Interactive Visualization for large-scale multi-factorial Research Designs

Andreas Friedrich[1,2][0000−0001−6843−8680], Luis de la Garza[1][0000−0002−1590−5430], Oliver Kohlbacher[1,2,3,4][0000−0003−1739−4598], and Sven Nahnsen[1][0000−0002−4375−0691]

[1] Quantitative Biology Center (QBiC), University of Tübingen,
Auf der Morgenstelle 10, 72076 Tübingen, Germany
[2] Center for Bioinformatics & Dept. of Computer Science, University of Tübingen,
Sand 14, 72076 Tübingen, Germany
[3] Biomolecular Interactions, Max Planck Institute for Developmental Biology,
MaxPlanckRing 5, 72076 Tübingen, Germany
[4] Institute for Translational Bioinformatics, University Hospital Tübingen

Abstract. Recent publications have shown that the majority of studies cannot be adequately reproduced. The underlying causes seem to be diverse. Usage of the wrong statistical tools can lead to the reporting of dubious correlations as significant results. Missing information from lab protocols or other metadata can make verification impossible. Especially with the advent of Big Data in the life sciences and the hereby-involved measurement of thousands of multi-omics samples, researchers depend more than ever on adequate metadata annotation. In recent years, the scientific community has created multiple experimental design standards, which try to define the minimum information necessary to make experiments reproducible. Tools help with creation or analysis of this abundance of metadata, but are often still based on spreadsheet formats and lack intuitive visualizations. We present an interactive graph visualization tailored to experiments using a factorial experimental design. Our solution summarizes sample sources and extracted samples based on similarity of independent variables, enabling a quick grasp of the scientific question at the core of the experiment even for large studies. We support the ISA-Tab standard, enabling visualization of diverse omics experiments. As part of our platform for data-driven biomedical research, our implementation offers additional features to detect the status of data generation and more.

Keywords: experimental design · aggregation graph · metadata · portal · reproducibility.

1 Introduction

The reproducibility crisis has revealed obvious shortcomings of modern biomedical experimental techniques. While outright fraud seems to be the exception, recent publications pinpoint many of the problems as based on missing statistical understanding when planning or performing scientific studies [4]. Even if enough data is available to draw significant conclusions, the interaction between different variables has to be reflected in the experimental design. Independent variables are usually the focus of a study and are controlled by the experimenter. However, it is rare that a variable like a disease state depends on only one single variable: regulatory networks commonly include proteins that act together to create a phenotype [12]. It is thus sensible to study multiple independent variables in a factorial experimental design. The concept of factorial experimental designs was popularized in crop research [5, 6] and allows experimenters to detect interactions, something not possible in one-factor-at-a-time (OFAAT) experiments.

The advancement of Big Data assists to conduct sophisticated experiments benefiting from these study designs. Yet, even well-designed studies can often not be reproduced, because crucial metadata is missing [23]. Convenient interfaces between experimenters' notes and online database systems are often missing. Excel spreadsheets are still the most widely used tool for research notes pertaining to assays and samples [16]: early efforts to standardize scientific reporting lead to formalized spreadsheet formats specifying the minimum required information to reproduce an experiment. MIAME, the Minimum Information About a Microarray Experiment [3, 2] standard and the microarray gene expression markup language MAGE-ML [20] aim at annotating experiments so they can be independently verified. Similarly, MIAPE, a standard describing the Minimum Information About a Proteomics Experiment, tries to specify the needed information to interpret analyses performed on proteins [21]. ISA-Tab combines these earlier approaches into an interoperable spreadsheet format relating information about research aims, other related studies and their associated assays [18, 19]. Different efforts have been undertaken to provide users with tools based on the ISA standard [9, 10]. linkedISA leverages the data provided to create a semantic, interoperable presentation and shows how implicitly defined study groups can be extracted from ISA-Tab. These groups are summarized and listed in Bio-GraphIIn, a graph-based repository for biological experimental data [8]. With the growing complexity of biological experiments and especially the communication thereof, efficient visualization are indispensable. However, most of the work has been focused on connecting experiments to ontology frameworks and making it machine-readable. While Bio-GraphIIn presents a list of study groups, this type of presentation can become difficult to grasp for huge experiments involving many experimental factors and other metadata. More information can only be obtained by displaying huge tables of samples.

