YYYY, 0–0
Advance Access Publication Date: DD Month YYYY

Systems biology

Vertical and horizontal integration of multi-omics data with miodin

Benjamin Ulfenborg[1,*]

[1]Department of Bioscience, University of Skövde, 544 28 Skövde, Sweden

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Studies on multiple modalities of omics data such as transcriptomics, genomics and proteomics are growing in popularity, since they allow us to investigate complex mechanisms across molecular layers. Integration of multi-omics data remains challenging and requires combination of several software tools to process data and generate biological insights. This paper presents miodin, a software package for streamlined data analysis and integration of omics data across experiments on the same samples, or across studies on the same variables.

Availability and implementation: The miodin package is implemented in R and is freely available for use and extension under the GPL-3 license. Package source, reference documentation and user manual are available at https://gitlab.com/algoromics/miodin.

Contact: benjamin.ulfenborg@his.se

1 Introduction

With the advances in high-throughput biotechnology over the past two decades, we now have access to an unprecedented wealth of data for many omics modalities. In this era of biomedical big data, the primary research challenges are how to integrate and analyze large-scale data of different types and sources to gain new insights into the complex mechanisms behind health and disease (Joyce and Palsson, 2006; Ebrahim *et al.*, 2016; Berger *et al.*, 2013; Karczewski and Snyder, 2018). Advanced bioinformatics tools and algorithms are being developed that can accommodate multi-omics data (Reich *et al.*, 2006; Fisch *et al.*, 2015; Afgan *et al.*, 2018), but performing transparent and reproducible integrated analysis remains a significant challenge, and considerable technical expertise is required to use many tools and combine them into a coherent pipeline.

To address these challenges, the miodin package was developed. This provides a software infrastructure to build data analysis workflows that import, process and analyze multi-omics data. Workflows accommodate data of different omics modalities, including transcriptomics, genomics, epigenomics and proteomics, from different experimental techniques (microarrays, sequencing and mass spectrometry). The package allows users to integrate omics data from different experiments on the same samples (vertical integration) or from studies on the same variables (horizontal integration). A key design goal when developing miodin was to streamline data analysis, by providing a clean syntax for building workflows and minimizing the extent of technical expertise required. The motivation behind this was to promote transparent and reproducible biomedical data science.

2 Package features

Data analysis with miodin follows an intuitive three-step process illustrated in Figure 1. The user first initializes a project, a study and a workflow. The project serves as a placeholder for all other objects and the study is used to declare the study design, including what samples and assays to analyze, and the experimental variables of interest, if any. The miodin package implements an expressive study design vocabulary and several convenience functions for common designs, such as case-control and time series experiments. These allow the user to declare all information required for data analysis in one place, thus reducing the risk of clerical errors in the analysis script and the amount of information the user must provide during the analysis itself. The workflow is used to define the data analysis steps as a set of sequentially connected workflow modules. Modules carry out specific steps, such as data import or processing, and can easily be combined using the pipe operator `%>%` in R.

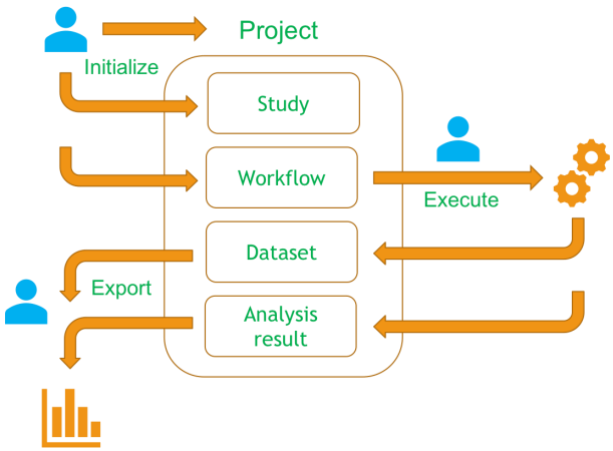


Fig. 1. Data analysis in miodin. The user starts by defining a project, a study and a workflow. The study contains the design of the experiment and the workflow is defined by instantiating analysis modules, which generate datasets and analysis results upon execution. The user can then inspect and export the results.

```
mw <- MiodinWorkflow("DataAnalysisFlow")
mw <- mw +
  importMicroarrayData(...) %>%
  processMicroarrayData(...) %>%
  analyzeOmicsModel(...)
mw <- execute(mw)
```

