

EDUCATION

Ten quick tips for avoiding pitfalls in multi-omics data integration analyses

Davide Chicco^{1*}, Fabio Cumbo², Claudio Angione³

1 Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada, **2** Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, United States of America, **3** School of Computing Engineering and Digital Technologies, Teesside University, Middlesbrough, United Kingdom

* davidechicco@davidechicco.it

Abstract

Data are the most important elements of bioinformatics: Computational analysis of bioinformatics data, in fact, can help researchers infer new knowledge about biology, chemistry, biophysics, and sometimes even medicine, influencing treatments and therapies for patients. Bioinformatics and high-throughput biological data coming from different sources can even be more helpful, because each of these different data chunks can provide alternative, complementary information about a specific biological phenomenon, similar to multiple photos of the same subject taken from different angles. In this context, the integration of bioinformatics and high-throughput biological data gets a pivotal role in running a successful bioinformatics study. In the last decades, data originating from proteomics, metabolomics, metagenomics, phenomics, transcriptomics, and epigenomics have been labelled *-omics* data, as a unique name to refer to them, and the integration of these omics data has gained importance in all biological areas. Even if this omics data integration is useful and relevant, due to its heterogeneity, it is not uncommon to make mistakes during the integration phases. We therefore decided to present these ten quick tips to perform an omics data integration correctly, avoiding common mistakes we experienced or noticed in published studies in the past. Even if we designed our ten guidelines for beginners, by using a simple language that (we hope) can be understood by anyone, we believe our ten recommendations should be taken into account by all the bioinformaticians performing omics data integration, including experts.

OPEN ACCESS

Citation: Chicco D, Cumbo F, Angione C (2023) Ten quick tips for avoiding pitfalls in multi-omics data integration analyses. *PLoS Comput Biol* 19(7): e1011224. <https://doi.org/10.1371/journal.pcbi.1011224>

Editor: Francis Ouellette, McGill University, CANADA

Published: July 6, 2023

Copyright: © 2023 Chicco et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: CA was funded by a Network Development Award and by Turing Network Funding from the Alan Turing Institute (grants TNDC2-10022 and D-ELA-013). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare they have no conflict of interest.

Introduction

Integration of omics data is a pillar of bioinformatics: Incorporating data of genomics, proteomics, metabolomics, metagenomics, phenomics, transcriptomics, epigenomics, and other *-omics* areas in a unique database, in fact, can provide a larger picture of a specific biological aspect and, therefore, facilitate the discovery of more relevant, interesting, and solid scientific results. Just like in photography, where photos of the same subject taken from different angles can provide different perspectives and pieces of information about the same phenomenon, bioinformatics data chunks of different types and coming from different sources can be more informative than a single-source dataset.

Multi-omics data integration, however, can bring several problems, especially if performed by beginners or apprentice researchers. For example, apprentice bioinformaticians sometimes produce multi-omics resources based solely on their perspective, without taking into account what the analysts would really need. The final database then becomes difficult to use and gets underutilized in the bioinformatics community, possibly generating a waste of effort, time, energy, and funds. Or, sometimes, apprentice bioinformaticians do not give enough importance to the metadata, only realizing too late that their biological data do not have enough descriptive metadata to be used broadly by the scientific community. To avoid these and other pitfalls and common mistakes, we propose this study where we describe a few guidelines to keep in mind when performing multi-omics data integration. We designed these quick tips for data curators, biomedical data scientists, machine learning analysts, computational biologists, bioinformaticians, and students who are going to perform a multi-omics data integration phase to produce an omics data resource to be used by analysts.

The potentials and advantages of multi-omics data integration were described in several studies. Sijia Huang and colleagues [1], for example, reported the recent progress in the field by explaining the most common and successful techniques for this scope. A more recent article by Indhupriya Subramanian and colleagues [2], on another hand, provided a detailed and thorough overview of the multi-omics general situation. Examples of effective multi-omics integration tools are mixOmics [3] in R and INTEGRATE [4] in Python. Sebastian Canzler and coauthors [5] proposed some perspective recommendations and challenges related to toxicology, while Mingon Kang and colleagues [6] proposed a roadmap for omics integration using deep learning.

In the *PLOS Computational Biology* education collection, no study provided recommendations for bioinformatics data integration so far. A study by Ramon Diaz-Uriarte and colleagues [7] proposed ten quick tips for biomarker discovery and its validation that, although interesting, does not regard multi-omics data integration directly. We fill this gap by presenting our current quick tips for multi-omics data integration as summarized in Fig 1.

Several studies follow the same pipeline described here while dealing with multi-omics data. A striking example is the research study by Eleonora Cappelli and colleagues [8] whose aim is to combine DNA methylation and RNA sequencing data that the authors used to train and test a supervised classification model for identifying disease-specific biomarker genes. In particular, they focused their analysis on three different types of cancer: breast invasive

