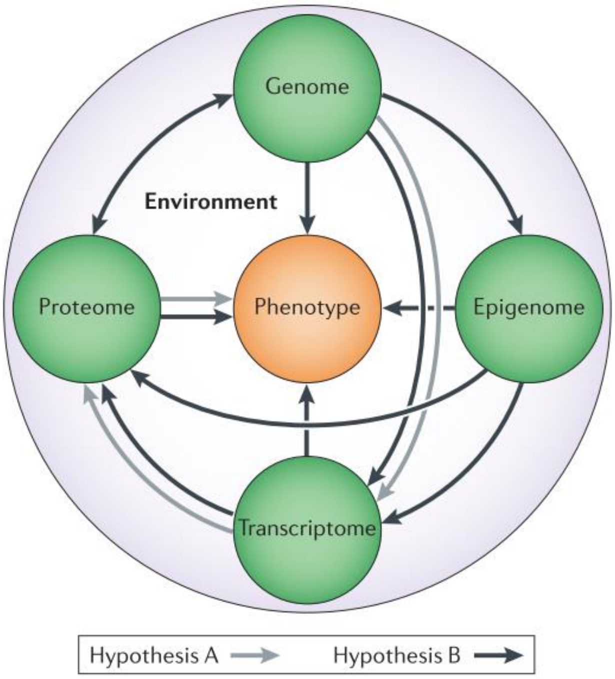


Fig. 1.2 Alternative hypotheses of interactions between molecular layers (from [175]). Hypothesis A (grey arrows) suggests a linear view of phenotype emergency. Hypothesis B (black arrows) assumes that the phenotype is given by simultaneous changes in the omics data.

cally solve this type of integration. This is especially true when only two data are considered: those models can predict cause-and-effect links between different data types. For example, in the work by Tapp *et al.* [202], methods such as Partial Least Squares (PLS) [132, 241], Least absolute shrinkage and selection operator (Lasso) [206] and Elastic net [254], were used to predict the concentration of proteins related to obesity. Protein concentrations were then integrated with the hepatic transcriptome, in order to elucidate the molecular mechanisms associated to adiposity and inflammation in high-fat fed mice.

1.2.2 Simultaneous multi-omics integration

Linear integration has the limitation that unknown inter-omics relationships are not considered: only assessed directions of interaction are explored. Moreover, in general, only part of the total inter-omics interactions are known from literature.

The complexity of the phenotypes suggests that phenotypic traits can be more exhaustively explored by the combination of simultaneous changes across all omics data. This view of inter-omics interactions, described by Hypothesis B in Figure 1.2, is also at the basis of alternative theories of the central dogma.

Multi-omics data integration based on Hypothesis B does not need prior knowledge to be included. Additionally, following this approach several omics data (i.e. three or more) can be combined at the same time, providing a more complete and realistic view of the problem at hand.

As pointed out by Huang *et. al* [79], the inter-omics interactions are major concerns for data integration strategies: this is why the more recent progresses in multi-omics data integration focus on the simultaneous approach, with the help of computational and mathematical tools.

1.3 Statistical multi-omics data integration

The field of multi-omics integration is developing from a linear to a simultaneous point of view. Statistical methodologies are improving to include more than two omics data types and to take into account inter-omics relationships.

In this section, we give an overview of the classical statistical methodologies that are at the basis of simultaneous integration techniques. Those methodologies were usually proposed to meet the requirements of pairwise omics integration. This implies the importance of linear integration, based on biological knowledge, in the improvement of the multi-omics data integration field.

Another important aspect to consider is whether information about the phenotype is used during the integration. With this in mind, statistical methodologies can be divided into either supervised or unsupervised approaches.

Supervised integration approaches consider the phenotype labels of samples (*e.g.* disease/normal, control/treatment) [79] and use this information to discover genotype-phenotype interactions [230] and to learn something more about the studied biological process.

Conversely, unsupervised integration approaches aim at drawing an inference from the considered omics data without having access to the labels of the response variables [79].

From the mathematical point of view, the k omics data measured from the same experiment are considered as matrices $X_1, \dots, X_k$ of dimension $n \times p_k$. n represents the number of common subjects while p_i $i = 1 \dots k$ is the number of biological features collected for omics X_i.

Latent variable factorization and networks are mathematical tools often used as a starting point for the developments of integrative methods.

1.3.1 Latent variable factorization

Latent variable factorization focuses on projecting variations occurring across different biological layers in a low-dimensional space. This goal is obtained by factorization of the matrices $X_1, \dots, X_k$ into the product of loadings F_i (of dimension $n \times r$) and factors Q_i (of dimension $p_i \times r$ and called latent variables) added to an error term:

$$X_i = F_i Q_i^T + E \quad i = 1 \dots k \tag{1.1}$$

The number of columns of F_i and Q_i, $r < p_i$, represents the number of latent components used to build the new low-dimensional space. Starting from this model, different methods can be developed accordingly to the inter-omics relationship that are searched. How these interactions are computed and which constraints should be imposed have to be considered in order to build a model.
Among all the possible methodologies developed starting from the model described in Equation 1.1, Partial Least Squares and Canonical Correlation Analysis have been largely used and ameliorated along the years to respond to new questions.

Partial Least Squares

Partial Least Squares (PLS) [107, 241] is a standard regression technique. It is used to identify a small set of features working as predictors for the response dataset [11] and strongly associated with them. Following Equation 1.1, two omics data, X and Y can be decomposed as:

$$\begin{aligned} X &= F_x Q_x^T + E \\ Y &= F_y Q_y^T + E \end{aligned} \tag{1.2}$$

For each latent component $j = 1 \dots r$, PLS finds the loading factors (q_x^j and q_y^j, respectively columns of Q_x and Q_y) maximizing the covariance between F_x and F_y. This is done by solving the equivalent problem:

$$\max_{||q_x^j||=1, ||q_y^j||=1} cov(X_{j-1} q_x^j, Y q_y^j) \quad j = 1 \dots r \tag{1.3}$$

where X_{j-1} is the residual matrix for each component [107]. To respond to the new challenges brought by technological advances, the method has been extended in 2012 to integrate more than two omics data by Li *et al.*. They proposed Multi-block PLS [114], which implies that different layers jointly contributes to a unique dataset used as a response.

Additionally, to focus only on important biomolecules and discard the others, sparsity has been introduced in the PLS model, for example adding to the maximization problem a Lasso penalty term [206], defined on the vector $x = (x_1 \dots x_n)$ as $P_L(x) = \sum |x_i|$.

Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) [67, 76] is another standard method to inspect interactions occurring between two data types. Differently to the PLS approach, for CCA it is not necessary to define which of the integrated omics data contains response variables. This makes the method more suitable for a total unsupervised integration.

Similarly to PLS, CCA searches for linear combinations of features. To find loading factors (F_x and F_y, from Equation 1.1), for each latent component $j = 1 \dots r$, Canonical Correlation Analysis maximizes the correlation among Xq_x^j and Yq_y^j, by solving

$$\max_{||q_x^j||=1, ||q_y^j||=1} corr(Xq_x^j, Yq_y^j) \quad j = 1 \dots r \tag{1.4}$$

The dimension of the maximization problem can be reduced also in the Canonical Correlation Analysis case, by applying regularization and penalization terms, such as the Lasso or the Elastic net ($P_E(x) = \sum |x_i| + \sum |x_i|^2$), to create sparse solutions.

To account for information coming from more than two omics data, in 2009 Witten and Tibshirani extended the sparse CCA version [239], by proposing Multiple Canonical Correlation Analysis: a detailed description of this method is given in Appendix A.

1.3.2 Network-based methods

Networks represent a powerful tool in the context of multi-omics data integration, since they are able to contain heterogeneous and high-dimensional information.

A network G is defined by the couple $G = (V, E)$, where $V = (v_1, \dots v_n)$ is a set of nodes while $E = \{e_{ij}, \; i, j = 1 \dots n\}$, is a set of edges connecting nodes in V, where e_{ij} represents the connection between nodes v_i and v_j. A weight w_{ij} can be associated to the edge e_{ij} to describe the importance of the link.

Networks can characterize complex interactions, thus identifying mechanism linked to different types of information [79] and associated to the phenotype of interest.

Networks can also be employed as a source of prior knowledge, describing biological processes and functions [230], helping in data interpretation. To this extent, databases of annotated pathways, such as KEGG [92] or Reactome [40], as well as gene networks or protein-protein interaction networks (PPI), have been collected to store inter-omics information obtained from literature. Omics datasets can be mapped to those databases to study over-representation or enrichment of molecules coming from different molecular layers.

Alternatively, networks can be generated on the basis of the omics data at hand. In this case, network-based approaches to multi-omics data integration take advantage of algorithms from graph theory, such as diffusion processes [42, 229] or sub-network extraction [115, 149]. According to what nodes and edges represent, as well as the type of inter-omics interactions searched, edges can be built in different ways [231]. Bayesian networks, for example, allow the use of informative priors to capture conditional dependencies between probabilistic events [167]: probabilities are used to define the relationships between nodes [6]. Text-mining approaches instead, build networks based on scientific publications, on the assumption that molecules likely to interact share contextual information [56, 215]. Finally, correlation networks such as those generated by the Weighted Correlation Network Analysis method (WGCNA) [103] are based on the correlation (or on significance of correlation) between nodes.

Bersanelli *et al.* in their review of statistical integration methods [11], distinguish two kinds of network formalism (heterogeneous and multiplex) to describe multiple layers of biological information and their interactions.

Given k omics data $X_1 \dots X_k$, heterogeneous networks consider k types of nodes, with each of them corresponding to a different omics. Edges are built to represent intra and inter-layers connections. This allows to extract information about the problem at hand from the same unique graph. An example of the use of heterogeneous network in multi-omics data integration is provided by Li and Patra [116]. They propose a random walk with restart algorithm to connect a gene network and a phenotype one: edges between genes and phenotypes represent the probability of the gene to be relevant for the phenotype. The gene network is built considering the PPI data from the Human Protein Reference Database [168], while genes-phenotype relations are obtained from the OMIM database [70]. Another approach based on heterogeneous networks is Lemon-Tree [16], which finds modules of co-expressed genes before combining them with one type of regulator data (such as methylation or miRNA) to infer regulatory scores, on the basis on decision trees assigned to regulators.

Multiplex networks [141] are instead defined as k different networks which store biological information on the same set of vertices (*e.g.* the samples/patients under study). In this case, an edge between two nodes in one of the networks represents the strength of intra-omics association among the two samples, such as their correlation or their similarity. To combine knowledge form the k data types, inter-omics links can be built. For instance, it can be determined, through degree correlation, if a hub in one of the considered layers has the same role in others [144]. Another example of use of multiplex networks in multi-omics data integration is provided by the algorithm developed by Wang *et al.* [229], called Similarity Network Fusion. Through an information diffusion-based strategy (described in details in Appendix A) the algorithm retrieves the strongest and most informative sample similarities across different omics.

1.4 Application of multi-omics integration to real data

On the basis of the two main approaches (latent variable factorization and networks) described in the previous sections, statistical methods have been proposed overcome the challenges of multi-omics data integration and to solve real biological questions.

1.4.1 Multi-omics data integration challenges

The main challenges of multi-omics data integration are data heterogeneity and the complexity of the inter/intra-omics interactions. Data heterogeneity refers to measurements from different platforms that are usually not taken on the same scale or have different distributions. Statistical methods should thus ensure that results are not biased towards the omics with larger dimension or larger variance. This issue can be overcome by scaling data or by reducing them to the most informative ones by means of feature selection approaches [74]. Inter-omics interactions and co-variations should be revealed without discarding relevant single omics patterns (intra-omics changes). Biologically significant results can be both supported by weaker signals involving more omics or strongly induced by a single data type. Thus multi-omics integration algorithms should be able to model either the significant intra and inter-omics relationships, to provide a complete view of the problem at hand.
The relevance of these challenges is intensified when integration methods are extended to support more than two omics data, cases that will be considered through this dissertation. Despite combining more biological levels implies that a more complete picture of the biological system under study is drawn [11], it also increases the amount of noise added to the model, intensifying false positive discovery and difficulties in interpretation.

Dealing with common and complementary information

One of the goals of multi-omics data integration is to reinforce the common signal coming from different platforms (Table 1.1), such as genomics and transcriptomics [62, 183], miRNA and transcriptomics [120, 233], transcriptomics and proteomics [202, 208], or proteomics and metabolomics [13, 152]. Several studies are available for the integration of these omics types. For example Wang *et al.* [233] integrated miRNAs and gene expression by meaningfully associating with Bayesian model analysis networks between the two omics and clinical outcomes in glioblastoma. Blanchet *et al.* [13] combined rat metabolomics and proteomics to divide samples at the onset of Experimental Autoimmune Encephalomyelitis, peripheral inflamed and healthy ones. A latent variable approach, Extended Canonical Variate Analysis [162], was separately applied to the data to reduce their dimension and extract relevant proteins and metabolites, which were then merged in a unique matrix and analysed through Principal Component Analysis (PCA) [85].

Strengthening common signal, however, is not the sole multi-omics data integration goal, since some levels like transcriptomics and metabolomics, do not interact directly. Their integration, however, could potentially bring complementary information [31, 36, 225]. Conesa *et al.* [36] integrated mice mRNAs and metabolites, measured on multiple time points and treatments, by N-PLS [19]. This method associates omics with a regression model on the latent space with maximum covariance between data and is an extension of the PLS algorithm to support structures with more factors (*e.g.*, different time points). Results were compared to those obtained from transcriptomics data with Tucker3 [194], an algorithm able to decompose multi-factorial data and study within-block relationships. This study revealed that, despite the high overlap of selected genes, those found by integration described the response to treatments more closely. Furthermore, N-PLS integration showed differences in time response between genes and metabolites, suggesting that this algorithm can be used for time course experimental design (*e.g.*, Dynamic Time Warping [30]).

1.4.2 Biological questions approached by multi-omics integration

The common or complementary information gained by multi-omics data integration can be used to solve several types of biological questions including the analysis of molecular mechanisms, sample classification and biomarker identification.

Exploring multilevel molecular mechanisms

To better understand molecular mechanisms underlying complex traits, interactions between biomolecules from different platforms and biological pathways are sought, especially by means of networks [32, 63, 102, 170, 209, 232]. In a study about breast cancer by Wang *et al.* [232], three sets of genes coming from pairwise integration of gene expression with one among somatic mutation, DNA copy number and DNA methylation were used to build a cross-talk network of risk pathways by random walk with restart [100] on a human protein interaction network. Networks were also built by the tool 3Omics [102] to unveil connections between transcriptomics, proteomics and metabolomics data under different experimental conditions. In such a case, the edge creation between biomolecules and the new node addition when information was missing were performed by correlation analysis and literature text-mining.

Sample classification

One of the biological questions more addressed by multi-omics data integration is to recognize, or correctly classify, different subtypes of the phenotype under study. For example, in studies concerning the effect of a treatment, responding subjects can be separated from those not responding to treatment. When dealing with diseases, it would be interesting to distinguish patients from healthy control subjects, for instance, to improve diagnosis and disease prognosis; or, in a more complex situation, to classify different subtypes of the same disease (like cancer). Because of the diversity in tumour types and the availability of patients, sample classification has been widely applied to oncology [118, 139, 147, 191, 252]. A method developed to solve this question is iCluster [191] which, after an initial estimation based on an optimized K-means clustering [250], computes integrated clusters with a likelihood-based solution of a joint Gaussian latent variable model. Applied to DNA copy number, methylation and gene expression, iCluster recognized three glioblastoma subtypes (*i.e.*, Proneural, Classical and Mesenchymal) [192].

Biomarker identification

Another important goal of multi-omics data integration is the identification of biomolecules characterizing a phenotype. Omics integration could be more effective than single omics analysis [16, 33, 97, 220, 225]: the interactions between biomolecules from different omics data cannot be modelled by separated analysis, leading to fragmented and incomplete information [231]. In a study on human metabolic disorders [225], metabolites and gene expression were clustered by means of WGCNA [103], which computes modules of highly correlated features through topological measures on correlation networks. The identified metabolite and gene clusters were then associated to external phenotypes such as variations of body weight, which revealed their connection with obesity and mitochondrial dysfunction.

1.5 DNA methylation

This section will provide an overview of epigenetics and DNA methylation main characteristics, as they will be the main focus of Chapter 2.

Epigenetics refers to any DNA modification that does not change its sequence of nucleic basis but regulates its transcription [84, 90, 198]. Although not directly discussed in the central dogma of molecular biology, epigenomics, which refers to the genome-wide distribution of epigenetic changes (Table 1.1), always played an important role in biology. In the last decades, great efforts have been made to characterize the epigenome, its regulation and its changes during the development of cells in normal and disease state [87]. In particular, it is well known that epigenetic is essential for developmental process and cellular differentiation, but can also occur randomly in mature cells, following environmental exposure [84].
The best known epigenetic process is perhaps DNA methylation. DNA methylation (see Figure 1.3 for a schematic representation) is the addition by DNA methyltransferases (DNMTs) of a methyl group (CH_3) to the 5 position of the cytosine in CpG dinucleotides (CpG sites), thus forming 5-methylcytosine [90].

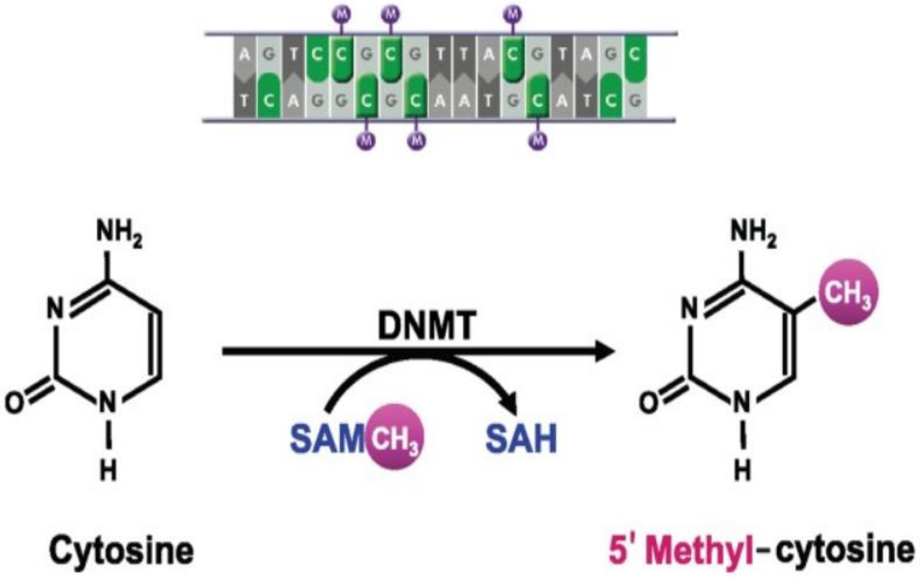

M
A G T C C G C G T T A C G T A G C
T C A G G C G C A A T G C A T C G
NH2
N
O
N
H
DNMT
SAM CH3
SAH
CH3
Cytosine
5' Methyl-cytosine

Fig. 1.3 Schematic representation of DNA methylation, from [249]: methylation usually occurs when cytosine is followed by guanine (CpG site). Cytosine is methylated by DNA methyltransferase (DNMT) through the addition of a methyl group CH_3, donated by S-adenosylmethionine (SAM).

The methyl group is donated by S-adenosylmethionine (SAM), which after the methyl transfer reaction forms S-adenosyl homocysteine (SAH), a potent DNMT inhibitor [249]. Methylation usually occurs in genomic regions with a high content of guanine and cytosine and rich of CpG dinucleotides, which are called CpG islands. The importance of DNA methylation is emphasized by the growing number of human diseases that occur when

methylation is not properly established or maintained [177]. For example, in cancer cells, methylation of CpG islands is known to contribute to silence tumor suppressor genes, while during carcinogenesis a genome-wide hypomethylation (low methylation levels) can be seen across the genome [84].
Methylation in CpG enriched regions provides regulatory mechanisms of gene expression: its effect on gene expression depends on where it occurs. High methylation levels (hypermethylation) in CpG islands associated with promoters normally repress gene transcription [112], although an increasing number of exceptions are identified [197]. Methylation within intronic and exonic regions of a gene body is instead positively correlated with expression [89]. Epigenetic changes, like DNA methylation, are thought to be involved in aging process and can be mediated by environmental factors, such as exposure to pollutants or diet and lifestyle. For example, methylation of CpG islands associated with estrogen receptor, undetectable in young individuals becomes progressively detectable with age [83]. Similarly, hypermethylation has been linked with atherosclerosis [235]. Also diet is known to affect variations in methylation levels: supplements of folate and vitamins affect the activity of the enzymes producing the methyl group [46], while methyl deficient diet induces liver cancer through hypomethylation and consequent higher expression of oncogenes [227]. Moreover, investigation on maternal diet in mice proved that feeding pregnant females with methyl donors induces changes in the offspring phenotypes (*e.g.* coat colour) [242]. Interestingly, such DNA modifications are for the most part reversible and can thus be modulated by optimizing, for instance, environments and daily habits. It is thus important to study modification in the DNA methylation connected with changes at other molecular layers, as well as to better study the impact of factors such as nutrition.

The hypotheses of linear and simultaneous omics interplay described in Figure 1.2 are tested in this thesis to take into account the challenges and the biological problem described above. We focus on differences and strengths of performing linear supervised and simultaneous unsupervised omics integration. More than two data types are always considered. The two methodologies extract different kind of information from the data: respectively, information based only on prior knowledge and that based on not evident/not yet studied interactions. Under these considerations, we first want to test the effect of considering only known inter-omics relationships. We want also to test which factors influence unsupervised three-omics integration and to prove that network-based methods are the best in recovering information from unknown interactions. Simultaneous integration is known to be a powerful statistical tool to combine omics data, but it is usually considered in its unsupervised version. In this thesis, we thus want to include prior knowledge to simultaneous methods, to improve their performances. In particular, since prior knowledge of interaction is not always available, we focus on computing it from the data with the help of multivariate statistics.

Chapter 2

A linear supervised three-omics integration study on human adipogenesis

In this Chapter we study the molecular mechanisms underlying adipogenesis, the biological process of creation and growth of adipocytes. Following the Hypothesis A (Figure 1.2), we perform a linear and supervised three-omics integration. We aim of the study is to inspect the role of DNA methylation (see Table 1.1 and Section 1.5) in the regulation gene expression during adipogenesis. DNA methylation levels, gene expression, and protein abundances were measured from adipocytes cultured along 16 days and with the addition of different doses of fructose. These omics are integrated with classical statistical tools. In this thesis framework, the novelty of this integration consists in the use of prior knowledge to reduce the amount of data to focus on. Indeed, integration is here performed by considering only genomic regions that encode differentially expressed genes (already known from a previous work [157]).
This step assures to focus on methylated DNA regions that are likely to affect gene expression. The result is a list of genes changing both at transcriptomic and epigenomic level. Those genes are involved in different biological processes linked to adipogenesis that were not found when only the transcriptomic level was analysed [157].
Additionally, some of the methylated/expressed genes change coordinately also at the proteomic level. This result highlights the importance of considering available inter-omics interactions when biomarkers are searched for the studied biological process.
The content of this chapter is the result of a collaboration with Nestlé Institute of Health Science (NIHS) and the U.S. Food and Drug Administration (FDA) and is included in a paper in preparation.

2.1 Introduction

Obesity and its comorbidities are growing worldwide epidemics [255]. A major hallmark of modern, westernised nutrition is increased consumption of highly refined sugar [154] that has been associated to the increasing incidence of metabolic disorders [94, 145]. Fructose consumption, as part of sugar-sweetened beverages and other processed foods, affects gut- and adipocyte-secreted hormones as well as the innate immune system [255]. Additionally, fructose causes deregulation of metabolic pathways in the hypothalamus and adipose tissue, both involved in mediation and regulation of the homeostatic maintenance of host energy balance resulting in the promotion of the development of metabolic syndrome [24].

Adipose tissue stores excess energy in the form of triglyceride through i) an increase of adipocyte size (hypertrophy) and ii) the promotion of differentiation or adipogenesis of pre-existing adipocytes (hyperplasia) [29]. Obesity occurs as consequence of a chronic positive energy intake which brings hypertrophy to a plateau resulting in the promotion of hyperplasia to cope with unbalance energy intakes [29].

Adipocyte differentiation mechanisms are regulated by a complex network of transcription factors responsible for expression of key proteins that induce mature adipocytes [157]. The main regulators of the early adipocyte differentiation process are the peroxisome proliferator-activated receptor (*PPAR*) and the CCAAT/enhancer binding proteins (*C/EBPs)* [111]. In the later stages, the process is regulated through the fatty acid binding protein 4 (*FABP4*), adiponectin, and fatty acid synthase (*FAS*) [201]. The understanding of adipogenesis was previously expanded and enriched by analyzing systems-wide transcriptomic profiles at specific time points from progenitor to mature adipocyte using a novel analytical tool for biological network activity, the Network Activity Score Finder (NASFinder) [157]. NASFinder identified high scoring networks in signaling pathways, transcription factors, metabolic, energy production, and membrane and cell structure functions that change simultaneously across the differentiation process.

Epigenetic modifications such as DNA methylation contribute to the control of gene expression and therefore participate in regulating these processes. For example, the promoter of lipoprotein lipase (*LPL*) that is expressed during adipose stem cell differentiation in culture [124] contains a hypomethylated *PPAR* responsive element [161]. Activation of *LPL* ultimately leads to induction of *FABP4* whose promoter harbors a *C/EBP*α site [142, 210]. Additionally, the promoter of *PPAR*γ2 was hypermethylated in 3T3-L1 mice preadipocytes and was progressively demethylated following the induction of differentiation with a concomitant increase of expression of its mRNA [58].

Methylation in CpG islands provides regulatory mechanisms of gene expression and is essential for cell differentiation and tissue integrity [4]. The effect of methylation on gene expression depends on where methylation occurs: high methylation levels in promoters normally repress gene transcription [112], while methylation within intronic and exonic regions of a gene body is positively correlated with expression [89]. Moreover, DNA methylation can be affected by environmental factors such as lifestyle and diet, particularly since choline, betaine, folate, riboflavin and vitamin B12 participate in the one carbon cycle that produces S-adenosylmethionine, the methyl donor [25].

Differential methylation has been associated with chronic diseases associated with improper diets such as obesity [4, 222] and increased BMI [48]. Changes in methylation of *FASN* in rats is associated with liver steatosis [37]. In addition, a relationship between methylation of genes involved in the circadian clock system and obesity, metabolic syndrome, and weight loss has been shown [143].

The controversial link between increased consumption of fructose in human diets and the obesity epidemic [18] stimulated research that tested the detrimental impact of this carbohydrate on insulin resistance and adipocyte differentiation, key processes to maintain metabolic health [105, 126]. The role of DNA methylation status in fructose-induced metabolic syndrome and DNA methylation status has not been well characterized. Consumption of high fructose has been shown to induce DNA methylation in *PPAR*α and *CPT1A* in rat liver [164], leading to reduced expression of these genes and then to a hepatic lipid accumulation. Fructose may alter adipocyte differentiation by increasing levels of *PPAR*γ, *C/EBP*α, and *FABP4*, at least in murine cells in culture [50]. Both fructose and glucose are substrates utilized to increase adiposity, but fructose contributes more to weight gain in humans [200].

Stable isotope tracer methods were used to show that fructose was metabolized to glutamate and fatty acids [218, 219] and it diverts glucose metabolites to the serine oxidative pathway producing additional metabolic energy [219]. To further characterize the effects of fructose on adipocyte biology, genome-wide transcriptomic and DNA methylation data were analyzed at multiple time-points during differentiation of the human Simpson-Golabi-Behmel Syndrome (SGBS) euploid progenitor cells.

We identified genomic regions of differentially expressed genes where methylation of CpG sites differed compared to undifferentiated adipocytes. The integrative DNA modification, transcriptomic, and proteomic analysis reported here revealed that adipocyte methylation changes are influenced by time and the state of differentiation, with the largest differences detected at 384 hours following the initiation of differentiation, when the adipocytes are in the fully differentiated state. Furthermore, a 1:2 ratio of fructose (2.5mM) to glucose (5mM) demonstrated the most altered DNA methylation patterns among the different concentrations of fructose examined. An overview of all the performed single omics analysis and of the consequent omics integration is represented in Figure 2.1.

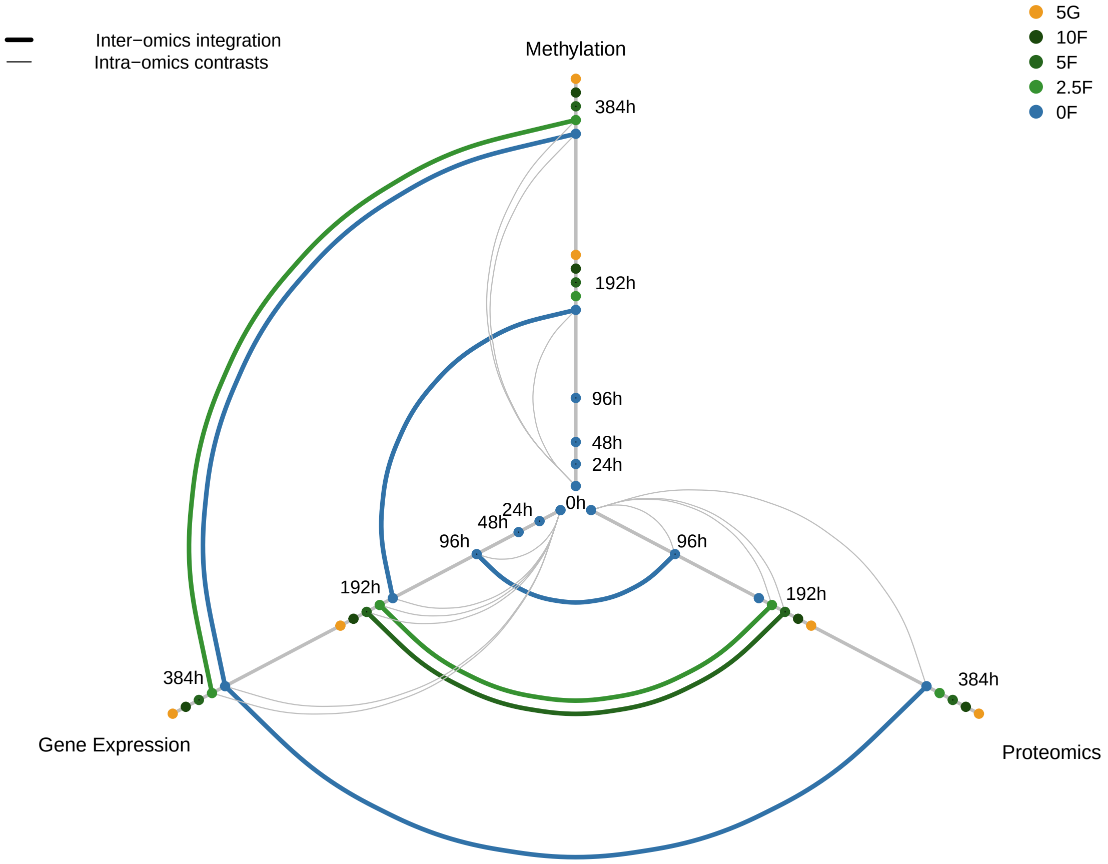


Fig. 2.1 Overview of the integration analysis of the three available omics data: methylation, gene expression and protein abundances. For each omics data, coloured dots indicate the different fructose/glucose doses available (0F, 2.5F, 5F, 10F, 5G) at each time point. To represent the significant integration among two different omics, for each fructose/glucose dose, a coloured thick line is drawn to connect the corresponding axes. The contrasts within the same omics providing the results are displayed as grey lines. Integration of gene expression and methylation resulted significant at 192 and 384 hours for control condition (0F) and for 2.5mM of fructose. Concordant integration results with protein abundances were found only at 384 hours in control condition.

2.2 Methods

2.2.1 Study design

Human Simpson-Golabi-Behmel syndrome (SGBS) preadipocytes, kindly provided by Martin Wabitsch, were used and cultured as described in [218]. The study design is summarized in Table 2.1. Triplicates of cells in culture were harvested for DNA methylation or transcriptomics assays at specific time points including 24, 48, 96, 192 and 384 hours for control adipocytes (six replicates were harvested at 0 hours). Cultures treated with different concentrations of fructose: 2.5, 5 and 10mM were harvested at 192 and 384 hours following the initiation of differentiation, as described in section 2.2.2. Importantly, samples for RNA and DNA platform were measured on cells plated at the same time and treated similarly.

For proteomic Somascan assays, cell lysates were obtained from 0, 96, 192, 384 hours following the induction of differentiation grown in 0, 2.5, 5, 10mM fructose.

Table 2.1 DNA methylation and gene expression were assayed at 6 different time points for control adipocytes, proteomics only at 4 time points (day 1 and day 2 excluded). DNA methylation, gene expression and proteins changes in the fructose-treated adipocytes were examined at 192 and 384 hours, following the addition of three different doses of fructose (2.5mM, 5mM, 10mM).

Hours	**Glucose (mM)**	**Fructose (mM)**	**Methylation**	**Gene Expression**	**Proteomics**
0	5	0	✓	✓	✓
24 (1 day)	5	0	✓	✓	–
48 (2 days)	5	0	✓	✓	–
96 (4 days)	5	0	✓	✓	✓
192 (8 days)	5	0, 2.5, 5, 10	✓	✓	✓
384 (16 days)	5	0, 2.5, 5, 10	✓	✓	✓

2.2.2 Fructose treatment of SGBS cells

For gene expression and DNA methylation studies, SGBS preadipocytes were plated at 2×10^5 cells in 100 mm dishes, supplemented with 10 ml growth medium, grown to confluence and initiated to differentiate as per [218]. All media used for the growth, differentiation and maintenance of adipocytes contained a basal amount of 5 mM glucose, equivalent to the normal blood glucose concentration. Cells were fully differentiated by day 8 (by oil red O staining, not shown). Cells for RNA and DNA isolations were collected at different time points across differentiation at 24, 48, 96, 192 and 384 hours. Fructose doses were based on reports found in the systemic circulation following exposure to fructose-rich food [80]. At the initiation of differentiation, 2.5, 5 and 10 mM fructose concentrations were added and maintained in the medium until the collection of cells and medium at end points of either day 8 (192 hours) or until day 16 (384 hours) of differentiation. Cell lysates from the control or fructose-treated adipocytes were collected for DNA and RNA isolations.

For RNA isolation, media was completely aspirated from cells and a total of 700 ul of QIAzol lysis reagent was added to the cells, and the lysed cells were scraped, collected in an Eppendorf vial, sheer disrupted by passing through a tuberculin syringe about 6 time, and the lysates flash frozen.

For obtaining samples for DNA isolation, media was removed and cells were washed with PBS and aspirated to remove all PBS. The cells were gently scraped in the presence of a total of 400 ul of PBS, collected using a pipette fitted with a wide mouth tip, transferred to an eppendorf vial, and flash frozen. Cells from triplicate wells were used for both RNA and DNA isolation respectively.

2.2.3 Adipocytes omics data analysis

DNA methylation data

Genome-wide methylation was assessed using Illumina Infinium HumanMethylation450k array platform (Illumina, San Diego, CA. USA) that contains a total of 485,512 CpG sites. Samples were distributed over four different BeadChips. CpG sites were then annotated with the R package ilmn12.hg.19 [72], which identified the gene and the region on the chromosome. Illumina GenomeStudio software was used to extract the raw signal intensities. Methylation data preprocessing was performed with the function preprocessIllumina from the R package Minfi [7]. This method was applied to reduce Infinium I/II type bias and correct for background. Absolute percentages of methylation (β-values) were then extracted and normalized with SWAN method [128].For each CpG site, averaged βvalues across the cell triplicates were considered for the following analysis.

Minfi package was used to detect differentially methylated positions (DMPs). Statistical significance of CpG sites was assessed with a moderated F-statistic implemented in the function dmpFinder. Since DMPs were used as starting point for further analysis, a loose FDR adjusted p-value threshold of 0.1 was chosen.

In addition to DMPs, differentially methylated regions (DMRs) were identified with R package COHCAP [234]. COHCAP functions take as input a list of annotated DMPs to compute average signals from CpG sites in the same region. A t-test with a FDR threshold of 0.05 was applied to find DMRs. The minimum number of sites needed to create a region was set at 1.

Transcriptomic data

The transcriptomic dataset was generated with 4 Illumina Human HT-12 v-4 BeadChips (Ilumina, Inc., San Diego, CA) hybridized with the RNA from 46 cell-cultures at different time points (0, 24, 48, 96, 192, 384 hours) and for different fructose concentration (0, 2.5mM, 5mM, 10mM). The RNA labeling and microarray hybridization was performed according to the manufacturer's recommendations.

The scanned data was acquired in R using the package illiminaio [195]. The non-normalized summarized bead-level data was then annotated with R package illuminaHumanv4.db [52]. Other labeling and analysis methods were performed with the preprocessing pipeline previously described in [157].

Differential expression analysis was carried out using the limma [196] R package. The probes were ranked by their log-odds scores given by empirical Bayesian moderation of sample variances with an FDR threshold of 0.01. The DEGs of fully differentiated adipocytes at 384h in controls were further processed to identify clusters of co-expressed genes. The clusters were decomposed according to the functional categories of their genes related to biological functions and pathways (DEG modules). The details of the procedures are described in [157]

Proteomic data

Regariding proteomics data, SGBS preadipocytes were plated at 1×10^5 cells/well in a 6-well plate and allowed to reach near confluence before adding differentiation medium. Samples were harvested from three replicate wells at 4 different time points including day 0 just before induction of differentiation, and then 96, 192 and 384 hours after the induction of differentiation.

The spent culture medium (supernatant) from respective wells was pipetted into an eppendorf vial, centrifuged at 13,000 RPM for 10 minutes at 4°C to pellet the cell debris. The supernatant was transferred to a fresh vial and stored at -80°C until used. Cells were then washed three times with ice cold PBS and then lysed by the addition of 125 μl Mammalian Protein Extraction Reagent M-PER® (Pierce biotechnology cat # 78503) containing supplemented with halt protease inhibitors (with EDTA) Pierce biotechnology cat # 87786) at 1x concentration and incubated for 5 minutes. Cell lysates were scraped and transferred to a microcentrifuge tube, centrifuged at 13,000 RPM for 10 minutes at 4°C to pellet the cell

debris. The clarified supernatant (lysate) obtained was transferred to a fresh tube and stored at -80°C. Protein concentrations in the supernatant and cell lysates were estimated using the micro BCA kit (Pierce biotechnology cat # 23235) as per the recommended protocol.
Cells lysates in triplicates were analysed with the SOMAscan platform (SomaLogic, Inc., Boulder, CO) consisting of 1,12909 aptamers at different time-points (0, 96, 192 and 384 hours) and for different doses of fructose (0, 2.5, 5 and 10mM). SomaLogic Inc. (Boulder, CO) performed all proteomic assessments and was blinded to the clinical characteristics of participants in this study. Samples were analysed as previously described [20, 64, 65, 165]. Differentially expressed proteins were found with robust linear model from R package limma [196]. A threshold of 0.05 on moderated empirical Bayesian FDR was set to select significant proteins.

2.2.4 Multi-omics data integration

Genome-wide methylation, transcriptomic and proteomic data were collected and analyzed as described above. Differentially methylated regions (DMRs), differentially expressed genes (DEGs) and significantly expressed proteins were searched for each time point and each different fructose concentration.
Integrative analysis of methylation and gene expression was then performed by determining the DMRs that were associated with differentially expressed genes, for each of the considered analysis. These genes show significant changes in both methylation and gene expression. The location of the DMRs in the corresponding gene (promoter, exon, intron or intergenic) were annotated with genomation R package [2]. Figure 2.2 illustrates the different steps used to perform the integration of these two omics data types.
Integration with protein expression was performed by determining the methylated/expressed genes that also associated with significantly expressed proteins. These genes show significant changes in methylation, gene and protein expression.

