nature computational science

PERSPECTIVE

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

Undisclosed, unmet and neglected challenges in multi-omics studies

Sonia Tarazona[1], Angeles Arzalluz-Luque[1] and Ana Conesa[2,3,4] ✉

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.

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 (for example, 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 results[1–3], a myriad of novel algorithms applying meta-analysis[4], Bayesian methods[5–7], factor analysis[8], and machine or deep learning[9–11] 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 globally[12–15] or with domain-specific applications[16–19]. 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. 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.

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 (for example, 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/CyTOF (fluorescence-activated cell sorting/cytometry by time of flight)[22] and large-scale phenotyping (phenomics)[23,24]. Considering this wide definition of the multi-omics datasets, we differentiate two major types of application, depending on whether the samples or the molecular features are the primary target of the analysis (Fig. 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 (Fig. 1). This approach is increasingly being adopted in medical studies aiming to improve patient stratification—in either a supervised or an 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 clustering[27] 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 variation[31–34] or consider alternative underlying distributions[7]. For a comprehensive review of unsupervised multi-omics analysis methods, refer to refs. [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 end[35–37]. 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 datasets[40]. For the latter, however, other analytical strategies are more suitable, among which dimension-reduction algorithms[8,26] and regularization techniques (such as Lasso[41–44] or Elastic Net[45] 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 biomarkers[46].

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

[1]Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Valencia, Spain. [2]Microbiology and Cell Science Department, Institute for Food and Agricultural Research, University of Florida, Gainesville, FL, USA. [3]Genetics Institute, University of Florida, Gainesville, FL, USA. [4]Present address: Institute for Integrative Systems Biology, Spanish National Research Council, Valencia, Spain. ✉e-mail: aconesa@ufl.edu

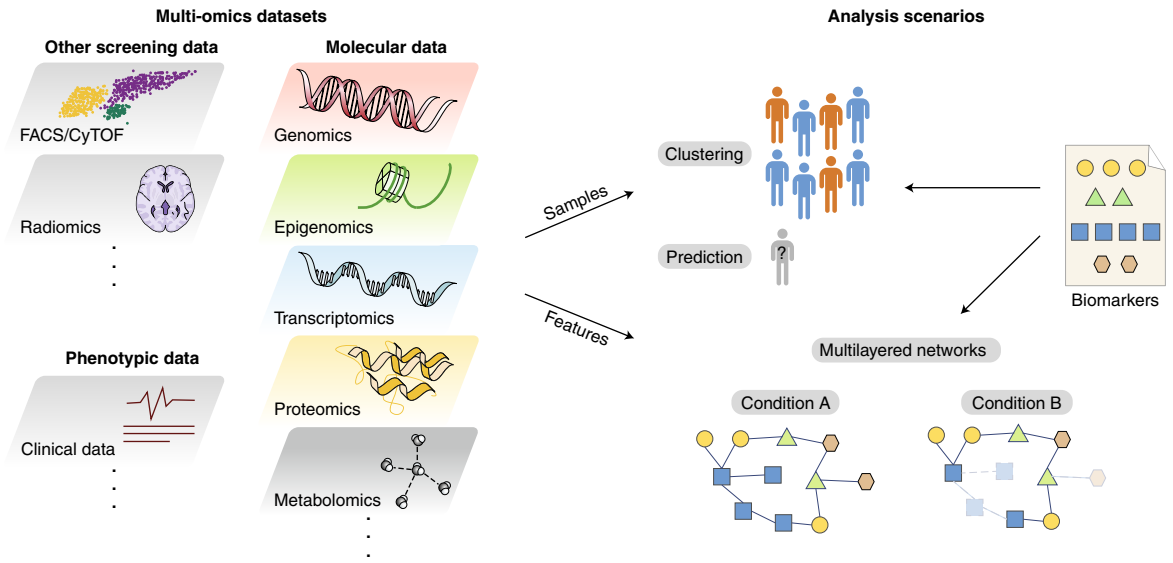

Fig. 1 | Schematic representation of analysis goals in multi-omics studies. Multi-omics datasets may be defined by any combination of molecular profiling data modalities, such as genomics, epigenomics, transcriptomics, proteomics or metabolomics, and include other types of high-throughput data such as 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.