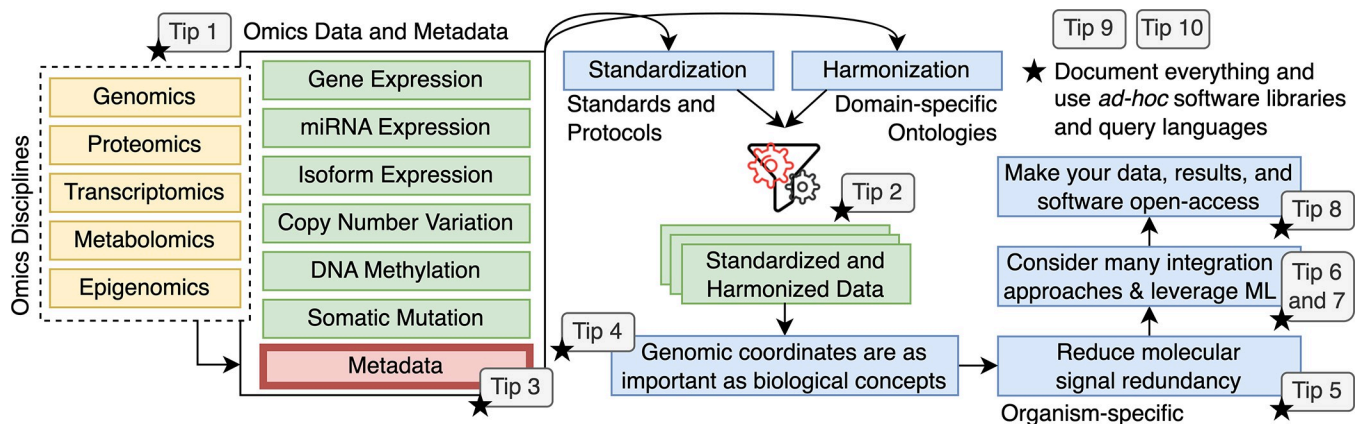


Fig 1. Schematization of the proposed tips as a flow chart that shows all the steps we suggest to follow in order to facilitate the integration and analysis of multi-omics data. Every tip, from the first up to the seventh, are marked with a black star, meaning that we suggest documenting everything and using ad hoc software libraries and query languages as suggested in Tip 9 and Tip 10.

<https://doi.org/10.1371/journal.pcbi.1011224.g001>

carcinoma (BRCA), thyroid carcinoma (THCA), and kidney renal papillary cell carcinoma (KIRP). The authors retrieved the data of these cancer types from the TCGA2BED [9] database, which provides all the publicly available data of the TCGA program standardized into the free-BED format. The authors integrated DNA methylation data with the beta values of the methylated CpG island among with RNA sequencing data with the expression of genes, by joining data based on common genomic coordinates. That is, if the RNA sequencing data contains the expression of specific genomic regions that refer to the genes, the specific single-nucleotide positions in the genomic regions of the genes can be methylated. The authors eventually analysed these data with different tree- and rule-based supervised classification algorithms (for example, C4.5 [10], Random Forests [11], RIPPER [12], and CAMUR [13]) producing over 15,000 classification models (in the form of gene sets) able to discriminate case and control samples with an accuracy of 95% on average. Most importantly, the authors documented every step of their analysis and made their software code openly available along with their results.

Tip 1: Design the integrated data resource from the perspective of the users, not from the perspective of the data curators

Integrating multi-omics data is a hard job, and several aspects should be taken into account. While you work on the integration of different data, coming from different data sources, perhaps having different formats and different origins, you might be tempted to see the whole project only from the point of view of the data curators: you and your colleagues working on it. But if you only consider the perspective of the data curators, the bioinformatics resource, when ready, would be eventually optimized only for data curators, and not for analysts and users who would take advantage of it. This would bring several drawbacks, making your bioinformatics data resource difficult to use for analysts and researchers, perhaps making it unnoticed or forgotten by the scientific community.

Instead, we suggest always keeping in mind the perspective and the point of view of the users and the analysis who would ultimately exploit the integrated bioinformatics data resource. Of course, we know that it is easier to write it than to do it, but this aspect can be the single most important aspect of a multi-omics data integration project, which can mean the difference between failure and success.

We therefore recommend designing some real use case scenarios in which users can exploit the bioinformatics data resource to solve a real scientific problem. Pretend you are the analyst who needs to solve a particular biomedical problem and would like to use your bioinformatics resource. What would you need? What is missing? What is difficult to do? What could be improved? Answering these questions precisely and thoroughly would help you make a better, improved multi-omics integrated resource.

ENCODE [14] is a great example of a popular, documented, and useful multi-omics data integration project designed from the perspective of the users. Moreover, we also suggest applying existing integrative methods to the data to get the full user experience.

Tip 2: Preprocess your data: Standardise and harmonise it

Standardising raw data helps to ensure that data from different omics technologies are compatible since they all have their own specific characteristics (for example, different measurement units, etc.). This process can involve a variety of different steps, such as normalizing data to account for differences in sample size or concentration, converting data to a common scale or unit of measurement, removing technical biases or artifacts, and filtering data to remove outliers or low-quality data points.

For small- and medium-scale studies, storing the raw data is important to ensure the full reproducibility of the results [15]. Giving access to the raw instrumentation data mitigates the issue that processing steps may vary, and allows researchers to make preprocessing assumptions that are appropriate for the selected downstream analysis.

When collecting multi-omics data, it is important to consider a sample size that can provide enough statistical power, and generate replicates, documentation, and project metadata, together with proper data management practices. In addition, the data need to be collected in a way that removes any possible sampling bias [16]. For preprocessed data, it is good practice to include full descriptions of the samples, equipment, and software used.

Data formats of multi-omics can vary widely, even within the same study. Therefore, for compatibility with machine learning or statistical analysis methods, further processing is often needed to unify the format, for example, n-by-k samples-by-feature matrix. This often needs also standard steps like normalization and batch effect correction [17,18].

Standardization and harmonization of data and metadata are key steps in multi-omics data integration because they help to ensure that data can be accurately and consistently interpreted and analyzed.

Standardization refers to the process of ensuring that data are collected, processed, and stored in a consistent manner, using agreed-upon standards and protocols. Lots of tools for standardising omics data have been developed over the last decade [9,19–21] in order to make the data comparable across different studies and platforms, in addition to make it easier to integrate and analyze data from multiple sources.

On the other hand, harmonization refers to the process of aligning data from different sources so that they can be integrated and analyzed together. This typically involves mapping data from different sources onto a common scale or reference and may involve the use of domain-specific ontologies or other standardized data formats [22–25]. Nikolai Russkikh and coauthors [26], for example, employed a style transfer method based on conditional variational autoencoders for RNA-seq data harmonization.

Moreover, it is important to describe precisely the preprocessing and normalisation techniques used in the project documentation and in the article associated with the project. This information would usually be inserted in the supplementary material of a scientific paper. If you have the authorization to release the data, we recommend releasing both the raw data and the preprocessed data in public repositories (Tip 8). Some users, in fact, might be interested in analyzing the raw data, depending on the aim of their projects.