Transcription Factors binding sites analysis

Binding sites of transcription factors (TFs) in DMRs were identified with the function get.enriched.motif of the R package ELMER [245, 246]. Binding sites were searched on 181 transcription factors identified in the ENCODE database [1]. The CpG sites of all DMPs were used as background. Motifs occurring at least 10 times and with an odd ratio higher than 1 in the 95% CI were considered significant. Significant motifs from the same family were summarized with the function motif.relevant.TFs data from the ELMER.data package [247].

[1] http://amp.pharm.mssm.edu/Harmonizome/dataset/ENCODE+Transcription+Factor+Targets

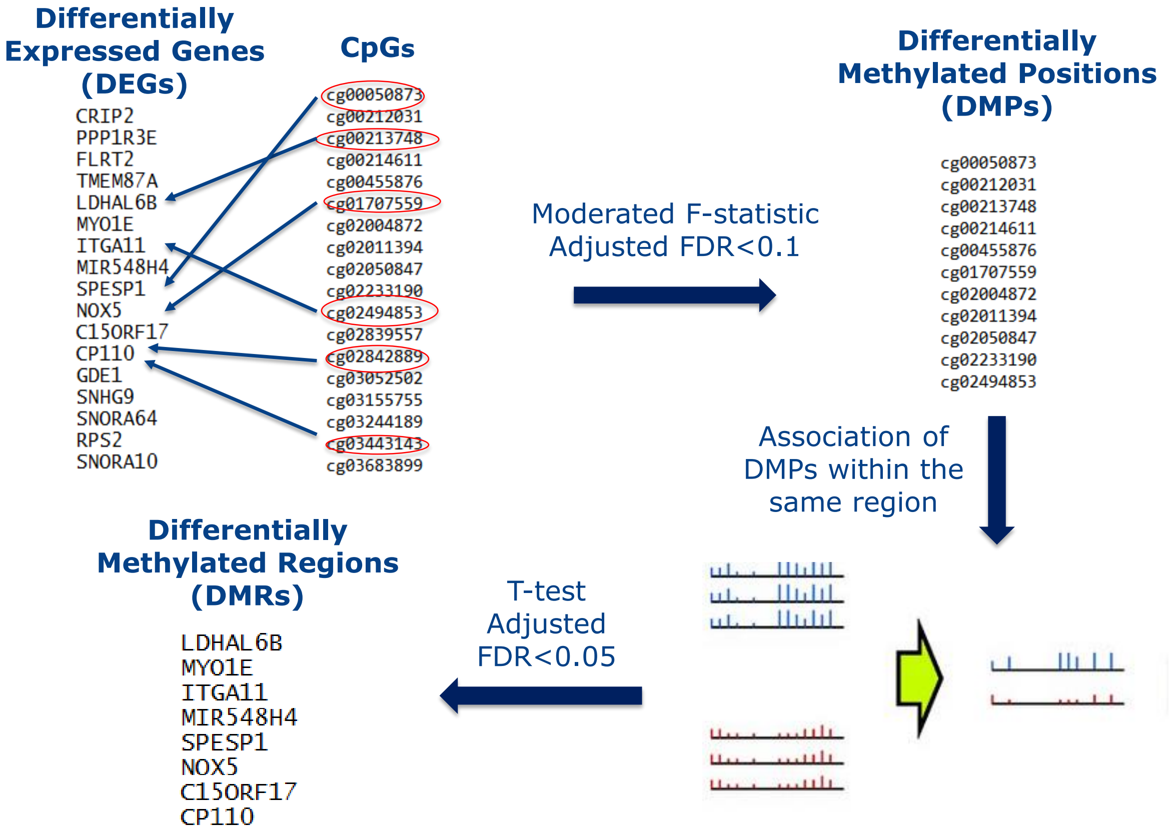


Fig. 2.2 Pipeline used to perform integration of DNA methylation and gene expression for each of the considered contrast. Starting from a given list of genes known to be differentially expressed (DEGs), CpG sites close to them are retrieved. DMPs are searched through the usage of F-test. The signal of DMPs found in the same genomic region are averaged, and a t-test is used to found differentially methylated genomic regions (DMRs). Those correspond to a sublist of the starting list f genes.

Pathway analysis

Pathways analysis was performed with NASFinder [157]. NASFinder identifies and scores statistically significant sub-networks of an interactome network connecting functionally related genes to its main regulator (e.g. receptors or transcription factors). The analysis described here were adipose-specific using transcription factors as regulators and transcripts that mapped to differentially methylated genes to find the most active pathways influenced by methylation. The p-value threshold used to select significant pathway was < 0.05. Functional pathway enrichment analysis was also performed with DAVID [47], using default parameters and a p-value threshold < 0.05 to analyse the fructose data.

2.3 Results

2.3.1 Integration of DNA methylation and gene expression during adipocyte differentiation

Genome-wide DNA methylation was measured using the Illumina 450K BeadChip at different time points during adipocyte differentiation to determine the changes in DNA methylation accompanying adipocyte differentiation compared to the baseline (0 hours in the absence of fructose without fructose). A total of 2, 2 ,4, 607 and 155,573 genome-wide DMRs were found at 24, 48, 96, 192 and 384 hours, respectively. The corresponding differential transcriptomic analysis identified 2007, 2473, 4977, 6594 and 5237 genes at the same respective time points.

The integration of DNA methylation and transcriptomic data identified DMRs in genes which were differentially expressed during adipocyte differentiation. The majority of methylation sites analysed did not change between pre-induction (0 hour) and 24, 48, and 96 hours after induction (Figure 2.3A). However, a large number of changes in methylation were apparent at 192 hours and 384 hours versus baseline (Figure 2.3 and Table 2.2).

Table 2.2 Differentially methylated regions (DMRs) and differentially expressed genes (DEGs) at different time points across adipocyte differentiation in control adipocytes. The number of genes differentially expressed used for the integration, and the number of regions with a significant change in methylation levels (both genome-wide and on DEGs) are displayed for each comparison. DMRs on DEGs are detected only at 192 and 384 hours following the initiation of differentiation.

Time-point (hours)	No. DEGs	Genome-wide DMRs	No. DMRs on DEGs
24	2007	2	–
48	2473	2	–
96	4977	4	–
192	6594	607	57
384	5237	155573	1437

At 192 hours, 57 of the 6,594 (0.8% of the DEG) differentially expressed genes showed significant changes also in methylation levels. At 384 hours, methylation changed in 1437 genomic regions (DMRs), in 1,254 of the 5,237 differentially expressed genes (23.95% of the DEGs). 130 of those 1,254 genes were differentially methylated in multiple regions. Hereafter, the differentially methylated/expressed genes (or regions) will be referred to as

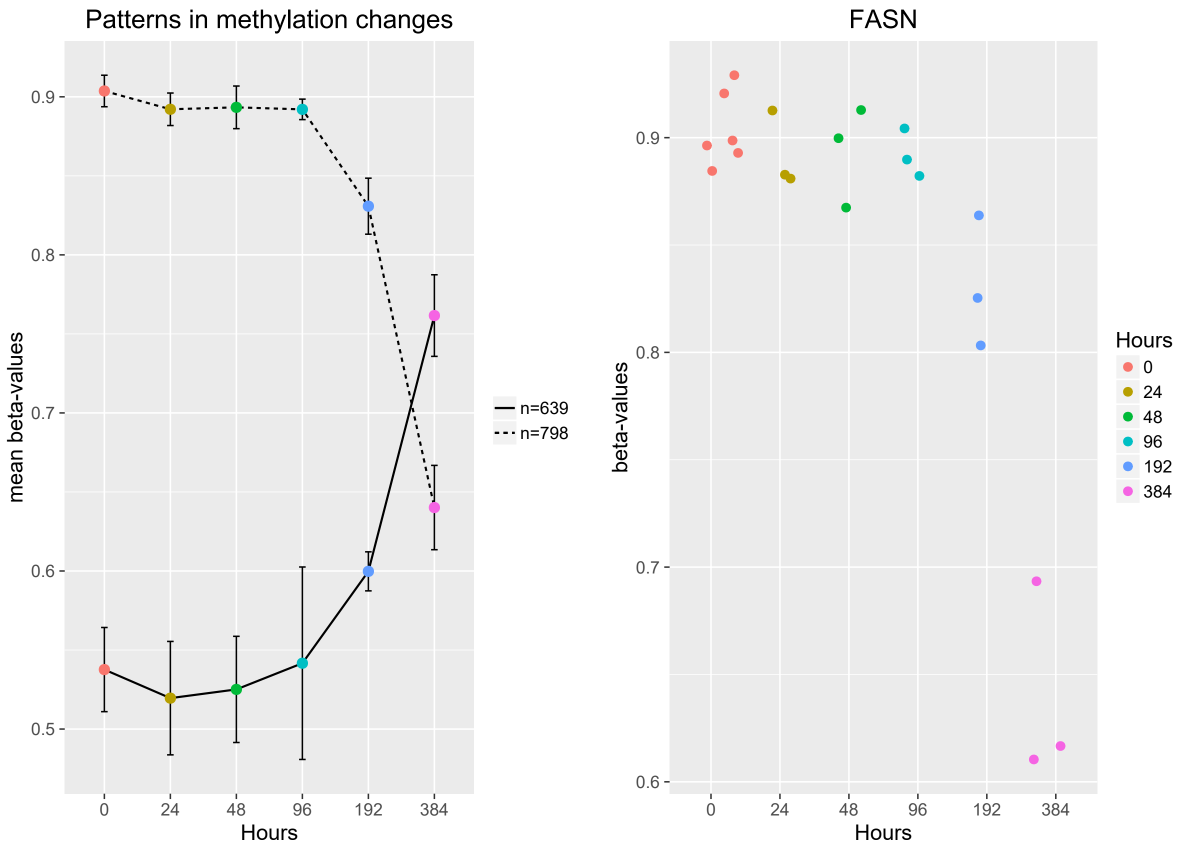


Fig. 2.3 DNA Methylation levels for probes in control adipocytes, reveals a general decreasing in DNA methylation. Different time-points across differentiation are represented by different colors. A) Methylation patterns during differentiation. Dots define averaged β-values for each time-point. 798 differentially methylated genes present the decreasing trend, but β-values for 639 genes increase with time. The biggest changes happen between 192 and 384 hours. B) Change of methylation levels during differentiation for the gene *FASN*. Dots represent the different replicates available for each time-point.

DMRs.

20 of the 57 genes DMRs at 192 hours maintained their methylation status at 384 hours (Figure 2.4), suggesting that methylation occurred before 192 hours and lasted at least until 384 hours. Gene enrichment analysis revealed that the top-ranked KEGG pathway was glyoxylate and dicarboxylate metabolism (*e.g.*, *SHMT1, GLUL*) and GO terms involved in morphogenesis, adhesion, and developmental process not corrected for multiple comparisons (Table 2.3), although none of these pathways was significant after correction for multiple comparisons. Eight (*BCOR, EBF3, ETS2, GLI2, ITGA7, NPC1, PDXK, SPON2*) of the 20 genes have been shown to be involved in adipocyte differentiation or function and all but four (*GLUL, PLEKHG6, SHMT1, TPD52L2*) are in involved in other types (*e.g.*, neurite,

intestinal, cardiac and other) of differentiation or development processes (see Table 2.10 in supplementary material section).

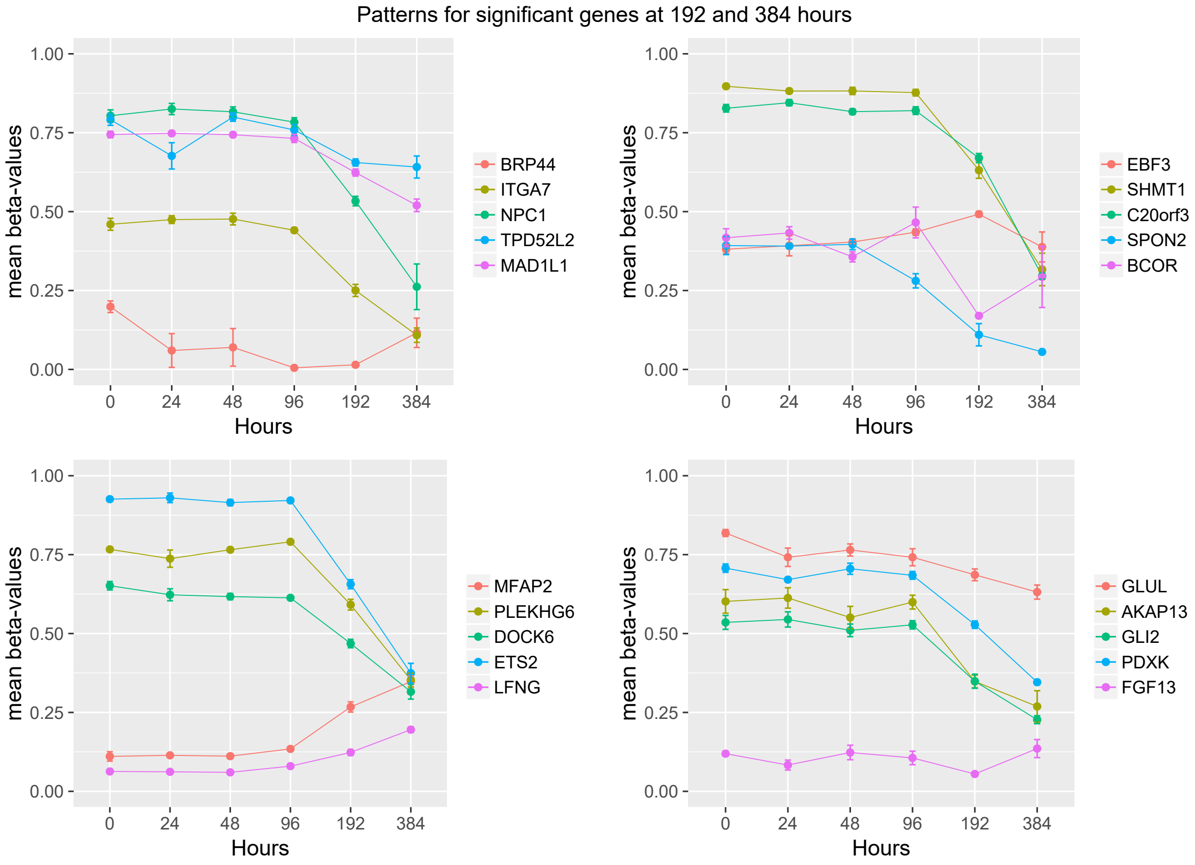


Fig. 2.4 Methylation patterns during differentiation for the 20 genes showing significant methylation/expression changes both at 192 and 384 hours. To better inspect the patterns, gene results are separated in four panels. Dots define averaged β-values for each time-point, while different colours represent different genes.

Table 2.3 Enriched pathways from the DAVID analysis of the 20 DMRs found to be significantly methylated/expressed both at 192 and 384 hours. P-value and pathway genes are also provided.

Category	**Term**	**P-value**	**Genes**
GOTERM_BP_FAT	GO:0007389∼pattern specification process	0.009	*ETS2, GLI2, BCOR, LFNG*
GOTERM_BP_FAT	GO:0009790∼embryo development	0.014	*ETS2, ITGA7, MFAP2, GLI2, LFNG*
GOTERM_BP_FAT	GO:0009952∼anterior/ posterior pattern specification	0.018	*ETS2, GLI2, LFNG*
GOTERM_BP_FAT	GO:0048598∼embryonic morphogenesis	0.019	*ETS2, ITGA7, MFAP2, GLI2*
GOTERM_BP_FAT	GO:0022603∼regulation of anatomical structure morphogenesis	0.020	*ITGA7, FGF13, AKAP13, BCOR, LFNG*
GOTERM_BP_FAT	GO:0045165∼cell fate commitment	0.025	*ETS2, FGF13, GLI2*
KEGG_PATHWAY	hsa00630: Glyoxylate and dicarboxylate metabolism	0.035	*SHMT1, GLUL*
GOTERM_BP_FAT	GO:0016337∼single organismal cell-cell adhesion	0.036	*MAD1L1, ITGA7, GLI2, LFNG*
GOTERM_BP_FAT	GO:0051093∼negative regulation of developmental process	0.039	*FGF13, GLI2, BCOR, LFNG*
GOTERM_BP_FAT	GO:0010639∼negative regulation of organelle organization	0.04	*MAD1L1, FGF13, BCOR*
GOTERM_BP_FAT	GO:0003002∼regionalization	0.043	*ETS2, GLI2, LFNG*
GOTERM_BP_FAT	GO:0098602∼single organism cell adhesion	0.043	*MAD1L1, ITGA7, GLI2, LFNG*

Another group of 20 DMRs were differentially methylated/expressed only at 192 hours (Figure 2.5) and returned to the baseline level at 384 hours suggesting that these were demethylated between these two time points. These genes are enriched in GO term for cell morphogenesis involved in differentiation processes (SPINT2, EFNA3, FN1, MBP) (Table 2.4)) although CTDSP2 and LSS have been individually studied for roles in neuronal or differentiation processes, respectively (supplementary material, Table 2.11).
The remaining 17 DMRs were differently methylated/expressed at 192 hours but were not significantly expressed or methylated at 384 hours.

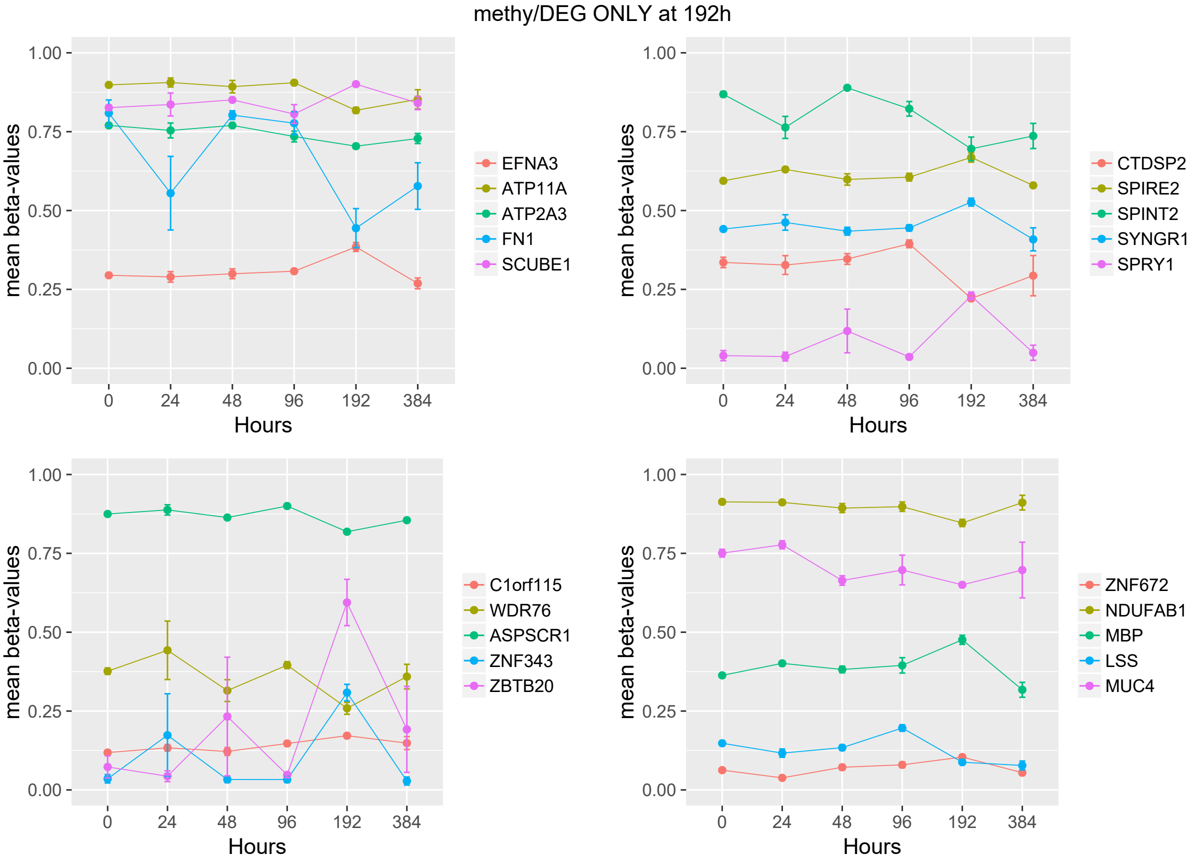


Fig. 2.5 Methylation patterns during differentiation for the 20 genes significantly methylated/expressed at 192 hours (but not at 384 hours) which methylation levels return to the baseline at 384 hours. For each panel, mean *beta*-values at each time point are represented by dots. Different colours represent different genes.

Table 2.4 Enriched pathways from the DAVID analysis of the 20 DMRs found to be significantly methylated/expressed both at 192 and returning to the baseline at 384 hours. P-value and pathway genes are also provided.

Category	Term	P-value	Genes
GOTERM_BP_FAT	GO:0046689 ∼response to mercury ion	0.017	*FN1, MBP*
GOTERM_BP_FAT	GO:0051293 ∼establishment of spindle localization	0.037	*SPRY1, SPIRE2*
GOTERM_BP_FAT	GO:0051653 ∼spindle localization	0.042	*SPRY1, SPIRE2*
GOTERM_BP_FAT	GO:0000904∼cell morphogenesis involved in differentiation	0.047	*SPINT2, EFNA3, FN1, MBP*

2.3.2 Gene location of methylated regions

The location in the gene where DNA methylation occurs differentially may influence gene expression [90]. At 192 hours, the majority of the changes in DNA methylation occurred in the promoter region (31 out of 57 DMRs (54.4%)), 18 (31.6%) were methylated in exons, and 8 (14.0%) in introns (Table 2.5A). At 384 hours, 987 DMRs (68.7% of the total) were found in the promoter of the genes, 272 (18.9%) in exons, 159 (11.1%) in introns, and 19 (1.3%) in intergenic regions of the genes (Table 2.5B). The majority of the genes with differentially methylated promoter regions had an opposite effect on gene expression (*e.g.* up-methylation of promoter region relative to 0 hour was associated with reduced gene expression), while most cases of methylation in exons or introns affected gene expression in the same direction at 384 hours.

One mechanism by which methylation can influence gene expression is by either positively or negatively altering access of transcription factor (TFs) to their binding sites [134]. Binding sites of differentially expressed transcription factors in promoter of differentially expressed genes were analyzed by data mining methods. Among the 181 transcription factors identified in the ENCODE database, 61 were found differentially expressed at 192 hours, and 56 at 384 hours. No consistent or significant binding motif in the promotors DMRs at 192 hours were found. However, 24 motifs were enriched in the differently methylated promoters at 384 hours. Of those, nine were binding sites for differentially expressed TFs (*TFAP2A, ELF1, ETS1, E2F4, E2F1, NR2C2, NR2F2, RXRA, FLI1*) which are involved in a large number of intracellular processes, such as *E2F4* [78] role in suppression of anti-proliferation associated genes and *E2F1* [180] mediated induction of the transcription factor *PPARγ*. The binding

Table 2.5 Differentially methylated regions can be separated by region within or near the gene and into four groups depending on gene expression (G) and methylation changes (M), which can be increasing (↑) or decreasing (↓). Most of the DMRs for A) 192 hours and B) 384 hours were found in the promoter of the genes. At 384 hours, the most common trend is decreasing gene expression and methylation. A combination of decreasing level of methylation and increased gene expression was observed at 192 hours.

A) 192 hours versus 0 hours

Location	**M↓ G↑**	**M↑ G↓**	**M↑ G↑**	**M↓ G↓**	**TOTAL**
Promoter	15	3	6	7	31
Exon	10	–	4	4	18
Intron	3	1	1	3	8
TOTAL	28	4	11	14	57

B) 384 hours versus 0 hours

Location	**M↓ G↑**	**M↑ G↓**	**M↑ G↑**	**M↓ G↓**	**TOTAL**
Promoter	218	386	171	230	987
Exon	63	48	16	145	272
Intron	40	25	7	87	159
Intergenic	8	1	3	7	19
TOTAL	329	442	197	469	1437

sites motifs for those TFs were found in a total set of 486 DMRs. Mapping these 486 DMRs to GO terms and pathways is however problematic since individual genes are regulated by multiple transcription factors.

2.3.3 Integration of methylation, gene and protein expression changes in fully differentiated adipocytes

Levels of 84 proteins were analysed at 384 hours, the only time point that had overlap between DMRs and these proteins in the SomaLogic Somscan V1.0 panel (total of 1,129 proteins). 73 of these proteins showed a significant change in amounts between baseline and 384 hours, consistent with promoter, intron, and exon methylation status. Eight proteins were up-regulated and 76 down-regulated. 63 of the down regulated proteins coordinated with the typical pattern of methylation up-regulation and down-regulation of gene expression. Thirteen of the down regulated proteins and 8 of the up-regulated proteins were however, found to to have up-regulated methylation and up-regulation of gene expression. DNA methylation in promoter regions typically silences genes while gene body methylation is reported to positively correlate to gene expression [244]. Analysis of the pathways of these proteins will be biased since the Somascan platform is based on primarily on secreted and membrane proteins found in the blood with subset derived from cellular contents.

Linked transcripts and proteins but not DMRs

Some genes (219) showed corresponding abundance changes at the protein and transcript level but not at the methylation level at 384 hours. DMRs at these sites were likely maintained at time points which were not analyzed in this study. Subsets of these genes mapped to 312 pathways (p-value< 0.05) with the secretory, membrane, and extracellular processes the most significant (p-value$< 10e-25$) (see Table 2.12 in Supplementary material).

2.3.4 Pathways enrichment analysis

To investigate how pathways involved in converting a pre-adipocyte to a mature adipocyte are influenced by methylation changes, network activity scores were calculated with NASFinder [157] for all DMRs at 384 hours. This tool identifies sub-networks by connecting a list of differently expressed genes to key regulators (in this case, transcription factors). NASFinder analysis revealed 29 significant pathways (shown in Table 2.6). For the transcription factors (TFs) analysis, the pathway with the highest activity score was phospholipase C D1 in phospholipid associated cell signaling. Several of the identified pathways are hallmarks of adipogenesis or function in adipocytes (*e.g.*, *RXR* and *RAR* heterodimerization, *FXR* and *LCR* regulation) but others contribute new information on the differentiation process. Four of the 18 DMRs (*ICAM1, PRKCA, RAC1, RAN*) involved in these 29 TFs pathways were shown to significantly change also at the protein level. *ICAM1* maps to the integrin signaling pathway and may contribute to priming inflammatory processes if misregulated [135]. *PRKCA* is involved in 11 of the 29 pathways demonstrating its central role in cell signaling. *RAC1* mapped to the semaphoring signaling and is a key signaling component for translocation of *GLUT4* to the cell surface [98]. *RAN* is in the noncanonical *WNT* signaling pathway and this GTPase is also involved in several intracellular transport processes necessary for cell fate determination, death, proliferation, differentiation, and transformation [155].

Table 2.6 Significant pathways of gene expression using NASFinder receptor analysis on the 1437 differentially methylated regions found at 384 hours. The p-value, the Network Activity Score (NAS), the DMRs involved in each pathway, and the receptor found by NASFinder are displayed. Genes in bold are up-regulated.

Pathways	**NAS**	**P-value**	**DMRs in the pathway**	**TF**
BIOCARTA PHOSPHOLIPASE C D1 IN PHOSPHOLIPID ASSOCIATED CELL SIGNALING	0.322	0.005	*PRKCA*	*JUNB*
BIOCARTA CBL MEDIATED LIGAND INDUCED DOWNREGULATION OF EGF RECEPTORS	0.191	0.008	*PRKCA*	*MET*
BIOCARTA ACTIVATION OF PKC THROUGH G PROTEIN COUPLED RECEPTOR	0.186	0.003	*PRKCA*	*NFKBIA*
BIOCARTA APOPTOTIC SIGNALING IN RESPONSE TO DNA DAMAGE	0.167	0.011	*APAF1, BID, PRKCA*	*TP53*
BIOCARTA ROLE OF MEF2D IN T CELL APOPTOSIS	0.163	0.018	*MEF2D, PRKCA*	*EP300*
BIOCARTA GROWTH HORMONE SIGNALING PATHWAY	0.154	0.032	*PRKCA*	*SRF*
REACTOME SEMA4D IN SEMAPHORIN SIGNALING	0.149	0.038	*ROCK1, RAC1*	*MET*
PID NONCANONICAL WNT SIGNALING PATHWAY	0.144	0.045	*ROCK1*	*MAPK9*
PID RXR AND RAR HETERODIMERIZATION WITH OTHER NUCLEAR RECEPTOR	0.122	0.037	***SREBF1, NR1H3, ABCA1,*** *NCOR2*	*RXRA*
BIOCARTA FXR AND LXR REGULATION OF CHOLESTEROL METABOLISM	0.115	0.010	***ABCA1, NR1H3***	*RXRA*
PID CANONICAL NF KAPPAB PATHWAY	0.107	0.033	*RAN*	*NFKBIA*
BIOCARTA P53 SIGNALING PATHWAY	0.101	0.030	*PCNA*	*PCNA*

Continued on next page

Table 2.6 – *Continued from previous page*

BIOCARTA OXIDATIVE STRESS INDUCED GENE EXPRESSION VIA NRF2	0.095	0.032	***HMOX1***	*CREB1*
BIOCARTA CADMIUM INDUCES DNA SYNTHESIS AND PROLIFERATION IN MACROPHAGES	0.088	0.013	*PRKCA*	*MYC*
BIOCARTA THE PRC2 COMPLEX SETS LONG TERM GENE SILENCING THROUGH MODIFICATION OF HISTONE TAILS	0.072	0.022	-	*YY1*
BIOCARTA KERATINOCYTE DIFFERENTIATION	0.069	0.046	*PRKCA, ETS2*	*ETS1*
BIOCARTA TPO SIGNALING PATHWAY	0.064	0.029	*PRKCA,* ***STAT5A***	*STAT3*
PID AMB2 INTEGRIN SIGNALING	0.063	0.044	*ROCK1, ICAM1*	*NFKB1*
BIOCARTA REGULATION OF CELL CYCLE PROGRESSION BY PLK3	0.061	0.014	-	*TP53*
BIOCARTA EFFECTS OF CALCINEURIN IN KERATINOCYTE DIFFERENTIATION	0.060	0.022	*PRKCA*	*SP3*
BIOCARTA VEGF HYPOXIA AND ANGIOGENESIS	0.060	0.030	*PRKCA*	*EIF2B1*
REACTOME REGULATION OF GENE EXPRESSION BY HYPOXIA INDUCIBLE FACTOR	0.056	0.012	*CITED2*	*HIF1A*
BIOCARTA TUMOR SUPPRESSOR ARF INHIBITS RIBOSOMAL BIOGENESIS	0.059	0.034	*RAC1*	*TWIST1*
PID HIF 2 ALPHA TRANSCRIPTION FACTOR NETWORK	0.049	0.047	*CITED2*	*ELK1*

Continued on next page

Table 2.6 – *Continued from previous page*

REACTOME SYNTHESIS OF BILE ACIDS AND BILE SALTS VIA 7ALPHA HYDROXYCHOLESTEROL	0.018	0.033	***HSD17B4***	*RXRA*
BIOCARTA OVERVIEW OF TELOMERASE PROTEIN COMPONENT GENE HTERT TRANSCRIPTIONAL REGULATION	0	0.014	-	*TP53*
BIOCARTA PHOSPHORYLATION OF MEK1 BY CDK5 P35 DOWN REGULATES THE MAP KINASE PATHWAY	0	0.021	-	*EGR1*
BIOCARTA NO2 DEPENDENT IL 12 PATHWAY IN NK CELLS	0	0.022	-	*STAT4*
BIOCARTA ROLE OF TOB IN T CELL ACTIVATION	0	0.032	-	*SMAD3*

2.3.5 Influence of fructose on DNA methylation

The status of DNA methylation sites was also assessed in SGBS adipocytes exposed to different concentrations of fructose (2.5, 5 and 10 mM) at 192 and 384 hours after induction. Methylation status of the same genes at each fructose concentration at both time points was compared to the control at both time points (without fructose at the same time point). Only 3 DMRs (*EDEM1, RNF145* and *SLC3A2*) had genes differentially expressed in 2.5mM fructose at 384 hours.

Since so few DMRs were found in genes differentially expressed at these time points and fructose concentrations, genome-wide analysis of DMRs was performed to find general effect of fructose on methylation. At 192 hours, three significant genome-wide DMRs were found only for the highest dose of fructose (10 mM). At 384 hours (Table 2.7) 26 genome-wide DMRs were detected with 2.5mM of fructose (Figure 2.6) and 7 genome-wide DMRs were detected with 5mM of fructose. Most of the genome-wide DMRs obtained in 2.5mM fructose at 384 hours occurred in the promoter regions of the genes (22 of the 26 DMRs). The addition of fructose resulted in up-methylation of the majority of these gene promoters (18 of the 22 DMRs) (Table 2.8). Functional analysis of these 26 significant DMRs resulted in 9 enriched pathways, for example, branched-chain amino acid metabolic process and oxoacid metabolic process (Table 2.9).

Table 2.7 A) Number of genome-wide differentially methylated regions (DMRs) identified at different doses of fructose at 192 and 384 hours. A fructose effect was found at 384 hours for 2.5mM and 5mM. A small effect at 192 hours was found for the highest dose of fructose (10mM), with 3 DMRs.

Fructose dose (mM)	**Time-point (hours)**	**Genome-wide DMRs**
2.5 vs 0	192	–
	384	26
5 vs 0	192	–
	384	7
10 vs 0	192	3
	384	–

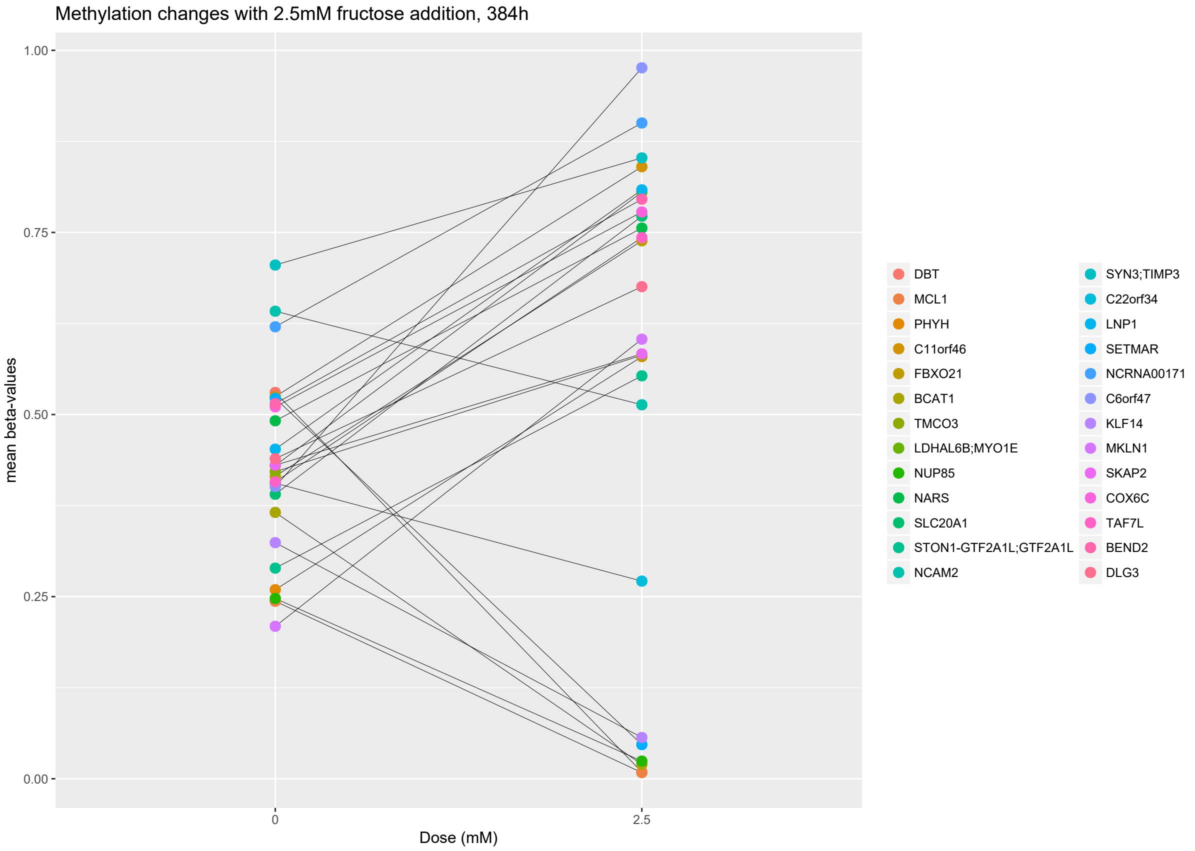


Fig. 2.6 Changes in methylation levels for the 26 DMRs most affected by addition of 2.5mM of fructose at 384 hours. Different colors represent different genes, while dots represent mean β-values without fructose and with 2.5mM fructose.

Table 2.8 Location on the genes and regulation of the 26 DMRs found at 384 hours for 2.5 mM fructose addition.

Location	M↓	M↑	TOTAL
Promoter	7	15	22
Exon	–	1	1
Intron	1	2	3
TOTAL	8	18	26

Table 2.9 Enriched pathways from the DAVID analysis of the significant 26 DMRs found in 2.5mM of fructose addition at 384 hours with p-value and pathway genes.

Category	Term	P-value	Genes
REACTOME_PATHWAY	R-HSA-70895: Branched-chain amino acid catabolism	0.023	*BCAT1, DBT*
GOTERM_BP_FAT	GO:0006367 ∼ transcription initiation from RNA polymerase II promoter	0.024	*STON1-GTF2A1L, GTF2A1L, TAF7L*
GOTERM_BP_FAT	GO:0019752 ∼ carboxylic acid metabolic process	0.025	*BCAT1, DBT, NARS, LDHAL6B, PHYH*
GOTERM_BP_FAT	GO:0043436 ∼ oxoacid metabolic process	0.026	*BCAT1, DBT, NARS, LDHAL6B, PHYH*
GOTERM_BP_FAT	GO:0009083 ∼ branched-chain amino acid catabolic process	0.026	*BCAT1, DBT*
GOTERM_BP_FAT	GO:0009081 ∼ branched-chain amino acid metabolic process	0.030	*BCAT1, DBT*
GOTERM_BP_FAT	GO:0006082 ∼ organic acid metabolic process	0.035	*BCAT1, DBT, NARS, LDHAL6B, PHYH*
GOTERM_BP_FAT	GO:0016054 ∼ organic acid catabolic process	0.037	*BCAT1, DBT, PHYH*
GOTERM_BP_FAT	GO:0006352 ∼ DNA-templated transcription, initiation	0.039	*STON1-GTF2A1L, GTF2A1L, TAF7L*

2.4 Discussion

Genome wide changes in DNA methylation of pre-adipocytes cells induced to differentiate into mature adipocytes were analysed to determine the role of DNA methylation in regulating gene and subsequently protein expression in control (5mM glucose) cells. The effect of varying concentrations of fructose on methylation status and transcriptional regulation was also analysed since fructose stimulates anabolic processes of glutamate and de novo fatty acid synthesis [218] and alters glucose metabolism to induce more energy [219].