The need to use computer-aided experimental design for large studies was previously discussed in early factorial design approaches in behavioral research, such as the online tool WEXTOR [17]. Here, the combination of every possible factor level pertaining to participants can be used to create specific webpages that guide the corresponding subjects to their questions or tests. In high-throughput biomedical science, analysis tools that make use of experimental design information are often limited to custom formats: MaxQuant allows users to edit an Experimental Design template to relate files with sample fractions [22].

Here, we build on our intuitive interface for experiment creation leveraging proven experimental design concepts like full-factorial study design [7]. To connect experimental designs with data integration, we provide an interactive visualization tool that can summarize complex study designs based on involved species, tissues, analytes and experimental factors into an intuitive experiment graph. In an effort to comply with existing standards while allowing easy options to manage high-throughput experiments, we provide interoperability with the ISA-Tab format, and suggest a format for simplified experiment creation. The highly modular structure makes our tool a good starting point for further developments in the area of quality control and statistical power estimation.

2 Methods

2.1 Factorial Experimental Designs

In a factorial design the influences of all independent experimental variables on the response are investigated. A factor of an experimental design is defined as one such variable that is being studied. A level is one possible variation of a factor. The number of levels denotes the total number of different variations for a single factor that was used in an experiment. Factorial designs are called full-factorial designs, if every possible combination of levels is tested. A full-factorial experiment with n factors and k levels for each factor is called a $k \times n$ factorial design and consists of k^n sub-experiments, as exemplified in Table 1. Each of these cases can then have multiple biological or technical replicates.

Table 1. Example of a 3x2 full-factorial experimental design. Two variables x_1 and x_2 containing three levels each are tested, leading to nine different experiments.

Variables	Experiment no.								
	1	2	3	4	5	6	7	8	9
x_1	-	-	-	+	+	+	0	0	0
x_2	-	+	0	-	+	0	-	+	0

2.2 Aggregation Graph

The hierarchical way in which omics experiments are typically performed leads to an intuitive sample graph connecting patient/model organism entities to those denoting tissue/cell extract and measured analyte entities as previously described [7, 14]. The example in Fig. 1 visualizes experiments on six mice. In each case a liver sample was taken and proteins prepared for mass spectrometry analysis.

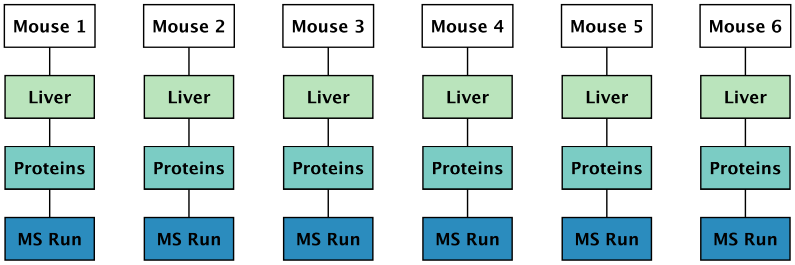


Fig. 1. Sample hierarchy of six sources with attached samples and measurements.

Let $G = (V_G, E_G)$ be a sample graph with vertices $v \in V_G$ denoting each of these entities in an experiment and edges $(v, w) \in E_G$ denoting the extraction of entity w from entity v in an experimental step.

Let further $f_1 \dots f_n$ be a set of experimental factors on a subset of these entities with factor level f_{iv} for factor f_i of vertex v and a similarity function on factor levels $s(f_{iv}, f_{iw}) = \{0, 1\}$.