Tip 3: Value your data with metadata

Metadata are simply data that describe the main data. When a photographer takes a photo with a modern camera, for example, the camera not only saves the photo itself, but also records additional details such as lenses used, time and date at which the picture was taken, focal length, image resolution, and color profiles [27]. All these data are the metadata of that photo. These metadata are not the photo, but rather describe the photo, and they can facilitate image processing, image search, and image retrieval [27]. Of course, the photo is still the main protagonist of photography, but the role of the photo's metadata is pivotal: They are the first pieces of the documentation of that photo and will be used by the photographer in several ways. If the metadata were absent, it would be complicated and almost impossible for the photographer to use it. Therefore, metadata are as important as data, not only in photography but in any field. Also in bioinformatics data integration, of course, metadata have an extremely important role. We therefore recommend paying particular attention to the curation of the metadata [28]. Any relevant information regarding a data element should be recorded in the metadata.

Examples of bioinformatics metadata can be found in multiple datasets. The GSE45255 dataset, for example, contains microarray gene expression of samples of patients with breast tumors [29,30] and is available on Gene Expression Omnibus (GEO). In bash, this dataset can be found through the software package `geoCancerPrognosticDatasetsRetriever` [31] and in R it can be downloaded through `geneExpressionFromGEO` [32].

This public dataset contains not only the gene expression data of the patients' samples, but also relevant information about each patient, recorded in metadata [30]: lymph node status, estrogen receptors status, progesterone receptor status, human epidermal growth factor receptor 2 (HER2) status, histological grade, size in millimetres, adjuvant treatment, chemotherapy, recurrence or death from breast cancer, distant metastasis or death from breast cancer, and death from breast cancer. Of course, these metadata help researchers perform better scientific analyses on these datasets, allowing them to make additional discoveries regarding this dataset and therefore regarding breast cancer.

Moreover, ready-to-use data might contain outliers or unexpected data instances [33], and in those cases, the availability of metadata can be necessary to understand what actually happened in the generation of those data elements. Several approaches can be used to include metadata in bioinformatics data [34]; an interesting technique is to use data of different types from the principle dataset. For example, in metabolomics [35], Oliver Fiehn and colleagues [36] added mass spectrometry metadata to the physiological, clinical, and genomic data. Some bioinformatics teams developed structured metadata management tools, such as `BioSamples` [37] and `medna-metadata` [38].

Regarding metadata curation and its importance, it is relevant to mention the initiatives of Susanna-Assunta Sansone, a researcher who has advocated for the standardization and the structured curation of metadata for several decades through multiple resources: FAIRsharing [39], Collaborative Open Plant Omics (COPO) [40], Investigation/Study/Assay (ISA) Metadata Framework [41], and machine-actionable metadata models [42], just to mention a few.

Tip 4: Take into account the genomic coordinates of the data, and not only the biological concepts

The integration of omics data is the process of combining data from different omics technologies (such as genomics, transcriptomics, proteomics, and metabolomics) to gain a more comprehensive understanding of a biological system. It is a powerful approach that can provide insights into the molecular mechanisms underlying complex diseases, identify potential therapeutic targets, and improve our understanding of fundamental biological processes.

We therefore recommend not to focus only on the anatomical sites of your samples or the disease that affects the host and always take into account the genomic regions of your data [43–46]. We refer to anatomical sites, tissues, organs, and diseases as biological concepts in this study. Considering the genomic regions helps in the identification of specific mechanisms underlying the regulation of gene expression and protein function, which can be crucial for understanding the molecular basis of diseases and identifying potential therapeutic targets.

Always use genome annotation tools that allow mapping your data onto specific genomic regions [43,47–49]. However, it is important to be careful when selecting genome annotation tools, as the choice of reference genome can have a significant impact on the results of your analyses. Different versions of the reference genome can have different sets of annotated features, such as genes, regulatory elements, and structural variations, and these differences can affect the interpretation of your data. For example, if you use a newer version of the reference genome that includes additional annotated features, you may identify different sets of genes or regulatory elements as being differentially expressed or altered in your samples. Therefore, it is

important to carefully consider the version of the reference genome that you use when selecting genome annotation tools. In general, it is recommended to use the most up-to-date version of the reference genome that is available, as this will provide the most complete and accurate set of annotated features. However, it is also important to ensure that the version of the reference genome is appropriate for your data.

Tip 5: Control for molecular signal redundancy via variable selection

The process of integrating data from different omic sources, commonly referred to as data fusion, has two major advantages: It preserves the original features of the data, and its flexibility allows for the mixing of data from multiple sources. However, omics signal redundancy needs to be carefully taken into account [3]. For instance, transcriptomics and proteomics are often (weakly) correlated [50], and methods that do not take into account (and filter appropriately) cross-omics information may perform the integration based on redundant signal. Normally, the redundancy depends on the organism being considered, and therefore the level of correction needs to be considered at an organism-specific level.

One solution is reducing the number of variables in an omic-specific or cross-modal way, which generates a more usable representation. With the advent of machine learning tools, supervised techniques can perform such filtering if appropriate clinical/output labels are available, for example, with a feature importance analysis. In general, selecting features decreases the level of noise and can be used as a method to balance the number of features across the omics. Furthermore, this also reduces the risk of overfitting. In the presence of a large dataset, it is also worth considering whether the redundancy can be addressed a priori, for instance, using a clinical stratification approach [51].

Variable selection can also be performed with unsupervised techniques, in a single-modal or multi-modal fashion, with tools like MOFA [52], JIVE [53], and sparse PLS [54]. Acharjee and colleagues [55], for example, took advantage of retention time-dependent clustering to remove the signal redundancy from the metabolite data involved in the omics integration. Similarly, Cao and Gao [56] recently proposed an integration method to mitigate the redundancy in single-cell data by explicitly modelling the regulatory interactions taking place across the omics layers.

Tip 6: Try different integration approaches

When integrating multi-omics data via machine learning models, it is important to note that the technique to be adopted depends on the dataset and the task at hand and cannot usually be decided a priori [57]. A simple concatenation of features across the omics (*early integration*) is a viable approach but is likely to generate enormous matrices, outliers, highly correlated variables, noise, and other difficulties.

