Review

Approaches to Integrating Metabolomics and Multi-Omics Data: A Primer

Takoua Jendoubi [1,†,‡]*

[1] University College London; t.jendoubi@ucl.ac.uk

Abstract: Metabolomics deals with multiple and complex chemical reactions within living organisms and how these are influenced by external or internal perturbations. It lies at the heart of omics profiling technologies not only as the underlying biochemical layer that reflects information expressed by the genome, the transcriptome and the proteome, but also as the closest layer to the phenome. The combination of metabolomics data with the information available from genomics, transcriptomics, and proteomics offers unprecedented possibilities to enhance current understanding of biological functions, elucidate their underlying mechanisms and uncover hidden associations between omics variables. As a result, a vast array of computational tools have been developed to assist with integrative analysis of metabolomics data with different omics. Here, we review and propose five criteria – hypothesis, data types, strategies, study design and study focus – to classify statistical multi-omics data integration approaches into state-of-the-art classes under which all existing statistical methods fall. The purpose of this review is to look at various aspects that lead the choice of the statistical integrative analysis pipeline in terms of the different classes. We will draw a particular attention to metabolomics and genomics data to assist those new to this field in the choice of the integrative analysis pipeline.

Keywords: Data integration; multi-omics; integration strategies; genomics

1. Introduction

Biological processes and functions are the result of multiple interactions between tens of thousands of molecules and are inherently complex. In the last 30 years, the parallel acquisition of high-throughput multi-omics datasets from the metabolome, genome, epigenome, proteome, and transcriptome has seen a tremendous boost. As a result, integrative analysis methods for multi-omics data are emerging and gaining popularity among researchers. Integrative analysis consists in the combination of the information available from multi-omics data to provide an enhanced readout of cellular processes and molecular programmes in multiple fields encompassing plant biology [1], animal science [2], toxicology [3,4], molecular epidemiology [5,6] and complex diseases [7,8].

The genome, metabolome, proteome, and transcriptome form different layers of the so-called omics cascade, each of which characterizes a biosystem or an organism at different biomolecular levels [9]. The terms "multi-omics" or "cross-omics" are often used to reflect the heterogeneity of biomolecular profiles and complexity of omics layers they try to measure. Integrating different omics profiles helps extract insightful information and appreciate more comprehensive snapshots of biological systems and molecular processes. Integrative analysis has been applied to associate omic entities to a phenotype of interest e.g. cardiovascular disease [10], cancer [11], or a given treatment or intervention [12]. Other applications of multi-omics analysis include biomarker discovery [13], cross-omics biomarker discovery [14,15], patient stratification [16,17], and functional analysis [18,19].

In fact, the need for data integration is naturally explained by the complex processes involving e.g. genetic variants, microorganisms, post-translational modifications, metabolic processes and the biological interrelationship between the different types of omic entities –the collection of which determines the biological state of a living organism [20,21]. In the early 2000s, multi-omic studies of genomic and metabolomic data have identified a number of alleles that explain a significant proportion of the variation in the metabolic profile [15,22,23]. Similarly, large population studies have linked sequence variations to changes in lipid profiles [24]. Conversely, metabolites can be involved in consequential reactions reaching as deep as cell building blocks [25]. For instance, metabolic fingerprints can help pinpoint genes that affect metabolism and provide functional insight by mapping back to the function of the gene [26]. Similarly, metabolites contribute into reinforcing gene annotations by identifying downstream targets from a specific gene [27].

An arsenal of mathematical and computational techniques has been developed to achieve integrative analyses ranging from least squares-based models to Bayesian models [28] and deep learning models [29]. In the era of high-throughput data, it becomes necessary to look into the fundamentals of integrating multi-omics data to provide early career researchers with a guidance to choose and develop multi-omics data integration methods. In this review, we focus on principles of integrative analysis using five criteria, hypothesis, data types, strategies, study design and study focus to assist early career researchers in the choice of options that integrative analysis offers. Based on these criteria, we also examine types of statistical data integration under which all existing methods fall. Table 1 provides a review of this primer on data integration and summarizes the different categories that we discuss later. We acknowledge that there are various surveys on statistical multi-omics integrative methods in the literature [30**?** –33]; however, these often focus on one specific aspect of data integration. By contrast, our review covers more comprehensive discussions on a higher level about the heuristics of data integration and the choices relevant to the integrative analysis process. In the following section we discuss the challenges that arise when combining multi-omics data. In the next sections we shall examine data integration methodologies according to the five criteria: study design, hypothesis, data types, strategies and study focus respectively.

2. Challenges in metabolomics and multi-omics data integration

When dealing with metabolomics data for integrative analysis, multiple challenges arise and these are in some cases shared with the other omics. On a first instance, omics are not fully characterized. Profiling technologies in metabolomics are subject to an incredibly complex chemical heterogeneity where metabolites are typically not easily identified. The metabolome, in fact, is characterized by a high diversity comprising thousands to hundreds of thousands of chemicals [51]. As a consequence, unknown metabolic entities might be entities that are not identified but might also represent chemicals that have not been reported in the literature. On the other hand, genomic variables are not usually fully characterized by the profiling technology and require an annotation step. Gene annotation is subject to two major bottlenecks: Identifying elements on the genome (Gene finding) and adding biological information to these elements (Gene function). Uncovering the function of genes is critical to understanding its biological roles and corresponding cellular mechanisms. This challenge in the characterization of variable identities is not only likely to induce biases in interpretability but can also lead to uninterpretable results [52].