This script initializes a workflow called *DataAnalysisFlow* with three workflow modules. Module parameters have been omitted for brevity. The first module imports microarray data, the second processes the output from the first module, and the final module performs statistical testing on the processed data. The analysis is carried out by calling the *execute* method. This syntax allows the user to define streamlined data analysis workflows, enhancing readability of the analysis script compared to longer chunks of code. Workflow modules have names starting with verbs denoting their function, making them easier to remember and improving analysis transparency.

To further improve transparency, the analysis workflow automatically documents each processing and analysis step, and these can be inspected and exported to reveal what was done to the data, what parameters were used and what result was generated. Another important strength of this workflow-based approach is data analysis reproducibility. Provided that the input data and module parameters are the same, the workflow will generate the same results. This makes it easy for different people to reproduce the analysis on their own system.

3 Sample workflows

The miodin package was applied to perform horizontal and vertical integration of omics data made available through Bioconductor. For horizontal integration, three breast cancer RNA-seq datasets from the metaSeq package (Sun *et al.*, 2011; Ha *et al.*, 2011; Tsuyuzaki and Nikaido, 2013) were used to identify differentially expressed genes between tumor samples and non-tumorigenic controls. The analysis script is available in Supplementary File S1 and the resulting tables of genes in Supplementary Table S1 and S2. Due to the small sample size genes were identified based on absolute log2 fold change > 6 only, resulting in five genes identified in all three datasets.

Vertical integration was carried out using breast cancer data from the curatedTCGAData package (Ramos *et al.*, 2018). RNA-seq gene and miRNA count data as well as 450k methylation data were integrated and used to fit a Multi Omics Factor Analysis (MOFA) model (Argelaguet *et al.*, 2018). The analysis script is supplied in Supplementary File S2 and the resulting MOFA-based sample ordination plot in Supplementary Figure S1.

4 Discussion

Multi-omics studies are becoming more prevalent in the biomedical literature. Two bioinformatics tools that can analyze multi-omics data are Galaxy (Afgan *et al.*, 2018) and GenePattern (Reich *et al.*, 2006), which provide users with a workflow-based interface and support an impressive number of methods for data analysis. Most tools focus on analysis of individual omics data types, however, and support for integrated analysis is currently lacking. The miodin package has been designed to fill this need and allows users to perform streamlined data analysis.

To ensure the research community benefits from miodin, the software package with extensive documentation is made freely available on GitLab under the GPL-3 license.

Funding

This work has been supported by the Knowledge Foundation [grand number 20160293] and the Systems Biology Research Centre at University of Skövde, Skövde, Sweden.

Conflict of Interest: none declared.

Acknowledgements

Vertical integration of omics data was performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

References

Afgan,E. et al. (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res., 46, W537--W544.

Argelaguet,R. et al. (2018) Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol., 14, 1–13.

Berger,B. et al. (2013) Computational solutions for omics data. Nat. Rev. Genet., 8, 1385–1395.

Ebrahim,A. et al. (2016) Multi-omic data integration enables discovery of hidden biological regularities. Nat. Commun., 7, 1–9.

Fisch,K.M. et al. (2015) Omics Pipe: A community-based framework for reproducible multi-omics data analysis. Bioinformatics, 31, 1724–1728.

Ha,K.C.H. et al. (2011) Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines. BMC Med. Genomics, 4, 75.

Joyce,A.R. and Palsson,B.Ø. (2006) The model organism as a system: integrating 'omics' data sets. Nat. Rev. Mol. Cell Biol., 7, 198–210.

