*Review*

# Integrating Heterogeneous omics Data via Statistical Inference and Learning Techniques

Ashar Ahmad[1,*], Holger Fröhlich[1,2]

[1]University of Bonn, Bonn-Aachen International Center for IT, Algorithmic Bioinformatics, Dahlmannstr. 2, 53113 Bonn, Germany
[2]UCB BioSciences GmbH, Alfred-Nobelstr. 10, 40789 Monheim, Germany

*Correspondence: ashar@bit.uni-bonn.de

Received 2016-04-13; Accepted 2016-09-09; Published 2016-09-27

## ABSTRACT

Multi-omics studies are believed to provide a more comprehensive picture of a complex biological system than traditional studies with one omics data source. However, from a statistical point of view data integration implies non-trivial challenges. In this review, we highlight recent statistical inference and learning techniques that have been devised in this context. In the first part of our article, we focus on techniques to identify a relevant biological sub-system based on combined omics data. In the second part of our article we ask, in which way integrated omics data could be used for better personalized patient treatment in a supervised as well as unsupervised learning setting. Different classes of algorithms are discussed for both application tasks. Existing and future challenges for data integration methods are pointed out.

## KEYWORDS

data integration; omics data; statistical learning

## INTRODUCTION

During the last years there has been an increasing interest to analyze multiple, heterogeneous omics data in an integrated manner in order to gain a more and more comprehensive picture on complex biological systems [1]. For example, large scale initiatives such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) [2] now provide transcriptomics, methylomics, proteomics and genomics data of hundreds of patients for several cancer entities, allowing novel insights into cancer biology [3]. Historically, expression QTL analysis (see systematic list of abbreviations in Table 1) can be seen as one of the first approaches combining two data modalities, namely information on genetic variations with gene expression [4, 5].

While most authors agree on the chances of omics data integration, the associated challenges have been discussed under a varying point of view over the last decade: While in 2006 data availability was seen as one of the big issues [6], later papers mentioned statistical challenges, such as the risk of overfitting [7], and the difficulties associated with different technical platforms, for example differing normalization protocols and batch effects [8].

Altogether the challenges for integrating heterogeneous omics data may be summarized as follows: omics data of different modality (e.g. transcriptomics vs. proteomics) are measured with different techniques. Hence, these data have differing numerical types (e.g. discrete counts vs. continuous signals) and scales, coupled with large differences in the number of measured features (several hundreds of thousands of SNPs vs. few hundreds of miRNAs). Furthermore, each technical platform has another noise level and sensitivity. Consequently, naive merging of heterogeneous omics data increases the dimensionality of the data and thus increases the chance to produce false positive hypothesis testing results. In a machine learning setting the chance increases to overfit the data. In order to circumvent these problems the key question is therefore, how to identify and combine relevant features from each data modality in a way that respects known biological dependencies.

The goal of this review is to give an overview about recent statistical inference and learning techniques that have been devised to address this issue. Previous reviews focused on specific applications of data integration in the cancer field [8] and genetics [9], on the relevance of data integration for personalized medicine [10], on technical aspects of network integration [11] and dimensionality reduction [12], or provided a high-level view on ongoing research projects [13]. A mathematically oriented review can be found in [14]. As opposed to our paper, the authors emphasize specific mathematical details of selected methods, whereas we try to characterize the overall methods landscape. More specifically, we here highlight two aspects: In the first part of our review we focus on learning and modelling dependencies between different data modalities on the level of individual features. That means the aim is to identify a relevant biological sub-system based on combined omics data. In the following section we ask, in which way integrated omics data could be used for personalized patient treatment in a supervised as well as unsupervised learning setting.

## BRIEF INTRODUCTION OF TECHNICAL TERMINOLOGY

Since this review focuses on technical aspects of omics data integration we would like to briefly introduce some necessary terminology, which is frequently used

| Abbreviation | Full form |
|---|---|
| CCA | Canonical Correlation Analysis |
| CN(V) | Copy Number (Variation) |
| LOH | Loss of Heterozyosity |
| miRNA / mRNA | micro / messenger Ribonucleic Acid |
| ODE | Ordinary differential equation (system) |
| PLS | Partial Least Squares |
| PCA | Principal Component Analysis |
| PPI | Protein Protein Interaction |
| SNP | Single Nucleotide Polymorphism |
| (e)QTL | (expression) Quantitative Trait Locus |

**Table 1:** List of common abbreviations

throughout this paper. While this terminology is commonly used in the machine learning and statistical literature it might not be entirely clear to all researchers with non-computational background. Furthermore, explanations could help to avoid misunderstandings due to partially ambiguous use of some terms in different scientific communities. Table 2 shows an overview about several frequently used terms together with explanations.

As already outlined in [13] "data integration" can have multiple meanings in different scientific contexts. Here we refer to data integration or data fusion as a statistical or machine learning based approach to extract information from multiple data sources / modalities (e.g. gene plus protein expression data). These information can be used to fit / train / learn statistical models. For example, such a model could be a classifier used to predict the disease state of a cancer patient. In general we distinguish between supervised and unsupervised model learning. In supervised learning we train a model based on a possibly large number of omics features coupled with the known outcome / phenotype for each training sample. In the previous cancer example, training samples would be patients with measured omics data plus known disease state. Notably, supervised learning is not restricted to classification, but also real valued outcomes (e.g. disease severity scores, survival) can be predicted via regression models.

As opposed to supervised learning, unsupervised learning aims at inferring patterns from data without having access to a phenotype. For example, unsupervised clustering of gene expression data from Glioblastoma Multiforme patients has been used to identify several disease subtypes [15].

A general concern with all machine learning models is overfitting, meaning a good fit to the training data, but poor prediction performance on further test data at model application phase. Overfitting can be attributed to overly complex models compared to the limited amount of available training data. Hence, model training should a priori favor simpler models over more complex models. For example, this can be achieved via a mathematical technique called "regularization". Furthermore, in the omics field variable / feature selection is essential, because data samples (e.g. representing patients) are sparsely distributed in an extremely high dimensional space, which is spanned by measured omics features. Hence, with typical number of training samples a huge set of different models could exist, which could equally well fit the data. The true model is thus typically non-identifiable without further background information. Reducing the number of features thus also reduces the set of possible models and thus lowers the chance to overfit the training data.

Notably, feature selection is not necessarily identical to differential expression analysis or similar techniques to identify most significant omics features. Hypothesis tests are typically conducted separately for each omics feature, yielding a p-value. However, non-significant features could be highly informative for a model in combination with other features, e.g. to separate two groups of patients with the help of a multi-variate classifier. Hence, it is important to think about feature selection from a multi-variate perspective.

Most machine learning models are trained by optimizing a certain objective function. However, some models also aim at describing the entire, multivariate statistical distribution, from which the training data has been drawn as samples. These models are known as probabilistic models. Many of these probabilistic models can be depicted as graphs, where nodes represent random variables and edges conditional statistical dependencies (graphical models). Examples thereof are Bayesian Networks.

Typical machine learning models are purely data driven, i.e. they extract statistical associations from training data without background information about the biological context. Therefore, extracted patterns may or may not correspond to any known biological mechanism. A conceptually different approach is thus to focus only on established biological knowledge, e.g. a given pathway or network structure, to which then biological data is mapped to qualitatively or quantitatively to better understand a biological system. Such an approach is called knowledge driven.

