Briefings in Bioinformatics, 17(5), 2016, 891–901

doi: 10.1093/bib/bbv090
Advance Access Publication Date: 14 October 2015
Paper

Transcriptomic and metabolomic data integration

Rachel Cavill, Danyel Jennen, Jos Kleinjans and Jacob Jan Briedé

Corresponding author: Rachel Cavill, Department of Knowledge Engineering, Maastricht University, PO Box 616, 6200 MD, Maastricht, The Netherlands.
E-mail: rachel.cavill@maastrichtuniversity.nl

Abstract

Many studies now produce parallel data sets from different omics technologies; however, the task of interpreting the acquired data in an integrated fashion is not trivial. This review covers those methods that have been used over the past decade to statistically integrate and interpret metabolomics and transcriptomic data sets. It defines four categories of approaches, correlation-based integration, concatenation-based integration, multivariate-based integration and pathway-based integration, into which all existing statistical methods fit. It also explores the choices in study design for generating samples for analysis by these omics technologies and the impact that these technical decisions have on the subsequent data analysis options.

Key words: transcriptomics; metabolomics; data integration; study design

Introduction

In recent years, an explosion has occurred in the acquisition of biological data through the use of so-called 'omics' techniques. Whilst many different omics technologies are now featured in the literature, the most frequently used omics are genomics, transcriptomics, proteomics and metabolomics.

Early studies tended to use only a single omics technology for the analysis of a set of samples, and there have been many efforts to standardize the reporting of preparation, data and data analysis pipelines for each individual data type [1–4]. However, increasingly, studies now include measurements from multiple omics techniques. The main impetus for data integration is that through these integrated data sets, an improved understanding of the underlying biology can be obtained, to be better able to predict a response variable and gain further insight into mechanistic aspects of the system. This review will show that metabolomics and transcriptomics produce data sets with complementary information, and to fully use these data, it is necessary to perform integrative steps in the data analysis, as opposed to separate analyses of the data sets and presentation of these separate analyses as a single study in the same publication.

When transcriptomic and metabolomic data are integrated, there is no direct association between metabolite and transcript. It is not possible to link each transcript to a metabolite or vice versa. This is in contrast with transcriptomics and proteomics, where most transcripts can be mapped to a single protein and then the obtained profiles compared (although there can still be significant deviations in the observed effects [5, 6] for a variety of biological reasons). This complicates the data analysis and implies that more complex statistical techniques must be used to find the relationships presumably inherent in the data set. Despite all this, transcriptomic–metabolomic integration is clearly a powerful combination, the metabolome providing phenotypic measurements to which we can anchor the global measurements of the transcriptome.

Taking into account these difficulties, it is unsurprising that the complexity and effort needed for effective data integration is often underestimated. An interesting estimate of the amount of effort required to integrate data sets has been performed by

Rachel Cavill is an assistant professor in Knowledge Engineering in Maastricht. She previously spent 8 years a post-doctoral bioinformatician, working with both transcriptomics and metabolomics data. She is particularly interested in developing methods for building interpretable integrated models of these data types.
Danyel Jennen is an assistant professor in the Toxicogenomics department in Maastricht. His research focuses on the development of integrated –omics applications using a systems biology/toxicology approach.
Jos Kleinjans is the head of the Toxicogenomics department in Maastricht. He has headed up many national and European Union projects around the use of omics data in Toxicology.
Jacco Jan Briedé is an assistant professor in the Toxicogenomics department in Maastricht. His main research subject concerns the applicability of electron spin resonance (ESR or EPR) spectrometry in the study of free radicals formation and their involvement in environmental health risk assessment and molecular toxicology. In addition, he is interested in the increased application of "omics" technologies in the field of toxicogenomics.
Submitted: 7 July 2015; **Received (in revised form):** 24 August 2015

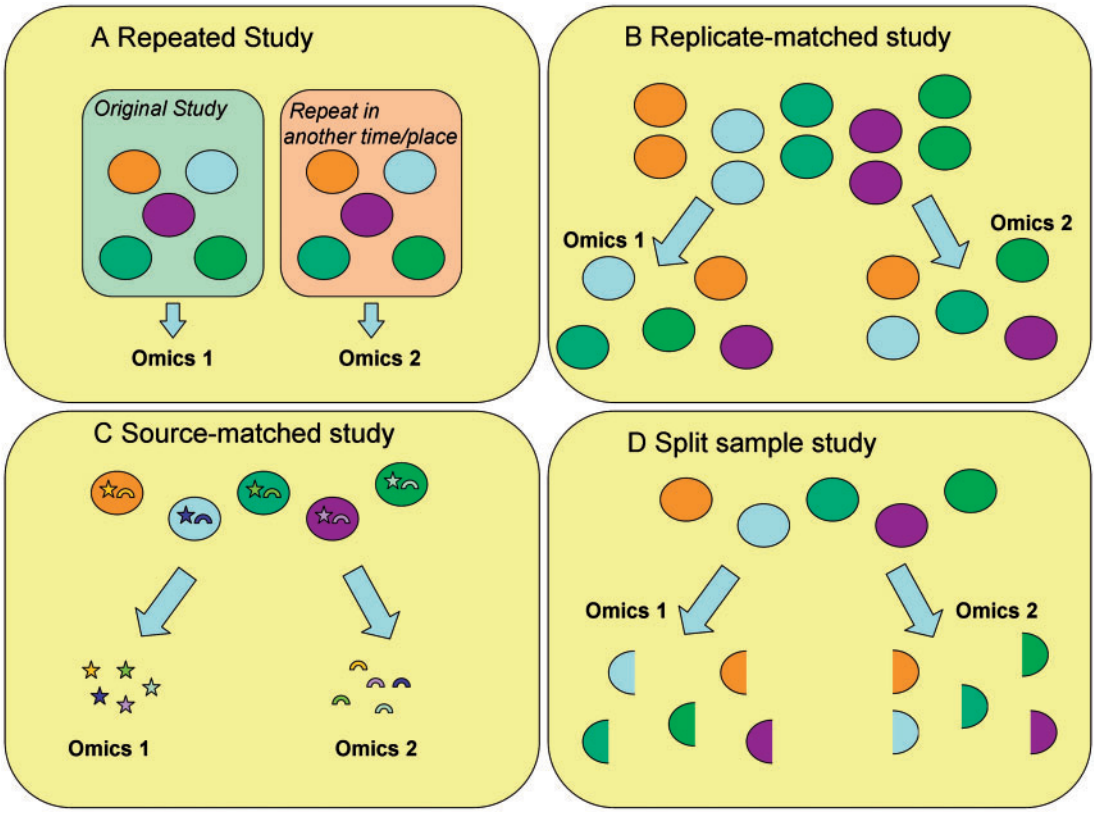

Figure 1. Four different study designs commonly used for multi-omics studies.

