JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

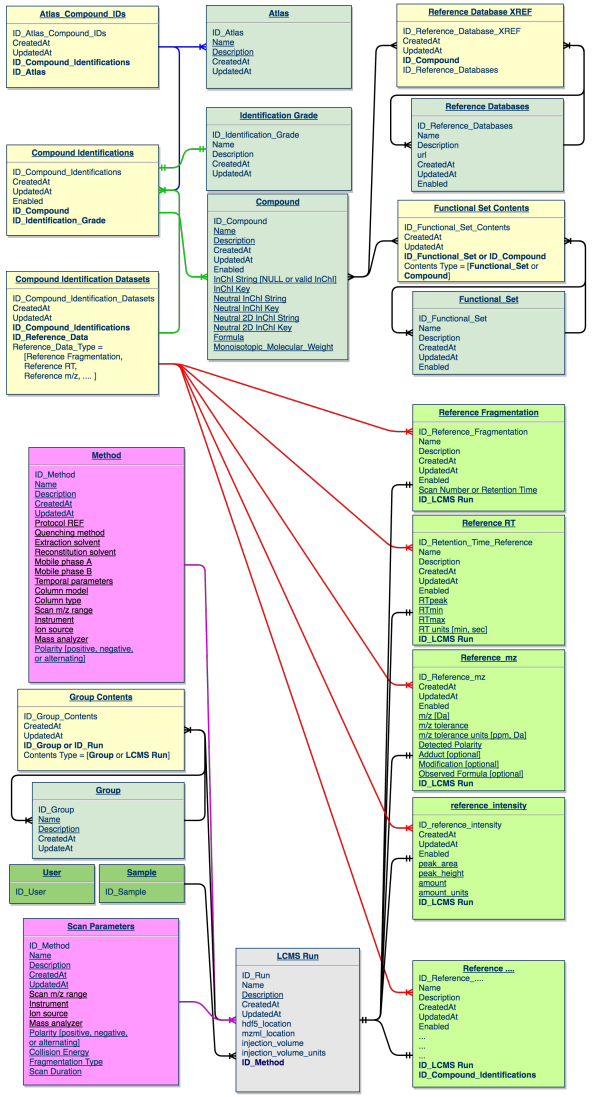
Linking Sequence to Metabolic Function

Informatics

March 22, 2016
Ben Bowen, PhD

Metabolomics & genomics integrated

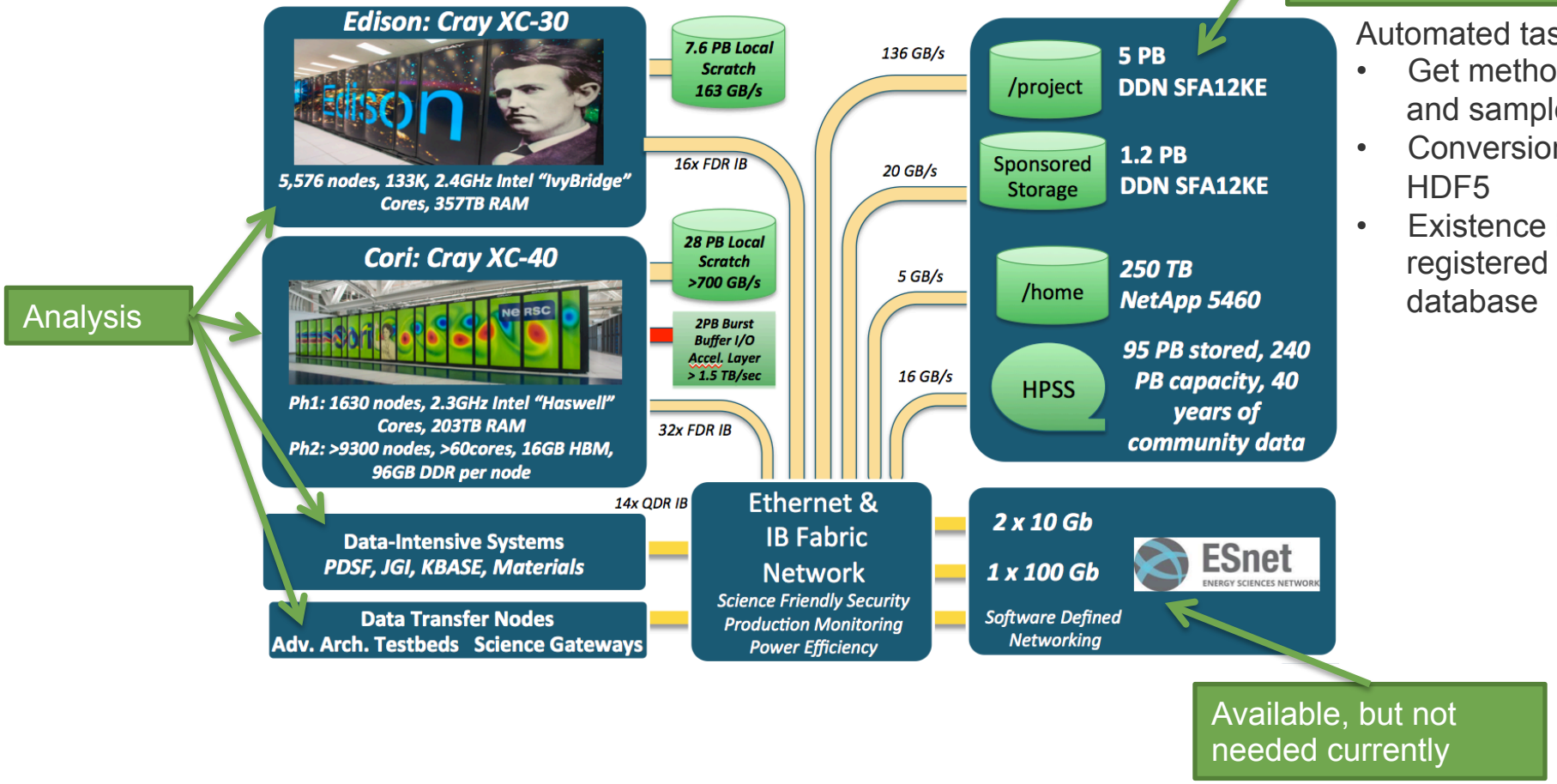
- Computational ecosystem is at the scale needed for JGI
- Maintain information to accelerate future experiments
- Metabolite discovery tools



1. Bowen, B. P., & Northen, T. R. (2010). Dealing with the unknown: metabolomics and metabolite atlases. *Journal of the American Society for Mass Spectrometry*, *21*(9), 1471–1476.
2. Fischer, C. R., Ruebel, O., & Bowen, B. P. (2016). An accessible, scalable ecosystem for enabling and sharing diverse mass spectrometry imaging analyses. *Archives of Biochemistry and Biophysics*, *589*, 18–26.
3. Wang, Y., Kora, G., Bowen, B. P., & Pan, C. (2014). MIDAS: a database-searching algorithm for metabolite identification in metabolomics. *Analytical Chemistry*, *86*(19), 9496–9503.
4. Yao, Y., Bowen, B. P., Baron, D., & Poznanski, D. (2015a). SciDB for High-Performance Array-Structured Science Data at NERSC. *Computing in Science & Engineering*, *17*(3), 44–52.
5. Yao, Y., Sun, T., Wang, T., Ruebel, O., Northen, T., & Bowen, B. P. (2015b). Analysis of Metabolomics Datasets with High-Performance Computing and Metabolite Atlases. *Metabolites*, *5*(3), 431–442.