We define a set of aggregation graphs $H_1 \dots H_n$, one for each factor f_i:

$$\begin{gathered} H_i = (V_H, E_H) \\ \forall\ v \in V_G : \sum_{w \in V_H} s(f_{iv}, f_{iw}) = 0 \rightarrow v \in V_H \\ \forall\ (v, w) \in E_G :\ v \in V_H \wedge w \in V_H \rightarrow (v, w) \in E_H \end{gathered} \tag{1}$$

Each graph H aggregates all entities of G with a similar factor level into a single vertex, while preserving connections between the hierarchy levels of the experiment. For nominal factors, similarity is best defined as the perfect match of both levels, while quantitative variables can be summarized using intervals.

2.3 Implementation

We use the Open Source Biology Information System (openBIS) to store datasets and annotating metadata. Our experimental design is represented both by interconnected entities denoting source organisms and samples as well as metadata properties of these entities. Experimental factors and other properties of sample source entities and samples are stored in a intermediary XML format in openBIS that is validated by an XML schema. This schema includes quantitative variables with or without units as well as variables on a nominal scale, for example different disease states. Metadata is read from and written to the system using a web portlet running on a Liferay portal.

The schematic integration of our experimental design visualization into the portal can be seen in Fig. 2. Users can create experiments using a wizard process or file import and browse information about existing experiments [7, 14]. Both imported and existing experiments can be translated into the aggregation graph

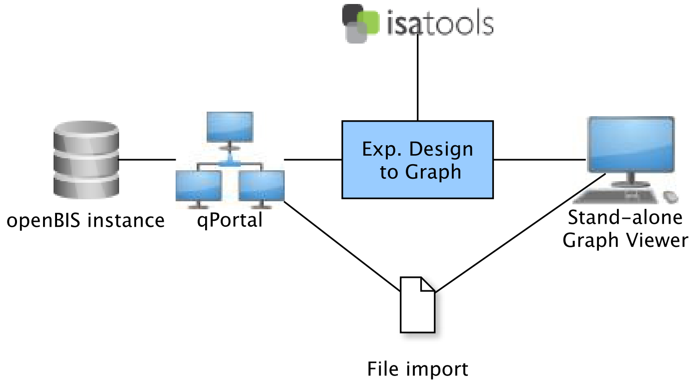
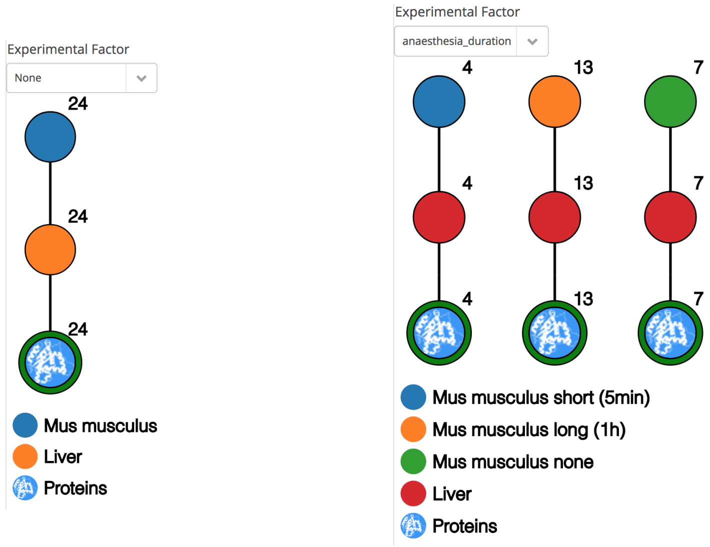


Fig. 2. Schematic diagram of our implementation: Existing experimental designs and their metadata stored in openBIS can be visualized in qPortal using our Java-based experimental design libraries. Users can import and view experiments using different formats, ISA-Tab being supported through isatools. A JavaFX implementation independent of portal or data source is available on Github.

and displayed using Javascript libraries [15, 1]. For existing experiments, meta information about attached datasets is leveraged from the data store. When importing ISA-Tab investigations, the open-source framework isatools is used in the translation process, using source and sample identifiers of the ISA study format as well as all defined experimental factors. The Javascript libraries dagre and Data-Driven Documents (D^3) are then used to compute graph coordinates and draw the the selected graph. A stand-alone version implemented in JavaFX can be used independently of the portal or openBIS.

