

Document downloaded from:

<http://hdl.handle.net/10251/190870>

This paper must be cited as:

Tarazona, S.; Arzalluz-Luque, Á.; Conesa, A. (2021). Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science*. 1(6):395-402.
<https://doi.org/10.1038/s43588-021-00086-z>



The final publication is available at

<https://doi.org/10.1038/s43588-021-00086-z>

Copyright Nature Publishing Group

Additional Information

Undisclosed, unmet and neglected challenges in multi-omics studies

Sonia Tarazona¹, Angeles Arzalluz-Luque¹, Ana Conesa^{2,3,4,*}

¹Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Valencia, Spain

²Microbiology and Cell Science Department, Institute for Food and Agricultural Research, University of Florida, Gainesville, Florida, United States

³Genetics Institute, University of Florida, Gainesville, Florida, United States

⁴ Present address: Institute for Integrative Systems Biology, Spanish National Research Council, Valencia, Spain

*Corresponding author, aconesa@ufl.edu

Abstract

Multi-omics approaches have become a reality in both large genomics projects and small laboratories. However, the multi-omics research community still faces a number of issues that have either not been sufficiently discussed or for which current solutions are still limited. In this Perspective, we elaborate on these limitations and suggest points of attention for future research. We finally discuss new opportunities and challenges brought to the field by the rapid development of single-cell high-throughput molecular technologies.

1. Introduction

Over the past decade, we have witnessed an increasing number of studies that have generated multi-omics data by different high-throughput molecular profiling technologies, each of them measuring a different layer of cellular regulation. Broadly, multi-omics approaches seek to either deliver more robust classifiers of biological samples (e.g. cancer subtypes) or to provide new insights into the circuits of molecular interactions that underlie life and disease. The strategies for the analysis of these datasets have greatly evolved in time. While initial studies analyzed data modalities independently to then combine results¹⁻³, a myriad of novel algorithms applying meta-analysis⁴, Bayesian methods⁵⁻⁷, factor analysis⁸, and machine or deep learning⁹⁻¹¹ has been developed to offer statistically sound solutions to the integrative analysis of the multi-faceted data. In addition, numerous reviews have described and compared these data analysis methodologies either globally¹²⁻¹⁵ or with domain-specific applications¹⁶⁻¹⁹. However, less attention has been paid to the discussion of the inherent, outstanding and sometimes neglected challenges presented by the integration of the broad diversity of multi-omics data types. In this Perspective, we present these challenges, discuss how they limit the potential of the multi-omics data paradigm and suggest points for consideration in future computational research. In order to provide context, we first briefly present the general scenario of multi-omics applications and main analysis strategies. We then discuss five aspects of multi-omics data analysis that we believe represent poorly addressed problems and outstanding computational needs. Finally, we address future challenges for the field, including those brought by the rapid development of multi-omics single-cell technologies.

2. Goals and strategies of multi-omics studies

The concept of "omic" has evolved over time, shifting from an initial definition that pertained the measurement of (nearly) all biomolecules or molecular events of a given type (e.g. genomics for DNA, transcriptomics for RNA, epigenomics for DNA modifications, proteomics for proteins or metabolomics for metabolites) to a more extended view that includes other types of high-dimensional data, frequently in combination with the molecular profiles, such as imaging (radiomics)^{20,21}, FACS/CyTOF²² and large-scale phenotyping (phenomics)^{23,24}. Considering this

wide definition of the multi-omics datasets, we differentiate two major types of applications, depending on whether the samples or the molecular features are the primary target of the analysis (Figure 1).

One of the main purposes of adopting a multi-omics strategy is to make use of the wealth of information provided by different types of molecular and non-molecular data to better classify biological samples (Figure 1). This approach is increasingly being adopted in medical studies aiming to improve patient stratification, either in a supervised or unsupervised manner, in order to develop precision treatments^{25,26}. Unsupervised multi-omics integration has the potential for capturing complex relationships within these data to reveal otherwise unnoticed groups of samples. Integrative clustering²⁷ and, in particular, latent-space analysis methods are suited for this goal as they are able to extract underlying variation patterns from large data structures. Through the years, a great variety of algorithmic adjustments have been proposed to accommodate complex experimental designs^{28,29}, preserve their multi-block structure in the analysis^{8,30}, dissect shared versus omics-specific variation³¹⁻³⁴, or consider alternative underlying distributions⁷. For a comprehensive review of unsupervised multi-omics analysis methods, refer to^{14,15}.

Multi-omics data can also be employed to predict response variables, such as clinical outcomes or other traits. Supervised classification methods and regression models are often used to this end³⁵⁻³⁷. Here, a distinction can be made depending on whether sample classification is of interest, or additionally, the study requires the identification of biomarkers associated with the response variable^{38,39}. Regarding the former, machine and deep learning methods are particularly powerful, as they achieve high prediction accuracy from large multi-omics datasets⁴⁰. For the latter, however, other analytical strategies are more suited, among which dimension reduction algorithms^{8,26} and regularization techniques (such as Lasso⁴¹⁻⁴⁴ or Elastic Net⁴⁵ for variable selection) are especially popular. Recently, interpretable machine learning methods applied to multi-omics data have emerged as a solution that combines both prediction power and the ability to identify biomarkers⁴⁶.

The second asset of the multi-omics approach is the often-claimed potential for unraveling regulatory mechanisms encompassing several molecular layers. In this case, the focus of the

analysis is the selection of relevant omics features followed by the inference of the relationships among them (Figure 1). Some methods developed to this end study specific connections using their associated data types, such as gene expression and methylation^{36,46} or gene expression and genome variation (i.e. transcriptome-wide association studies, TWAS)⁴⁷, while other approaches attempt to sequentially expand across omics layers, performing a chained analysis of regulatory relationships⁴⁸. Regression models have also been used in this context, including, but not limited to, regularized regression¹³, regression based on latent variables⁸ and structural equation models⁴⁹. In many of these applications, some level of a-priori information of the intra-feature relationships, such as genome coordinates proximity or trans-acting binding patterns, may be employed. Mechanistic models provided by the multi-omics adaptation of Flux Balance Analysis (FBA)^{50,51} have extended this concept to the analysis of metabolic regulation. While FBA initially focused on metabolic flux modelling, the development of variants that integrate transcriptomics⁵² data into the FBA network has shifted the paradigm, and there is also growing interest in studying the chromatin-metabolism interplay and its implications in human disease using similar approaches⁵³.

Finally, some modeling strategies have been proposed to achieve both of the goals described here: outcome prediction and inference of the co-regulation networks driving the outcome. mixOmics⁸ is a good example of a dimension reduction approach that creates an outcome prediction model, performs a Lasso variable selection on the predictors and returns a co-regulation network of predictors. Similarly, the recently published CANTARE⁵⁴ strategy first fits linear models to build the interaction network and then uses logistic regression to predict the outcome from highly connected subnetworks and other non-omics variables.

