iHUB - An Information and Collaborative Management Platform for Life Sciences

David Salt[1] Mourad Ouzzani[2] Eduard Dragut[3] Peter Baker[4] Srivathsava Rangarajan[4]
[1]University of Aberdeen, Institute of Biological and Environmental Sciences
[2]Qatar Computing Research Institute
[3]Temple University, Computer and Information Sciences Department
[4]Purdue University, Cyber Center, Discovery Park
[1]david.salt@abdn.ac.uk [2]mouzzani@qf.org.qa [3]edragut@temple.edu
[4]{pnbaker, rangars}@purdue.edu

1. INTRODUCTION

We describe **ionomicshub**, iHUB for short, a large scale cyber-infrastructure to support end-to-end research that aims to improve our understanding of how plants take up, transport and store their nutrient and toxic elements.

One of the biggest challenges in the life sciences is *functional genomics*, which is the study of the molecular mechanistic functional contribution of each of the genes in a genome to the overall biology of an organism. *Ionomics* is one example of a functional genomics approach for the study of the genes and gene-networks that control mineral nutrient and trace element *homeostasis* (which is the property of a system to regulate its variables so that internal conditions remain stable and relatively constant in response to external conditions, e.g., temperature regulation). Quantitative and qualitative high-throughput molecular phenotyping tools have been developed, which, when coupled to similarly high-throughput genotyping tools, allow gene-to-function connections to be made rapidly. The advent of these high-throughput technologies has created a deluge of information that is challenging to deal with, not only because of its sheer volume, but also because of the difficulties in interpreting the various measurements in the context of different genotypes and organismal physiologies.

iHUB provides integrated workflow control, data storage, and analysis to facilitate high-throughput data acquisition, along with integrated tools for data search, retrieval, and visualization for hypothesis development. It also consists of tools for data discovery, data sharing, data annotation, work groups, communication, literature filtering, ranking and sharing, and data collection. iHUB is deployed as a Web-enabled system, allowing for integration of distributed workflow processes and open access to raw data for analysis by numerous laboratories. iHUB is accessible at `www.ionomicshub.org`.

Contribution. iHUB is an example of a cyber-infrastructure collaboratively developed by Web, database, knowledge management, visualization and biology communities for scientists in areas such as biology, agriculture and medicine. iHUB integrates cyber-infrastructure with human interactions to maximize community access to ionomic resources, and knowledge extraction from these resources. Specifically, iHUB is equipped with (1) a laboratory information management system (from planting stage to sample analysis stages), (2) enhance collaborative activity tools (e.g., filtering and sharing of literature) and (3) access to shared ionomic resources. iHUB allows laboratories without the capacity to perform ionomic analysis to access ionomic data on many plants, such as rice, soybeans, yeast and maize to directly test gene function, perform analysis and identify potentially interesting mutants for further study.

iHUB Usage. iHUB has been used by 10,496 people from 100 countries. Many scientific findings have been possible because of iHUB: there are at least 10 scientific publications from independent research outside of our group that used iHUB and 18 from our group (`www.ionomicshub.org/mediawiki/index.php/PiiMS_Publications#Review_Publications`).