3 Results

To compare our implementation to the usual, complete sample graph of a project, we demonstrate both visualizations on a simple proteomics experiment. 24 mice were anesthetized for different periods of time, liver tissue was extracted and proteins from those tissue samples were measured using mass spectrometry. Fig. 1 shows a subset of the full sample graph. In contrast, the summarized experimental design graph of the same experiment seen in Fig. 3 gives a quick, condensed overview of the experiment hierarchy, if no factor is chosen. Metadata like sample identifiers can be shown by clicking on nodes of the graph. When the factor *anaesthesia_duration* is selected, our algorithm splits the graph into three groups of mice and descendant samples according to the three levels of this factor. Colors and legend are entirely dependent on the graph and inform users about species, tissues, analytes and different factor values. Furthermore, the green outline of the protein nodes show that data generation has been completed for all samples.



Experimental Factor
None

Experimental Factor
anaesthesia_duration

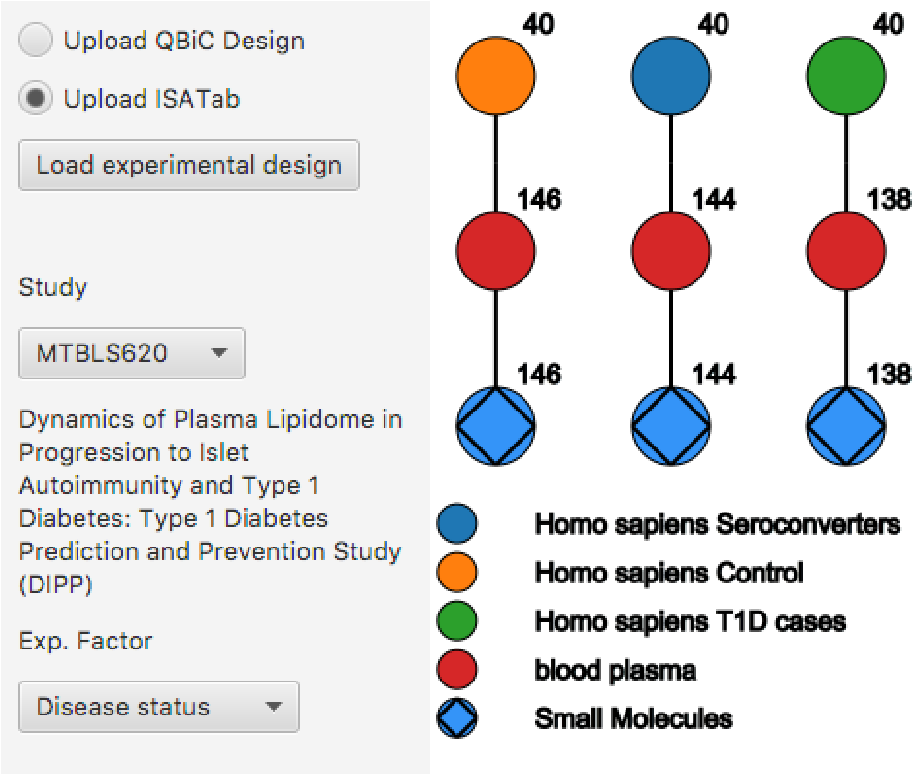
Fig. 3. Experimental design graphs of the same experiment as seen in qPortal. Numbers denote the amount of summarized samples for each factor and hierarchy level. Left: no experimental factor chosen. Right: nodes denoting mice and child nodes in the graph are split by experimental factor *anesthesia duration*.

We evaluate our stand-alone implementation using a recent lipidomics study on the progression to islet autoimmunity and type 1 diabetes taken from the MetaboLights database for metabolomics experiments [13, 11]. Our application shows a description of the imported ISA-study and lists every experimental factor that the authors have annotated in a drop-down menu. Selecting *disease status* as seen in Fig. 4 shows that data generated from the blood plasma samples of 40 patients of a control group were compared to those of 40 type 1 diabetes (T1D) cases, as well as 40 cases of autoimmunity against islet cells, that had not yet progressed to diabetes. Selecting the *age* factor reveals that this is a time series study, where blood was taken at different ages of patients. In this case the authors failed to include units in their metadata, so it is only clear from their publication that the ages are measured in months.