2.4.1 DMRs in control conditions

A previous work [213] concluded that DNA methylation of 84 genes was relatively stable between 0 and 240 hours during adipocyte (hMSC cells) differentiation and that changes in DNA methylation were not an underlying mechanism regulating gene expression during adipocyte differentiation. Our results largely confirm these findings since DNA methylation did not change appreciably at 24, 48, and 96 hours with less than 1% at 192 hours. The majority of methylation regions at 192 hours were conserved at 384 hours. The subset of genes with transient methylation (methylated at 192, not at 384, Figure 2.5) occurred in genes associated to cell morphogenesis involved in the differentiation process (*SPINT2, EFNA3, FN1, MBP*, see Table 2.4). In addition, significant changes in methylation linked to changes in gene expression occurred almost 25% (1,254 of 5,237) of the sites analysed at 384 hours compared to pre-induction. These results are consistent with previous metabolic [218, 219] and transcriptomic analysis [157] that showed adipocyte specific metabolism and gene regulation at 192 hours (8 days) of differentiation which became more "robust" between 192 and 384 hours. That is, a complex set of interactions between metabolic pathways, transcriptional regulation, and DNA methylation finalizes maturation of the adipocyte and/or "locks" in its fully differentiated state.

Changes in DNA methylation occurred in genes and pathways known to be involved in adipogenesis (*e.g.*, a *PPARG* receptor linked to *RAR* and *RXR*), as well as pathways not previously associated with adipogenesis (Table 2.6). In addition, pathways identified by DAVID functional annotation of genes with all genes with DMRs down methylated but transcriptionally up regulated (the expected pattern) were similar to those genes that had DMRs up methylated with genes up-regulated (the unexpected pattern). That is, the up-regulated genes mapped to lipid metabolism, mitochondria, oxidoreductase and other pathways regardless of the state of methylation. Similar observations were found for genes that were down regulated regardless of methylation status. Down regulated genes mapped to cell adhesion, cell cycle, cell division, and cytoskeleton (among others) pathways. These results suggest that methylation is a consequence and not a driver of the final maturation stage of adipogenesis. Genes and pathways identified in this in vitro study play roles in obesity and related conditions. An epigenome-wide association study has recently shown that body mass index was associated with widespread changes in DNA methylation in 210 candidate genes in blood cells. Twenty-five of those candidate genes [226] were also differentially methylated at 384 hours in the results presented here. Among these DMRs, *SREBF1, HOXA5, CPT1A, LPIN1 and PHGDH* have established roles in adipose tissue biology and insulin resistance.

2.4.2 Fructose effect on methylation

Fructose has been implicated in the obesity epidemic and specifically in altering the physiology of adipocytes. Methylation changes of differentiating and differentiated adipocytes under the exposition to different fructose concentration can help to better understand the link between fructose and metabolic syndrome. Indeed, modifications in DNA methylation levels are for the most part reversible and can be modulated by optimizing daily habits, such as diet. Although studying the effects of nutrients in cell culture experiments is controversial because concentrations have to be estimated, the doses used in this in vitro experiment (2.5, 5 and 10mM) were modelled on levels following fructose ingestion in humans [153, 224] and considered that local concentrations (*e.g.*, adipose tissue associated with the intestinal tract) could be higher than reported plasma levels. Moreover, the different concentrations of fructose were added to a medium containing 5mM glucose, to better resemble the normal physiological blood glucose concentration [218].

An inverse correlation was found between the number of DMRs assayed and fructose levels (Table 2.7). At 384 hours, 26 genome-wide DMRs were detected in cells grown in 2.5mM of fructose in the presence of 5mM glucose. Pathway analysis mapped these genes to transcription factor processes and branched-chain amino acid (BCAA) catabolism at uncorrected p-values of < 0.05. BCAA catabolism has a functional role in adipocyte differentiation [68] and decreased catabolism of BCAA may be related to insulin resistance, impairment of sub-cutaneous adipocyte hypertrophy, and associated pathologies [169, 236]. Higher circulating BCAA levels were observed in obese and diabetic patients [169].

Only 7 and 3 DMRs were found at 5mM and 10mM fructose in the presence of 5mM glucose at 384 hours post-induction. These small number of genes precluded pathway analysis but no apparent pattern was observed. Individual genes can be annotated and associated with whole body phenotypes. For example, the endoplasmic reticulum degradation-enhancing alphamannosidase-likeprotein1 (*EDEM1*) found to be differentially methylated at 10mM fructose is an endoplasmic reticulum stress (ERS) marker [179]. Acute ERS can weaken the capacity of mature adipocytes to store lipids and chronic ERS can impair the adipogenic potential of preadipocytes [99]. Disruption of these pathways could contribute to obesity-associated morbidities such as lipid spillover, ectopic fat deposition and ultimately insulin resistance [29]. Epigenetic modifications in ring finger protein 145 (*RNF145*, a metal binding proteins), also found at 10mM fructose were associated with BMI, waist circumference and changes in BMI in African American adults [45].

The results presented here suggest that fructose has moderate effects on methylation levels at 2.5mM fructose, a 1:2 ratio with the 5mM glucose in the culture media. However, changes in methylation decreased at equimolar doses (5mM fructose added to basal 5mM glucose) or 2:1 (10mM fructose with 5mM glucose). We speculate that fructose may play a regulatory role at doses less than 1:1 fructose to glucose, but at equimolar or greater levels, fructose is “shunted” to metabolic pathways to produce stored (oleate) and released fatty acids (palmitate) as demonstrated by Varma *et al.* in their previous work [218]. A study in sheep indicated that hepatic DNMT3A mRNA levels decreased with increasing consumption of high-fat-sucrose diets [38] consistent with the observations at high fructose described for adipocytes in culture.

2.4.3 Integration of methylation, gene expression and proteomics

A novel feature of this study was the analysis of DNA methylation, mRNA levels, and selected proteins at 384 hours after differentiation. Seventy-three of the DMRs (out of the 84 SomaLogic proteins with corresponding DMRs) showed significant proteomic changes indicating that DNA methylation changes were transmitted to protein levels. Of the 63 down-regulated proteins and genes, DNA methylation of majority of genes occurred in the promoter region. Methylation occurred at more than one region in some genes, for example, up-methylation of *COL18A1* occurred both in the intronic and exonic regions of the gene. On the other hand, methylation of some genes including *MRC2, CRLF1, RAC1* occurred in the promoter or either intronic/exonic regions. DKK [69] and other genes that have well established roles in adipogenesis have been shown to have methylation consistent with direction of expression [214].

More than half of the 73 genes that were methylated and coordinated with protein have been identified as methylated regions in clinical samples of human tissue adipose tissue that have been previously linked to obesity and diabetes [159]. Although identified to occur in adipose tissue, many of these genes still remain to be fully characterized in adipose tissue and a functional role in the adipogenic process, remains to be defined (*e.g. CRLF1*) [253]. Methylation of several genes and their protein changes are here reported for the first time and no literature exists describing their expression in adipose tissue (*e.g., RPS7, COLEC12*) or rolein adipogenesis. The integration of gene methylation, and their mRNA and protein levels make these likely targets and biomarkers of adipocytes differentiation and contribute to improving our understanding of this process.

A limitation of this analysis was the use of somamers (*e.g.*, Somalogic platform) that is based on a subset of proteins that necessarily need to be found in blood. Nevertheless, these diverse high throughput methods allow for identifying and linking changes at the chromosome level through protein levels. The integration of these diverse data types was possible by the use of a highly characterized euploid cell line, as to more complex adipose tissue which would have multiple cell types.

To be noted, the transcriptomic and methylation data used for this analysis were obtained from different cells plates cultured at the same time and under identical conditions. We also analyzed the integration of methylation data with a gene expression dataset obtained from an experiment conducted on a different day but under identical conditions [157]. This secondary analysis led to similar but less significant results. For example, significant DMRs were found only at 384 hours for a total of 198 genes compared with the 8279 DEGs described here. Sixty-eight of these genes were found to be differentially methylated when using data from the same experiment. This replication study highlights the difficulties that could be encountered when integrating data produced under similar conditions but not collected at the same time, as suggested by Cavill *et. al* [31].

2.5 Conclusion

In summary, the results presented in this Chapter provide additional insight into the molecular process of preadipocyte differentiation to mature adipocytes, through the use of supervised linear omics integration. To reduce the size of the problem, with respect to the genome-wide study, we used prior knowledge to focus only on genomic regions known to contain differentially expressed genes.

The transcriptomic and methylation data integration indicated that DNA methylation and resultant gene expression patterns are "pre-programmed" since up or down gene regulation overlap DMRs that could have the expected (down methylation, up gene expression) or the reverse (up methylation up gene regulation) pattern.

Importantly, three-omics integration (methylation, gene expression and protein concentrations) identified coordinated changes across the omics data in several genes, indicating that DNA methylation changes were transmitted to protein levels.

2.6 Supplementary material

Table 2.10 Information about genes found to be differentially methylated/expressed at 192 and 384 hours (from Figure 2.4). The * indicates direction of mean beta value

Gene	Protein	Regulation Post 96 hr*	Gene Card Info	Notes
AKAP13	A-Kinase Anchoring Protein 13	Down	Binds to the regulatory subunit of protein kinase A (PKA) and confining the holoenzyme to discrete locations within the cell. Also enhance ligand-dependent activity of estrogen receptors alpha and beta	Cardiac development & cardiomyocyte development [133]
BCOR	BCL6 Corepressor	Down	Interacting corepressor of BCL6, a POZ/zinc finger transcription repressor that is required for germinal center formation and may influence apoptosis. Specific class I and II histone deacetylases (HDACs) have been shown to interact with this protein	Required for adipocyte differentiation [81]
BRP44	Mitochondrial Pyruvate Carrier 2	Up	Mediates the uptake of pyruvate into mitochondria	Expression increased between intestinal stem cells and fully differentiated cells [184]
C20orf3	Adipocyte Plasma Membrane Associated Protein	Down	Exhibits strong arylesterase activity with beta-naphthyl acetate and phenyl acetate. May play a role in adipocyte differentiation.	Key node at day 10 of adipogensis connected with PPAR-g node [14]
DOCK6	Dedicator Of Cytokinesis 6	Down	Plays a role in actin cytoskeletal reorganization by activating the Rho GTPases Cdc42 and Rac1	Needed for neurite outgrowth [146]
EBF3	Early B-Cell Factor 3	Down	Inhibits cell survival through the regulation of genes involved in cell cycle arrest and apoptosis.	Expressed during adipogenesis but may not be essential [86]
ETS2	ETS Proto-Oncogene 2, Transcription Factor	Down	Transcription factor which regulates genes involved in development and apoptosis. Involved in regulation of telomerase.	Regulates adipogenesis in vitro through a role in clonal expansion [12]

Continued on next page

Table 2.10 – *Continued from previous page*

FGF13	Fibroblast Growth Factor 13	Up	Embryonic development, cell growth, morphogenesis, tissue repair, tumor growth, and invasion	Neuronal migration and polarization, negative regulatory role in muscle differentiation [125, 251]
GLI2	Glioma-Associated Oncogene Family Zinc Finger 2	Down	Gli family zinc finger proteins are mediators of Sonic hedgehog (Shh) signaling.	Counterbalances osteogenesis and adipogenesis [172]
GLUL	Glutamate-Ammonia Ligase	Down	Catalyzes the synthesis of glutamine from glutamate and ammonia in an ATP-dependent reaction (detoxification), acid-base homeostasis, cell signaling, and cell proliferation.	
ITGA7	Integrin, Alpha 7	Down	Wide spectrum of cell-cell and cell-matrix interactions involved in cell migration, morphologic development, differentiation, and metastasis	Upregulation of ITG7 during adipogenesis at 15 days [211]
LFNG	O-Fucosylpeptide 3-Beta-N-Acetylglucosaminyltransferase	Up	Glycosyltransferases that act in the Notch signaling pathway to define boundaries during embryonic development	Promote T and B cell development [198]
MAD1L1	Mitotic Checkpoint MAD1 Protein Homolog	Down	Component of the mitotic spindle-assembly checkpoint that prevents the onset of anaphase until all chromosome are properly aligned at the metaphase plate	Mad1 expression reduces diffferentiation [171]
MFAP2	Microfibril Associated Protein 2	Up	Elastin-associated microfibrils	Up regulated during osteoblastic differentiation of human mesenchymal cells [23]
NPC1	NPC Intracellular Cholesterol Transporter 1	Down	Limiting membrane of endosomes and lysosomes and mediates intracellular cholesterol trafficking via binding of cholesterol to its N-terminal domain	No regulation observed in SGBS at day 10 of differentiation [10]

Continued on next page

Table 2.10 – *Continued from previous page*

PDXK	Pyridoxal Kinase	Down	Phosphorylates vitamin B6, a step required for the conversion of vitamin B6 to pyridoxal-5-phosphate, an important cofactor in intermediary metabolism	PDXK mRNA level associated with adipogenic, lipid-droplet-related and lipogenic genes. Insulin sensitivity positively associated with PDXK expression. In human pre-adipocytes, PDXK mRNA levels increased during adipocyte differentiation at day 7 [148]
PLEKHG6	Pleckstrin Homology And RhoGEF Domain Containing G6	Down	Induces myosin filament formation. At the cleavage furrow to advance furrow ingression during cytokinesis. In epithelial cells, required for the formation of microvilli and membrane ruffles on the apical pole	
SHMT1	Serine Hydroxymethyltransferase 1	Down	Cytosolic form of serine hydroxymethyltransferase, a pyridoxal phosphate-containing enzyme that catalyzes the reversible conversion of serine and tetrahydrofolate to glycine and 5,10-methylene tetrahydrofolate	
SPON2	Spondin 2	Down	Cell adhesion protein that promotes adhesion and outgrowth of hippocampal embryonic neurons. Binds directly to bacteria and their components and functions as an opsonin for macrophage phagocytosis of bacteria. Essential in the initiation of the innate immune response and represents a unique pattern-recognition molecule in the ECM for microbial pathogens (By similarity). Binds bacterial lipopolysaccharide (LPS).	20(S)-hydroxycholesterol induces SPON2 at 96 hours adipocyte differentiation [173]
TPD52L2	Tumor Protein D52 Like 2	Down	GO annotations related to this gene include poly(A) RNA binding and protein heterodimerization activity.	

Table 2.11 Information about genes found to be differentially methylated/expressed at 192 but not at 384 hours (from Figure 2.5). The * indicates direction of methylation mean beta value

Gene	Protein	Gene Card Info	Notes
ASPSCR1	Alveolar Soft Part Sarcoma Chromosomal Region Candidate Gene 1 Protein	Contains a UBX domain and interacts with glucose transporter type 4 (GLUT4) as a tether, sequestering GLUT4 in intracellular vesicles in muscle and fat cells in the absence of insulin. Redistributes the GLUT4 to the plasma membrane within minutes of insulin stimulation.	
ATP11A	ATPase Phospholipid Transporting 11A	Probably phosphorylated in its intermediate state and likely drives the transport of ions (e.g., calcium) across membranes	
ATP2A3	ATPase Sarcoplasmic/Endoplasmic Reticulum Ca2+ Transporting 3	Catalyzes the hydrolysis of ATP coupled with the translocation of calcium from the cytosol to the sarcoplasmic reticulum lumen, and is involved in calcium sequestration associated with muscular excitation and contraction	
C1orf115			
CTDSP2	CTD (Carboxy-Terminal Domain, RNA Polymerase II, Polypeptide A) Small Phosphatase 2	Negatively regulates RNA polymerase II transcription, possibly by controlling the transition from initiation/capping to processive transcript elongation	Knock-down induces neuronal development [49, 71]
EFNA3	Eph-Related Receptor Tyrosine Kinase Ligand 3	Implicated in mediating developmental events, especially in the nervous system and in erythropoiesis	Expression in early fetal but inhibited in adult muscle stem cells [5]
FN1	Fibronectin 1	Fibronectin involved in cell adhesion and migration processes including embryogenesis, wound healing, blood coagulation, host defense, and metastasis	Dysregulated in obesity [9]

Continued on next page

Table 2.11 – *Continued from previous page*

LSS	Lanosterol Synthase (2,3-Oxidosqualene-Lanosterol Cyclase)	Member of the terpene cyclase/mutase family and catalyzes the first step in the biosynthesis of cholesterol, steroid hormones, and vitamin D.	Up-regulated in self-renewing cells by blocking apoptosis and differentiation program [137]
MBP	Myelin Basic Protein	MBP is a major constituent of the myelin sheath of oligodendrocytes and Schwann cells in the nervous system	
MUC4	Mucin 4, Cell Surface Associated	highly glycosylated proteins called mucins are part of the viscous secretion that covers epithelial surfaces such as those in the trachea, colon, and cervix.	
NDUFAB1	NADH:Ubiquinone Oxidoreductase Subunit AB1	Carrier of the growing fatty acid chain in fatty acid biosynthesis. Accessory and non-catalytic subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase (Complex I)	
SCUBE1	Signal Peptide, CUB Domain And EGF Like Domain Containing 1	Expressed in platelets and endothelial cells and may play an important role in vascular biology	
SPINT2	Serine Peptidase Inhibitor, Kunitz Type 2	Inhibits HGF activator which prevents the formation of active hepatocyte growth factor	
SPIRE2	Spire Type Actin Nucleation Factor 2	Actin nucleation factor, remains associated with the slow-growing pointed end of the new filament. Involved in intracellular vesicle transport along actin fibers, providing a novel link between actin cytoskeleton dynamics and intracellular transport.	

Continued on next page

Table 2.11 – *Continued from previous page*

SPRY1	Sprouty RTK Signaling Antagonist 1	May function as an antagonist of fibroblast growth factor (FGF) pathways and may negatively modulate respiratory organogenesis	Spry1 is a critical regulator of adipocyte differentiation and mesenchymal stem cell (MSC) lineage allocation, potentially acting through regulation of CEBP-beta and TAZ [212]
SYNGR1	Synaptogyrin 1	Studies of a similar murine protein suggest that it functions in synaptic plasticity without being required for synaptic transmission	
WDR76	WD Repeat Domain 76	Specifically binds 5-hydroxymethylcytosine (5hmC), suggesting that it acts as a specific reader of 5hmC	
ZBTB20	Zinc Finger And BTB Domain Containing 20	Acts as a transcriptional repressor and plays a role in many processes including neurogenesis, glucose homeostasis, and postnatal growth.	Required for de novo lipogenesis perhaps through ChREBP [119]
ZNF343	Zinc Finger Protein 343	May be involved in transcriptional regulation	
ZNF672	Zinc Finger Protein 672	May be involved in transcriptional regulation	

Table 2.12 Most significant enriched pathways from the DAVID analysis of the 218 genes found to be significantly expressed at the proteomic and transcriptomic level at 384 hours. Those genes di not show methylation changes. The number of genes in the pathway, the p-value and FDR are shown in the table.

Category	**Term**	**No. Genes**	**P-value**	**FDR**
UP_KEYWORDS	Secreted	116	1.07e-59	1.43e-56
UP_SEQ_FEATURE	signal peptide	141	1.57e-56	2.48e-53
GOTERM_CC_DIRECT	GO:0005615∼ extracellular space	99	3.91e-53	5.28e-50
UP_KEYWORDS	Disulfide bond	136	9.95e-52	1.33e-48
GOTERM_CC_DIRECT	GO:0005576∼ extracellular region	102	5.39e-49	7.29e-46
UP_KEYWORDS	Signal	144	1.08e-48	1.44e-45
UP_SEQ_FEATURE	disulfide bond	120	3.99e-44	6.31e-41
UP_KEYWORDS	Glycoprotein	129	3.94e-32	5.28e-29
UP_SEQ_FEATURE	glycosylation site: N-linked (GlcNAc)	121	1.28e-28	2.03e-25

Chapter 3

Comparison of simultaneous and unsupervised clustering methodologies

The complexity of multi-omics data integration increases as far as more than two molecular layers are considered. The addition of an higher number of omics data implies more interactions to be modelled. Those relationships can moreover occur simultaneously, as stated by the Hypothesis B of interaction (Figure 1.2). To take this into account, multi-omics integration is approached by the use of specific statistical and mathematical tools. Recently, there has been an advancement in the development of mathematical methodologies to simultaneously combine more than two data types. With respect to biological (linear) integration, simultaneous methods allow researchers to use more information, but have as a counterpart an increase in their complexity.

In this Chapter, we thus study the impact of several factors on unsupervised simultaneous integration. We consider the problem of correctly separating sample subtypes. We categorized those methods with regards to how they handle data before combining them, then we study the influence of this classification on simulated and real world datasets. Additionally, to better understand the general problem of multi-omics integration, we study how integration performance is affected by the data at hand. We consider the number of omics integrated, the number of subtypes to be recovered, presence of noise and signal strength across different data types. Finally, since not all the biomolecules in a molecular layer are related to the phenotype of interest, we study the effect of data pre-processing, such as feature selection, and show that noise reduction can improve the quality of the retrieved information.

The content of this chapter has been published in 2017 in *Briefings in Bionformatics* [207].

3.1 Introduction

Technological advances in high-throughput biological data generation such as Next Generation Sequencing [217], Mass Spectrometry [54] and Nuclear Magnetic Resonance Spectroscopy [59], now allow the simultaneous collection of information from multiple molecular levels and biological systems. Usually, molecular levels (*i.e.*, -omics) have been investigated in isolation for their association with a phenotypic trait of interest. This concept is, however, challenged by many since it views biology linearly and does not consider the interactions between different molecular levels at the basis of the central dogma of biology [39]. Denis Noble recently proposed a multi-level causality theory with feedback cycles among biochemical layers [160] where interactions within and across different omics are acknowledged. The growing availability of multi-omics data and the emerging biological phenotypes originating from complex traits and interactions increased the need for adequate multi-omics integration methods [156].

Some reviews and theoretical classifications have recently defined general pipelines to combine omics data. They focused on specific data types or biological systems [28, 31, 152, 175] and computational differences among methods [11, 230]. In this work we will focus on statistical methods that simultaneously combine more than two different omics [53, 66], which is in line with the hypothesis that multiple biomolecular levels interact non-linearly to contribute to a given phenotype [175]. We will provide a classification of those methods based on how data are handled before performing integration, and we will explore the effects of factors such as data pre-processing, number of considered omics, signal strength on resulting omics integration.

Statistical integration methods can be used to solve several types of biological questions by reinforcing common signal from different platforms (*e.g.* genomics and transcriptomics, miRNA and transcriptomics, transcriptomics and proteomics, or proteomics and metabolomics) or by combining complementary information potentially carried by data that do not interact directly (*e.g.* transcriptomics and metabolomics). Multi-omics integration has been used for the discovery of molecular mechanisms [63, 170, 232], biomarkers [33, 193, 220, 225] and sample/patient classification [118, 140, 147, 191, 192, 252]. New methods are constantly developed to challenge these biological questions: recently, Singh et al. have introduced DIABLO [193], an expansion to more than two data types of the integrOmics supervised integration method [108], which found biomarkers for three different Breast Cancer subtypes (Basal, Her2, Luminal A).

This chapter will focus on the sample classification case by comparing statistical unsupervised multi-omics integration methods that deal simultaneously with more than two data types.

Rewiev of statistical multi-omics integration approaches

Statistical integration approaches can be classified as multivariate, concatenation-based, and transformation-based methods according to how data are manipulated before applying the algorithm.

Multivariate methods [31] are usually based on Partial Least Squares (PLS) [107, 241] or Canonical Correlation Analysis (CCA) [67, 76] and they treat different omics separately to find associations between them. We focus here on CCA-based approaches [108, 117, 239], which, differently to PLS-based methods [17, 36, 108, 114, 121, 122, 193], do not imply any hierarchy between data. An example of multivariate CCA-based approach is the Multiple Canonical Correlation Analysis (MCCA) [239], an extended sparse CCA [240].

Concatenation-based integration [175] is performed by combining omics data in a single matrix, used as input for low-rank based approximation [131] or latent factor analysis [88] in order to combine the data into a single low-dimensional space [120, 138, 139, 186, 191, 243]. Lock et al. proposed Joint and Individual Variation Explained (JIVE) [120], a method based on the decomposition of omics data in the sum of three terms: a low-rank joint variation matrix, a low-rank individual matrix and the residual noise. This method applied to gene expression and miRNA from Glioblastoma Multiforme (GBM) samples, revealed differences in GBM subtypes involving both miRNA and gene expression. Another concatenation-based method is the Multiple Co-Inertia Analysis (MCIA) [138], an extension of Co-Inertia Analysis [41] to more than two data types. Following covariance optimization between the global score derived from the concatenated matrix and single omics scores, this method was applied to mRNA, miRNA and proteomics data and succeeded in distinguishing profiles from melanoma, leukemia and CNS cell lines [140]. Furthermore, Multiple Factor Analysis (MFA) [44, 166] is a concatenation-based method whose strategy is instead based on the principal component analysis (PCA) of the concatenated matrix. MFA was applied in [44] to copy-number measurements and gene expression from a glioma dataset to study differences between different tumor subtypes.

Finally, the transformation-based methods integrate omics data after their transformation into an intermediate and common form, like a graph or a kernel matrix [33, 43, 115, 130, 199, 229]. The main advantage of a transformation step is to preserve individual omics characteristics that can be lost otherwise [175]. For example, the Similarity Network Fusion (SNF), described by Wang et al. [229], creates patient similarity networks from the omics data of interest. The method recognized three GBM subtypes with different survival profiles from the integration of DNA methylation, mRNA and miRNA expression.

The methods selected for the comparison are Multiple Canonical Correlation Analysis (MCCA) [239], Joint and Individual Variation Explained (JIVE) [120], Multiple Co-Inertia

Analysis (MCIA) [138], Multiple Factor Analysis (MFA) [44] and Similarity Network Fusion (SNF) [229] (see Table 3.1 and Appendix A for a more detailed description of the methods).

Table 3.1 Summary of the multi-omics integration methods reviewed. The column "Data scaling" indicates which scaling has been applied to data before integration.

Method	Integration Approach	Description	Data Scaling	R package
MCCA [239]	Multivariate	Seeks linear combination of correlated features from different data	Columns normalization (mean=0; sd=1)	PMA
JIVE [120]	Concatenation	Separates signal common to all data from individual one	Columns normalization (mean=0; sd=1)	r.jive
MCIA [138]	Concatenation	Projects data on a common lower dimensional space	Non-symmetric correspondence analysis	omicade4
MFA [44]	Concatenation	Projects data on a common lower dimensional space	Columns normalization (mean=0; sd=1)	FactoMineR
SNF [229]	Transformation	Builds a fused network from single ones	Columns normalization (mean=0; sd=1)	SNFtool

These chosen methods are well-known unsupervised algorithms, representative of the different classes of statistical integration approaches and already considered in reviews focused on specific theoretical characteristics of the methods (unsupervised/supervised [79], use of networks [11], cluster computation [230], dimension reduction [140]). Our classification, based on how methods handle data, take into account all these aspects by providing a direct comparison of the methodologies, which, although suggested in [79], has never been presented in literature. Moreover, these methods can be applied to different types of omics without any required previous knowledge about the phenotype of interest. Interestingly as well, these methods are all provided as R packages, making them suitable for a direct comparison inside the same computing environment. Finally, this paper will also address the impact of experimental design, data pre-processing and parameter training on the multi-omics integration outcomes.

A graphical overview of the chapter structure is presented in Figure 3.1 describing the comparison pipeline, method classification, the tested datasets and result organization.

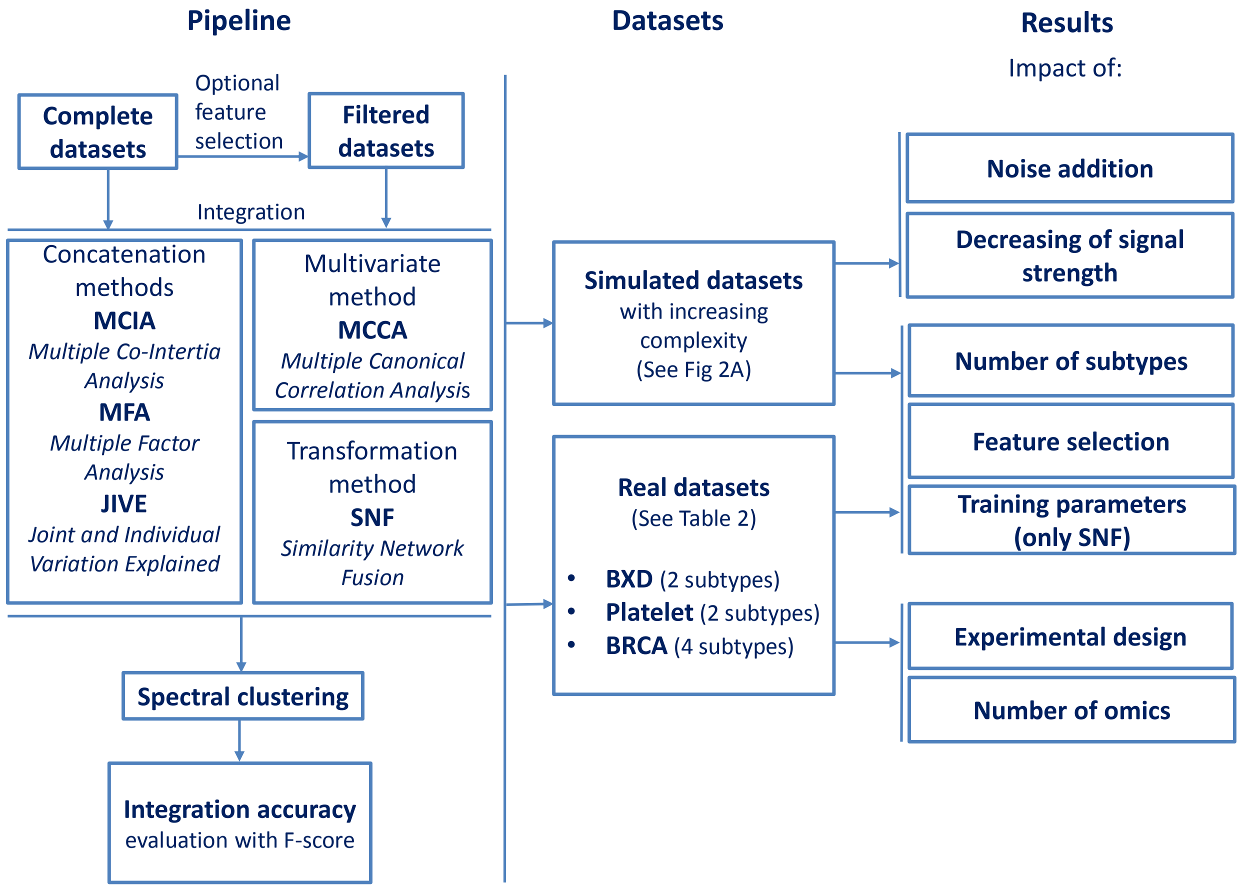


Fig. 3.1 Graphical overview of the comparison of the multi-omics integration methods. The schema is divided in three areas: i) method classification and comparison pipeline, ii) datasets iii) result organization. In the left side of the schema, the different pipeline steps are presented, together with the method classification. This is represented in the pipeline step called “Integration”, where each block collects the methods belonging to the same integration approach. Datasets are represented in the middle according to the division simulated/real datasets (BXD: murine liver dataset; Platelet: platelet reactivity dataset; BRCA: breast cancer dataset). Finally, results are organized in the right part of the diagram. Arrows linking datasets and results indicate which dataset has been used to produce the corresponding result.

3.2 Datasets

Methods were tested on three real datasets and on several simulated dataset, each composed by three different data types. Pre-processing on the former and generation of the latter are described in the sections 3.2.1 and 3.2.2.

3.2.1 Real datasets

Methods were tested on three real datasets (murine liver (BXD) [238], platelet reactivity [256] and Breast Cancer (BRCA) [203] datasets), each one composed by three different data types including transcriptomics, proteomics, metabolomics, miRNA and epigenomics (see details in Table 3.2; PCA data visualization and correlation analysis in Figures 3.6-3.8 and Tables 3.4-3.6).

Table 3.2 Overview of the three real datasets used to compare integration methods. Columns provide the studied phenotype, the number of subjects (total and for each subtype) and the omics data included in each dataset.

Dataset	Phenotype	No. Subjects	Subtypes	Omics	Platform
BXD [238]	Mitochondrial metabolism	66	High Fat Diet (31) Chow Diet (35)	Transcriptomics	Affymetrix Mouse Gene 1.0 ST microarrays
				Proteomics	SWATH-MS quantification
				Metabolomic	MS signatures
Platelet [256]	Platelet Reactivity	12	High (6) Low (6)	Transcriptomics	Affymetrix GeneChip Human Genome U133 Plus 2.0 arrays
				Proteomics	MS quantification
				miRNA	NanoString
BRCA [203]	Breast Cancer	491	Luminal A (225) Luminal B (120) HER2 enriched (56) Basal-like (90)	Epigenomics	Illumina Infinium
				Transcriptomics	Agilent microarrays
				miRNA	Illumina sequencing

Some data pre-processing was performed prior to integration. The averaged values across the measured probes of the BXD dataset were retained for each gene. Missing values for proteins and metabolites measurements were substituted by their median over all the cohorts. To obtain the proteomics data for the Platelet dataset, the averaged ratio of all peptides for a given protein was considered (available for download from the Omics Discovery Index, http://www.omicsdi.org/). Then, quantile normalization was applied to reduce the batch effect (function normalizeBetweenArrays from limma R package [196]).
For the BRCA dataset [203], the 8 subjects with a Normal-like cancer subtype were excluded from the analysis, due to the small number of samples. Moreover, 80 randomly selected subjects were considered for the two largest subtypes, Luminal A and B, to avoid a bias with respect to these cancer subtypes: missing values for gene expression and methylation data were then substituted with their median values across the subjects.

3.2.2 Simulated datasets

Several simulated datasets (Figure 3.2), composed of three matrices 60×500 (60 subjects and 500 features), were created to evaluate the performances of the considered algorithms in a more controlled context. In each dataset, sample profiles were generated following normal distributions with mean ($m_1 \dots m_{500}$) and standard deviation ($sd_1 \dots sd_{500}$) derived from randomly selected gene expressions, methylation levels and miRNA from the Breast Cancer dataset [203].
Samples were then associated to a class and their profiles were generated according to the following normal distributions:

- $\sim \mathcal{N}(m_i, sd_i)$ if the sample belonged to the first class;
- $\sim \mathcal{N}(m_i - \frac{m_i}{2}, sd_i)$ if it was assumed to belong to the second one;
- $\sim \mathcal{N}(m_i + \frac{m_i}{2}, sd_i + \frac{sd_i}{10})$ if it belonged to the third one.

The parameters used in the second and third distributions were chosen to generate well separated groups of samples. Independent noise $\sim \mathcal{N}(0, 0.4)$ was also added to the data matrices during the generation.
Finally, since not all the molecules in a system usually contribute to a significant signal, the considered methods were also tested after adding noisy columns to the matrices. Sample profiles on those columns were again generated as normal distributions with means derived from the Breast Cancer dataset features and standard deviation equal to 2. To reproduce the diverse quantity of noise that different omics data often exhibit in real studies, 100 features

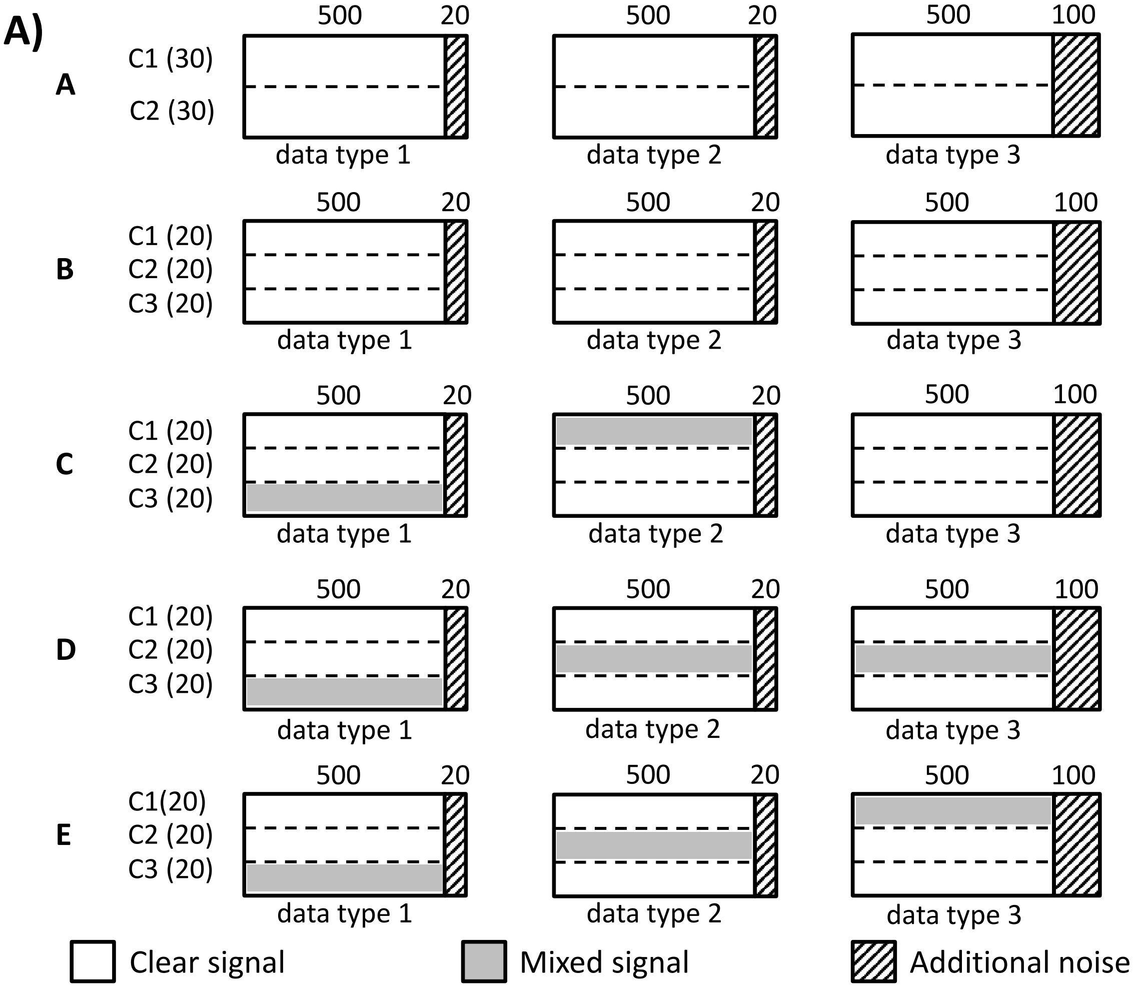


Fig. 3.2 Visualization of the simulated scenarios: for each of them, three data types (d1, d2 and d3) of 500 features were generated by creating 60 samples divided in two (case A) or three groups (cases B-E). When the group is colored in white, it has been generated to be clearly detectable in the data matrix; otherwise it is colored in gray. 20 (d1 and d2) or 100 (d3) columns of noise were also added to the data.

(the 20% of the total feature number in a single data matrix) were generated for one matrix, while 20 extra features (the 4%) were generated for the others. This allowed decreasing the uncertainty of the simulated scenarios since the number of noisy features was reduced in two over three matrices. This does not affect the accuracy of the work since three real datasets, whose noise was not under control, were also considered.

Samples have been generated to reproduce the following scenarios, represented in Figure 3.2: A) two groups of 30 samples clearly distinguishable in each data matrix; B) three groups of 20 samples clearly distinguishable in each data matrix; C) three groups of 20 samples, with only one matrix over three generated to distinguish all of them. One group is created to be detectable in all the three matrices; D) two groups over three clearly distinguishable in each matrix, with one of them common to the three matrices. One group is created to be detectable only in one matrix; E) two groups over three clearly distinguishable in each matrix, without a common detectable one. PCA data visualization and correlation analysis can be found in Figures 3.9-3.13 and Tables 3.7-3.11).