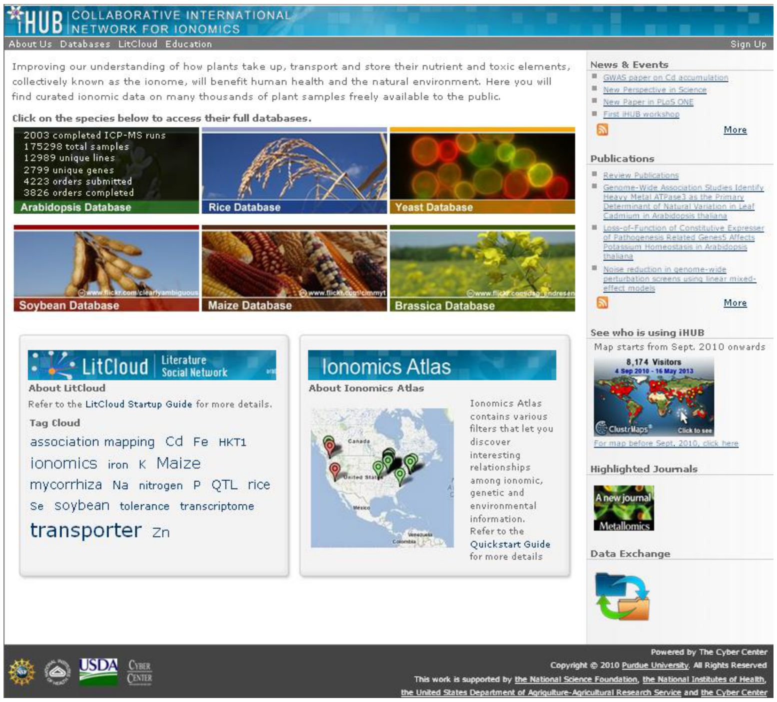
Additional Demo Material. Additional demo material can be accessed from the main Web page of iHUB by clicking on the menu item *Education* (Figure 2). We provide videos and documentation about using iHUB for specific tasks such as submitting seeds for analysis and database search.

Numerous researchers from the Web, information retrieval and knowledge management communities are actively involved in life science projects and we believe that our tool is of interest to them.

2. IHUB KEY COMPONENTS

The key components of iHUB are organized into two subsystems (Figure 1): the Laboratory Information Management System (LIMS) and Community Information Management System (CIMS). LIMS models the physical work flow in a laboratory and divides it into stages, based on the physical activity, information needs and data generated throughout the experimental life cycle. CIMS provides a web-enabled communication infrastructure for the ionomics

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577033.



Improving our understanding of how plants take up, transport and store their nutrient and toxic elements, collectively known as the ionome, will benefit human health and the natural environment. Here you will find curated ionomic data on many thousands of plant samples freely available to the public.

Click on the species below to access their full databases.

2003 completed ICP-MS runs
175298 total samples
12989 unique lines
2799 unique genes
4223 orders submitted
3826 orders completed

About LitCloud

Refer to the LitCloud Startup Guide for more details.

Tag Cloud

association mapping Cd Fe HKT1 ionomics iron K Maize mycorrhiza Na nitrogen P QTL rice Se soybean tolerance transcriptome transporter Zn

About Ionomics Atlas

Ionomics Atlas contains various filters that let you discover interesting relationships among ionomic, genetic and environmental information. Refer to the Quickstart Guide for more details

News & Events

- GWAS paper on Cd accumulation
- New Perspective in Science
- New Paper in PLoS ONE
- First iHUB workshop

More

Publications

- Review Publications
- Genome-Wide Association Studies Identify Heavy Metal ATPase3 as the Primary Determinant of Natural Variation in Leaf Cadmium in Arabidopsis thaliana
- Loss-of-Function of Constitutive Expresser of Pathogenesis Related Genes5 Affects Potassium Homeostasis in Arabidopsis thaliana
- Noise reduction in genome-wide perturbation screens using linear mixed-effect models

More

See who is using iHUB

Map starts from Sept. 2010 onwards

For map before Sept. 2010, click here

Highlighted Journals

Data Exchange

Powered by The Cyber Center
Copyright © 2010 Purdue University. All Rights Reserved
This work is supported by the National Science Foundation, the National Institutes of Health, the United States Department of Agriculture-Agricultural Research Service and the Cyber Center

Figure 2: The main Web page of the iHUB. iHUB hosts ionomics data about plants such as rice, yeast, soybean, and maize. It allows easy access to data and provides simple community building tools. A summary of each of the datasets is displayed when the mouse pointer is moved over the images (e.g., A. thaliana summary)

.