## LEARNING AND MODELLING MOLECULAR FEATURE DEPENDENCIES BETWEEN DIFFERENT DATA SOURCES

### Sequential Analysis

After having introduced basic terminology in the preceding Section we now focus on computational approaches for learning and modelling feature

| Vocabulary used | Explanation |
|---|---|
| data integration / data fusion | Statistical approaches to extract information from multiple data sources |
| data source / data modality | Individual omics data source (e.g. gene expression, DNA methylation) |
| supervised, unsupervised learning | Supervised learning techniques employ omics data together with known biological phenotypes or outcomes (e.g. disease, healthy). Unsupervised learning methods aim at discovering patterns in data without information about a phenotype. An example is clustering |
| classification, regression | Both are supervised approaches. While classification deals with discrete phenotypes, regression usually takes into account continuous phenotypes (e.g. disease severity scores, survival) |
| clustering | Unsupervised statistical learning techniques to discover groupings in the data |
| overfitting | Tendency of statistical models to fit training data well, but perform poorly on further test data at model application phase. Overfitting can be attributed to overly complex models compared to the limited amount of available training data |
| Regularization | Mathematical technique to favor simpler over more complex models |
| model identifiability | Ambiguity in the unique determination of the model based on available training data |
| hypothesis testing | A statistical framework to evaluate the unlikeliness of an observation due to pure chance |
| feature / variable selection | Selection of most informative subset of omics features |
| probabilistic model | Model described via dependencies of random variables. Aim is to describe and model a (multivariate) statistical distribution |
| data driven, knowledge driven modelling | Data driven machine learning techniques capture statistical associations in data, which may or may not correspond to anything biologically known. Knowledge driven approaches focus on representing established biological knowledge |

**Table 2:** Explanation of Technical Terminology

dependencies between different data modalities. An overview about the techniques discussed in this Section can be found in Table 3.

Probably the most straight forward approach to integrate heterogeneous omics data is an independent analysis of each omics data modality. As a second step, correspondences between relevant features based on biological background knowledge are investigated. For example, as a first step significant CNVs and gene expression changes can be determined. In the second step significant CNVs can be mapped to genes, which are then overlapped with differentially expressed genes [16, 17]. This type of analysis can be extended to more than two data modalities: Sun et al. looked for overlaps between differentially expressed and differentially methylated genes as well as for regions with statistically significant copy number variations in breast cancer cells [18]. Chari et al. also integrated loss of heterozygosity profiles and investigated, in how far the direction of observed expression changes matched with CNV and DNA methylation status in different breast cancer cell lines [19].

A principal limitation of sequential analysis is that relevant features from each dataset are only determined with respect to the phenotype. Hence, there can be non-significant features, which nonetheless correlate well with features from another data modality. For example, a certain SNP could by itself not demonstrate a clear separation between two clinical groups, but nonetheless shows a clear statistical effect on gene expression. Such a feature is missed in sequential analysis. Moreover, significance cutoffs in each omics data modality are defined independently. Hence, features in the overlap can be biologically non-concordant just because truly relevant features are above the significant cutoff in one of the data sources. Likewise, false positive features can yield non-concordance.

### Correlation, Covariance and Regression Based Techniques

Another approach is to directly look for significant correlations between features from different data modalities. For example, CNVs and gene expression can be correlated with each other [20, 21] as well as expression of miRNAs and genes [22]. Correlation is closely related to regression analysis, which allows for potential inclusion of further covariates. For example, linear regression is frequently employed to identify eQTLs from combined SNP and gene expression data [23, 24].

Classical linear regression techniques require fewer predictor variables than samples. Penalized linear regression techniques overcome this limitation: In consequence lasso, group lasso and elastic net penalized methods have been proposed, e.g. to model the combinatorial effect of miRNA on gene expression [25–28]. Also non-linear machine learning techniques such as gradient boosting and Random Forests have recently been used to model influences of the transcriptome on protein expression [29] and to perform eQTL mapping [30].

The above mentioned methods essentially assume a dependency of one particular feature in one data modality A on one or several other features in a second data modality B. For example, the expression level of one particular gene, in eQTL studies, is modeled as a function of one or several SNPs. This type of analysis fails, if clear dependencies only exist between feature *combinations* in data modality A and feature combinations in data modality B. Detecting these types of correlations is essentially the motivation behind canonical correlation analysis (CCA) [31]. Briefly, the idea in classical CCA is to construct in data modality A a canonical variable, which is a linear combination of existing features and as much as possible correlated with the canonical variable in data modality B. Similar to PCA, further canonical variables can be found under the additional constraint of being uncorrelated with the preceding ones. In order to use CCA for integration of different omics data modalities, sparse variants have been recently developed [32, 33]. Lasso and elastic net penalized CCA variants have been successfully applied in several studies to integrate CNVs and gene expression data [32, 34, 35] as well as SNP information with fMRI measurements [36].

Whereas CCA focuses on directions of maximal correlation, partial least squares regression (PLS) aims for modelling latent variables, which explain maximal covariance. Sparse PLS variants have been used to combine different omics data modalities [37] and to map genetic markers to complex phenotypes [38]. Sparse PLS has also been compared to sparse CCA methods, indicating overall similar results with the elastic net penalized CCA [39].

Other approaches for omics data integration, which do not fall into one of the aforementioned categories include independent component analysis [40], generalized singular value decomposition [41], co-inertia analysis [42], sparse factor analysis [43] and kernel PCA [44].

### Bayesian Networks

Bayesian Networks (BNs) belong to the family of graphical models and offer a completely probabilistic view on data integration. BNs explicitly model conditional statistical dependencies between random variables [45]. Typically, each random variable represents one molecular feature (e.g. a protein). Integration of different omics data modalities (e.g. gene and protein expression) can be performed by discretizing each dataset while including additional auxiliary indicator variables describing the biological context [46]. Another option is to use e.g. gene expression as primary data source and employ further data (such as protein-protein interactions, DNA-protein interactions, histone modifications or combinations of several sources) to construct informative network priors, which facilitate the identification of the true biological network from data [47–49].

Arguably, one of the first applications of BNs for biological data integration can be found in Huttenhower and Troyanskaya [50], where the authors presented a BN integrating gene expression and various functional and relational/interaction data in order to predict functional relationships between proteins. As usual in BN modelling, the suggested method was based on a data discretization in order to harmonize differing data types and to allow for non-linear relationships between variables. Later work showed the possibility to integrate heterogeneous modalities on very large scale via a naive Bayesian classifier system with parameter regularization in order to predict protein functions [51].

Another line of research within the BN framework focuses on extending the module network approach in order to decipher regulatory programs [52]. The essential idea behind the module network algorithm is to group genes based on co-expression. Within the BN framework random variables falling into one module share the same parameters and parents, hence yielding a significant reduction of model complexity and improvement of prediction performance compared to traditional BNs [53]. While the original module network algorithm was based on a greedy strategy to assign network nodes to modules, later variants introduced Gibbs sampling [54] and ensemble learning [55]. More recently, module network variants integrating gene expression data with SNPs [56], CNVs [57], miRNA and clinical [58] as well as potential further data [59] have been proposed. The key idea in all of these modifications is to employ data different from gene expression for selecting candidate drivers/regulators of specific gene modules.

Specific BN variants have also been employed for predicting miRNA and transcription factor combinations explaining gene expression changes based on joint gene and miRNA expression data [60, 61]. The authors of these papers treated regulator activities as hidden binary variables in a special kind of BN, in which the topology was defined by target gene prediction methods. Markov Chain Monte Carlo (MCMC) was then performed for Bayesian inference of these latent variables. miRNA and gene expression measurements were included as observed Gaussian variables. In [61] the author extended the approach further by Bayesian learning of the regulator-regulator dependency network.

### Using Molecular Networks for Data Integration

Another line of research focuses on using biological networks as backbone for heterogeneous data integration. Information about canonical pathways, protein-protein, protein-DNA as well as predicted miRNA-gene interactions can nowadays be found in large scale databases [62–67]. Based on these resources molecular networks can be reconstructed and employed for mapping statistics (e.g. z-scores) of omics data. Accordingly, graph based algorithms can be employed for identifying relevant sub-networks [68]. Dittrich and co-workers interpreted this task as an instance of the Price Collecting Steiner Tree (PCST) problem and came up with a provably optimal solution via integer linear programming [69]. At the same time they demonstrated that in this way gene expression and other data modalities (e.g. clinical information) could be combined.

Another example is the work of Nibbe et al., who integrated gene and protein expression data with the

help of a protein-protein interaction (PPI) network [70]. The authors first looked for differentially expressed proteins at the proteome level. In a second step they scanned potential interaction partners for synergistic dysregulation on mRNA level using a modification of Google's Page Rank algorithm in order to define candidate sub-network. In a last step these candidate sub-networks were then scored via a mutual information based approach.