On a second instance, distinct omics datasets have their own limitations and require complex analysis pipelines prior to performing data integration. For instance, analysis of methylation data is complicated by the uneven distribution of methylation target sequences across the genome requiring specific normalization and scaling strategies [53]. Each omics platform faces unique challenges such as experimental and inherent biological noise, differences among experimental platforms and detection bias [54]. In a similar vein to processing genomic data, a supplementary step is critical to ensure

Integrative analysis	Description	Examples
hypothesis		
Multi-staged	Inter-omic variation is hierarchical	Nicholson *et al.* [15], Gieger *et al.* [22], Krumsiek *et al.* [34]
Meta-dimensional	Inter-omic variation is simultaneous	Smolinska *et al.* [35], Witten and Tibshirani [36], Daemen *et al.* [37]
data types		
Horizontal or homogeneous	Same omic entities are measured across various cohorts, labs or studies	Richardson *et al.* [30], Yuan *et al.* [38]
Vertical or heterogeneous	Entities from different omic levels are measured using different platforms	Richardson *et al.* [30], Evangelou and Ioannidis [39]
strategy		
Early integration	Concatenation-based integration	Fridley *et al.* [40], Le Van *et al.* [41]
Intermediate integration	Data transformation needs to be applied prior to modeling	Smolinska *et al.* [35], Lanckriet *et al.* [42], Guo *et al.* [43]
Late integration	Based on combining single data models into a high level model	Acharjee *et al.* [44]
study design		
Repeated study	Study is repeated in another time or place to generate a second type of data	Cavill *et al.* [45]
Replicate matched study	Biological replicates are used to generate additional data	Cavill *et al.* [45]
Source matched study	Samples from the same organism e.g. animal are extracted	Cavill *et al.* [45]
Split sample study	Samples e.g. tissues are split for profiling with different omics technologies	Cavill *et al.* [45]
study focus		
Sequential analysis	Does the additional data type enhance understanding of the first data type?	Yuan *et al.* [38], Le Van *et al.* [41], Shen *et al.* [46]
Biological analysis	What are the underlying processes leading to phenotypical changes?	Hirai *et al.* [47], Cavill *et al.* [48], Safo *et al.* [49], Hong *et al.* [50]
Model-based analysis	Which variables are phenotypically relevant? significantly associated? Can predictive ability be improved?	Smolinska *et al.* [35], Witten and Tibshirani [36], Daemen *et al.* [37]

Table 1: Classification of different data integration approaches. The examples list is by no means exhaustive.

interpretability in metabolomics data: metabolite identification. In an agnostic approach, where metabolites are putatively annotated, integrative analysis can be performed regardless of the metabolite identification step. However in a more specific approach, integrative analysis needs to be performed with regards to whether or not metabolite identification has been realized beforehand. For example, if metabolites have not been identified, data integration would be rather limited to almost purely statistical analysis i.e. classification purposes, prediction purposes or inference of significant variables whereas when metabolite identities are known enrichment analysis methods can be applied. Additional challenges arise since there is often not a one to one relationship between genes and metabolites.

On a third instance, the metabolome is intrinsically different between individual samples due to its rich chemical diversity and hence some observed patterns in metabolic profiles might not be the result of perturbations in the biosystem or phenotype of interest. Therefore, integrative analysis ideally needs robust models to account for individual variations.

On a fourth instance, metabolomic datasets are characterized by high correlation structures in that many peaks can arise from the same metabolites and metabolites operate within networks of chemical reactions. Hence, two highly correlated metabolites might not be directly related but high correlation can be the result of complex interactions with other metabolites in common. Finally, additional fundamental challenges are similar to typical challenges that usually arise in data integration frameworks, for example, incompleteness of each data type (i.e. missing values), high dimensionality and heterogeneity of data obtained from multiple sources. It is noteworthy to acknowledge that all data analysis steps need to be performed in account of the technical and experimental limitations of each omics platform including metabolomics. Metabolomics requires relatively high-cost instrumentation, complex data analysis and still suffers from issues of variable sensitivity, high volatility and sample-to-sample variability [55]. As a result, reproducibility is one of the significant hurdles in metabolomics [56] . In fact, the delicate stability of metabolites implies that biological samples need to be handled quickly and uniformly. Minor changes in the experimental conditions or procedure (e.g. different laboratories, external conditions ...) can cause critical changes in the observed metabolome which might impact results. Hence, statistical data integration approaches should be appropriately selected based on study design amongst other criteria we will discuss in the following sections [1].

3. Study design

Study design including sample and data collection needs to be selected in conjunction with the study research questions and hypotheses (see section 5). Several scenarios need to be considered with respect to the experimental design such as: Is there an intervention effect and is the effect evaluated in different groups e.g. cases vs controls, or there is no intervention, treatment, or exposure administered to participants e.g. cohort study? Are measurements recorded on intervention only, or before and after intervention or at multiple time points? Which types of omic platforms potentially provide more insight? How samples will be collected? Ideally, samples would be collected from the same biological source for omic platforms.

Cavill *et al.* [45] have identified four types of sample collection that should be considered when performing data integration: *repeated study*, *replicate matched study*, *split sample study* and *source matched study*. Briefly speaking, in a repeated study, one omic dataset is first generated following a specific experimental protocol. A second dataset is then obtained by repeating the same experimental protocol usually on a different time or lab. This study design is likely to introduce batch effects into the integrative analysis and is usually complicated to correct for. In certain studies, separate samples from different

[1] Integrative models based on metabolic fluxes and kinetic models are not in the scope of this article.

biological replicates (in the same experiment) are needed to generate metabolomics and transcriptomics data depending on the experimental extraction protocol. This is referred to as replicate matched study. The split sample study is subject to less variation than the repeated or replicate matched study. As its name suggests, it consists in splitting the same biological sample e.g. tissue or biofluid into two samples profiled with different omics technologies. The last case is the source matched study where different fractions of the biological system are used for different analyses for example urine, plasma or faeces. For example Yusufi *et al.* [57], Gulston *et al.* [58], Kaluarachchi *et al.* [59] advantageously used either the source matched study design or the split sample study design to benefit from reduced batch effects.

Ultimately, the experimental design should be informed by the data analysis to limit confounding and batch effects which could be introduced during preparation and storing. Although technical artifacts cannot be completely removed, they can be significantly reduced via a randomized study. It is still important however to recognize these limitations as early as possible in the sample and data collection process and acknowledge their aftermaths.