Palsson [7], the outcome of whi[illegible]ourages further consideration of this key stumbling block.

We here review the techniques that have been applied for statistical data integration of metabolomics and transcriptomics data.

Study design for integrated data analysis

When multiple omics technologies are applied within a single study, there are several possible study designs. The choice of study design is important, as it will have many implicati[illegible] throughout the study, in particular with regard to the data a[illegible]lysis. The most common study designs are shown in Figure 1.

The repeated study (Figure 1A) can often be the most convenient approach, producing a set of samples for the first omics and then repeating the experimental protocol to generate the samples for a second omics (or conducting the same protocol in multiple laboratories to generate different sample sets). However, this will introduce batch effects into the data analysis as noted by Tillinghast [8]. Batch effects, when occurring in large studies with a single omics technology, can be corrected for by using statistical methods, such as simply centring the data from each batch separately before combination or other more complex approaches as discussed by Leek *et al.* [9]. However, when multiple omics technologies are applied through separate experiments, there is no possibility to apply batch correction. The repeated study design does, however, have some[illegible] vantages, in that measurements could be considered statistic[illegible] independent, or independent tests of the same hypothesis [10].

The other designs shown in Figur[illegible] are more similar, in that they all are assumed to produce d[illegible]thout batch effects between the different omics data sets (within omics batch effects may still be present, especially if large numbers of samples are run over a period of time). In terms of data integration, the ideal situation is to have samples originating from the same biological source material. For instance, a piece of tissue may be cut into two sections and one used for metabolomics analysis, whilst the other is used to extrac[illegible]. We will term this a split sample study design (Figure 1B).

Usually, the deciding factor concerning the choice of[illegible]udy design, is feasibility, e.g. whether samples for multip[illegible]ics can be taken from a single biological replicate or source. For instance, in case of an *in vitro* study, if the intracellular extract is to be analysed by metabolomics and the isolated RNA by transcriptomics, a separate sample is needed for the metabolomics analysis, as the RNA extraction process and the sample preparation protocol of a cell pellet for metabolomics are mutually exclusive. In this case, a replicate-matched study design (Figure 1C) may be used to generate the matched samples from different biological replicates from within the same experiment. When performing a replicate-matched study, it is important to randomize every aspect of the study between the samples being prepared for each omics technology, so as not to introduce any bias between the profiles obtained by each technology. A replicate-matched study differs critically from a repeat stud[illegible] the samples for both omics are produced/obtained at the [illegible] time, and thus the introduction of batch effects is avoided.

Alternatively, one may choose to use different fractions of the biological system for different analyses. We term this a source-matched study (Figure 1D); the clearest example of this is seen in animal studies, where blood or urine may be taken for metabolomics to be matched with RNA profiles from the affected tissue. Source-matched studies can also be performed *in vitro*, for instance, if cell material is used fo[illegible]scriptomics analysis and the cell media for metabolomics.

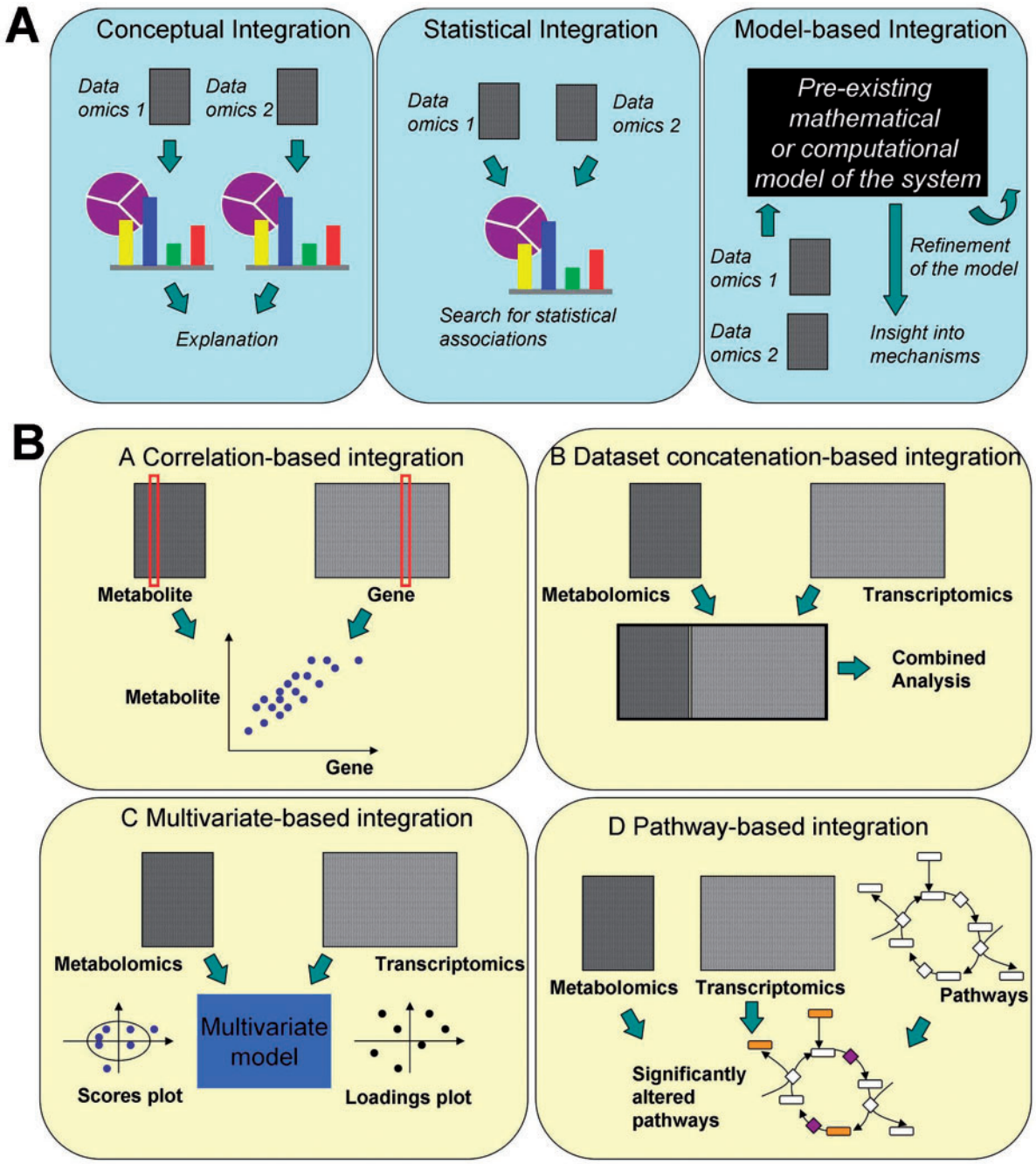

Figure 2. Methods for omics data integration: Panel A shows the three main types of integration as per [11]. Panel B elaborates on the different methods for statistical data integration, dividing them into four approaches, which are reviewed separately in this article.