Rather then generating and scoring sub-networks other researchers have focused on predefined pathways and gene sets. Following this line of research Tyekucheva and colleagues proposed a meta analysis approach following separate gene set investigation for each data modality [71]. As an alternative method the same authors suggested a model based integration of gene-to-phenotype association scores using all data sources together, which is followed by gene set analysis of these scores. Tyekucheva and co-workers in this way were able to integrate gene expression and CNV data from glioblastoma multiforme patients.

In line with the meta analysis idea Sun et al. [72] and Kamborov et al. [73] developed tools, which combine gene set statistics of different data modalities via rank aggregation and consensus p-value method, respectively.

In contrast to the above described gene set analysis methods PARADIGM is a fully Bayesian approach, which takes into account the structure of a molecular pathway as well as further biological background knowledge [74]. PARADIGM allows for annotating molecular network nodes with further functional information (e.g. "apoptosis") as well as information on molecule type (e.g. "DNA", "mRNA", "protein"). In PARADIGM activities of node are viewed as latent variables in a factor graph model. Given observed data, inference about the state of these latent variables is then performed via belief propagation. As a result pathway activities are estimated. The authors have demonstrated that their method is able to integrate mRNA and CNV data as well as additional miRNA and DNA methylation information [75].

Wachter et al. recently published an R-package, which specifically focuses on the integration of transcriptomics and proteomics data via pathways, protein-protein and protein-DNA interactions [76]. These information are used to link differentially expressed proteins to transcription factors as well as pathways. At the same time differentially expressed transcripts are linked to transcription factors and pathways. In a consensus step all results are combined either via simple overlap analysis, via a shortest-path based approximate Steiner tree algorithm [77, 78] or via a dynamic Bayesian Network structure learning algorithm [79].

## UTILIZING INTEGRATED OMICS DATA FOR PERSONALIZED MEDICINE

### Clinical Outcome Prediction

One of the primary goals of personalized medicine is to stratify patients into clinically relevant sub-populations based on suitable biomarker signatures. An overview about the associated statistical learning techniques discussed in this section can be found in Table 4.

During the last decade computational research in the personalized medicine area has mainly focused on learning predictive models based on one data modality (e.g. gene expression), possibly also in combination with biological background knowledge (see [80] for a review). The advent of multiple, heterogeneous omics data modalities from the same patient (e.g. somatic mutations plus gene expression data) now raises the question, whether predictive models utilizing several combined data sources could improve prediction performance. Hence, the primary objective for omics data integration in personalized medicine is not to identify biologically relevant molecular networks, as discussed in the last Section, but to enhance model learning and prediction performance.

In the machine learning community traditionally three general strategies for data integration are distinguished [81, 82]: Early integration methods focus on extraction of common features from several data modalities, resulting into one integrated data matrix. In a second step conventional machine learning methods can then be applied. Late integration algorithms first learn separate models for each data modality and then combine predictions made by these models, for example with the help of a meta-model trained on the outputs of data source specific sub-models. The latter strategy is called stacking [83]. Intermediate integration algorithms are the youngest branch of data fusion approaches. The idea is to join data sources while building the predictive model. An example of this strategy is Support Vector Machine (SVM) learning with linear combinations of multiple kernel functions [84].

All three data integration strategies have been applied in the area of personalized medicine: Pittman et al. [85] integrated clinical and gene expression data into a Bayesian decision tree classifier to predict breast cancer prognosis. Following an early integration approach the authors first summarized gene expression data into meta-genes [86], which were then joined with clinical variables. Selection of relevant variables was carried out via forward selection.

Boulesteix et al. first used partial least squares (PLS) regression to extract features from gene expression data [87]. These features were then combined with clinical variables to train a Random Forest classifier for predicting breast and colorectal cancer outcome. In a similar vein Cao et al. [88] proposed a mixture of experts model to jointly model the effect of gene expression and patient clinical data to predict patient outcomes, they concluded that using gene expression data can provide valuable insights to understanding survival mechanisms by identifying prognostic biomarkers.

Gevaert et al. [89] employed a Bayesian Network to combine clinical and gene expression based on the 70 gene breast cancer signature by van't Veer et al. [90]. The authors compared an early integration strategy based on simple pasting of data matrices with an intermediate and a late strategy. In the intermediate integration the authors first learned separate BN structures for each data sources and then

join these networks based on the node representing the clinical outcome they had in common. In the late strategy only predictions by the two separate BN models were weighted and aggregated. The authors found the intermediate strategy to be most promising.

Daemen and co-workers suggested the use of a multiple kernel learning (MKL) framework to predict disease outcome of rectal cancer based on gene and protein expression data, and of prostate cancer based on transcriptome and CNV data [91]. Within the MKL framework separate kernel functions were defined for each omics data modality. A linear combination of these kernels was then employed to train a least squares SVM (LS-SVM). The authors reported a better prediction performance of this intermediate data integration strategy compared to model stacking.

Following again the idea of MKL, Thomas et al. suggested a weighted LS-SVM classifier to combine gene expression and clinical data [92]. Compared to models built on each individual data modality as well as compared to an early integration strategy using generalized singular value decomposition, the authors found a significant improvement of their approach for predicting breast cancer outcome.

Wang et al. [93] developed an integration scheme based on probabilistic graphical models and Bayesian inference. Their iBAG algorithm (integrated Bayesian Analysis) combines miRNA, DNA methylation and mRNA data to predict patient survival of Glioblastoma Multiforme (GBM) patients. Their approach explicitly takes into account the biological relationship between different data modalities. The authors identified separate gene sets related to disease outcome and demonstrated better prediction power to detect disease related genes than non-integrative methods.

Gade et al. first constructed a correlation weighted bipartite miRNA-target gene graph [94]. This graph was then used to guide feature selection with a component-wise likelihood boosting algorithm for predicting prostate cancer outcome [95]. Going one step further other authors also considered protein-protein interaction information [96]. Their method first smoothes marginal t-statistics of genes and miRNAs over the structure of the integrated PPI and miRNA-target gene network via random walk kernels. Most relevant features are then determined via a permutation test. Subsequently a conventional SVM classifier is trained. The authors demonstrated the benefit of this approach compared to stacking for predicting disease prognosis in several cancers.

Arguably one of the most advanced but also computationally costly approaches for intermediate data integration in the field of personalized medicine has recently been suggested by Zitnik and Zupan [97]. The authors combined gene expression and histological data from animals and human with protein-protein interactions and GO annotation to predict liver injury induced by chemicals. This was done based on a constrained matrix tri-factorization algorithm suggested by the same authors [98].

Vliet et al. made a comparison of several integration strategies (pasting of feature matrices, linear combination of distance measures or kernel functions, stacking) and classifiers to predict breast cancer outcome [99]. The authors reported most success via an intermediate strategy using a nearest mean classifier our via a late strategy using a logical OR function.

### Unsupervised Patient Subgroup Detection

Apart from supervised patient stratification using defined clinical endpoints (e.g. survival times), a lot of effort has been made to detect patient sub-populations in a completely unsupervised manner based on molecular data. An example of this approach is the detection of four different molecular subtypes of Glioblastoma Multiforme (GBM) patients based on gene expression data by Verhaak et al. [3]. As more molecular data modalities from the same patient become available now, many authors explored the possibility of fusing these data for discovering stronger patterns (see [100] for a review)

Akin to the case of supervised learning for patient stratification, unsupervised data fusion approaches can be broadly classified into three groups, which involve early, late and intermediate integration schemes. Early integration methods work with a joint feature matrix and modify traditional clustering algorithms, such as k-means, to calculate a weight for each data source [101]. Late integration combines patient similarity matrices obtained from independent clusterings of distinct data types. Intermediate integration methods typically aim for extracting common features from different data modalities combined with clustering of patients.

An example of an intermediate integration strategy is non-negative matrix factorization (NMF) [102]. The idea behind NMF is to factorize a data matrix into a product of two matrices, one indicating discriminative feature combinations between clusters and one indicating cluster assignments of patients. While originally NMF was designed to work with one data modality only, later work has extended the approach to simultaneous clustering of several data types. For example, Zhang et al. used an extended NMF framework to cluster 385 ovarian cancer patients based on joint gene expression, DNA methylation and miRNA profiles [103].