Our implementation is not limited to single-omics experiments. Fig. 6 shows a complex multi-omics study imported via the ISA-tab format. Since the experimental factor levels only differ between cell cultures, the graph stays connected on the species level. For studies of this complexity, zoom functionality can be used to show details.

3.1 Availability and License

The study aggregation graph is available through qPortal:
https://portal.qbic.uni-tuebingen.de/portal/web/qbic/software
A stand-alone JavaFX implementation and example studies are available on



○ Upload QBiC Design

● Upload ISATab

Study

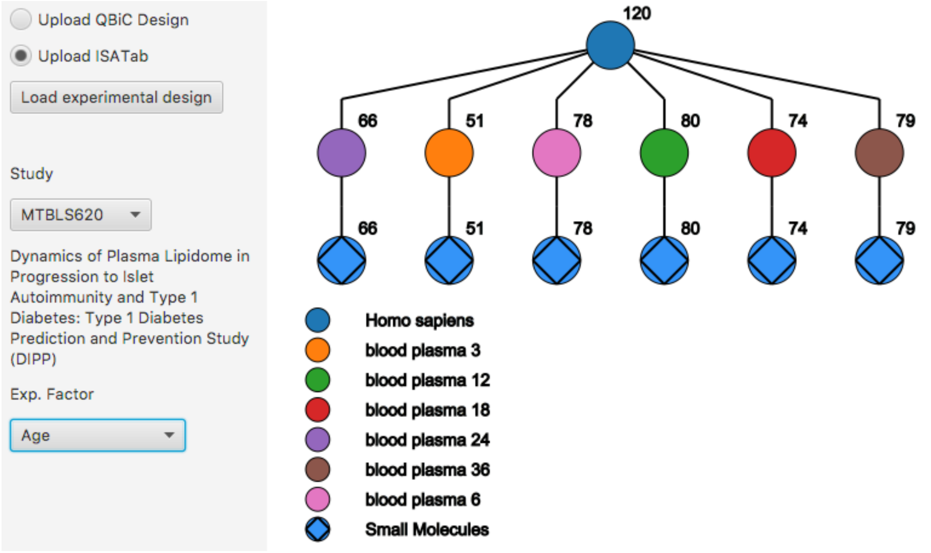
MTBLS620

Dynamics of Plasma Lipidome in Progression to Islet Autoimmunity and Type 1 Diabetes: Type 1 Diabetes Prediction and Prevention Study (DIPP)

Exp. Factor

Disease status

Fig. 4. View of the stand-alone application after importing an ISA-Tab folder and selecting a study as well as the experimental factor *disease status*. Information about the selected study is displayed. Users can change experimental factors from a drop-down menu. Our aggregation graph shows extraction of blood plasma from 120 patients belonging to the three groups control, type 1 diabetes (T1D) and seropositivity (for islet cell autoantibodies). The metabolome of those samples was measured.



○ Upload QBiC Design

● Upload ISATab

Study: MTBLS620

Dynamics of Plasma Lipidome in Progression to Islet Autoimmunity and Type 1 Diabetes: Type 1 Diabetes Prediction and Prevention Study (DIPP)

Exp. Factor: Age

Fig. 5. View of the stand-alone application after selection of the experimental factor *age*. The levels of this factor show that blood plasma samples were taken at different ages of the same patients, since the experimental levels are defined at the second level.