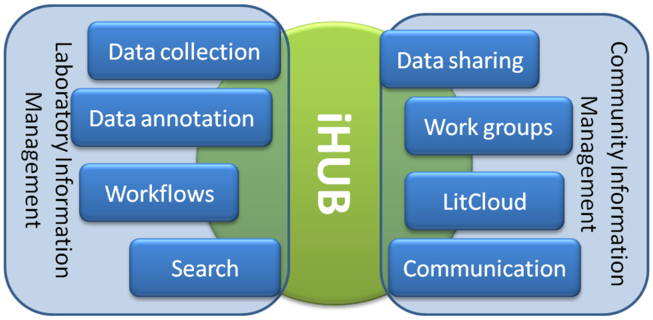


Figure 1: iHUB architecture view

.

community to efficiently probe function on a genomic and comparative genomic scale. We describe them below.

2.1 LIMS Components

We describe the key components of LIMS in this section.

2.1.1 *Data Collection*

iHUB is used daily to manage a continuous functional genomics pipeline containing over 250,000 active samples. iHUB manages plant growth, harvest, sample preparation, and ICP-MS analysis and is used daily to analyze and store over 2,000 fully quantified elemental concentrations.

iHUB allows scientists to submit lines for ionomic analysis within the pipeline [2]. (An account is needed.) iHUB supplies data input portal for the entry of critical information defining each line to be submitted, including line name, genetic structure, mutation type, background accession, and gene mutated. A *quality control component* assists the customer along the way. For instance, to prevent typographic or transposition errors in entering these lines, the system checks the user's submission against standardized resources (e.g., the T-DNA insertional mutant collection curated at SIGnAL and the Arabidopsis accessions at the ABRC), alerts the user if there is a discrepancy, and suggests lines from the locally stored catalog that have fields that match those submitted by the user. Submitted customer orders are tracked and reviewed by the iHUB system administrator, who checks that submitted lines are appropriate for the system and that the submitted metadata are correctly formatted. Upon completion of the analysis the data is ingested in the iHUB database [2, 1] and the customer is notified.

2.1.2 *Search*

iHUB allows users with diverse backgrounds to retrieve data based on familiar parameters: e.g., a biologist can query the data by genetic associations and/or soil properties, while an environmentalist can query by climate parameters. iHUB provides three search means: Basic, Advanced and Map-based. Public users have open access to the Basic Search mode. In the basic mode a user can search the database based on ionomic phenotype, gene, line, experiment number, or order number. This mode also allows scientists to view the data from their completed sample submissions. The Advanced Search mode allows the construction of extensive Boolean queries using multiple indexes, including Mutant Type, Ecotype, Gene Name, Gene ID, and Experiment Number and Range. Returned data are formatted for download onto a local machine as a commadelimited text file. Currently, the Advanced Search is only accessible to expert users who have log-on privileges. The Map-based mode is accessible from the Ionomic Atlas [3] (Figure 2, bottom). This is a complex query interface (`www.ionomicshub.org/ionomicsatlas/`) that allows a user to query the database based on a combination of criteria including: Accession, Gene, Ionomic profile, Geographic location, Growth conditions, Climatic conditions. The data is displayed on a map. The user can click on one of the markers on the map, to obtain a description of the accession residing at those coordinates on the map.

2.1.3 Data Annotation

We use *bdbms*, an extensible database engine for biological databases [5, 4] to support flexible annotations of iHUB data. We have implemented several features in iHUB, such as annotation and provenance management, local dependency tracking, and update authorization. In particular, *bdbms* treats annotations as first class objects and allows adding annotations at multiple granularities of a relational schema (table, tuple, column, and cell levels), archiving and restoring annotations, and querying the data based on the annotation values.