Another popular intermediate integration approach is the iCluster method by Shen et al. [104, 105]. This technique combines ideas from sparse matrix decomposition and latent factor models and has also remarkable similarities to probabilistic PCA [106] and k-means [107]. Furthermore, the iCluster method can be seen as a special case of Bayesian canonical correlation analysis with a sparsity prior for the coefficient matrix [108], facilitating model identifiability and interpretability. In [105] the authors used iCluster to integrate gene expression, DNA methylation as well as CNV data of Glioblastoma Multiforme (GBM) patients. The iCluster method treats information from all patients with the same confidence, which may lead to erroneous results, if there are patients with dis-concordant information from different omics data modalities. The latter issue was taken up by Yuan et al. [109], who developed a Patient Specific Data Fusion (PSDF) model,

which gives different patients separate weights within a non-parametric Bayesian framework. A unique aspect of PSDF is that it allows for the separation of concordant and dis-concordant signals from patients and unlike iCluster does not force patients to cluster together. The obtained disease subtypes via PSDF were reported to be prognostically relevant by the authors. A limitation of the PSDF method is in the required data discretization, which may lead to considerable loss of information. Similar to the PSDF method Kormaksson et al. [110] proposed a mixture-model for integrative clustering of gene expression and DNA methylation data. Unlike PSDF, the method does not require data discretization. However, a limitation is the assumption of statistical independence of molecular features.

Another recent mixture model approach is the MDI (Multiple Data Integration) approach by Kirk et al. [111] and Savage et al. [112]. Following a Bayesian non-parametric clustering approach MDI assumes a Dirichlet Process Prior over cluster assignments. Moreover, and in contrast to PSDF, MDI learns exact dependencies between the different data sources as a directed acyclic graph. This implicitly results in a preference to put patients into the same cluster, if they tend to group together in each of the different data sources. However, at the same time each data source still retains its own clustering, reflecting the fact that different molecular data may express partially non-concordant patient groupings. Savage et al. [112] used the MDI model to integrate genomic, epigenomic and transcriptomic information of GBM patients and reported clinically relevant disease sub-types. MDI is flexible in modelling continuous (e.g. gene expression) as well as discrete (e.g. CNVs) data. A limitation is the assumption of statistical independence of molecular features.

Generative modelling approach, such as MDI and PSDF, require to express explicitly the joint statistical distribution over different data modalities. This complication is avoided in late integration techniques. Examples are Similarity Network Fusion (SNF) [113] and Multiview Genomic Data Integration (MVDA) [114]. These techniques use independent clustering algorithms for each data modality and aggregate results of patient similarity matrices from each data source. Thus, late integration potentially allows for incorporating thousands of features for each data modality. Furthermore, late integration techniques are typically more robust to small sample sizes. A limitation is the difficulty to explicitly model dependencies between data modalities. The SNF method models patient similarities as networks with nodes representing patients. Each data modality generates its own network, and these networks are then fused into a consensus network using a message-passing algorithm. The authors in this way integrated gene expression, DNA methylation as well as miRNA profiles over five cancer datasets and performed graph clustering on the consensus network to identify disease subtypes. The MVDA approach [114] concatenates patient-patient similarity matrices obtained from different data sources and then uses matrix factorization of the concatenated matrix to come up with a consensus clustering.

Biclustering is yet another popular statistical technique for simultaneous clustering of the rows and columns of a data matrix and has recently also been employed for data fusion. The original method along with its modifications has since many years found several applications in biological data analysis (see [115] for a comprehensive review). Recently, Bunte et al. [116] developed a novel bi-clustering algorithm to cluster cancer cell lines treated with different drugs while including CNV, DNA methylation, mRNA, protein abundance and exome sequencing information. The model is based on the previous work of the same group of authors on the Group Factor Analysis Model [117]. Another technique based on biclustering has been proposed by Sun et al [118, 119]. Their method is based on sparse singular value decomposition (SSVD) and was applied to combine SNP information with clinical data for disease subtyping and identification of subtype-specific genotype variations.

## CONCLUSION

Fusion of heterogeneous omics data modalities is widely believed to improve our understanding of biological systems and to enable better personalized medicine. However, statistical data integration is associated with non-trivial challenges resulting from differing numerical and statistical properties of individual omics data modalities. These challenges come in addition to all the well known issues with individual omics data, namely high dimensionality at low sample size and high noise level. As multi-omics studies become more and more common practice, there is a growing need for appropriate statistical learning and inference techniques. The goal of this review is to shed light on the current state of methodology in the field. We are aware of the fact that our review is limited at this point. Techniques not covered here include e.g. ODE based models in systems biology [120]. Moreover, we restricted ourselves to the question of integrating multiple -omics data modalities. Hence, we did not address genomic data fusion approaches, which have been e.g. applied in the context of disease gene prioritization [121].

Multi-omics data are believed to reflect more information about a biological sub-system than a single data modality. Hence, a considerable number of approaches focus on learning and modelling dependencies between molecular features of different modalities. Most techniques within that family follow a knowledge driven approach, which employs biological networks as a backbone. These methods integrate data and biological knowledge in a far better and more consistent way than purely sequential analysis approaches. However, despite the success of methods such as PARADIGM the knowledge driven framework is limited by the incompleteness of current biological knowledge. Moreover, biological networks are in principle cell type and biological condition dependent, which is often ignored in practice. On the other hand methods within a BN learning scheme have been developed, such as extensions of the module network algorithm, which allow for a greater flexibility

| Method | Modelling Approach | Input | Output | Assumptions | Advantages | Limitations |
|---|---|---|---|---|---|---|
| Chari et al. [19] | Sequential Analysis | CNV, LOH, methylation, mRNA | disease genes and pathways | independent analysis of different data modalities yields biologically consistent results | conceptually simple; computationally cheap | treats data sources as independent |
| Sun et al. [18] | Sequential Analysis | CNV, methylation, mRNA | disease genes | independent analysis of different data modalities yields biologically consistent results | conceptually simple; computationally cheap | treats data sources as independent |
| Wandaliz et al. [29] | Random Forests | mRNA, protein concentrations | abundance of undetected proteins | gene expression can explain a relevant fraction of the variance in protein expression | feature selection; computationally moderate | limited to specific biological question |
| Waajinborg et al. [34] | Penalized CCA | mRNA, CN | disease genes | biologically relevant information exists in a linear subspace of the data | flexible and extend-able framework | number of latent variables needs to be determined; computationally costly |
| Le Cao et al. [37] | Sparse PLS, sparse CCA | mRNA, metabolites | disease genes and pathways | biologically relevant information exists in a linear subspace of the data | flexible and extend-able framework | number of latent variables needs to be determined; computationally costly |
| CONEXIC, Akavia et al. [57] | Bayesian Network (Module Network) | CNV, mRNA | cancer drivers | model consistent with biological data and at least partially identifiable | probabilistic model | limited to specific biological question; computationally costly; true model typically not fully identifiable |
| LeMoNe, Bonnet et al. [59] | Bayesian Network (Module Network) | mRNA, miRNA, Clinical data | gene regulatory network | model consistent with biological data and at least partially identifiable | integrates many information sources | limited to specific biological question; computationally costly; true model typically not fully identifiable |
| BiRTe , Fröhlich [61] | Bayesian Network | mRNA, miRNA | context-specific gene regulatory network | model consistent with measured data and at least partially identifiable | combines inference of active transcriptional regulators with regulatory network learning | limited to specific biological question; computationally costly |
| Nibbe et al. [70] | graph based (modified Page Rank) | PPI network, mRNA, protein expression | functional disease network | PPI network largely consistent with reality | works with large scale networks and data | relies on quality of PPI network; only mRNA + protein expression |
| Tyekucheva et al. [71] | statistical meta-analysis | CNV, mRNA, gene sets (pathways) | phenotype related pathways | similar statistical power across all data modalities | flexible and extend-able framework | relies on predefined pathways |
| PARADIGM, Vaske et al. [74] | graphical model (Markov Random Field) | pathways, mRNA, CNV | pathway activities | defined pathway activity score captures relevant biological information; pathway structure largely in agreement with biological reality | flexible and extend-able framework; probabilistic approach | relies on predefined pathways |
| pWOMICS, Wachter et al. [76] | graph based | pathways, TF-targets, PPI network, proteomics, mRNA | molecular network | existing biological knowledge largely consistent with reality | combines and integrates existing biological knowledge | limited to protein and gene expression data; relies on quality of PPI networks and TF-targets |