Given a free choice of study design, the split-sample study would be the best to use, followed by the replicate-matched study. Source-matched studies can be useful as shown later, but it must be carefully considered how any links found between the two sets will be interpreted. Repeated study designs are best avoided if possible to circumvent the inherent batch effects.

Taxonomies of data integration

In [11], Ebbels suggests that there are three levels of data integration: conceptual integration, statistical integration and model-based integration. These are shown in Figure 2, panel A. Conceptual integration refers to the situation where multiple omics data sets are analysed separately, and then, the resulting conclusions are matched without any further analysis of the data set as a whole, such as in [12, 13]. This approach can produce valuable insights; however, it will also miss associations that could only be found when both data sets are analysed together. Statistical integration is the most common form of integration applied to transcriptomic and metabolomic data sets, where statistical associations are sought between the elements from the different data sets. Model-based integration is currently an unobtainable ideal in most situations in which global

omics data are acquired, as there are not yet enough data to determine all the model the system based on a priori biological knowledge would pre-exist before the experiment, and then the data obtained from the experiment could be compared parameters needed (concentrations, rate constants, interactions, etc). In this approach, a computational or mathematical model of with that predicted from the model. The model might, for instance, be a fully parameterized metabolic network for the biological system under study.

In this article, we will focus on examples of statistical data integration, and further classify them into four subgroups (see Figure 2, panel B). Firstly, correlation-based integration seeks to find correlative links between elements of one data set and elements from the other. Secondly, data set concatenation-based integration methods group together those approaches that concatenate the gene and metabolite measurements into a single table to perform an integrated analysis. However, both techniques are severely limited in their scopes. Correlation often fails because of different time scales of change, and concatenation-based methods have to deal with the different underlying distributions of the data from the contributing technologies. Thirdly, multivariate-based integration uses variations on standard multivariate techniques such as partial least square (PLS) and principal component analysis (PCA) to model the data and either to perform predictions or to find relationships between variables/samples and visualize the variation inherent in the data. The last group is a class of methods that attempt to integrate the data using existing biological knowledge through metabolic pathways as predefined in databases such as KEGG [14] and Wikipathways [15]. We call these techniques pathway-based integration. These pathway-based integration techniques are distinct from modelling efforts, as rather than studying reaction fluxes, they instead map, in an automated manner, the measured metabolites and transcripts to pathways and find those where there is statistical evidence of a significant change in their behaviour between two conditions, or a correlation between pathway behaviour and a phenotypic end point of interest.

These four categories can be related to the split between 'biology-driven strategies' and 'data-driven strategies' identified by Thomas and Ganji [16], in that their 'biology-driven strategies' cover studies using a purely conceptual analysis, which are not reviewed here, as well as those using pathway-based integration. Meanwhile, their 'data-driven strategies' are equivalent to the other three categories presented here.

Correlation-based integration

One of the simplest ways to explore multi-omics data in an integrative way is to explicitly set out to look for correlative links between the data sets. Correlations have frequently been used to examine the associations between transcriptomic data and metabolomics measurements. There are many ways to assess the correlation of two sets of measurements, the most common of these being Pearson's and Spearman's correlation for parametric and non-parametric data, respectively.

Naively, one expects metabolites to correlate with those genes with which they have associations; however, this is not always the case. While on the one hand, Urbanczyk-Wochniak found more than double the number of significantly correlated metabolite–transcript pairs than would be expected by chance in potato tubers [17], and Fendt *et al.* [18] quantified transcripts, proteins and metabolites in yeast and other species on perturbations (for instance single enzyme modulations) and showed an inverse relationship between the log fold change of a metabolite and the log fold change of the protein or transcript catalyzing the reaction. They also found a great deal of variation in the correlation's strengths, with a more significant trend in the correlations between substrates and enzymes than between reaction products and enzymes. On the other hand, ter Kuile and Westerhoff [19] found that fluxes through steps in the biochemical pathways did not correlate proportionally with the concentrations of the corresponding biochemical enzymes, and Moxley *et al.* [20] also report low correlation coefficients between transcripts of metabolic enzymes and related metabolite fluxes ($r = 0.07–0.8$) in yeast.

More concerningly for pure correlation-based approaches, Bradley [21] noted that both the direction and magnitude of correlation between metabolites and related genes could vary significantly between experimental conditions.

As well as using standard correlation coefficients, such as Pearson's or Spearman's, there are also other methods for measuring correlation; the Goodman and Kruskal gamma test, which only takes into account the up/down regulation of each metabolite/gene, e.g. [22]; robust linear models, which look at each transcripts' ability to predict each metabolite [23]; and partial correlations [24, 25], which evaluate those correlations that are independent of the other colinear measurements. For instance, what is the independent correlation of gene A and metabolite B, given that they are both correlated to gene C, can be calculated through a partial correlation computation.

In many experimental designs, we know that the changes in the metabolome and the transcriptome will not be simultaneous, and therefore it will be best to obtain a time course of samples for each omics. In these cases, it has been shown that by firstly aligning the data through techniques such as Dynamic Time Warping will help the detection of associated metabolites and transcripts [26].

In mammalian systems, data integration is often performed on source-matched data sets, where the RNA and metabolomics samples were acquired from different tissues. This complicates the expected correlation patterns. For instance, linking plasma metabolic changes to liver transcript changes in a source-matched study of fenofibrate and fish oil treatments in mice [27]. Lu *et al.* found that the expression of genes involved in fatty acid metabolism were associated with levels of plasma cholesterol and phosphatidycholine.

This section shows that a straightforward application of Pearson or Spearman correlation has many potential problems and is not overly suited to the task of metabolomic–transcriptomic data integration. Those elements that are known to be closely related in the pathways, often do not show a correlative link, while correlations can also occur at great distances across the network. More work needs to be undertaken to see if partial correlations can aid the identification of the most direct connections. There are also issues of time that obscure the correlative links between transcriptomic and metabolomic samples taken at matched time points, with metabolomic changes being connected to a transcriptomic changes at a much earlier time points and vice versa. In cases where the time course data exist, alignment will be an important step before the correlative associations are evaluated.

Data set concatenation-based integration

Data set concatenation-based integration methods are some of the earliest and conceptually some of the simplest methods for combining data sets across multiple omics platforms into a

single model. By concatenating the data tables produced by each omics technology into a single data table, standard techniques such as Self-Organizing Maps [28, 29], K-means cluster analysis [30] or random forests [31] can be applied. Daub [32] presented MetaGeneAlyse, a web service that can run many of these standard methods on concatenated metabolomics–transcriptomics data sets.

One complication in concatenative data integration occurs when the data sets that are to be integrated have vastly different sizes. For instance, if the metabolomics measured 100 metabolites, compared with 10 000 transcripts, any concatenative model would be dominated by the patterns seen in the gene data. To overcome this problem, one can apply block scaling factors to each block, weighting each variable by the inverse of the number of variables in the block [33], introducing the assumption that every block should have equal weight in the final model.