2.1.4 Workflows

iHUB supports the workflow defined by the different stages that take place in an ionomics laboratory. These different stages correspond to order submission, order approval, planting, harvesting, drying, mass spectrum analysis (i.e., biologists and organic chemists use it to elucidate the structure of chemical compounds), internal review, and data release. At any given point in time, concurrent users are provided with different tools to carry out those tasks and check their status. Data entering the system undergoes multiple levels of translation and validation before finally being stored in a relational format. iHUB captures all necessary data and metadata, including laboratory management data, customer orders, experiments, including mass spectrometry and analysis data.

2.2 CIMS Components

We present the components of CIMS in this section.

2.2.1 Data Sharing

To foster the collaborative process among scientists iHUB allows compiling data, making it available to the community (via DOI - Digital Object Identifier), and using it in experiments. In iHUB, the shared data appears in the collaborators "InBox" and arrival of shared data is signaled by both email and visual notification for users within the iHUB environment. iHUB allows collaborators to discuss the shared data. Collaborators can attach text to the shared experiment describing their thoughts on the data; this text could then be viewed anytime. All text communication (IM and attached annotation) about a particular experiment is saved and be viewable by the individuals that are directly collaborating. iHUB contains the infrastructure to control access privileges and to define secure collaborative workspaces.

2.2.2 Literature Cloud (LitCloud)

Thousands of papers are published every month that are relevant to a scientist. Most scientists can only scan a small fraction of those and as such miss a significant number of papers that could be of use to their work. RSS (Really Simple Syndication) feeds provide a granular way to track a myriad of information sources on the web and can be collated with programs like Google Reader. **LitCloud** is an innovative platform to harvest community-contributed literature references and utilize the "social engagement" the community has with these references, through the RSS actions of a community researchers, to filter, rank, and facilitate their discovery and use. A person using LitCloud can evaluate and annotate the content of publications by adding a star to specify that a reference important, a tag to assign a descriptive keyword to the reference, by commenting or by emailing references to colleagues. LitCloud aims to harvest the references and this annotation information; capturing the users' actions and aggregating these actions across multiple users to provide high quality literature ranking and filtering services for the whole iHUB community. Because of the annotation activities of the multiple members of the LitCloud, the literature content shared through LitCloud is considerably more focused and of a better quality than the original RSS feeds, providing better, more relevant literature recommendations to members.

2.2.3 Work Groups

iHUB allows the self formation of online work groups and online communication. Users with common interests or whom are working on the same project are able to create their own groups within iHUB. For example, a user can use the search tools (Section 2.1.2) to create an experiment dataset. Around the experiment dataset the user can initiate a number activities: (1) a wiki space where members can easily create private or public wiki pages and link them together based on their needs; (2) a discussion forum where group members can discuss topics of interest to the experiment; and (3) a group calendar to coordinate events and meetings.

2.2.4 Data Display

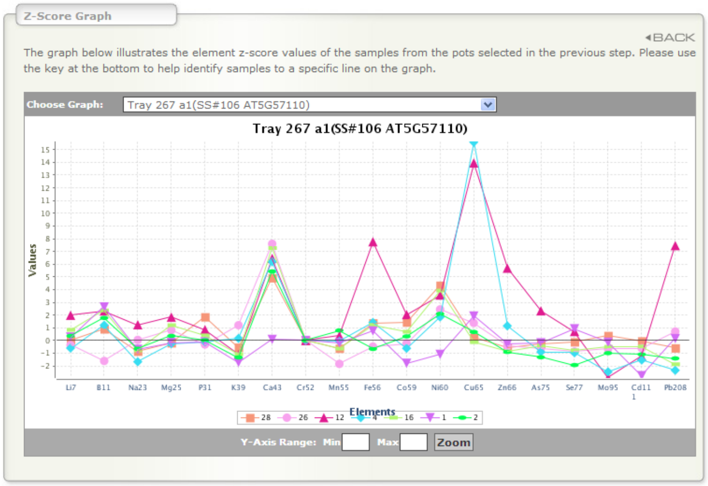
iHUB has various tools to summarize, visualize, and download the data. It is equipped with access to plotting and descriptive statistical analysis tools (Figure 3): e.g., z-scores (number of standard deviations from the mean) and percentage difference (from the mean). For example, for each query, iHUB can display data in tabular and chart views or even on a map [3]. The table displayed is quite flexible. One can sort by any of the fields. The charts are mostly design to contrast visually two sets of data, e.g., the data retrieved in response to a query versus the entire data.