**Table 3:** Selected Statistical Techniques for Learning Feature Dependencies from Multiple Data Sources

| Method | Objective | modelling Approach | Input | Output | Assumptions | Advantages | Limitations |
|---|---|---|---|---|---|---|---|
| Daemen et al. [91] | supervised clinical outcome prediction | multiple kernel learning | mRNA, CNV, clinical data | clinical outcome | linear kernel combination can enhance prediction performance | flexible and extend-able framework | computationally costly |
| iBAG, Wang et al. [93] | supervised clinical outcome prediction | graphical model | miRNA, mRNA, methylation | patient survival | model consistent with biological data and at least partially identifiable | fully probabilistic approach | framework not easy to extend; computationally costly |
| Gade et al. [94] | supervised clinical outcome prediction | correlation, statistical meta-analysis, boosting | miRNA, mRNA | patient survival | miRNA-target predictions largely consistent with biological reality | conceptually simple | framework not easy to extend; computationally costly |
| Zitnik et al. [97] | supervised clinical outcome prediction | matrix factorization | miRNA, PPI, GO annotation, histological data | chemical induced liver injury | biologically relevant information can be extracted from linear subspace of the data | flexible and extend-able framework, can integrate various types of information | relies on relations between biological entities (e.g. GO terms and genes), computationally costly |
| Zhang et al. [103] | unsupervised disease subgroup identification | matrix factorization | mRNA, miRNA, methylation | disease subtypes | biologically relevant information can be extracted from linear subspace of the data | flexible and extend-able framework | same influence of each data source |
| iCLUSTER, Shen et al. [104, 105] | unsupervised disease subgroup identification | matrix factorization | mRNA, miRNA, methylation | disease subtypes | biologically relevant information can be extracted from linear subspace of the data | flexible and extend-able framework | same influence of each data source |
| PSDF, Yuan et al. [109] | unsupervised disease subgroup identification | Bayesian non-parametric (Dirichlet process mixture model) | mRNA, CNV | disease subtypes | model consistent with biological data and at least partially identifiable | fully probabilistic, flexible and extend-able framework | data discretization, computationally costly |
| MDI, Kirk et al. [111, 112] | unsupervised disease subgroup identification | Bayesian non-parametric (Dirichlet process mixture model) | mRNA, DNA methylation, CNV | disease subtypes | model consistent with biological data and at least partially identifiable | fully probabilistic, flexible and extend-able framework | assumes statistical feature independence; computationally costly |
| SNF [113] | unsupervised disease subgroup identification | patient similarity, message passing | mRNA, miRNA, DNA methylation | disease subtypes | disease subgroups can be identified from thresholded patient-patient similarities defined for individual data modalities | flexible and extend-able framework; can be applied to large number of features | neglects biological dependencies between data modalities |

**Table 4:** Selected Statistical Learning Techniques for Personalized Medicine using Multiple Data Sources

at this point, but are confronted with the non-trivial challenges (including non-identifiability) of network structure learning. Correlation and covariance based methods, such as sparse CCA, make considerably less assumptions than knowledge driven methods, while being at the same time significantly less computationally demanding than BN based approaches. Moreover, these methods do not face identifiability problems up to the same degree than BN based approaches. However, interpretation of feature combinations extracted by sparse CCA is typically more difficult and may require in addition biological networks in a secondary analysis step, resulting in similar limitations than mentioned above for knowledge driven approaches.

In the area of personalized medicine two goals of multi-omics data integration are better supervised prediction of clinical outcomes and better unsupervised identification of so far unknown patient sub-populations. For both types of machine learning tasks available approaches can be categorized as early, intermediate or late phase integration. The majority of methods follow the intermediate phase integration scheme, because it is believed that in this way most of the dependencies between different data modalities can be captured. Specifically in the unsupervised setting many approaches follow a probabilistic modelling scheme, whereas for supervised clinical outcome prediction the picture of applied techniques is more diverse. Arguably one of the most advanced methods in the field is the matrix tri-factorization method by Zitnik and Zupan [97], which makes use of known relationships (physical, functional, semantic, ...) between molecular entities, patients and diseases.

Of course, each of the methods in the two above discussed application domains of multi-omics data integration makes specific assumptions, which are often difficult to verify or falsify in practice, partially due to the high dimensionality of the data. Moreover, apart from supervised machine techniques, for most of the above mentioned models there is no clear and objective performance metric for assessing their actual success. Typically, authors thus rely on simulations and biological interpretation of results on real data to validate their methods. Altogether the choice of an appropriate integration approach should thus depend on different factors, such as the amount and quality of biological background knowledge for the particular research question and the amount and quality of available data. Fewer data with lower quality will generally require less complex models than high quality datasets with larger sample size. Furthermore, the level of expertise of the modeler is in practice a non-negligible factor [122].

A general issue with all multi-omics approaches, which is specifically true in the personalized medicine area, is that for one and the same biological sample not always all omics data types have been systematically measured. For example, in TCGA for many patients and cancer types gene expression data is not always matching with somatic mutations and DNA methylation. Since most current integration strategies focus only on those samples, for which all data modalities are available, there is a loss of information. Hence, new integration methods, possibly utilizing ensemble learning techniques, should focus on reducing this loss of information, while at the same time appropriately handling biases resulting from unequally balanced data types.

The above mentioned aspect may be viewed in the light of the current Big Data discussion: While there is on one hand an evident data explosion in modern biology and medicine – think e.g. about data volume for one whole genome patient sequencing – on the other hand in many cases we still lack the data to cover a single relevant biological phenomenon up to sufficient level. This paradoxical situation in modern biology and medicine can be partially attributed to the enormous complexity of biological systems coupled with still existing limitations of measurement techniques and costs. Future developments in biotechnology may help to overcome some of these limitations and allow for obtaining a more and more comprehensive picture of biology. It is likely that these developments in turn will increase the relevance of heterogeneous data integration methods.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## REFERENCES


1. Hawkins RD, Hon GC, Ren B. **Next-generation genomics: an integrative approach**. Nature Reviews Genetics. 2011 Jan;doi:10.1038/nrg2795.
2. Hudson (Chairperson) TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. **International network of cancer genome projects**. Nature. 2010 Apr;464(7291):993–998. doi:10.1038/nature08987.
3. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, M Mastrogianakis G, et al. **Comprehensive genomic characterization defines human glioblastoma genes and core pathways**. Nature. 2008 Oct;455(7216):1061–1068. doi:10.1038/nature07385.
4. Brem RB, Yvert G, Clinton R, Kruglyak L. **Genetic dissection of transcriptional regulation in budding yeast**. Science (New York, NY). 2002 Apr;296(5568):752–755. doi:10.1126/science.1069516.
5. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, et al. **Genetics of gene expression surveyed in maize, mouse and man**. Nature. 2003 Mar;422(6929):297–302. doi:10.1038/nature01434.
6. Joyce AR, Palsson BØ. **The model organism as a system: integrating 'omics' data sets**. Nature Reviews Molecular Cell Biology. 2006 Mar;7(3):198–210. doi:10.1038/nrm1857.
7. Choi H, Pavelka N. **When One and One Gives More than Two: Challenges and Opportunities of Integrative Omics**. Frontiers in Genetics. 2012 Jan;2. doi:10.3389/fgene.2011.00105.
8. Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale AL. **Principles and methods of integrative genomic analyses in cancer**. Nature Reviews Cancer. 2014 May;14(5):299–313. doi:10.1038/nrc3721.
9. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. **Methods of integrating data to uncover genotype-phenotype interactions.** Nat Rev Genet. 2015 Feb;16(2):85–97. doi:10.1038/nrg3868.
10. Alyass A, Turcotte M, Meyre D. **From big data analysis to personalized medicine for all: challenges and opportunities.** BMC Med Genomics. 2015;8:33. doi:10.1186/s12920-015-0108-y.
11. Gligorijević V, Pržulj N. **Methods for biological data integration: perspectives and challenges.** J R Soc Interface. 2015 Nov;12(112). doi:10.1098/rsif.2015.0571.

12. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. **Dimension reduction techniques for the integrative analysis of multi-omics data.** Brief Bioinform. 2016 Jul;17(4):628–641. doi:10.1093/bib/bbv108.
13. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. **Data integration in the era of omics: current and future challenges.** BMC Syst Biol. 2014;8 Suppl 2:I1. doi:10.1186/1752-0509-8-S2-I1.
14. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. **Methods for the integration of multi-omics data: mathematical aspects.** BMC Bioinformatics. 2016;17 Suppl 2:15. doi:10.1186/s12859-015-0857-9.
15. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.** Cancer Cell. 2010 Jan;17(1):98–110. doi:10.1016/j.ccr.2009.12.020.
16. Hyman E, Kauraniemi P, Hautaniemi S, Wolf M, Mousses S, Rozenblum E, et al. **Impact of DNA amplification on gene expression patterns in breast cancer**. Cancer Research. 2002 Nov;62(21):6240–6245.
17. Bicciato S, Spinelli R, Zampieri M, Mangano E, Ferrari F, Beltrame L, et al. **A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets**. Nucleic Acids Research. 2009 Aug;37(15):5057–5070. doi:10.1093/nar/gkp520.
18. Sun Z, Asmann YW, Kalari KR, Bot B, Eckel-Passow JE, Baker TR, et al. **Integrated Analysis of Gene Expression, CpG Island Methylation, and Gene Copy Number in Breast Cancer Cells by Deep Sequencing**. PLoS ONE. 2011 Feb;6(2):e17490. doi:10.1371/journal.pone.0017490.
19. Chari R, Coe BP, Vucic EA, Lockwood WW, Lam WL. **An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer**. BMC Systems Biology. 2010;4(1):67. doi:10.1186/1752-0509-4-67.
20. Lipson D, Ben-Dor A, Dehan E, Yakhini Z. **Joint Analysis of DNA Copy Numbers and Gene Expression Levels**. Springer Berlin Heidelberg; 2004. p. 135–146. doi:10.1007/978-3-540-30219-3_12.
21. Ortiz-Estevez M, De Las Rivas J, Fontanillo C, Rubio A. **Segmentation of genomic and transcriptomic microarrays data reveals major correlation between DNA copy number aberrations and gene-loci expression**. Genomics. 2011 Feb;97(2):86–93. doi:10.1016/j.ygeno.2010.10.008.
22. Nunez-Iglesias J, Liu CC, Morgan TE, Finch CE, Zhou XJ. **Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer's disease cortex reveals altered miRNA regulation.** PLoS One. 2010;5(2):e8898. doi:10.1371/journal.pone.0008898.
23. Kendziorski CM, Chen M, Yuan M, Lan H, Attie AD. **Statistical methods for expression quantitative trait loci (eQTL) mapping**. Biometrics. 2006 Mar;62(1):19–27. doi:10.1111/j.1541-0420.2005.00437.x.
24. Shabalin AA. **Matrix eQTL: ultra fast eQTL analysis via large matrix operations**. Bioinformatics. 2012 May;28(10):1353–1358. doi:10.1093/bioinformatics/bts163.
25. Beck D, Ayers S, Wen J, Brandl MB, Pham TD, Webb P, et al. **Integrative analysis of next generation sequencing for small non-coding RNAs and transcriptional regulation in Myelodysplastic Syndromes**. BMC Medical Genomics. 2011 Feb;4:19. doi:10.1186/1755-8794-4-19.
26. Muniategui A, Nogales-Cadenas R, Vázquez M, L Aranguren X, Agirre X, Luttun A, et al. **Quantification of miRNA-mRNA Interactions**. PLoS ONE. 2012 Feb;7(2):e30766. doi:10.1371/journal.pone.0030766.
27. Setty M, Helmy K, Khan AA, Silber J, Arvey A, Neezen F, et al. **Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma.** Mol Syst Biol. 2012;8:605. doi:10.1038/msb.2012.37.
28. Le HS, Bar-Joseph Z. **Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation**. Bioinformatics. 2013 Jan;29(13):i89–i97. doi:10.1093/bioinformatics/btt231.
29. Torres-García W, Zhang W, Runger GC, Johnson RH, Meldrum DR. **Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins**. Bioinformatics (Oxford, England). 2009 Aug;25(15):1905–1914. doi:10.1093/bioinformatics/btp325.
30. Michaelson JJ, Alberts R, Schughart K, Beyer A. **Data-driven assessment of eQTL mapping methods**. BMC Genomics. 2010;11:502. doi:10.1186/1471-2164-11-502.
31. Hotelling H. **Relations between two sets of variables**. Biometrika. 1936;28:321–377. doi:10.2307/2333955.
32. Witten DM, Tibshirani RJ. **Extensions of sparse canonical correlation analysis with applications to genomic data**. Statistical applications in genetics and molecular biology. 2009;8(1):1–27. doi:10.2202/1544-6115.1470.
33. Chalise P, Fridley BL. **Comparison of penalty functions for sparse canonical correlation analysis**. Computational statistics & data analysis. 2012;56(2):245–254. doi:10.1016/j.csda.2011.07.012.
34. Waaijenborg S, Verselewel de Witt Hamer PC, Zwinderman AH. **Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis**. Statistical Applications in Genetics and Molecular Biology. 2008;7(1). doi:10.2202/1544-6115.1329.
35. Lin D, Zhang J, Li J, Calhoun VD, Deng HW, Wang YP. **Group sparse canonical correlation analysis for genomic data integration**. BMC Bioinformatics. 2013 Aug;14:245. doi:10.1186/1471-2105-14-245.
36. Lin D, Calhoun VD, Wang YP. **Correspondence between fMRI and SNP data by group sparse canonical correlation analysis**. Medical image analysis. 2014;18(6):891–902. doi:10.1016/j.media.2013.10.010.
37. Lê Cao KA, Rossouw D, Robert-Granié C, Besse P. **A sparse PLS for variable selection when integrating omics data**. Statistical applications in genetics and molecular biology. 2008;7(1). doi:10.2202/1544-6115.1390.
38. Le Floch É, Guillemot V, Frouin V, Pinel P, Lalanne C, Trinchera L, et al. **Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares**. NeuroImage. 2012 Oct;63(1):11–24. doi:10.1016/j.neuroimage.2012.06.061.
39. Lê Cao KA, Martin PG, Robert-Granié C, Besse P. **Sparse canonical methods for biological data integration: application to a cross-platform study**. BMC bioinformatics. 2009;10(1):34. doi:10.1186/1471-2105-10-34.
40. Sheng J, Deng HW, Calhoun VD, Wang YP. **Integrated analysis of gene expression and copy number data on gene shaving using independent component analysis**. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM. 2011 Nov-Dec;8(6):1568–1579. doi:10.1109/TCBB.2011.71.
41. Tomescu OA, Mattanovich D, Thallinger GG. **Integrative omics analysis. A study based on Plasmodium falciparum mRNA and protein data**. BMC Systems Biology. 2014 Mar;8(Suppl 2):S4. doi:10.1186/1752-0509-8-S2-S4.
42. Fagan A, Culhane AC, Higgins DG. **A multivariate analysis approach to the integration of proteomic and gene expression data**. PROTEOMICS. 2007 Jun;7(13):2162–2171. doi:10.1002/pmic.200600898.
43. Ray P, Zheng L, Lucas J, Carin L. **Bayesian Joint Analysis of Heterogeneous Genomics Data**. Bioinformatics. 2014 Jan;p. btu064. doi:10.1093/bioinformatics/btu064.
44. Reverter F, Vegas E, Oller JM. **Kernel-PCA data integration with enhanced interpretability**. BMC Systems Biology. 2014 Mar;8(Suppl 2):S6. doi:10.1186/1752-0509-8-S2-S6.
45. Koller D, Friedman N. Probabilistic Graphical Models: Principles and Technique. MIT Press; 2009.
46. Fröhlich H, Bahamondez G, Götschel F, Korf U. **Dynamic Bayesian Network Modeling of the Interplay between EGFR and Hedgehog Signaling**. PLoS ONE. 2015 Nov;10(11):e0142646. doi:10.1371/journal.pone.0142646.
47. Imoto, S H, T , Goto T, Tashiro K, Kuhara S, Miyano S. **Combining Microarrays and Biological Knowledge for Estimating Gene Networks via Bayesian Networks**. In: Proc. 2nd Computational Systems Bioinformatics; 2003. p. 104–113. doi:10.1109/csb.2003.1227309.
48. Zheng J, Chaturvedi I, Rajapakse JC. **Integration of Epigenetic Data in Bayesian Network Modeling of Gene Regulatory Network**. In: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al., editors. Pattern Recognition in Bioinformatics. vol. 7036. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 87–96.