As well as having vastly different sizes, metabolomics and transcriptomics data sets are obtained from vastly different technologies, which implies that the data sets have different structures, different patterns of expected values, different distributions of underlying noise and different variances. Consequently, obtaining integrative links between metabolomics and transcriptomics data from a simple concatenated data set is not straightforward. It is thus normal that when clustering a concatenated data set, elements from each data set will tend to cluster with other elements of their own data set, obscuring any inter-omics associations. This problem can be particularly acute, if the concatenated data set is being processed by PCA, PLS or another variance maximizing algorithm, as sources of analytical variation will typically be selected in the first components. As the sources of analytical variation will vary between the two halves of the concatenated data set, this will lead to a fragmented model, which will not inform about the joint variation seen in the data sets. More advanced methods, as described in the multivariate models section, are needed to counter this problem.

Using tools such as iCluster [34], which cluster the data sets both together and independently, may allow for some of these restrictions to be circumvented, although more research needs to be done to evaluate this. The effect of these underlying distribution differences can be minimized by removing the elements with little or no signal, where the structure of the noise will have a much larger impact, and through applying a common normalization and scaling to the data. Even so, it is not possible through preprocessing to remove these effects entirely. In summary, whilst concatenated approaches are simple to implement, their added value is inherently limited by the concerns documented above, and therefore, we believe that the other options presented in this review are more suited to the task of metabolomic–transcriptomic data integration.

Multivariate-based integration

Moving beyond the relatively simple and straightforward methods described in the previous two sections, multivariate modelling will now be explored. These methods are simultaneously much more powerful for data integration and much more complex, and many of them have been applied in the chemometrics domain before coming to the attention of bioinformaticians.

The two most common multivariate techniques are PCA and PLSs, which are both well covered in [35]. Both PCA and PLS are particularly useful for data sets with high levels of colinearity, as is the case with omics data, where many genes or metabolites will have similar (colinear) profiles. When using multivariate models for omics data integration, one may seek to use one data set to predict aspects of another data set or to find the 'covariance' associations between the two data sets. Unlike the work in the previous section, here the data sets are used in a non-concatenative way, keeping the metabolomics and transcriptomics data sets in separate blocks or dimensions within the model.

The earliest example of PLS models being used for metabolomic–transcriptomic data integration is from Griffin *et al.* [36], who integrated metabolomics and transcriptomics data from two different strains of rats treated with orotic acid. They used the metabolomic nuclear magnetic resonance (NMR) spectral regions as the *x* for the model, and the transcriptomics data as the *y* to be predicted. Once a model had been constructed, they examined the metabolites that were associated with various genes through the loadings from the model. Alongside other analyses, these models helped them to explain why one strain of rats was more susceptible to fatty liver disease following dosing. In a later paper [37], these authors reversed this process by using transcriptomics profiles as the *x* block and the NMR-integrated peaks as the *y*, to study the links between metabolomic and transcriptomic changes in rats following phenobarbital exposure, finding changes in hepatic metabolism, oxidative stress and cyctochrome P450 induction. There are also several examples where PLS is used with transcriptomic profiles as the *x* and measurements of two or three targeted metabolites as the *y* [38, 39]. Sparse regression models have also been used by Jauhiainen *et al.* to predict the metabolic profile from the transcriptomic profile in a similar manner [40].

Using Griffin's approach, it is necessary to define which of the data sets will be used as the *x* and which as the *y*, and these two are then non-equivalent in the model. This is not ideal for those data integration studies, where the mutual relationships between metabolites and transcripts are to be studied. Therefore, a more natural method to use in these situations is an extension of PLS, called O2PLS. In this algorithm, the *x* and the *y* are symmetrical, so the assignment of which data set will be *x* and which will be *y* is inconsequential. O2PLS extends the standard PLS model with one block modelling the associations between *x* and *y*, then one block each modelling the remaining modelable parts of *x* and *y* separately and then as before the unmodelable residuals (see Figure 3).

O2PLS has been applied by Rantalainen *et al.* to metabolomic and proteomic data sets from a human prostate xenograft model in mice [41]; by Eveillard *et al.* for examining liver transcripts and blood plasma metabolites after di(2-ethylhexyl)-phthalate exposure in humans [42]; by Grimplet *et al.* for comparing metabolomics, transcriptomics and proteomics data sets from grapevines [43]; and by Bylsojö *et al.* for integrating metabolomics and transcriptomics data sets from populus tree crosses and with the addition of proteomics data from hybrid aspen trees [44, 45].

Through O2PLS modelling, it becomes possible to obtain the percentage of variance of each omics data set that is modelable by the other data set. Using this approach with metabolomics, transcriptomics and proteomics data, Bylesjö *et al.* [44, 45] found that approximately one-third of each data set was modelable from the other data sets. This work reinforces that there is much in each data set that is unique to that data set and cannot be simply linked to parts of the other data sets. This helps us to understand why it is often difficult to find as many associations between data sets as one might naively expect.

Related to these methods is another subclass called 'multi-block methods', where the different omics (or other data) are

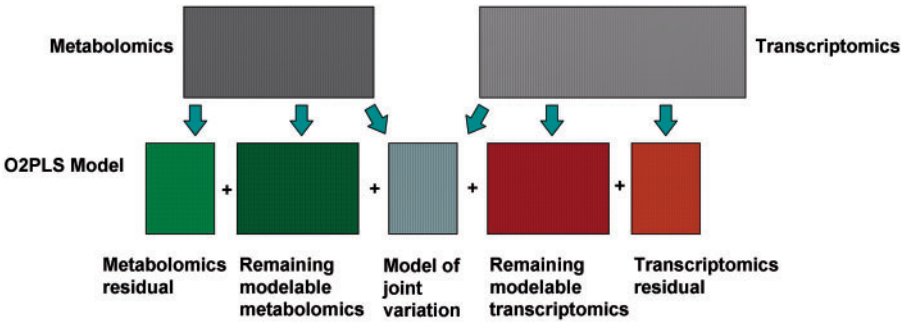

Figure 3. O2PLS models—the makeup of a standard O2PLS model for metabolomic–transcriptomic data integration, showing the different matrices involved.

treated as different blocks in the model and thereupon combined. The simplest of these methods is consensus PCA, which was first applied to metabolomics and transcriptomics data integration by Heijne *et al.* [46]. Consensus PCA performs a standard PCA analysis on data that is split into blocks of variables (for instance, a block of transcript measurements and a block of metabolite measurements) and undergoes a block normalization beforehand. In addition to the standard scores and loadings, this method produces super weights, which indicate how the blocks are related to each other [47]. These super weights generated extend this method beyond a simple PCA on a concatenated data set, as was seen previously. Heijne *et al.* [46] uses this method to integrate data sets obtained from rats with chemically induced hepatic necrosis in a source-matched study examining blood and urine metabolomics alongside liver transcriptomics. The consensus PCA model produces a scores plot showing the variation of the samples across the combined data set, and then evaluates the metabolites and genes that are most important to the separation. The interpretation and visualization of these models was extended by Hassani *et al.* [48], who developed a range of plots and illustrated these on a five-block data set containing genomic (in this case, not transcriptomic but Amplified Fragment Length Polymorphism genetic fingerprinting) and Fourier Transform Infrared metabolomics data along with phenotypic end points.