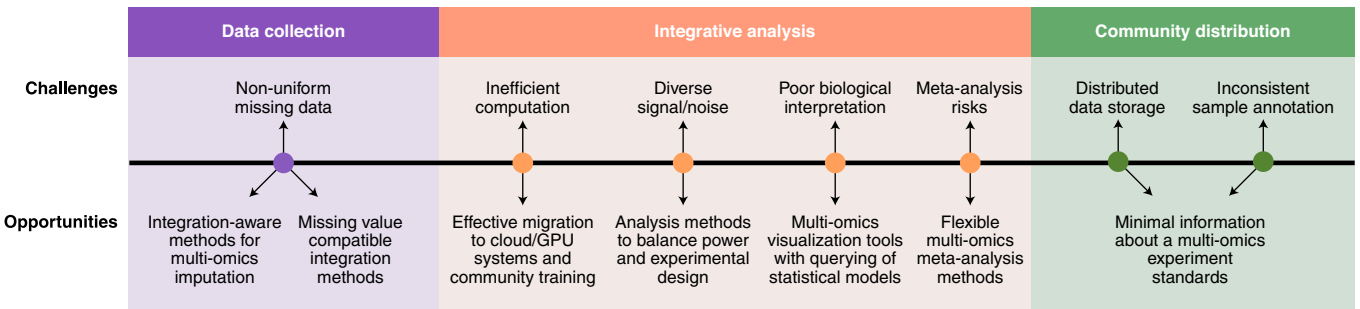

Fig. 2 | Challenges and opportunities in multi-omics data-integration. Analysis challenges together with possible solutions are presented for three different aspects—data collection, integrative analysis and community distribution—of the multi-omic data-integration paradigm.

relationships among them (Fig. 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 (that is, transcriptome-wide association studies, TWAS)[47], while other approaches attempt to sequentially expand across omics layers, performing a chained analysis of regulatory relationships[48]. Regression models have also been used in this context, including, but not limited to, regularized regression[13], regression based on latent variables[8] and structural equation models[49]. In many of these applications, some level of a priori information of the intrafeature relationships, such as genome coordinate 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 modeling, the development of variants that integrate transcriptomics[52] 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 approaches[53].

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. mixOmics[8] 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 CANTARE[54] 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.

Challenges in multi-omics data analysis

Having established the general goals and analysis frameworks in multi-omics studies, we next reflect on 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 multilayered systems models,