3.3 Methods

As input for statistical methods, omics data measured on a common set of n samples, are thought as matrices $X_i, i >= 2$, of dimensions $n \times p_i$, where p_i is the number of features of omics i (for example, the number of genes, proteins, etc.).

Feature selection

To check the effect of pre-processing on integration accuracy, a feature selection step was performed for each dataset. The features showing a coefficient of variation (CV) [113] lower than a selected threshold, were removed from the analysis (see Table 3.3). Coefficient of variation is a standard way to compare variability in different data types since it is independent from the scaling. According to [113] the coefficient of variation has been computed by

$$CV(x) = \frac{sd(x+k)}{mean(x+k)}, \quad k = \begin{cases} |\min(x)|, & \text{if} \min(x) < 0 \\ 0, \text{otherwise} \end{cases}$$

where x represents the values of a feature across the samples without its maximum and minimum value. The sum $x+k$ in the formula makes all feature values positive [57], when the minimum value of a feature is negative. To remove only features carrying less variation, a threshold common to the different data types was selected for each dataset. The threshold for a given dataset was chosen according to the omics-specific distributions of the coefficient of variation. The common CV value describing low variation across all the omics was selected. In some cases (BXD dataset) the selected value does not remove features from all the available omics, however the selection of a greater threshold was observed to cause the deletion of features carrying signals.

Table 3.3 Number of features for each dataset before (Complete) and after (Filtered) feature selection step, with the corresponding threshold on the coefficient of variation (CV) used to filter the data.

Dataset	**Omics**	**Complete (No. features)**	**CV threshold**	**Filtered (No. features)**
BXD	Transcriptomics	21836	0.015	17036
	Proteomics	976		976
	Metabolomic	2607		2607
Platelet	Transcriptomics	54675	0.02	39888
	Proteomics	663		661
	miRNA	490		407
BRCA	Epigenomics	14443	0.2	12474
	Transcriptomics	17814		16419
	miRNA	1010		942

Dataset multi-omics integration

Default parameters were selected to apply integration methods to the data matrices of the selected datasets. For SNF, the parameter σ of the function affinityMatrix was set to 0.5. The number of neighbours for each sample, K, was set to n/C, where C is the number of expected clusters [229], corresponding to the number of real subtypes. In the PMA function MultiCCA, the parameter ncomponents was set to 3 to compute the first three canonical variates that can still show high correlation between features. The parameters "penalty" and "ws" were set respectively to the values "bestpenalties" and "ws.init "previously computed by the function perm.out. Three-omics and pairwise integration for all the different omics couples were then computed for each dataset with the five considered methods, both before and after the feature selection step.

Spectral clustering analysis

After performing integration, similarity matrices have been computed to cluster samples by spectral clustering, a method known to outperform other clustering algorithms[223]. Similarity matrices were obtained in all cases by means of the affinityMatrix function of the SNFtool R package [229], with parameters selected according to those of the SNF method. The function spectralClustering (SNFtool package), was applied to the obtained similarity matrices to perform samples clustering.

Since the functions used to compute samples clustering are included in the SNFtool, the SNF algorithm natively supports this analysis. For the other considered methods, the following preliminary operations have been performed before applying the affinityMatrix function. (i) JIVE: the function was applied to the joint matrix produced by the method. (ii) MCCA: the vectors providing feature weights for each canonical variate are put one after the other to obtain a matrix of dimension n_variates × total_feature_number. The affinityMatrix function was applied to the product of this matrix by the concatenated dataset. (iii-iv) MCIA and MFA: the function was applied to the subject coordinates on the bi-dimensional space computed by the method.

Clustering evaluation

Once clusters were identified, their agreement with real subtypes was computed with the F-score index [216]:

$$Fscore = 2\frac{P*R}{P+R} \in [0,1]$$

$$\text{with } P = \frac{TruePositives}{TruePositives + FalsePositives};\ R = \frac{TruePositives}{TruePositives + FalseNegatives}$$

a standard measure assessing the optimality in binary classifications [104, 188]. Real subtypes were matched with the cluster giving the highest F-score, by avoiding assigning the same cluster to more than one subtype. A 0 F-score was assigned to subtypes for which no cluster has been identified. To assess the performance of the overall classification, both the minimum F-score (worst case) and the averaged F-score were considered.
Although F-score is a more sensible index, also accuracy

$$Acc = \frac{TruePositives + TrueNegatives}{total\ of\ samples}$$

is used in the literature to assess the performance of the methods. Barplots with the resulting averaged accuracies are reported in Figures 3.14 and 3.15.

Tuning parameters with a training/validation procedure

The SNF method has also been applied after performing a training/validation procedure on its parameters, to explore the gain of this procedure on the classification performances. Here the value of SNF parameters σ and K were trained to obtain the highest minimum F-score on 80% of the samples. The trained parameters were then validated on the remaining 20% of subjects. F-scores from integration performances before and after feature selection with default parameters were also computed on the validation set. Due to the small number of subjects (12), this analysis was not applied to the Platelet reactivity dataset.

3.4 Results

3.4.1 Simulated scenarios

Influence of signal strength

Regarding the influence of signal strength, we observed a general decrease in classification performance in all the simulated scenarios when the signal strength across the data types diminished (Figure 3.3, from A to E, light-shaded bars without noise addition). As expected, all methods obtained the highest classification accuracy in scenario A (easiest situation, see Figures 3.2 and 3.9), with averaged F-scores ranging from a minimum of 0.833 to the best value of 1 obtained by MFA. Classification performances decreased step by step in scenarios from B to E, where only SNF was able to distinguish all the sample groups in all the considered scenarios (see Figure 3.3: only SNF has minimum F-score (solid black lines) higher than zero). The method with the worst performance resulted to be JIVE, especially in scenario E where no clear signal was common to the three data matrices and therefore no joint pattern was found by the method.

Influence of noise addition

All methods exhibited a general decrease in performance when noise has been added to the datasets without applying a feature selection step (Figure 3.3, from A to E, light-shaded bars with noise addition). MFA was the method less affected by noise in the simpler scenarios, however, in the most complex case (scenario E), only SNF was still able to distinguish all the sample groups). JIVE resulted to be the method most affected by noise due to its inability to detect common signal in scenarios B, C and E. As discussed in [120], noise can overwhelm the low-rank signal, affecting the permutation testing approach employed by JIVE.

Influence of feature selection

To understand the impact of data pre-processing on multi-omics integration, a preliminary feature selection step was also applied to all the simulated scenarios with and without noise addition (Figure 3.3, from A to E, dark-shaded bars). After feature selection, 25 out of the 50 considered trials did not change F-score; 19 improved and 6 diminished. Feature selection did

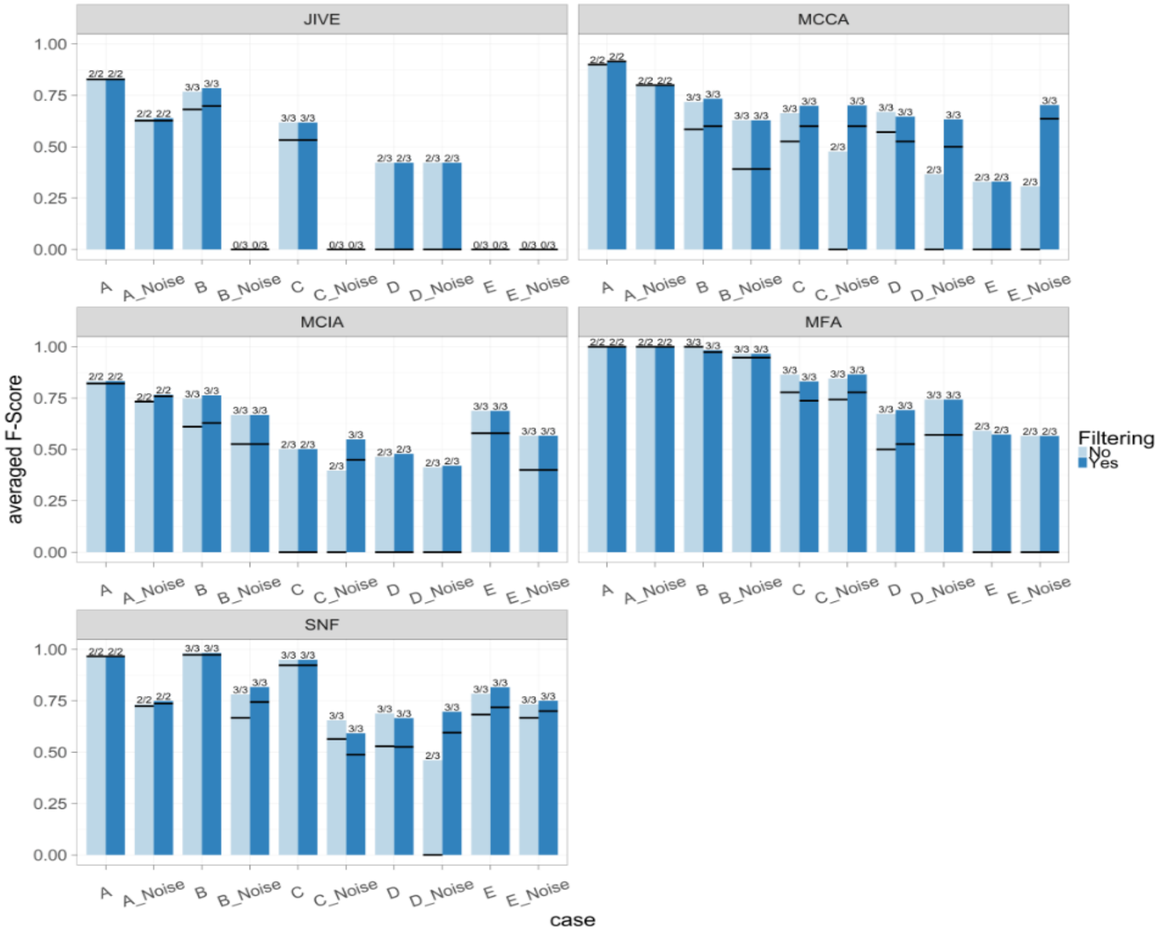


Fig. 3.3 Comparison of the integration methods (JIVE, MCCA, MCIA, MFA, SNF) applied to the simulated scenarios, with and without noise addition. Light and dark-shaded bars represent the averaged F-scores obtained before and after the feature selection step, respectively. The solid black lines represent the minimum F-scores. A minimum F-score equal to 0 means that not all the groups have been recognized. The number of subtypes recognized for each trial is added above the bars.

not improve the accuracy for JIVE, in line with the method description, because its strategy is natively able to separate residual noise in an additional matrix without influencing the joint signal. Conversely, performances of MCCA were the most positively affected by feature selection, with 6 improved trials over 10. This result is in line with the paper by Witten and Tibshirani [239], where a fused lasso penalty has been employed to reduce samples noise before applying MCCA. Although not all the performances benefited from feature sections, the classification accuracy lost by adding noise to the dataset have been generally recovered by applying this pre-processing step.

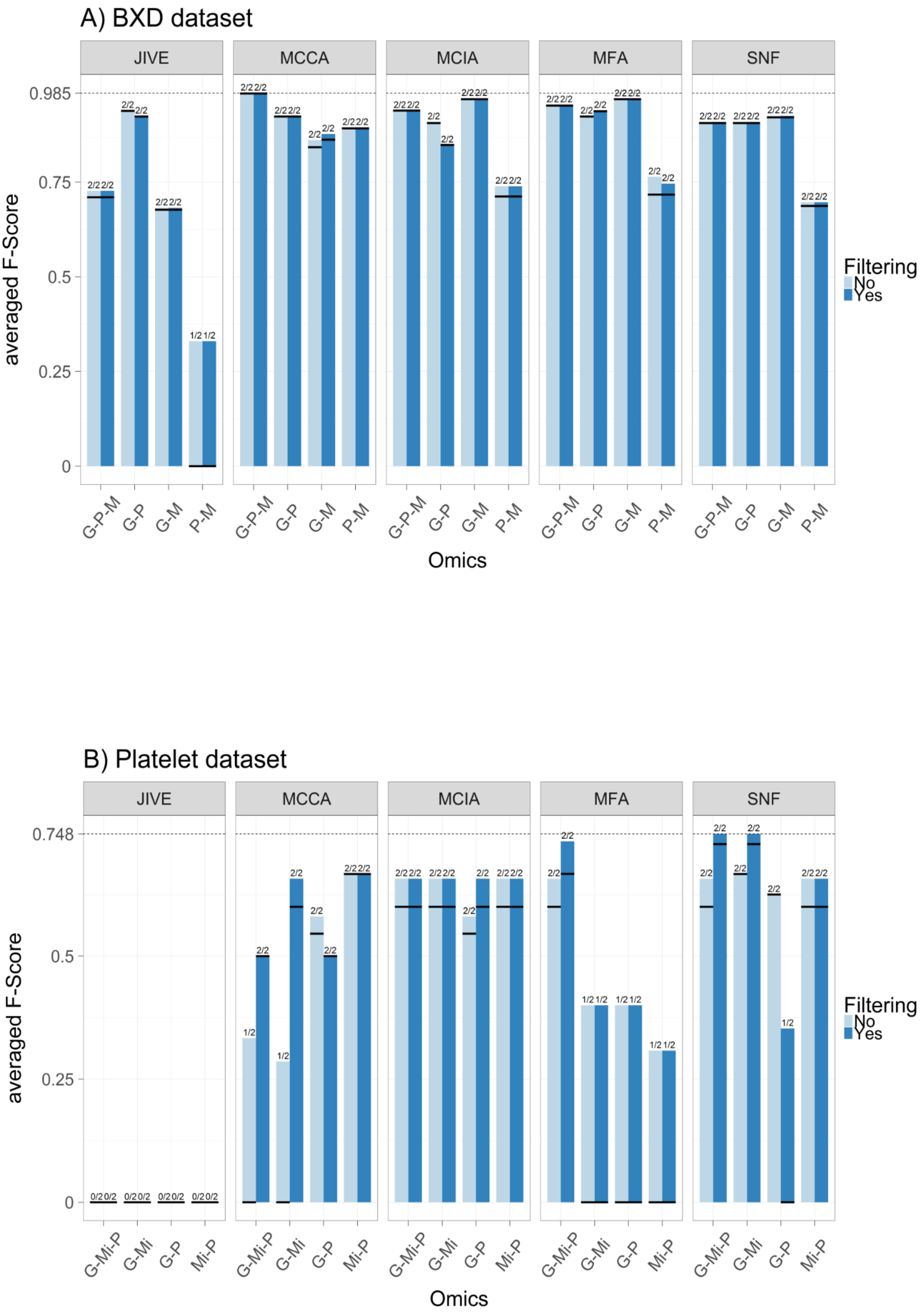
3.4.2 Real datasets

Three-omics integration versus pairwise integration

Methods comparison was repeated on the real datasets described in Table 3.2: the BXD (Figure 3.4A), the Platelet (Figure 3.4B) and the Breast Cancer (Figure 3.4C) datasets. The best classifications were obtained by three-omics integration in all the three datasets even if this result was obtained by different methods (MCCA for BXD, SNF for Platelet and BRCA). This highlights the importance of considering additional omics when possible. Interestingly, we also observed a sort of general agreement on the omics couple more difficult to integrate: proteins and metabolites for the BXD dataset; gene expression and proteins for the Platelet; miRNA and methylation for the BRCA.

Applied to the BXD dataset, three-omics integration allowed a good separation of mice with different diets (Figure 3.4A) with an averaged F-score ranging from a minimum of 0.727 to the best value of 0.985 obtained by MCCA. JIVE obtained the best classification result by considering the omics couple genes-proteins (F-score= 0.939) while SNF, MCIA and MFA by considering genes-metabolites (F-scores of 0.925, 0.97 and 0.97, respectively). The omics couple proteins-metabolites was the most difficult to integrate for all the approaches except for MCCA. This result could be related to the fact that some cross-dimensional correlations have been observed between metabolites and adjacent enzymes in known metabolic pathways (*e.g.* TCA cycle) [238].

SNF resulted the method performing better (Figure 3.4B) in the Platelet dataset. It reached the same F-score of 0.748 both with three-omics integration and with the omics couple genes-miRNAs. SNF obtained the worst performance with the omics couple genes-proteins. This could be explained by the weak Spearman correlation observed between the platelet transcriptome and proteome [123]. MFA obtained the highest F-score of 0.657 for three-omics integration. Except for genes-proteins integration, MCIA always obtained the same F-score of 0.657, while MCCA reached the highest F-score by integrating miRNAs and proteins (F-score= 0.667). JIVE could not recognize common signal in any of the tested omics sets. This indicates that the amount of signal dividing the two extreme phenotypes across the different data types is not strong, in line with what observed by Zufferey *et al.* [256].


A) BXD dataset
JIVE
MCCA
MCIA
MFA
SNF
averaged F-Score
0.985
0.75
0.5
0.25
0
G-P-M
G-P
G-M
P-M
Omics
Filtering
No
Yes
B) Platelet dataset
JIVE
MCCA
MCIA
MFA
SNF
averaged F-Score
0.748
0.5
0.25
0
G-Mi-P
G-Mi
G-P
Mi-P
Omics
Filtering
No
Yes

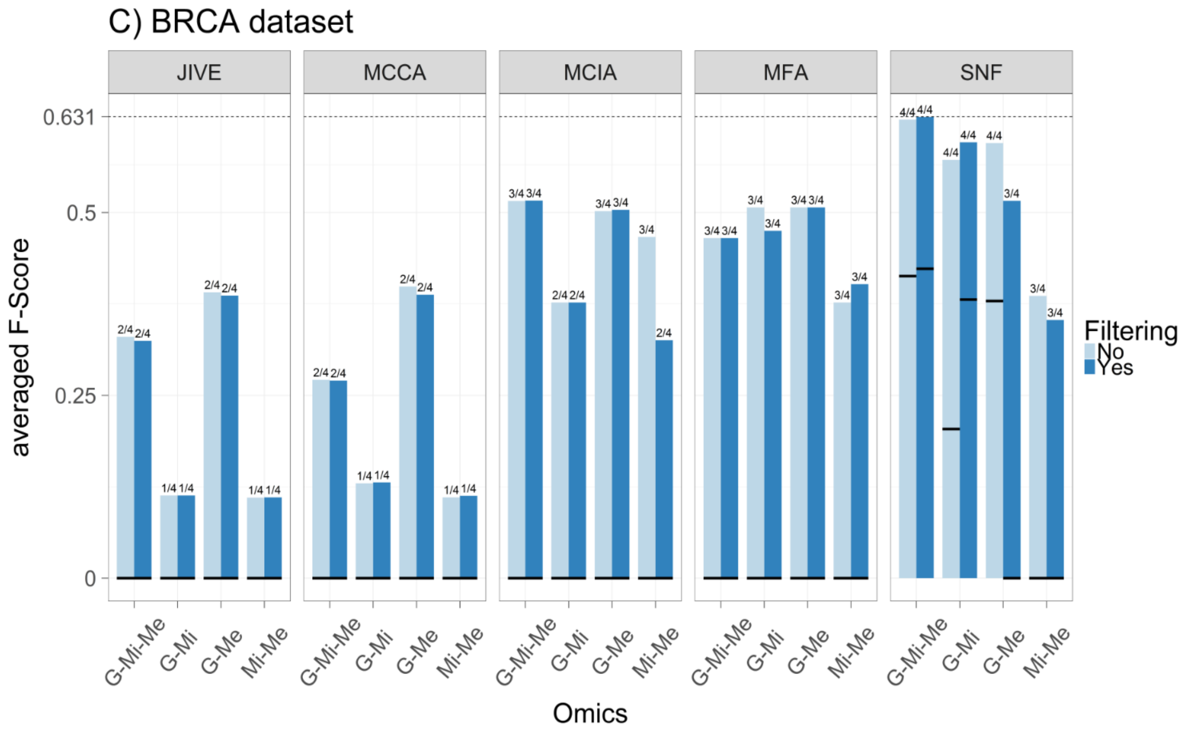


Fig. 3.4 Comparison of the different integration methods (JIVE, MCCA, MCIA, MFA and SNF) applied to the real datasets on all the possible omics combinations (provided in the x-axis). The light and dark-shaded bars represent the averaged F-scores obtained before and after feature selection, respectively. For each method, the first two bars represent the results from three-omics integration. The thick black lines represent the minimum F-score obtained for each trial: a minimum value equal to zero means that not all the subtypes have been recognized. The number of subtypes recognized for each trial is added above the bars. The horizontal dashed lines give the highest F-scores reached for the dataset. **Panel A)** BXD dataset (G: gene expression, P: proteins, M: metabolites. **Panel B)** Platelet dataset (G: gene expression, Mi: miRNA, P: proteins). **Panel C)** BRCA dataset (G: gene expression, Mi: miRNA, Me: methylation).

Differently to previous cases, where a binary classification was required, samples of the BRCA dataet have been classified in four clinical subtypes: Luminal A, Luminal B, Basal-like and HER2-enriched (Figure 3.4C). As expected, the obtained F-scores were generally lower than those of the other studies, confirming the increased level of uncertainty with respect to binary classification.

As for the Platelet dataset, SNF was the method providing the highest performance (best result with three-omics integration, F-score= 0.631). SNF was also the only method able to recognize all the four clinical subtypes. The classification accuracy of MCIA, the second-best method after SNF, was also highest in three-omics integration (averaged F-score of 0.516), but the method failed to recognize the HER2-enriched subtype, which resulted to be the most difficult subtype to recognize, in line with [203]. Also MFA recognized three subtypes (HER2-enriched excluded), but with lower F-scores: the highest (0.507) was obtained for genes-miRNA and genes-methylation integration. The latter case provided the highest, but indeed very low, F-scores for JIVE and MCCA, which could distinguish only Basal-like and Luminal A. The poor result obtained with JIVE could be motivated by the fact that, since the HER2-enriched subtype does not provide a signal common to the three omics, the method could not recognize a shared pattern. This is also in agreement with [163], where JIVE applied on mRNA, methylation and miRNA from another breast cancer dataset, separated Basal-like and Luminal A samples from the others.

Basal-like and Luminal A subtypes were recognized by all the methods, especially with three-omics integration. This agrees with the literature, since the Basal-like subtype is known to be clearly separated from the Luminal one [203].

Influence of feature selection

The effect of feature selection on the real datasets was also evaluated (Figure 3.4 A, B and C, dark-shaded bars). A different threshold for the coefficient of variation was selected for each dataset (see Table 3.3), to reduce data dimensions without losing too much signal. Three-omics integration performances were not diminished by feature selection (BXD dataset, due to the mild filtering which reduced only transcriptomic features), and in some cases were improved by it, as in the Platelet and BRCA datasets (Figure 3.4 B and C). In the latter case, although, F-scores were only slightly improved.

3.4.3 Influence of parameter training

All classification results presented so far have been computed by applying the reviewed methods with default parameters. Here we investigate the gain in classification accuracy

obtained by training parameters according to the training/validation procedure described in Section 3.3. In this analysis, all classification results have been computed by SNF, the method that performed better, on average, in all previous results.
The training procedure was applied to all the datasets (Figure 3.5), both simulated and real, with the exception of the Platelet dataset, where the very limited number of samples prevents the reliability of the analysis.

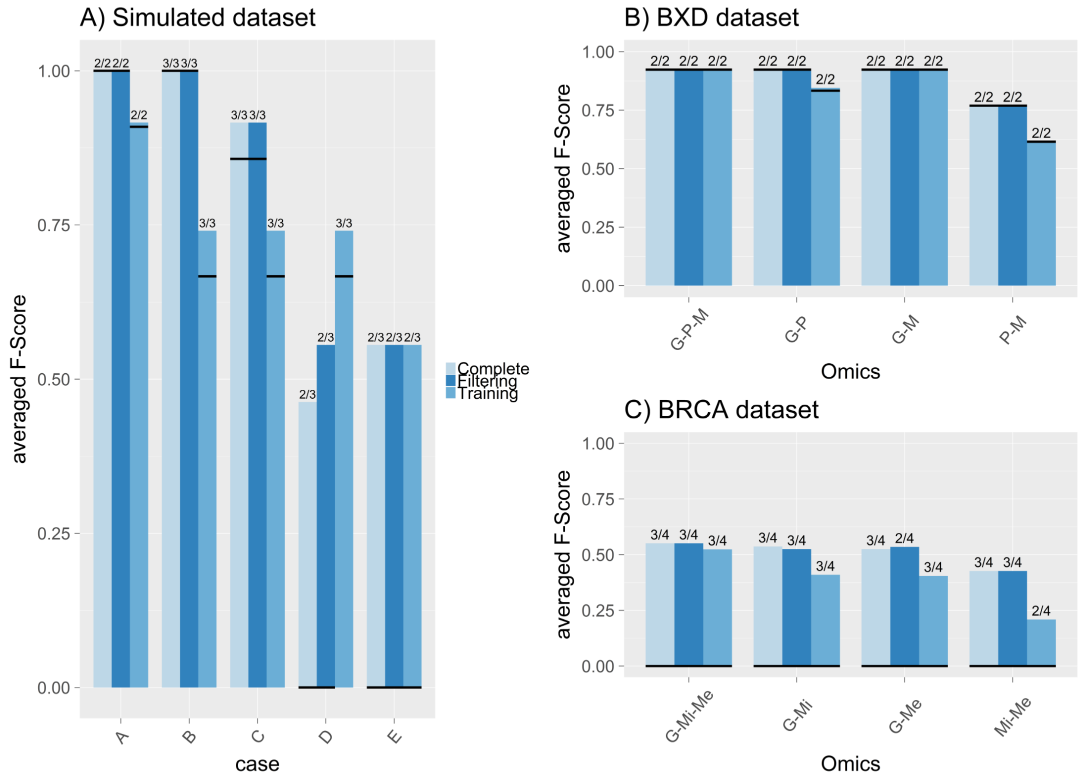


Fig. 3.5 Comparison of SNF results on the validation sets by using default parameters before and after feature selection and by using trained parameters without feature selection. Averaged F-scores obtained from the three analyses are represented with light, dark and medium-shaded bars respectively. Minimum F-scores are represented with black lines. A minimum F-score equal to zero indicates that not all the subtypes have been recognized. The number of subtypes recognized for each trial is added above the bars. **Panel A)** Simulated scenarios. **Panel B)** BXD dataset (G: gene expression, P: proteins, M: metabolites). **Panel C)** BRCA dataset (G: gene expression, Mi: miRNA, Me: methylation).

According to the literature, all classification results refer to the sample subset devoted to validation (20% of the dataset samples). This is an important aspect to consider, because in all the considered scenarios the classification performances obtained by training parameters outperformed the one obtained by using default parameters when the samples included in the training set have been also considered (data not shown). This result, however, has been rarely confirmed in the validation set. Indeed, the training/validation procedure demonstrated an obvious advantage with respect to the standard unsupervised procedure only in simulated scenario D.

Integration with default parameters on the simulated datasets outperformed training/validation in cases A, B and C (Figure 3.5A). An effective gain in training parameters was observed in scenario D, emphasizing the advantage of training parameters when the signal in the dataset becomes weak (see Supplementary Figure 7). Such a result has not been confirmed in scenario E, but this could be motivated by the high complexity of classifying samples when a clear common signal is missing between the three data matrices (see Figure 3.2 and Supplementary Figure 8).

On real datasets, integration with default parameters outperformed training/validation in all cases (Figure 3.5B and C). On the validation set of the BXD dataset (7 CD and 6 HFD samples), three-omics and genes-metabolites integration after parameter training ($K = 12$, $\sigma = 0.59$ and $K = 7$, $\sigma = 0.8$ respectively) reached the same averaged *F-score* (0.923) of integration with default parameters (Figure 3.5B). Training parameters for the other omics couples resulted in lower accuracies. On the validation set of the BRCA dataset (11 HER2-enriched, 45 Luminal A, 24 Luminal B and 18 Basal-like samples), training parameters never improved integration results (Figure 3.5C), but here the analysis could be influenced by the different number of samples of the clinical subtypes, which can affect the estimation of some parameters of the method.

3.4.4 Influence of multiclass classification and experimental design

The results presented so far indicate that data characteristics, such as the number of sample subtypes and the experimental design, could influence multi-omics integration performances.

Multiclass classification

When samples belong to two subtypes (simulated scenario A, BXD and Platelet datasets) all the methods, excluded JIVE for Platelet, identified the two sample groups. This confirms the relative simplicity of recovering information when the signal is given by two phenotypes and strong across all the data. Similarly, the two different diets in the BXD dataset induced a strong signal when single omics were individually assessed for significance. Conversely, multi-omics integration of datasets with multiple subtypes (BRCA dataset and simulated scenarios from B to E) resulted more challenging, with generally low F-scores.

Experimental design

The high classification performances observed for the BXD samples can also be explained in terms of experimental design. In fact, transcriptomics, proteomics and metabolomics were assessed on the same mice livers, with a split-sample study, which Cavill et al. [31] suggested being the best experimental design for multi-omics integration. Also in the Platelet dataset, the omics couple providing the best classification performance was genes-miRNAs and this could be motivated by the fact that both gene expression and miRNAs were assessed from the same RNA. This provides an important advantage with respect to the way in which proteomics data have been obtained: proteins were quantified with different preparations and three technical replicates for each patient. Moreover, they were separated in two groups, thus presenting a batch effect on sampling timing that needed to be corrected, and which could have negatively influenced proteomics integration with the other omics.

3.5 Discussion

Five multi omics integration methods, representative of multivariate, concatenation-based and transformation-based approaches, were selected for comparison of their ability to integrate more than two omics data in unsupervised way.

In general, our analysis showed that the integration of three different omics results in better sample classification than pairwise omics integration. This demonstrates that the additional knowledge brought by considering multiple omics data at once is essential to increase the understanding of the mechanisms underlying the characteristics of sample subtypes. Furthermore, F-scores obtained with SNF, the transformation-based method, were the highest in 9 of the 22 trials considered and among the highest in the other cases. MFA performed the best in 6 trials (simulated datasets), MCIA and MCCA in 3 and JIVE only in one. Additionally, by also considering the accuracy index (Figures 3.14 and 3.15), SNF demonstrated to be the best method when the dataset complexity increases.

In addition, the comparisons revealed that the outcome of multi-omics integration is data-dependent and influenced by the experimental design as suggested by Cavill et al. [31]. This could thus warrant some preliminary examination of the data at hand to determine the appropriate integration method to use. Omics data should be separately analysed and visualized (*e.g.* with PCA) to quantify how much signal is carried by each omics and how much of it is shared across omics types. Recently, Ciucci *et al.* [35] proposed an algorithm able to detect the optimal normalization method to be applied and the most discriminative dimensions. In cases where PCA is not powerful enough to segregate samples, more advanced techniques (such as Minimum Curvilinear Embedding [26, 27]) based on non-linear dimension reduction can be tested to inspect each omics data type at a time [3]. Computing correlation between the omics could also help in evaluating the strength of inter-omics relationships.

If preliminary analysis reveals shared signals across data (simulated scenario A), a method like JIVE able to separate noise and to provide the common pattern in an already computed matrix could be a good choice. The multivariate method MCCA could instead reinforce visible intra-omics signal, when no evident inter-omics signal (necessary for JIVE) is present (simulated scenarios C, D). Since MCCA is correlation-based, it could also be applied when datasets show well correlated features across omics: it obtained the most precise sample classification for the BXD dataset, where 25% of transcript-proteins pairs correlated significantly in the CD subtypes (P-value< 0.05), 137 of those with Spearman's $\rho = |0.65|$ ([238]).

For more complex cases, like simulated scenario E (multiple subtypes and noisy dataset), SNF, MCIA and MFA, methods based on subjects' similarities and dissimilarities, can be better options. Indeed, these methods recognized all the subtypes demonstrating their ability to recover not only shared but also complementary signals across omics. Moreover, the data transformation step applied by SNF succeeded in distinguishing all the tumour subtypes of the BRCA dataset, including the HER2-enriched (weakest signal), while MCIA and MFA could distinguish three of the four subtypes.

The method comparison performed in this Chapter highlighted the strengths and the weaknesses of the selected methods.

As said above, the concatenation method JIVE is, for instance, able to provide the common signal across omics data in a computed matrix and to separate noise from the data signal. Thus JIVE resulted useful to integrate omics sharing common patterns (such as genes and proteins in the BXD dataset). However, in general, this integration method performed worse than the others. This is probably due to the fact that PCA is used for matrix factorization, which makes the method suffering from the presence of outliers [79]. Moreover, JIVE can deal only with Gaussian distributed data. Similar concatenation methods present in literature, such as iCluster [191] are instead able to integrate also binary and sequential data, but they require a strong pre-selection of features critical for clustering, which is not a necessary step for JIVE.

Methods like the multivariate MCCA and the concatenation-based MCIA do not factorize the input matrices using Principal Component Analysis. To reduce the dimension of the problem they rely on other approaches namely canonical correlation and co-inertia. With respect to

JIVE, those methods were shown to be more robust to noise addition and complexity of the datasets. Despite MCCA obtained the highest performance in classifying the samples of the BXD dataset, this approach is based on the level of correlation across the omics data: low levels of correlation would not make it a good choice. In particular, when dealing with more than two omics data types, only some of them could show high correlation. This makes difficult for MCCA to gain good three-omics classification: for the real Platelet and BRCA datasets, three-omics MCCA integration obtained lower precision than some pairwise integration.

The method comparison presented in this Chapter revealed that MCIA is able to recover not only shared but also complementary signals across omics. Indeed, the strength of MCIA relies on the computation of the similarity of sample profiles in the new low-dimensional space. The MCIA optimization of the covariance among the omics data types, instead of the optimization of the correlation, allows to obtain more precise information of complex datasets. It is important to notice that other approaches, based on covariance optimization, have been recently proposed. One of those method is MINT [178] which however needs to select a response matrix among the integrated ones. However, with respect to MCIA, it considers a penalty term in the optimization problem. This additional term, missing in MCIA could decrease the unwanted systematic variation (*e.g.* batch effect) that can be present when integrating data coming from different platforms.

The concatenation method MFA is also based on Principal Component Analysis, which is used to weight the single omics matrices prior to integration. The method comparison shows that this use of the PCA is better than the one considered for JIVE. Moreover, MFA obtained results comparable to MCIA and MCCA in the simulated scenarios, as well as in the BXD and BRCA real datasets. This is due to the main characteristic of the method: providing a higher weight to data in the most informative matrices, for example those sharing more information with the others. Concerning SNF, it resulted the overall best method in the comparison. However, it was not always able to correctly classify the samples. This is probably due to the parameter selection: default parameters were chosen to perform all the omics integration. A more specific choice would have improved the overall classification. Moreover, one of the SNF parameters is the number of sample neighbours, important for clustering. Cases with subtypes showing great differences in the number of samples may suffer from the wrong choice of this parameter. Instead, the transformation of the matrices in graphs and the use of an information passing algorithms are the main strengths of the method: they makes possible to reinforce weak, but important, signals across omics. SNF was indeed the only method able to recognize all the subtypes in both the most complex simulated dataset (case E) and in the BRCA real dataset. Interestingly, the strengths of

SNF and MFA (transforming the input matrices in intermediate forms and providing them a weight) are used in combination by methods mostly based on multiple kernel learning. A recent example is proposed in [129] where, after transforming omics data in kernels, these are weighed to obtain the best consensus kernel. Weights provided to the kernels are instead optimized to preserve the topology of the data in the feature space, making the method well designed to integrate also sparse omics data.

Another aspect that we studied was feature selection. Feature selection is a popular pre-processing step and, according to our analysis, it can be useful to integrate omics not showing a strongly shared signal. However, the thresholds for feature selection need to be selected carefully. Features can carry signals not detected in single omics analysis but that can make the difference when more omics types are integrated. To be able to define a unique threshold of low variability across data types, we used a general method to filter out noise. This can be substituted by more specific methods considering the data and the problem at hand (*e.g.* supervised/unsupervised) or whether the relationships among features should be considered important while filtering (see [74] for a review of feature selection methods).

3.6 Conclusion

The addition of biological knowledge obtained by considering multiple molecular levels (omics) to the analysis increases the knowledge extracted from the available data, in the present case, sample classification precision. Simultaneous omics integration should thus be considered in future studies with more omics data available. Noise was also shown to influence integration results; an effect that can be mitigated by adding a feature selection step before proceeding with data integration. This is especially recommended when dealing with complex design (such as those having more than two different omics data, or with low signal strength, or multiple cellular subtypes). However, we believe that statistical integration methods could still be improved, for example by adding a priori information about relationships between the different omics data, which could diminish false positive results, while enhancing the relevance of true molecular interactions.

3.7 Supplementary material

Dataset exploration

To have a better insight into the real and the simulated datasets used for the multi-omics data integration methods comparison, we visualized the PCA of the single data types with the analytical tool PC_corr [35] (Figures 3.6-3.8).

To obtain a better sample segregation, this tool suggests the best normalization methodology to be applied, together with the most discriminative component. In each of the figures representing the single omics visualization, black colour is always assigned to the sample group more to the left, red colour is assigned to the class more on the right.

Additionally, for each dataset we computed Spearman correlation among all the possible data pairs and their corresponding False Discovery Rate (FDR) with the function corr.test (R package psych [174]). Real omics datasets were previously filtered due to their high dimensionality by retaining only features with variance within the fourth quartile. This allowed the correlation analysis to be also performed on the real datasets. The number of features from real datasets with Spearman's $|\rho| > 0.5$ are listed in Tables 3.4-3.6 of this section.

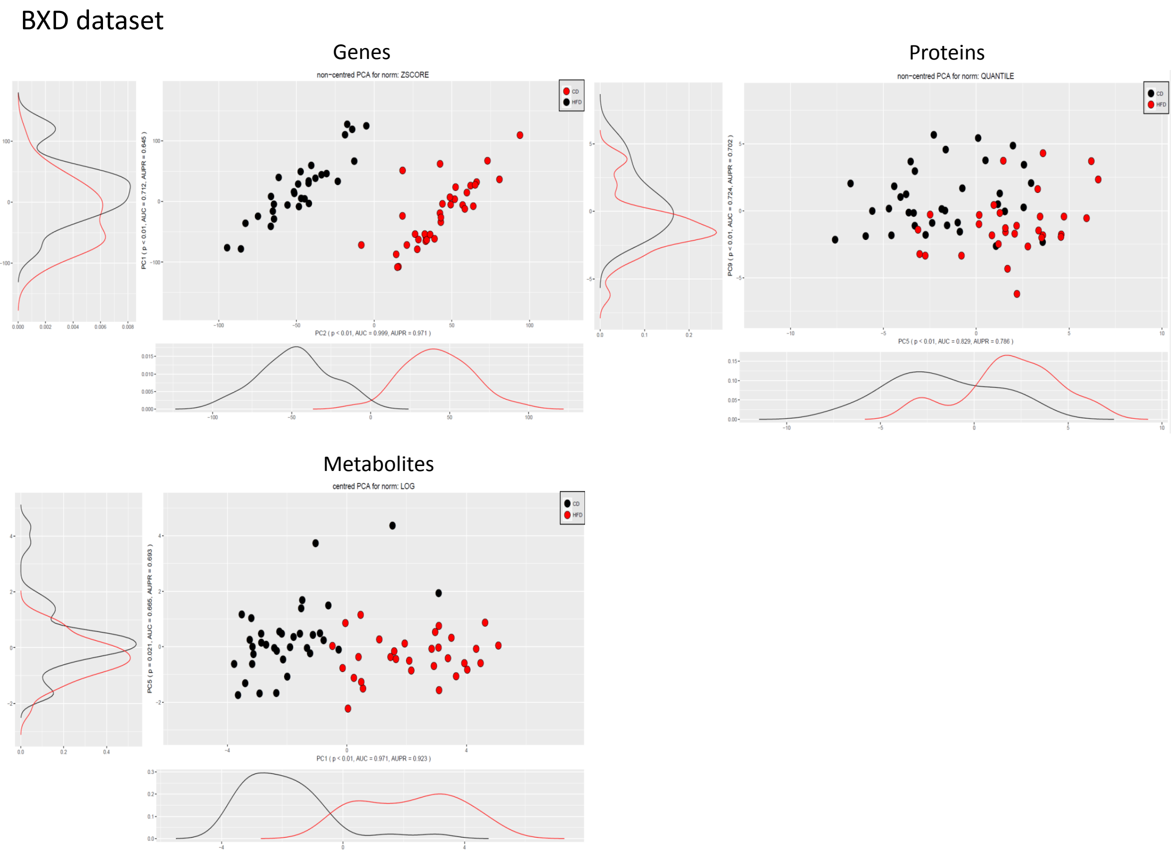


Fig. 3.6 PCA visualization of the omics data composing the BXD dataset. Each block represents a different omics type: transcriptomic, proteomics and metabolomics. Dots represent the 66 mice samples with colors representing the Chow Fat Diet (CFD, 35) and the High Fat Diet (HFD, 31) subtypes. The normalization methods selected are respectively ZSCORE, QUANTILE and LOG

Table 3.4 Spearman correlation among different omics in the BXD dataset. The percentage of pairs with Spearman's $|\rho| > 0.5$ with respect to the total number of correlated pairs is shown in parenthesis.