Metabolite Atlas

Quick and Painless: Manage raw LCMS data

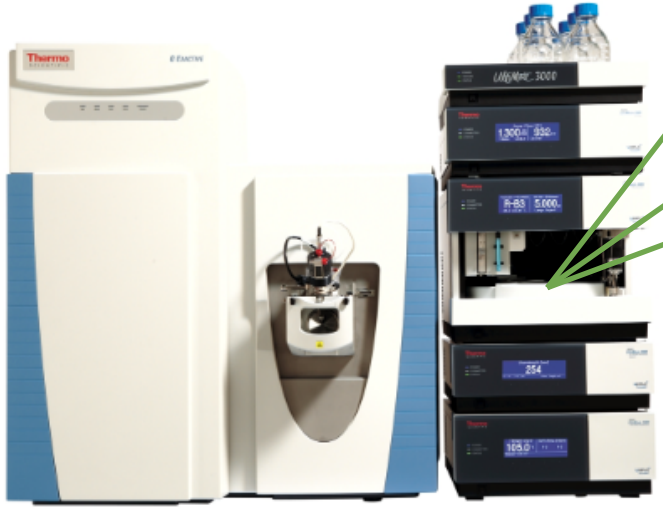


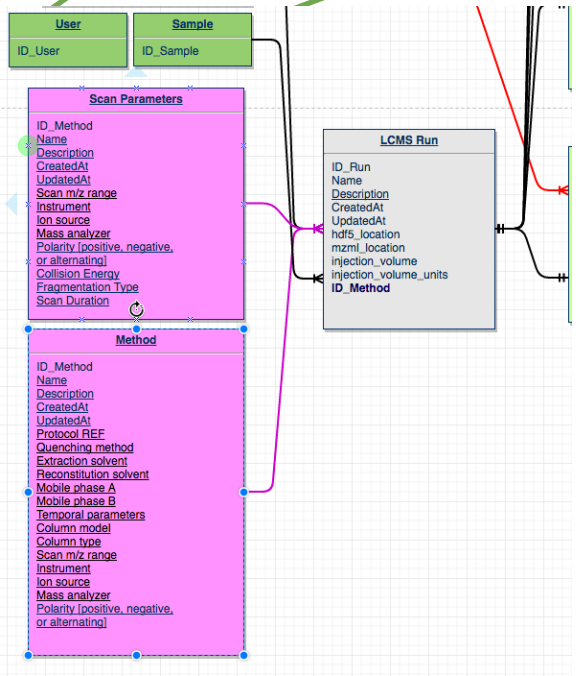
Automated tasks:

- Get method and sample
- Conversion to HDF5
- Existence is registered in database