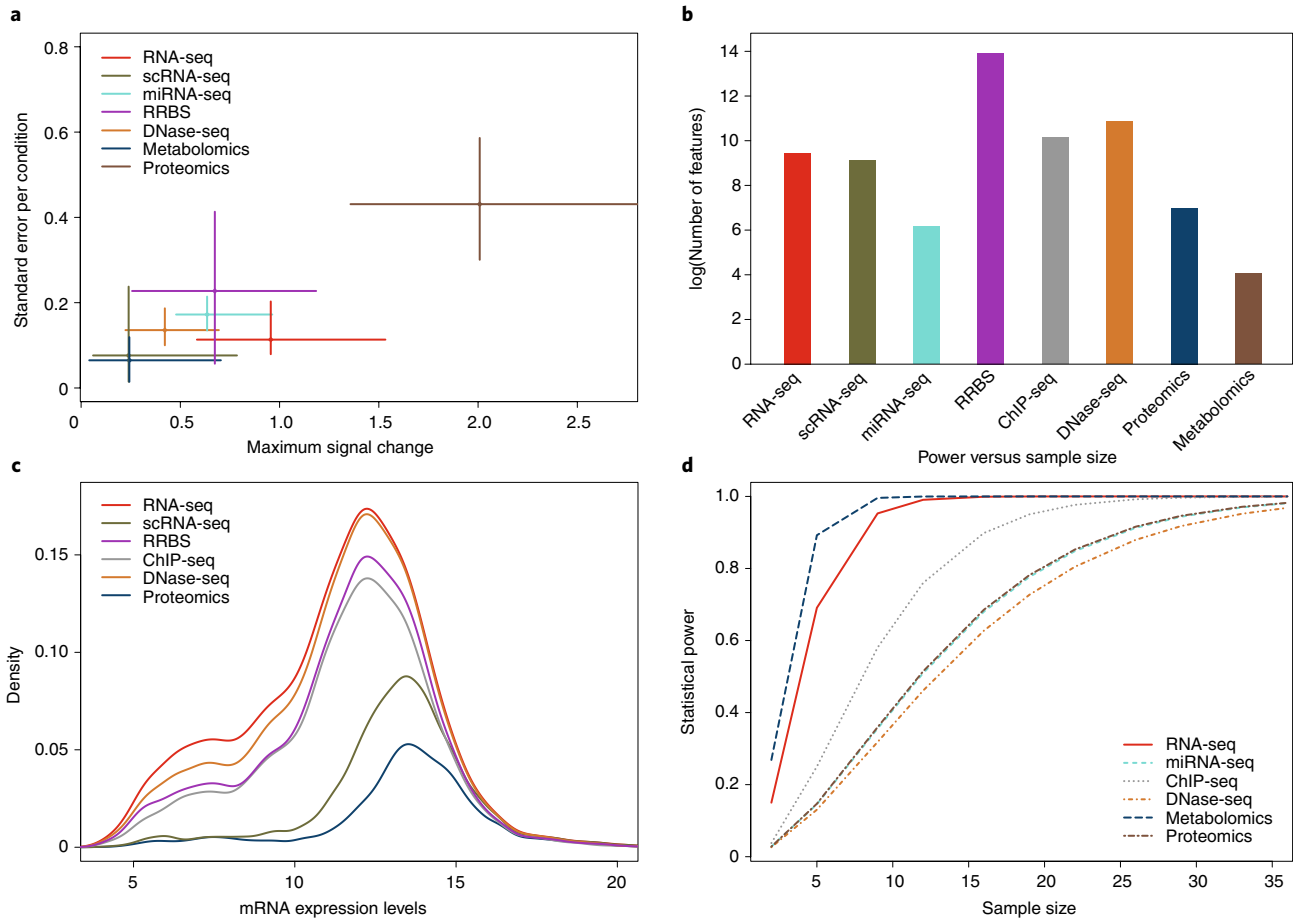

Fig. 3 | Comparison of the properties of omics data types. The STATegra dataset, representing a transcription factor Ikaros-induced mouse B-cell differentiation time course[125], is used as an example. **a**, Signal-to-noise plot. Segments represent the interquartile ranges of the maximum signal change between all pairs of experimental conditions (that is, time points and Ikaros-induced versus control; *x* axis) and the 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 software[55] for statistical power calculations in multi-omics experiments. mRNA, messenger RNA; miRNA-seq, microRNA sequencing; RRBS, reduced representation bisulfite sequencing; DNase-seq, DNase I sequencing.

and the problems pertaining data annotation, storage and computational resources (Fig. 2).

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 (Fig. 3a) is often ignored during multimodal 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 efforts[55]. 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, as 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, chromatin immunoprecipitation sequencing (ChIP-seq) has less sensitivity than RNA sequencing (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. In addition, omics platforms also differ in the number of detected features (Fig. 3b) and degree of completeness achieved when measuring the targeted molecular space (Fig. 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 power[55], which can vary substantially across omics modalities (Fig. 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 developed[55]. 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, where the former is related to the above-described false-negative-association problem, and the latter is related 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.

As discussed above, statistical power refers to the ability to detect a given effect size with sufficient statistical significance using each omic modality. However, integrating different types of measurement 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 single nucleotide polymorphisms (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 potential of the genome variants. Similarly, meta-analysis methods combine *P* values associated with instances of the 'same feature' measured by different omics modalities, such as promoter methylation, gene expression and protein abundance of the same coding gene[58] or pathway[59]. 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 with 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.

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 (for example, 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-read DNA-sequencing methods). While the former can be easily addressed by imputation methods, that is, by leveraging the correlation structure across omics modalities[60], the latter is related to analytical platform coverage[55] and therefore can solely be alleviated by employing alternative technologies, such as long-read sequencing[61]. 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 dataset[62]. 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, MOFA[7] analyzes the latent space across omics types to impute missing samples, while MultiBaC[65] 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 dataset[66] 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.

Interpretability of multi-omics models. One of the most attractive characteristics of multi-omics approaches is their capacity to deliver multilayered 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 independently[67–70] 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 (that is Cytoscape[75], 3Omics[76] 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 multilayered 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 modeled molecular systems. Graphical databases such as Neo4j[77] 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 data[78].