Intermediate integration is a viable alternative, in which the idea is to jointly integrate the features across the omics without prior omic-specific processing. The advantage of this approach is the possibility to process the features based on their redundancy or complementarity both within each omic and across the different omics [58]. Intermediate integration is based on the assumption that all the omics can be mapped onto a shared latent space. Therefore, disparities among the omics may bring challenges and lead to an imbalanced learning process. Furthermore, it often depends on unsupervised matrix factorization, which has difficulty incorporating substantial amounts of preexisting biological knowledge.

Late integration is another option, often based on ensemble machine learning methods. Specifically, a model is first trained for each omic to perform the prediction independently, and then the predictions achieved from each omic are combined via averaging or voting. Late

integration may be appropriate when the predictive performance is unbalanced across the omics, for instance, if one omic is significantly more predictive than others, but the integration with the other omics still improves the overall performance. However, it does not directly integrate the data and may overlook cross-omics relationships [59]. Finally, hierarchical approaches or mixed integration strategies are also potential alternatives [60].

Tip 7: Prepare your data for multi-omics data integration with machine learning

Several supervised and unsupervised machine learning techniques have been successfully employed for multi-omic data analysis and integration [61]. Supervised learning trains a model by using preassigned labels for each sample, for instance, the subtype of a given cancer (classification problem) or the overall survival probability of a patient (regression problem). The model is trained using the data at hand, but in a way that does not overfit the same data, as the goal is then to use the trained model for achieving accurate predictions when new or “unseen” samples are considered. Conversely, unsupervised learning can be applied when labels are not available and can be used to reduce the dimensionality of the dataset or detect patterns or clusters within the samples. However, in complex phenotypes like cancer, where the interaction between events spanning multiple omics layers is likely to be the main cause of progression, traditional data-driven multi-omics methods based on machine learning are only able to uncover associations among genes, proteins, or other omics components, without offering a mechanistic interpretation [62].

In this regard, it is possible to use systems biology techniques for omics data integration and to provide further mechanistic knowledge to be incorporated into machine learning approaches. While recognising the role of individual components within a biological system, systems biology follows the notion that “the whole is greater than the sum of its parts.” Specifically, it aims at investigating a biological system as a whole and with an integrated approach, namely by considering the interactions between different components as a way to enrich and explain the behaviour of each component [63].

Among such tools, genome-scale metabolic models (GSMMs) are mathematical reconstructions of metabolic networks that can be used as scaffolds for further omics data integration, therefore generating patient- or condition-specific models that achieve more accurate predictions of disease phenotypes [25,64–66].

In the machine learning era, the effective integration of data-driven and knowledge-driven approaches is increasingly being recognized as key to improving the outcome of omics integration studies, for example, biomarker prediction or phenotype characterisation [67]. For instance, adding features derived or processed through modelling techniques can incorporate knowledge into the model and allow omics data interpretation on a mechanistic or phenotypic level, rather than merely on a data-driven level [68,69].

To ensure the biological interpretability of the results, it is important to focus on methodological advances that can combine multi-omics integration and knowledge extraction with modelling techniques [70,71]. Importantly, since omics data can be quantified numerically and in a condition-, tissue-, and patient-specific way (for example, transcriptomic profiles, protein levels, and metabolite concentrations), using such models within machine learning pipelines can filter out some of the redundancy inherently present in genome-scale omics data. The long-term goal is to find the trade-off between selecting features via machine learning or data-driven approaches only (which has no biological rationale), and using biology-informed approaches (which is likely to lead to suboptimal machine learning performance) [72,73].

Tip 8: Use open science best practices

When starting a new computational biology project, one often has the possibility to decide which programming languages, software platforms, and data query languages to use. Similarly to what we recommended for machine learning [74], pathway enrichment analysis [75], data cleaning and feature engineering [33], and medical image analysis [76], we advocate for using only open-source computer languages and software programs.

Open-source programming languages (such as R or Python), open-source software platforms (Bioconductor [77], Bioconda [78] and Galaxy [79], Anvi'o [80]), open-source data query languages (SQL), and open-source relational database management systems (PostgreSQL and SQLite) can bring several advantages to your multi-omics project, compared to proprietary software.

Open-source material, in fact, can be shared easily among colleagues and collaborators, without worrying about licenses. Moreover, open-source technology can be updated and upgraded easily and most of the time free of cost and can be reutilized in several other projects. If one needs to switch labs, institutes, or jobs, one can take their software and code with themselves in the new environment.

For operating systems, we suggest Linux Ubuntu; for distributed systems, we recommend Apache Spark; for office productivity software, we advocate for LibreOffice. Regarding software code sharing, we advise sharing your code online openly on public epositories such as GitHub, GitLab, and Bitbucket, for example. You can take all the scripts that you developed for your multi-omics integration and release them publicly on the internet. This practice would allow the reproducibility of your work and let other users around the world find possible mistakes in your analyses, allowing you to correct them and ultimately generate better, more solid results.

Once one has spent several months and energy working on a multi-omics integrated data resource, of course we suggest releasing it publicly online, following the findability, accessibility, interoperability, and reusability (FAIR) principles [81]. A researcher can release their data on free open platforms such as Kaggle [82], University of California Irvine Machine Learning Repository [83], FigShare [84], or Zenodo [85]. One can also consider creating and releasing their own data repository [86,87].

Releasing data online would permit other researchers in the world to analyze them and therefore to make new scientific discoveries through secondary analyses [88,89]. The more available the data are, the more secondary studies will be carried out, the higher impact a dataset can have, also in terms of article citations.

Regarding the publication, once the study manuscript is ready for submission to a scientific journal, we suggest releasing it as a preprint on bioRxiv, medRxiv, or arXiv. Moreover, if one has the chance to choose which scientific journals to submit their articles to, we advocate for open-access journals. Open-access articles, in fact, can be freely read and accessed by anyone in the world, including high school students and researchers from developing countries. Open-access journals can be found on the Scimago Journal Ranking website [90].