There also exists a selection of multiblock PLS methods that have been applied to transcriptomic–metabolomic data [49], such as Li *et al.* used in a complex pipeline to identify liver toxicity pathways in HepG2 cells after exposure to free fatty acids and tumour necrosis factor-alpha. However, as yet, no work has compared this approach with other more standard approaches for identifying pertinent pathways.

More recently, a consensus form of the Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) method was introduced by Boccard *et al.* [50] and tested on among others, data from metabolomics, transcriptomics and proteomics experiments with the NCI60 cell lines, with the results in this case validated through bootstraping. In all three test cases, the combination of the different omics data sets was found to be useful to achieve coherent biological interpretations. The main advantage of consensus OPLS-DA over O2PLS is that it can cope with more than two blocks of data and treats all these blocks identically; however, the model does not provide information about the interrelated features between the data sets, but instead regresses all the data against a class variable (discriminant analysis).

Another alternative to the multiblock methods are multiway methods. Here, the multiple blocks are stacked to produce an *N*-dimensional data set. One example of such a multiway method is N-PLS, which Conesa *et al.* [51] applied to the data from [46]. As well as integrating transcriptomics and metabolomics data, they also include physiological end points, which then form an additional dimension. The authors are able to interpret the toxicological meaning of the components in the models built, and between 61 and 70 of the 99 genes determined in the original paper were identified by their method.

Hierarchical models also allow for the integration of different data types. In these models, the individual data sets first undergo a standard PCA or PLS-DA modelling step, before the scores from the first *n* components of these models are combined as the inputs to a further PCA or PLS-DA modelling step. In terms of metabolomic and transcriptomic data integration, this approach has been applied by Spicker [33] to combine three data sets; clinical chemistry, transcriptomics and metabolomics from rats dosed with three toxicants. They compared the models built through this hierarchical process with those built from a concatenative approach, and concluded that the hierarchical approach was better, in that it led to improved class separation.

The final multivariate model-based method we review, is canonical correlation analysis (CCA), which has been applied to transcriptomics and gas chromatography–mass spectroscopy metabolomics data from *Escherichia coli* undergoing a stress response by Jozefczuk *et al.* [30]. In CCA, one searches for linear combinations of variables from one block that correlate with linear combinations of variables from the other block. The canonical variate pairs are then reported, along with their significance, and the linear combinations can be visualized similarly to a PCA loadings plot. Jozefczuk, having limited the inputs to those genes and metabolites involved in primary metabolism found, with one exception (the mqo gene which grouped with TCA-cycle metabolites), the strongest associations grouped only metabolites from the same pathway without linking them to any genes. A free R package, called integrOmics [52], implements CCA.

In summary, there are many options available to researchers wishing to use multivariate modelling for data integration, however, the choice of an appropriate method and the interpretation of the resulting models is complex. In our opinion, the field is lacking in comparison work, which evaluates these options against each other and compares the results obtainable from each. A key challenge is to produce tools that will make the methods not only usable, but also interpretable and easily applicable by non-experts, alongside further investigation into the limitations and advantages of each of the methodologies and their accompanying visualizations. If these objectives are met, then tools such as O2PLS could become powerful parts of the standard toolkits for integrative omics data analysis.

Pathway-based integration

When integrating transcriptomic and metabolomic data, we often wish to interpret these data against the backdrop of the

existing biological knowledge pertaining to how these entities are linked. Much of this information is contained in pathway databases available online. The use of these pathway databases to examine the connections between genes that respond in a particular way to a stimulus is well established, with a wide range of statistical tools to examine the over-representation and enrichment of pathways available.

The extension of these methods to metabolomics data is increasingly available, through tools such as Metabolite Set Enrichment Analysis [53] or the metabolite set analysis option in ConsensusPathDB [54].

The integration of transcriptomics and metabolomics data through pathways can be approached through several tools and methods. Some early work by Cavill *et al.* [10] took publicly available data from the NCI60 set of 57 cancer cell lines, where baseline profiles of untreated cells had been measured using metabolomics and transcriptomics through a repeated study. It was examined how the baseline, untreated profiles of the cells correlated to the cell line's sensitivity to a range of 118 potential chemotherapeutics. Focusing firstly on the genes and metabolites that were correlated to a set of platinum drugs, it was found that there was a great increase in sensitivity to pathways by combining both omics data sets. The original integration approach applied took advantage of the repeated study design, as it could be assumed that the metabolite measurements are statistically independent of the gene measurements. Integrated Molecular Pathway Level Analysis (IMPaLA) [55] implements this method as a freely available web-based tool. In addition to this joint over-representation analysis, IMPaLA also offers Wilcoxon enrichment, where two values are supplied for all measured entities, and then pathways are examined based on whether the mean difference between the two values is significantly different from 0. More recently, IMPaLA has been updated to use Fisher's method for combining *P*-values from multiple tests of the same hypothesis [56], making it more accurate when applied to non-repeated study designs. Kaever *et al.* [57] have also explored using this and similar methods for the combination of *P*-values from independent (repeated study) and dependent (all other study designs) data sets and developed a tool MarVis pathway that imbeds this into metabolomic/transcriptomic pathway-based data integration [58].

One commercially available tool MetaCore (GeneGo Inc.) has been used in several studies [59–61]. In MetaCore, the pathway ranking with metabolites and genes is calculated by taking the minimum *P*-value from the separate omics analyses.

Another commercially available tool Ingenuity IPA also performs integrated analysis, where it sorts the pathways based on the sum of the –log (*P*-value) across all data sets. This is equivalent to the independent combination of *P*-values seen in the original IMPaLA.

Other tools available for joint pathway analysis are PathVisio [62], Paintomics [63], InCroMAP [64] and INtegrative Meta-analysis of Expression data (INMEX) [65], all of which produce a joint pathway *P*-value by totalling the number of differentially expressed genes and metabolites and combining this with the total number of measured genes and metabolites. INMEX also produces a topology-based ranking of pathways, where certain genes/metabolites are evaluated as more crucial because of their position in the network and are therefore weighted more heavily in the calculation.