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 them[79–81]. 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 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 the Short Read Archive (SRA)[83], the database of Genotypes and Phenotypes (dbGAP)[84] or the European Genome Archive (EGA) [85] for sequencing data, MetaboLights[86] for metabolomics or ProteomeXchange[87] for proteomics data. Unfortunately, the multi-omics nature of these

Table 1 | Public projects with linked multi-omics data

Study	Available omic types	Scope	Magnitude	Matching level
TCGA[126]	Gene expression, methylation, protein, miRNA	Cancer types	>10,000 samples	Tumor sample
CCLE[127]	Genomics, transcriptomics, miRNAs, DNA methylation, proteomics, histone modifications	Cancer cell lines	~1,000 cell lines	Cell line
GTEx[128]	Genome variation, transcriptomics	Human tissues	~1,000 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, miRNA-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

Non-exhaustive list of public resources that provide multi-omics data measured on the same set of bio-entities. The last column indicates the type of sample where matching data are generated. TCGA, The Cancer Genome Atlas; HiC, genome-wide chromatin conformation capture; WGBS, whole-genome bisulfite sequencing; CCLE, Cancer Cell Line Encyclopedia; GTEx, Genotype-Tissue Expression; TOPMed, Trans-Omics for Precision Medicine; ENCODE, Encyclopedia of DNA Elements; iHEC, International Human Epigenome Consortium.

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 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, and then making them available on data commons[88], although this is not common practice yet. While software is available for the integrated annotation of multi-omics datasets, either for small laboratories[89] or for technology service providers[90,91], these tools are not intended for public data sharing. Efforts such as the European Bioinformatics Institute OmicsDI[92] 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, Zenodo or Lifebit, 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 established[93] and minimal information guidelines have been defined for most omics types, the concept of 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 preprocessing 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.

Performance and scalability challenges. Multi-omics datasets are intrinsically large in terms of both features and samples and will probably 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 studies[97], while single-cell omics[98], 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 demand[100], 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 clusters. However, the broad deployment of high-performance computing 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, as 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 central processing units (CPUs) in cloud computing infrastructures has enabled the development of parallelized algorithmic strategies, but performing an efficient implementation of parallel computation is far from trivial[101]. 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 omic[102]. 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 data[103]. GPUs have become a particularly interesting alternative given the popularization of deep learning methods for the analysis of multi-omics data[9]. 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 parallelization[104]. 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.

Towards a cell-level model of molecular regulation

One of the most promising recent advances in multi-omics has been the development of single-cell multi-omics technologies, which opens up exciting opportunities for the inference of multilayered 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 progression[105], and understand the functionality

of tissues and organs with high levels of cellular specialization, such as the brain[106]. 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 single-cell RNA sequencing (scRNA-seq) with the measurement of one or more additional molecular layers, including DNA variability, chromatin accessibility, DNA methylation and protein abundance (see refs. [98,107] 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 (see ref. [108] 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 clustered regularly interspaced short palindromic repeats (CRISPR) perturbations[109], spatial information[110], lineage reconstruction[111] and time-dependent trajectory data[112].

While single-cell multi-omics data constitutes an exciting opportunity due to the wealth of biological information it provides, integrative analysis faces the same types of challenge 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 manner[113–115], while scATAC-seq (single-cell assay for transposase-accessible chromatin sequencing) suffers from very low read coverage, leading to high sparsity and making data analysis and interpretation challenging[116]. 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 refs. [98,117] 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 refs. [108,117]) have been developed, a number of which are dedicated to single-cell specific data modalities such as trajectory[118] and lineage data[119]. 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[120], eQTLs[121] 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 coming 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 catalog of available single-cell approaches. Furthermore, the potential of combining single-cell omics with cell-level imaging and morphological profiling technologies, such as Cell Painting[123], which 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 (for example, 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, for example, PARADIGM[124]). 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 methodology, provided that other problems—such as sparsity—are addressed. Finally, the incorporation of clinical data from multiple patients into single-cell studies and statistical modeling 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

The data used for Fig. 3 were taken from the publicly available STATegra dataset[125]. Pre-processed values used to generate graphs are provided together with the code.