Github under the MIT license:
https://github.com/qbicsoftware/experiment-graph-gui
ISA-Tab files of the type 1 diabetes study are available on MetaboLights:
https://www.ebi.ac.uk/metabolights/MTBLS620

4 Discussion

We present tools to visualize large biomedical studies by their most important experimental aspects. Building on our graphical interface for the creation of factorial experimental designs and our hierarchical data model, we create graphs summarizing complex hierarchies of experimental variables, allowing users to quickly familiarize themselves with the important aspects of a study. When used in a platform integrating experiment data and metadata, like qPortal, additional information about datasets can be leveraged, marking missing data or the status of a project. Our aggregation graph gives a concise and intuitive overview in cases where representation of experiments was previously only possible using large tables.

The lack of statistical power and sound experimental design has lead to the so-called reproducibility crisis. Extensive work has been done to standardize metadata annotation and storage, leading to multi-omics standards like ISA-Tab, which not only stores metadata, but also provides a foundation to search, display and use these annotations. These methods are clearly required due to the

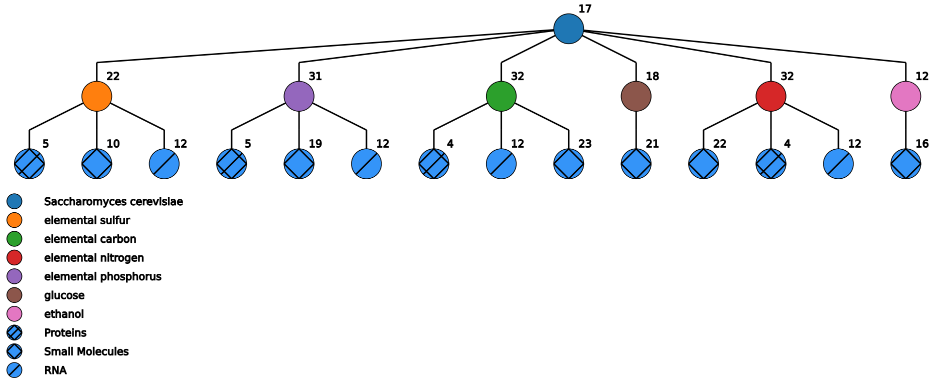


Fig. 6. Aggregation graph of one study of an imported ISA-tab investigation. Yeast cultures are grown lacking different nutrients and proteome, transcriptome and metabolome are measured.

size of modern biomedical experiments and their metadata: a simple overview a of study often leads to huge tables or cluttered graphs. Some approaches are examples of successful, interactive uses of study design visualization, yet they address very specific questions. Bio-GraphIIn [8] focuses on listing the replicates of each study group. By contrast, our approach supports the ISA-Tab format, provides an interactive visualization of a large number of experiments and is able to summarize replicates (with respect to one factor) into a single node to display a concise representation with which users can interact to control the displayed level of detail.

We have shown that our approach can display current studies including several hundred entities. Since ISA-Tab is not a minimum information standard, the amount of actual information beyond the sample hierarchy that can be drawn from its format depends on the annotations provided by researchers, as our example shows. We have taken the first steps towards a fully modular solution that will allow integration of our tool, enforcing standards that fit with their experimental data model.

Experimental factors are one of the most important type of study annotation, since they are at the core of the question scientists want to answer. However, our concept is not necessarily bound to the aggregation of different factor levels. Any property that can split subjects or samples in different groups, can be useful to find out more about a study. In large studies involving multiple groups, sharing information about the status of the project and data generation is often important. Provided this information is available, future work could include a time-component, displaying the history of a study.