4. Data types

Generally speaking, the research community is following two kinds of integrative analysis : *horizontal* or *homogeneous meta-analysis* versus *vertical* or *heterogeneous multi-omics analysis* [30,31,39]. Horizontal analysis concerns integration of data where similar entities are measured across different studies, cohorts or labs. On the contrary, vertical analysis deals specifically with different entities such as metabolites and proteins measured on the same set of samples. In the following, we will mainly discuss aspects relevant to vertical data integration.

5. Hypothesis

Ritchie *et al.* [60] defines multi-omic data integration as "the incorporation of multi-omic information in a *meaningful* way to provide a more comprehensive analysis of a biological point of interest". Hence, data integration in omics does not only concern data concatenation, linking, coupling or correlation but most importantly the biological consistency of the combined information. Biological consistency is hence a major driver in integrative analysis. For instance, biological consistency is of crucial importance if the integrative analysis method adopted is conceptual, that is based on conclusions mostly synthesized by the researcher or the method is model-based, that is the biological system can be justly mathematically described, to ensure biological model assumptions are valid [61].

To ensure biological consistency, researchers should question their hypothesis at early stages prior to integrative analysis. In the context of multi-omics integration, one of the main biological hypotheses to think about is whether variation between omics is unidirectional or multi-directional. For instance, if variation is assumed to be unidirectional that is hierarchical from the genome to the metabolome, a multi-staged integrative analysis should be privileged [60]. Multi-staged analysis stands for the process of combining data in consecutive steps where, for example, genomic variables are first associated with transcriptomic variables. Significant transcriptomic variables are then associated with metabotypes. An additional example is where genomic and metabolomic data are separately filtered and associated with a specific phenotype e.g. via GWAS (Genome-Wide Association Studies) and MWAS (Metabolome-Wide Association Studies). The resulting datasets are then tested for mutual association e.g. via metabolome GWAS [7,15,22,34]. This approach is generally carried out to identify changes in phenotypic traits that are induced by changes in the metabolome which in turn are caused by variation in the genome (Figure 1). On the contrary, meta-dimensional analysis supports the hypothesis of simultaneous variation in the genome, transcriptome, proteome and metabolome leading to the phenotype. In other words, the meta-dimensional approach assumes that it is the combination of multiple variables from various data types that results in the phenotype [35,36,48]. In this case, concatenation-based or transformation-based statistical methods can be used to analyze the data simultaneously (See Section 6).

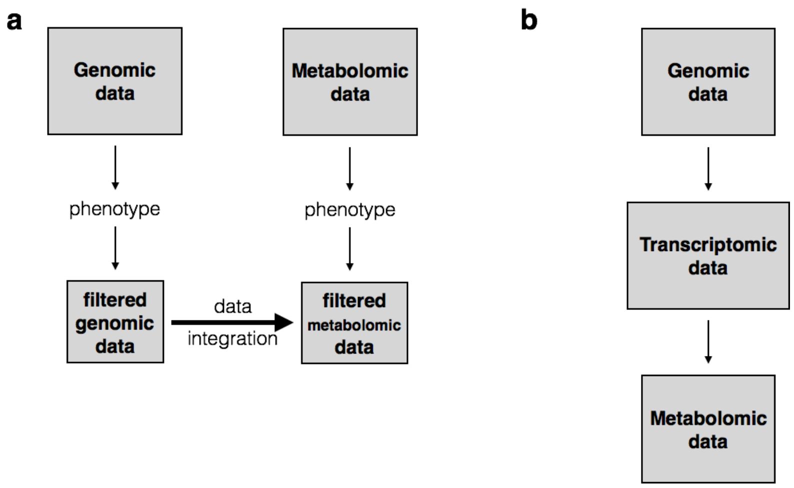

Figure 1. Examples of a multi-stage integrative analysis approach. Example (a) illustrates a three-step framework where genomic and metabolomic datasets are concurrently tested for association with the phenotype resulting in smaller datasets. These datasets are then investigated to infer linked variables. Example (b) illustrates a typical scenario where genomic variables are tested for association with transcripts which are in turn associated to metabotypes. These metabotypes might for instance explain the expression of phenotype.

6. Data integration strategies

Integrative analysis can also be broadly categorized with respect to different strategies : *low level* or *early integration*, *intermediate integration* and *high level* or *late integration* (Figure 2). In early integration, all inputs are concatenated at raw or processed level to form a single dataset, gathering all the provided information with minimal loss. Hence, one major challenge in early integration is to use an appropriate common representation for datasets from different scales [40,42]. Nonetheless, early integration benefits from two main advantages. First, statistical methods as used for a single data can be applied with slight modifications to the obtained data matrix. Second, it usually preserves information of interaction between omic variables of the input datasets. However, this strategy is subject to increasing the high dimensionality of the data by concatenating the total number of variables from all input data into a single one. Hence, dimensionality reduction techniques might be required before performing early integration.

Whereas in early integration transformation shouldn't change the nature of the data, intermediate integration deals with finding a suitable mapping into another format prior to data combination. This approach covers models that make use of kernel functions or network representation of the data. Kernels have been widely used to capture and transform implicit patterns into explicit schemes by embedding data items into feature space [35,42,62]. By contrast to their superior predictive accuracy, a major disadvantage of kernel-based methods is that they are often difficult to interpret. On another hand, network based methods are popular in omics data integration as they offer easy integration (by merging edges for example) and enhanced interpretability [43,63]. In metabolomics, ease of interpretability is a major concern and ensures high functionality of the model. Ease of interpretability is, thus, one of the key aspects to consider when developing integrative models.

In late integration, each data type is modeled independently then, the resulting models are used to build an integrative or high level model. In the metabolomics literature, latent variable models, namely PCA and PLS variants, are very popular. They can be used for integrative analysis according to the following procedure [64] : Separate models are fitted for each dataset and score matrices are extracted. These score matrices are then concatenated and used as input for an additional model. The latter is termed as high level model. The main limitation of late integration lies in the fact that information about mutual interactions between different data might be lost as the models are first fitted separately [44,65].