Code availability

The code and data used to generate Fig. 3 are available on GitHub at https://github.com/ConesaLab/Perspective_Multi-Omics_Integration.

Received: 18 March 2021; Accepted: 17 May 2021;
Published online: 21 June 2021

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 (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 (2010).
3. Song, L. et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* **21**, 1757–1767 (2011).
4. Kim, S., Jhong, J.-H., Lee, J. & Koo, J.-Y. Meta-analytic support vector machine for integrating multiple omics data. *BioData Min.* **10**, 2 (2017).
5. Vaske, C. J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (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 (2017).
7. Argelaguet, R. et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (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 Comput. Biol.* **13**, e1005752 (2017).
9. Zhang, L. et al. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Front. Genet.* **9**, 477 (2018).
10. Ma, T. & Zhang, A. Integrate multi-omics data with biological interaction networks using multi-view factorization autoencoder (MAE). *BMC Genomics* **20**, 944 (2019).
11. Huang, Z. et al. SALMON: survival analysis learning with multi-omics neural networks on breast cancer. *Front. Genet.* **10**, 166 (2019).
12. Bersanelli, M. et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* **17**, 15 (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. *Brief. Bioinform.* **21**, 1904–1919 (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. *Brief. Bioinform.* **21**, 2011–2030 (2020).
15. Meng, C. et al. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* **17**, 628–641 (2016).
16. Buescher, J. M. & Driggers, E. M. Integration of omics: more than the sum of its parts. *Cancer Metab.* **4**, 4 (2016).
17. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).
18. Kristensen, V. N. et al. Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* **14**, 299–313 (2014).
19. Sathyanarayanan, A. et al. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. *Brief. Bioinform.* **21**, 1920–1936 (2020).