Tip 9: Use ad hoc software libraries and query languages, do not develop new scripts on your own

There are several reasons why you might consider using ad hoc software libraries and query languages for integrating omics data than developing new scripts on your own:

1. Time and resources: developing your own scripts for integrating omics data can be a time-consuming and resource-intensive process, especially if you are not familiar with the

specific programming languages and tools that are commonly used in the bioinformatics domain;

2. Accuracy and reliability: ad hoc software libraries and query languages are usually developed and maintained by experts in the field and have been extensively tested and validated. This means that you can usually be confident in the accuracy and reliability of the results produced by these tools;
3. Community support: when you use consolidated tools, you can benefit from the support of a larger community of users who are familiar with the tools and can provide guidance and help if you face any issues. This can be extremely helpful especially if you are new to these tools;
4. Compatibility with other tools: ad hoc software libraries and query languages are often designed to be compatible with other tools and resources in the field. This can make it easier to integrate your results with other data sources or to use the tools in conjunction with other software;
5. Easy of use: ad hoc software libraries and query languages are often designed with usability in mind and may offer user-friendly interfaces and documentation to help you get started. This can make it easier for you to use the tools, even if you are not an experienced programmer.

Overall, keep in mind the aforementioned points before writing your own scripts. There are tons of open-source libraries and tools for roughly every kind of omics data integration and analysis distributed over public repositories like the Python Package Index [91] and Bioconductor [77] with Python and R packages, respectively, as well as Bioconda [78] and the Galaxy ToolShed [92], which extremely simplify the process of distributing and effectively using such kind of software tools for managing, integrating, and analysing multi-omics data. These factors can make it more efficient and effective for you to integrate omics data and can help to ensure the quality and reliability of your results.

Tip 10: Document everything

We already mentioned the importance of metadata, data that describe the principle data. Metadata can also be seen as a structured form of documentation.

In scientific research, well-written documentation and reporting are as important as the scientific discoveries themselves [93]. Also in software development, well-written documentation is as important as the software itself [94].

Documentation should be as detailed as possible: addressed both to developers who want to redo the integration analysis and to users that just want to learn more about the whole bioinformatics process.

The same rule for scientific research and software development is true for bioinformatics data integration: document everything. Document how you obtained the data, how you integrated them and why, which data sources you selected and why, which technologies you decided to use and why, and so on. This documentation will be invaluable for your colleagues and collaborators, but also for your future self that will need to recover some information about the data integration resource. The documentation will also be pivotal for writing the article on the study.

The documentation should be written following some precise standards [95] and not only at the end of the project, but during its development. Ongoing documentation should be recorded in a notebook or scientific diary [96]. Good examples of documentation can be

found on the Bioconductor website: for example, the tutorial for the usage of OMICsPCA R software package [97].

Conclusions

Integration of multi-omics data is a key aspect of bioinformatics, since assembling data from different sources can surely enrich the scientific description of a phenomenon. Integrated data derived from different sources can then be used for computational analysis through machine learning or biostatistics methods and eventually lead to better, more solid results and outcomes. Multi-omics data integration, however, can suffer from many pitfalls and common mistakes, which sometimes might even go unnoticed, and which might undermine or even corrupt the final scientific results of the analysis.

To prevent these common mistakes, we present here our ten quick tips for multi-omics data integration, which we designed for any data curator, biomedical data scientist, machine learning analyst, computational biologist, bioinformatician, and student who wants to perform these steps to generate an omics data resource. We believe that our simple guidelines, if followed correctly, can improve the quality of the multi-omics data integration phase and therefore help generate better results, which can help us better understand the underlying biology of the system being studied.

Author Contributions

Conceptualization: Davide Chicco, Fabio Cumbo, Claudio Angione.

Data curation: Davide Chicco.

Formal analysis: Davide Chicco, Fabio Cumbo, Claudio Angione.

Investigation: Davide Chicco, Claudio Angione.

Methodology: Davide Chicco, Fabio Cumbo, Claudio Angione.

Project administration: Davide Chicco.

Resources: Davide Chicco.

Supervision: Davide Chicco.

Writing – original draft: Davide Chicco, Fabio Cumbo, Claudio Angione.

References

1. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet.* 2017; 8:84. <https://doi.org/10.3389/fgene.2017.00084> PMID: 28670325
2. Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights.* 2020; 14:1177932219899051. <https://doi.org/10.1177/1177932219899051> PMID: 32076369
3. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLoS Comput Biol.* 2017; 13(11):e1005752. <https://doi.org/10.1371/journal.pcbi.1005752> PMID: 29099853
4. Di Filippo M, Pescini D, Galuzzi BG, Bonanomi M, Gaglio D, Mangano E, et al. INTEGRATE: model-based multi-omics data integration to characterize multi-level metabolic regulation. *PLoS Comput Biol.* 2022; 18(2):e1009337. <https://doi.org/10.1371/journal.pcbi.1009337> PMID: 35130273
5. Canzler S, Schor J, Busch W, Schubert K, Rolle-Kampczyk UE, Seitz H, et al. Prospects and challenges of multi-omics data integration in toxicology. *Arch Toxicol.* 2020; 94(2):371–388. <https://doi.org/10.1007/s00204-020-02656-y> PMID: 32034435
6. Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. *Brief Bioinform.* 2022; 23(1):bbab454. <https://doi.org/10.1093/bib/bbab454> PMID: 34791014