3. Challenges in multi-omics data analysis

Having established the general goals and analysis frameworks in multi-omics studies, we next reflect upon some of the pending issues in the multi-omics data analysis field that require attention to realize the full potential of combining high-throughput data obtained from different molecular layers. These challenges include the heterogeneity across omics technologies, the

treatment of missing values, the difficulty of interpreting multi-layered systems models, and the problems pertaining data annotation, storage and computational resources (Figure 2).

a) Heterogeneity in signal-to noise ratio among omics technologies

Although there is broad agreement over the fact that different omics technologies have disparate precision levels, the different signal-to-noise ratio across multi-omics measurements (Figure 3a) is often ignored during multi-modal study design and method development. There is a general lack of studies that compare variance figures across omics modalities and how they affect data integration efforts⁵⁵. This is particularly relevant when multi-omics methods are used to explain regulatory mechanisms and discover feature connections across molecular layers. In particular, combining omics data with different detection power may result in false conclusions, since a lack of detection of an association among two features can easily be interpreted as an absence of molecular relationship instead of as a detection problem. For instance, ChIP-seq possesses less sensitivity than RNA-seq, which may result in significant gene expression changes not being mirrored by chromatin modifications due to the lower resolution of the chromatin data. Additionally, omics platforms also differ in the number of detected features (Figure 3b) and degree of completeness achieved when measuring the targeted molecular space (Figure 3c). Proteomics constitutes a representative example, given its inherent bias to detect abundant proteins or to target only those with specific chemical properties, while this problem is largely absent in transcriptomics. As a result of this differential feature coverage, analyses of the relationship between gene and protein expression are constrained by the proteomics data. Differences in both signal-to-noise ratio and number of measured features ultimately affect statistical power⁵⁵, which can vary substantially across omics modalities (Figure 3d). To mitigate this, algorithms for estimating the number of samples per omic required to achieve a given statistical power or sample classification accuracy have been developed⁵⁵. However, these methods have also evidenced the underlying dichotomy of multi-omics experimental design, where complete designs (same number of samples per data modality) result in different power across omics, whereas equivalent power requirements result in unbalanced (different number of samples per data modality) designs. Both scenarios constitute important data integration pitfalls, the former related to the above-described false negative association problem, and the latter due to the fact that many

integrative statistical methods require complete experimental designs. The field therefore strives for new analysis methods that can successfully balance both aspects, achieving homogeneous discovery power across multimodal data while also enabling easy-to-implement experimental designs.

Power as discussed above refers to the ability to detect a given effect size with sufficient statistical significance using each omic modality. However, integrating different types of measurements can also increase the power to identify a targeted effect, or as mentioned earlier, improve sample classification. This is exemplified by the extended TWAS framework, where power for the detection of SNPs of phenotypic relevance is increased by the incorporation of gene expression and DNA methylation data^{47,56,57}, as both expression quantitative trait loci (eQTLs) and methylation quantitative trait loci (mQTLs) inform on the functional significance of the genome variants. Similarly, meta-analysis methods combine p-values associated to instances of the “same feature” measured by different omics modalities, such as promoter methylation, gene expression and protein abundance of the same coding gene⁵⁸ or pathway⁵⁹. These power enhancement approaches rely on the assumption that a coordinated effect acts across molecular layers following biological principles. Although likely to be true in a majority of settings, this strategy might be overly simplistic in some cases, and fail to capture complex regulatory patterns or to highlight effects that are exclusively associated to one molecular layer, such as genome variants affecting protein function but not expression, or protein activity levels regulated by post-translational modifications. A sound integrative analysis of multi-omics data should be aware of these possibilities and carefully ponder how the assortment in molecular interdependence may affect analysis results.

b) Missing value imputation in multi-omics data

A critical issue in the integrative analysis of multi-omics data is, without a doubt, the problem of missing values. Missing values can be caused by features failing to be measured in some of the samples, a frequent scenario in proteomics and metabolomics data, or by detection difficulties inherent to the corresponding technology (e.g. very long transcripts might be difficult to capture by RNA-seq due to 5' degradation, or repetitive sequences may be hard to assemble by short-reads DNA sequencing methods). While the former can be easily be addressed by imputation

methods, i.e. leveraging the correlation structure across omics modalities⁶⁰, the latter is related to analytical platform coverage⁵⁵ and therefore can solely be alleviated by employing alternative technologies, such as long-read sequencing⁶¹. A more frequent and disruptive category of missing values, however, arises when a complete set of measurements is unavailable for some of the samples in the multi-omics dataset. This is typically encountered in cohort studies, in which not all individuals will have all omics data types collected, and thus the number patients with complete records can be a small fraction of the total dataset⁶². Missing samples can also be the result of quality control procedures that may discard complete sets of measurements for some samples, or, as previously discussed, that may originate due to power considerations leading to unbalanced experimental designs. This particular kind of missing values represents a problem for many integrative analysis methods, which require matching observations across data matrices. Fortunately, some approaches have already addressed the problem, usually taking advantage of the fraction of observations that provide complete measurements to transfer information across samples^{7,63-65}. Among them, MOFA⁷ analyzes the latent space across omics types to impute missing samples, while MultiBac⁶⁵ creates a multivariate predictive model of the incomplete omics types as a function of a shared omics modality. However, in spite of their success in creating complete multi-omics datasets and unlocking the utilization of powerful data integration approaches, missing value imputation methods risk reducing the variability of the resulting dataset⁶⁶ and creating a data structure that violates independence assumptions required by many statistical frameworks. As an example of this, when DNA methylation is used to impute missing gene expression values, both datasets become dependent and therefore are flawed for a subsequent study of the correlation between gene expression and methylation levels. Strikingly, these issues are frequently overlooked by the multi-omics data analysis community, as missing data imputation is still considered to be a pre-processing step bearing no relationship to downstream statistical analysis. Instead, data dependence as well as the distribution requirements imposed by multi-omics data integration methods should be re-evaluated and verified in those cases where missing value imputation is a part of the pre-processing pipeline, and sensitivity analysis should be performed to assess the impact of imputation in the downstream analyses.

c) Interpretability of multi-omics models

One of the most attractive characteristics of multi-omics approaches is their capacity to deliver multi-layered systems biology models that bring us closer to understanding the complexity of life at the molecular level. However, this idyllic view is still far from reality. In most cases, interpretability is restricted to the identification of biomarkers and the biological processes they represent. Some of the strategies adopted to interpret multi-omics data include functional enrichment analyses that are either applied to each omics type independently⁶⁷⁻⁷⁰ or in combination^{59,71}. Even though these methods can successfully report the biological processes affected in the study, they also have inherent limitations when it comes to providing mechanistic insights across molecular layers. The mechanistic view is best achieved when the multi-omics data are mapped onto existing pathway models^{59,72-74}. However, representation of different types of molecular data on the pathway maps is difficult and these representations lack flexibility for displaying novel regulatory relationships. Data-driven network representations (i.e. Cytoscape⁷⁵, 3Omics⁷⁶, and so forth) where each type of omics feature is represented by a different symbol, although popular and versatile, are prone to become too large, complex and hard to interpret. In general, extant approaches for the visual interpretation of multi-omics models tend to be rather static and represent a final analysis result instead of a dynamic mathematical model. Computational methods that combine the mathematical formulation of multi-layered regulation with powerful and interactive visualization solutions are yet to be developed. Such tools would enable both a high-level overview and a detailed mechanistic understanding of the modelled molecular systems. Graphical databases such as Neo4j⁷⁷ are promising solutions, as they are flexible to navigate and query, although they still require incorporation of statistical tools to permit model-based interrogation of the multi-omics data⁷⁸.

d) Annotation and storage of multi-omics datasets.