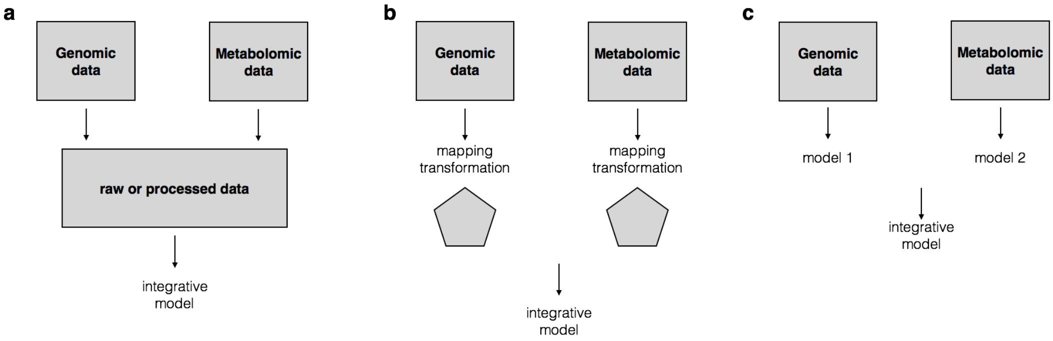

Figure 2. Different data integration strategies. Figure (a) illustrates early integration where data is combined into a single data matrix before modeling. Figure (b) depicts the intermediate data integration level where data matrices are transformed or mapped into a common meaningful representation before modeling. In figure (c), each data model is generated separately and is then combined with models based on other data types to generate the integrative or high-level model.

7. Study focus

It goes without saying that the integrative analysis process is greatly influenced by the primary statistical or biological focus of the study. Three non exhaustive categories of integrative analysis types according to study objectives have been identified by Daemen *et al.* [37], Wang *et al.* [66]. The first category, called *sequential integration*, attempts to uncover the development of a phenotype e.g disease as opposed to its prediction. This category mainly answers questions on how does data fusion deepen our understanding of the disease? Does the additional data type confirm the findings of the first data type? Does the additional data type enhance our understanding of the first data type? Such analyses were conducted by e.g. Kleemann *et al.* [67], Santos *et al.* [68], Verhoeckx *et al.* [69]. In general, the authors firstly identify genes that are associated to external perturbations or disease. Secondly, genes are linked to metabolites and corresponding enriched pathways. Metabolomics is a highly suitable source for deriving phenotype biomarkers as well as cross-omics biomarkers since it integrates genetic as well as non-genetic factors. Regression is one of the elementary methods used for cross-omics biomarker discovery in sequential analysis. In a similar fashion to GWAS and MWAS, metabolome GWAS was widely applied to integrate of metabolomic and genomic data but is criticized as analysis is performed in a univariate way. Due the correlation structures inherent to omics data, multivariate regression can be achieved by introducing penalty terms in the frequentist setting or shrinkage priors in the Bayesian setting. Yet, these methods ignore dependence between metabolites in favor of genotype-metabotype dependence. Hence, one of the fundamental challenges that arise in this class of models it to simultaneously model metabolite-metabolite associations and metabolite associations with other omics entities. Biological entities are complex by nature and are arguably regulated by sequences of actions and complex interactions. In this sense, modeling a sequence of observations naturally regulated by chemical processes has proven successful in computational biology [12].

Sequential analysis also comprises disease subtype discovery, commonly achieved via clustering approaches [38,41,46,70,71]. For instance, Le Van *et al.* [41] propose a clustering model that simultaneously identifies features related to each subtype. In this approach data is integrated via ranked transformation. Clustering for functional analysis was explored by Manikandan *et al.* [72], Becker *et al.* [73], Yi *et al.* [74] in the Bayesian parametric setting and proven to provide more understanding of the forces underlying cellular processes and an unbiased method for researchers to identify related functional clusters. In the Bayesian nonparametric setting more flexible models were implemented by Kirk *et al.* [28], Yuan *et al.* [38], Savage *et al.* [75] via hierarchical models where the notion of "fusion" state was introduced. Hierarchical models offer more degrees of freedom than one-level models and thus allows to define for each data its own parameters that might (or not) be shared. Nonetheless, these models are only applicable for homogeneous integrative analysis, that is features that represent the same omic entity (e.g. copy number and expression data). One of the fundamental challenges in this context, hence, lies in the heterogeneity of multiple data types. In [38], the designed model, termed as PSDF, allows clustering of different types of discrete genomic data to identify cancer subtypes, feature selection and infer whether patients exhibit similar profiles across data types. The PSDF model makes use of the Dirichlet process to infer probabilistic cluster assignments and Bayesian hierarchical modeling to integrate genomic data. As it uses discrete data with similar scales, initial data transformation is not required for the PSDF model, however, typically data transformation needs to be realized before applying integrative clustering models.

An important literature body in multi-omics analysis involves two additional types of integrative analysis: *biological integration* and *model-based integration*. The aim of biological integration is to uncover the biological mechanisms of interaction between heterogeneous variables including metabolic pathways, regulatory mechanisms and signaling mechanisms [47–50] [2]. Metabolomics data are characterized by a high number of metabolic profiles compared to the number of biological samples. Moreover, metabolomic variables are also regulated by complex and strong correlation patterns. Henceforth, dimensionality reduction techniques are of fundamental importance in chemometrics for ease of visualization and interpretation. Dimensionality reduction techniques such as CCA, PCA and PLS derived techniques usually involve maximizing a covariance function under orthogonality constraints. Although these methods offer rich interpretation in terms of shared and orthogonal components, it is not straight-forward to quantify associations between the different variables and thus limits interpretation with regards to variables' mutual interrelationships. In biological integration one might use available prior knowledge such as metabolic pathways to reinforce interpretability in dimensionality reduction techniques. In particular, Safo *et al.* [49] developed a sparse canonical correlation analysis (CCA) model to uncover hidden association patterns between heterogenous data where sparsity is adjusted based on structural information of biological networks.

