



SetupX – a study design database compliant to the Metabolomics Standards Initiative (MSI)



UCDAVIS

Martin Scholz, Oliver Fiehn

UC Davis Genome Center, Davis, CA

Initial idea

Metabolomic data can only be interpreted on the basis of background information of the experimental design of the biological parameters that were studied, and the details of data acquisition and data processing. Metabolites, unlike proteins, genes or RNA molecules, do not commonly carry specific information content. Instead, the role of metabolites in biological processes needs to be unraveled by their changes in levels, turnover rates and location in response to influence factors such as perturbation of the genetic constitution or external stress treatments.

SetupX allows to capture all the relevant metadata / it actually enforces each researcher to design the experiment before running the samples. By capturing the data for a large number of studies we created a repository that combines all relevant data (metadata, annotation data) in one single location.

The system also works as a lab wide tracking system:

- Status of experiment / sample
- File tracking / repository for the samples
- Feeding samples into the machines
- Scheduling samples for post processing

Current approaches in metabolomic databases do not support such details in machine-readable and systematic formats, rendering SetupX a novelty in the field.



Standards required

In order to be able to query the entire dataset unique identifiers are required for the experimental design.

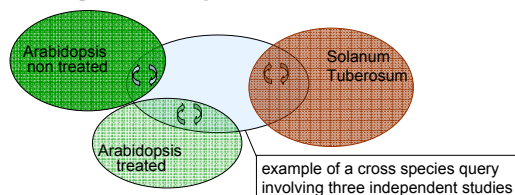
The NCBI Taxonomy is used for defining the species. Based on this the system adopts the user interface to the species - allowing different sets of required fields for each of the kingdoms. Plant ontology is used to define the organ of each plant while the system utilizes different ontologies for humans and animals.

A large number of the attributes which define an experiment are mandatory while a number of attributes is optional. Even custom attributes describing the sample can be added.

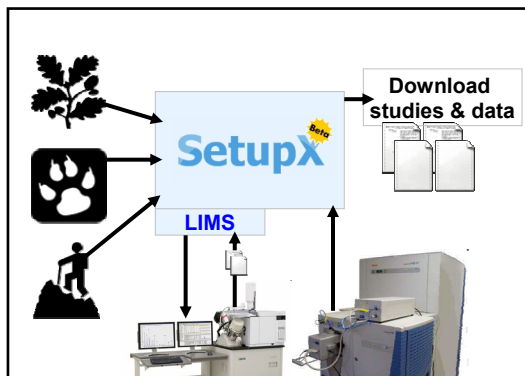
ID	class	label	comment	Gender	MSI
951472	arabid	ex1284	loc 1-45	male	Cell type
951473	arabid	ex1285	loc 1-32	female	Subcellular location
951480	arabid	ex1286	loc 1-414	male	Biobiosource annotat
951484	arabid	ex1287	loc 1-231	male	Growth support
951489	arabid	ex1288	loc 1-43	male	Growth location
951492	arabid	ex1289	loc 1-312	male	Growth pilot design
951496	arabid	ex1290	loc 1-432	female	Humidity regime
951500	arabid	ex1291	loc 1-424	male	Temperature at night
951504	arabid	ex1292	year comment	male	Working regime
951508	arabid	ex1293	year comment	male	Natural region
951512	arabid	ex1294	year comment	male	Country/continent/transporting
951516	arabid	ex1295	year comment	male	Harvest date, time
951520	arabid	ex1296	year comment	male	Plant growth stage
951524	arabid	ex1297	year comment	male	Metabolism quenching methods
951528	arabid	ex1298	year comment	male	Harvest method
951532	arabid	ex1299	year comment	male	Sample storage

Studies stored in the system can be downloaded. The system provides access to the data via XML, XLS, well formatted reports or even RSS. The XML data is structured using an XML Schema. All the download functionality is also available for a portion of the datasets stored in the repository that are declared as publicly available.

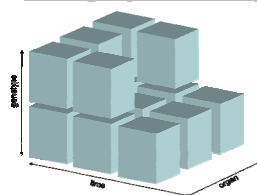
Query of multiple studies



Well annotated studies enable researchers to reuse data from multiple independent studies.