3. IHUB IMPLEMENTATION

The project is implemented using open source software and free Web APIs. The back end is a PostgreSQL database. We use Apache Tomcat Java for the application server. Most of the code to access the database is developed using a combination of Java and scripting languages. The implementation of the front end utilizes a number of open source APIs. We develop the plotting tools with Open Flash Chart[1]. The geographic map browser is developed with the Google Earth API[2], which allows displaying a Google Earth interface in a web browser while enabling powerful rendering capabilities.

[1] `http://teethgrinder.co.uk/open-flash-chart-2`

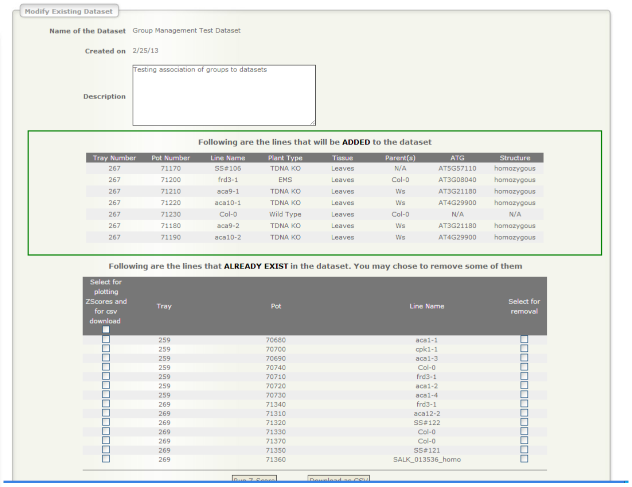
[2] `developers.google.com/earth/`



The graph below illustrates the element z-score values of the samples from the pots selected in the previous step. Please use the key at the bottom to help identify samples to a specific line on the graph.

Choose Graph: Tray 267 a1(SS#106 AT5G57110)

Figure 3: A screen shot of the plotting functionality.



Name of the Dataset Group Management Test Dataset

Created on 2/25/13

Description Testing association of groups to datasets

Following are the lines that will be **ADDED** to the dataset

Tray Number	Pot Number	Line Name	Plant Type	Tissue	Parent(s)	ATG	Structure
267	71170	SS#106	TDNA KO	Leaves	N/A	AT5G57110	homozygous
267	71200	frd3-1	EMS	Leaves	Col-0	AT3G08040	homozygous
267	71210	aca9-1	TDNA KO	Leaves	Ws	AT3G21180	homozygous
267	71220	aca10-1	TDNA KO	Leaves	Ws	AT4G29900	homozygous
267	71230	Col-0	Wild Type	Leaves	Col-0	N/A	N/A
267	71180	aca9-2	TDNA KO	Leaves	Ws	AT3G21180	homozygous
267	71190	aca10-2	TDNA KO	Leaves	Ws	AT4G29900	homozygous

Following are the lines that **ALREADY EXIST** in the dataset. You may chose to remove some of them

Select for plotting ZScores and for csv download ☐	Tray	Pot	Line Name	Select for removal
☐	259	70680	aca1-1	☐
☐	259	70700	cpk1-1	☐
☐	259	70690	aca1-3	☐
☐	259	70740	Col-0	☐
☐	259	70710	frd3-1	☐
☐	259	70720	aca1-2	☐
☐	259	70730	aca1-4	☐
☐	269	71340	frd3-1	☐
☐	269	71310	aca12-2	☐
☐	269	71320	SS#122	☐
☐	269	71330	Col-0	☐
☐	269	71370	Col-0	☐
☐	269	71350	SS#121	☐
☐	269	71360	SALK_013536_homo	☐

Figure 4: A screen shot of the data sharing.