The method for generating a joint *P*-value in the pathway analysis will have a large impact on the pathways reported as differentially expressed. The different approaches currently used are compared in Table 1. If, as per PathVisio and

Table 1. Example pathways and the equivalent *P*-value calculations by the methods currently in use

Pathway	Metabolite significance					Gene significance					Combined significance			
	Metabolites measured	DE metabolites	Metabolites in pathway	DE metabolites in pathway	P (metabolites)	Genes measured	DE genes	Genes in pathway	DE genes in pathway	P(genes)	Minimum P MetaCore	P P (Independent combination) Original IMPaLA, IIngenuity	P (All entities equal) Paintomics, PathVisio, INMEX, InCroMAP	P (Fisher's method) Updated IMPaLA, MarVis
A	50	5	5	5	**4.7e-7**	5000	500	500	0	1	**4.7e-7**	**4.7e-7**	1	**7.3e-6**
B	50	10	10	4	0.097	5000	500	500	60	0.070	0.070	**0.0068**	**0.034**	**0.041**
C	200	25	10	5	**0.003**	5000	500	500	70	**0.002**	**0.002**	**5.4e-5**	**3.4e-3**	**7.8e-5**

Pathway A is only over-represented in the metabolites. Pathway B is not significant for either omics singularly, whilst it does see limited changes in both. Pathway C is significantly changed in both omics. The table shows the different levels of significance given by the integration strategies in these cases. DE = Differentially Expressed and significant pathways ($P < 0.05$ without multiple testing correction) are shown in bold. The software that uses each method for calculating combined significance is also listed in normal type after the combination method.

Paintomics, the genes and metabolites are treated as equivalent entities, then if a pathway contains many more genes than metabolites, the significance of the genes will overwhelm the significance of the metabolites. This is shown in Pathway A in Table 1, where it is clear that the all entities equal strategy misses the significance of this pathway in the metabolomic data. Because many pathways are indeed imbalanced in one direction or the other, this is an important consideration. MetaCore uses the minimum *P*-value from either transcriptomics or metabolomics analyses, while the original IMPaLA and Ingenuity calculate the significance of each pathway in the genes and the metabolites separately and then use the assumed statistical independence to combine these values, whereas the updated IMPaLA and MarVis [58] use Fisher's method to combine the *P*-values, treating the tests as repeated tests of the same hypothesis. Pathway C in Table 1 shows a situation where the pathway is significantly changed in both omics data sets. Here, all measures will declare the pathway significant; however, the level of significance varies by some orders of magnitude.

Another key consideration in performing over-representation-based pathway analysis, which most of these tools provide, is the composition of the background list. The background list should contain all the entities that were measured, irrespective of whether they were deemed statistically significant between the conditions. These background lists are crucial to the analysis, and omitting a background list will result in a different list of significant pathways. With transcriptomics, such a background list is easy to obtain during the processing of the data. However, for metabolomics data, full assignments are rarely performed, and so many of the peaks that are tested for significance will be unassigned. In addition, there are two key biases that will affect the pathway analysis results. Firstly, the measured metabolites will be biased towards certain groups and pathways, depending on the analytical method used. For instance, amino acids are easy to assign from NMR spectra, and therefore a larger proportion of the identified (and potentially significant) metabolites will be amino acids than would be expected if one picked metabolites to assign at random. Without a background list stating this bias, pathways involving amino acids will always be significantly over-represented. Secondly, in metabolomics, identifying metabolites in the spectra can take a considerable amount of time from a skilled researcher, and often studies will only invest the time to identify the metabolites that are altered between experimental conditions, leading to further bias in the list of identified metabolites from a data set. One initiative that would meet this need, would be a public resource of well-assigned examples from a selection of frequently used sample types, cell lines, tissues, biofluids, etc, measured on a range of metabolomic platforms and prepared under different protocols, so that baseline lists of measurable metabolites would be available for many situations. As an intermediate step, we encourage researchers to publish the background lists used in their metabolomics pathway analysis as supplementary information, to not only enable the community to reach a consensus as to the optimal background list for each sample type—analytical platform pair—but also to provide adequate information for the analysis to be reproduced.

Finally, the results from the pathway analysis will also be affected by the mapping between identifiers. Different identifiers of genes/metabolites have many-to-many relationships with identifiers in other naming systems, and this problem can be acute with metabolomics data. For instance, in the case of lactate and KEGG identifiers (see Table 2), there are three KEGG identifiers that relate to lactate, C00186 the identifier for L-lactate (S-lactate), C00256 the identifier for D-lactate (R-lactate) and C01432 the identifier for Lactate. Whether these chiral metabolites are distinguishable depends on the details of the analytical platform. Because they will have the same mass, often specially selected columns need to be used in mass spectrometry to make sure they elute at different times. Imagine that a lactate peak has been measured without the capability to discriminate between the L and D form and found to be significantly different between conditions. In humans, the lactate pool contains mostly L-lactate, but there is also a measurable amount of bacterial produced D-lactate [66] that can be metabolized by endogenous human enzymes found in the pyruvate metabolism pathway in KEGG. Because, in this case, the measurement is known to be of a pool of both isomers, one may choose to map the lactate peak measurement to the identifier for Lactate (C01432); however, this identifier is not present in any KEGG pathways (see Table 2), so then this peak will then have no influence on the pathway analysis. If one maps the measurement to both D-lactate and L-lactate, then the pyruvate metabolism pathway that contains both of these isomers, will appear to have two significant metabolites from one significant peak, and therefore will have a chance of appearing falsely significant. In this case, mapping to L-lactate alone is the best option, as L-lactate is present in all the KEGG pathways in which D-lactate is present, and additionally, this matches the biological reality that the lactate pool is mostly L-lactate. However, there is no optimal rule for the general case, and so, for each similar case, the possibilities must be evaluated by hand, taking into account in which pathways a particular identifier is present (and re-evaluated when pathway databases are updated). An additional complication is caused by peaks that are obtained from overlapping metabolites, or which have ambiguous assignments. Again, publication of these mapping decisions as supplementary information will greatly benefit the community and the reproducibility of work.

Future prospects

There have been two recurring messages throughout this review. Firstly, that the direction and strength of associations between metabolites and gene expressions vary strongly between experimental conditions, because of the nature of the complex networks in which the gene expressions and metabolites are embedded. Secondly, that these associations are hidden by the

Table 2. Lactate identifier mappings with the KEGG pathways containing each identifier

Measured	Metabolite	KEGG id	KEGG pathways
All lactate	L-lactate	C00186	Glycolysis/gluconeogensis; pyruvate metabolism; propanoate metabolism; styrene metabolism; metabolic pathways; biosynthesis of secondary metabolites; microbial metabolism in diverse environments; HIF-1 signalling pathway
	D-lactate	C00256	Pyruvate metabolism
	Lactate	C01432	None

noise and variance structure inherent in each data set and often not as abundant as one would naively expect. This conversely leads to the positive observation that the different data sets have much complementary information content, and therefore integrated analyses have the potential to reveal much more of the biology than the total of the separate analyses.