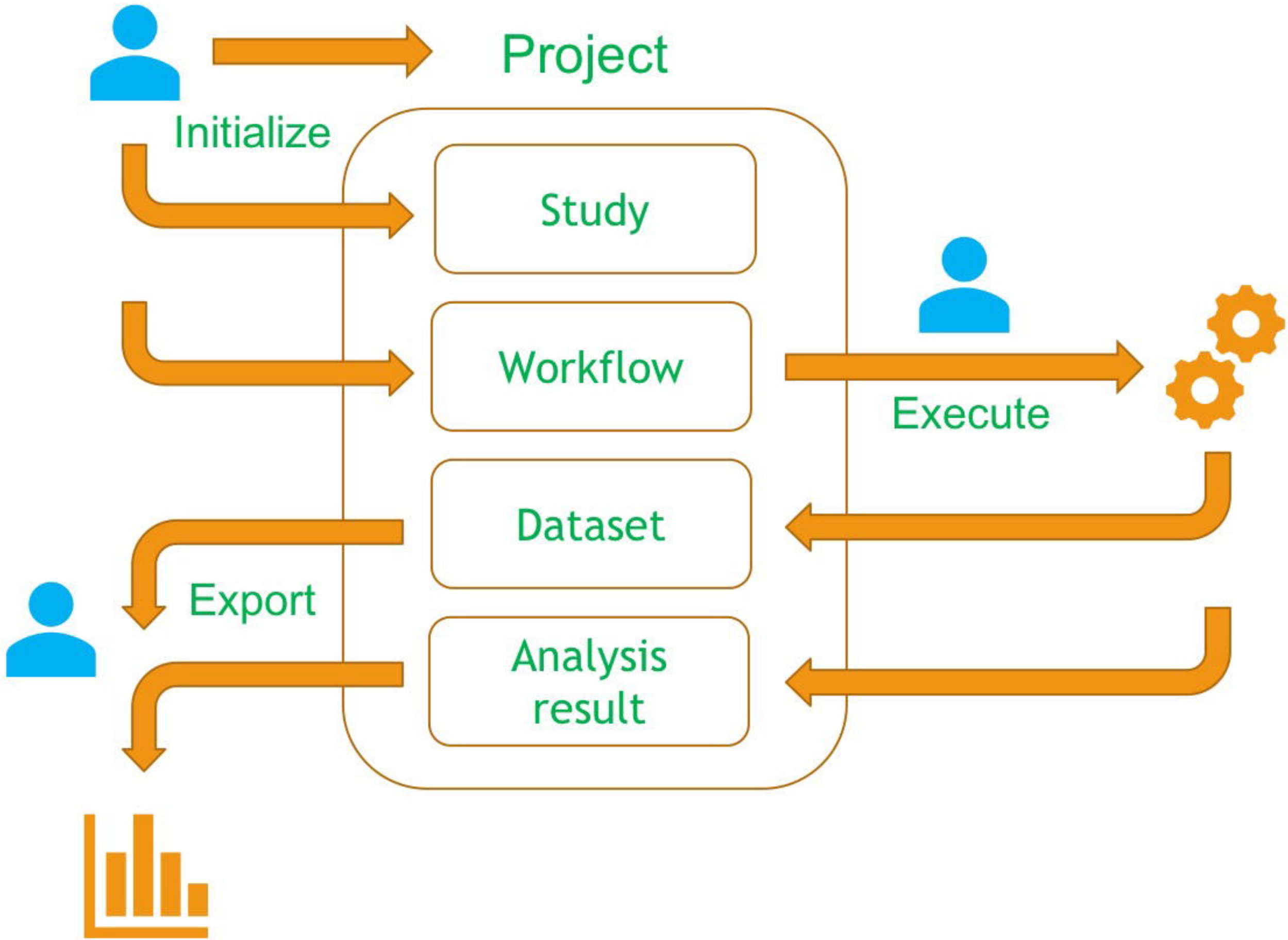
Karczewski,K.J. and Snyder,M.P. (2018) Integrative omics for health and disease. Nat. Rev. Genet., 19, 29–39.

Ramos,M. et al. (2018) curatedTCGAData: Curated Data From The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects. R Package. version 1.2.0.

Reich,M. et al. (2006) GenePattern 2.0. Nat. Genet., 38, 500.

Sun,Z. et al. (2011) Integrated analysis of gene expression, CpG Island methylation, and gene copy number in breast cancer cells by deep sequencing. PLoS One, 6.

Tsuyuzaki,K. and Nikaido,I. (2013) metaSeq: Meta-analysis of RNA-Seq count data in multiple studies. R Package. version 1.20.0.


Project
Initialize
Study
Workflow
Execute
Dataset
Export
Analysis result