Design your study



SetupX aids researchers to carefully design and completely document their studies with the added benefit that studies are laid out in a systematic manner and can be queried on different levels (e.g. species, treatment or other search terms).

In order to capture the design the system has to determine what the variations in the study are.

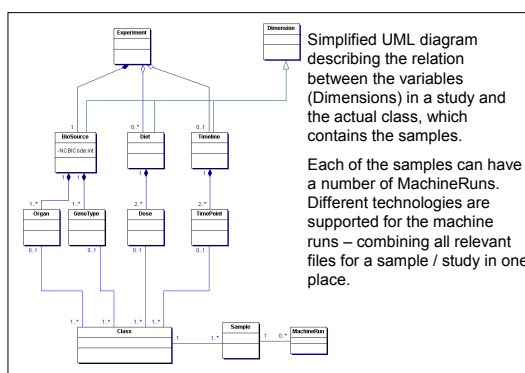
The numbers of each variations (represented by a dimension above) span a multidimensional cube.

(Genotype x organ) x (treatment)

Biosource

The number of dimensions in the cube is not limited and varies according to the experimental design of each study. Objects that cannot be distinguished by any of the vectors are called bioreplicates and belong into the same class yet have unique object identifiers.

BioSource and treatment define the most important vectors that span the dimensions of a metabolomic study. In principle, further dimensions can be spanned by using different chemical treatments or data acquisition methods, but mostly this is not intended in metabolomic studies.



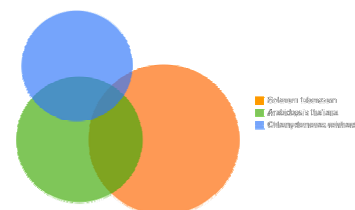
open source for download:
<http://setupx.googlecode.com/>

Repository



Kingdoms				
Number of Samples and Experiments per Kingdoms				
Kingdom	Samples	Experiments	Experiments	
plant	6692	58%	70	46%
microorganism	1685	14%	37	24%
animal	2368	20%	28	18%
human	711	6%	15	10%

A subset of 1,614 out of the total 11,102 samples is publicly available, including the complete experimental design, raw and annotated GC-TOF result data. Additional samples and studies become publicly available as laboratory collaborators publish their findings.



The Metabolites found in three different species. The diagram is generated in real-time and can be requested for every possible combination of species.

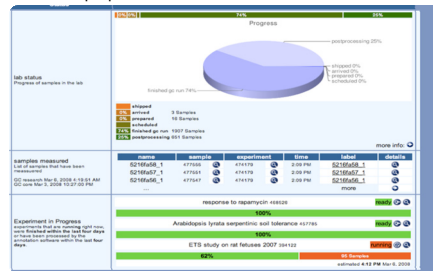
LIMS

Laboratory Information Management System

SetupX acts as a LIMS. It combines all files relevant to a study in a central place, enabling the operator to submit result data from different technologies (GCMS, LCMS or direct infusion FTMS) in a single environment. Scheduling and randomizing samples is directly connected to the Leco GCTOF mass spectrometer to aid laboratory assistants.



SetupX further enables the lab managers and administrators to gain overview of scheduled experiments, current data acquisition and status of data processing in real time. SetupX is seamlessly integrated to the BinBase GCTOF database, which uses the study design for statistical purposes.



Conclusions

SetupX therefore presents a fully functional and public database system integrating metabolomic workflows from conceptual design over laboratory practice to steering data processing tasks and result queries.

SetupX is an open source project and can be downloaded and installed under the LGPL license.

Funding:

This work was funded by NIEHS R01 ES13932.

References:

- Fiehn O, Wohlgemuth G, Scholz M, Kind T, Lee DY, Lu Y, Moon S, Nikolau BJ (2008) Quality control for plant metabolomics: Reporting MSI-compliant studies. *Plant Journal* 53, 691-704
- Scholz M, Fiehn O (2007) SetupX - A Public Study Design Database for Metabolomic Projects. *Pacific Symp. Biocomp.* 12, 169-180
- Fiehn O, Wohlgemuth G, Scholz M (2005) Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata. *Proc. Lect. Notes Bioinformatics* 3615, 224-239