20. Zeng, H. et al. Integrative radiogenomics analysis for predicting molecular features and survival in clear cell renal cell carcinoma. *Aging* **13**, 9960–9975 (2021).
21. Kirienko, M. et al. Radiomics and gene expression profile to characterise the disease and predict outcome in patients with lung cancer. *Eur. J. Nucl. Med. Mol. Imaging* https://doi.org/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. *Front. Immunol.* **12**, 590742 (2021).
23. Houle, D., Govindaraju, D. R. & Omholt, S. Phenomics: the next challenge. *Nat. Rev. Genet.* **11**, 855–866 (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. *Plant J. Cell Mol. Biol.* **97**, 112–133 (2019).
25. Zhu, R., Zhao, Q., Zhao, H. & Ma, S. Integrating multidimensional omics data for cancer outcome. *Biostatistics* **17**, 605–618 (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 Biol.* **22**, 39 (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 (2009).
28. Yener, B. et al. Multiway modeling and analysis in stem cell systems biology. *BMC Syst. Biol.* **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. *Chemom. Intell. Lab. Syst.* **104**, 101–111 (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 (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 (2016).
32. O'Connell, M. J. & Lock, E. F. R.JIVE for exploration of multi-source molecular data. *Bioinformatics* **32**, 2877–2879 (2016).
33. Bouhaddani, S. E. et al. Integrating omics datasets with the OmicsPLS package. *BMC Bioinformatics* **19**, 371 (2018).
34. Planell, N. et al. STATegra: multi-omics data integration—a conceptual scheme with a bioinformatics pipeline. *Front. Genet.* **12**, 143 (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. *Comput. Math. Methods Med.* **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 (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 (2016).
38. Wu, C. et al. A selective review of multi-level omics data integration using variable selection. *High Throughput* **8**, 4 (2019).
39. Lagani, V., Kortas, G. & Tsamardinos, I. Biomarker signature identification in 'omics' data with multi-class outcome. *Comput. Struct. Biotechnol. J.* **6**, e201303004 (2013).
40. Le, D.-H. Machine learning-based approaches for disease gene prediction. *Brief. Funct. Genomics* **19**, 350–363 (2020).
41. Fang, H., Huang, C., Zhao, H. & Deng, M. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31**, 3172–3180 (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 (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 (2020).
44. Park, H., Niida, A., Miyano, S. & Imoto, S. Sparse overlapping group lasso for integrative multi-omics analysis. *J. Comput. Biol.* **22**, 73–84 (2015).
45. Singh, A. et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).
46. Patel-Murray, N. L. et al. A multi-omics interpretable machine learning model reveals modes of action of small molecules. *Sci. Rep.* **10**, 954 (2020).
47. Gusev, A. et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.* **50**, 538–548 (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. *Sci. Rep.* **11**, 1907 (2021).
49. Cai, X., Bazerque, J. A. & Giannakis, G. B. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Comput. Biol.* **9**, e1003068 (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 (2009).
51. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248 (2010).
52. Covert, M. W., Schilling, C. H. & Palsson, B. Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.* **213**, 73–88 (2001).
53. Tzika, E., Dreker, T. & Imhof, A. Epigenetics and metabolism in health and disease. *Front. Genet.* **9**, 361 (2018).
54. Siebert, J. C. et al. CANTARE: finding and visualizing network-based multi-omic predictive models. *BMC Bioinformatics* **22**, 80 (2021).
55. Tarazona, S. et al. Harmonization of quality metrics and power calculation in multi-omic studies. *Nat. Commun.* **11**, 3092 (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. *Front. Neurosci.* **14**, https://doi.org/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. *Front. Genet.* **10**, https://doi.org/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 (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 (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 (2016).
61. Kuo, R. I. et al. Normalized long read RNA sequencing in chicken reveals transcriptome complexity similar to human. *BMC Genomics* **18**, 323 (2017).
62. Conesa, A. & Beck, S. Making multi-omics data accessible to researchers. *Sci. Data* **6**, 251 (2019).
63. Dong, X. et al. TOBMI: *trans*-omics block missing data imputation using a *k*-nearest neighbor weighted approach. *Bioinformatics* **35**, 1278–1283 (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**, https://doi.org/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 (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. *Contemp. Clin. Trials* **27**, 506–517 (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. *Hum. Genet.* **126**, 289–301 (2009).
68. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (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 (2016).
70. Welch, R. P. et al. ChIP-enrich: gene set enrichment testing for ChIP-seq data. *Nucleic Acids Res.* **42**, e105 (2014).
71. Canzler, S. & Hackermüller, J. multiGSEA: a GSEA-based pathway enrichment analysis for multi-omics data. *BMC Bioinformatics* **21**, 561 (2020).
72. Long, Y., Lu, M., Cheng, T., Zhan, X. & Zhan, X. Multiomics-based signaling pathway network alterations in human non-functional pituitary adenomas. *Front. Endocrinol.* **10**, https://doi.org/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 Res.* **46**, W503–W509 (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 Res.* **39**, D677–D684 (2011).
75. Su, G., Morris, J. H., Demchak, B. & Bader, G. D. Biological network exploration with Cytoscape 3. *Curr. Protoc. Bioinformatics* **47**, 8.13.11–18.13.24 (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**, https://doi.org/10.1186/1752-0509-7-64 (2013).