In model-based integration, researchers are faced with a range of statistical questions such as which omics variables are associated with the phenotypic changes? Which groups of variables from the different datasets are interacting? Does data fusion improve predictive accuracy of phenotype, disease, temporal behaviour? Is information expressed by the different data types redundant? In this context, an important range of statistical and machine learning methods have been developed in the literature. By way of illustration, kernel-based approaches where proposed to integrate multi-platform metabolomics data such as NMR and GC-MS [35] and multi-platform genomics data [37]. Both authors show that predictive ability of the integrative model significantly outperforms predictive ability of models based on a single type of data. Žitnik and Zupan [65] used matrix factorization to integrate 11 data types to predict gene function in D.discoideum and similarly shows that the integrative model significantly improves prediction compared to single models and is more robust to technical and methodological biases. In order to

[2] By contrast to sequential analysis where one data is analyzed then a second one is used to confirm or deepen results from the first analysis (the emphasis is not specifically about cellular mechanisms), the focus in biological integration is directly related to underlying cellular mechanisms.

find significant linear combinations between copy number and gene expression data, Witten and Tibshirani [36] developed a supervised sparse canonical correlation model where datasets were linked to a clinical outcome.

8. Discussion

To conclude, different multi-omics integration approaches can be further classified according to multiple dimensions. Broadly speaking, *data types* and *study design* are parts of the experimental dimension whereas the *strategy* types are parts of the methodological dimension. Finally, *study focus* and underlying variation *hypothesis* reflect the biological dimension regardless of the adopted statistical method. Table 1 summarizes different data integration classes depending on *hypothesis*, *data types*, *strategy*, *study design* and most importantly *study focus*. We also acknowledge that there three types of multi-omics data integration as identified by Ebbels and Cavill [61]: conceptual, statistical and model-based where the involvement of mathematical procedures in integrative analysis is different.

Study focus is of crucial importance to performing *meaningful* and efficient integrative analysis. In this thesis we mainly focus on *biological analysis*, *model-based analysis* and *sequential analysis* frameworks. Metabolomics is a highly suitable source for deriving biomarkers under these frameworks as it is the closest layer of the omics cascade that is related to the phenotype. In fact, metabolic profiling is widely used to study genotype-metabotype interactions or metabotype-phenotype interactions such as disease-relevant phenotypes or external stimuli. Identifying interactions between omic variables either in terms of significant statistical associations, biomarker discovery or biological networks enhances data interpretability and represents the end goal of many studies. For the sake of interpretability, an arsenal of mathematical and computational techniques has been developed to achieve such analyses. These techniques include, amongst others, correlation analysis [36,49], integrative regression models [15,22,34] and Bayesian integrative clustering of gene profiles [28,38,75].

It is worthy to note that a preliminary examination of the literature at the time of research shed light on two substantial shortcomings. Most of the current integrative analysis approaches are conducted separately from the main stream of the analysis, i.e. as a supplementary step. These two-step integrative approaches are very informative to prioritize data signals, nevertheless, they are not optimal. The heterogeneity of biosystems suggests that interrelationships between the various omics entities is key to exhibiting specific phenotype implying that data integration plays an important role into deciphering mechanisms of biological functions in living organisms [76]. As a consequence, integrative analysis should be part of the main analysis pipeline. On top of that, a close survey of the literature reveals that applications of probabilistic models for integrative analysis in metabolomics are very scarce. This is mostly ascribable to the limited number of available software on probabilistic models in the field which restricted their popularity.

Funding: “This research received no external funding”

Informed Consent Statement: “Not applicable”

Data Availability Statement: “Not applicable”

Acknowledgments: Thank you to Prof K.Strimmer, Dr T. Ebbels and Dr M. Chadeau to useful discussions.

Conflicts of Interest: “The authors declare no conflict of interest.”

References

1. Bylesjö, M.; Eriksson, D.; Kusano, M.; Moritz, T.; Trygg, J. Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *The Plant Journal* **2007**, *52*, 1181–1191.

