Contact : walshs@ethz.ch

Adaptable integration, analysis and data-sharing of diverse multi-omic and low-throughput data: computational platforms from European consortia.

Walsh S.[1], Baerenfaller K. [1], Graf A. [1], Coman D. [1], Hirsch-Hoffmann M. [1], Önder K. [1], Sulpice R. [2], Szakonyi D. [4], Zielinski T. [5], Agron-omics and TiMet data contributors[1,2,3,4,5,+], Granier C. [3], Stitt M. [2], Millar A. [5], Hilson P. [4], Gruissem W. [1]

1. Department of Biology, ETH Zurich, Switzerland (add more)
2. Max Planck Institute of Molecular Plant Physiology, Golm, Germany
3. Laboratoire d'Ecophysiologie des Plantes sous Stress Environnementaux (LEPSE), INRA-AGRO-M, Montpellier Cedex, France
4. Department of Plant Systems Biology, VIB, Gent, Belgium
5. SynthSys, University of Edinburgh, The Kings Buildings, Mayfield Road, EH9 3JD, Edinburgh, UK

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

Plant Biotechnology

A substantial part of the **AGRON-OMICS** consortium is devoted to profiling the growing Arabidopsis leaf in a number of environmental conditions. The **TiMet** consortium studies the link between circadian clock and metabolism, focused on both primary- and isoprenoid-metabolism. These international multi-institute projects generate a **diverse range of quantitative molecular and phenotypic data**. Vital to our analytical pipelines are **adaptable database integrations** that exploit standard and advanced features of the MySQL database engine and tools. These implementations are utilized for the **processes of data and meta-data capture, validation, the tracking of provenance, for certain statistical-, mathematical-, and structural data transformations, for integration with R and for generating visualizations**. Our systems provide access controlled **user workspaces** and the ability to **run high performance queries** across multiple and some high volume data sets. Interpreting novel datasets also requires the integration of **pre-existing knowledge** and consequently a range of annotations and classifications are included. Where detailed annotations were lacking, the **Knowtator** tool was used for curating **phenotype-genotype-environment relations using ontologies**. A number of scientific use-cases are presented that demonstrate the pivotal role that coherent integration can play in data quality control, project management and data analysis. Since the database engine and tools are freely available, the data, code and documentation can be simply and rapidly replicated for community dissemination and/or extension. These developments provide a useful template for a computational platform that has analytical value during a project and beyond.

Distributed research consortia need:

- Shared access to versioned, structured and validated data
- Ability to query across all data and meta-data
- Flexibility to rapidly encode new questions

Data Overview

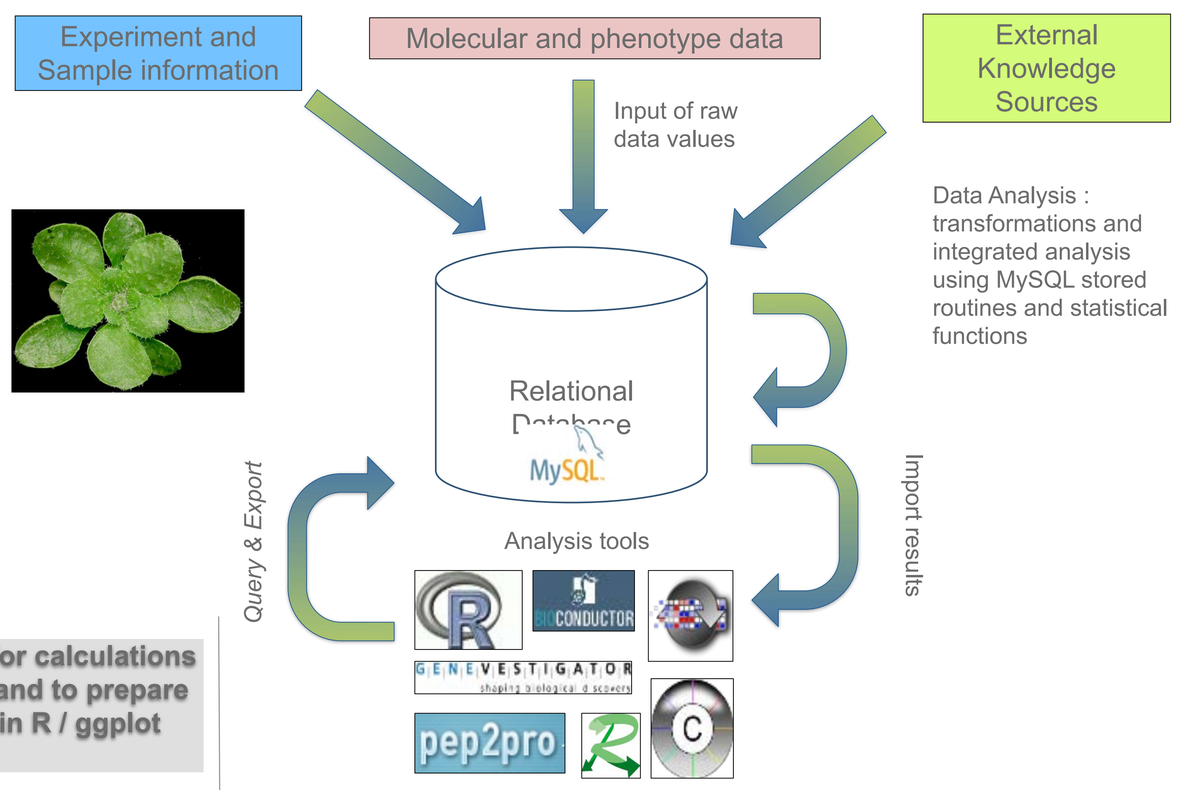
High-throughput data-sets include:

- Transcriptome (ATH1 and AGRONOMICS1 tilling arrays)
- Proteome (iTRAQ and label free MS-MS)
- Primary, secondary and lipids metabolites (GC-MS/MS)
- Epigenome methylation states

Additionally, a diverse range of lower throughput data are generated including :

Visible phenotypes, enzyme activities, metabolite quantifications by spectrophotometry, plant dry weight/fresh weight, endoreduplication number, photosynthesis and respiration rates, a variety of quantitative PCR data-sets (rRNAs, mitochondrial/chloroplast transcripts, genes).

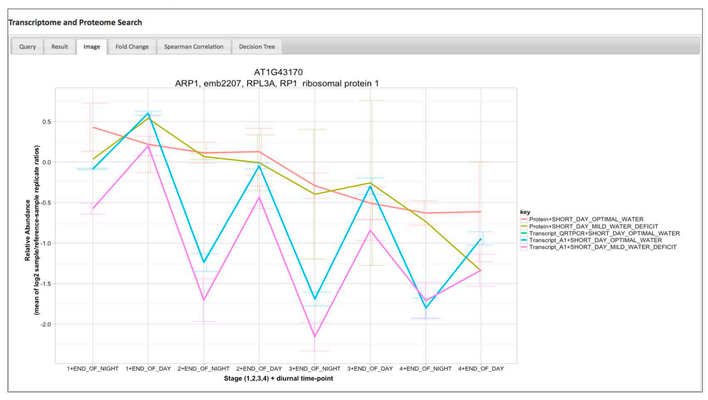
Strategy overview



Integration of proteome, transcriptome and phenotype

Integration of quantitative protein and transcript abundances together with phenotype data enables a systems understanding of leaf development through growth, expansion and diurnal time in two environmental conditions.

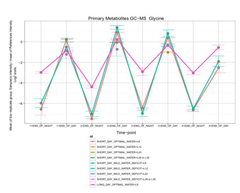
Systems-based analysis of Arabidopsis leaf growth reveals adaptation to water deficit. Molecular Systems Biology (2012) 8
https://www.agronomics.ethz.ch/

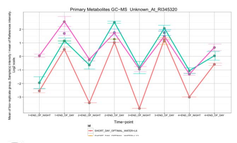


Integrating metabolite quantifications (from spectrophotometry and MS) with enzyme activities

Having multiple integrated datasets available to all project partners through a project data warehouse enables and encourages a variety of data mining strategies. For example, metabolites and enzymes activities were analyzed using a variety of methods including profile plots, ANOVA, PCA, correlation matrices and with a variety of clustering methods. Availability of consistent summary statistics (biological replicate values/variances) together with individual sample measures for data-sets is essential to allow the comparison of results based on a single version of the data.