Omics (No. considered features)	**No. of total Features pairs**	**No. pairs with $|\rho| > 0.5$**	**No. pairs with FDR < 0.05**
Genes (5459) - Metabolites (244)	1331996	26249 (1.97 %)	26249 (100 %)
Genes (5459) – Proteins (652)	3559268	10625 (0.3 %)	10625 (100 %)
Proteins (652) – Metabolites (244)	159088	255 (0.16 %)	255 (100 %)

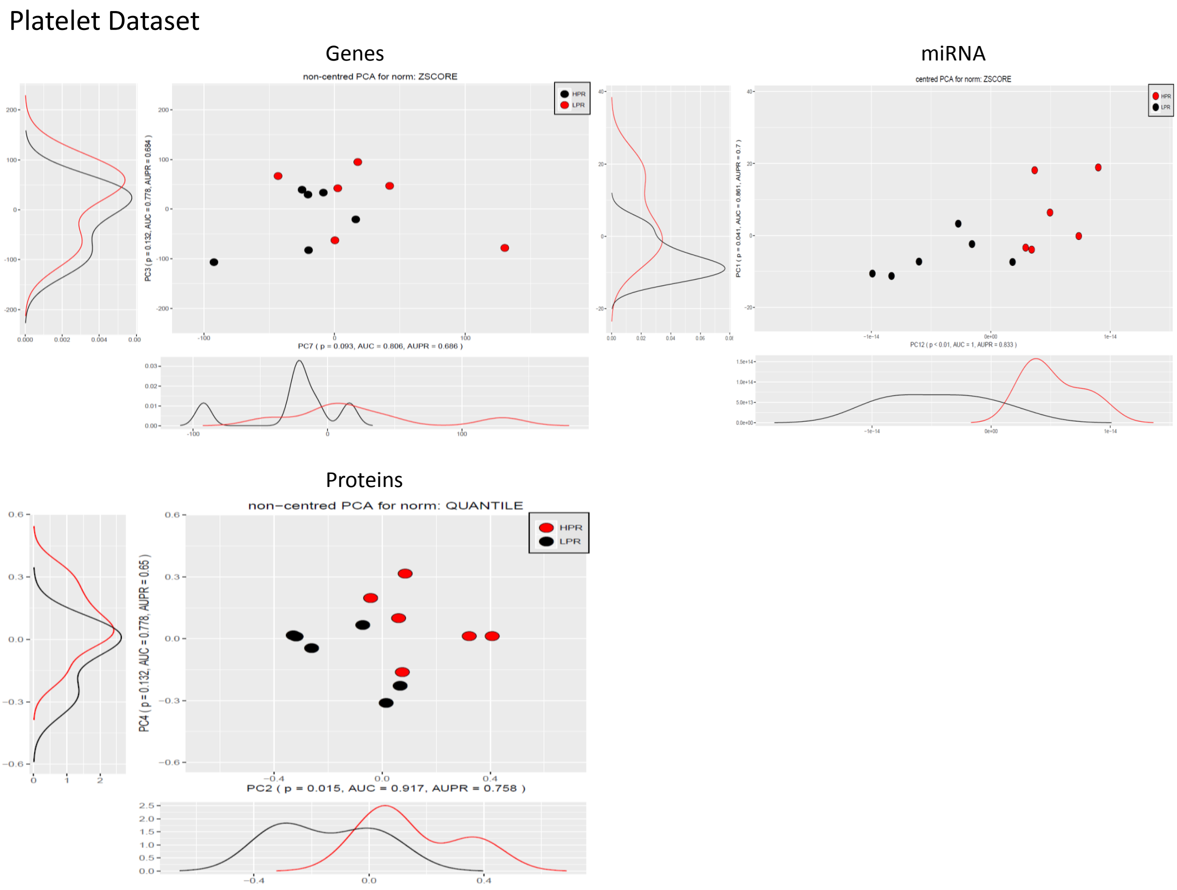


Fig. 3.7 PCA visualization of the omics data composing the Platelet dataset. Each block represents a different omics type: transcriptomics, miRNA and proteomics. Points correspond to the high platelet reactivity patients (HPR) and low platelet reactivity patients (LPR). The selected normalization method is displayed on the top (ZSCORE, ZSCORE and QUANTILE respectively).

Table 3.5 Spearman correlation among different omics in the Platelet dataset. The percentage of pairs with Spearman's $|\rho| > 0.5$ with respect to the total number of correlated pairs is shown in parenthesis.

Omics (No. considered features)	**No. of total Features pairs**	**No. pairs with $|\rho| > 0.5$**	**No. pairs with FDR < 0.05**
Genes (13669) – miRNA (123)	1681287	189345 (11.24 %)	0
Genes (13669) - Proteins (166)	2269054	221657 (9.75 %)	3 (0.001 %)
Proteins (166) – miRNA (123)	20418	2314 (11.33 %)	2 (0.086 %)

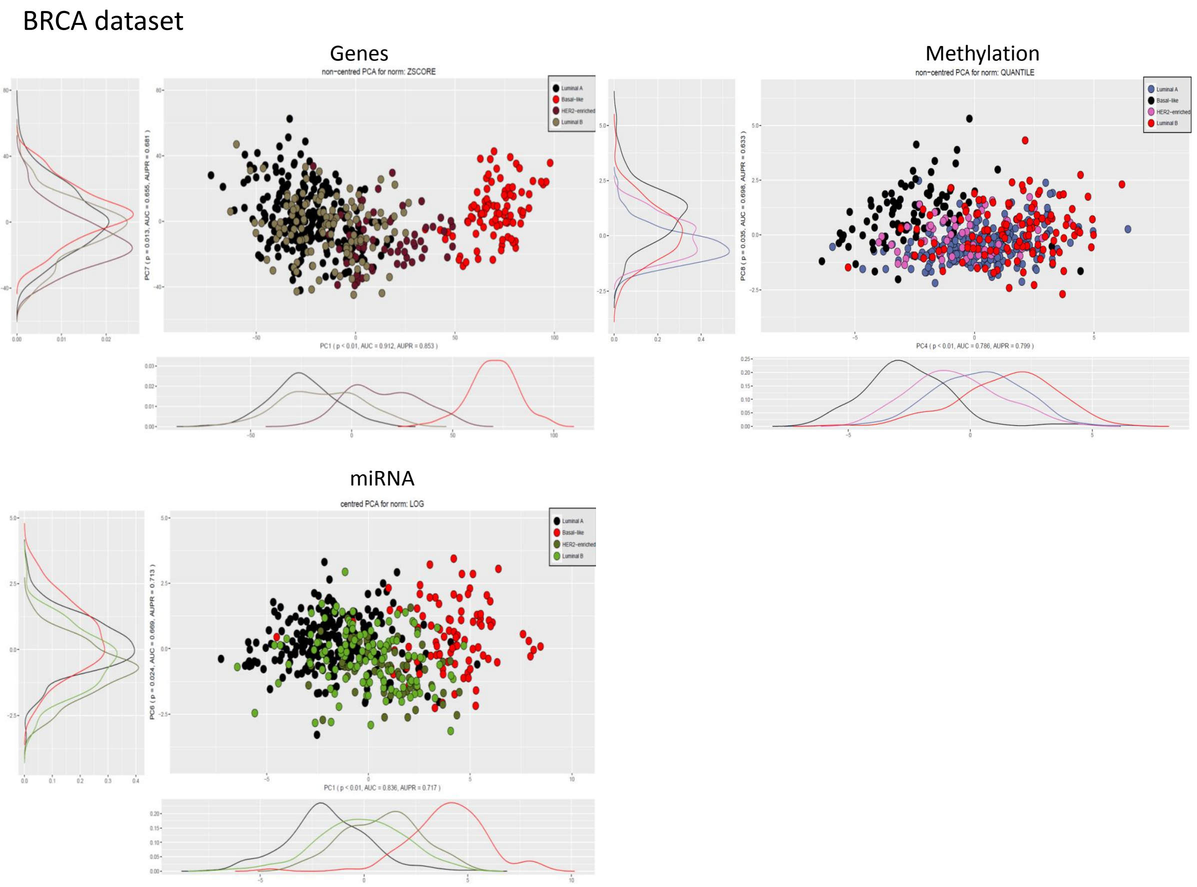


Fig. 3.8 PCA visualization of the omics data in the BRCA dataset. Each block represents a different omics type: gene expression, methylation and miRNA. Dots represent the 491 patients and are coloured according to their subtype. The normalization methods used to find the best separation are ZSCORE, QUANTILE and LOG.

Table 3.6 Spearman correlation among different omics in the BRCA dataset. The percentage of pairs with Spearman's $|\rho| > 0.5$ with respect to the total number of correlated pairs is shown in parenthesis.

Omics (No. considered features)	**No. of total Features pairs**	**No. pairs with $\|\rho\| > 0.5$**	**No. pairs with FDR < 0.05**
Genes (4454) – Methylation (3611)	16083394	15395 (0.096 %)	15395 (100 %)
Genes (4454) – miRNA (253)	1126862	2895 (0.26 %)	2895 (100 %)
miRNA (253) – Methylation (3611)	913583	391 (0.04 %)	391 (100 %)

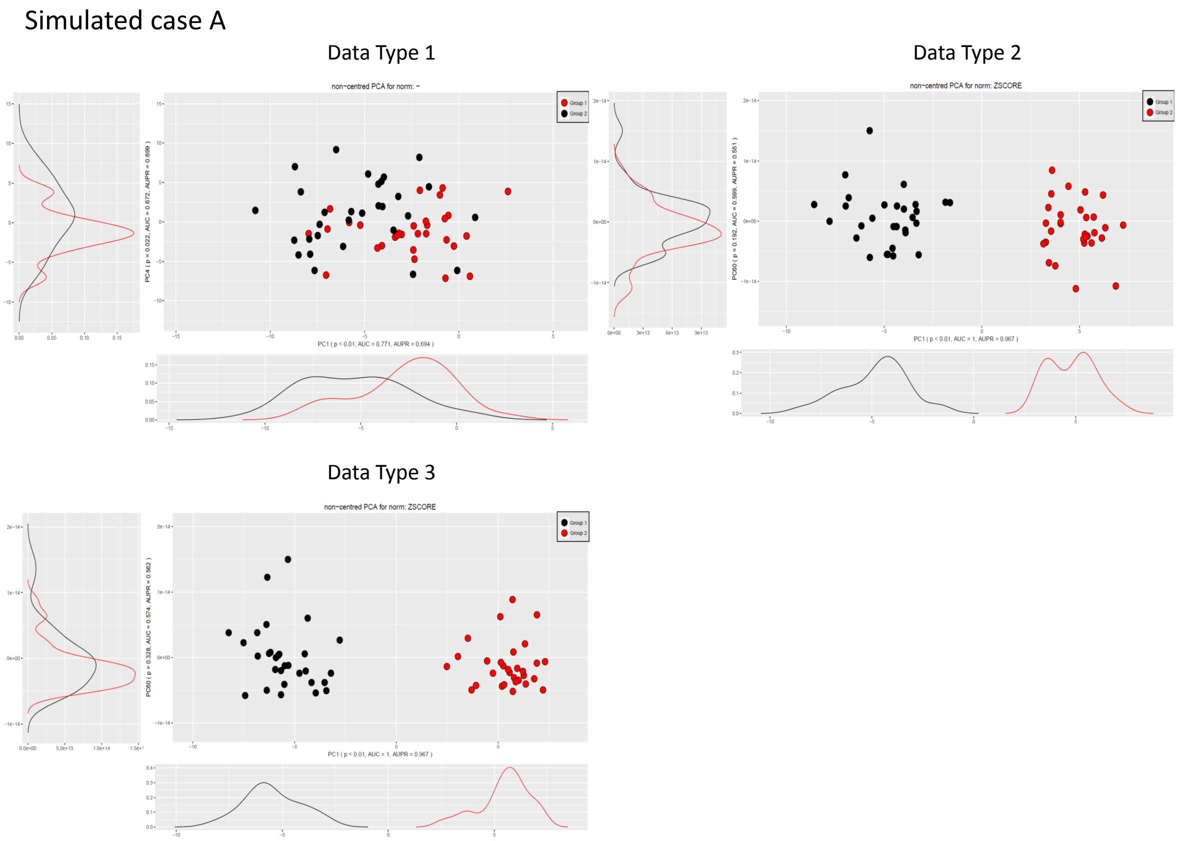


Fig. 3.9 PCA visualization of the data types generated for the simulated scenario A. Blocks represent different data types. Dots represent the 60 generated samples, which were divided in two groups (of 30 samples), coloured in black and red. Only data type 2 and 3 were normalized, both with ZSCORE.

Table 3.7 Spearman correlation among different omics in the simualted scenario A. The percentage of pairs with Spearman's $|\rho| > 0.5$ with respect to the total number of correlated pairs is shown in parenthesis.

Data (No. considered features)	**No. of total Features pairs**	**No. pairs with $\|\rho\| > 0.5$**	**No. pairs with FDR < 0.05**
Data Type 1 (500) - Data Type 2 (500)	250000	17 (0.007 %)	0
Data Type 1 (500) - Data Type 3 (500)	250000	405 (0.162 %)	405 (100 %)
Data Type 2 (500) - Data Type 3 (500)	250000	36 (0.014 %)	0

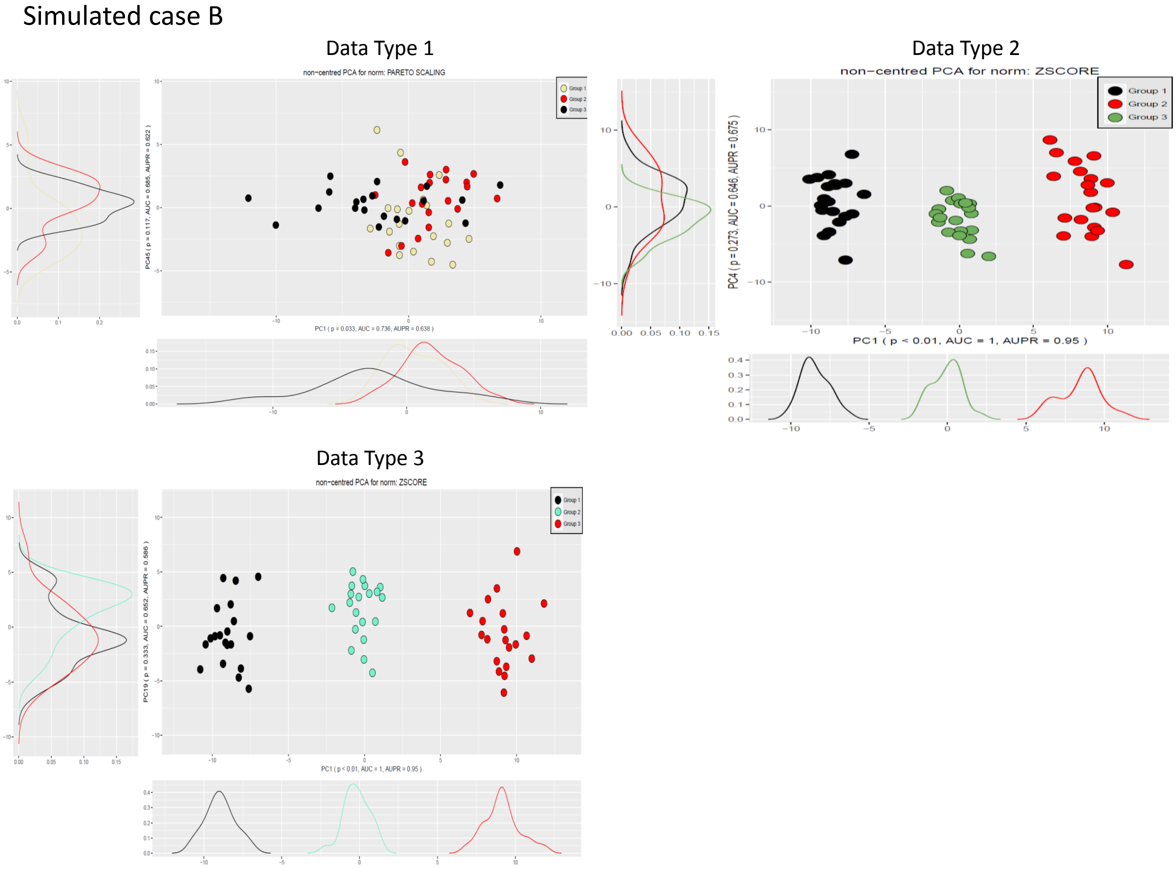


Fig. 3.10 PCA visualization of the data types generated for the simulated scenario B. Blocks represent PCA for the different data types. Dots represent the 60 generated samples, that were divided in three groups, and represented by different colours. Normalization methods selected were PARETO SCALING for data type 1 and ZSCORE for the others.

Table 3.8 Spearman correlation among different omics in the BXD dataset. The percentage of pairs with Spearman's $|\rho| > 0.5$ with respect to the total number of correlated pairs is shown in parenthesis.

Data (No. considered features)	**No. of total Features pairs**	**No. pairs with $|\rho| > 0.5$**	**No. pairs with FDR < 0.05**
Data Type 1 (500) - Data Type 2 (500)	250000	94 (0.038 %)	67 (71.3 %)
Data Type 1 (500) - Data Type 3 (500)	250000	230 (0.092 %)	228 (99.1 %)
Data Type 2 (500) - Data Type 3 (500)	250000	213 (0.085 %)	208 (97.7 %)

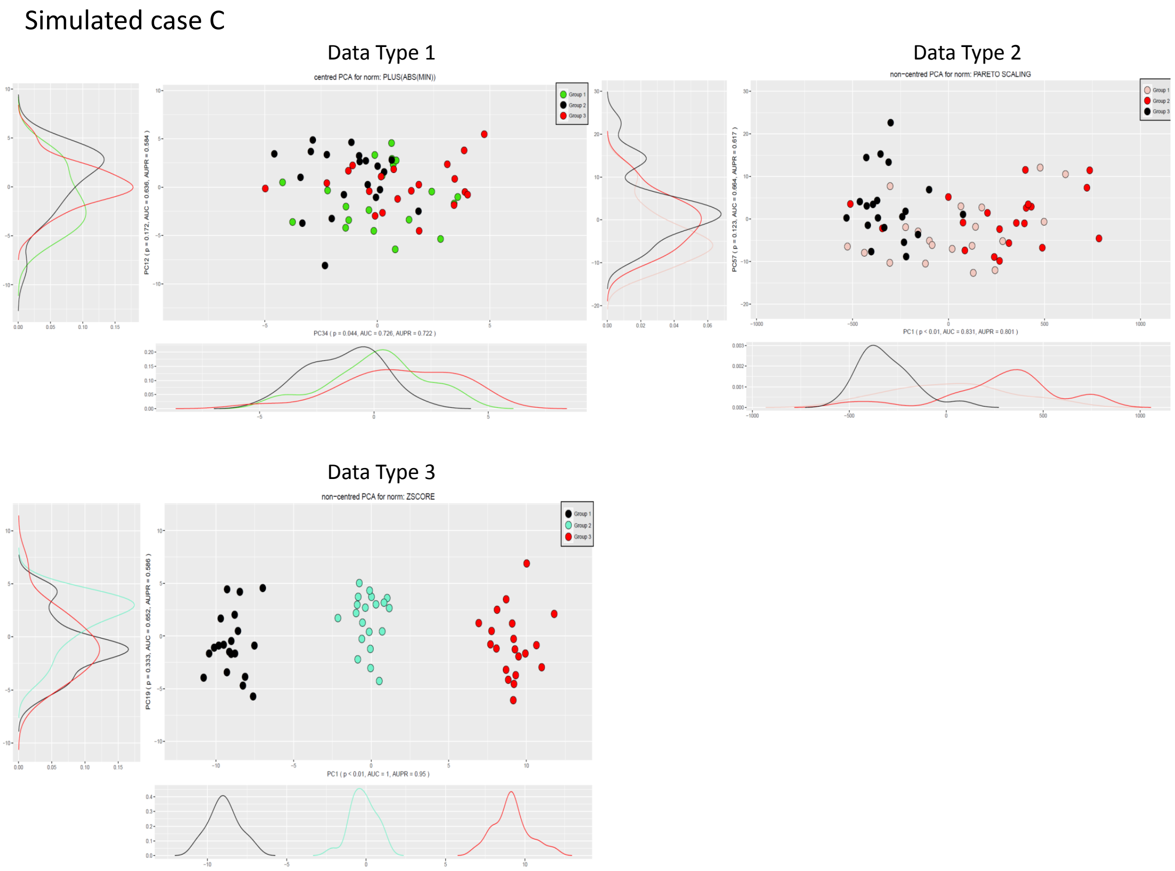


Fig. 3.11 PCA visualization of the data types generated for the simulated scenario C. The PCA of the different data types are shown in different blocks. The 60 generated samples are represented by different colours. PLUS(ABS(MIN)), PARETO SCALING and ZSCORE were used to normalize the data.

Table 3.9 Spearman correlation among different omics in the simulated scenario C. The percentage of pairs with Spearman's $|\rho| > 0.5$ with respect to the total number of correlated pairs is shown in parenthesis.

Data (No. considered features)	**No. of total Features pairs**	**No. pairs with $|\rho| > 0.5$**	**No. pairs with FDR < 0.05**
Data Type 1 (500) - Data Type 2 (500)	250000	92 (0.037 %)	72 (78.3 %)
Data Type 1 (500) - Data Type 3 (500)	250000	10 (0.004 %)	0
Data Type 2 (500) - Data Type 3 (500)	250000	23 (0.009 %)	0

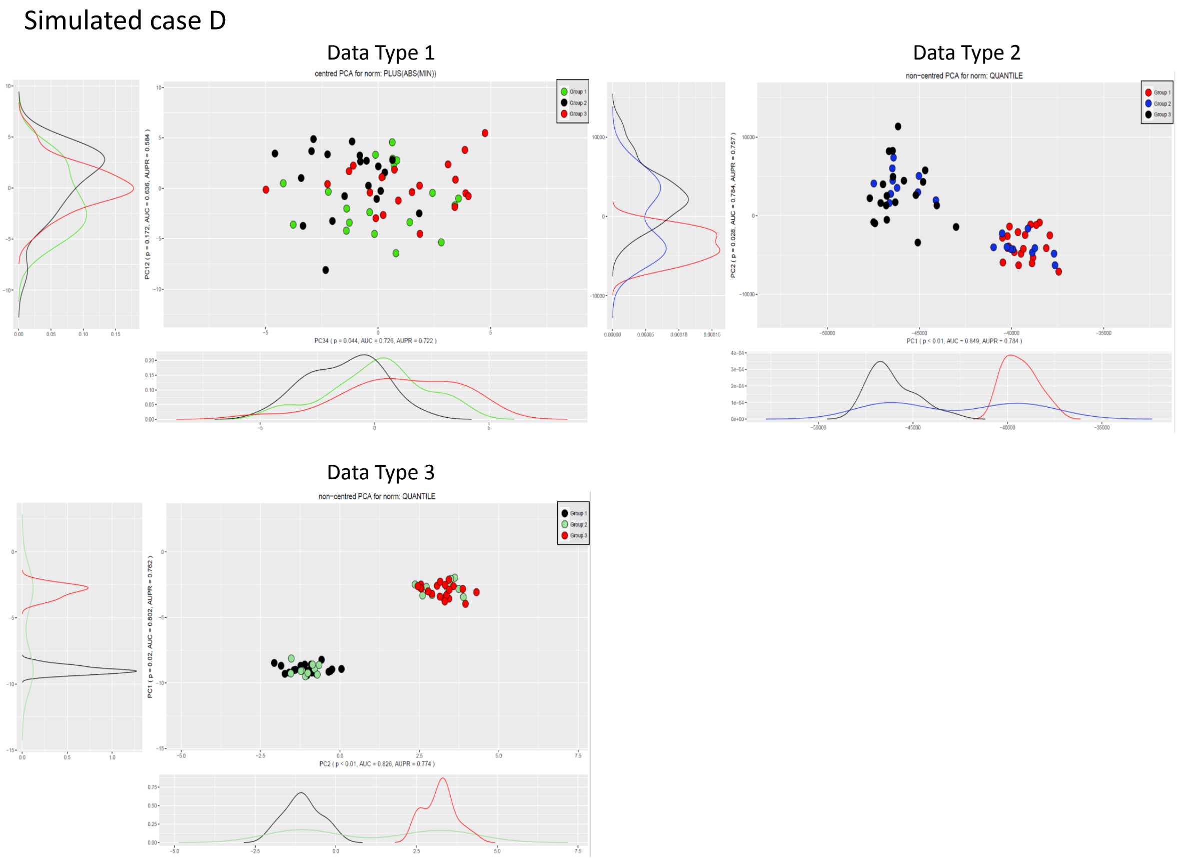


Fig. 3.12 PCA visualization of the data types generated for the simulated scenario D. Blocks represent different data types while dots represent the 60 generated samples. They were generated in three groups, and coloured accordingly. PLUS(ABS(MIN)) was used to normalize data type 1, QUANTILE was used in the other cases.

Table 3.10 Spearman correlation among different omics in the simulated scenario D. The percentage of pairs with Spearman's $|\rho| > 0.5$ with respect to the total number of correlated pairs is shown in parenthesis.

Data (No. considered features)	**No. of total Features pairs**	**No. pairs with $\|\rho\| > 0.5$**	**No. pairs with FDR < 0.05**
Data Type 1 (500) - Data Type 2 (500)	250000	163 (0.065 %)	153 (93.9 %)
Data Type 1 (500) - Data Type 3 (500)	250000	225 (0.09 %)	223 (99.1 %)
Data Type 2 (500) - Data Type 3 (500)	250000	207 (0.083 %)	205 (99 %)

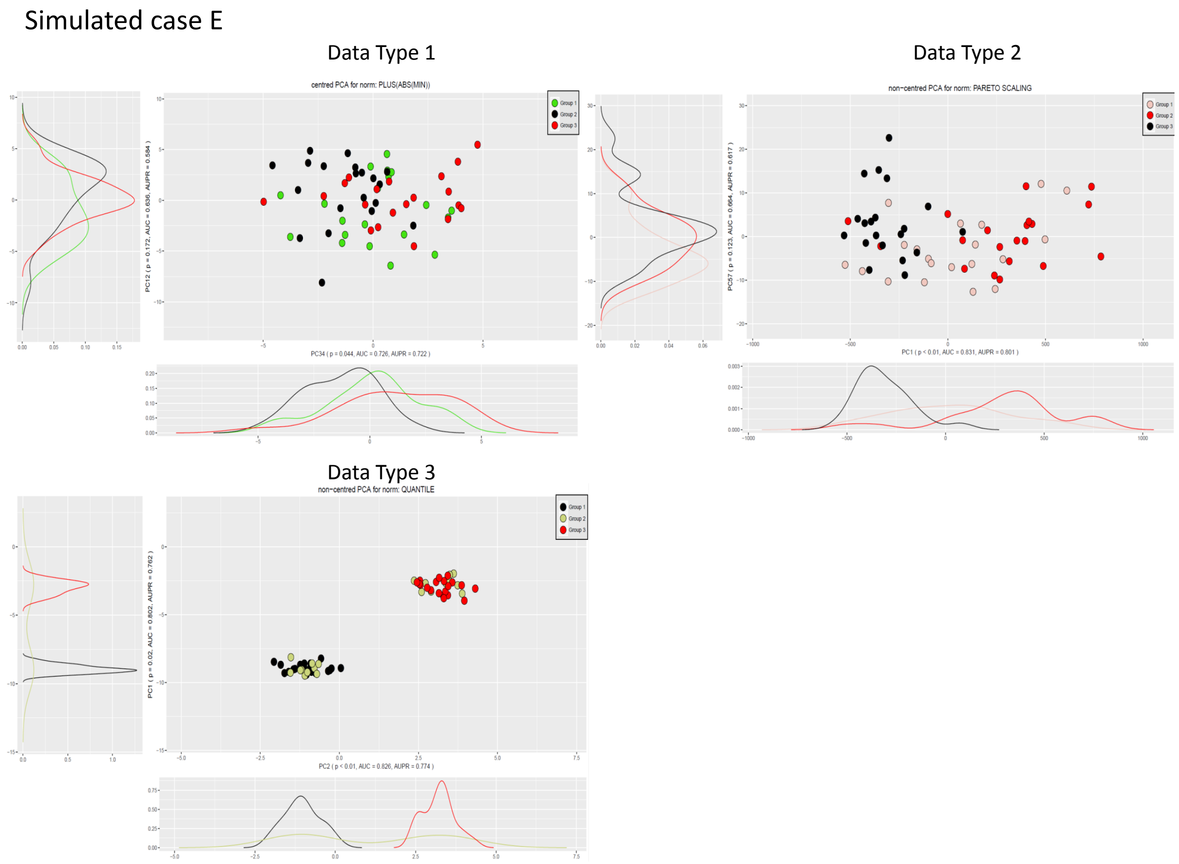


Fig. 3.13 PCA visualization of the data types generated for the simulated scenario E. Blocks represent PCA of the different data types. The 60 generated samples were divided into three groups, and are represented by differently colored dots. Normalization methods selected from PC_corr were respectively PLUS(ABS(MIN)), PARETO SCALING and QUANTILE.

Table 3.11 Spearman correlation among different omics in the simulated scenario E. The percentage of pairs with Spearman's $|\rho| > 0.5$ with respect to the total number of correlated pairs is shown in parenthesis.

Data (No. considered features)	**No. of total Features pairs**	**No. pairs with $|\rho| > 0.5$**	**No. pairs with FDR < 0.05**
Data Type 1 (500) - Data Type 2 (500)	250000	92 (0.037 %)	72 (78.3 %)
Data Type 1 (500) - Data Type 3 (500)	250000	225 (0.09 %)	223 (99.1 %)
Data Type 2 (500) - Data Type 3 (500)	250000	207 (0.083 %)	204 (98.6 %)

Accuracy results

To better characterize the obtained results, in this section we provide barplots showing the averaged accuracy of the different methods (Figures 3.14 and 3.15). Similarly to F-score, a null accuracy was assigned to classes for which no cluster has been identified, as stated in the Method section (3.3).

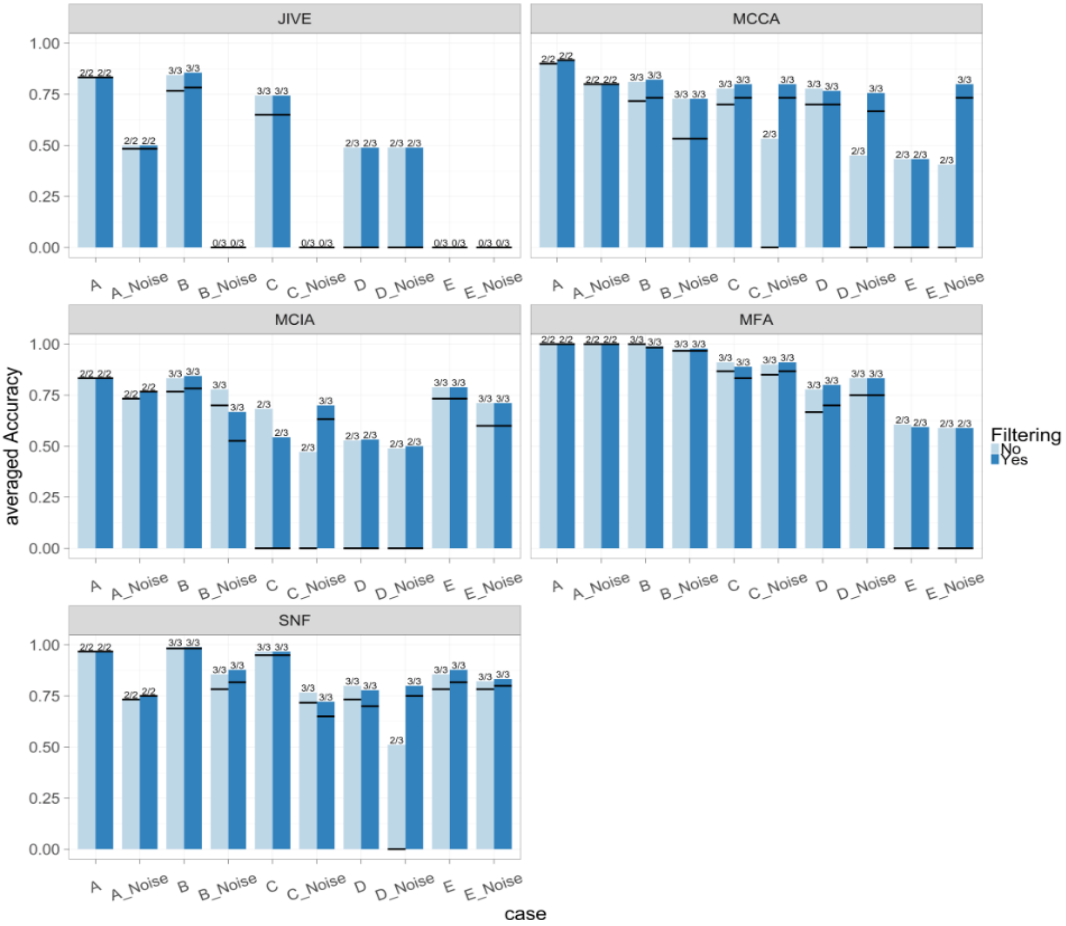
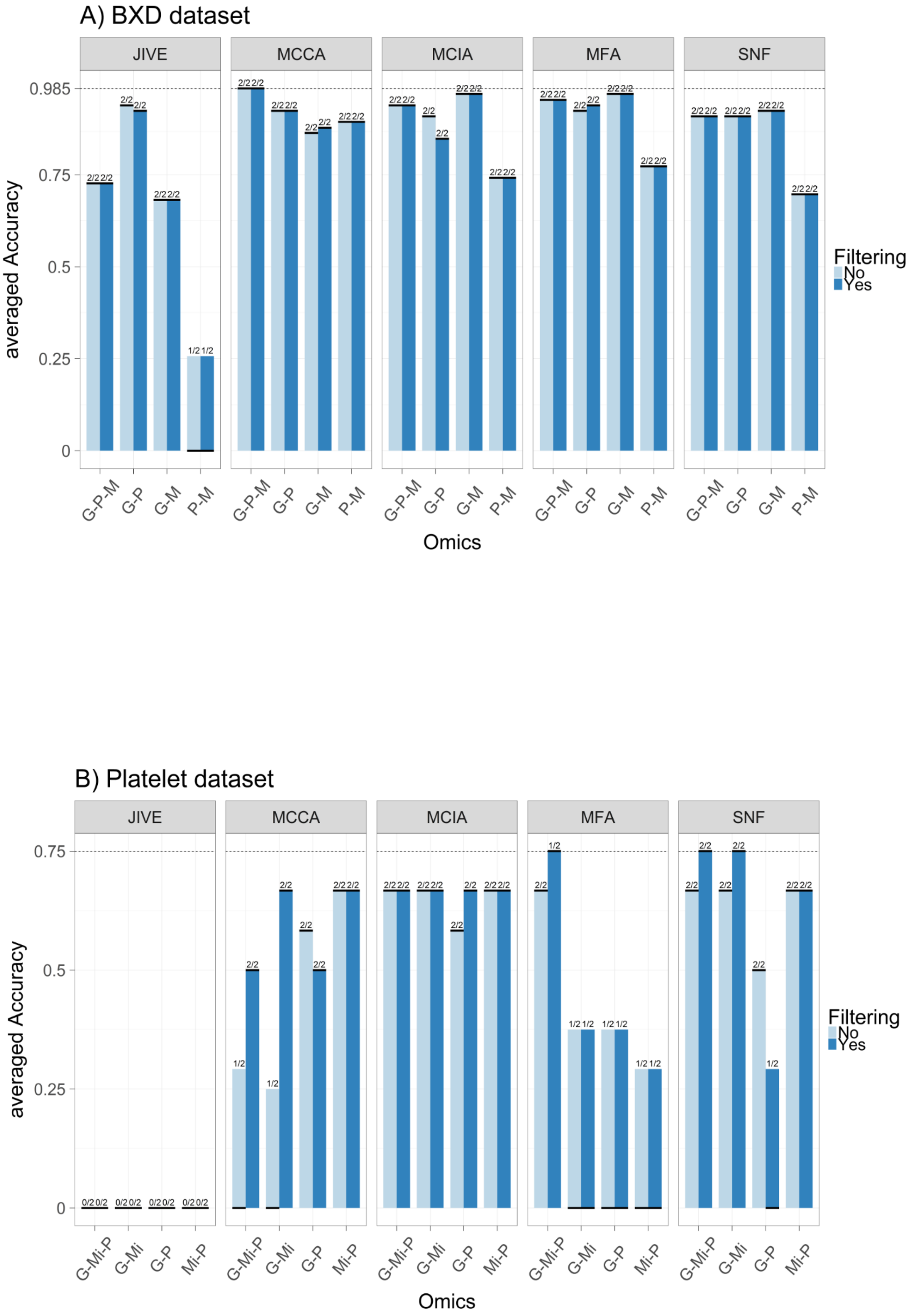


Fig. 3.14 Comparison of accuracy of the integration methods (JIVE, MCCA, MCIA, MFA, SNF) applied to the simulated scenarios, with and without noise addition. Light and dark-shaded bars represent the averaged accuracies obtained before and after the feature selection step, respectively. The solid black lines represent the minimum accuracy. A minimum accuracy equal to 0 means that not all the groups have been recognized. The number of subtypes recognized for each trial is added above the bars.