As multi-omics studies become increasingly financially and technically feasible, the application of these methods will continue to expand and will need to develop. We imagine that the correlation and concatenative methods will continue to be the first port of call for many researchers, because of the simplicity and familiarity of the methods from single-omics data analysis—however, as observed above, these analyses have many failings and can only give limited insights into these data sets. Multivariate methods can generate valuable insights into the amount and content of the overlapping information contained in complementary omics data sets, along with strong predictive models for phenotypic end points. However, the tools are not often designed to be user-friendly, and furthermore, the interpretation of these models is complex. Choosing which model to best apply can be bewildering, and there exists a myriad of factors (such as scaling) that can significantly affect the results, making their application even harder. However, we feel that these methods have incredible potential, and the pay-off from investing the time to become an expert will be substantial for scientists with the inclination. For biologists, the pathway-based methods are the most intuitive and give ample relevant information towards interpretations of the data. However, it is important to note, that because these methods are based on existing pathway knowledge, they can never be used to discover *de novo* gene-metabolite associations, and so there will continue to be a need for other techniques to explore these links, especially outside the relatively well-explored domains of model species.

Returning to the study designs presented at the start of the article, it is clear that the study design chosen impacts the optimal analysis type. For instance, O2PLS is best suited towards split sample data, where the joint information content between the data sets is maximized, while in our experience with repeated studies, the overlapping information is often insufficient to generate a robust model. Importantly, in writing this review, it was often noted that it was difficult to determine the study design used from the descriptions of the methods in the literature. Given the strong impact of study design choice on the optimal analysis, this is concerning. We believe that the impact of the study design on multi-omics integration methods should be further explored to fully understand how the different methods behave and where misleading results may appear.

Overall, the outlook for the application of multi-omics studies is bright, because of a growing set of methodologies and tools in place for the analysis and interpretation. This will undoubtedly lead to many new advances in systems understanding of biology.

Key Points

- Transcriptomic and metabolomic data sets contain complementary information, so it is important that statistical data integration is performed.
- Some simpler data integration methods do not work because of the inherent differences in the distributions of data points in the data sets because of the underlying analytical technologies.
- Study design matters for the data integration steps.

Funding

RC was funded by BE-Basic (grant no. FES0905).

References

1. Shi L, Reid LH, Jones WD, *et al*. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006;**24**:1151–61.
2. Brazma A, Hingamp P, Quackenbush J, *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;**29**:365–71.
3. Fiehn O, Robertson D, Griffin J, *et al*. The metabolomics standards initiative (MSI). *Metabolomics* 2007;**3**:175–8.
4. Shi L, Campbell G, Jones WD, *et al*. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol* 2010;**28**:827–38.
5. Gygi SP, Rochon Y, Franza BR, *et al*. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 1999;**19**:1720–30.
6. Chen G. Discordant protein and mRNA expression in Lung Adenocarcinomas. *Mol Cell Proteomics* 2002;**1**:304–13.
7. Palsson BO, Zengler K. The challenges of integrating multi-omics data sets. *Nat Chem Biol* 2010;**6**:787–98.
8. Tillinghast GW. Microarrays in the clinic. *Nat Biotechnol* 2010;**28**:810–2.
9. Leek JT, Scharpf RB, Bravo HC, *et al*. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;**11**:733–9.
10. Cavill R, Kamburov A, Ellis JK, *et al*. Consensus-phenotype integration of transcriptomic and metabolomic data implies a role for metabolism in the chemosensitivity of tumour cells. *PLoS Comput Biol* 2011;**7**:12.
11. Ebbels TMD, Cavill R. Bioinformatic methods in NMR-based metabolic profiling. *Prog Nucl Magn Reson Spectrosc* 2009;**55**:361–74.
12. Cavill R, Sidhu JK, Kilarski W, *et al*. A combined Metabonomic and Transcriptomic approach to investigate metabolism during development in the Chick Chorioallantoic Membrane research articles. *J Proteome Res* 2010;**9**:3126–34.
13. Fan TWM, Bandura LL, Higashi RM, *et al*. Metabolomics-edited transcriptomics analysis of Se anticancer action in human lung cancer cells. *Metabolomics* 2005;**1**:325–39.
14. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
15. Kelder T, Pico AR, Hanspers K, *et al*. Mining biological pathways using WikiPathways web services. *PLoS One* 2009;**4**:e6447.
16. Thomas CE, Ganji G. Integration of genomic and metabonomic data in systems biology-are we 'there' yet? *Curr Opin Drug Discov Devel* 2006;**9**:92–100.
17. Urbanczyk-Wochniak E, Luedemann A, Kopka J, *et al*. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep* 2003;**4**:989–93.
18. Fendt SM, Buescher JM, Rudroff F, *et al*. Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. *Mol Syst Biol* 2010;**6**:356.
19. Ter Kuile BH, Westerhoff HV. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett* 2001;**500**:169–71.
20. Moxley JF, Jewett MC, Antoniewicz MR, *et al*. Linking high-resolution metabolic flux phenotypes and transcriptional