77. Miller, J. J. Graph database applications and concepts with Neo4j. In *Proc. Southern Association for Information Systems Conference* (AIS, 2013).
78. Yoon, B.-H., Kim, S.-K. & Kim, S.-Y. Use of graph database for the integration of heterogeneous biological data. *Genomics Inform.* **15**, 19–27 (2017).
79. Consortium, T. I. Hi. 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 (2014).
80. *ICGC Data Portal* (The International Cancer Genome Consortium, 2021); https://dcc.icgc.org/
81. *Human Microbiome Project Data Portal* (Human Microbiome Project, 2021); https://portal.hmpdacc.org/
82. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
83. Kodama, Y., Shumway, M. & Leinonen, R. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).
84. Tryka, K. A. et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.* **42**, D975–D979 (2014).
85. Lappalainen, I. et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* **47**, 692–695 (2015).
86. Haug, K. et al. MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.* **48**, D440–D444 (2020).
87. Deutsch, E. W. et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* **45**, D1100–D1106 (2017).
88. Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, practical genomic data sharing that accelerates research. *Nat. Rev. Genet.* **21**, 615–629 (2020).
89. Hernandez-de-Diego, R. et al. STATegra EMS: an experiment management system for complex next-generation omics experiments. *BMC Syst. Biol.* **8**, S9 (2014).
90. Lin, K. et al. MADMAX—management and analysis database for multiple ~omics experiments. *J. Integr. Bioinform.* **8**, 59–74 (2011).
91. Venco, F., Vaskin, Y., Ceol, A. & Muller, H. SMITH: a LIMS for handling next-generation sequencing workflows. *BMC Bioinformatics* **15**, S3 (2014).
92. Perez-Riverol, Y. et al. Discovering and linking public omics data sets using the Omics Discovery index. *Nat. Biotechnol.* **35**, 406–409 (2017).
93. Chervitz, S. A. et al. in *Bioinformatics for Omics Data: Methods and Protocols* (ed. Mayer, B.) 31–69 (Humana Press, 2011).
94. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nat. Rev. Genet.* **19**, 299–310 (2018).
95. van Karnebeek, C. D. M. et al. The role of the clinician in the multi-omics era: are you ready? *J. Inherit. Metab. Dis.* **41**, 571–582 (2018).
96. Angione, C. Human systems biology and metabolic modelling: a review—from disease metabolism to precision medicine. *Biomed. Res. Int.* **2019**, 8304260 (2019).
97. Hériché, J.-K., Alexander, S. & Ellenberg, J. Integrating imaging and omics: computational methods and challenges. *Annu. Rev. Biomed. Data Sci.* **2**, 175–197 (2019).
98. Lee, J., Hyeon, D. Y. & Hwang, D. Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.* **52**, 1428–1442 (2020).
99. Stein, L. D. The case for cloud computing in genome informatics. *Genome Biol.* **11**, 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. *Brief. Bioinform.* **22**, 66–76 (2020).
101. Solomonik, E., Carson, E., Knight, N. & Demmel, J. Trade-offs between synchronization, communication, and computation in parallel linear algebra computations. *ACM Trans. Parallel Comput.* **3**, 1–47 (2016).
102. Berger, B., Peng, J. & Singh, M. Computational solutions for omics data. *Nat. Rev. Genet.* **14**, 333–346 (2013).
103. Alyass, A., Turcotte, M. & Meyre, D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med. Genet.* **8**, 33 (2015).
104. Chen, X.-W. & Lin, X. Big data deep learning: challenges and perspectives. *IEEE Access* **2**, 514–525 (2014).
105. Fan, J., Slowikowski, K. & Zhang, F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp. Mol. Med.* **52**, 1452–1465 (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 (2021).
107. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nat. Methods* **17**, 11–14 (2020).
108. Forcato, M., Romano, O. & Bicciato, S. Computational methods for the integrative analysis of single-cell data. *Brief. Bioinform.* **22**, 20–29 (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 (2016).
110. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
111. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (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**, 381–386 (2014).
113. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
114. Darmanis, S. et al. Simultaneous multiplexed measurement of RNA and proteins in single cells. *Cell Rep.* **14**, 380–389 (2016).
115. Peterson, V. M. et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35**, 936–939 (2017).
116. Chen, H. et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **20**, 241 (2019).
117. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
118. Welch, J. D., Hartemink, A. J. & Prins, J. F. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol.* **18**, 138 (2017).
119. Campbell, K. R. et al. clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biol.* **20**, 54 (2019).
120. Fan, J. et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* **28**, 1217–1227 (2018).
121. Van Der Wijst, M. G. P. et al. Single-cell RNA sequencing identifies celltype-specific *cis*-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
122. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
123. Bray, M.-A. et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (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* **29**, i62–i70 (2013).
125. Gomez-Cabrero, D. et al. STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Sci. Data* **6**, 256 (2019).
126. Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: creating lasting value beyond its data. *Cell* **173**, 283–285 (2018).
127. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
128. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
129. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
130. Moore, J. E. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
131. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
132. Mergner, J. et al. Mass-spectrometry-based draft of the *Arabidopsis* proteome. *Nature* **579**, 409–414 (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 (2005).

Acknowledgements

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

Author contributions

A.C. drafted the structure of the manuscript and integrated author contributions. A.A.-L., S.T. and A.C. drafted the manuscript, reviewed the literature, contributed to figures and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to A.C.

Peer review information *Nature Computational Science* thanks Casey Greene and Terry Speed for their contribution to the peer review of this work. Handling editor: Fernando Chirigati, in collaboration with the *Nature Computational Science* team.