7. Diaz-Uriarte R, Gómez de Lope E, Giugno R, Fröhlich H, Nazarov PV, Nepomuceno-Chamorro IA, et al. Ten quick tips for biomarker discovery and validation analyses using machine learning. *PLoS Comput Biol*. 2022; 18(8):e1010357. <https://doi.org/10.1371/journal.pcbi.1010357> PMID: 35951526
8. Cappelli E, Felici G, Weitschek E. Combining DNA methylation and RNA sequencing data of cancer for supervised knowledge extraction. *BioData Mining* 2018; 11(1):1–23. <https://doi.org/10.1186/s13040-018-0184-6> PMID: 30386434
9. Cumbo F, Fiscon G, Ceri S, Masseroli M, Weitschek E. TCGA2BED: extracting, extending, integrating, and querying The Cancer Genome Atlas. *BMC Bioinformatics*. 2017; 18(1):1–9.
10. Quinlan JR. *C4.5: programs for machine learning*. Elsevier; 2014.
11. Breiman L. *Random Forests*. *Mach Learn*. 2001; 45:5–32.
12. Asadi S, Shahrabi J. RipMC: RIPPER for multiclass classification. *Neurocomputing*. 2016; 191:19–33.
13. Cestarelli V, Fiscon G, Felici G, Bertolazzi P, Weitschek E. CAMUR: Knowledge extraction from RNA-seq cancer data through equivalent classification rules. *Bioinformatics*. 2016; 32(5):697–704. <https://doi.org/10.1093/bioinformatics/btv635> PMID: 26519501
14. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*. 2011; 9(4):e1001046. <https://doi.org/10.1371/journal.pbio.1001046> PMID: 21526222
15. Yamada R, Okada D, Wang J, Basak T, Koyama S. Interpretation of omics data analyses. *J Hum Genet*. 2021; 66(1):93–102. <https://doi.org/10.1038/s10038-020-0763-5> PMID: 32385339
16. López de Maturana E, Alonso L, Alarcón P, et al. Challenges in the integration of omics and non-omics data. *Gene*. 2019; 10(3):238. <https://doi.org/10.3390/genes10030238> PMID: 30897838
17. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS ONE*. 2011; 6(2): e17238. <https://doi.org/10.1371/journal.pone.0017238> PMID: 21386892
18. Sprang M, Andrade-Navarro MA, Fontaine JF. Batch effect detection and correction in RNA-seq data using machine-learning-based automated assessment of quality. *BMC Bioinformatics*. 2022; 23(6):1–15. <https://doi.org/10.1186/s12859-022-04775-y> PMID: 35836114
19. Cappelli E, Cumbo F, Bernasconi A, Canakoglu A, Ceri S, Masseroli M, et al. OpenGDC: unifying, modeling, integrating cancer genomic data and clinical metadata. *Appl Sci*. 2020; 10(18):6367.
20. Wei L, Jin Z, Yang S, Xu Y, Zhu Y, Ji Y. TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. *Bioinformatics*. 2018; 34(9):1615–1617. <https://doi.org/10.1093/bioinformatics/btx812> PMID: 29272348
21. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016; 44(8):e71–e71. <https://doi.org/10.1093/nar/gkv1507> PMID: 26704973
22. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 2007; 25(11):1251–1255. <https://doi.org/10.1038/nbt1346> PMID: 17989687
23. Galeota E, Kishore K, Pelizzola M. Ontology-driven integrative analysis of omics data through Onassis. *Sci Rep*. 2020; 10(1):1–9.
24. Chicco D, Masseroli M. Ontology-based prediction and prioritization of gene functional annotations. *IEEE/ACM Trans Comput Biol Bioinform*. 2015; 13(2):248–260.
25. Pinoli P, Chicco D, Masseroli M. Enhanced probabilistic latent semantic analysis with weighting schemes to predict genomic annotations. *Proceedings of IEEE BIBE 2013 –the 13th IEEE International Conference on Bioinformatics and Bioengineering*; 2013; IEEE. p. 1–4.
26. Russkikh N, Antonets D, Shtokalo D, Makarov A, Vyatkin Y, Zakharov A, et al. Style transfer with variational autoencoders is a promising approach to RNA-seq data harmonization and analysis. *Bioinformatics*. 2020; 36(20):5076–5085. <https://doi.org/10.1093/bioinformatics/btaa624> PMID: 33026062
27. Atlan. What is metadata: definition, examples, and types; 2022. Available from: <https://atlan.com/what-is-metadata/> [cited 2023 Apr 17].
28. Vuong P, Wise MJ, Whiteley AS, Kaur P. Ten simple rules for investigating (meta)genomic data from environmental ecosystems. *PLoS Comput Biol*. 2022; 18(12):e1010675. <https://doi.org/10.1371/journal.pcbi.1010675> PMID: 36480496
29. Nagalla S, Chou JW, Willingham MC, Ruiz J, Vaughn JP, Dubey P, et al. Interactions between immunity, proliferation and molecular subtype in breast cancer prognosis. *Genome Biol*. 2013; 14(4):1–18. <https://doi.org/10.1186/gb-2013-14-4-r34> PMID: 23618380
30. Gene Expression Omnibus. GEO2R –GSE45255: Expression profiles of breast tumors from Singapore and Europe; 2022. Available from: <https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE45255> [cited 2023 Apr 17].