A) BXD dataset
JIVE
MCCA
MCIA
MFA
SNF
averaged Accuracy
0.985
0.75
0.5
0.25
0
2/2 2/2
2/2 2/2
2/2 2/2
1/2 1/2
G-P-M
G-P
G-M
P-M
Omics
Filtering
No
Yes
B) Platelet dataset
JIVE
MCCA
MCIA
MFA
SNF
averaged Accuracy
0.75
0.5
0.25
0
0/2 0/2
1/2 2/2
2/2 1/2
G-Mi-P
G-Mi
G-P
Mi-P
Omics
Filtering
No
Yes

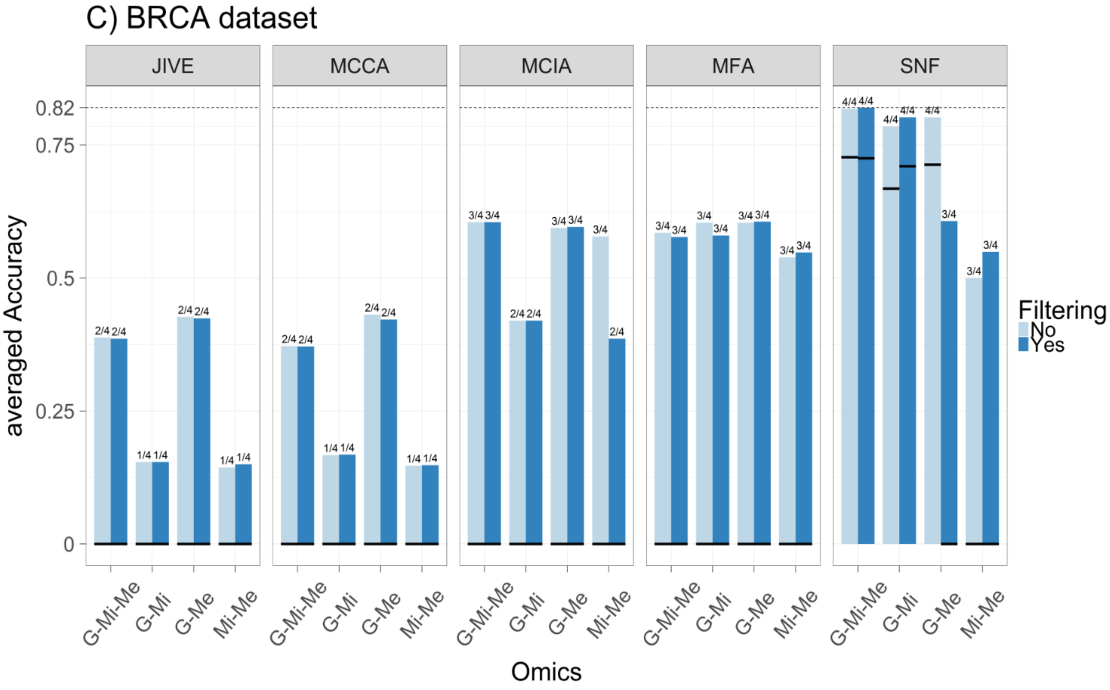


Fig. 3.15 Comparison of accuracy of the integration methods (JIVE, MCCA, MCIA, MFA and SNF) applied to the real datasets on all the possible omics combinations (provided in the x-axis). The light and dark-shaded bars represent the averaged accuracy obtained before and after feature selection, respectively. For each method, the first two bars represent the results from three-omics integration. The thick black lines represent the minimum accuracy obtained for each trial: a minimum value equal to zero means that not all the subtypes have been recognized. The number of subtypes recognized for each trial is added above the bars. The horizontal dashed lines give the highest accuracy reached for the dataset. **Panel A)** BXD dataset (G: gene expression, P: proteins, M: metabolites. **Panel B)** Platelet dataset (G: gene expression, Mi: miRNA, P: proteins). **Panel C)** BRCA dataset (G: gene expression, Mi: miRNA, Me: methylation).

Chapter 4

Addition of prior knowledge for multi-omics sample classification

To take advantage of the main strengths of the approaches discussed in the previous chapters those methodologies can be joined in a unique framework. On one hand, the use of prior knowledge from linear supervised integration, is beneficial to include information from recognized biological interactions. Unsupervised simultaneous integration, on the other hand, allows to inform on unknown omics relations, that would be missed otherwise.

Following these considerations, we propose in this Chapter a strategy aimed at simultaneously integrate multi-omics data after retrieving information on inter-omics interactions. Building on the ability of networks to contain high-dimensional information, a prior-knowledge network is generated. Links between the features (*e.g.* genes, proteins, metabolites) composing the datasets are derived by a sparse multivariate regression model, which is able to compute the probabilities of association among variables of two different data types. The algorithm that obtained best results for sample classification in Chapter 3, SNF, is then used to integrate the data. We address the problem of correctly classify sample subtypes, to be able to assess the accuracy of the obtained results by comparing them with the real sample labels.

Moreover, since we previously showed that addition of molecular layers provides more accurate classifications, we focus on datasets composed by more than two omics data. The results obtained in this Chapter on simulated datasets indicate that the inclusion of inter-omics relationships can improve the classification accuracy of unsupervised simultaneous methods, which simultaneously combine omics data without having access to the sample labels.

The content of this Chapter has been developed during my internship at the Nestlé Institute of Health Science (NIHS) in Lausanne, under the supervision and help of Dr. Jorg Hager and Hélène Ruffieux.

4.1 Introduction

The emergence of a disease is often due to a complex interplay of different molecular layers (for instance, genome, transcriptome, proteome) [176]. In the last decades, several studies [34, 110] have started to incorporate in their models prior biological knowledge, such as known protein-protein interactions or assessed molecular pathways information to increase accuracy and reliability. With prior biological knowledge, the predictive power of gene expression data was increased with respect to classical studies [95] and this enhanced the understanding of the mechanisms of the studied process [176], for instance the presence of different disease subtypes. As suggested in the recent work by Boluki *et al.*, [15], the integration of prior knowledge with omics data is critical to classify samples. Genes and proteins connected to diseases with similar phenotypes are more likely to interact with each other [61]: without considering these interactions, analysis would yield biased results and lead to a fragmented picture of the sample classification [257]. The relevance of prior biological knowledge inclusion is especially useful for classification of subtypes with weak signal, such as those composed by a low number of samples [15].

Prior knowledge inclusion is also used in multi-omics integration. For example, Multiple Factor Analysis (MFA) [166], described in Appendix A, was extended to include information external to the data [44]. To ease the interpretation of the performed analysis, Gene Ontology terms [205] are used to assemble modules of genes, which are then superimposed on the low-dimensional space obtained by the standard algorithm. Applied to copy-number measurements and gene expression from brain cancer datasets, the method found reliable markers for glioma diagnostic. The multi-omics integration algorithm proposed by Kim *et al.* [96] extends instead a previous approach focused only on gene expression [95]. Its aim is to search for meta-dimensional knowledge-driven genomic interactions (MKGIs) associated with clinical outcomes in cancer. Omics are transformed in pathway-based datasets, composed of patients-pathways matrices, which are then used as input for grammatical neural networks [150]. Integrated knowledge-based model for clinical response prediction are finally extracted. Applied to copy number alteration, methylation and gene expression data from ovarian cancer, the method found pathways associated with cancer prognosis and with the potential development of therapies targeting the genes in the pathways [96].

Prior knowledge inclusion could help to focus on important dataset features (such as genes, proteins and other molecules) and on their relationships during the analysis [79], thus reducing the complexity of the problem at hand. It can moreover reduce the presence of noise and the number of false positive results. Additionally, inclusion of inter-omics relationships can simplify the interpretation of the integration results by retaining only findings that are biologically meaningful. Nevertheless, some issues have to be overcome to include prior knowledge in multi-omics integration algorithms, such as: i) differences in the type of information given by measurement and knowledge (quantitative and qualitative, respectively); ii) amount of data added to already high-dimensional problems; iii) reliability of the source considered to retrieve prior knowledge.

Networks-based approaches provide a natural way to solve the first two issues. Information related to inter-omics interactions can be collected in a prior-knowledge network, and successively, investigated and included into multi-omics integration models. To describe the inter-omics connections, nodes of the network can represent the features of the datasets, while edges can be weighted according to the strength of the relationship. This step allows to transform the qualitative nature of prior knowledge to quantitative information.

We previously demonstrated that integration is highly data-dependent [207]. Therefore, to solve the problem of source reliability, relationships between features can be inferred from the data themselves. Interactions coming from databases can in fact be specific to a certain disease, not always reflecting the population/species/tissue under study [257]. Data-driven information can be retrieved by multivariate regression models. Among those model, *locus*, a sparse multivariate regression model, allows simultaneous selection of predictors and associated responses belonging to two different omics data [181]. The output of *locus* are the posterior probabilities of association γ_{st}, which describe the likelihood that two features s and t are associated. One of the advantages of this method compared to other similar methodologies is that it avoids sampling by implementing a deterministic variational inference strategy. Moreover, regression and variance parameters are exchangeable: features considered predictors and those used as responses have the same prior probability to be involved in the associations [181] computed by the method.Applied to a metabolite quantitative trait locus (mQTL) analysis, it successfully recovered both known and new SNPs-metabolite associations [181].

In this Chapter we thus propose a strategy to improve multi-omics sample classification by including information about inter-omics interactions in the analysis. The pipeline is based on the SNF method [229], indicated as the best one for clustering [207], and includes the sparse multivariate regression model *locus* to integrate the prior knowledge coming from inter-omics relationships. This prior knowledge is then used to identify biomolecules more related to the phenotype under study. The pipeline was tested on simulated datasets and results show that including in the analysis features more associated with the phenotype has a positive effect on classification. To obtain more precise classification, we also studied the impact on integration of tuning parameters used by the SNF algorithm.

4.2 Methods

The pipeline for multi-omics sample classification consists in two main steps: i) the investigation of inter-omics associations (see section 4.2.1); ii) the inclusion of the computed inter-omics relationships in the three-omics integration, to find the features more related to the phenotype at hand (section 4.2.2). Figure 4.1 provides an overview of the general pipeline.

The first step of the pipeline uses a multivariate statistical method to find pairwise omics interactions. Those are then used to build a prior-knowledge network (see section 4.2.1). In the second step, features are weighted according to their importance in the network. The qualitative nature of prior knowledge is thus transformed in a quantitative value. The measurements of each feature are multiplied by the correspondent weight. The unsupervised Similarity Network Fusion (SNF) method is finally applied to the weighted datasets to retrieve subtype classification.

4.2.1 Prior knowledge computation

Multivariate and linear pairwise integration

Let's consider a dataset $\mathbf{X} = (X_1, X_2, X_3)$ composed of three different data types X_i. The sparse multivariate regression model *locus* [181] was applied to all the possible data couples (that is, (X_1,X_2), (X_1,X_3) and (X_2,X_3)). Three matrices $\Gamma^i = \left\{\gamma_{st}{}^i\right\},\ i = 1,2,3$ were obtained, where $\gamma_{st}{}^i$ represents the posterior probabilities of association between features s and t, which belong to two different data types.

To obtain more reliable feature interactions we computed False Discovery Rates (FDR). FDRs also allow to avoid too many false positive associations. As pointed out by Efron [55], False Discovery Rates (firstly introduced by Benjamini and Hochberg in 1995 [8]), represent an important tool to solve the large-scale hypothesis testing situations, intensified by the high-throughput technologies development. The Bayesian interpretation of the False Discovery Rate proposed by Efron [55] and recalled in [181], was used here.

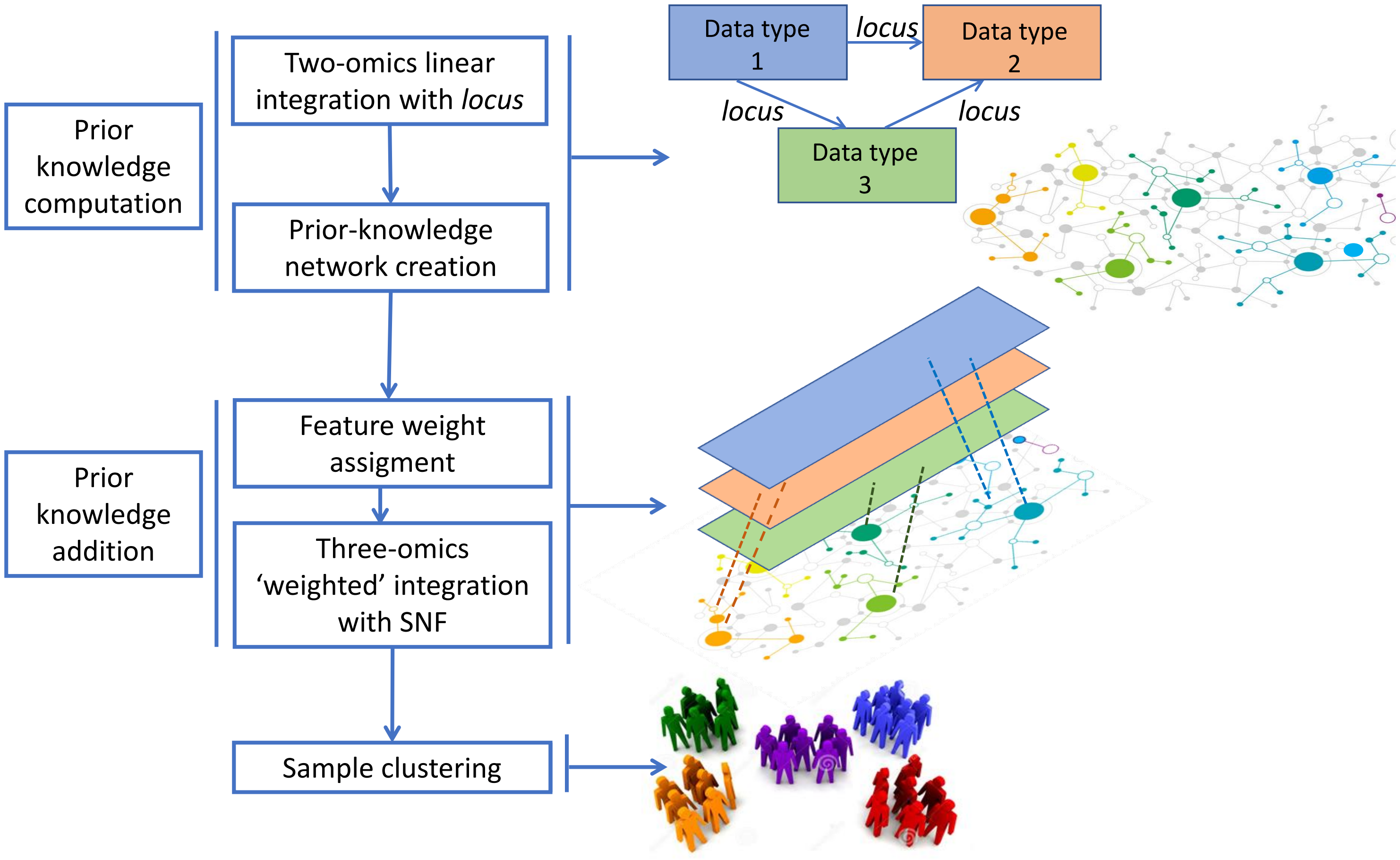

Prior knowledge computation
Two-omics linear integration with *locus*
Prior-knowledge network creation
Data type 1
Data type 2
Data type 3
locus
locus
locus
Prior knowledge addition
Feature weight assigment
Three-omics 'weighted' integration with SNF
Sample clustering

Fig. 4.1 Graphical representation of the pipeline performed to include prior knowledge into unsupervised multi-omics integration. The pipeline is composed by two main steps: i) prior knowledge computation and ii) prior knowledge addition. The former is accomplished by computing pairwise probabilities, with the multivariate method *locus*. A network containing this information is then generated. A weight is assigned to each node (feature) according to its importance in the network. The weighted omics data are then integrated with the SNF method, which provides sample classification.

On this basis, for each matrix Γ^i, we computed empirical False Discovery Rates. Given a threshold $\tau \in [0,1]$, the FDR takes into account the likelihood that associations with posterior probability ${\gamma_{st}}^i \geq \tau$ are not active:

$$FDR(\tau) = \frac{\sum_{s,t}(1-{\gamma_{st}}^i)\mathbb{1}\left\{{\gamma_{st}}^i \geq \tau\right\}}{\sum_{s,t}\mathbb{1}\left\{{\gamma_{st}}^i \geq \tau\right\}}, \quad i = 1,2,3 \tag{4.1}$$

where

$$\mathbb{1}\left\{{\gamma_{st}}^i \geq \tau\right\} = \begin{cases} 1 \text{ if } {\gamma_{st}}^i \geq \tau, & 0 \leq \tau \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

Given a False Discovery Rate value T (for example 5%), the formula in 4.1 can be computed for a grid of probabilities $\tau \in [0,1]$, in order to select the $\bar{\tau}_i$ which provides only associations with $FDR(\bar{\tau}_i) \leq T$.

However, the dimensions of the different data types could influence the selection of the different value of $\bar{\tau}_i \;\; i = 1,2,3$. If the same grid of τ values is used, a significant larger number of links giving the FDR value T will come from the highest-dimension data types. Links obtained by other data would be penalized. To avoid these calibration issues, for each couple of features s' and t' belonging to two different data types, we computed:

$$FDR\left({\gamma_{s't'}}^i\right) = \frac{\sum_{s,t}(1-{\gamma_{st}}^i)\mathbb{1}\left\{{\gamma_{st}}^i \geq {\gamma_{s't'}}^i\right\}}{\sum_{s,t}\mathbb{1}\left\{{\gamma_{st}}^i \geq {\gamma_{s't'}}^i\right\}}, \quad i = 1,2,3 \tag{4.2}$$

where

$$\mathbb{1}\left\{{\gamma_{st}}^i \geq {\gamma_{s't'}}^i\right\} = \begin{cases} 1 \text{ if } {\gamma_{st}}^i \geq {\gamma_{s't'}}^i \\ 0 \text{ otherwise} \end{cases}$$

The threshold term τ in Equation 4.1 is substituted by each single probability of posterior association $\gamma_{s't'}$. It is thus possible to consider each feature couple independently from all the others. The term $\mathbb{1}\left\{{\gamma_{st}}^i \geq {\gamma_{s't'}}^i\right\}$ in equation 4.2 counts how many associations ${\gamma_{st}}^i$ are more likely to exist with respect to the one under study, ${\gamma_{s't'}}^i$. The term $(1-{\gamma_{st}}^i)$ provides the probability that the two elements s and t are not connected. In this way, for each matrix $\Gamma^i = \left\{{\gamma_{st}}^i\right\}$, False Discovery Rate computation is tailored on its own elements. This still allows the selection of a final general FDR value T, but the posterior probabilities considered do not suffer from the effect of data dimension.

Creation of the prior-knowledge network

Once the $FDR\left(\gamma_{s't'}{}^{i}\right)$ have been computed for all the data type couples, it is possible to generate a network $G = (V,E)$. Nodes in V represent the features, while the edges in E describe the connections between features coming from different data types. As pointed out in Chapter 1, an edge can be weighted according to the importance of the connection it represents. In this case, the weight w_{st} of the link connecting the features s and t is given by $w_{st} = \dfrac{1}{FDR(\gamma_{st}{}^{i})}$. In this way, associations with a low False Discovery Rate (more reliable) are considered more important in the network.
The network $G = (V,E)$, generated following the rules above, contains information about inter-omics connections, and can be considered as another layer of knowledge to be integrated. The prior-knowledge network can be visualized (with tools such as Cytoscape [189]) to study relationships between features from different data types. To improve network visualization and results interpretation, a threshold can be set on the *FDRs* (*e.g.* $T = 5\%$): only edges describing False Discovery Rate lower than the selected value T will be visualized. This step can also be useful to perform feature selection, an important step to remove unwanted noise (see Chapter 3 and [207]).

4.2.2 Prior knowledge inclusion

Feature weights assignment

Once that the prior-knowledge network $G = (V,E)$ is generated, the aim of the pipeline is to assign a weight to the nodes (*i.e*, the features), rather than to the edges.
For each node $s \in V$, the posterior probability $\tilde{\gamma}_s{}^{i}$ that the node s is connected to at least another element $t \in V$ is given by:

$$\begin{aligned} \tilde{\gamma}_s{}^{i} &= \mathbb{P}\left(\bigcup_t \{\gamma_{st}{}^{i} > 0\}\right) = \\ &= 1 - \prod_t \mathbb{P}\left(\gamma_{s,t}{}^{i} = 0\right) = \\ &= 1 - \prod_t \left(1 - \gamma_{s,t}{}^{i}\right) \end{aligned} \tag{4.3}$$

where the index i refers to the matrix $\Gamma^{i} = \{\gamma_{st}{}^{i}\}$, which stores information about connections between the data types containing s and t.
A higher value of $\tilde{\gamma}_s{}^{i}$ is assigned to the nodes more likely to connect to other nodes. Importantly, we should remember that the edges considered in this case are built across different

data types: inter-omics relationships are modelled.
As previously done for the network generation, a False Discovery Rate $FDR(s')$ was computed with equation 4.2 for each node $s' \in V$ to avoid calibration issues.

To give more importance to the features showing low False Discovery Rate, feature weights were computed as $\omega_s' = \frac{1}{FDR(s')}, \quad \forall s' \in V$. Weights computed for the same feature s' but obtained from different matrices were summed to a obtain a unique value. Weights were then normalized with the following formula to lay in the range $(0,1]$:

$$\bar{x} = \frac{x - \min x + 0.01}{\max x - \min x + 0.01} \tag{4.4}$$

where the addition of 0.01 helps avoiding weights exactly equal to 0. The inverse of the assigned weights are then considered as multiplying factor. This shrinks the distance between samples with similar profiles. Features more related to the subtype provide higher effect in the SNF similarity matrix (see Appendix A).

Three-omics and prior knowledge integration

The Similarity Network Fusion (SNF) method [229] has been considered to perform the three-omics integration step of the pipeline, since results obtained in Chapter 3 indicate it as the best unsupervised clustering method [207].

As described in the Appendix 1, the first step of SNF is the construction of a similarity matrix $W^i = \left\{ w_{hj}^i \right\}$ for each single data type X_i. Those matrices are obtained by computing a kernel in Gaussian form, which is known to automatically provide a vectorial representation of the data in the feature space. Gaussian kernels, moreover are among the most used in classification algorithms, thanks to their ability to generate non-parametric classification functions [185]. Given two samples x_h and x_j, the kernel used in the SNF method is computed as describe in [228]:

$$w_{hj}^i = \frac{e^{-\rho^2(x_h,x_j)}}{\sigma \varepsilon_{hj}} \tag{4.5}$$

where ρ represents a metric (*e.g.* the Euclidean distance) between the samples x_h and x_j. The term ε_{hj} in the denominator is instead given by:

$$\varepsilon_{hj} = \frac{\overline{\rho(x_h,N_h)} + \overline{\rho(x_j,N_j)} + \rho(x_h,x_j)}{3}$$

where $\overline{\rho(x_i,N_i)}$ is the averaged metric between the sample x_i and its K nearest neighbours, N_i. K and $\sigma \in [0.3,0.8]$ are input parameters of the method. The number of neighbours K was always set to n/C, where n is the number of samples and C is the number of searched subtypes.

Despite the default metric suggested in the work by Wang *et al.* [229] is the Euclidean distance, also correlation or other types of distances can be used as ρ. This is due to the fact that the input of the SNF method can be feature vectors, sample pairwise distances, or sample pairwise similarities [229]. The used distance can influence integration results: we explored the use of generalized Euclidean metric in computing the pairwise distances in equation 4.5. With this aim, we applied the SNF method to the simulated datasets by substituting the Euclidean distance with the more general Minkowski distance.

Given two vectors in $\mathbb{R}^n$, $X = (x_1,\ldots,x_n)$ and $Y = (y_1,\ldots,y_n)$, the Miknowski distance of order p is defined as:

$$\rho_p(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

It is possible to notice that, when $p = 2$, the Minkowski distance coincides with the Euclidean distance.

As already did in Chapter 3, integration performances of the SNF method, before and after weighting the features, were evaluated by applying spectral clustering. The obtained clusters were compared to the real subtypes: their agreement was evaluated with the F-score. The averaged F-score across the subtypes was used as index of integration accuracy.

We used the paired Wilcoxon signed-rank test [237] to understand whether the classification abilities of the unsupervised method are improved by weighting the features. The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test. It is used to check the significance of the difference between distributions of results obtained by two methodologies. We tested the alternative hypothesis of obtaining higher F-scores after that feature weights were provided. Results with p-value$\leq$ 0.05 were considered statistically significant.

4.2.3 Datasets and simulations

The pipeline presented in the previous sections was tested on simulated datasets. We generated them with an increasing level of complexity, given by the number of subtypes to be recognized and by the signal strength across omics.

Since we showed that the integration of more layers increases multi-omics integration performances [207], we always generated datasets composed by three different data types.

Simulation 1: assigned feature weights

As a first analysis, we wanted to test the second step of the pipeline. We wanted to understand if assigning an higher weight to features that are important for the phenotype emergence provides useful information. With this aim, we run a simulation on a set of simulated datasets. We generated 100 datasets

$$\mathbf{X^j} = (X_1^j, X_2^j, X_3^j), \quad j = 1, \ldots, 100$$

composed by three different data types, $X_i^j, \quad i = 1,2,3$, built as matrices of dimension $n \times p_i$. The data types had the same number of rows ($n = 200$), representing samples, but different numbers of columns, respectively $p_1 = 300$, $p_2 = 1000$ and $p_3 = 700$.

The data types were created with R package *echoseq* [181], which allows the generation of a response matrix associated to the predictors used as input. As response matrix, we produced a binary vector $y = (y_1, \ldots, y_n)$ of dimension $n = 200$. Each element $y_h, \quad h = 1 \ldots n$ represents the subtype (0 or 1) associated to the corresponding sample.

Additionally, for each dataset $\mathbf{X^j} \quad j = 1 \ldots 100$, a fixed number of "active" features ($p_0 = 30$) was generated to be responsible of the different states of the response vector. Those features were randomly selected across the three data types composing the dataset.

To be consistent with the simulated datasets considered in the previous chapters, the functions of the *echoseq* package, created to reproduce SNPs values, were modified to generate Gaussian distributed features. The input parameter *max_tot_pve* (maximum phenotypic variance explained by all active molecules) was set to 0.9.

For each dataset $\mathbf{X^j} = (X_1^j, X_2^j, X_3^j), \quad j = 1, \ldots, 100$, weights $\omega^i = (\omega_1^i, \ldots, \omega_{p_i}^i)$ were generated for the p_i features contained in $X_i \quad i = 1,2,3$. For this analysis, weights were assigned to features according to their role in the subtype emergence. Weights were drawn from two different Beta distributions, which are frequently used in Bayesian analysis to describe the initial knowledge regarding the probability of success [127]. Weights for the 30 features that had been associated to the phenotype were drawn from Beta(α_1, β_1) with $50 \leq \alpha_1 \leq 200$ and $2 \leq \beta_1 \leq 15$, in order to obtain values close to 1. On the other hand, weights for the other features were taken from Beta(α_2, β_2) with $2 \leq \alpha_2 \leq 15$ and $50 \leq \beta_2 \leq 200$. This choice allowed to obtain positive weights distributed close to 0.

Parameters $(\alpha_1, \beta_1, \alpha_2, \beta_2)$ for the Beta distributions were randomly selected in the specified ranges. The ranges were chosen to draw values for the important/not important features from clearly separated distributions.

Once that weights $\omega^i = (\omega_1^i, \ldots, \omega_{p_i}^i)$ were drawn, feature values belonging to X_i^j were multiplied by the factor $\frac{1}{\omega_h^i}$, $h = 1 \ldots p_i$.

The 100 generated datasets $\mathbf{X^j}$ were integrated with SNF, before and after weighting the features. In this case, the Minkowski distances ρ_p with $p = 1, 2$ were considered.

Simulation 2: feature weights from inter-omics relationships

To test the complete pipeline on datasets with more reliable inter-omics relationships, we considered the simulated scenarios C, D and E. As described in Chapter 3, they come from Breast Cancer dataset [203]. They have three subtypes to be recognized and an increased level of complexity (see Figure 3.2). SNF was applied to those simulated scenarios before and after the computation of the feature weights from the prior-knowledge network (section 4.2.1). For this analysis we considered the Minkowski distances ρ_p for $1 \leq p \leq 5$, in order to study the effect of the selected metric.

To validate the results obtained for datasets C, D and E, we additionally generated 100 datasets (composed of three data types of dimension 60×900). For each of them, we randomly selected a total of 900 features from the Breast Cancer dataset used in Chapter 3. Profiles of 60 samples, divided in three subtypes, were generated following the creation of simulated scenario C (for each dataset, three different subtypes were generated, with only one data type able to distinguish all of them, see Figure 3.2). In this simulation, the Minkowski distance with $p = 1$ was used to compute the SNF similarity matrices for the data types, since results on datasets C, D and E indicate it as the most powerful (see Section 4.3.2).

4.3 Results and Discussion

4.3.1 Effect of prior-knowledge inclusion

We first tested the effect of prior knowledge inclusion to unsupervised multi-omics integration methods, with a simulation on 100 datasets (Simulation 1).

In Figure 4.2 the distribution of the simulated weights for one of the generated data types is presented. To better reflect real biological cases, where only a small number of biomolecules (with respect to the total number of measurements) are responsible for the emergence of the phenotype, only 1.5% of the generated features has been assigned a high weight (close to 1). This percentage is represented by the right peak in Figure 4.2. The lower weights, close to 0, which were assigned to the remaining elements, are instead represented in the peak on the left.

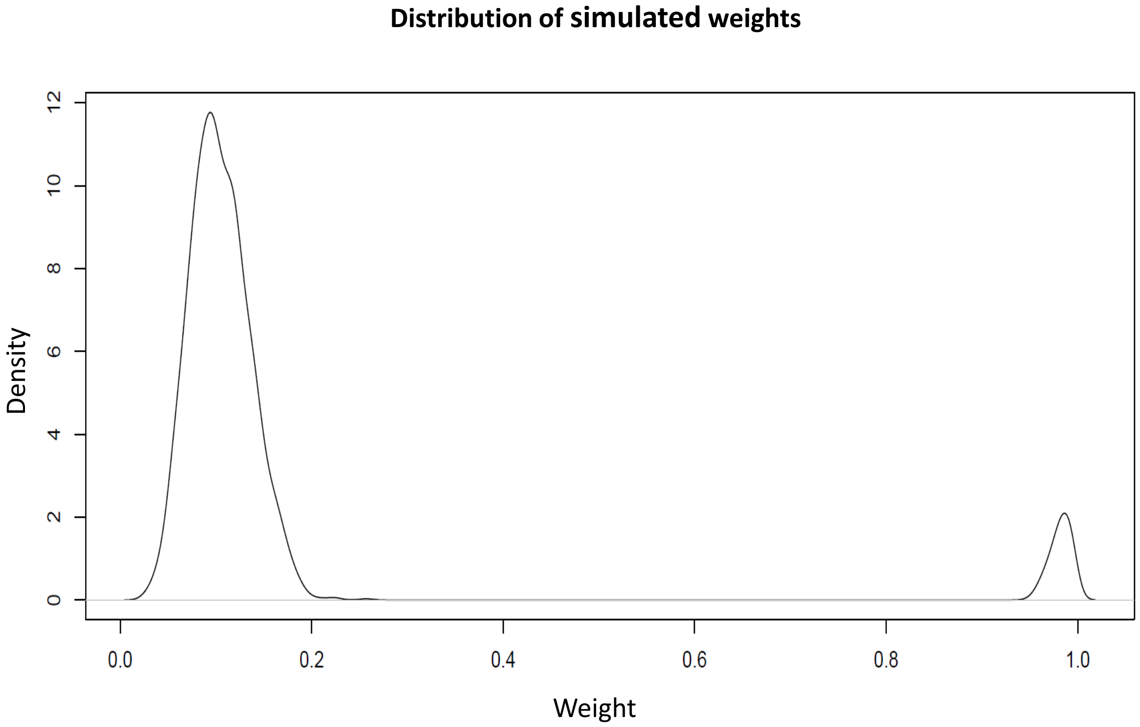


Fig. 4.2 Example of weight distribution for one of the 100 generated datasets (Simulation 1). Weights were drawn from beta distributions: only few elements (related to the phenotype) had been assigned to a weight close to 1 (right peak). The majority of the features was instead assigned to a weight close to 0 (left peak), since they were not related to the generated phenotype.

The 100 generated datasets from Simulation 1 were analysed with the SNF algorithm, with the aim to correctly classify the 200 samples, which were divided in two subtypes. SNF was applied both before and after weighting the dataset features.

The performed analysis provided higher F-scores when weighted features were integrated. This result was obtained not only with the classical Euclidean distance (Minkowski distance of order $p = 2$), but also when the similarity matrix (equation 4.5) were computed with Minkowski distances of different orders.

More in detail, results of the simulation for the Minkowski distance of order $p = 1$ are presented in Figure 4.3. With respect to the classical Similarity Network Fusion method (orange boxplot), SNF integration of weighted datasets (distribution of average F-score in the green boxplot) provided, on average, more accurate classification.

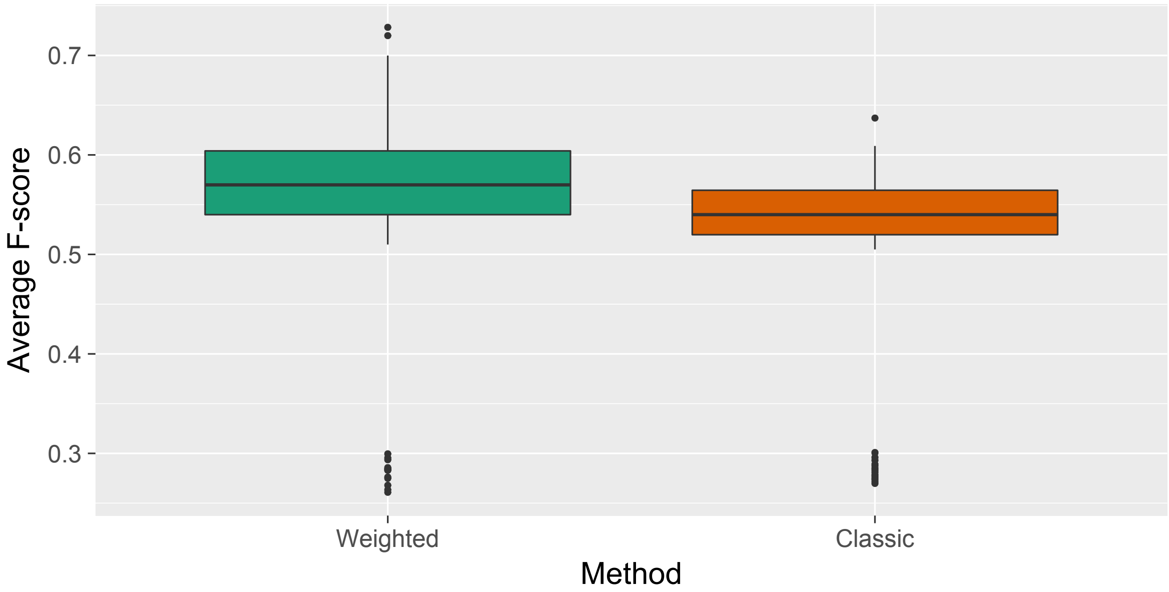


Fig. 4.3 Boxplots of the averaged F-scores obtained for the Simulation 1. The orange and green boxplots represents results obtained before (orange) and after (green) weighting the features. The statistical significance of this result was p-value=10^{-5}.

The statistical significance of this result was assessed with the paired Wilcoxon signed-rank test. We tested the alternative hypothesis of obtaining higher values in the F-scores distribution when considering weighted data types. The Wilcoxon test provided a significant p-value= 10^{-5}, which allowed to accept the alternative hypothesis. The statistical significance of the result indicate that prioritizing features known to be related to the phenotype improves the precision of unsupervised integration methods.

4.3.2 Effect of data-driven computation of inter-omics relationship

We then tested the impact of computing the feature weights according to the inter-omics relationships occurring between the features. With this aim, we applied the complete pipeline to the simulated scenarios described in Chapter 3. Thus, the relationships between features belonging to different data types were already established, while those occurring between features and subtypes were built during the datasets creation.

For each tested dataset, we generated the prior-knowledge network and assigned weights to features as described in section 4.2.1 and 4.2.2. We then classified samples from the weighted and not-weighted datasets with SNF.

For this analysis, we computed the kernel matrix described in equation 4.5 starting from

Minkowski distances of different orders. For each distance we computed the F-score for values of the SNF parameter σ in the range $[0.3, 0.8]$, as suggested in [229]. Results of this analysis for the simulated scenario C, D and E are shown in Figures 4.4, 4.5 and 4.6 respectively.

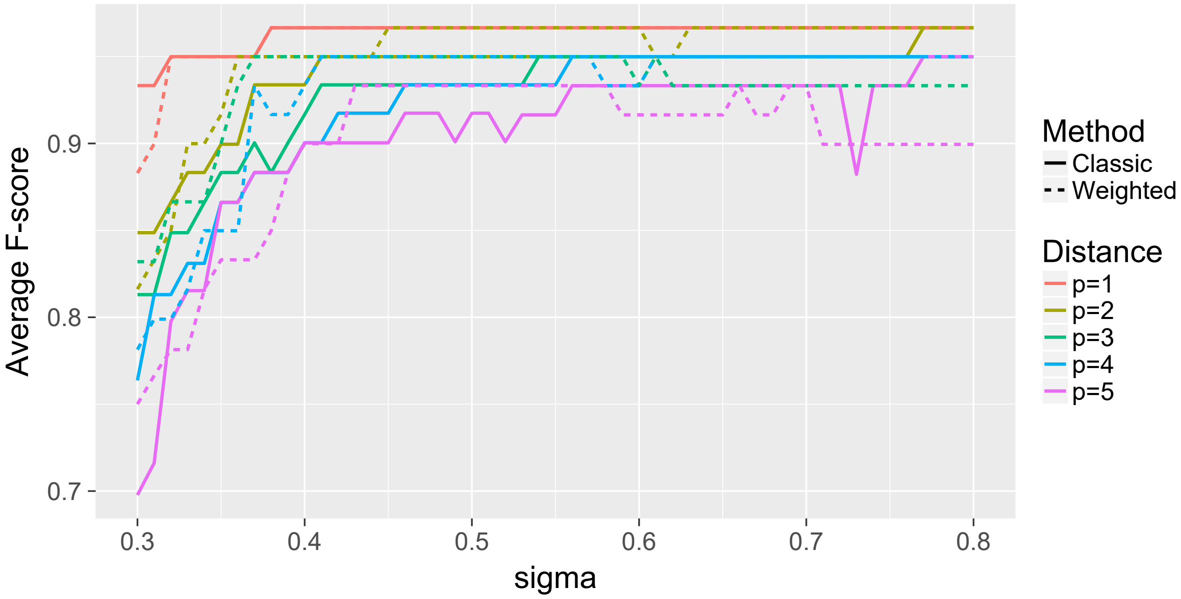


Fig. 4.4 Classification results using SNF data integration on simulated scenario C before (solid lines) and after (dashed lines) weighting the data. Colours represent Minkowski distances of different order. On the x-axis there are the values for the parameter σ (sigma) of the SNF method, on the y-axis the averaged F-score obtained for each considered case.

