Research Ideas and Outcomes 3: e11624
doi: 10.3897/rio.3.e11624

Data Management Plan

Data Management Plan for a Biotechnology and Biological Sciences Research Council (BBSRC) Tools and Resources Development Fund (TRDF) Grant

Laurent Gatto ‡

‡ University of Cambridge, Cambridge, United Kingdom

Corresponding author: Laurent Gatto (lg390@cam.ac.uk)

Reviewable v1

Received: 23 Dec 2016 | Published: 05 Jan 2017

Citation: Gatto L (2017) Data Management Plan for a Biotechnology and Biological Sciences Research Council (BBSRC) Tools and Resources Development Fund (TRDF) Grant. Research Ideas and Outcomes 3: e11624. https://doi.org/10.3897/rio.3.e11624

Abstract

Background

This Data Management Plan (DMP) was created for Laurent Gatto's BBSRC Tools and Resources Development Fund award (BB/N023129/1).

New information

The DMP describes the management and sharing of all data and code associated with the grant, including software dissemination and release schedule, source code development and open source licensing, software documentation, reproducible framework and data annotation and dissemination.

Keywords

Spatial proteomics, Bioconductor, machine learning, mass spectrometry, proteomics, software

Products of research

The participants have a long history of successful collaboration and open source development and are fully committed to abiding by the BBSRC's policy on data management. Specific outputs of this project and how they will be made available to the community are listed below.

Software

All software infrastructure and statistical routines developed in this project will be submitted to the Bioconductor project (Huber et al. 2015). We will continue to follow the well established standards for packaging, versioning, documentation, updating and installation. Bioconductor will be the official distribution channel for the software, thus benefiting from existing dissemination infrastructure, support channels, user base and the developer community. In addition to biannual official releases in March and October, tested and documented development versions of the software will also be available through the dedicated Bioconductor development branch.

Source code

We understand the value of open source development practices within the scientific community. The source code of the software will be freely available in code repositories under permissive open source licenses and hosted on the Bioconductor subversion server. In addition, we will continue to use the GitHub *social coding* infrastructure to facilitate collaboration within the team and promote contributions from the community. The two repositories will be clearly documented (for example by using software versions) to avoid any confusion and kept in sync using dedicate tools such as git-svn. As well as being good practice, open source and collaborative development of our software will enhance the visibility and sustainability of what we produce.

Documentation

All software that will be released as part of this project will be thoroughly documented in multiple ways. Individual functions and data containers will be described in detail to allow users and developers to understand and use them in their own pipelines. In addition, we will produce vignettes, dynamically generated documents that offer a general overview of the functionality of the software and flexibility of the pipelines, advise on how to explore the data and understand the results, information on data preparation and import into the R environment and links to relevant resources. We will also produce educational material that will be broadly distributed independently of the software through workshops and courses to maximise visibility of the software and analysis methodologies and facilitate adoption by new users less familiar with the R/Bioconductor environment and community. In particular, the material for our second workshop dedicated to the analysis and interpretation of spatial proteomics will be made publicly available.

Reproducible framework

End users will gain access to accurate, biologically relevant results and experimental data through existing resources, dedicated data packages and wider databases, and experts interested in the analytical process will gain open access to relevant elements of a key proteomics methodology. The combined distribution of annotated data and well-documented software bundled in analysis scripts will offer users and developers a complete reproducible environment.

Data

While no new data will be generated specifically in the frame of this project, statistically sound (re-)analysis and reliable (re-)interpretation of published or private data will be produced. These data will be made available through multiple existing community resources using established standards and annotated withample meta data. They will be distributed as dedicated R object (in well-established data structures defined in MSnbase Gatto and Lilley 2011), as used and manipulated through the pRoloc (Gatto et al. 2014) and pRolocGUI software and included in the open pRolocdata (Gatto et al. 2014) data package. All datasets will be thoroughly annotated with meta data to provide users with all necessary details on the origin or manipulation of the data in order to favour and facilitate re-use and reproducibility. Several exporters are already available, to offer these same data as spreadsheets or in the mzTab (Griss et al. 2014) format. When available, raw and identification data will be distributed using the mzML and mzIdentML Proteomics Standards Initiative (PSI) community formats and disseminated through the ProteomeXchange (PX) project (Vizcaíno et al. 2014) and the PRoteomics IDEntifications (PRIDE) resource. We will also distribute the data and results through the online resource *SpatialMap.org* that we are currently developing, which will enable users to interactively visualise, explore and search the data and annotated results stemming of our state-of-the-art statistical learning pipelines.

The refined and novel protein sub-cellular localisations will be communicated to the wider proteomics community via relevant protein databases and annotation providers like Swiss-Prot, the Gene Ontology Annotation database as well as more specialised resources. The improved localisation information will be distributed with all technical details regarding the analysis and interpretation/evidence, including algorithm specifications and parameters and assignment probabilities.

Data will be made available as soon as it has been quality controlled and converted into usable computational objects. Once validated on various datasets, the algorithms will be included and distributed through the relevant software packages. The multiple sources and formats will be cross-referenced to maximise utility and availability to the research community.

Acknowledgements

The author would like to thank Dr Marta Teperek and Dr Ross Mounce for their enrouragements to publish this DMP, as well as the Research Data Management team at the University of Cambridge for their efforts in promoting open data and good data management practice.

Grant title

Understanding protein multi- and trans-localisation at the full proteome level

Hosting institution

University of Cambridge

Author contributions

Laurent Gatto wrote the Data Management Plan.

References

- Gatto L, Lilley KS (2011) MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. Bioinformatics 28 (2): 288-289. https://doi.org/10.1093/bioinformatics/btr645
- Gatto L, Breckels LM, Wieczorek S, Burger T, Lilley KS (2014) Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. Bioinformatics 30 (9): 1322-1324. https://doi.org/10.1093/bioinformatics/btu013
- Griss J, Jones AR, Sachsenberg T, Walzer M, Gatto L, Hartler J, Thallinger GG, Salek RM, Steinbeck C, Neuhauser N, Cox J, Neumann S, Fan J, Reisinger F, Xu Q-, Toro Nd, Perez-Riverol Y, Ghali F, Bandeira N, Xenarios I, Kohlbacher O, Vizcaino JA, Hermjakob H (2014) The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. Molecular & Cellular Proteomics 13 (10): 2765-2775. https://doi.org/10.1074/mcp.o113.036681
- Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M (2015) Orchestrating high-throughput genomic analysis with Bioconductor. Nature methods 12 (2): 115-21. https://doi.org/10.1038/nmeth.3252

- Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz P, Xenarios I, Eisenacher M, Mayer G, Gatto L, Campos A, Chalkley RJ, Kraus H, Albar JP, Martinez-Bartolomé S, Apweiler R, Omenn GS, Martens L, Jones AR, Hermjakob H (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nature Biotechnology 32 (3): 223-226. https://doi.org/10.1038/nbt.2839