2. Griffin, J.L.; Blenkiron, C.; Valonen, P.K.; Caldas, C.; Kauppinen, R.A. High-resolution magic angle spinning 1H NMR spectroscopy and reverse transcription-PCR analysis of apoptosis in a rat glioma. *Analytical chemistry* **2006**, *78*, 1546–1552.
3. Lindon, J.C.; Nicholson, J.K.; Holmes, E.; Antti, H.; Bollard, M.E.; Keun, H.; Beckonert, O.; Ebbels, T.M.; Reily, M.D.; Robertson, D.; others. Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicology and applied pharmacology* **2003**, *187*, 137–146.
4. Li, S.; Todor, A.; Luo, R. Blood transcriptomics and metabolomics for personalized medicine. *Computational and structural biotechnology journal* **2016**, *14*, 1–7.
5. Robinson, O.; Chadeau, H.M.; Karaman, I.; Climaco, P.R.; Ala-Korpela, M.; Handakas, E.; Fiorito, G.; Gao, H.; Heard, A.; Jarvelin, M.R.; Lewis, M.; Pazoki, R.; Polidoro, S.; Tzoulaki, I.; Wielscher, M.; Elliott, P.; Vineis, P. Determinants of accelerated metabolomic and epigenetic ageing in a UK cohort. *Aging Cell* **2020**, *19*, 1–13. doi:10.1111/acel.13149.
6. Karaman, I.; Ferreira, D.; Boulange, C.; Kaluarachchi, M.; Herrington, D.; Dona, A.; Castagné, R.; Moayyeri, A.; Lehne, B.; Loh, M.; de, V.P.; Dehghan, A.; Franco, O.; Hofman, A.; Evangelou, E.; Tzoulaki, I.; Elliott, P.; Lindon, J.; Ebbels, T. A workflow for integrated processing of multi-cohort untargeted 1H NMR metabolomics data in large scale metabolic epidemiology. *Journal of Proteome Research* **2016**, *15*, 4188–4194. doi:10.1021/acs.jproteome.6b00125.
7. Valcárcel, B.; Ebbels, T.M.; Kangas, A.J.; Soininen, P.; Elliot, P.; Ala-Korpela, M.; Järvelin, M.R.; de Iorio, M. Genome metabolome integrated network analysis to uncover connections between genetic variants and complex traits: an application to obesity. *Journal of The Royal Society Interface* **2014**, *11*, 20130908.
8. Nicholson, J.K.; Wilson, I.D. Understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nature Reviews Drug Discovery* **2003**, *2*, 668.
9. Haukaas, T.H.; Euceda, L.R.; Giskeødegård, G.F.; Bathen, T.F. Metabolic portraits of breast cancer by HR MAS MR spectroscopy of intact tissue samples. *Metabolites* **2017**, *7*, 18.
10. Pazoki, R.; Evangelou, E.; Mosen-Ansorena, D.; Pinto, R.; Karaman, I.; Blakeley, P.; Gill, D.; Zuber, V.; Elliott, P.; Tzoulaki, I.; Dehghan, A. PATHWAYS UNDERLYING URINARY SODIUM AND POTASSIUM EXCRETION AND THE LINK TO BLOOD PRESSURE AND CARDIOVASCULAR DISEASE. *Journal of Hypertension* **2019**, *37*, e74–e74. doi:10.1097/01.hjh.0000571108.82708.c0.
11. Rantalainen, M.; Cloarec, O.; Beckonert, O.; Wilson, I.; Jackson, D.; Tonge, R.; Rowlinson, R.; Rayner, S.; Nickson, J.; Wilkinson, R.W.; others. Statistically integrated metabonomic- proteomic studies on a human prostate cancer xenograft model in mice. *Journal of proteome research* **2006**, *5*, 2642–2655.
12. Jendoubi, T.; Ebbels, T.M. Integrative analysis of time course metabolic data and biomarker discovery. *BMC bioinformatics* **2020**, *21*, 1–16.
13. Ruepp, S.U.; Tonge, R.P.; Shaw, J.; Wallis, N.; Pognan, F. Genomics and proteomics analysis of acetaminophen toxicity in mouse liver. *Toxicological sciences* **2002**, *65*, 135–150.
14. Dumas, M.E.; Wilder, S.P.; Bihoreau, M.T.; Barton, R.H.; Fearnside, J.F.; Argoud, K.; D'Amato, L.; Wallis, R.H.; Blancher, C.; Keun, H.C.; others. Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models. *Nature genetics* **2007**, *39*, 666–672.
15. Nicholson, G.; Rantalainen, M.; Li, J.V.; Maher, A.D.; Malmodin, D.; Ahmadi, K.R.; Faber, J.H.; Barrett, A.; Min, J.L.; Rayner, N.W.; others. A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS genetics* **2011**, *7*, e1002270.
16. Clayton, T.A.; Lindon, J.C.; Cloarec, O.; Antti, H.; Charuel, C.; Hanton, G.; Provost, J.P.; Le Net, J.L.; Baker, D.; Walley, R.J.; others. Pharmaco-metabonomic phenotyping and personalized drug treatment. *Nature* **2006**, *440*, 1073–1077.
17. Teitsma, X.M.; Yang, W.; Jacobs, J.W.; Pethö-Schramm, A.; Borm, M.E.; Harms, A.C.; Hankemeier, T.; van Laar, J.M.; Bijlsma, J.W.; Lafeber, F.P. Baseline metabolic profiles of early rheumatoid arthritis patients achieving sustained drug-free remission after initiating treat-to-target tocilizumab, methotrexate, or the combination: insights from systems biology. *Arthritis research & therapy* **2018**, *20*, 230.
18. Griffin, J.L.; Bonney, S.A.; Mann, C.; Hebbachi, A.M.; Gibbons, G.F.; Nicholson, J.K.; Shoulders, C.C.; Scott, J. An integrated reverse functional genomic and metabolic approach to understanding orotic acid-induced fatty liver. *Physiological Genomics* **2004**, *17*, 140–149.
19. Raamsdonk, L.M.; Teusink, B.; Broadhurst, D.; Zhang, N.; Hayes, A.; Walsh, M.C.; Berden, J.A.; Brindle, K.M.; Kell, D.B.; Rowland, J.J.; others. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature biotechnology* **2001**, *19*, 45–50.
20. Lindon, J.C.; Nicholson, J.K. Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics. *Annu. Rev. Anal. Chem.* **2008**, *1*, 45–69.
21. Nicholson, J.K.; Holmes, E.; Lindon, J.C.; Wilson, I.D. The challenges of modeling mammalian biocomplexity. *Nature biotechnology* **2004**, *22*, 1268–1274.
22. Gieger, C.; Geistlinger, L.; Altmaier, E.; De Angelis, M.H.; Kronenberg, F.; Meitinger, T.; Mewes, H.W.; Wichmann, H.E.; Weinberger, K.M.; Adamski, J.; others. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS genetics* **2008**, *4*, e1000282.
23. Kathiresan, S.; Manning, A.K.; Demissie, S.; D'agostino, R.B.; Surti, A.; Guiducci, C.; Gianniny, L.; Burtt, N.P.; Melander, O.; Orho-Melander, M.; others. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC medical genetics* **2007**, *8*, 1–10.