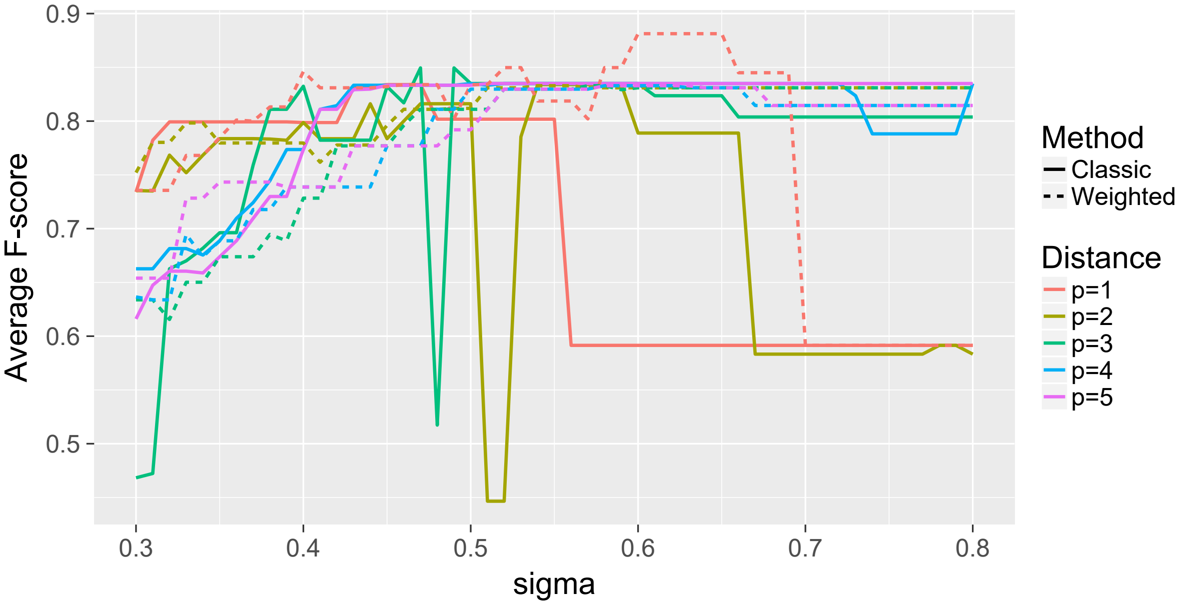


Fig. 4.5 Classification results using SNF data integration on simulated scenario D before (solid lines) and after (dashed lines) weighting the data. Colours represent Minkowski distances of different order. On the x-axis there are the values for the parameter σ (sigma) of the SNF method, on the y-axis the averaged F-score obtained for each considered case.

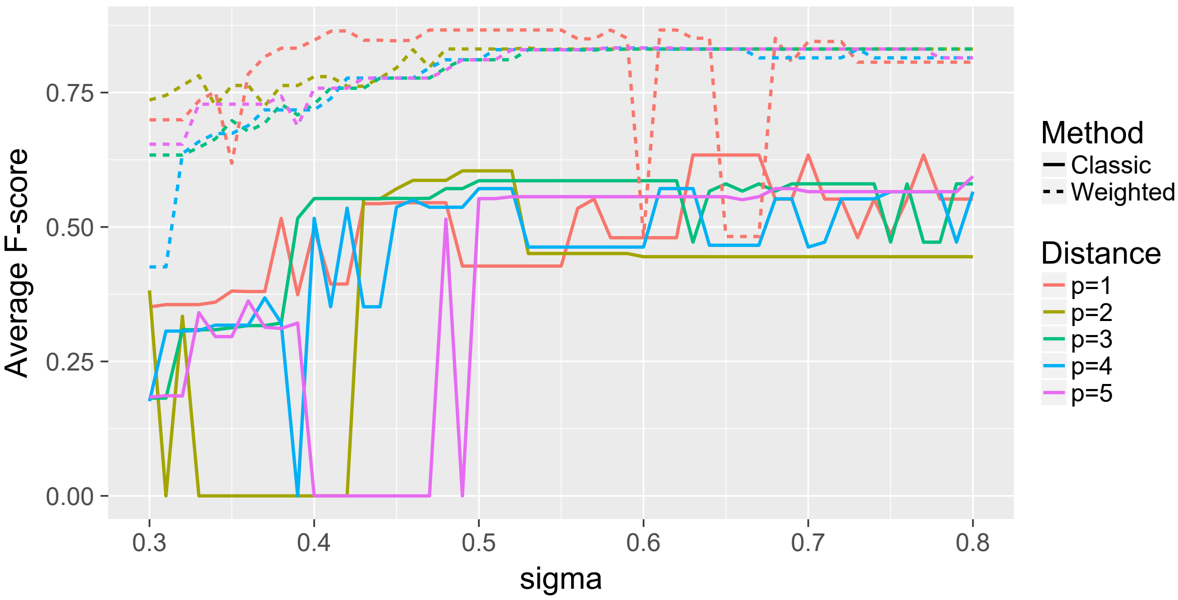


Fig. 4.6 Classification results using SNF data integration on simulated scenario E before (solid lines) and after (dashed lines) weighting the data. Colours represent Minkowski distances of different order. On the x-axis there are the values for the parameter σ (sigma) of the SNF method, on the y-axis the averaged F-score obtained for each considered case.

For the simplest among the considered cases, scenario C (Figure 4.4), the same maximum F-scores (0.967) were obtained for Minkowski distance of order $p = 1$ and $p = 2$. Distances of higher orders obtained lower accuracies. The use of Minkowski distance of order $p = 1$ provided instead the highest accuracies in cases D and E. In particular, it performed better than the Euclidean distance ($p = 2$). The maximum F-score obtained for scenario D (Figure 4.5) was equal to 0.88 ($p = 1$, weighted data). In the same case, for $p = 2$, the maximum F-score for not-weighted data was equal to 0.83. Similarly, for case E (Figure 4.6), the distance of order $p = 1$ provided the highest F-scores either in the weighted (0.866) and not-weighted (0.634) case. This is confirmed also by the literature: low values of the distance order p are preferable when similarity between objects coming from high-dimensional data [1] is measured. Additionally, increasing p was proved to make similarities more susceptible to the problem dimension. In particular, when $p \geq 1$, the Minkowski distance of order $p = 1$, (also called Manhattan Distance metric) was shown to perform better than the Euclidean one [1].

Moreover, from Figures 4.4, 4.5 and 4.6, it is possible to notice that integration performed on weighted data (dashed lines), resulted on average more robust against the changes of the SNF parameter σ, with respect to the classic SNF method (solid lines). This trend can be seen also for Minkowski distances that did not reach the maximum F-score (with F-scores never being lower than 0.6 for case D and on average close to 0.75 for case E), especially with the standard Euclidean distance. This suggests that the inclusion of prior knowledge in the integration pipeline increases the stability of the method. Indeed, in this case, integration relies on more information. Since σ is an input parameter of the SNF method, an improved stability against its choice is useful especially when complex datasets are integrated: optimization on the parameter value, that requires more time for complex data, could be avoided or simplified.

In agreement with the results obtained in section 4.3.1, weighting the features provided F-scores higher than those gained with the completely unsupervised approach. However, it should be noted that the feature weights here considered were obtained the retrieved inter-omics relationships (see section 4.2.1). The importance of prior-knowledge inclusion in multi-omics integration is underlined by the fact that classification accuracy obtained on weighted datasets increased together with the datasets complexity (from C to E). In simulated scenario C, for example, the obtained F-scores are similar for the two approaches. In case D weighted data obtained higher classification for some of the considered distances. On the other hand, the application of the complete pipeline to case E (the most complex case, generated with a low shared signal), had an high impact on classification results. F-scores computed for weighed data were, on average, higher than those obtained with not weighted data.

To statistically confirm the results obtained for these three datasets, we applied the integration pipeline to a set of 100 datasets (Simulation 2). These datasets were generated as simulated scenario C (see Figure 3.2).

The distribution of the weights obtained for one of those datasets is presented in Figure 4.7: similarly to the distribution of the simulated weights (see Figure 4.2), a high number of features was assigned to a low value, while a decreasing number of elements was assigned to high weights. On the other hand, the transition between high-weighted and low-weighted features is smoother with respect to the previous case (Figure 4.2). This can be justified by the fact that the data used to generate weights displayed in Figure 4.7 come from real omics measurements: features can be connected to the phenotype emergence with several levels of strength.

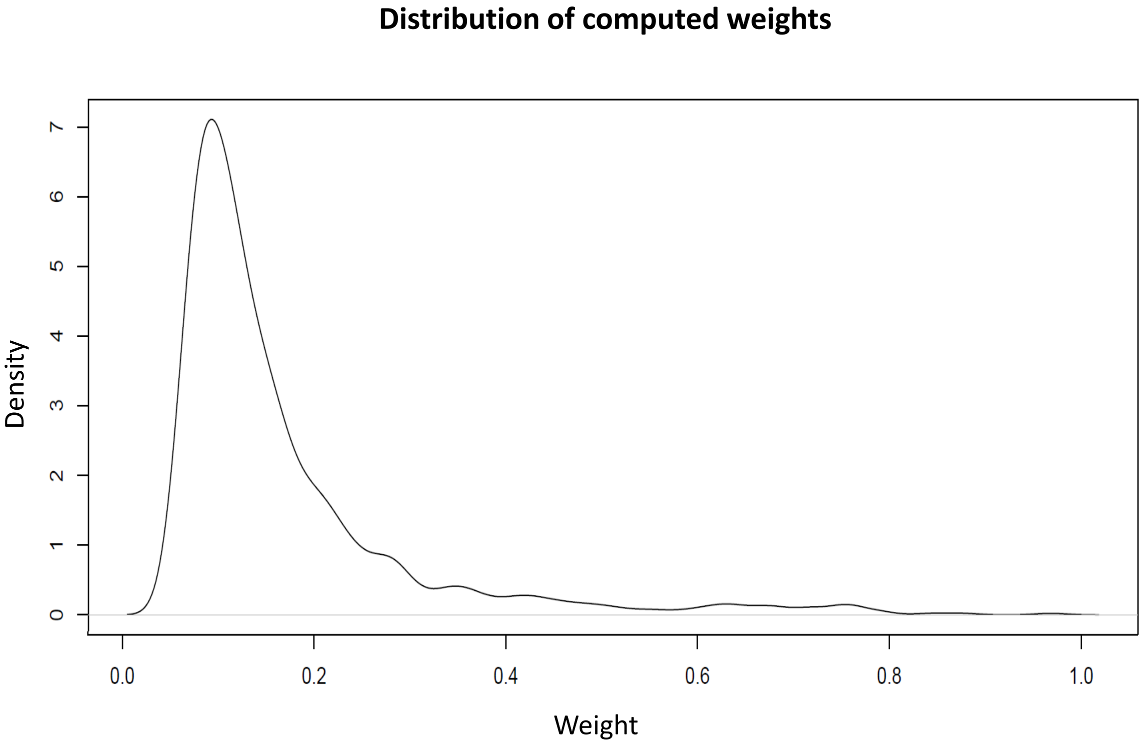


Fig. 4.7 Distribution of the feature weights obtained by the use of the prior-knowledge network (Simulation 2). These weights were computed for one of the 100 generated datasets. The majority of the features obtained a weight close to 0, while less elements obtained higher weights.

Moreover, the simulated weights considered in section 4.2.3 were provided individually for each feature. For this analysis, feature weights were instead computed taking into account relationships with elements coming from other data types.
The results obtained for simulated case C, D and E by the use of a prior-knowledge network to compute feature weights were confirmed by Simulation 2. Following the results from the three datasets, Minkowski distance of order $p = 1$ was used in the simulation, both for the weighted and not weighted data. As presented in Figure 4.8, weighted data resulted in an improved classification precision.

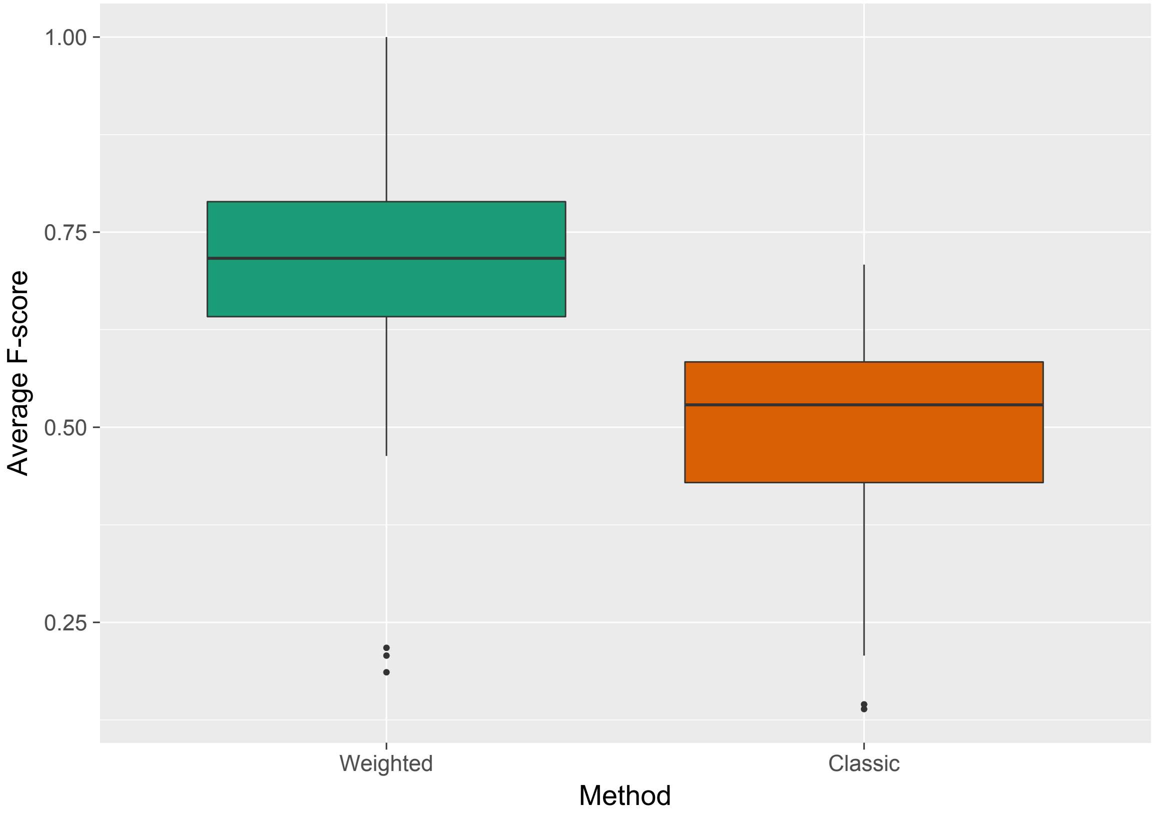


Fig. 4.8 Boxplots of the averaged F-scores obtained for Simulation 2. The different boxplots represent the results obtained before (orange) and after (green) weighting the features belonging to the different datasets. The distribution of F-scores with weighted data is higher than the other with a statistical significance of p-value=4.95×10^{-14}

The distribution of the averaged F-scores obtained after weighting the features on the basis of the prior-knowledge networks is represented as a green boxplot in Figure 4.8. F-scores resulted significantly higher with respect to the distribution of those obtained with not weighted data (orange boxplot, Figure 4.8).

The averaged F-scores computed by the use of simulated weights (green boxplot in Figure 4.3) obtained a maximum value of 0.73, with a median of the distribution close to 0.57. On the other hand, when weights were computed on the basis of feature interactions, integration obtained higher F-scores (green boxplot in Figure 4.8), with a maximum of 1 (perfect classification) and a median of the distribution equal to 0.72. The statistical significance of the results described in Figure 4.8 was again assessed with the paired Wilcoxon signed-rank test, which provided a p-value equal to 4.95×10^{-14}. While only two subtypes were generated for the previous case (4.3), it should be considered that, in this case, samples were divided in three subtypes (more complex case). Additionally, with respect to the case where weights were simulated (Figure 4.3), computation of weights from observed inter-omics interactions provided not only higher F-scores, but also higher statistical significance. This indicated that, at least on simulated datasets, the inclusion of knowledge about inter-omics relationships is beneficial for unsupervised simultaneous methods.

4.4 Conclusion

In this Chapter we tested the effect of the inclusion of prior knowledge to unsupervised simultaneous integration methods using simulated datasets. We computed inter-omics relationships by the use of a sparse multivariate statistical model, which helped us to build a prior-knowledge network. Features were then weighted according to their importance in the network. Weighting the features of simulated datasets improved the precision of the multi-omics integration method to classify samples. This suggests that integration methods should consider inter-omics connections, such as those coming from linear integration methodologies, to increase their power. We should underline that we considered knowledge retrieved from the data themselves: including into the analysis external prior knowledge, such as that coming from pathways repositories or protein-protein interactions may strongly bias the results: other experiments and simulations should be conducted in the future to better understand the effects of adding external knowledge. We also tested the influence of the type of distance used to compute the similarity network in the unsupervised method. Although the Euclidean distance is often selected, other distances such as the Manhattan one should be considered when dealing with high-dimensional spaces. In general, our results suggest to carefully select the metric used to measure sample distance. Importantly, on the tested datasets, prior knowledge inclusion showed a high positive effect on the classification of samples from complex datasets (*e.g.* lower signal strength). This suggests that the combination of unknown interactions (modelled by unsupervised methods) and of statistically significant computed relations coming from the data themselves can provide a better description of the subtypes under study. The results on simulated datasets here described should be considered as an indication for further research on the importance of prior biological knowledge addition to multi-omics integration models.

Chapter 5

Discussion

Mathematical and statistical models are necessary tools in the development of accurate methods to perform multi-omics data integration. The results summarized in this thesis highlight the differences and strengths of different mathematical approaches to tackle particular biological problems involving inter-omics interactions. In particular, we focused on the influence of the inclusion of linear prior knowledge in simultaneous unsupervised integration of three different omics data. The models and studies we described here may be further extended: in the following paragraphs we discuss some of those possible developments that could improve the understanding of multi-omics data integration methods and their application.

The results presented in Chapter 2 were obtained by applying a linear supervised approach to real data. The multi-omics integration of methylation, gene expression and protein levels performed in this study provides additional insights into the molecular process of preadipocyte differentiation to mature adipocytes. It focuses in particular on the influence of DNA methylation changes on the other omics during the adipogenetic process. The impact of the addition of different doses of fructose was also explored.

The transcriptomic and methylation data integration indicated that DNA methylation and resultant gene expression patterns are "pre-programmed" since up or down gene regulation overlap differently methylated regions (DMRs) that could have the expected (down methylation, up gene expression) or the reverse (up methylation up gene regulation) pattern. In addition, the differentiation program over-rides differences in type and level of nutrients (fructose versus glucose) consistent with previous studies of the metabolic fate of fructose [218] and the effect of fructose on glucose metabolism [219] in these cells.

One important result of the study is that, for a set of genes, we observed related changes of methylation, gene expression and protein levels. This ensures that the integration of more omics data provides a deeper understanding of biological processes. This fact should also be considered when addressing other biological problems. Since obesity and its related comorbidities are among the major problems caused by westernization of diet, the results that we obtained suggest to further research on how nutrients affect methylation, gene expression and protein levels.

Additionally, the influence of the experimental design (factor also addressed in Chapter 3) on this three-omics integration suggests that, whenever different omics come from experiments taken at the same time and under identical conditions, their combination provides stronger results, with higher statistical significance. This is in accordance with the study of experimental design of Cavill *et al.* [31], and reflects the practical importance of this factor: researchers should pay attention when designing the study, if the final aim is to integrate datasets. Moreover, this factor should always be considered when interpreting the results of integration.

Results obtained by the linear three-omics integration performed in Chapter 2 highlight the ability of the linear approach of solving molecular mechanism discovery problems in a supervised way. This approach can also easily model direct and causal inter-omics associations and thus describe related changes across the different molecular layers (*e.g.* genes varying at the epigenomic, transcriptomic and proteomic level). However, the relationships considered by linear approaches may not reflect the complete set of biological interactions underneath the process under study. Indeed, only known relationships are taken into account, forgetting the unknown, but potentially relevant, inter-omics interactions. Moreover, linear approaches often model only one direction of causality and are not able to describe changes happening at the same time in different molecular layers.

The simultaneous unsupervised approach, able to deal with the limitations of linear methodologies, was instead studied in deeper in Chapter 3. Several factors that could influence multi-omics data integration, such as pre-processing, number of omics integrated, and, as already pointed out, experimental design were explored. We investigated and compared results related to only one of the possible applications of multi-omics data integration (sample classification) due to the possibility to clearly assess classification results against real subtypes. It has to be underlined that the methodologies tested in this Chapter can be applied to solve other questions with potentially different outcomes. For example, they can be used to identify biomarkers, which are biomolecules characterizing the considered phenotype. In this framework, it would be interesting to compare the set sets of biomarkers found by the different methodologies. This, however, would require quantitative methods able to assess the power of a selected set of biomolecules to describe the phenotype under study.

With regard to the biological question studied in Chapter 3, instead, considering multiple molecular levels (more than two), increased the amount of knowledge extracted from the available data and resulted in a more accurate sample classification. Nevertheless, it should be considered that the available omics data are not always suitable for integration: one of the data types could, for instance, display a high level of noise. Thus, pre-processing and visualization of the data are important steps to be performed prior than integration. If the level of noise is too high, the noisy omic data should be removed from the analysis, not to bias the results. To expand the research on how much multi-omics integration is data dependent, a possibility would be to consider not only gaussian distributed data, as we did through this dissertation, but also more heterogeneous data (*e.g.* microbiome). This would require the study of more specific methods (like the recent one proposed by Gabasova *et al.* in [60]), able to deal with differences in data distributions and inter-omics cluster agreement. We tested methods in situations known to potentially affect results of multi-omics data integration such as when considering omics that bring reinforced or complementary signal, with increasing numbers of subtypes, when noise is present. Among the tested factors, we found that noise influences integration results; an effect that can be mitigated by the addition of a feature selection step before proceeding with data integration. The results of our study especially recommends it when dealing with complex design (such as those having more than two different omics data, or with low signal strength, or multiple cellular subtypes). However, we tested the influence of a general type of noise that could affect omics data (that is, gaussian distributed noise with fixed standard deviation) and of a general feature selection approach. Thus, a further step to better understand the problem of multi-omics data integration, could be to test the influence of other types of noise and filtering methods. For example, it would be interesting to consider datasets generated with different levels of noise, such as with an increasing number of noisy features or generated with diverse variances. Noise could also be generated following different distributions (*e.g.* gaussian, binormal) to better reflect the distribution of real omics measurements. Similarly, specific filtering method should be considered for each of the considered data types (like those proposed in [74]) and to compare the effects of *ad hoc* feature selection and feature prioritization methods.

The study and results presented in Chapter 3 suggest that simultaneous omics integration should be considered in future studies with more omics data available. This is due to their ability to deal with unknown relationships not in a hierarchical way. However, from the methodology comparison, we also realized that statistical integration methods could still be improved, for example by the addition of prior information about inter-omics relationships. This step could diminish false positive results, while enhancing the relevance of true molecular interactions.

For this reason, we proposed in Chapter 4, a combined pipeline to take advantage of the main characteristics of the linear supervised and simultaneous unsupervised methodologies. Both of them can extract useful information from the data, respectively by: i) relying on known interactions between different omics data and ii) modelling relationships without considering whether they are already present in the literature.

We tested the effect of adding prior knowledge to a unsupervised integration methodology by focusing on sample classification. We computed relationships among elements coming from different data types by the use of a sparse multivariate statistical model, and weighted features according to the importance of their connections. On simulated datasets, this last step improved the classification precision of the multi-omics data integration. This fact suggests that multi-omics integration methods should consider inter-omics relationships, coming from linear multi-omics data integration methodologies, to increase their power. Interestingly, the positive effect of adding prior-knowledge to the simulated scenarios increased together with the datasets complexity (*e.g.* lower signal strength). However, further research should be done to better understand to which extent prior biological knowledge addition is necessary in multi-omics integration. For example, it would be important to test the effect of prior knowledge inclusion when dealing with complex scenarios, such as noisy datasets, or when samples are divided in heterogeneous clusters and not in clusters consistent across the omics. Despite the classification accuracy obtained with the proposed pipeline were higher than those gained with the unsupervised methodology alone, some additional steps could be explored. For example, we only considered interactions computed from the data themselves. Although this can help in the extraction of data-driven relationships, still the use of already established connections is missing. Thus, a further step in studying the influence of prior-knowledge on multi-omics data integration could be the inclusion of information about inter-omics relationships coming both from the data themselves and from external sources of knowledge, such as pathways repositories.

However, it should be taken into account that external knowledge external to the experimental data, for example information coming from protein-protein interactions, is usually tissue or disease specific. Thus, its inclusion in the analysis may strongly bias the results, leading to false or incorrect conclusions. As an example, although some cancers share genetic signatures across individuals, driver mutations are highly diverse: considering information coming from the wrong tissue could lead to differences in prognosis and therapies [93]. This suggests that further research should be done to quantify the strength of the bias generated by the inclusion of external sources of information. Moreover, the origin and the phenotype described by the prior konwledge should be clearly assessed before its inclusion.

The correct use of external sources of knowledge could instead further decrease the false discovery rates while increasing the power of knowledge inclusion. To this end, the use of networks, will still be desirable. An interesting approach in this direction is provided by weighted multiplex networks [141]: specific indices developed to describe across-omics links and their weights can be used to evaluate network properties. In particular, entropy can be considered to quantify the amount of information encoded in the inter-layers connections. The indices suggested by this method could give a deeper insight also in other aspects of the biological problem of sample classification, that is, it would be possible not only to compute clusters of patients but also to understand the molecular mechanisms underlying the obtained classification.

Related to the prior-knowledge network computation, we explored methods able to find a linear predictor/response relationship between two omics data. However, it would be interesting to test other statistical methods to weight biomolecules and their connections, like methods relating biomolecules without imposing a hierarchical direction.

In this dissertation we explored some of the most relevant aspects of multi-omics data integration when considering more than two data types. We focused especially on the effect of simultaneous and linear combination of data to solve different problems and we proposed a pipeline to combine them. Among the other factor, we also evaluated the effect of biomolecule selection and the impact of the experimental design on the results reliability.

The topics studied in this thesis and the obtained results are on the cutting edge of research in multi-omics integration. Several reviews published in the last years pointed out the importance of better understanding the existing integration algorithms [11, 31] and to critically compare them [79]. The results obtained with the methodology comparison, which as main result indicate network-based approaches as the most appropriate to perform sample classification, are in line with these requests and might aid decisions of researchers working with omics data.

At the same time, the need to include biological knowledge when combining different molecular layers is constantly remarked in literature [79, 93, 156]. In this framework, the importance of our study relies in the positive effect which the inclusion of knowledge was shown to have, both in multi-omics discovery of mechanisms underlying biological processes and in subtype classification. In particular, the integration of prior knowledge coming from the data themselves responds to the need of considering interactions less well described in the literature.

This aspect acquires relevance since omics data measured from the patients or the samples at hand could not always be comparable to existing reference datasets. Those are in fact usually developed for a specific disease or tissue. For example, if applied to real omics data, our approach could be useful to help the diagnosis of diseases whose genetic aetiology (genetic factors causing a specific condition) is not well known [93].

Additionally, the use of heterogeneous networks (which consider a different types of nodes for each type of elements and inter-layers connections), provide a valuable indication to overcome the problem of having information from diverse sources, as it is with molecular layers. Indeed, those networks collect signals that may come from fundamentally different data, such as discrete genetic variation or continuous gene expression measurements.

Moreover, the linear modelling of two-omics interactions that we used to build the network gives evidence to support the signals and takes already into account dimensionality problems. One of the advantages of using networks to collect and study those inter-omics connections is that, through visualization and further analysis, modules of biomolecules can be found which may explain the differences in the retrieved subtypes. From this point of view, these aspects of our approach can be important especially applied to research in cancer profiling and diagnosis. It is well known that tumor subtypes may have varying survival prognoses and therapeutic strategies [93]: a more precise subtype classification, together with information on causal groups of biomolecules, could help researchers and physicians to recommend or develop more specific therapies.

Nevertheless, high-throughput technologies able to collect and measure different biological layers keep improving, together with the awareness of strengths, weakness and challenges of combining data. In the next future, this will lead research towards exploration of possible new questions to be answered and to not yet considered perspectives. For example, being able to use results obtained by multi-omics data integration to perform prediction and missing data imputation would be useful to relate theoretical and data-based discoveries to preventive medicine, nutrition and other sciences important for improving the quality of life.

Appendix A

Overview of multi-omics integration statistical methods

Details about the statistical methods used in Chapter 3 to perform multi-omics integration wll be provided in this Appendix.
The k different omics included in a given dataset (*e.g.*, gene expression or protein abundances) can be represented as matrices $X_1 \dots X_k$, of dimensions $n \times p_i,\ (i = 1 \dots k)$, where n is the number of subjects while p_i is the number of variables measured in the i-th matrix.

The first two methods considered for the performed comparison, Multiple Canonical Correlation Analysis and Multiple Co-Inertia Analysis (see following sections), are based on a latent factor decomposition to reduce the problem of high-dimensional data [140].
Following this approach, each omics data X_i can be decomposed by means of r latent components ($r < p_i$) as:

$$X_i = F Q_i^T + E_i,\ \ i = 1 \dots k$$

where F is an $n \times r$ matrix whose columns contain the co-structures between the different omics data; Q_i is a $p_i \times r$ matrix whose columns $(q_i^1 \dots q_i^r)$ contain the loadings of the p_i variables on the r latent components; E_i is an $n \times p_i$ matrix storing error terms [140]. The constraints imposed to find the latent components, and the procedure to compute them, are different in the two methods (MCCA and MCIA).

Multiple Canonical Correlation Analysis (MCCA)

Multiple Canonical Correlation Analysis [239] is a multivariate based method, implemented in the R package PMA [240]. Before integration, data columns are scaled to have mean equal

to zero and standard deviation equal to one.
The method computes $h = 1 \dots r$ latent components, where r is an input parameter of the algorithm, by searching for linear combinations of maximally correlated canonical variates from multiple omics, defined as $X_i q_i^h$ ($i = 1 \dots k$ and $h = 1 \dots r$). The maximization problem to be solved for the first latent component ($h = 1$) is the following:

$$\max_{q_1^1 \cdots q_k^1} \sum_{i<j} \text{cor}(X_i q_i^1, X_j q_j^1), \quad \text{with } ||q_i^1||^2 < 1 \quad \forall = 1 \dots k$$

where $q_1^1 \dots q_k^1$ are the k vectors providing the first column of the matrices Q_i ($i = 1 \dots k$). To compute the other latent components ($h = 2 \dots r$), the maximization problem above is solved again, but in such cases, the omics data $X_1 \dots X_k$ are replaced by *ad hoc* matrices that are orthogonal to the canonical variates computed in the previous steps. This allows discarding the already processed information.

The implementation of MCCA we considered [240], extends to more than two omics data a sparse CCA implementation, computed by adding a lasso penalty term $||q_i^h||_1$ to the maximization problem defined above. This allows setting to zero the loadings of the biomolecules not showing correlation across the data.

Multiple Co-Inertia Analysis (MCIA)

Multiple Co-Inertia Analysis [138] is a concatenation-based method that projects different omics data in the same lower dimensional space, where similar subjects are located close to each other. The algorithm is implemented in the R package omicade4 [138].

MCIA first pre-processes the omics data by non-symmetric correspondence analysis [101] to scale the measurements x_{ij} of each omics data $X_1 \dots X_k$. Each x_{ij} is firstly made positive by adding the absolute value of the smallest element in the correspondent matrix. Then, for each x_{ij} the method computes three values: (i) the relative contribution $r_i = \frac{\sum_j x_{ij}}{\sum_{ij} x_{ij}}$ of row i over the total variation in the matrix; (ii) the relative contribution $c_j = \frac{\sum_i x_{ij}}{\sum_{ij} x_{ij}}$ of column j over the total variation and (iii) the single element contribution $p_{ij} = \frac{x_{ij}}{\sum_{ij} x_{ij}}$. Finally, omics data elements are scaled as $x_{ij} = \frac{p_{ij}}{r_i} - c_j$.

Once the dataset has been preprocessed, the omics data matrices are concatenated to obtain the matrix $X = [X_1 \dots X_k]$. The algorithm computes, as for MCCA, the $h = \dots r$ latent components, where r is an input parameter of the method. With this aim, a maximization problem is solved by extending co-inertia analysis [41]. For the first latent component the problem is defined as:

$$\max_{q_1^1 \dots q_k^1} \sum_{i=1}^{k} \mathrm{cov}^2(X_i q_i^1, X q^1), \quad \text{with } ||q_i^1|| = 1 \quad \forall = 1 \dots k$$

where the obtained $X_i q_i^1$ are called "block scores". To compute the other latent components, the maximization problem above is solved for $h = 2 \dots r$. In those cases, the problem is applied to residual matrices computed by subtracting, from the matrices X_i, the variance induced by the loadings of the block scores computed in the previous iteration. As for MCCA, this step allows discarding the already processed information.

Multiple Factor Analysis (MFA)

Multiple Factor Analysis [44] is another concatenation-based method, whose aim, similarly to MCIA, is to project data in a lower dimensional space. It is implemented in the FactoMineR [106] R package.

Data columns are firstly scaled to have mean equal to zero and standard deviation equal to one. Separate analyses are then performed by using PCA on each single omics data $X_i, \ \ i = 1 \dots k$ to obtain the eigenvalues of the matrix of covariance-variance associated to each X_i. The first eigenvalue λ_1^i of X_i is finally used to scale the corresponding omics matrix according to the weight $\frac{1}{\lambda_1^i}$ in such a way that the information included in each omics matrix becomes comparable between different omics. Once the single omics data have been scaled, omics matrices are concatenated to obtain the matrix $X = [X_1 \dots X_k]$, which is used as input of the final global analysis, again based on PCA, that projects the concatenated data in a lower dimensional space.

Joint and Individual Variation Explained (JIVE)

Joint and Individual Variation Explained [120] is a concatenation-based method, which assumes that the different k omics data $X_1 \dots X_k$ can be defined in terms of a joint variation structure plus individual patterns. JIVE is implemented in the R package r.jive [163].

Data are firstly scaled to have mean equal to zero and standard deviation equal to one. Then, the concatenated matrix $X = [X_1 \dots X_k]$ of dimension $n \times p$, with $p = p1 + \dots + p_k$ is computed. Finally, the method decomposes each omics data X_i in the sum of three terms:

$$X_i = J_i + A_i + \varepsilon_i, \quad i = 1 \dots k$$

where J is the matrix storing the joint variation structure (J_i is the $n \times p_i$ submatrix of J related to the i-th omics), A_i is the individual pattern of X_i and ε_i gives the residual noise.

JIVE decomposes data by an iterative procedure: at each step, it firstly estimates the rank of J and a_i ($i = 1 \dots k$) and then computes those matrices by low-rank approximations [131] of the concatenated matrix. This step requires minimizing the sum of squared errors of the residuals ε_i with the additional constraint $JA_i^T = 0$ to maintain the orthogonality between the joint and the individual structures. The iterative process ends when convergence is reached, that is, when the estimated ranks of J and A_i ($i = 1 \dots k$) remain equal in two consecutive iterations.

Similarity Network Fusion (SNF)

Similarity Network Fusion [229] is a transformation-based approach implemented in the R package SNFtool [229].

After normalizing data columns to have mean equal to zero and standard deviation equal to one, omics data $X_1 \dots X_k$ are separately transformed in similarity graphs $W^i = V, E$. For each W^i, the vertices V correspond to the samples $x_1, x_2 \dots x_n$.

The weight of the edge in E between two subjects x_h and x_j is computed by scaled exponential kernels (symmetrical and positive semi-definite matrices) as

$$w_{hj}^i = \frac{e^{-\rho^2(x_h, x_j)}}{\sigma \varepsilon_{hj}}$$

In the formula, ρ represents the Euclidean distance between the subjects and ε_{hj} depends on the average distance of x_h and x_j from their K (input parameter of the method) closest neighbors. The parameter σ is another input of the method, taken in the range $[0.3, 0.8]$.

A global similarity matrix $P_{i,0}$, and a local one S_i, are then derived from the W^i. Starting from the k different $P_{i,0}$, the SNF iterative procedure computes, at each step t, an updated set of global matrices:

$$P_{i,t} = S_i \times \left(\frac{\sum_{j \neq i} P_{j,t-1}}{k-1} \right) \times S_i^T \quad \forall i = 1 \dots k$$

After T iterations (input of the method), the weighted adjacency matrix of the similarity fused network is computed as $\frac{\sum_i P_{i,T}}{k}$. The usage of global similarity matrices allows the computed fused network to also include edges with low weight when they are represented in several similarity graphs W^i.

Declaration

I, Giulia Tini, hereby declare that this Ph.D. thesis was carried out by me for the degree of Doctor of Philosophy in Mathematics under the guidance and supervision of Prof. Corrado Priami and Doctor Marie-Pier Scott-Boyer and Doctor Luca Marchetti, Department of Mathematics, University of Trento, Italy and COSBI-The Microsoft Research – University of Trento Centre for Computational Systems Biology, Rovereto, Italy.

I certify that this thesis has not been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Trento and where applicable, any partner institution responsible for the joint-award of this degree.

This work is the intellectual property of the author. You may copy up to 5% of this work for private study, or personal, non-commercial research. Any re-use of the information contained within this document should be fully referenced, quoting the author, title, university, degree level and pagination. Queries or requests for any other use, or if a more substantial copy is required, should be directed in the owner(s) of the Intellectual Property Rights.

I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Giulia Tini
May 2018

Acknowledgements

Grazie a Corrado, Marie Pier e Luca che mi hanno dato l'opportunità di lavorare a COSBI e che mi hanno seguito durante questi anni. Grazie per tutto quello che mi avete insegnato, per il supporto, la pazienza e il tempo speso per me. Grazie anche agli altri ricercatori e amici di COSBI, specialmente a Silvia, Rosario, Nikola, Samanta, Giulia, Verena e Matteo per le chiacchiere scientifiche e quelle più rilassate.

Grazie alla mia famiglia che mi ha insegnato a pensare liberamente, ad essere curiosa e a impegnarmi sempre al meglio in quel che faccio. Grazie a mia mamma Emma, mia zia Mari e mio zio Massi che sono i miei punti di riferimento; a Noemi e Viola che mi fanno sentire sempre felice quando siamo insieme; ai miei nonni Carla e Ennio che mi coccolano e consigliano. A papà, zio Paolo, zio Davide e Giovanni: pensare a voi mi rende sempre un po' più coraggiosa.

Grazie a tutti i miei compagni di ufficio e amici, sia a COSBI che a Povo, che hanno condiviso con me questa avventura. Senza di voi sarebbe stato tutto più difficile e meno divertente. In particolare ad Albe perché ormai ha imparato a capire cosa non mangio e per le lunghe chiacchierate serali; a Cansu per i "Friday-Saturday movies" e i racconti sulle disavventure quotidiane in cucina; a Chia per le nostre playlist indie e per i pomeriggi passati a lavorare, chiacchierare, scambiarci (ottimi) consigli; a Fede per aver sopportato le merende non salutari e per i buoni consigli; a Giù per essere la mia piccola parte di allegra Emilia a Trento e per gli abbracci rincuoranti; a Karla per aver cantato mille volte con me "Mr. Brightside" e per le chiacchiere sul futuro; a Marco per il nostro umorismo così raffinato e per essere sempre pronto ad ascoltarmi; a Rob per il suo divano comodissimo e per aver condiviso con me tante passioni. Grazie a Flavia, che mi ha fatto sentire a casa anche a Trento.

Grazie alle amiche di una vita, che mi sono sempre vicine e perché stare con loro vuol dire essere a casa. Grazie a Feddy, Marti e Monic, che nonostante la lontananza sono sempre pronte ad ascoltare le mie disavventure e grazie a Moni, che mi incoraggia sempre a seguire i miei sogni e le mie idee anche quando vuol dire scegliere la strada più difficile.

Grazie a Lorenzo, che è capace di trasformare i miei insormontabili dubbi in piccole opportunità, che ride ancora alle mie battute e che cerca ogni giorno di farmi vedere la vita come una fantastica avventura.