Metabolite Atlas

How a feature (peak) becomes a compound with an intensity from your sample

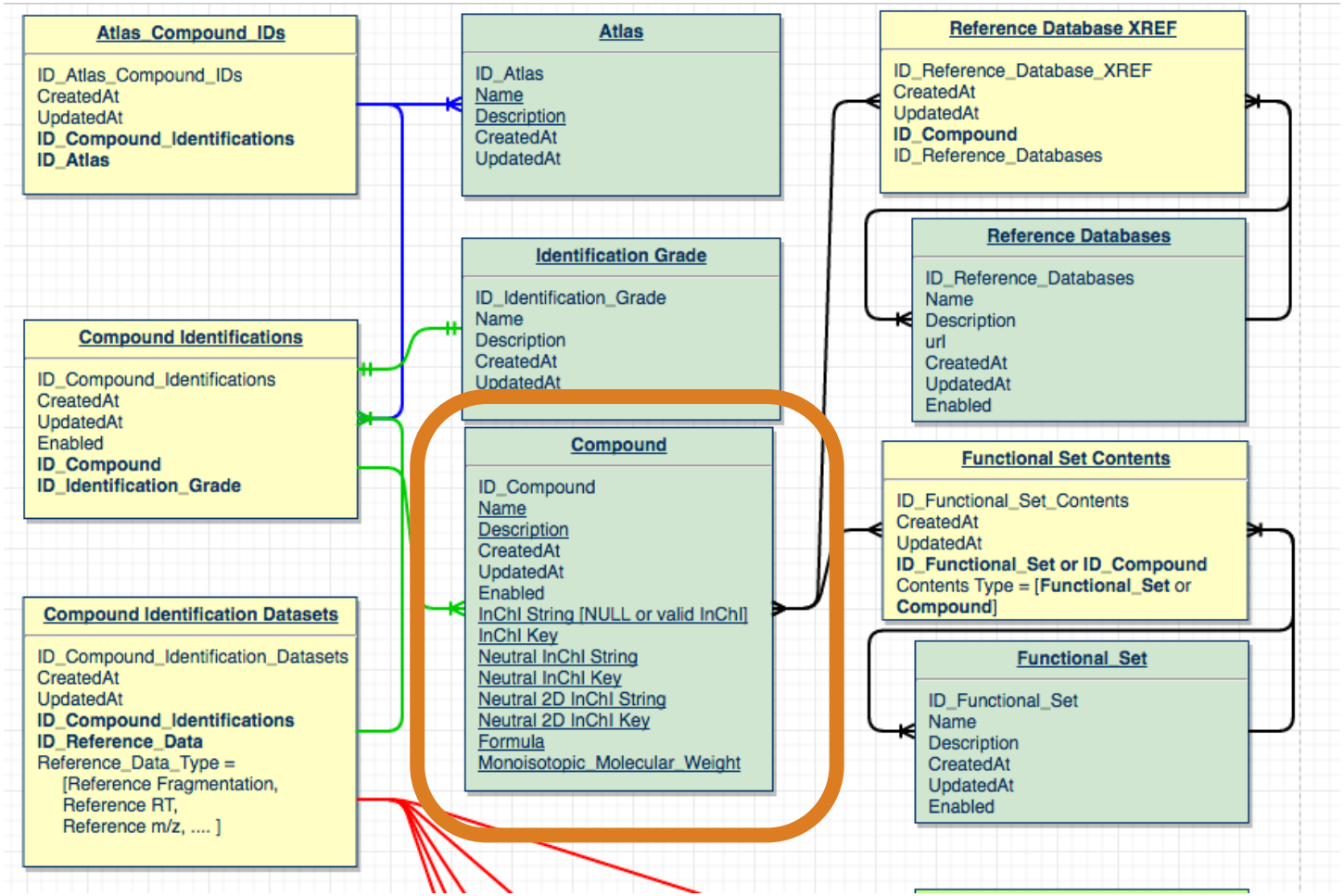




Each sample ID for an LCMS run can map to a genome in IMG

Metabolite Atlas

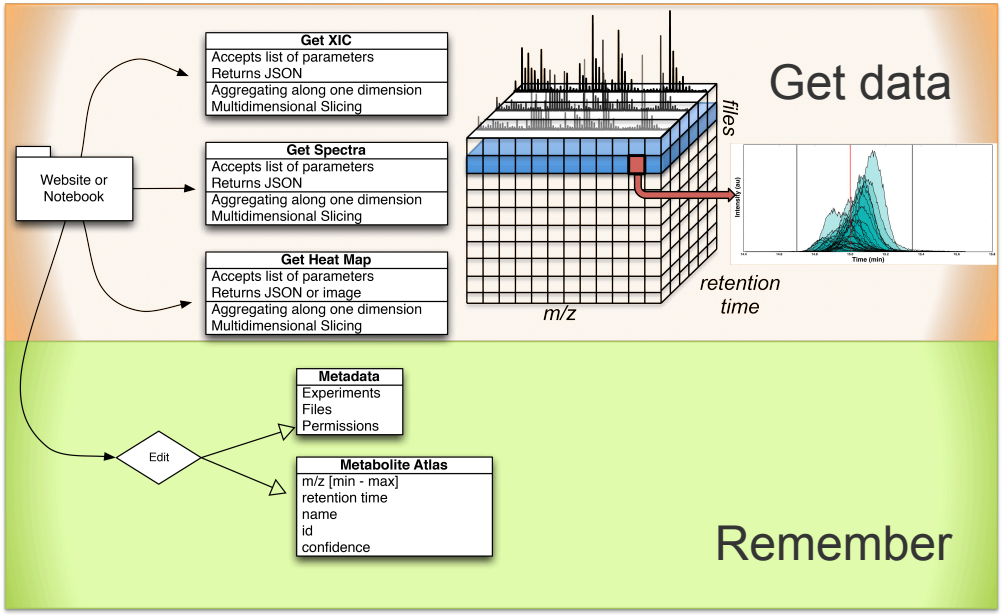
Provide compounds that can link genes and reactions