24. Teslovich, T.M.; Musunuru, K.; Smith, A.V.; Edmondson, A.C.; Stylianou, I.M.; Koseki, M.; Pirruccello, J.P.; Ripatti, S.; Chasman, D.I.; Willer, C.J.; others. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **2010**, *466*, 707–713.
25. Vineis, P.; van Veldhoven, K.; Chadeau-Hyam, M.; Athersuch, T.J. Advancing the application of omics-based biomarkers in environmental epidemiology. *Environmental and molecular mutagenesis* **2013**, *54*, 461–467.
26. Suhre, K.; Wallaschofski, H.; Raffler, J.; Friedrich, N.; Haring, R.; Michael, K.; Wasner, C.; Krebs, A.; Kronenberg, F.; Chang, D.; others. A genome-wide association study of metabolic traits in human urine. *Nature genetics* **2011**, *43*, 565.
27. Rattray, N.J.; Deziel, N.C.; Wallach, J.D.; Khan, S.A.; Vasiliou, V.; Ioannidis, J.P.; Johnson, C.H. Beyond genomics: understanding exposotypes through metabolomics. *Human genomics* **2018**, *12*, 4.
28. Kirk, P.; Griffin, J.E.; Savage, R.S.; Ghahramani, Z.; Wild, D.L. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **2012**, *28*, 3290–3297.
29. Liang, M.; Li, Z.; Chen, T.; Zeng, J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics* **2014**, *12*, 928–937.
30. Richardson, S.; Tseng, G.C.; Sun, W. Statistical methods in integrative genomics. *Annual review of statistics and its application* **2016**, *3*, 181–209.
31. Hamid, J.S.; Hu, P.; Roslin, N.M.; Ling, V.; Greenwood, C.M.; Beyene, J. Data integration in genetics and genomics: methods and challenges. *Human genomics and proteomics: HGP* **2009**, *2009*.
32. Tseng, G.; Ghosh, D.; Zhou, X.J. *Integrating Omics Data*; Cambridge University Press, 2015.
33. O'Shea, K.; Misra, B.B. Software tools, databases and resources in metabolomics: Updates from 2018 to 2019. *Metabolomics* **2020**, *16*, 1–23.
34. Krumsiek, J.; Suhre, K.; Illig, T.; Adamski, J.; Theis, F.J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology* **2011**, *5*, 21.
35. Smolinska, A.; Blanchet, L.; Coulier, L.; Ampt, K.A.M.; Luider, T.; Hintzen, R.Q.; Wijmenga, S.S.; Buydens, L.M.C. Interpretation and visualization of non-linear data fusion in kernel space: Study on metabolomic characterization of progression of multiple sclerosis. *PLOS One* **2012**, *7*.
36. Witten, D.M.; Tibshirani, R.J. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology* **2009**, *8*, 1–27.
37. Daemen, A.; Gevaert, O.; Ojeda, F.; Debucquoy, A.; Suykens, J.A.; Sempoux, C.; Machiels, J.P.; Haustermans, K.; De Moor, B. A kernel-based integration of genome-wide data for clinical decision support. *Genome medicine* **2009**, *1*, 39.
38. Yuan, Y.; Savage, R.S.; Markowetz, F. Patient-specific data fusion defines prognostic cancer subtypes. *PLoS computational biology* **2011**, *7*, e1002227.
39. Evangelou, E.; Ioannidis, J.P. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics* **2013**, *14*, 379.
40. Fridley, B.L.; Lund, S.; Jenkins, G.D.; Wang, L. A Bayesian Integrative Genomic Model for Pathway Analysis of Complex Traits. *Genetic epidemiology* **2012**, *36*, 352–359.
41. Le Van, T.; van Leeuwen, M.; Carolina Fierro, A.; De Maeyer, D.; Van den Eynden, J.; Verbeke, L.; De Raedt, L.; Marchal, K.; Nijssen, S. Simultaneous discovery of cancer subtypes and subtype features by molecular data integration. *Bioinformatics* **2016**, *32*, i445–i454.
42. Lanckriet, G.R.; De Bie, T.; Cristianini, N.; Jordan, M.I.; Noble, W.S. A statistical framework for genomic data fusion. *Bioinformatics* **2004**, *20*, 2626–2635.
43. Guo, X.; Gao, L.; Wei, C.; Yang, X.; Zhao, Y.; Dong, A. A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. *PloS one* **2011**, *6*, e24171.
44. Acharjee, A.; Ament, Z.; West, J.A.; Stanley, E.; Griffin, J.L. Integration of metabolomics, lipidomics and clinical data using a machine learning method. *BMC bioinformatics* **2016**, *17*, 440.
45. Cavill, R.; Jennen, D.; Kleinjans, J.; Briedé, J.J. Transcriptomic and metabolomic data integration. *Briefings in bioinformatics* **2015**, *17*, 891–901.
46. Shen, R.; Mo, Q.; Schultz, N.; Seshan, V.E.; Olshen, A.B.; Huse, J.; Ladanyi, M.; Sander, C. Integrative subtype discovery in glioblastoma using iCluster. *PloS one* **2012**, *7*, e35236.
47. Hirai, M.Y.; Klein, M.; Fujikawa, Y.; Yano, M.; Goodenowe, D.B.; Yamazaki, Y.; Kanaya, S.; Nakamura, Y.; Kitayama, M.; Suzuki, H.; others. Elucidation of gene-to-gene and metabolite-to-gene networks in Arabidopsis by integration of metabolomics and transcriptomics. *Journal of Biological Chemistry* **2005**.
48. Cavill, R.; Kamburov, A.; Ellis, J.K.; Athersuch, T.J.; Blagrove, M.S.; Herwig, R.; Ebbels, T.M.; Keun, H.C. Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS computational biology* **2011**, *7*, e1001113.
49. Safo, S.E.; Li, S.; Long, Q. Integrative analysis of transcriptomic and metabolomic data via sparse canonical correlation analysis with incorporation of biological information. *Biometrics* **2018**, *74*, 300–312.
50. Hong, S.; Chen, X.; Jin, L.; Xiong, M. Canonical correlation analysis for RNA-seq co-expression networks. *Nucleic acids research* **2013**, *41*, e95–e95.
51. Devlin, T.M. *Textbook of biochemistry*; John Wiley & Sons, 2011.