Despite the large amount of multi-omics studies and databases available in the public domain, the retrieval of multi-omics data for any given set of biological entities is still an open problem. Repositories that collect data from different high throughput experiments around the same topic include search options for accessing specific studies or data platforms, but hardly connect data across them⁷⁹⁻⁸¹. Multi-omics projects that do generate multi-omics data on a related set of samples may be of very different scope and magnitude, and this hinders their possibilities for

sharing following FAIR (findability, accessibility, interoperability, and reproducibility) principles^{62,82}. Large multi-omics projects usually develop *ad-hoc* online portals with consolidated metadata annotations that facilitate access to omics data files of the same samples (Table 1). However, this is not as straightforward when multi-omics projects make use of public repositories for data sharing, which is the best available option for small studies. In these cases, datasets are hosted at omic-type specific repositories such as SRA⁸³, dbGAP⁸⁴ or EGA⁸⁵ for sequencing data, MetaboLights⁸⁶ for metabolomics or ProteomeXchange⁸⁷ for proteomics data. Unfortunately, the multi-omics nature of these datasets is marginally taken into account and hyperlinks are the only option to connect the same study across databases. As a consequence, the retrieval of the sample-level matched multi-omics dataset is pretty much a manual and cumbersome task, when not impossible. One possible solution when distributing data across specialized repositories would be to construct metadata summaries connecting multi-omics data files from the same samples, then making them available on data commons⁸⁸, although this is not common practice yet. While software is available for the integrated annotation of multi-omics datasets, either for small laboratories⁸⁹ or technology service providers^{90,91}, these tools are not intended for public data sharing. Efforts such as the EBI OmicsDI⁹² developed a unified querying system that rapidly finds omics datasets responding to the same keyword across repositories, although linking does not reach the sample level. Other initiatives, such as Figshare (<https://figshare.com/>), Zenodo (<http://zenodo.org>) or Lifebit (<https://lifebit.ai/>), have also succeeded in offering a common storage place. Still, recovery of the multi-omics dataset structure is not better supported by these resources, mainly due to the lack of a common framework for multi-omics metadata annotation. While the standards for omics data sharing have long been established⁹³ and minimal information guidelines have been defined for most omics types, the *Minimal Information about a Multi-Omics Experiment* has not yet been developed. Such a standard should provide a metadata description framework in which the experimental conditions, batches, replication and pre-processing are defined. Ultimately, this would enable the meaningful connection of samples across omics data files and the effective reutilization of the multi-omics data structure.

e) Performance and scalability challenges

Multi-omics datasets are intrinsically large in terms of both features and samples and will likely

grow in complexity in the future, especially as multi-omics studies steadily expand from cross-condition comparisons to precision medicine studies. In this novel setting, multi-omics datasets from large cohorts of patients will need to be integrated^{17,94-96}. Furthermore, intrinsically large data types such as imaging data are also being incorporated into multi-omics studies⁹⁷, while single-cell omics⁹⁸, in which the number of available observations increases exponentially, are likely to eventually replace bulk strategies and increase the complexity of this scenario. Luckily, the popularization of cloud-based services is already alleviating this burden for research groups, who rely on cloud servers for storage to «make the software come to the data»⁹⁹. This facilitates data sharing and provides access to computational resources in a scalable manner on demand¹⁰⁰, while also saving the cost of set-up and maintenance of these computational infrastructures. These two considerations make a particularly strong case for fully cloud-based environments, as they constitute major advantages to the extensively used High Performance Computing (HPC) clusters. However, the broad deployment of HPC may raise reluctance among members of the bioinformatics community. Suitable training programs will therefore be essential to provide the necessary skills to make a long-lasting transition, since cloud computing is only to become popular as it evolves into a more powerful, accessible and cost-effective service.

Handling large amounts of data also creates new challenges in computational power. Regarding software, algorithms need to be designed to effortlessly operate on big datasets. The availability of multicore CPUs in cloud computing infrastructures has enabled the development of parallelized algorithmic strategies, but performing an efficient implementation of parallel computation is far from trivial¹⁰¹. Furthermore, the nature of the data is very diverse across omics, which precludes the design of one-size-fits-all solutions for efficient data handling and may require computational solutions to be tailored to each omic¹⁰². Regarding hardware, fast multi-omics data processing can be assisted by the use of sufficiently powerful computing units. Ever since the technology emerged, CPUs have begun to be replaced by Graphics Processing Units (GPUs), which enable faster processing of omics data¹⁰³. GPUs have become a particularly interesting alternative given the popularization of deep learning methods for the analysis of multi-omics data⁹. Deep learning can effectively be used to extract patterns and help gain insight from big data, but these methods require frameworks that can be up to the task of providing high computational power and massive

parallelization¹⁰⁴. All in all, multi-omics data integration could be boosted by a smart combination of efficient software design and the popularization of new, more powerful processing technologies as a part of cloud-based services.

4. Future perspectives: towards a cell-level model of molecular regulation

One of the most promising among recent advances in multi-omics has been the development of single-cell multi-omics technologies, which opens up exciting opportunities for the inference of multi-layered models of molecular regulation at a better resolution. Indeed, accounting for cell type heterogeneity can help refine models of complex biological processes such as tumor progression¹⁰⁵ and understand the functionality of tissues and organs with high levels of cellular specialization, such as the brain¹⁰⁶. Especially interesting within single-cell multi-omics are the increasing number of parallel methods, where several omics modalities can be generated from the same cell. Currently published studies have combined scRNA-seq with the measurement of one or more additional molecular layers, including DNA variability, chromatin accessibility, DNA methylation and protein abundance (see¹⁰⁷ and⁹⁸ for a comprehensive list of technologies). Given that cell identities are shared within the dataset, features from different modalities can be matched using a wide range of methods (see¹⁰⁸ for a review), unlocking downstream integrative analyses that could take multi-omics studies to the next level. Even more interestingly, single-cell multimodal omics could not only refine bulk models, but also expand them with new layers whose monitoring has been enabled specifically by single-cell approaches, such as CRISPR perturbations¹⁰⁹, spatial information¹¹⁰, lineage reconstruction¹¹¹ and time-dependent trajectory data¹¹².

While single-cell multi-omics data constitutes an exciting opportunity due to the wealth of biological information it provides, integrative analysis faces the same type of challenges as described above, which are sometimes exacerbated by the very nature of single-cell data. Regarding missing values, for instance, methods coupling scRNA-seq with protein abundance can only capture a reduced number of proteins, and in a targeted manner¹¹³⁻¹¹⁵ while scATAC-seq suffers from very low read coverage, leading to high sparsity and making data analysis and interpretation challenging¹¹⁶. In addition, access to technologies that can generate multimodal

data is still limited regarding both infrastructure and cost, while some published protocols have important drawbacks such as very limited cell throughputs or lack of automation (see ¹¹⁷ and ⁹⁸ for a summary). As a result, current integrative studies usually rely on non-parallel methodologies integrating modalities from different datasets. This strategy poses the additional challenge of matching cell types or states across omics, a non-trivial problem that relies on strong assumptions, but could greatly boost the number of single-cell multi-omics studies by enabling the exploitation of publicly available data from different modalities. To achieve this, complex integration strategies (recently reviewed in ¹¹⁷ and ¹⁰⁸) have been developed, a number of which are dedicated to single-cell specific data modalities such as trajectory¹¹⁸ and lineage data¹¹⁹. Another interesting way around these limitations is the development of methods for the extraction of multiple types of information from scRNA-seq data, namely the detection of copy number variation (CNV)¹²⁰, eQTLs¹²¹ or trajectory information (RNA velocity)¹²². Finally, single-cell integrative models need to contemplate technology-specific limitations such as cell-level noise and sparsity, which precludes the usage of approaches developed for bulk data and calls for innovative strategies, while specific methods for the biological interpretation of cell and cell-type resolved data need to be developed.