References

1. Bostock, M., Ogievetsky, V., Heer, J.: D^3 data-driven documents. IEEE transactions on visualization and computer graphics **17**(12), 2301–2309 (2011)
2. Brazma, A.: Minimum information about a microarray experiment (miame)–successes, failures, challenges. The Scientific World Journal **9**, 420–423 (2009)
3. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., et al.: Minimum information about a microarray experiment (miame)toward standards for microarray data. Nature genetics **29**(4), 365–371 (2001)
4. Collins, F.S., Tabak, L.A.: Nih plans to enhance reproducibility. Nature **505**(7485), 612 (2014)
5. Fisher, R.: Introduction to the arrangement of field experiments. J Minist Agric G B **33**, 503–13 (1926)
6. Fisher, R.A.: The design of experiments. Oliver And Boyd; Edinburgh; London (1937)
7. Friedrich, A., Kenar, E., Kohlbacher, O., Nahnsen, S.: Intuitive web-based experimental design for high-throughput biomedical data. BioMed research international **2015** (2015)
8. Gonzalez-Beltran, A., Maguire, E., Georgiou, P., Sansone, S.A., Rocca-Serra, P.: Bio-graphiin: a graph-based, integrative and semantically-enabled repository for life science experimental data. EMBnet. journal **19**(B), pp–46 (2013)
9. González-Beltrán, A., Maguire, E., Rocca-Serra, P., Sansone, S.A.: The open source isa software suite and its international user community: knowledge management of experimental data. EMBnet. journal **18**(B), pp–35 (2012)
10. González-Beltrán, A., Maguire, E., Sansone, S.A., Rocca-Serra, P.: linkedisa: semantic representation of isa-tab experimental metadata. BMC bioinformatics **15**(14), S4 (2014)
11. Haug, K., Salek, R.M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., et al.: Metabolightsan open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic acids research **41**(D1), D781–D786 (2012)
12. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M.: Kegg for representation and analysis of molecular networks involving diseases and drugs. Nucleic acids research **38**(suppl_1), D355–D360 (2009)
13. Lamichhane, S., Ahonen, L., Dyrlund, T.S., Kemppainen, E., Siljander, H., Hyoty, H., Ilonen, J., Toppari, J., Veijola, R., Hyotylainen, T., et al.: Dynamics of plasma lipidome in progression to islet autoimmunity and type 1 diabetes: Type 1 diabetes prediction and prevention study (dipp). bioRxiv p. 294033 (2018)
14. Mohr, C., Friedrich, A., Wojnar, D., Kenar, E., Polatkan, A.C., Codrea, M.C., Czemmel, S., Kohlbacher, O., Nahnsen, S.: qportal: A platform for data-driven biomedical research. PloS one **13**(1), e0191603 (2018)
15. Pettitt, C.: dagre - graph layout for javascript. https://github.com/dagrejs/dagre (2014)
16. Rayner, T.F., Rocca-Serra, P., Spellman, P.T., Causton, H.C., Farne, A., Holloway, E., Irizarry, R.A., Liu, J., Maier, D.S., Miller, M., et al.: A simple spreadsheet-based, miame-supportive format for microarray data: Mage-tab. BMC bioinformatics **7**(1), 489 (2006)
17. Reips, U.D., Neuhaus, C.: Wextor: A web-based tool for generating and visualizing experimental designs and procedures. Behavior Research Methods, Instruments, & Computers **34**(2), 234–240 (2002)

18. Sansone, S.A., Rocca-Serra, P., Brandizi, M., Brazma, A., Field, D., Fostel, J., Garrow, A.G., Gilbert, J., Goodsaid, F., Hardy, N., et al.: The first rsbi (isa-tab) workshop:can a simple format work for complex studies?. OMICS A Journal of Integrative Biology **12**(2), 143–149 (2008)
19. Sansone, S.A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., et al.: Toward interoperable bioscience data. Nature genetics **44**(2), 121 (2012)
20. Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., et al.: Design and implementation of microarray gene expression markup language (mage-ml). Genome biology **3**(9), research0046–1 (2002)
21. Taylor, C.F., Paton, N.W., Lilley, K.S., Binz, P.A., Julian, R.K., Jones, A.R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E.W., et al.: The minimum information about a proteomics experiment (miape). Nature biotechnology **25**(8), 887–893 (2007)
22. Tyanova, S., Mann, M., Cox, J.: Maxquant for in-depth analysis of large silac datasets. In: Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC), pp. 351–364. Springer (2014)
23. Vasilevsky, N.A., Brush, M.H., Paddock, H., Ponting, L., Tripathy, S.J., LaRocca, G.M., Haendel, M.A.: On the reproducibility of science: unique identification of research resources in the biomedical literature. PeerJ **1**, e148 (2013)