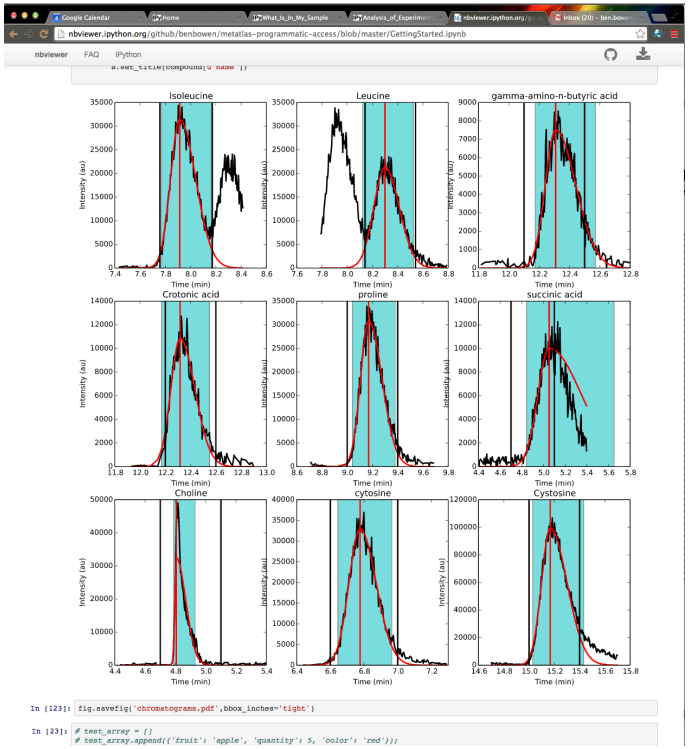


http://tinyurl.com/jgi2016-bowen

Metabolite Atlas

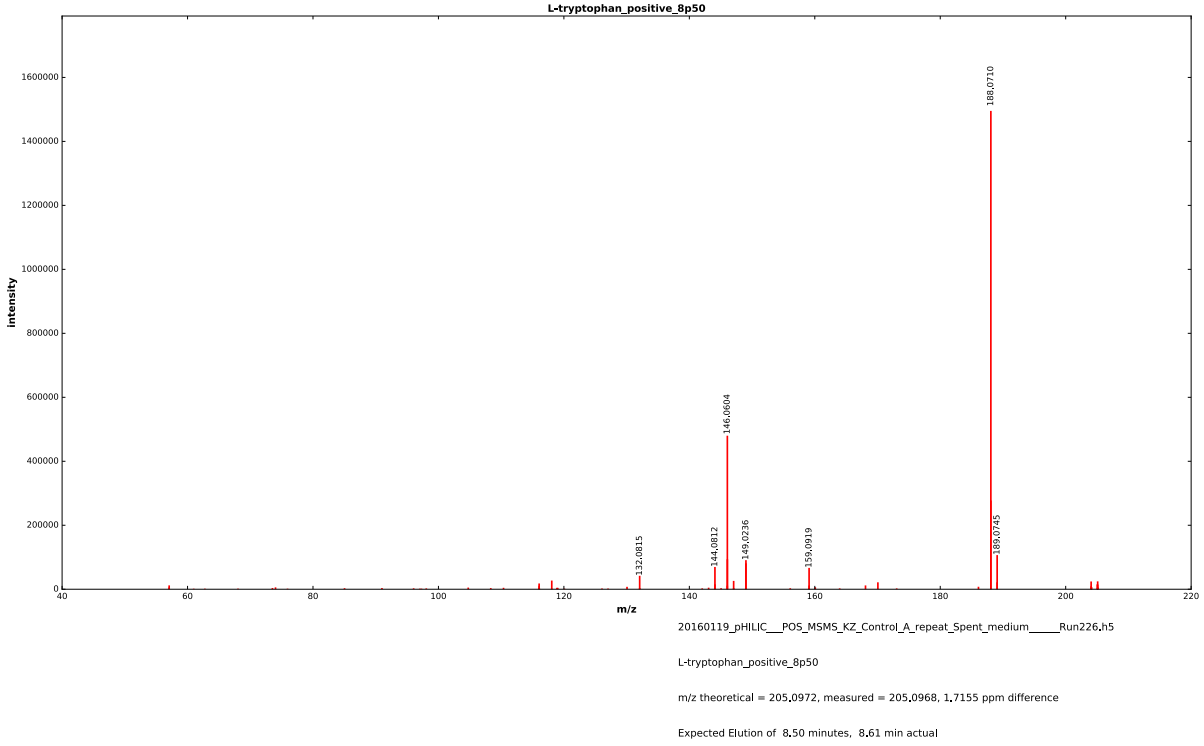
Sample and method specific atlases let us find expected molecules