All in all, multiple advances are to be expected in the single-cell multi-omics field in the upcoming years. In addition to solving current technological limitations to improve the quality of single-cell multimodal data, the development of missing modalities such as ChIP-seq, metabolomics and large-scale proteomics would be a valuable addition to complete the catalogue of available single-cell approaches. Furthermore, the potential of combining single-cell omics with cell-level imaging and morphological profiling technologies, such as Cell Painting¹²³, that represents the single-cell counterpart of bulk imaging data, is yet to be explored. Another key aspect for the future of single-cell multi-omics will be the development of methods that can decipher the intricate signals connecting modalities (e.g. scATAC-seq and scRNA-seq) to generate biologically interpretable hypotheses and models of molecular regulation. In line with this, the application of causal statistical models such as structural equation models or graphical models are a promising field in multi-omics regulatory analysis (see e.g. PARADIGM¹²⁴). However, these tools rely on strong assumptions and require a graph model to be built based on previous knowledge, as well as enough data for the model to be adjusted. Single-cell data might represent an opportunity for

this type of methodologies, provided that other problems -such as sparsity- are addressed. Finally, the incorporation of clinical data from multiple patients into single-cell studies and statistical modelling efforts would be the cherry on top of the cake for the career towards a cell-level model of molecular regulation that can explain both health and disease.

Data Availability

Data used for figure 3 were taken from the publicly available STATegra dataset¹²⁵. Pre-processed values used to generate graphs are provided together with the code.

Code Availability

The code and data used to generate figure 3 is available on GitHub: [https://github.com/ConesaLab/Perspective Multi-Omics Integration](https://github.com/ConesaLab/Perspective-Multi-Omics-Integration).

Acknowledgements

This work has been funded by the Spanish Ministry of Science and Innovation with grant number BES-2016-076994 to AAL.

Competing Interests

Authors declare no competing interests.

Author contributions

AC drafted the structure of the manuscript and integrated author contributions. AAL, ST and AC drafted the manuscript, reviewed the literature, contributed to figures and approved the final version of the manuscript.

References

- 1 Fan, T. W. M., Bandura, L. L., Higashi, R. M. & Lane, A. N. Metabolomics-edited transcriptomics analysis of Se anticancer action in human lung cancer cells. *Metabolomics* **1**, 325-339, doi:10.1007/s11306-005-0012-0 (2005).

- 2 Panguluri, S. K. *et al.* Genomic profiling of messenger RNAs and microRNAs reveals potential mechanisms of TWEAK-induced skeletal muscle wasting in mice. *PloS one* **5**, e8760-e8760, doi:10.1371/journal.pone.0008760 (2010).
- 3 Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research* **21**, 1757-1767, doi:10.1101/gr.121541.111 (2011).
- 4 Kim, S., Jhong, J.-H., Lee, J. & Koo, J.-Y. Meta-analytic support vector machine for integrating multiple omics data. *BioData Mining* **10**, 2, doi:10.1186/s13040-017-0126-8 (2017).
- 5 Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics (Oxford, England)* **26**, i237-i245, doi:10.1093/bioinformatics/btq182 (2010).
- 6 Mo, Q. *et al.* A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* **19**, 71-86, doi:10.1093/biostatistics/kxx017 (2017).
- 7 Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology* **14**, e8124-e8124, doi:10.15252/msb.20178124 (2018).
- 8 Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology* **13**, e1005752-e1005752, doi:10.1371/journal.pcbi.1005752 (2017).
- 9 Zhang, L. *et al.* Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. *Frontiers in genetics* **9**, 477-477, doi:10.3389/fgene.2018.00477 (2018).
- 10 Ma, T. & Zhang, A. Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). *BMC genomics* **20**, 944-944, doi:10.1186/s12864-019-6285-x (2019).
- 11 Huang, Z. *et al.* SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. *Frontiers in genetics* **10**, 166-166, doi:10.3389/fgene.2019.00166 (2019).
- 12 Bersanelli, M. *et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics* **17 Suppl 2**, 15-15, doi:10.1186/s12859-015-0857-9 (2016).
- 13 De Bin, R., Boulesteix, A.-L., Benner, A., Becker, N. & Sauerbrei, W. Combining clinical and molecular data in regression prediction models: insights from a simulation study. *Briefings in Bioinformatics* **21**, 1904-1919, doi:10.1093/bib/bbz136 (2020).
- 14 Pierre-Jean, M., Deleuze, J.-F., Le Floch, E. & Mauger, F. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Briefings in Bioinformatics* **21**, 2011-2030, doi:10.1093/bib/bbz138 (2020).
- 15 Meng, C. *et al.* Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in bioinformatics* **17**, 628-641, doi:10.1093/bib/bbv108 (2016).
- 16 Buescher, J. M. & Driggers, E. M. Integration of omics: more than the sum of its parts. *Cancer Metab* **4**, 4-4, doi:10.1186/s40170-016-0143-y (2016).
- 17 Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome biology* **18**, 83-83, doi:10.1186/s13059-017-1215-1 (2017).
- 18 Kristensen, V. N. *et al.* Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer* **14**, 299-313, doi:10.1038/nrc3721 (2014).

- 19 Sathyanarayanan, A. *et al.* A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. *Briefings in Bioinformatics* **21**, 1920-1936, doi:10.1093/bib/bbz121 (2020).
- 20 Zeng, H. *et al.* Integrative radiogenomics analysis for predicting molecular features and survival in clear cell renal cell carcinoma. *Aging (Albany NY)* **13**, 9960-9975, doi:10.18632/aging.202752 (2021).
- 21 Kirienko, M. *et al.* Radiomics and gene expression profile to characterise the disease and predict outcome in patients with lung cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, doi:10.1007/s00259-021-05371-7 (2021).
- 22 Zielinski, J. M., Luke, J. J., Guglietta, S. & Krieg, C. High Throughput Multi-Omics Approaches for Clinical Trial Evaluation and Drug Discovery. *Frontiers in immunology* **12**, 590742-590742, doi:10.3389/fimmu.2021.590742 (2021).
- 23 Houle, D., Govindaraju, D. R. & Omholt, S. Phenomics: the next challenge. *Nature Reviews Genetics* **11**, 855-866, doi:10.1038/nrg2897 (2010).
- 24 van Bezouw, R. F. H. M., Keurentjes, J. J. B., Harbinson, J. & Aarts, M. G. M. Converging phenomics and genomics to study natural variation in plant photosynthetic efficiency. *The Plant journal : for cell and molecular biology* **97**, 112-133, doi:10.1111/tpj.14190 (2019).
- 25 Zhu, R., Zhao, Q., Zhao, H. & Ma, S. Integrating multidimensional omics data for cancer outcome. *Biostatistics (Oxford, England)* **17**, 605-618, doi:10.1093/biostatistics/kxw010 (2016).
- 26 Balzano-Nogueira, L. *et al.* Integrative analyses of TEDDY Omics data reveal lipid metabolism abnormalities, increased intracellular ROS and heightened inflammation prior to autoimmunity for type 1 diabetes. *Genome biology* **22**, 39-39, doi:10.1186/s13059-021-02262-w (2021).
- 27 Shen, R., Olshen, A. B. & 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* **25**, 2906-2912, doi:10.1093/bioinformatics/btp543 (2009).
- 28 Yener, B. *et al.* Multiway modeling and analysis in stem cell systems biology. *BMC Syst Biol* **2**, 63, doi:10.1186/1752-0509-2-63 (2008).
- 29 Conesa, A., Prats-Montalbán, J. M., Tarazona, S., Nueda, M. J. & Ferrer, A. A multiway approach to data integration in systems biology based on Tucker3 and N-PLS. *Chemometrics and Intelligent Laboratory Systems* **104**, 101-111, doi:10.1016/j.chemolab.2010.06.004 (2010).
- 30 Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. M. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* **15**, 162, doi:10.1186/1471-2105-15-162 (2014).
- 31 van der Kloet, F. M., Sebastián-León, P., Conesa, A., Smilde, A. K. & Westerhuis, J. A. Separating common from distinctive variation. *BMC Bioinformatics* **17**, 195, doi:10.1186/s12859-016-1037-2 (2016).
- 32 O'Connell, M. J. & Lock, E. F. R.JIVE for exploration of multi-source molecular data. *Bioinformatics (Oxford, England)* **32**, 2877-2879, doi:10.1093/bioinformatics/btw324 (2016).
- 33 Bouhaddani, S. e. *et al.* Integrating omics datasets with the OmicsPLS package. *BMC Bioinformatics* **19**, 371, doi:10.1186/s12859-018-2371-3 (2018).