Bibliography

[1] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science, Springer*, pages 420–434, 2001.

[2] A. Akalin, V. Franke, K. Vlahoviček, C. E. Mason, and D. Schübeler. Genomation: A toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics*, 31(7): 1127–1129, 2015.

[3] G. Alanis-Lobato, C. V. Cannistraci, A. Eriksson, A. Manica, and T. Ravasi. Highlighting nonlinear patterns in population genetics datasets. *Scientific Reports*, 5(1): 8140, 2015.

[4] M. S. Almén et al. Genome-wide analysis reveals DNA methylation markers that vary with both age and obesity. *Gene*, 548(1):61–67, 2014.

[5] S. Alonso-Martin et al. Gene Expression Profiling of Muscle Stem Cells Identifies Novel Regulators of Postnatal Myogenesis. *Frontiers in Cell and Developmental Biology*, 4(June):1–20, 2016.

[6] G. Alterovitz, J. Liu, E. Afkhami, and M. F. Ramoni. Bayesian methods for proteomics. *Proteomics*, 7(16):2843–2855, 2007.

[7] M. J. Aryee et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[8] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300, 1995.

[9] E. Berger et al. Pathways commonly dysregulated in mouse and human obese adipose tissue: FAT/CD36 modulates differentiation and lipogenesis. *Adipocyte*, 4(3):161–180, 2015.

[10] F. Bernhard et al. Functional relevance of genes implicated by obesity genome-wide association study signals for human adipocyte biology. *Diabetologia*, 56(2):311–322, 2013.

[11] M. Bersanelli et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*, 17(S2):S15, dec 2016.

[12] K. Birsoy et al. Analysis of gene networks in white adipose tissue development reveals a role for ETS2 in adipogenesis. *Development*, 138(21):4709–4719, nov 2011.

[13] L. Blanchet et al. Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis. *BMC Bioinformatics*, 12(1):254, 2011.

[14] J. G. Bogner-Strauss et al. Reconstruction of gene association network reveals a transmembrane protein required for adipogenesis and targeted by PPARγ. *Cellular and Molecular Life Sciences*, 67(23):4049–4064, dec 2010.

[15] S. Boluki, M. S. Esfahani, X. Qian, and E. R. Dougherty. Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors. *BMC Bioinformatics*, 18(Suppl 14), 2017.

[16] E. Bonnet, L. Calzone, and T. Michoel. Integrative Multi-omics Module Network Inference with Lemon-Tree. *PLOS Computational Biology*, 11(2):e1003983, feb 2015.

[17] S. E. Bouhaddani et al. Evaluation of O2PLS in Omics data integration. *BMC Bioinformatics*, 17(S2):S11, dec 2016.

[18] G. A. Bray, S. J. Nielsen, and B. M. Popkin. Consumption of high-fructose corn syrup in beverages may play a role in the epidemic of obesity. *Am J Clin Nutr*, 79:537–543, 2004.

[19] R. Bro. Multiway calibration. Multilinear PLS. *Journal of Chemometrics*, 10(1): 47–61, jan 1996.

[20] E. Brody et al. Life's simple measures: Unlocking the proteome. *Journal of Molecular Biology*, 422(5):595–606, 2012.

[21] J. Brosius. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene*, 238(1):115–134, 1999.

[22] J. Brosius. The contribution of RNAs and retroposition to evolutionary novelties. *Genetica*, 118(2-3):99–116, 2003.

[23] J. S. Burns, P. L. Rasmussen, K. H. Larsen, H. D. Schrøder, and M. Kassem. Parameters in Three-Dimensional Osteospheroids of Telomerized Human Mesenchymal (Stromal) Stem Cells Grown on Osteoconductive Scaffolds That Predict In Vivo Bone-Forming Potential. *Tissue Engineering Part A*, 16(7):2331–2342, jul 2010.

[24] B. N. Bursać et al. High-fructose diet leads to visceral adiposity and hypothalamic leptin resistance in male rats - do glucocorticoids play a role? *Journal of Nutritional Biochemistry*, 25(4):446–455, 2014.

[25] J. Campión, F. Milagro, and J. A. Martínez. Epigenetics and Obesity. *Progress in Molecular Biology and Translational Science*, 94:291–347, 2010.

[26] C. V. Cannistraci, T. Ravasi, F. M. Montevecchi, T. Ideker, and M. Alessio. Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes. *Bioinformatics*, 27(13):i531–i539, 2011.

[27] C. V. Cannistraci, G. Alanis-Lobato, and T. Ravasi. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics*, 29(13):199–209, 2013.

[28] G. C. Castellani et al. Systems medicine of inflammaging. *Briefings in Bioinformatics*, 17(3):527–540, may 2016.

[29] A. V. B. Castro, C. M. Kolka, S. P. Kim, and R. N. Bergman. Obesity, insulin resistance and comorbidities - Mechanisms of association. *Arquivos brasileiros de endocrinologia e metabologia*, 58(6):600–609, 2014.

[30] R. Cavill, J. Kleinjans, and J.-J. Briede. DTW4Omics: Comparing Patterns in Biological Time Series. *PLoS ONE*, 8(8):e71823, 2013.

[31] R. Cavill, D. Jennen, J. Kleinjans, and J. J. Briedé. Transcriptomic and metabolomic data integration. *Briefings in Bioinformatics*, 17(5):891–901, sep 2016.

[32] R. Chari, B. P. Coe, E. A. Vucic, W. W. Lockwood, and W. L. Lam. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Systems Biology*, 4(1):67, 2010.

[33] Y. Chen, X. Wu, and R. Jiang. Integrating human omics data to prioritize candidate genes. *BMC Medical Genomics*, 6(1):57, dec 2013.

[34] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(140):1–10, 2007.

[35] S. Ciucci et al. Enlightening discriminative network functional modules behind Principal Component Analysis separation in differential-omic science studies. *Scientific Reports*, 7(March):43946, 2017.

[36] A. Conesa, J. M. Prats-Montalbán, S. Tarazona, M. J. Nueda, and A. Ferrer. A multiway approach to data integration in systems biology based on Tucker3 and N-PLS. *Chemometrics and Intelligent Laboratory Systems*, 104(1):101–111, nov 2010.

[37] P. Cordero, A. M. Gomez-Uriz, J. Campion, F. I. Milagro, and J. A. Martinez. Dietary supplementation with methyl donors reduces fatty liver and modifies the fatty acid synthase DNA methylation profile in rats fed an obesogenic diet. *Genes and Nutrition*, 8(1):105–113, 2013.

[38] P. Cordero, F. I. Milagro, J. Campion, and J. A. Martinez. Maternal methyl donors supplementation during lactation prevents the hyperhomocysteinemia induced by a high-fat-sucrose intake by Dams. *International Journal of Molecular Sciences*, 14 (12):24422–24437, 2013.

[39] F. Crick. Central Dogma of Molecualr Biology. *Nature*, 227(8):561–563, 1970.

[40] D. Croft et al. Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(SUPPL. 1):691–697, 2011.

[41] A. C. Culhane, G. Perrière, and D. G. Higgins. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4(1):59, 2003.

[42] Y. Cun and H. Fröhlich. NetClass: An R-package for network based, integrative biomarker signature discovery. *Bioinformatics*, 30(9):1325–1326, 2014.

[43] A. Daemen et al. A kernel-based integration of genome-wide data for clinical decision support. *Genome Medicine*, 1(4):39, 2009.

[44] M. de Tayrac, S. Le, M. Aubry, J. Mosser, and F. Husson. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics*, 10(1):32, 2009.

[45] E. W. Demerath et al. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Human molecular genetics*, 24(15):4464–79, 2015.

[46] I. B. V. den Veyver. Genetic effects of methylation diets. *Annual Review of Nutrition*, 22(1):255–282, 2002.

[47] G. Dennis Jr et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology*, 4(9):R60, 2003.

[48] K. J. Dick et al. DNA methylation and body-mass index: A genome-wide analysis. *The Lancet*, 383:1990–1998, 2014.

[49] H. Dill, B. Linder, A. Fehr, and U. Fischer. Intronic miR-26b controls neuronal differentiation by repressing its host transcript, ctdsp2. *Genes and Development*, 26 (1):25–30, 2012.

[50] L. Du and A. P. Heaney. Regulation of adipose differentiation by fructose and GluT5. *Molecular endocrinology*, 26(10):1773–82, 2012.

[51] I. Dunham et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[52] M. Dunning, A. Lynch, and M. Eldridge. illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4). R package version 1.26.0., 2016.

[53] T. M. Ebbels and R. Cavill. Bioinformatic methods in NMR-based metabolic profiling. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 55(4):361–374, nov 2009.

[54] J. R. Edwards, H. Ruparel, and J. Ju. Mass-spectrometry DNA sequencing. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 573(1-2):3–12, jun 2005.

[55] B. Efron. Microarrays, Empirical Bayes and the Two-Groups Model. *Statistical Science*, 23(1):1–22, 2008.

[56] P. L. Elkin, M. S. Tuttle, B. E. Trusko, and S. H. Brown. BioProspecting: novel marker discovery obtained by mining the bibleome. *BMC bioinformatics*, 10 Suppl 2:1–8, 2009.

[57] M.-A. Félix and M. Barkoulas. Pervasive robustness in biological systems. *Nature Reviews Genetics*, 16(8):483–496, jul 2015.

[58] K. Fujiki, F. Kano, K. Shiota, and M. Murata. Expression of the peroxisome proliferator activated receptor gamma gene is repressed by DNA methylation in visceral adipose tissue of mouse models of diabetes. *BMC Biology*, 7(1):38, 2009.

[59] B. Fürtig, C. Richter, J. Wöhnert, and H. Schwalbe. NMR Spectroscopy of RNA. *ChemBioChem*, 4(10):936–962, oct 2003.

[60] E. Gabasova, J. Reid, and L. Wernisch. Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLoS Computational Biology*, 13(10):1–29, 2017.

[61] T. Gandhi et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3):285–293, 2006.

[62] L. A. Garraway et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*, 436(7047):117–122, jul 2005.

[63] K. Glass, C. Huttenhower, J. Quackenbush, and G.-C. Yuan. Passing Messages between Biological Networks to Refine Predicted Interactions. *PLoS ONE*, 8(5): e64832, may 2013.

[64] L. Gold, J. J. Walker, S. K. Wilcox, and S. Williams. Advances in human proteomics at high scale with the SOMAscan proteomics platform. *New Biotechnology*, 29(5): 543–549, 2012.

[65] L. Gold et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS ONE*, 5(12), 2010.

[66] D. Gomez-Cabrero et al. Data integration in the era of omics: current and future challenges. *BMC Systems Biology*, 8(Suppl 2):I1, 2014.

[67] I. Gonzalez et al. Highlighting Relationships Between Heteregeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis. *Journal of Biological Systems*, 17(02):173–199, jun 2009.

[68] C. R. Green et al. Branched-chain amino acid catabolism fuels adipocyte differentiation and lipogenesis. *Nature Chemical Biology*, 12(1):15–21, 2016.

[69] B. Gustafson and U. Smith. The WNT inhibitor dickkopf 1 and bone morphogenetic protein 4 rescue adipogenesis in hypertrophic obesity in humans. *Diabetes*, 61(5): 1217–1224, 2012.

[70] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–D517, dec 2004.

[71] J. Han, A. M. Denli, and F. H. Gage. The enemy within: intronic miR-26b represses its host gene, ctdsp2, to regulate neurogenesis. *Genes & Development*, 26(1):6–10, jan 2012.

[72] Hansen KD. IlluminaHumanMethylation450kanno.ilmn12.hg19, 2016.

[73] Y. Hasin, M. Seldin, and A. Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, 2017.

[74] Z. M. Hira and D. F. Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in Bioinformatics*, 2015(1), 2015.

[75] B. P. Hodkinson and E. A. Grice. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Advances in Wound Care*, 4(1): 50–58, 2015.

[76] H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321–377, 1936.

[77] A. E. House and K. W. Lynch. Regulation of alternative splicing: More than just the ABCs. *Journal of Biological Chemistry*, 283(3):1217–1221, 2008.

[78] J. Hsu and J. Sage. Novel functions for the transcription factor E2F4 in development and disease. *Cell Cycle*, 15(23):3183–3190, 2016.

[79] S. Huang, K. Chaudhary, and L. X. Garmire. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in Genetics*, 8(June):84, jun 2017.

[80] H. Hui et al. Direct Spectrophotometric Determination of Serum Fructose in Pancreatic Cancer Patients. *Pancreas*, 38(6):706–712, aug 2009.

[81] T. Inagaki et al. The FBXL10/KDM2B Scaffolding Protein Associates with Novel Polycomb Repressive Complex-1 to Regulate Adipogenesis. *Journal of Biological Chemistry*, 290(7):4163–4177, feb 2015.

[82] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[83] J.-P. Issa. Cpg-island methylation in aging and cancer. *Current topics in microbiology and immunology*, 249:101–18, 02 2000.

[84] R. Jaenisch and A. Bird. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(3S):245–254, 2003.

[85] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, 2013.

[86] M. A. Jimenez, P. Akerblad, M. Sigvardsson, and E. D. Rosen. Critical Role for Ebf1 and Ebf2 in the Adipogenic Transcriptional Cascade. *Molecular and Cellular Biology*, 27(2):743–757, jan 2007.

[87] B. Jin, Y. Li, and K. D. Robertson. DNA methylation: Superior or subordinate in the epigenetic hierarchy? *Genes and Cancer*, 2(6):607–617, 2011.

[88] I. Jolliffe. *Principal component analysis*. Springer, second edition, 2002.

[89] P. A. Jones. The DNA methylation paradox The methylation of CpG islands is often equated with transcriptional inactivity and there is overwhelming. *Trends in genetics : TIG*, 15(1):34–37, 1999.

[90] P. A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, 13(7):484–92, 2012.

[91] R. P. Kandpal, B. Saviola, and J. Felton. The era of 'omics unlimited. *BioTechniques*, 46(5 SPEC. ISSUE):351–355, 2009.

[92] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1): 109–114, 2012.

[93] K. J. Karczewski and M. P. Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 2018.

[94] Z. Khitan and D. H. Kim. Fructose : A Key Factor in the Development of Metabolic Syndrome and Hypertension. *Journal of Nutrition and Metabolism*, 2013:1–12, 2013.

[95] D. Kim et al. Knowledge-driven genomic interactions: an application in ovarian cancer. *BioData Mining*, 7(1):20, 2014.

[96] D. Kim et al. Using knowledge-driven genomic interactions for multi-omics data analysis: Metadimensional models for predicting clinical outcomes in ovarian carcinoma. *Journal of the American Medical Informatics Association*, 24(3):577–587, 2017.

[97] M. Kim, N. Rai, V. Zorraquino, and I. Tagkopoulos. Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. *Nature Communications*, 7:13090, oct 2016.

[98] A. Klip, Y. Sun, T. T. Chiu, and K. P. Foley. Signal transduction meets vesicle traffic: the software and hardware of GLUT4 translocation. *AJP: Cell Physiology*, 306(10): C879–C886, 2014.

[99] M. Koc et al. Stress of endoplasmic reticulum modulates differentiation and lipogenesis of human adipocytes. *Biochemical and Biophysical Research Communications*, 460(3):684–690, 2015.

[100] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson. Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics*, 82(April):949–958, 2008.

[101] P. M. Kroonenberg and R. Lombardo. Nonsymmetric Correspondence Analysis: A Tool for Analysing Contingency TablesWith a Dependence Structure. *Multivariate Behavioral Research*, 34(3):367–396, jul 1999.

[102] T.-C. Kuo, T.-F. Tian, and Y. Tseng. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Systems Biology*, 7(1):64, 2013.

[103] P. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, 2008.

[104] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99*, pages 16–22, New York, New York, USA, 1999. ACM Press.

[105] K. A. Lê et al. Fructose overconsumption causes dyslipidemia and ectopic lipid deposition in healthy subjects with and without a family history of type 2 diabetes. *American Journal of Clinical Nutrition*, 89(6):1760–1765, 2009.

[106] S. Lê, J. Josse, and F. Husson. FactoMineR : An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1):1–18, 2008.

[107] K.-A. Le Cao, D. Rossow, C. Robert-Granié, and P. Besse. A Sparse PLS for Variable Selection when Integrating Omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1):pp. 35, 2008.

[108] K.-A. Le Cao, I. Gonzalez, and S. Dejean. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, 25(21):2855–2856, nov 2009.

[109] D. Y. Lee, C. Teyssier, B. D. Strahl, and M. R. Stallcup. Role of protein methylation in regulation of transcription. *Endocrine Reviews*, 26(2):147–170, 2005.

[110] E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, and D. Lee. Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, 4(11), 2008.

[111] M. I. Lefterova and M. A. Lazar. New developments in adipogenesis. *Trends in Endocrinology and Metabolism*, 20(3):107–114, 2009.

[112] G. Lev Maor, A. Yearim, and G. Ast. The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, 31(5):274–280, 2015.

[113] W. Li, M. Fan, and M. Xiong. SamCluster: an integrated scheme for automatic discovery of sample classes using gene expression profile. *Bioinformatics*, 19(7): 811–817, may 2003.

[114] W. Li, S. Zhang, C.-C. Liu, and X. J. Zhou. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, 28(19):2458–2466, oct 2012.

[115] W. Li et al. Integrative Analysis of Many Weighted Co-Expression Networks Using Tensor Computation. *PLoS Computational Biology*, 7(6):e1001106, jun 2011.

[116] Y. Li and J. C. Patra. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224, 2010.

[117] D. Lin et al. Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics*, 14(1):245, 2013.

[118] B. Liu, X. Shen, and W. Pan. Integrative and regularized principal component analysis of multiple sources of data. *Statistics in Medicine*, 35(13):2235–2250, jun 2016.

[119] G. Liu et al. Regulation of hepatic lipogenesis by the zinc finger protein Zbtb20. *Nature Communications*, 8:14824, mar 2017.

[120] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542, mar 2013.

[121] T. Löfstedt, M. Hanafi, G. Mazerolles, and J. Trygg. OnPLS path modelling. *Chemometrics and Intelligent Laboratory Systems*, 118:139–149, aug 2012.

[122] T. Löfstedt, D. Hoffman, and J. Trygg. Global, local and unique decompositions in OnPLS for multiblock data analysis. *Analytica Chimica Acta*, 791:13–24, aug 2013.

[123] E. R. Londin et al. The human platelet: strong transcriptome correlations among individuals associate weakly with the platelet proteome. *Biology Direct*, 9(1):3, 2014.

[124] C. E. Lowe, S. O'Rahilly, and J. J. Rochford. Adipogenesis at a glance. *Journal of Cell Science*, 124(16):2681–2686, aug 2011.

[125] H. Lu et al. FGF13 regulates proliferation and differentiation of skeletal muscle by down-regulating Spry1. *Cell Proliferation*, 48(5):550–560, oct 2015.

[126] R. H. Lustig. Fructose: Metabolic, Hedonic, and Societal Parallels with Ethanol. *Journal of the American Dietetic Associaiton*, 110(9):1307–1321, 2010.

[127] D. J. C. MacKay. *Information Theory and Learning Algorithms*, volume 22. Cambridge University Press, 2003.

[128] J. Maksimovic, L. Gordon, and A. Oshlack. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome biology*, 13(6):R44, 2012.

[129] J. Mariette and N. Villa-vialaneix. Integrating TARA Oceans datasets using unsupervised multiple kernel learning. pages 1–16, 2017.

[130] J. Mariette and N. Villa-Vialaneix. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, (December):1–7, 2017.

[131] I. Markovsky. *Low Rank Approximation*. Springer, 2008.

[132] H. Martens and T. Naes. *Multivariate calibration*. Wiley, Chichester, 2nd edition, 1989.

[133] C. M. Mayers et al. The Rho Guanine Nucleotide Exchange Factor AKAP13 (BRX) Is Essential for Cardiac Development in Mice. *Journal of Biological Chemistry*, 285 (16):12344–12354, apr 2010.

[134] Y. A. Medvedeva et al. Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics*, 15(1):1–12, 2014.

[135] K. Meijer et al. Human primary adipocytes exhibit immune cell function: Adipocytes prime inflammation independent of macrophages. *PLoS ONE*, 6(3), 2011.

[136] A. Meissner et al. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877, 2005.

[137] C. Mejia-Pous, F. Damiola, and O. Gandrillon. Cholesterol synthesis-related enzyme oxidosqualene cyclase is required to maintain self-renewal in primary erythroid progenitors. *Cell Proliferation*, 44(5):441–452, 2011.

[138] C. Meng, B. Kuster, A. C. Culhane, and A. Gholami. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, 15(1):162, 2014.

[139] C. Meng, D. Helm, M. Frejno, and B. Kuster. moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. *Journal of Proteome Research*, 15(3):755–765, mar 2016.

[140] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4):628–641, jul 2016.

[141] G. Menichetti, D. Remondini, P. Panzarasa, R. J. Mondragón, and G. Bianconi. Weighted multiplex networks. *PLoS ONE*, 9(6):6–13, 2014.

[142] M. Merkel, R. H. Eckel, and I. J. Goldberg. Lipoprotein lipase: genetics, lipid uptake, and regulation. *The Journal of Lipid Research*, 43(12):1997–2006, 2002.

[143] F. I. Milagro et al. CLOCK, PER2 and BMAL1 DNA Methylation: Association with Obesity and Metabolic Syndrome Characteristics and Monounsaturated Fat Intake. *Chronobiology International*, 29(9):1180–1194, 2012.

[144] B. Min, S. D. Yi, K.-M. Lee, and K.-I. Goh. Network robustness of multiplex networks with interlayer degree correlations. *Physical Review E*, 89(4):042811, apr 2014.

[145] P. Mirmiran, E. Yuzbashian, G. Asghari, S. Hosseinpour-Niazi, and F. Azizi. Consumption of sugar sweetened beverage is associated with incidence of metabolic syndrome in Tehranian children and adolescents. *Nutrition & Metabolism*, 12:25, 2015.

[146] Y. Miyamoto, J. Yamauchi, A. Sanbe, and A. Tanoue. Dock6, a Dock-C subfamily guanine nucleotide exchanger, has the dual specificity for Rac1 and Cdc42 and regulates neurite outgrowth. *Experimental Cell Research*, 313(4):791–804, feb 2007.

[147] Q. Mo et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, mar 2013.

[148] J. M. Moreno-Navarrete et al. Metabolomics uncovers the role of adipose tissue PDXK in adipogenesis and systemic insulin sensitivity. *Diabetologia*, 59(4):822–832, apr 2016.

[149] E. Mosca and L. Milanesi. Network-based analysis of omics with multi-objective optimization. *Mol. BioSyst.*, 9:2971–2980, 2013.

[150] A. A. Motsinger, S. M. Dudek, L. W. Hahn, and M. D. Ritchie. Comparison of neural network optimization approaches for studies of human genetics. *Applications of Evolutionary Computing, Proceedings*, 3907:103–114, 2006.

[151] L. Mourino-Alvarez, C. Laborde, and M. Barderas. Proteomics and metabolomics in aortic stenosis: Studying healthy valves for a better understanding of the disease. In E. Aikawa, editor, *Calcific Aortic Valve Disease*, chapter 6. InTech, Rijeka, 2013.

[152] T. Moyon et al. Statistical strategies for relating metabolomics and proteomics data: a real case study in nutrition research area. *Metabolomics*, 8(6):1090–1101, dec 2012.

[153] K. Münstedt, M. Böhme, A. Hauenschild, and I. Hrgovic. Consumption of rapeseed honey leads to higher serum fructose levels compared with analogue glucose/fructose solutions. *European Journal of Clinical Nutrition*, 65(1):77–80, 2011.

[154] I. A. Myles. Fast food fever : reviewing the impacts of the Western diet on immunity. *Nutrition Journal*, 13(1):61, 2014.

[155] M. Nagai and Y. Yoneda. Small GTPase Ran and Ran-binding proteins. *Biomolecular Concepts*, 3(4):307–318, 2012.

[156] C. Nardini, J. Dent, and P. Tieri. Editorial: Multi-omic data integration. *Frontiers in Cell and Developmental Biology*, 3(July):2014–2015, 2015.

[157] I. Nassiri et al. Systems view of adipogenesis via novel omics-driven and tissue-specific activity scoring of network unctional modules. *Scientific Reports*, 6:28851, 2016.

[158] S. B. Ng et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, 2009.

[159] E. Nilsson et al. Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes*, 63(9):2962–2976, 2014.

[160] D. Noble. A theory of biological relativity: no privileged level of causation. *Interface Focus*, 2(1):55–64, feb 2012.

[161] A. Noer, A. C. Boquest, and P. Collas. Dynamics of adipogenic promoter DNA methylation during clonal culture of human adipose stem cells to senescence. *BMC Cell Biology*, 8:18, 2007.

[162] L. Nørgaard, R. Bro, F. Westad, and S. B. Engelsen. A modification of canonical variates analysis to handle highly collinear multivariate data. *Journal of Chemometrics*, 20(8-10):425–435, aug 2006.

[163] M. J. O'Connell and E. F. Lock. R.JIVE for exploration of multi-source molecular data. *Bioinformatics*, 32(18):2877–2879, sep 2016.

[164] K. Ohashi et al. High fructose consumption induces DNA methylation at PPARα and CPT1A promoter regions in the rat liver. *Biochemical and Biophysical Research Communications*, 468:185–189, 2015.

[165] R. Ostroff et al. The stability of the circulating human proteome to variations in sample collection and handling procedures measured with an aptamer-based proteomics array. *Journal of Proteomics*, 73(3):649–666, 2010.

[166] J. Pages. Multiple Factor Analysis: Main Features and Application to Sensory Data. *Revista Colombian de Estadistica*, 27(1):1–26, 2004.

[167] D. Pe'er. Bayesian network analysis of signaling networks: a primer. *Signal Transduction Knowledge Environment*, 2005(281):pl4, 2005.

[168] S. Peri et al. Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans. *Genome Research*, 13(10):2363–2371, oct 2003.

[169] K. H. Pietiläinen et al. Global transcript profiles of fat in monozygotic twins discordant for BMI: Pathways behind acquired obesity. *PLoS Medicine*, 5(3):0472–0483, 2008.

[170] M. Piwowar and W. Jurkowski. ONION: Functional Approach for Integration of Lipidomics and Transcriptomics Data. *PLOS ONE*, 10(6):e0128854, jun 2015.

[171] B. Pulverer, A. Sommer, G. A. McArthur, R. N. Eisenman, and B. Luscher. Analysis of Myc/Max/Mad network members in adipogenesis: Inhibition of the proliferative burst and differentiation by ectopically expressed Mad1. *Journal of Cellular Physiology*, 183(3):399–410, jun 2000.

[172] N. Qiu, L. Cao, V. David, L. D. Quarles, and Z. Xiao. Kif3a Deficiency Reverses the Skeletal Abnormalities in Pkd1 Deficient Mice by Restoring the Balance Between Osteogenesis and Adipogenesis. *PLoS ONE*, 5(12):e15240, dec 2010.

[173] A. Regassa and W. Kim. Transcriptome analysis of hen preadipocytes treated with an adipogenic cocktail (DMIOA) with or without 20(S)-hydroxylcholesterol. *BMC Genomics*, 16(1):91, 2015.

[174] M. W. Revelle. Package ' psych. *October*, pages 1–250, 2011.

[175] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, jan 2015.

[176] M. D. Ritchie, J. R. Davis, H. Aschard, A. Battle, D. Conti, M. Du, E. Eskin, M. D. Fallin, L. Hsu, P. Kraft, J. H. Moore, B. L. Pierce, S. A. Bien, D. C. Blair Thomas, P. Wei, and S. B. Montgomery. Incorporation of Biological Knowledge into the Study of Gene-Environment Interactions. *American Journal of Epidemiology*, 186(7): 771–777, 2017.

[177] K. D. Robertson. DNA methylation and human disease. *Nature Reviews Genetics*, 6 (8):597–610, 2005.

[178] F. Rohart, A. Eslami, N. Matigian, S. Bougeard, and K.-A. L. Cao. MINT: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *bioRxiv*, 2:070813, 2016.

[179] E. Ron et al. Bypass of glycan-dependent glycoprotein delivery to ERAD by upregulated EDEM1. *Molecular biology of the cell*, 22(21):3945–54, 2011.

[180] P. Roworth, F. Ghari, and N. B. La Thangue. To live or let die – complexity within the E2F1 pathway. *Molecular & Cellular Oncology*, 2(1):e970480, 2015.

[181] H. Ruffieux, A. C. Davison, J. Hager, and I. Irincheeva. Efficient inference for genetic association studies with multiple outcomes. *Biostatistics (Oxford, England)*, 45(2): 846–860, 2017.

[182] L. Saleh and F. B. Perler. Protein splicing in cis and in trans. *Chemical Record*, 6(4): 183–193, 2006.

[183] E. E. Schadt et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717, jul 2005.

[184] J. C. Schell et al. Control of intestinal stem cell function and proliferation by mitochondrial pyruvate metabolism. *Nature Cell Biology*, 19(9):1027–1036, aug 2017.

[185] B. Scholkopf, K. Tsuda, and J.-P. Vert. *Kernel methods in Computational Biology*. MIT Press, 2004.

[186] M. Schouteden, K. Van Deun, T. F. Wilderjans, and I. Van Mechelen. Performing DISCO-SCA to search for distinctive and common information in linked data. *Behavior Research Methods*, 46(2):576–87, nov 2013.

[187] A. Schulze and J. Downward. Navigating gene expression using microarrays — a technology review. *Nature Cell Biology*, 3(8):E190–E195, 2001.

[188] N. Shah and S. Mahajan. Document Clustering : A Detailed Review. *International Journal of Applied Information Systems*, 4(5):30–38, 2012.

[189] P. Shannon et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[190] J. A. Shapiro. Revisiting the central dogma in the 21st century. *Annals of the New York Academy of Sciences*, 1178:6–28, 2009.

[191] R. Shen, A. B. Olshen, and M. Ladanyi. 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, nov 2009.

[192] R. Shen et al. Integrative Subtype Discovery in Glioblastoma Using iCluster. *PLoS ONE*, 7(4):e35236, apr 2012.

[193] A. Singh et al. DIABLO – an integrative, multi-omics, multivariate method for multi-group classification. *bioRxiv*, page 067611, 2016.

[194] A. Smilde, R. Bro, and P. Geladi. *Multi-Way Analysis with Applications in the Chemical Sciences*. Number September. John Wiley & Sons, Ltd, Chichester, UK, aug 2004.

[195] M. L. Smith, K. A. Baggerly, H. Bengtsson, M. E. Ritchie, and K. D. Hansen. illuminaio: An open source IDAT parsing tool for Illumina microarrays. *F1000Research*, 2: 264, 2013.

[196] G. K. Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–26, 2004.

[197] F. Song et al. Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proceedings of the National Academy of Sciences*, 102(9):3336–3341, 2005.

[198] Y. Song, V. Kumar, H.-X. Wei, J. Qiu, and P. S. Kumar. Lunatic, Manic and Radical Fringe Each Promote T and B Cell Development. *J Immunology*, 196(1):232 – 243, 2016.

[199] N. K. Speicher and N. Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12):i268–i275, jun 2015.

[200] K. L. Stanhope et al. Consuming fructose-sweetened, not glucose-sweetened, beverages increase visceral adiposity and lipids and decrease insulin sensitivity in overweight/obese men. *Journal of Clinical Investigation*, 1334(5):1322–1334, 2009.

[201] T. Tanaka, N. Yoshida, T. Kishimoto, and S. Akira. Defective adipocyte differentiation in mice lacking the C / EBP β and / or C / EBP δ gene. *The EMBO Journal*, 16(24): 7432–7443, 1997.

[202] H. S. Tapp, M. Radonjic, E. Kate Kemsley, and U. Thissen. Evaluation of multiple variate selection methods from a biological perspective: a nutrigenomics case study. *Genes & Nutrition*, 7(3):387–397, jul 2012.

[203] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, sep 2012.

[204] The Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[205] The Gene Ontology Consortium. Gene ontologie: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[206] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996.

[207] G. Tini, L. Marchetti, C. Priami, and M.-P. Scott-Boyer. Multi-omics integration—a comparison of unsupervised clustering methodologies. *Briefings in Bioinformatics*, (November):1–11, 2017.

[208] O. A. Tomescu, D. Mattanovich, and G. G. Thallinger. Integrative omics analysis. A study based on Plasmodium falciparum mRNA and protein data. *BMC Systems Biology*, 8(Suppl 2):S4, 2014.

[209] N. Tuncbag et al. Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLOS Computational Biology*, 12(4): e1004879, apr 2016.

[210] G. Tuncman et al. A genetic variant at the fatty acid-binding protein aP2 locus reduces the risk for hypertriglyceridemia, type 2 diabetes, and cardiovascular disease. *PNAS*, 103(18):6970–6975, 2006.

[211] M. Ullah, M. Sittinger, and J. Ringe. Extracellular matrix of adipogenically differentiated mesenchymal stem cells reveals a network of collagen filaments, mostly interwoven by hexagonal structural units. *Matrix Biology*, 32(7):452 – 465, 2013.

[212] S. Urs et al. Sprouty1 is a critical regulatory switch of mesenchymal stem cell lineage allocation. *FASEB Journal*, 24(9):3264 – 3272, 2010.

[213] M. van den Dungen, A. Murk, D. Kok, and W. Steegenga. Comprehensive DNA Methylation and Gene Expression Profiling in Differentiating Human Adipocytes. *Journal of Cellular Biochemistry*, 12(April):1–12, 2016.

[214] M. W. van den Dungen, A. J. Murk, D. E. Kok, and W. T. Steegenga. Persistent organic pollutants alter DNA methylation during human adipocyte differentiation. *Toxicology in Vitro*, 40:79–87, 2017.

[215] H. H. van Haagen et al. Novel protein-protein interactions inferred from literature context. *PLoS ONE*, 4(11), 2009.

[216] C. J. Van Rijsbergen. Foundation of Evaluation. *Journal of Documentation*, 22(3): 266–268, mar 1966.

[217] A. H. van Vliet. Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiology Letters*, 302(1):1–7, jan 2010.

[218] V. Varma et al. Metabolic fate of fructose in human adipocytes: a targeted 13 C tracer fate association study. *Metabolomics*, 11:529–544, 2015.

[219] V. Varma et al. Fructose alters intermediary metabolism of glucose in human adipocytes and diverts glucose to serine oxidation in the one-carbon cycle energy producing pathway. *Metabolites*, 5(2):364–385, 2015.

[220] C. J. Vaske et al. Inference of patient-specific pathway activities from multidimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–i245, jun 2010.

[221] B. F. Voight et al. The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLoS Genetics*, 8(8):1–12, 2012.

[222] S. Voisin et al. Many obesity-associated SNPs strongly associate with DNA methylation changes at proximal promoters and enhancers. *Genome medicine*, 7(1):103, 2015.

[223] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, dec 2007.

[224] P. N. Wahjudi et al. Measurement of glucose and fructose in clinical samples using gas chromatography/mass spectrometry. *Clinical Biochemistry*, 43(1-2):198–207, 2010.

[225] S. Wahl et al. Multi-omic signature of body weight change: results from a population-based cohort study. *BMC Medicine*, 13(1):48, dec 2015.

[226] S. Wahl et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*, 541:81–86, 2016.

[227] E. Wainfan and L. a. Poirier. Methyl Groups in Carcinogenesis : Effects on DNA Methylation and Gene. pages 2071–2078, 1992.

[228] B. Wang, J. Jiang, W. Wang, Z. H. Zhou, and Z. Tu. Unsupervised metric fusion by cross diffusion. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2997–3004, 2012.

[229] B. Wang et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, jan 2014.

[230] D. Wang and J. Gu. Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quantitative Biology*, 4(1):58–67, mar 2016.

[231] J. Wang et al. Pathway and Network Approaches for Identification of Cancer Signature Markers from Omics Data. *Journal of Cancer*, 6(1):54–65, 2015.

[232] L. Wang et al. Integrating Multi-Omics for Uncovering the Architecture of Cross-Talking Pathways in Breast Cancer. *PLoS ONE*, 9(8):e104282, aug 2014.

[233] W. Wang, V. Baladandayuthapani, C. C. Holmes, and K.-A. Do. Integrative network-based Bayesian analysis of diverse genomics data. *BMC Bioinformatics*, 14(Suppl 13):S8, 2013.

[234] C. D. Warden et al. COHCAP: An integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Research*, 41(11):e117, 2013.

[235] B. Weinhold. Epigenetics: the science of change. *Environmental Health Perspectives*, 114(3):160–167, 2006.

[236] P. Wiklund et al. Insulin resistance is associated with altered amino acid metabolism and adipose tissue dysfunction in normoglycemic women. *Scientific Reports*, 6(April): 1–11, 2016.

[237] F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6): 80, 1945.

[238] E. G. Williams et al. Systems proteomics of liver mitochondria function. *Science*, 352 (6291):aad0189–aad0189, jun 2016.

[239] D. M. Witten and R. J. Tibshirani. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.

[240] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, jul 2009.

[241] H. Wold. Estimation of Principal Components and Related Models by Iterative Least squares. *Academic Press, New York*, pages 391–420, 1966.

[242] G. L. Wolff, R. L. Kodell, S. R. Moore, and C. A. Cooney. Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. *Faseb J.*, 12(11): 949–957, 1998.

[243] D. Wu, D. Wang, M. Q. Zhang, and J. Gu. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics*, 16(1):1022, dec 2015.

[244] X. Yang et al. Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. *Cancer Cell*, 26(4):577–590, oct 2014.

[245] L. Yao, H. Shen, P. W. Laird, P. J. Farnham, and B. P. Berman. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biology*, 16:105, 2015.

[246] L. Yao et al. ELMER : An R / Bioconductor Tool Inferring Regulatory Element Landscapes and Transcription Factor Networks Using Methylomes. pages 1–19, 2016.

[247] L. Yao et al. ELMER . data : Supporting data for the ELMER package. pages 1–5, 2016.

[248] B. C. Yoo, K. H. Kim, S. M. Woo, and J. K. Myung. Clinical multi-omics strategies for the effective cancer management. *Journal of Proteomics*, (July):0–1, 2017.

[249] S. Zakhari. Alcohol metabolism and epigenetics changes. *Alcohol research : current reviews*, 35(1):6–16, 2013.

[250] H. Zha, X. He, C. Ding, and H. Simon. Spectral Relaxation for K-means Clustering. *MIT Press*, pages 1057—-1064, 2001.

[251] X. Zhang, L. Bao, L. Yang, Q. Wu, and S. Li. Roles of intracellular fibroblast growth factors in neural development and functions. *Science China Life Sciences*, 55(12): 1038–1044, dec 2012.

[252] Z. Zhang et al. Molecular Subtyping of Serous Ovarian Cancer Based on Multi-omics Data. *Scientific Reports*, 6(1):26001, sep 2016.

[253] J. Zhong et al. Temporal profiling of the secretome during adipogenesis in humans. *Journal of Proteome Research*, 9(10):5228–5238, 2010.

[254] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, apr 2005.

[255] M. G. Zubiría et al. Long-Term Fructose Intake Increases Adipogenic Potential: Evidence of Direct Effects of Fructose on Adipocyte Precursor Cells. *Nutrients*, 8(4): 198, jan 2016.

[256] A. Zufferey et al. New molecular insights into modulation of platelet reactivity in aspirin-treated patients using a network-based approach. *Human Genetics*, 135(4): 403–414, apr 2016.

[257] Y. Zuo, Y. Cui, G. Yu, R. Li, and H. W. Ressom. Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO. *BMC Bioinformatics*, 18(1):1–14, 2017.