regulation in yeast modulated by the global regulator Gcn4p. *Proc Natl Acad Sci USA* 2009;**106**:6477–82.
21. Bradley PH, Brauer MJ, Rabinowitz JD, *et al*. Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae. *PLoS Comput Biol* 2009;**5**:e1000270.
22. Askenazi M, Driggers EM, Holtzman DA, *et al*. Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat Biotechnol* 2003;**21**:150–6.
23. Rantalainen MJ, Bjerrum JT, Olsen J, *et al*. Integrative transcriptomic and metabonomic molecular profiling of colonic mucosal biopsies indicates a unique molecular phenotype for ulcerative colitis. *J Proteome Res* 2014;**14**:479–90.
24. De la Fuente A, Bing N, Hoeschele I, *et al*. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinform* 2004;**20**:3565–74.
25. Kayano M, Imoto S, Yamaguchi RUI, *et al*. Multi-omics approach for estimating metabolic networks using low-order partial correlations. *J Comput Biol* 2013;**20**:571–82.
26. Cavill R, Kleinjans JCS, Briedé JJ. Dynamic time warping for omics. *PLoS One* 2013;**8**:e71823.
27. Lu Y, Boekschoten MV, Wopereis S, *et al*. Comparative transcriptomic and metabolomic analysis of fenofibrate and fish oil treatments in mice. *Physiol Genomics* 2011;**43**:1307–18.
28. Hirai MY, Yano M, Goodenowe DB, *et al*. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc Natl Acad Sci USA* 2004;**101**:10205–10.
29. Hirai MY, Klein M, Fujikawa Y, *et al*. Elucidation of gene-to-gene and metabolite-to-gene networks in arabidopsis by integration of metabolomics and transcriptomics. *J Biol Chem* 2005;**280**:25590–5.
30. Jozefczuk S, Klie S, Catchpole G, *et al*. Metabolomic and transcriptomic stress response of Escherichia coli. *Mol Syst Biol* 2010;**6**:364.
31. Acharjee A, Kloosterman B, de Vos RCH, *et al*. Data integration and network reconstruction with ~ omics data using random forest regression in potato. *Anal Chim Acta* 2011;**705**:56–63.
32. Daub CO, Kloska S, Selbig J. MetaGeneAlyse: analysis of integrated transcriptional and metabolite data. *Bioinformatics* 2003;**19**:2332–3.
33. Spicker JS, Brunak S, Frederiksen KS, *et al*. Integration of clinical chemistry, expression, and metabolite data leads to better toxicological class separation. *Toxicol Sci* 2008;**102**:444–54.
34. Shen R, Mo Q, Schultz N, *et al*. Integrative subtype discovery in glioblastoma using iCluster. *PLoS One* 2012;**7**:e35236.
35. Lindon JC, Holmes E, Nicholson JK. Pattern recognition methods and applications in biomedical magnetic resonance. *Prog Nucl Magn Reson Spectrosc* 2001;**39**:1–40.
36. Griffin JL, Bonney SA, Mann C, *et al*. An integrated reverse functional genomic and metabolic approach to understanding orotic acid-induced fatty liver. *Physiol Genomics* 2004;**17**:140–9.
37. Waterman CL, Currie RA, Cottrell LA, *et al*. An integrated functional genomic study of acute phenobarbital exposure in the rat. *BMC Genomics* 2010;**11**:9.
38. Pir P, Kirdar B, Hayes A, *et al*. Integrative investigation of metabolic and transcriptomic data. *BMC Bioinformatics* 2006;**7**:203.
39. Li Z, Chan C. Integrating gene expression and metabolic profiles. *J Biol Chem* 2004;**279**:27124–37.
40. Jauhiainen A, Nerman O, Michailidis G, *et al*. Transcriptional and metabolic data integration and modeling for identification of active pathways. *Biostatistics* 2012;**13**:748–61.
41. Rantalainen M, Cloarec O, Beckonert O, *et al*. Statistically integrated metabonomic-proteomic studies on a human prostate cancer xenograft model in mice. *J Proteome Res* 2006;**5**:2642–55.
42. Eveillard A, Lasserre F, de Tayrac M, *et al*. Identification of potential mechanisms of toxicity after di-(2-ethylhexyl)-phthalate (DEHP) adult exposure in the liver using a systems biology approach. *Toxicol Appl Pharmacol* 2009;**236**:282–92.
43. Grimplet J, Cramer GR, Dickerson JA, *et al*. VitisNet: 'Omics' integration through grapevine molecular networks. *PLoS One* 2009;**4**:e8365.
44. Bylesjö M, Eriksson D, Kusano M, *et al*. Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data. *Plant J* 2007;**52**:1181–91.
45. Bylesjö M, Nilsson R, Srivastava V, *et al*. Integrated analysis of transcript, protein and metabolite data to study lignin biosynthesis in hybrid aspen. *J Proteome Res* 2009;**8**:199–210.
46. Heijne WHM, Lamers RJAN, van Bladeren PJ, *et al*. Profiles of metabolites and gene expression in rats with chemically induced hepatic necrosis. *Toxicol Pathol* 2005;**33**:425–33.
47. Westerhuis JA., Kourti T, MacGregor JF. Analysis of multiblock and hierarchical PCA and PLS models. *J Chemom* 1998;**12**:301–21.
48. Hassani S, Martens H, Qannari EM, *et al*. Analysis of -omics data: graphical interpretation- and validation tools in multiblock methods. *Chemom Intell Lab Syst* 2010;**104**:140–53.
49. Li Z, Chan C. Systems biology for identifying liver toxicity pathways. *BMC Proc* 2009;**3**(S2):S2.
50. Boccard J, Rutledge DN. A consensus OPLS-DA strategy for multiblock Omics data fusion. *Anal Chim Acta* 2013;**769**:30–9.
51. Conesa A, Prats-Montalbán JM, Tarazona S, *et al*. A multiway approach to data integration in systems biology based on Tucker3 and N-PLS. *Chemom Intell Lab Syst* 2010;**104**:101–11.
52. Lê Cao KA, González I, Déjean S. IntegrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* 2009;**25**:2855–6.
53. Xia J, Wishart DS. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res* 2010;**38**:W71–7.
54. Kamburov A, Pentchev K, Galicka H, *et al*. ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* 2011;**39**:D712–7.
55. Kamburov A, Cavill R, Ebbels T, *et al*. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* 2011;**27**:2917–8.
56. Fisher RA. *Statistical methods for research workers*. *Biological Monographs and Manuals*, Chapter 4, Eleventh revised edition 1950, pp. 99–100.
57. Kaever A, Landesfeind M, Feussner K, *et al*. Meta-analysis of pathway enrichment: combining independent and dependent omics data sets. *PLoS One* 2014;**9**:e89297.
58. Kaever A, Landesfeind M, Feussner K, *et al*. MarVis-Pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data. *Metabolomics* 2014;**11**:764–77.
59. Jennen D, Ruiz-Aracama A, Magkoufopoulou C, *et al*. Integrating transcriptomics and metabonomics to unravel modes-of-action of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) in HepG2 cells. *BMC Syst Biol* 2011;**5**:139.
60. Xu EY, Perlina A, Vu H, *et al*. Integrated pathway analysis of rat urine metabolic profiles and kidney transcriptomic profiles to elucidate the systems toxicology of model nephrotoxicants. *Chem Res Toxicol* 2008;**21**:1548–61.

61. Kleemann R, Verschuren L, van Erk MJ, *et al*. Atherosclerosis and liver inflammation induced by increased dietary cholesterol intake: a combined transcriptomics and metabolomics analysis. *Genome Biol* 2007;**8**:R200.
62. Van Iersel MP, Kelder T, Pico AR, *et al*. Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics* 2008;**9**:399.
63. García-Alcalde F, García-López F, Dopazo J, *et al*. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics* 2011;**27**:137–9.
64. Eichner J, Rosenbaum L, Wrzodek C, *et al*. Integrated enrichment analysis and pathway-centeredvisualization of metabolomics, proteomics, transcriptomics, andgenomics data by using the InCroMAP software. *J Chromatogr B* 2014;**966**:77–82.
65. Xia J, Fjell CD, Mayer ML, *et al*. INMEX—a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res* 2013;**41**:W63–70.
66. Scheijen JLJM, Hanssen NMJ, Van De Waarenburg MPH, *et al*. L(+) and D(-) lactate are increased in plasma and urine samples of type 2 diabetes as measured by a simultaneous quantification of L(+) and D(-) lactate by reversed-phase liquid chromatography tandem mass spectrometry. *Exp Diabetes Res* 2012;**2012**:234812.