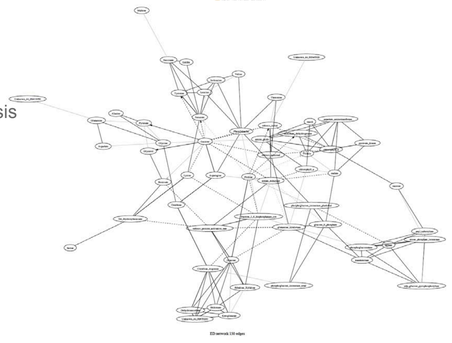
Primary metabolism: metabolites and enzyme activities. ED only. distance=manhattan cluster=complete.





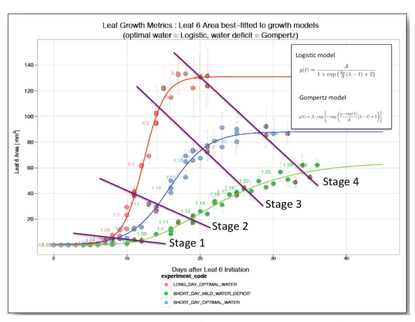
Agron-omics metabolite profiles in 3 environmental conditions

Agron-omics : principle component analysis

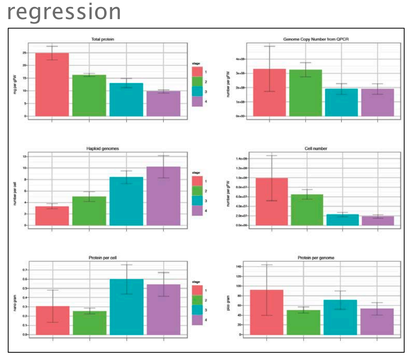


Agron-omics : causal network inference with GeneNet

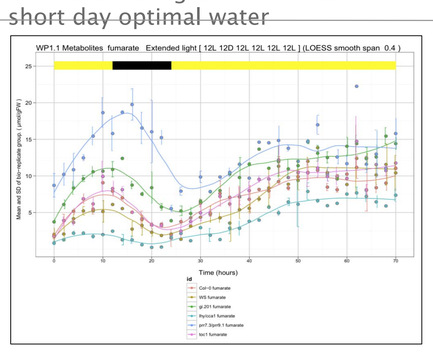
Stored routines are used for calculations across multiple datasets and to prepare data for plotting with thin R / ggplot clients.



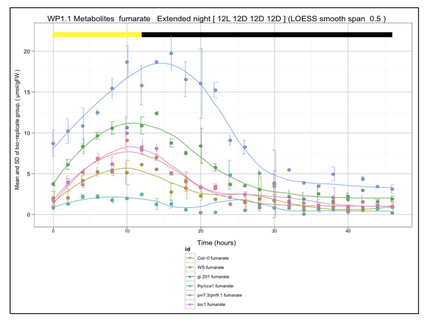
Agron-omics Leaf 6 area non linear regression



Agron-omics observed and calculated leaf growth metrics in short day optimal water