31. Alameer A, Chicco D. geoCancerPrognosticDatasetsRetriever: a bioinformatics tool to easily identify cancer prognostic datasets on Gene Expression Omnibus (GEO). *Bioinformatics*. 2022; 38(6):1761–1763. <https://doi.org/10.1093/bioinformatics/btab852> PMID: 34935889
32. Chicco D. geneExpressionFromGEO: an R package to facilitate data reading from Gene Expression Omnibus (GEO). *Microarray Data Analysis*. Springer; 2022. p. 187–194.
33. Chicco D, Oneto L, Tavazzi E. Eleven quick tips for data cleaning and feature engineering. *PLoS Comput Biol*. 2022; 18(12):e1010718. <https://doi.org/10.1371/journal.pcbi.1010718> PMID: 36520712
34. Garcia Santa Cruz B, Vega C, Hertel F. The need of standardised metadata to encode causal relationships: towards safer data-driven machine learning biological solutions. *Proceedings of CIBB 2021 –the 17th International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*. vol. 13483 of *Lecture Notes in Computer Science*. Springer; 2022. p. 200–216.
35. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, et al. Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res*. 2016; 44(D1):D463–D470. <https://doi.org/10.1093/nar/gkv1042> PMID: 26467476
36. Fiehn O, Wohlgemuth G, Scholz M. Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata. *Proceedings of the 2nd International Workshop on Data Integration in the Life Sciences at DILS 2005*. Springer; 2005. p. 224–239.
37. Courtot M, Gupta D, Liyanage I, Xu F, Burdett T. BioSamples database: FAIRer samples metadata to accelerate research data management. *Nucleic Acids Res*. 2022; 50(D1):D1500–D1507. <https://doi.org/10.1093/nar/gkab1046> PMID: 34747489
38. Kimble M, Allers S, Campbell K, Chen C, Jackson L, King B, et al. medna-metadata: an open-source data management system for tracking environmental DNA samples and metadata. *Bioinformatics*. 2022; 38(19):4589–4597. <https://doi.org/10.1093/bioinformatics/btac556> PMID: 35960154
39. Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, et al. FAIRsharing as a community approach to standards, repositories and policies. *Nat Biotechnol*. 2019; 37(4):358–367. <https://doi.org/10.1038/s41587-019-0080-8> PMID: 30940948
40. Shaw F, Etuk A, Minotto A, Gonzalez-Beltran A, Johnson D, Rocca-Serra P, et al. COPO: a metadata platform for brokering FAIR data in the life sciences. *F1000Res*. 2020; 9(495):495.
41. Johnson D, Batista D, Cochrane K, Davey RP, Etuk A, Gonzalez-Beltran A, et al. ISA API: an open platform for interoperable life science experimental metadata. *GigaScience*. 2021; 10(9):giab060. <https://doi.org/10.1093/gigascience/giab060> PMID: 34528664
42. Batista D, Gonzalez-Beltran A, Sansone SA, Rocca-Serra P. Machine actionable metadata models. *Sci Data*. 2022; 9(1):1–8.
43. Chicco D, Jurman G. A brief survey of tools for genomic regions enrichment analysis. *Front Bioinform*. 2022; 2:968327. <https://doi.org/10.3389/fbinf.2022.968327> PMID: 36388843
44. Masseroli M, Pinoli P, Venco F, Kaitoua A, Jalili V, Palluzzi F, et al. GenoMetric Query Language: a novel approach to large-scale genomic data management. *Bioinformatics*. 2015; 31(12):1881–1888. <https://doi.org/10.1093/bioinformatics/btv048> PMID: 25649616
45. Masseroli M, Kaitoua A, Pinoli P, Ceri S. Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying. *Methods*. 2016; 111:3–11. <https://doi.org/10.1016/j.ymeth.2016.09.002> PMID: 27637471
46. Weitschek E, Cumbo F, Cappelli E, Felici G. Genomic data integration: a case study on next generation sequencing of cancer. *Proceedings of DEXA 2016 –the 27th International Workshop on Database and Expert Systems Applications*. IEEE. 2016:49–53.
47. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu Y, et al. TheUCSC Genome Browser Database. *Nucleic Acids Res*. 2003; 31(1):51–54. <https://doi.org/10.1093/nar/gkg129> PMID: 12519945
48. McLean CY, Bristol D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*. 2010; 28(5):495–501. <https://doi.org/10.1038/nbt.1630> PMID: 20436461
49. Fang H, Knezevic B, Burnham KL, Knight JC. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. *Genome Med*. 2016; 8(1):1–20.
50. Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, Mungrue IN, et al. Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet*. 2011; 7(6):e1001393. <https://doi.org/10.1371/journal.pgen.1001393> PMID: 21695224
51. Magazzù G, Zampieri G, Angione C. Clinical stratification improves the diagnostic accuracy of small omics datasets within machine learning and genome-scale metabolic modelling methods. *Comput Biol Med*. 2022; 151:106244. <https://doi.org/10.1016/j.combiomed.2022.106244> PMID: 36343407

52. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* 2020; 21(1):1–17. <https://doi.org/10.1186/s13059-020-02015-1> PMID: 32393329
53. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat.* 2013; 7(1):523. <https://doi.org/10.1214/12-AOAS597> PMID: 23745156
54. Lê Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics.* 2011; 12(1):1–17. <https://doi.org/10.1186/1471-2105-12-253> PMID: 21693065
55. Acharjee A, Kloosterman B, Visser RG, Maliepaard C. Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics.* 2016; 17(5):363–373. <https://doi.org/10.1186/s12859-016-1043-4> PMID: 27295212
56. Cao ZJ, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol.* 2022; 40(10):1458–1466. <https://doi.org/10.1038/s41587-022-01284-4> PMID: 35501393
57. Zampieri G, Vijayakumar S, Yaneske E, Angione C. Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput Biol.* 2019; 15(7):e1007084. <https://doi.org/10.1371/journal.pcbi.1007084> PMID: 31295267
58. El-Manzalawy Y, Hsieh TY, Shivakumar M, Kim D, Honavar V. Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data. *BMC Med Genet.* 2018; 11(3):19–31. <https://doi.org/10.1186/s12920-018-0388-0> PMID: 30255801
59. Poirion OB, Jing Z, Chaudhary K, Huang S, Garmire LX. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med.* 2021; 13(1):1–15.
60. Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J.* 2021; 19:3735–3746. <https://doi.org/10.1016/j.csbj.2021.06.030> PMID: 34285775
61. Feldner-Busztin D, Firbas Nisantzis P, Edmunds SJ, Boza G, Racimo F, Gopalakrishnan S, et al. Dealing with dimensionality: the application of machine learning to multi-omics data. *Bioinformatics.* 2023;39(2):btad021.
62. Xu C, Jackson SA. Machine learning and complex biological data. *Genome Biol.* 2019; 20(1):1–4.
63. Angione C. Human systems biology and metabolic modelling: a review—from disease metabolism to precision medicine. *Biomed Res Int.* 2019;2019. <https://doi.org/10.1155/2019/8304260> PMID: 31281846
64. Vijayakumar S, Conway M, Lió P, Angione C. Optimization of multi-omic genome-scale models: methodologies, hands-on tutorial, and perspectives. In: Fondi M, editor. *Metabolic Network Reconstruction and Modeling*, vol. 1716. New York, NY: Humana Press; 2018. p. 389–408.
65. Rohatgi N, Ghoshdastider U, Baruah P, Kulshrestha T, Skanderup AJ. A pan-cancer metabolic atlas of the tumor microenvironment. *Cell Rep.* 2022; 39(6):110800. <https://doi.org/10.1016/j.celrep.2022.110800> PMID: 35545044
66. Martorell-Marugán J, Chierici M, Jurman G, Alarcón-Riquelme ME, Carmona-Sáez P. Differential diagnosis of systemic lupus erythematosus and Sjögren's syndrome using machine learning and multi-omics data. *Comput Biol Med.* 2023; 152:106373.
67. Culley C, Vijayakumar S, Zampieri G, Angione C. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. *Proc Natl Acad Sci U S A.* 2020; 117(31):18869–18879. <https://doi.org/10.1073/pnas.2002959117> PMID: 32675233
68. Lewis JE, Kemp ML. Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance. *Nat Commun.* 2021; 12(1):1–14.
69. Magazzù G, Zampieri G, Angione C. Multimodal regularized linear models with flux balance analysis for mechanistic integration of omics data. *Bioinformatics.* 2021; 37(20):3546–3552. <https://doi.org/10.1093/bioinformatics/btab324> PMID: 33974036
70. Gopalakrishnan S, Joshi CJ, Gomez MV, Icten E, Rolandi P, Johnson W, et al. Guidelines for extracting biologically relevant context-specific metabolic models using gene expression data. *Metab Eng.* 2023; 75:181–191. <https://doi.org/10.1016/j.ymben.2022.12.003> PMID: 36566974
71. Baldwin E, Han J, Luo W, Zhou J, An L, Liu J, et al. On fusion methods for knowledge discovery from multi-omics datasets. *Comput Struct Biotechnol J.* 2020; 18:509–517. <https://doi.org/10.1016/j.csbj.2020.02.011> PMID: 32206210