- 34 Planell, N. *et al.* STATegra: Multi-Omics Data Integration – A Conceptual Scheme With a Bioinformatics Pipeline. *Frontiers in Genetics* **12**, doi:10.3389/fgene.2021.620453 (2021).
- 35 Boulesteix, A.-L., De Bin, R., Jiang, X. & Fuchs, M. IPF-LASSO: Integrative L(1)-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Computational and mathematical methods in medicine* **2017**, 7691937-7691937, doi:10.1155/2017/7691937 (2017).
- 36 Kennedy, E. M. *et al.* An integrated -omics analysis of the epigenetic landscape of gene expression in human blood cells. *BMC Genomics* **19**, 476, doi:10.1186/s12864-018-4842-3 (2018).
- 37 Wu, M.-Y. *et al.* Regularized logistic regression with network-based pairwise interaction for biomarker identification in breast cancer. *BMC Bioinformatics* **17**, 108, doi:10.1186/s12859-016-0951-7 (2016).
- 38 Wu, C. *et al.* A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. *High Throughput* **8**, 4, doi:10.3390/ht8010004 (2019).
- 39 Lagani, V., Kortas, G. & Tsamardinos, I. BIOMARKER SIGNATURE IDENTIFICATION IN "OMICS" DATA WITH MULTI-CLASS OUTCOME. *Computational and Structural Biotechnology Journal* **6**, e201303004, doi:<https://doi.org/10.5936/csbj.201303004> (2013).
- 40 Le, D.-H. Machine learning-based approaches for disease gene prediction. *Briefings in Functional Genomics* **19**, 350-363, doi:10.1093/bfgp/elaa013 (2020).
- 41 Fang, H., Huang, C., Zhao, H. & Deng, M. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31**, 3172-3180, doi:10.1093/bioinformatics/btv349 (2015).
- 42 Klau, S., Jurinovic, V., Hornung, R., Herold, T. & Boulesteix, A.-L. Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics* **19**, 322, doi:10.1186/s12859-018-2344-6 (2018).
- 43 Li, J., Lu, Q. & Wen, Y. Multi-kernel linear mixed model with adaptive lasso for prediction analysis on high-dimensional multi-omics data. *Bioinformatics* **36**, 1785-1794, doi:10.1093/bioinformatics/btz822 (2020).
- 44 Park, H., Niida, A., Miyano, S. & Imoto, S. Sparse Overlapping Group Lasso for Integrative Multi-Omics Analysis. *Journal of Computational Biology* **22**, 73-84, doi:10.1089/cmb.2014.0197 (2015).
- 45 Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055-3062, doi:10.1093/bioinformatics/bty1054 (2019).
- 46 Patel-Murray, N. L. *et al.* A Multi-Omics Interpretable Machine Learning Model Reveals Modes of Action of Small Molecules. *Scientific reports* **10**, 954-954, doi:10.1038/s41598-020-57691-7 (2020).
- 47 Gusev, A. *et al.* Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat Genet* **50**, 538-548, doi:10.1038/s41588-018-0092-1 (2018).
- 48 Rubio, T. *et al.* Multi-omic analysis unveils biological pathways in peripheral immune system associated to minimal hepatic encephalopathy appearance in cirrhotic patients. *Scientific reports* **11**, 1907-1907, doi:10.1038/s41598-020-80941-7 (2021).

- 49 Cai, X., Bazerque, J. A. & Giannakis, G. B. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS computational biology* **9**, e1003068-e1003068, doi:10.1371/journal.pcbi.1003068 (2013).
- 50 Oberhardt, M. A., Chavali, A. K. & Papin, J. A. Flux balance analysis: interrogating genome-scale metabolic networks. *Methods Mol Biol* **500**, 61-80, doi:10.1007/978-1-59745-525-1_3 (2009).
- 51 Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nature biotechnology* **28**, 245-248, doi:10.1038/nbt.1614 (2010).
- 52 Covert, M. W., Schilling, C. H. & Palsson, B. Regulation of Gene Expression in Flux Balance Models of Metabolism. *Journal of Theoretical Biology* **213**, 73-88, doi:<https://doi.org/10.1006/jtbi.2001.2405> (2001).
- 53 Tzika, E., Dreker, T. & Imhof, A. Epigenetics and Metabolism in Health and Disease. *Frontiers in genetics* **9**, 361-361, doi:10.3389/fgene.2018.00361 (2018).
- 54 Siebert, J. C. *et al.* CANTARE: finding and visualizing network-based multi-omic predictive models. *BMC bioinformatics* **22**, 80-80, doi:10.1186/s12859-021-04016-8 (2021).
- 55 Tarazona, S. *et al.* Harmonization of quality metrics and power calculation in multi-omic studies. *Nature communications* **11**, 3092-3092, doi:10.1038/s41467-020-16937-8 (2020).
- 56 Soerensen, M. *et al.* A Genome-Wide Integrative Association Study of DNA Methylation and Gene Expression Data and Later Life Cognitive Functioning in Monozygotic Twins. *Frontiers in Neuroscience* **14**, doi:10.3389/fnins.2020.00233 (2020).
- 57 Dai, Y., Pei, G., Zhao, Z. & Jia, P. A Convergent Study of Genetic Variants Associated With Crohn's Disease: Evidence From GWAS, Gene Expression, Methylation, eQTL and TWAS. *Frontiers in Genetics* **10**, doi:10.3389/fgene.2019.00318 (2019).
- 58 Karathanasis, N., Tsamardinos, I. & Lagani, V. omicsNPC: Applying the Non-Parametric Combination Methodology to the Integrative Analysis of Heterogeneous Omics Data. *PLoS one* **11**, e0165545-e0165545, doi:10.1371/journal.pone.0165545 (2016).
- 59 Garcia-Alcalde, F., Garcia-Lopez, F., Dopazo, J. & Conesa, A. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics* **27**, 137-139, doi:10.1093/bioinformatics/btq594 (2011).
- 60 Voillet, V., Besse, P., Liaubet, L., San Cristobal, M. & González, I. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. *BMC Bioinformatics* **17**, 402, doi:10.1186/s12859-016-1273-5 (2016).
- 61 Kuo, R. I. *et al.* Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* **18**, 323, doi:10.1186/s12864-017-3691-9 (2017).
- 62 Conesa, A. & Beck, S. Making multi-omics data accessible to researchers. *Sci Data* **6**, 251-251, doi:10.1038/s41597-019-0258-4 (2019).
- 63 Dong, X. *et al.* TOBML: trans-omics block missing data imputation using a k-nearest neighbor weighted approach. *Bioinformatics* **35**, 1278-1283, doi:10.1093/bioinformatics/bty796 (2019).
- 64 Zhou, X., Chai, H., Zhao, H., Luo, C.-H. & Yang, Y. Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning-based neural network. *GigaScience* **9**, doi:10.1093/gigascience/giaa076 (2020).