49. Praveen P, Fröhlich H. **Boosting probabilistic graphical model inference by incorporating prior knowledge from multiple sources**. PloS one. 2013;8(6):e67410. doi:10.1371/journal.pone.0067410.
50. Huttenhower C, Troyanskaya OG. **Bayesian data integration: a functional perspective**. Comput Syst Bioinformatics Conf. 2006;p. 341–351.
51. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, Coller HA, et al. **Exploring the human genome with functional maps**. Genome Research. 2009 Jun;19(6):1093–1106. doi:10.1101/gr.082214.108.
52. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** Nat Genet. 2003 Jun;34(2):166–176. doi:10.1038/ng1165.
53. Segal E, Pe'er D, Regev A, Koller D, Friedman N. **Learning module networks**. In: Advances in Neural Information Processing Systems. vol. 578; 2004. p. 297–304.
54. Joshi A, Van de Peer Y, Michoel T. **Analysis of a Gibbs sampler method for model-based clustering of gene expression data**. Bioinformatics (Oxford, England). 2008 Jan;24(2):176–183. doi:10.1093/bioinformatics/btm562.
55. Joshi A, Smet RD, Marchal K, de Peer YV, Michoel T. **Module networks revisited: computational assessment and prioritization of model predictions**. Bioinformatics. 2009 Feb;25(4):490–496. doi:10.1093/bioinformatics/btn658.
56. Zhang W, Zhu J, Schadt EE, Liu JS. **A Bayesian Partition Method for Detecting Pleiotropic and Epistatic eQTL Modules**. PLoS Comput Biol. 2010 Jan;6(1):e1000642. doi:10.1371/journal.pcbi.1000642.
57. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, et al. **An Integrated Approach to Uncover Drivers of Cancer**. Cell. 2010 Dec;143(6):1005–1017. doi:10.1016/j.cell.2010.11.013.
58. Bonnet E, Michoel T, de Peer YV. **Prediction of a gene regulatory network linked to prostate cancer from gene expression, microRNA and clinical data.** Bioinformatics. 2010 Sep;26(18):i638–i644. doi:10.1093/bioinformatics/btq395.
59. Bonnet E, Calzone L, Michoel T. **Integrative Multi-omics Module Network Inference with Lemon-Tree**. PLoS Computational Biology. 2015 Feb;11(2). doi:10.1371/journal.pcbi.1003983.
60. Zacher B, Abnaof K, Gade S, Younesi E, Tresch A, Fröhlich H. **Joint Bayesian inference of condition-specific miRNA and transcription factor activities from combined gene and microRNA expression data.** Bioinformatics. 2012 Jul;28(13):1714–1720. doi:10.1093/bioinformatics/bts257.
61. Fröhlich H. **biRte: Bayesian inference of context-specific regulator activities and transcriptional networks**. Bioinformatics. 2015 Jun;p. btv379. doi:10.1093/bioinformatics/btv379.
62. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. **KEGG for linking genomes to life and the environment**. Nucleic Acids Res. 2008;36:480–484. doi:10.1093/nar/gkm882.
63. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. **Pathway Commons, a web resource for biological pathway data.** Nucleic Acids Res. 2011 Jan;39(Database issue):D685–D690. doi:10.1093/nar/gkq1039.
64. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, et al. **The BioGRID interaction database: 2013 update.** Nucleic Acids Res. 2013 Jan;41(Database issue):D816–D823. doi:10.1093/nar/gks1158.
65. Vejnar CE, Zdobnov EM. **MiRmap: comprehensive prediction of microRNA target repression strength.** Nucleic Acids Res. 2012 Dec;40(22):11673–11683. doi:10.1093/nar/gks901.
66. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. **Human MicroRNA targets**. PLoS Biol. 2004 Nov;2(11):e363. doi:10.1371/journal.pbio.0020363.
67. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. **miRBase: tools for microRNA genomics.** Nucleic Acids Res. 2008 Jan;36(Database issue):D154–D158. doi:10.1093/nar/gkm952.
68. Ideker T, Ozier O, Schwikowski B, Siegel AF. **Discovering regulatory and signalling circuits in molecular interaction networks**. Bioinformatics (Oxford, England). 2002;18 Suppl 1:S233–240.
69. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Müller T. **Identifying functional modules in protein-protein interaction networks: an integrated exact approach.** Bioinformatics (Oxford, England). 2008 Jul;24(13):i223–31. doi:10.1093/bioinformatics/btn161.
70. Nibbe RK, Koyutürk M, Chance MR. **An Integrative -omics Approach to Identify Functional Sub-Networks in Human Colorectal Cancer**. PLoS Comput Biol. 2010 Jan;6(1):e1000639. doi:10.1371/journal.pcbi.1000639.
71. Tyekucheva S, Marchionni L, Karchin R, Parmigiani G. **Integrating diverse genomic data using gene sets**. Genome Biol. 2011;12(10):R105. doi:10.1186/gb-2011-12-10-r105.
72. Sun H, Wang H, Zhu R, Tang K, Gong Q, Cui J, et al. **iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis**. Bioinformatics (Oxford, England). 2014 Mar;30(5):737–739. doi:10.1093/bioinformatics/btt576.
73. Kamburov A, Cavill R, Ebbels TMD, Herwig R, Keun HC. **Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA**. Bioinformatics. 2011 Oct;27(20):2917–2918. doi:10.1093/bioinformatics/btr499.
74. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. **Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM**. Bioinformatics. 2010 Jun;26(12):i237–i245. doi:10.1093/bioinformatics/btq182.
75. Kristensen VN, Vaske CJ, Ursini-Siegel J, Van Loo P, Nordgard SH, Sachidanandam R, et al. **Integrated molecular profiles of invasive breast tumors and ductal carcinoma in situ (DCIS) reveal differential vascular and interleukin signaling**. Proceedings of the National Academy of Sciences of the United States of America. 2012 Feb;109(8):2802–2807. doi:10.1073/pnas.1108781108.
76. Wachter A, Beissbarth T. **pwOmics: An R package for pathway-based integration of time-series omics data using public database knowledge**. Bioinformatics. 2015;p. btv323. doi:10.1093/bioinformatics/btv323.
77. Takahashi H, Matsuyama A. **An approximate solution for the Steiner problem in graphs**. Math Jap. 1980;24:573 – 577.
78. Sadeghi A, Fröhlich H. **Steiner tree methods for optimal sub-network identification: an empirical study**. BMC Bioinformatics. 2013;14:144. doi:10.1186/1471-2105-14-144.
79. Rau A, Jaffrézic F, Foulley JL, Doerge RW. **An Empirical Bayesian Method for Estimating Biological Networks from Temporal Microarray Data**. Statistical Applications in Genetics and Molecular Biology. 2010;9(1). doi:10.2202/1544-6115.1513.
80. Cun Y, Fröhlich H. **Biomarker Gene Signature Discovery Integrating Network Knowledge**. Biology. 2012 Feb;1(1):5–17. doi:10.3390/biology1010005.
81. Pavlidis P, Weston J, Cai J, Grundy W. **Gene functional classification from heteregoneous data**. In: Proc. 5th Int. Conf. Computational Molecular Biology; 2001. p. 242–248. doi:10.1145/369133.369228.
82. Maragos P, Gros P, Katsamanis A, Papandreou G. **Cross-Modal Integration for Performance Improving in Multimedia: A Review**. In: Maragos P, Potamianos A, Gros P, editors. Multimodal Processing and Interaction. Boston, MA: Springer US; 2008. p. 1–46.
83. Wolpert DH. **Stacked Generalization**. Neural Networks. 1992;5:241 – 259. doi:10.1016/s0893-6080(05)80023-1.
84. Lanckriet G, Cristianini N, Bartlett P, Ghaoui LE, Jordan M. **Learning the Kernel Matrix with Semidefinite Programming**. J Machine Learning Research. 2004;5:27–72.
85. Pittman J, Huang E, Dressman H, Horng CF, Cheng SH, Tsou MH, et al. **Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes**. Proceedings of the National Academy of Sciences of the United States of America. 2004 Jan;101(22):8431–8436. doi:10.1073/pnas.0401736101.
86. Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, et al. **Gene expression phenotypic models that predict the activity of oncogenic pathways**. Nature Genetics. 2003 Jun;34(2):226–230. doi:10.1038/ng1167.
87. Boulesteix AL, Porzelius C, Daumer M. **Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value**. Bioinformatics. 2008 Jan;24(15):1698–1706. doi:10.1093/bioinformatics/btn262.
88. Lê Cao KA, Meugnier E, McLachlan GJ. **Integrative mixture of experts to combine clinical factors and gene markers**. Bioinformatics. 2010;26(9):1192–1198. doi:10.1093/bioinformatics/btq107.