72. Doan LMT, Angione C, Occhipinti A. Machine learning methods for survival analysis with clinical and transcriptomics data of breast cancer. *Computational Biology and Machine Learning for Metabolic Engineering and Synthetic Biology*. Springer; 2023. p. 325–393.
73. Cuperlovic-Culf M, Nguyen-Tran T, Bennett SA. Machine learning and hybrid methods for metabolic pathway modeling. *Computational Biology and Machine Learning for Metabolic Engineering and Synthetic Biology*. Springer; 2023. p. 417–439. https://doi.org/10.1007/978-1-0716-2617-7_18 PMID: 36227553
74. Chicco D. Ten quick tips for machine learning in computational biology. *BioData Mining*. 2017; 10(1):1–17.
75. Chicco D, Agapito G. Nine quick tips for pathway enrichment analysis. *PLoS Comput Biol*. 2022; 18(8): e1010348. <https://doi.org/10.1371/journal.pcbi.1010348> PMID: 35951505
76. Chicco D, Shiradkar R. Ten quick tips for computational analysis of medical images. *PLoS Comput Biol*. 2023; 19(1):e1010778. <https://doi.org/10.1371/journal.pcbi.1010778> PMID: 36602952
77. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004; 5(10):1–16. <https://doi.org/10.1186/gb-2004-5-10-r80> PMID: 15461798
78. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods*. 2018; 15(7):475–476. <https://doi.org/10.1038/s41592-018-0046-7> PMID: 29967506
79. Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res*. 2020; 48(W1):W395–W402. <https://doi.org/10.1093/nar/gkaa434> PMID: 32479607
80. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*. 2015; 3:e1319. <https://doi.org/10.7717/peerj.1319> PMID: 26500826
81. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016; 3(1):1–9. <https://doi.org/10.1038/sdata.2016.18> PMID: 26978244
82. Kaggle. Kaggle datasets—Explore, analyze, and share quality data; 2022. Available from: <https://www.kaggle.com/datasets> [cited 2023 Apr 17].
83. University of California Irvine. Machine Learning Repository; 1987. Available from: <https://archive.ics.uci.edu/> [cited 2023 Apr 17].
84. FigShare. Store, share, discover research; 2011. Available from: <https://www.figshare.com> [cited 2023 Apr 17].
85. Zenodo. Research, shared; 2013. Available from: <https://www.zenodo.org> [cited 2023 Apr 17].
86. Chicco D, Ceroni G, Cangelosi D. A survey on publicly available open datasets derived from electronic health records (EHRs) of patients with neuroblastoma. *Data Sci J*. 2022; 21(1).
87. Chicco D. Neuroblastoma Electronic Health Records Open Data Repository; 2022. Available from: https://davechicco.github.io/neuroblastoma_EHRs_data/ [cited 2023 Apr 17].
88. Brown AV, Campbell JD, Assefa T, Grant D, Nelson RT, Weeks NT, et al. Ten quick tips for sharing open genomic data. *PLoS Comput Biol*. 2018; 14(12):e1006472. <https://doi.org/10.1371/journal.pcbi.1006472> PMID: 30589835
89. Pisani E, Aaby P, Breugelmans JG, Carr D, Groves T, Helinski M, et al. Beyond open data: realising the health benefits of sharing data. *BMJ*. 2016;355. <https://doi.org/10.1136/bmj.i5295> PMID: 27758792
90. Scimago Journal Ranking. Molecular biology open access journals; 2022. Available from: <https://www.scimagojr.com/journalrank.php?category=1312> [cited 2023 Apr 17].
91. Python Software Foundation. The Python Package Index; 2003. Available from: <https://pypi.org> [cited 2023 Apr 17].
92. Blankenberg D, von Kuster G, Bouvier E, Baker D, Afgan E, Stoler N, et al. Dissemination of scientific software with Galaxy ToolShed. *Genome Biol*. 2014; 15(2):1–3. <https://doi.org/10.1186/gb4161> PMID: 25001293
93. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. Ten simple rules for the care and feeding of scientific data. *PLoS Comput Biol*. 2014; 10(4):e1003542. <https://doi.org/10.1371/journal.pcbi.1003542> PMID: 24763340
94. Lee BD. Ten simple rules for documenting scientific software. *PLoS Comput Biol*. 2018; 14(12): e1006561. <https://doi.org/10.1371/journal.pcbi.1006561> PMID: 30571677
95. Karimzadeh M, Hoffman MM. Top considerations for creating bioinformatics software documentation. *Brief Bioinform*. 2018; 19(4):693–699. <https://doi.org/10.1093/bib/bbw134> PMID: 28088754

96. Schnell S. Ten simple rules for a computational biologist's laboratory notebook. *PLoS Comput Biol.* 2015; 11(9):e1004385.
97. Das S. OMICsPCA: an R package for quantitative integration and analysis of multiple omics assays from heterogeneous samples; 2022. Available from: <https://www.bioconductor.org/packages/release/bioc/vignettes/OMICsPCA/inst/doc/vignettes.html> [cited 2023 Mar 21].