- 65 Ugidos, M., Tarazona, S., Prats-Montalbán, J. M., Ferrer, A. & Conesa, A. MultiBaC: A strategy to remove batch effects between different omic data types. *Stat Methods Med Res* **29**, 2851-2864, doi:10.1177/0962280220907365 (2020).
- 66 Messer, K., Vaida, F. & Hogan, C. Robust analysis of biomarker data with informative missingness using a two-stage hypothesis test in an HIV treatment interruption trial: AIEDRP AIN503/ACTG A5217. *Contemporary Clinical Trials* **27**, 506-517, doi:<https://doi.org/10.1016/j.cct.2006.07.003> (2006).
- 67 Hong, M.-G., Pawitan, Y., Magnusson, P. K. E. & Prince, J. A. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Human genetics* **126**, 289-301, doi:10.1007/s00439-009-0676-z (2009).
- 68 Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
- 69 Arneson, D., Bhattacharya, A., Shu, L., Mäkinen, V.-P. & Yang, X. Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration. *BMC Genomics* **17**, 722, doi:10.1186/s12864-016-3057-8 (2016).
- 70 Welch, R. P. *et al.* ChIP-Enrich: gene set enrichment testing for ChIP-seq data. *Nucleic Acids Research* **42**, e105-e105, doi:10.1093/nar/gku463 (2014).
- 71 Canzler, S. & Hackermüller, J. multiGSEA: a GSEA-based pathway enrichment analysis for multi-omics data. *BMC bioinformatics* **21**, 561-561, doi:10.1186/s12859-020-03910-x (2020).
- 72 Long, Y., Lu, M., Cheng, T., Zhan, X. & Zhan, X. Multiomics-Based Signaling Pathway Network Alterations in Human Non-functional Pituitary Adenomas. *Frontiers in Endocrinology* **10**, doi:10.3389/fendo.2019.00835 (2019).
- 73 Hernández-de-Diego, R. *et al.* PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Research* **46**, W503-W509, doi:10.1093/nar/gky466 (2018).
- 74 Sakurai, N. *et al.* KaPPA-View4: a metabolic pathway database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data. *Nucleic Acids Research* **39**, D677-D684, doi:10.1093/nar/gkq989 (2011).
- 75 Su, G., Morris, J. H., Demchak, B. & Bader, G. D. Biological network exploration with Cytoscape 3. *Current protocols in bioinformatics* **47**, 8.13.11-18.13.24, doi:10.1002/0471250953.bi0813s47 (2014).
- 76 Kuo, T. C., Tian, T. F. & Tseng, Y. J. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Syst Biol* **7**, doi:10.1186/1752-0509-7-64 (2013).
- 77 Miller, J. J. in *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA*.
- 78 Yoon, B.-H., Kim, S.-K. & Kim, S.-Y. Use of Graph Database for the Integration of Heterogeneous Biological Data. *Genomics & informatics* **15**, 19-27, doi:10.5808/GI.2017.15.1.19 (2017).

79 Consortium, T. I. H. i. R. N. The Integrative Human Microbiome Project: Dynamic Analysis
of Microbiome-Host Omics Profiles during Periods of Human Health and Disease. *Cell Host
& Microbe* **16**, 276-289, doi:<https://doi.org/10.1016/j.chom.2014.08.014> (2014).

80 Project, T. I. C. G. *ICGC Data Portal*, <<https://dcc.icgc.org/>> (
81 Project, T. i. H. M. Human Microbiome Project Data Portal.

82 Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and
stewardship. *Sci Data* **3**, 160018-160018, doi:10.1038/sdata.2016.18 (2016).

83 Kodama, Y., Shumway, M., Leinonen, R. & International Nucleotide Sequence Database, C.
The Sequence Read Archive: explosive growth of sequencing data. *Nucleic acids research*
40, D54-D56, doi:10.1093/nar/gkr854 (2012).

84 Tryka, K. A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic acids
research* **42**, D975-D979, doi:10.1093/nar/gkt1211 (2014).

85 Lappalainen, I. *et al.* The European Genome-phenome Archive of human data consented
for biomedical research. *Nature Genetics* **47**, 692-695, doi:10.1038/ng.3312 (2015).

86 Haug, K. *et al.* MetaboLights: a resource evolving in response to the needs of its scientific
community. *Nucleic acids research* **48**, D440-D444, doi:10.1093/nar/gkz1019 (2020).

87 Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2017: supporting the cultural
change in proteomics public data deposition. *Nucleic Acids Research* **45**, D1100-D1106,
doi:10.1093/nar/gkw936 (2017).

88 Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, practical
genomic data sharing that accelerates research. *Nature Reviews Genetics* **21**, 615-629,
doi:10.1038/s41576-020-0257-5 (2020).

89 Hernandez-de-Diego, R. *et al.* STATegra EMS: an Experiment Management System for
complex next-generation omics experiments. *BMC Syst Biol* **8 Suppl 2**, S9,
doi:10.1186/1752-0509-8-S2-S9 (2014).

90 Lin, K. *et al.* MADMAX – Management and analysis database for multiple ~omics
experiments. *Journal of Integrative Bioinformatics* **8**, 59-74, doi:doi:10.1515/jib-2011-160
(2011).

91 Venco, F., Vaskin, Y., Ceol, A. & Muller, H. SMITH: a LIMS for handling next-generation
sequencing workflows. *BMC bioinformatics* **15 Suppl 14**, S3-S3, doi:10.1186/1471-2105-
15-S14-S3 (2014).

92 Perez-Riverol, Y. *et al.* Discovering and linking public omics data sets using the Omics
Discovery Index. *Nature biotechnology* **35**, 406-409, doi:10.1038/nbt.3790 (2017).

93 Chervitz, S. A. *et al.* in *Bioinformatics for Omics Data: Methods and Protocols* (ed Bernd
Mayer) 31-69 (Humana Press, 2011).

94 Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nature Reviews
Genetics* **19**, 299-310, doi:10.1038/nrg.2018.4 (2018).

95 van Karnebeek, C. D. M. *et al.* The role of the clinician in the multi-omics era: are you ready?
Journal of Inherited Metabolic Disease **41**, 571-582, doi:10.1007/s10545-017-0128-1 (2018).

96 Angione, C. Human Systems Biology and Metabolic Modelling: A Review—From Disease
Metabolism to Precision Medicine. *BioMed Research International* **2019**, 8304260,
doi:10.1155/2019/8304260 (2019).