4. DEMO PLAN

The goal of our demo plan is to demonstrate to visitors that indeed iHUB is a step forward in the ongoing endeavour of the scientific community to build tools that allow scientific discovery through data exploration and community collaboration [6]. We describe here several demo scenarios we will prepare for our audience that showcase iHUB as an enabling platform for science that requires the collaboration of people with different backgrounds and the integration of several Web and data management technologies.

4.1 Datasets

We will first explain to our audience the kind of datasets iHUB is currently hosting: accessions, ionomics, climate, soil and annotations. For that part of our audience that is not familiar with this kind of data, we will give them details about the datasets. Our main goal however is to convey to our audience the combined scientific potential of these datasets when they can be studied together under a coherent cyber-infrastructure, which would otherwise be missed.

4.2 Data Sharing

In this demo scenario we will show how users of iHUB can search the database, find data they can use for research or publication purposes and bookmark this data in the system (Figure 4). The bookmarked data may be modified and used as sample data in publications. The datasets can be created from the search results, have annotations on them, exchange these datasets amongst registered users of the iHUB, allow changes to them and finally publish them using DOIs for citation purposes. Compiling a new experiment dataset follows the shopping cart paradigm in iHUB. A user starts by searching for the needed data and whenever a piece of data meets user's requirements, it is added to the "shopping cart". A user may also remove entries from a dataset. Alternatively, a user may start creating a new experiment data from an existing dataset. (iHUB allows both public and private users to explore ionomic data collected in iHUB, which are cited in journals or part of research endeavors. This is accessible at `www.ionomicshub.org/home/PiiMS/dataexchange`). In this case, a copy of the original data is created and the user builds the new dataset from the copy. Finally, a dataset may be "frozen" so that it may be used as supporting documentation to any publication. This process requires getting a DOI for the data and maintaining the pdf file containing this data at a location referenced by dx.doi.org.

4.3 Work Groups

We will show how work groups are created and managed in iHUB. Groups can be created only by admin and advanced users. A group is designated an owner of a dataset. A group may be the owner of several datasets. We will show how work groups are applied to (experiment) datasets. We will also illustrate how the members of a work group associated to an experiment data are notified about the changes made to the data, including the annotations and comments, and who in the group made the changes.

We will encourage an interactive and inquisitive demo session for each visitor. We believe that the amount of rich features of the iHUB will results in a lively demo session.

Acknowledgements: This work was supported in part by the US NSF grants 0953433, 0701119, 0916614, and 1117766, and by Cyber Center, Purdue University.

5. REFERENCES

[1] I. Baxter and et al. A coastal cline in sodium accumulation in arabidopsis thaliana is driven by natural variation of the sodium transporter athkt1;1. *PLoS Genet*, 6, Nov. 2010.

[2] I. Baxter, M. Ouzzani, S. Orcun, B. Kennedy, S. S. Jandhyala, and D. E. Salt. Purdue ionomics information management system. an integrated functional genomics platform. *Plant Physiology*, 143, Feb. 2007.

[3] E. C. Dragut, M. Ouzzani, A. Madkour, N. Mohamed, P. Baker, and D. E. Salt. Ionomics Atlas: a tool to explore interconnected ionomic, genomic and environmental data. In *CIKM*, 2012.

[4] M. Y. Eltabakh, W. G. Aref, A. K. Elmagarmid, M. Ouzzani, and Y. N. Silva. Supporting annotations on relations. In *EDBT*, 2009.

[5] M. Y. Eltabakh, M. Ouzzani, W. G. Aref, A. K. Elmagarmid, Y. Laura-Silva, M. U. Arshad, D. E. Salt, and I. Baxter. Managing biological data using bdbms. In *ICDE*, pages 1600–1603, 2008.

[6] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.