Concepts:

Data availability for users, community and public (including processed data and metadata)

We will ensure that we follow the principles of open science to make our data findable, accessible, interoperable and reusable (FAIR). Since 2020 we are part of the DFG funded consortium DataPLANT (https://www.nfdi4plants.de/, funded by DFG project 442077441). DataPlant provides a platform for data publication including the Annotated Research Context (ARC) according to the FAIR principles. All generated data are assigned a unique Digital Object Identifier (DOI) and annotated with ontology-based metadata. This is controlled by the DataPlant repository, which is based on specific formats, checklists and terminologies. Data sharing and the creation of comprehensive datasets are essential for collaborative research, and DataPlant is ideally suited for data sharing with users and within the community. Data can be stored and shared on DataPlant before the final data construct is submitted to an endpoint repository. This allows multiple scientists to simultaneously add and logically connect their data to existing sets. Git's versioning feature allows each step to be tracked at any time, preserving the provenance of each data set. The collection of ARCs is available in a searchable form. Based on specific data or metadata attributes, each scientist can have quick access to individual measurements, information or entire ARCs.

In addition, the Faculty of Biology is part of the LMU's Open Science Center, which offers workshops to train users in data management. It also provides access to a large community of researchers interested in RDM and open science.

Data storage

We expect to generate 25-35 TB of data per year. To ensure storage and backup of the data for at least 10 years, we use in-house servers hosted by the Biology Department's IT group. These servers have 300 TB of space spread over 3 RAID arrays, and the data is synchronised to an identical server hosted at the BMC in case of any local network/physical/device problems. A 10 GB connection ensures timely data transfer. Data is additionally backed up by the Tivoli replacement "IBM Spectrum Protect" hosted by the LRZ (Leibnitz Rechenzentrum). This "belt and braces" approach allows for minimal downtime in the event of a problem.

Available and required hardware and software

Standard data analysis (data dependent analysis, DDA) will be performed with the open source MaxQuant (Tyanova et al. 2016a), which is a widely used software tool for the analysis of mass spectrometry-based proteomics data. It supports label-free quantification approaches (LFQ) as well the analysis of TMT labeled peptides and post-translational modifications (PTMs).

Data independent analysis (DIA) will be performed with DIA-NN (Demichev et. al. 2020), a recently developed open-source software suite that uses deep neural networks. Thereby, DIA-NN provides the possibility to analyses large datasets with high speed. It does not require a spectral library and reliable quantifications can especially be obtained when analyzing PTMs or using low sample volumes.

In addition, we will use ProteomIQon. ProteomIQon (doi:10.5281/zenodo.6335068) is a collection of open source computational proteomics tools to build pipelines for the evaluation of MS-derived proteomics data written in F#. This tool also allows the quantification of 15N-labeled peptides, a method we have successfully used in the past and will continue to use in the future.

However, these programs can be computationally intensive, especially when processing large proteomics datasets. Therefore, we will use the High Performance Compute Cluster recently established by the Faculty of Biology (BioHPC, funded by DFG project 450674345 to Jochen Wolf). The

BioHPC provides access to all compute resources and the additional use of a shared storage system (currently DSS). The BioHPC consists of compute nodes with access to large amounts of RAM, which are connected to a storage system via high-speed connections, tailored to the need for high I/O performance of high-throughput data. The cluster is physically hosted by the LRZ, which is also responsible for the setup and maintenance of all hardware. Due to the modular design of the BioHPC cluster, we will be able to implement and run our data analysis using Galaxy Europe, an open source platform for FAIR data analysis. Both MaxQuant and ProteomIQon are implemented on Galaxy, which provides an easy-to-use platform. Although standard analysis will be performed by the MSBioLMU, the Galaxy interface is intuitive and does not require advanced programming skills, making it easy to train users.

Downstream statistical analysis

Basic statistical analysis will be performed using Perseus (Tyanova et al. 2016b). For more complex analysis ask in advance.

References

Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M. (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. Nat Methods. 1:41-44. doi: 10.1038/s41592-019-0638-x

[a]Tyanova S, Temu T, Cox J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc. 12:2301-2319. doi: 10.1038/nprot.2016.136

[b]Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 9:731-40. doi: 10.1038/nmeth.3901