- 97 Hériché, J.-K., Alexander, S. & Ellenberg, J. Integrating Imaging and Omics: Computational Methods and Challenges. *Annual Review of Biomedical Data Science* **2**, 175-197, doi:10.1146/annurev-biodatasci-080917-013328 (2019).
- 98 Lee, J., Hyeon, D. Y. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Experimental and Molecular Medicine* **52**, 1428-1442, doi:10.1038/s12276-020-0420-2 (2020).
- 99 Stein, L. D. The case for cloud computing in genome informatics. *Genome Biology* **11**, 207-207, doi:10.1186/gb-2010-11-5-207 (2010).
- 100 Oh, M., Park, S., Kim, S. & Chae, H. Machine learning-based analysis of multi-omics data on the cloud for investigating gene regulations. *Briefings in Bioinformatics* **22**, 66-76, doi:10.1093/bib/bbaa032 (2020).
- 101 Solomonik, E., Carson, E., Knight, N. & Demmel, J. Trade-Offs Between Synchronization, Communication, and Computation in Parallel Linear Algebra Computations. *ACM Transactions on Parallel Computing* **3**, 1-47, doi:10.1145/2897188 (2016).
- 102 Berger, B., Peng, J. & Singh, M. Computational solutions for omics data. *Nature Reviews Genetics* **14**, 333-346, doi:10.1038/nrg3433 (2013).
- 103 Alyass, A., Turcotte, M. & Meyre, D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC medical genomics* **8**, 33-33, doi:10.1186/s12920-015-0108-y (2015).
- 104 Chen, X.-W. & Lin, X. Big Data Deep Learning: Challenges and Perspectives. *IEEE Access* **2**, 514-525, doi:10.1109/ACCESS.2014.2325029 (2014).
- 105 Fan, J., Slowikowski, K. & Zhang, F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine* **52**, 1452-1465, doi:10.1038/s12276-020-0422-0 (2020).
- 106 Armand, E. J., Li, J., Xie, F., Luo, C. & Mukamel, E. A. Single-Cell Sequencing of Brain Cell Transcriptomes and Epigenomes. *Neuron* **109**, 11-26, doi:10.1016/j.neuron.2020.12.010 (2021).
- 107 Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nature Methods* **17**, 11-14, doi:10.1038/s41592-019-0691-5 (2020).
- 108 Forcato, M., Romano, O. & Bicciato, S. Computational methods for the integrative analysis of single-cell data. *Briefings in bioinformatics* **22**, 20-29, doi:10.1093/bib/bbaa042 (2021).
- 109 Adamson, B. *et al.* A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882.e1821, doi:<https://doi.org/10.1016/j.cell.2016.11.048> (2016).
- 110 Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, 5691-5691, doi:10.1126/science.aat5691 (2018).
- 111 Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology* **36**, 442-450, doi:10.1038/nbt.4103 (2018).
- 112 Trapnell, C. & Cacchiarelli, D. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, doi:10.1038/nbt.2859 (2014).
- 113 Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14**, 865-868, doi:10.1038/nmeth.4380 (2017).

- 114 Darmanis, S. *et al.* Simultaneous Multiplexed Measurement of RNA and Proteins in Single
Cells. *Cell Reports* **14**, 380-389, doi:10.1016/j.celrep.2015.12.021 (2016).
- 115 Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells.
Nature Biotechnology **35**, 936-939, doi:10.1038/nbt.3973 (2017).
- 116 Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-
seq data. *Genome Biology* **20**, 241-241, doi:10.1186/s13059-019-1854-5 (2019).
- 117 Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177(7)**, doi:doi:
10.1016/j.cell.2019.05.031 (2019).
- 118 Welch, J. D., Hartemink, A. J. & Prins, J. F. MATCHER: manifold alignment reveals
correspondence between single cell transcriptome and epigenome dynamics. *Genome
Biology* **18**, 138-138, doi:10.1186/s13059-017-1269-0 (2017).
- 119 Campbell, K. R. *et al.* clonealign: statistical integration of independent single-cell RNA and
DNA sequencing data from human cancers. *Genome Biology* **20**, 54-54,
doi:10.1186/s13059-019-1645-z (2019).
- 120 Fan, J. *et al.* Linking transcriptional and genetic tumor heterogeneity through allele analysis
of single-cell RNA-seq data. *Genome Research* **28**, 1217-1227, doi:10.1101/gr.228080.117
(2018).
- 121 Van Der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-
eQTLs and co-expression QTLs. *Nature Genetics* **50**, 493-497, doi:10.1038/s41588-018-
0089-9 (2018).
- 122 La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494-498, doi:10.1038/s41586-
018-0414-6 (2018).
- 123 Bray, M.-A. *et al.* Cell Painting, a high-content image-based assay for morphological
profiling using multiplexed fluorescent dyes. *Nature Protocols* **11**, 1757-1774,
doi:10.1038/nprot.2016.105 (2016).
- 124 Sedgewick, A. J., Benz, S. C., Rabizadeh, S., Soon-Shiong, P. & Vaske, C. J. Learning
subgroup-specific regulatory interactions and regulator independence with PARADIGM.
Bioinformatics (Oxford, England) **29**, i62-i70, doi:10.1093/bioinformatics/btt229 (2013).
- 125 Gomez-Cabrero, D. *et al.* STATegra, a comprehensive multi-omics dataset of B-cell
differentiation in mouse. *Sci Data* **6**, 256, doi:10.1038/s41597-019-0202-7 (2019).
- 126 Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: Creating Lasting Value beyond Its
Data. *Cell* **173**, 283-285, doi:10.1016/j.cell.2018.03.042 (2018).
- 127 Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia.
Nature **569**, 503-508, doi:10.1038/s41586-019-1186-3 (2019).
- 128 The Genotype-Tissue expression (GTEx) project. *Nat Genet* **45**, doi:10.1038/ng.2653 (2013).
- 129 Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.
Nature **590**, 290-299, doi:10.1038/s41586-021-03205-y (2021).
- 130 Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse
genomes. *Nature* **583**, 699-710, doi:10.1038/s41586-020-2493-4 (2020).
- 131 Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499-
504, doi:10.1038/s41586-019-0965-1 (2019).
- 132 Mergner, J. *et al.* Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature*
579, 409-414, doi:10.1038/s41586-020-2094-2 (2020).

133 O'Connor, T. R., Dyreson, C. & Wyrick, J. J. Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. *Bioinformatics* **21**, 4411-4413, doi:10.1093/bioinformatics/bti714 (2005).

Figures

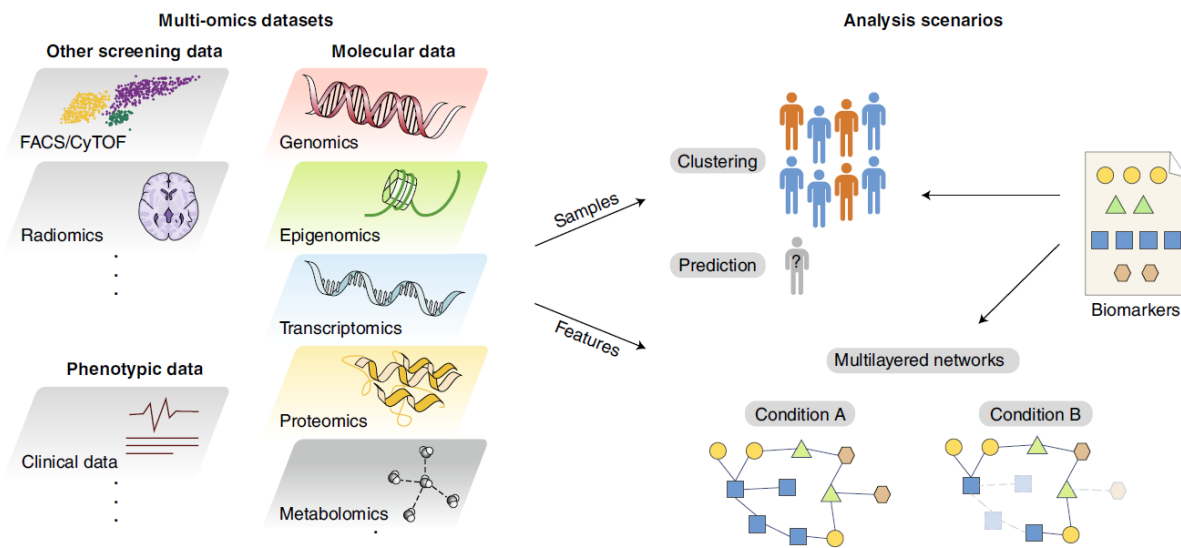


Fig 1. Schematic representation of analysis goals and challenges in multi-omics studies. Multi-omics datasets may be defined by any combination of molecular profiling data modalities such as genomics, epigenomics, transcriptomics, proteomics or metabolomics, and include other types of high-throughput data such as FACS/CyTOF or radiomics measurements, as well as phenotypic or clinical co-variates. Multi-omics studies can be classified depending on the primary goal of the analysis. When the focus is on the samples, these may undergo unsupervised clustering to reveal the underlying structure of the dataset, or supervised analysis to predict class allocation of new samples. Features involved in these groupings may be extracted and used as biomarkers. When the focus is on the features, the analysis seeks to identify significant relationships among omic variables belonging to different omics types, which can be represented as a network.