52. Sumner, L.W.; Amberg, A.; Barrett, D.; Beale, M.H.; Beger, R.; Daykin, C.A.; Fan, T.W.M.; Fiehn, O.; Goodacre, R.; Griffin, J.L.; others. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **2007**, *3*, 211–221.
53. Holliday, R. DNA methylation and epigenetic inheritance. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **1990**, *326*, 329–338.
54. Richelle, A.; Joshi, C.; Lewis, N.E. Assessing key decisions for transcriptomic data integration in biochemical networks. *PLoS computational biology* **2019**, *15*, e1007185.
55. Riekeberg, E.; Powers, R. New frontiers in metabolomics: from measurement to insight. *F1000Research* **2017**, *6*.
56. Keun, H.C.; Ebbels, T.M.; Antti, H.; Bollard, M.E.; Beckonert, O.; Schlotterbeck, G.; Senn, H.; Niederhauser, U.; Holmes, E.; Lindon, J.C.; others. Analytical reproducibility in 1H NMR-based metabonomic urinalysis. *Chemical research in toxicology* **2002**, *15*, 1380–1386.
57. Yusufi, F.N.K.; Lakshmanan, M.; Ho, Y.S.; Loo, B.L.W.; Ariyaratne, P.; Yang, Y.; Ng, S.K.; Tan, T.R.M.; Yeo, H.C.; Lim, H.L.; others. Mammalian systems biotechnology reveals global cellular adaptations in a recombinant CHO cell line. *Cell systems* **2017**, *4*, 530–542.
58. Gulston, M.K.; Rubtsov, D.V.; Atherton, H.J.; Clarke, K.; Davies, K.E.; Lilley, K.S.; Griffin, J.L. A combined metabolomic and proteomic investigation of the effects of a failure to express dystrophin in the mouse heart. *Journal of proteome research* **2008**, *7*, 2069–2077.
59. Kaluarachchi, M.R.; Boulangé, C.L.; Garcia-Perez, I.; Lindon, J.C.; Minet, E.F. Multiplatform serum metabolic phenotyping combined with pathway mapping to identify biochemical differences in smokers. *Bioanalysis* **2016**, *8*, 2023–2043.
60. Ritchie, M.D.; Holzinger, E.R.; Li, R.; Pendergrass, S.A.; Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* **2015**, *16*, 85.
61. Ebbels, T.M.; Cavill, R. Bioinformatic methods in NMR-based metabolic profiling. *Progress in nuclear magnetic resonance spectroscopy* **2009**, *4*, 361–374.
62. Lanckriet, G.; Deng, M.; Cristianini, N.; Jordan, M.; Noble, W. Kernel-based data fusion and its application to protein function prediction in yeast. *Biocomputing 2004, Proceedings of the Pacific Symposium, Hawaii, USA* **2004**, pp. 300–311.
63. Davis, D.A.; Chawla, N.V. Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PloS one* **2011**, *6*, e22670.
64. Forshed, J.; Idborg, H.; Jacobsson, S.P. Evaluation of different techniques for data fusion of LC/MS and 1H-NMR. *Chemometrics and intelligent laboratory systems* **2007**, *85*, 102–109.
65. Žitnik, M.; Zupan, B. Data fusion by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence* **2015**, *37*, 41–53.
66. Wang, W.; Baladandayuthapani, V.; Morris, J.S.; Broom, B.M.; Manyam, G.; Do, K.A. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **2012**, *29*, 149–159.
67. Kleemann, R.; Verschuren, L.; van Erk, M.J.; Nikolsky, Y.; Cnubben, N.H.; Verheij, E.R.; Smilde, A.K.; Hendriks, H.F.; Zadelaar, S.; Smith, G.J.; others. Atherosclerosis and liver inflammation induced by increased dietary cholesterol intake: a combined transcriptomics and metabolomics analysis. *Genome biology* **2007**, *8*, R200.
68. Santos, E.M.; Ball, J.S.; Williams, T.D.; Wu, H.; Ortega, F.; Van Aerle, R.; Katsiadaki, I.; Falciani, F.; Viant, M.R.; Chipman, J.K.; others. Identifying health impacts of exposure to copper using transcriptomics and metabolomics in a fish model. *Environmental science & technology* **2009**, *44*, 820–826.
69. Verhoeckx, K.C.; Bijlsma, S.; Jespersen, S.; Ramaker, R.; Verheij, E.R.; Witkamp, R.F.; van der Greef, J.; Rodenburg, R.J. Characterization of anti-inflammatory compounds using transcriptomics, proteomics, and metabolomics in combination with multivariate data analysis. *International Immunopharmacology* **2004**, *4*, 1499–1514.
70. Sun, X.; Stewart, D.A.; Sandhu, R.; Kirk, E.L.; Pathmasiri, W.W.; McRitchie, S.L.; Clark, R.F.; Troester, M.A.; Sumner, S.J. Correlated metabolomic, genomic, and histologic phenotypes in histologically normal breast tissue. *PloS one* **2018**, *13*, e0193792.
71. Tang, X.; Lin, C.C.; Spasojevic, I.; Iversen, E.S.; Chi, J.T.; Marks, J.R. A joint analysis of metabolomics and genetics of breast cancer. *Breast cancer research* **2014**, *16*, 415.
72. Manikandan, P.; Ramyachitra, D.; Banupriya, D. Detection of overlapping protein complexes in gene expression, phenotype and pathways of Saccharomyces cerevisiae using Prorank based Fuzzy algorithm. *Gene* **2016**, *580*, 144–158.
73. Becker, E.; Robisson, B.; Chapple, C.E.; Guénoche, A.; Brun, C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics* **2012**, *28*, 84–90.
74. Yi, G.; Sze, S.H.; Thon, M.R. Identifying clusters of functionally related genes in genomes. *Bioinformatics* **2007**, *23*, 1053–1060.
75. Savage, R.S.; Ghahramani, Z.; Griffin, J.E.; De La Cruz, B.J.; Wild, D.L. Discovering transcriptional modules by Bayesian data integration. *Bioinformatics* **2010**, *26*, i158–i167.
76. Joyce, A.R.; Palsson, B.Ø. The model organism as a system: integrating'omics' data sets. *Nature reviews Molecular cell biology* **2006**, *7*, 198–210.