89. Gevaert O, Smet FD, Timmerman D, Moreau Y, Moor BD. **Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks**. Bioinformatics. 2006 Jul;22(14):e184–e190. doi:10.1093/bioinformatics/btl230.
90. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. **Gene expression profiling predicts clinical outcome of breast cancer.** Nature. 2002 Jan;415(6871):530–536. doi:10.1038/415530a.
91. Daemen A, Gevaert O, Ojeda F, Debucquoy A, Suykens JA, Sempoux C, et al. **A kernel-based integration of genome-wide data for clinical decision support**. Genome Medicine. 2009 Apr;1(4):39. doi:10.1186/gm39.
92. Thomas M, Brabanter KD, Suykens JA, Moor BD. **Predicting breast cancer using an expression values weighted clinical classifier**. BMC Bioinformatics. 2014;15:411. doi:10.1186/s12859-014-0411-1.
93. Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, Do KA. **iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data**. Bioinformatics. 2013;29(2):149–159. doi:10.1093/bioinformatics/bts655.
94. Gade S, Porzelius C, Faelth M, Brase J, Wuttig D, Kuner R, et al. **Graph based fusion of miRNA and mRNA expression data improves clinical outcome prediction in prostate cancer**. BMC Bioinformatics. 2011;12(1):488. doi:10.1186/1471-2105-12-488.
95. Binder H, Schumacher M. **Incorporating pathway information into boosting estimation of high-dimensional risk prediction models.** BMC Bioinformatics. 2009;10:18. doi:10.1186/1471-2105-10-18.
96. Cun Y, Fröhlich H. **Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics**. PLoS ONE. 2013 Sep;8(9):e73074. doi:10.1371/journal.pone.0073074.
97. Zitnik M, Zupan B. **Survival regression by data fusion**. Systems Biomedicine. 2014;2(3):49–55. doi:10.1080/21628130.2015.1016702.
98. Zitnik M, Zupan B. **Data Fusion by Matrix Factorization**. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015 Jan;37(1):41–53. doi:10.1109/TPAMI.2014.2343973.
99. van Vliet MH, Horlings HM, van de Vijver MJ, Reinders MJT, Wessels LFA. **Integration of Clinical and Gene Expression Data Has a Synergetic Effect on Predicting Breast Cancer Outcome**. PLoS ONE. 2012 Jul;7(7):e40358. doi:10.1371/journal.pone.0040358.
100. Chalise P, Koestler DC, Bimali M, Yu Q, Fridley BL. **Integrative clustering methods for high-dimensional molecular data**. Translational cancer research. 2014;3(3):202.
101. Chen X, Xu X, Huang JZ, Ye Y. **TW-k-means: automated two-level variable weighting clustering algorithm for multiview data**. Knowledge and Data Engineering, IEEE Transactions on. 2013;25(4):932–944. doi:10.1109/tkde.2011.262.
102. Lee DD, Seung HS. **Algorithms for non-negative matrix factorization**. In: Advances in neural information processing systems; 2001. p. 556–562.
103. Zhang S, Liu CC, Li W, Shen H, Laird PW, Zhou XJ. **Discovery of multi-dimensional modules by integrative analysis of cancer genomic data**. Nucleic acids research. 2012;p. gks725. doi:10.1093/nar/gks725.
104. Shen R, Olshen AB, Ladanyi M. **Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis**. Bioinformatics. 2009;25(22):2906–2912. doi:10.1093/bioinformatics/btp659.
105. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, et al. **Integrative subtype discovery in glioblastoma using iCluster**. PloS one. 2012;7(4):e35236. doi:10.1371/journal.pone.0035236.
106. Tipping ME, Bishop CM. **Probabilistic principal component analysis**. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1999;61(3):611–622. doi:10.1111/1467-9868.00196.
107. Ding C, He X. **K-means clustering via principal component analysis**. In: Proceedings of the twenty-first international conference on Machine learning. ACM; 2004. p. 29. doi:10.1145/1015330.1015408.
108. Klami A, Virtanen S, Kaski S. **Bayesian canonical correlation analysis**. The Journal of Machine Learning Research. 2013;14(1):965–1003.
109. Yuan Y, Savage RS, Markowetz F. **Patient-specific data fusion defines prognostic cancer subtypes**. PLoS Comput Biol. 2011;7(10):e1002227. doi:10.1371/journal.pcbi.1002227.
110. Kormaksson M, Booth JG, Figueroa ME, Melnick A, et al. **Integrative model-based clustering of microarray methylation and expression data**. The Annals of Applied Statistics. 2012;6(3):1327–1347. doi:10.1214/11-aoas533.
111. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. **Bayesian correlated clustering to integrate multiple datasets**. Bioinformatics. 2012;28(24):3290–3297. doi:10.1093/bioinformatics/bts595.
112. Savage RS, Ghahramani Z, Griffin JE, Kirk P, Wild DL. **Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data**. In: International Conference on Machine Learning (ICML) 2012: Workshop on Machine Learning in Genetics and Genomics; 2012. .
113. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. **Similarity network fusion for aggregating data types on a genomic scale**. Nature methods. 2014;11(3):333–337. doi:10.1038/nmeth.2810.
114. Serra A, Fratello M, Fortino V, Raiconi G, Tagliaferri R, Greco D. **MVDA: a multi-view genomic data integration methodology**. BMC bioinformatics. 2015;16(1):1. doi:10.1186/s12859-015-0680-3.
115. Madeira SC, Oliveira AL. **Biclustering algorithms for biological data analysis: a survey**. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2004;1(1):24–45. doi:10.1109/tcbb.2004.2.
116. Bunte K, Leppaaho E, Saarinen I, Kaski S. **Sparse group factor analysis for biclustering of multiple data sources**. Bioinformatics. 2016;32(16):2457–2463. doi:10.1093/bioinformatics/btw207.
117. Klami A, Virtanen S, Leppaaho E, Kaski S. **Group Factor Analysis**. Neural Networks and Learning Systems, IEEE Transactions on. 2015;26(9):2136–2147. doi:10.1109/TNNLS.2014.2376974.
118. Sun J, Bi J, Kranzler HR. **Multi-view biclustering for genotype-phenotype association studies of complex diseases**. In: Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on. IEEE; 2013. p. 316–321. doi:10.1109/bibm.2013.6732509.
119. Sun J, Bi J, Kranzler HR. **Multi-view singular value decomposition for disease subtyping and genetic associations**. BMC genetics. 2014;15(1):73. doi:10.1186/1471-2156-15-73.
120. Birtwistle MR, Hatakeyama M, Yumoto N, Ogunnaike BA, Hoek JB, Kholodenko BN. **Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses**. Molecular Systems Biology. 2007 Nov;3. doi:10.1038/msb4100188.
121. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. **Gene prioritization through genomic data fusion**. Nature Biotechnology. 2006 May;24(5):537–544. doi:10.1038/nbt1203.
122. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, 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 Aug;28(8):827–838. doi:10.1038/nbt.1665.