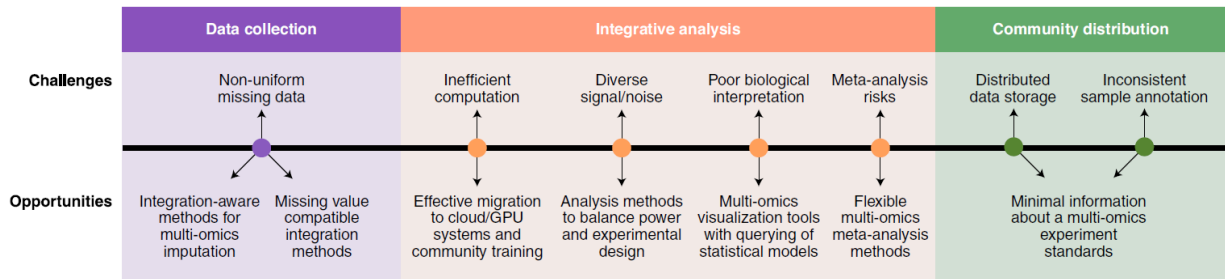


Fig 2. Multi-omics data integration challenges and opportunities. Analysis challenges together with possible solutions are presented at three different aspects -data collection, integrative analysis and community distribution- of the multi-omic data integration paradigm.

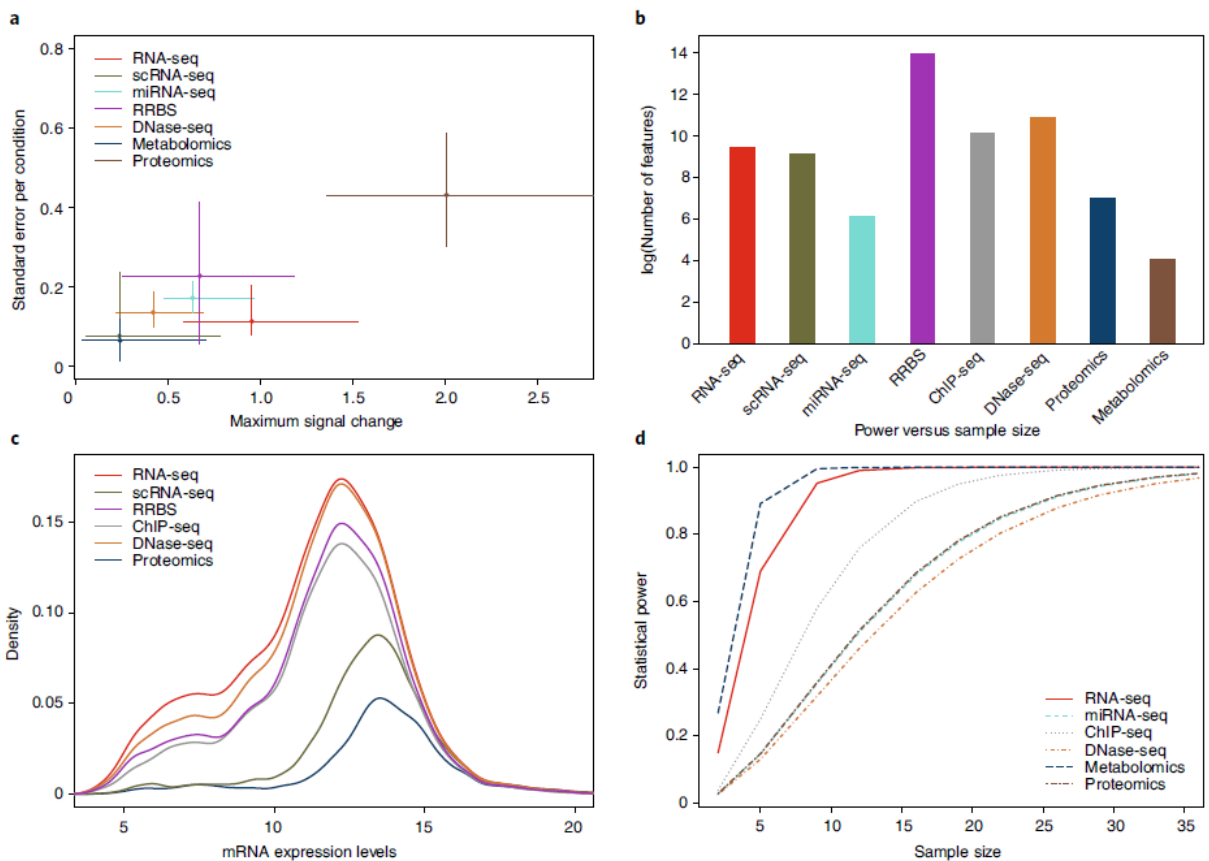


Fig 3. Comparison of omics data-types properties. The STATegra dataset representing a transcription factor Ikaros-induced mouse B-cell differentiation time course¹²⁵ is used as an example. **a** Signal-to-noise plot. Segments represent the interquartile ranges of maximum signal change between all pairs of experimental conditions (i.e. time points and Ikaros-induced versus control; X axis) and average standard

error across conditions (Y axis). **b** Number of features detected by each technology. **c** Differential coverage of the feature space. Each density line represents the distribution of the fraction of expressed genes captured by each method. **d** Statistical power curves across omics data types as a function of sample size calculated using MultiPower software⁵⁵ for statistical power calculations in multi-omics experiments. RNA-seq: RNA sequencing; ChIP-seq: Chromatin immunoprecipitation sequencing; RRBS: Reduced Representation Bisulfite Sequencing; DNase-seq: DNaseI sequencing.

Table 1: Public projects with linked multi-omics data. Non exhaustive list of public resources that provide multi-omics data measured on the same set of bio-entities. Last column indicates the type of sample where matching data are generated.

Study	Available omic types	Scope	Magnitude	Matching level
TCGA ¹²⁶	Gene expression, methylation, protein, microRNA	Cancer types	> 10,000 samples	Tumor sample
CCLE ¹²⁷	Genomics, transcriptomics, miRNAs, DNA methylation, proteomics, histone modifications	Cancer cell lines	~ 1000 cell lines	Cell line
GTEX ¹²⁸	Genome variation, transcriptomics	Human tissues	~ 1000 donors	Donor and sample
TopMED ¹²⁹	Genomics, transcriptomics, metabolomics, proteomics	Lung, heart, blood, sleep disorders. Multi-ethnic	> 50,000 individuals 16 multi-omics studies	Donor and sample
ENCODE ¹³⁰	RNA-seq, DNase-seq, microRNA-seq, ATAC-seq, ChIP-seq, HiC, WGBS, RRBS,	Human and mouse cell lines and tissues	> 10,000 experiments	Cell line, tissue, primary cell type
BLUEPRINT iHEC ¹³¹	ChIP-seq, DNase-seq, RNA-seq, Bisulfite-seq	Human cell types	> 7,500 datasets	Cell types
ATHENA ^{132,133}	Transcriptomics, proteomics	Arabidopsis tissues	30 plant tissues	Tissue type