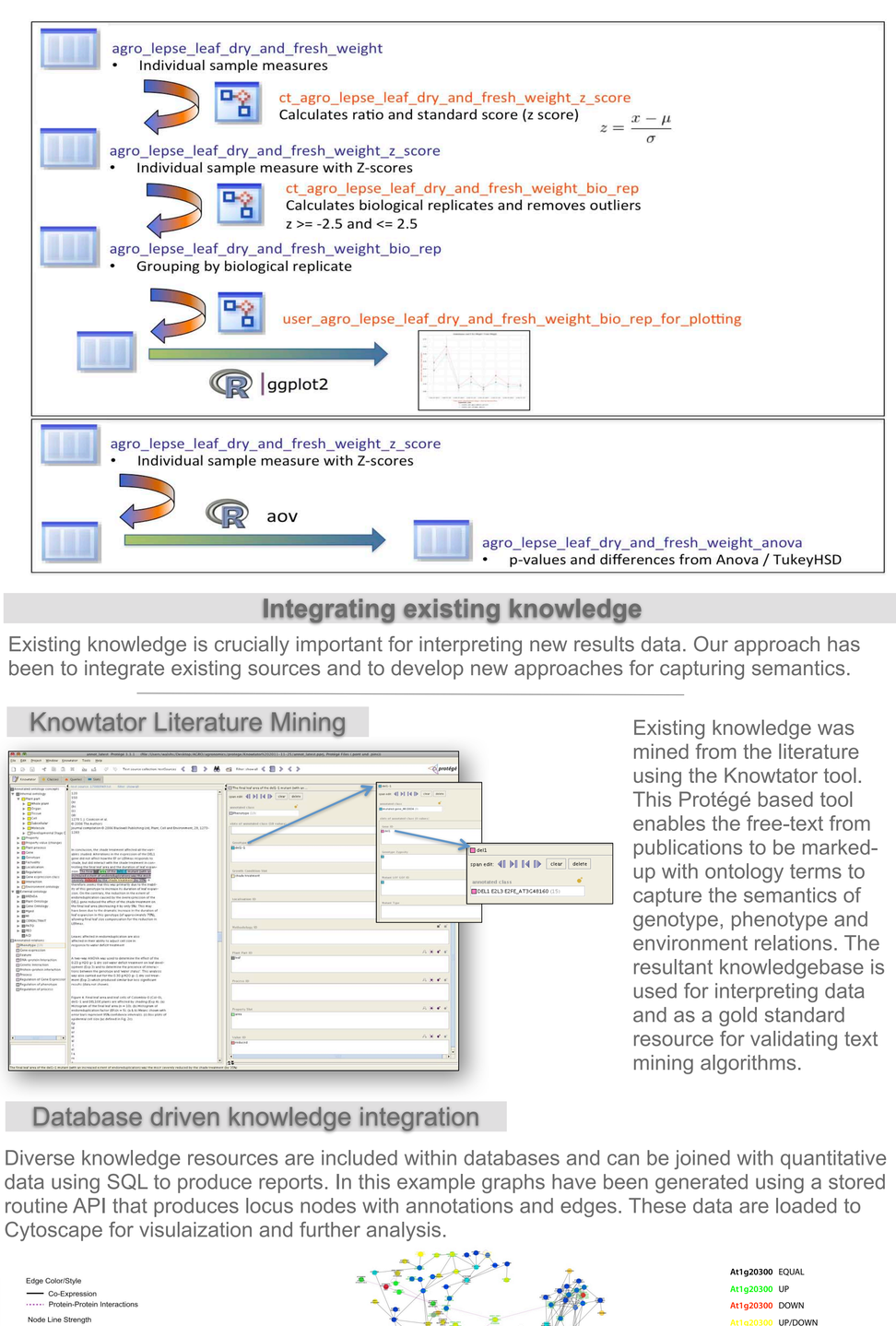


TiMet: metabolite profiles under variable light regimes



These databases use freely available tools to integrate all data and meta-data from the project. The approach provides user workspaces and interfaces to enable data analysis and bespoke visualizations within the project and for data re-use by the community.

Example data processing pipeline using database stored routines



Integrating existing knowledge

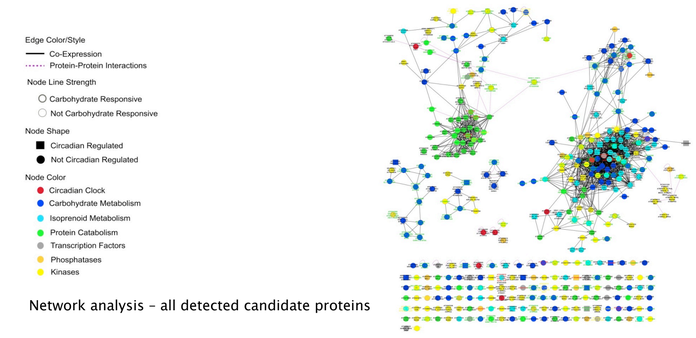
Existing knowledge is crucially important for interpreting new results data. Our approach has been to integrate existing sources and to develop new approaches for capturing semantics.

Knowtator Literature Mining

Existing knowledge was mined from the literature using the Knowtator tool. This Protégé based tool enables the free-text from publications to be marked-up with ontology terms to capture the semantics of genotype, phenotype and environment relations. The resultant knowledgebase is used for interpreting data and as a gold standard resource for validating text mining algorithms.

Database driven knowledge integration

Diverse knowledge resources are included within databases and can be joined with quantitative data using SQL to produce reports. In this example graphs have been generated using a stored routine API that produces locus nodes with annotations and edges. These data are loaded to Cytoscape for visulaization and further analysis.



Network analysis - all detected candidate proteins