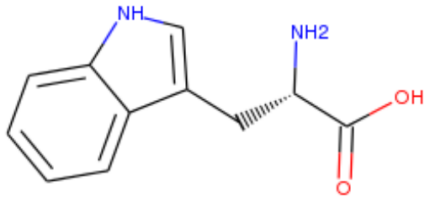


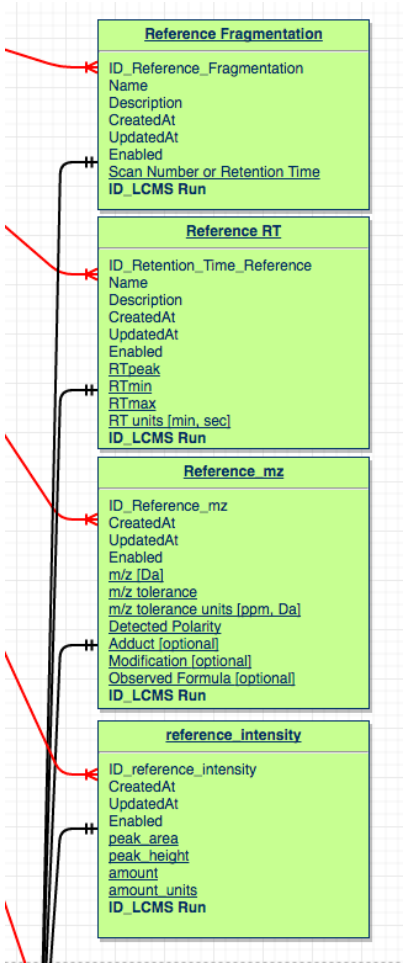


Metabolite Atlas

Layered identification to facilitate untargeted detection of intense ions







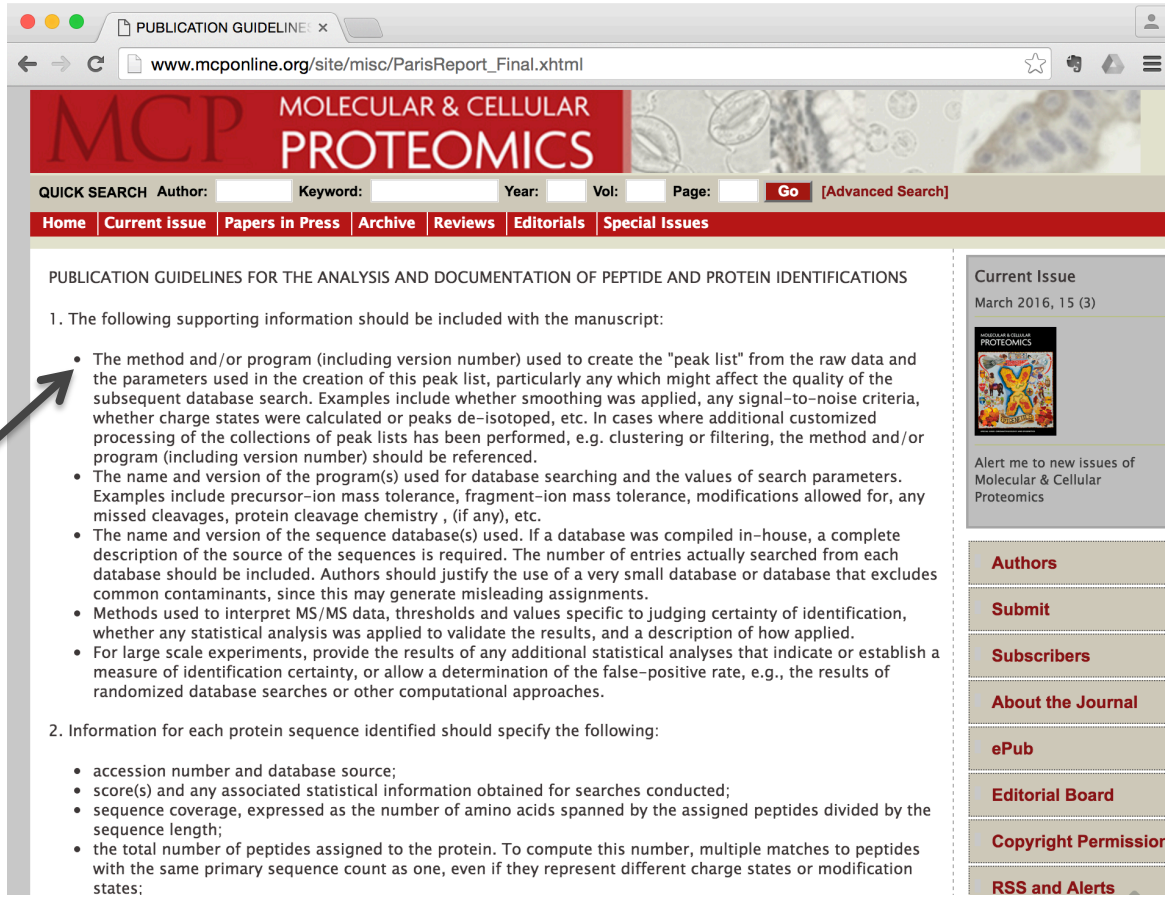
Metabolite Atlas

Provide a useful output that is sufficient for the most rigorous peer review

- MSMS
- Integration bounds
- Expected RT
- Each file meets QC and ISTD
- Compound information

In 2007, Proteomics requirements emerged.

In 2017, you will see more metabolomics requirements.



Each metabolite tells its own story

Book-keeping is essential for scalable identification

nature biotechnology

PERSPECTIVE

A proposed framework for the description of plant metabolomics experiments and their results

Helen Jenkins[1], Nigel Hardy[1], Manfred Beckmann[2], John Draper[2], Aileen R Smith[2], Janet Taylor[1,21], Oliver Fiehn[3], Royston Goodacre[4], Raoul J Bino[5,6], Robert Hall[5], Joachim Kopka[3], Geoffrey A Lane[7], B Markus Lange[8], Jang R Liu[9], Pedro Mendes[10], Basil J Nikolau[11], Stephen G Oliver[12], Norman W Paton[13], Sue Rhee[14], Ute Roessner-Tunali[15], Kazuki Saito[16], Jørn Smedsgaard[17], Lloyd W Sumner[18], Trevor Wang[19], Sean Walsh[19], Eve Syrkin Wurtele[20] & Douglas B Kell[4]

The study of the metabolite complement of biological samples, known as metabolomics, is creating large amounts of data, and support for handling these data sets is required to facilitate meaningful analyses that will answer biological questions. We present a data model for plant metabolomics known as ArMet (architecture for metabolomics). It encompasses the entire experimental time line from experiment definition and description of biological source material, through sample growth and preparation to the results of chemical analysis. Such formal data descriptions, which specify the full design and example implementations are freely available (http://www.armet.org/). We seek to advance discussion and community adoption of a standard for metabolomics, which would promote principled collection, storage and transmission of experiment data.

Functional genomic research is generating large amounts of data. These must be transmitted, stored safely with adequate curation and made available in convenient and supportive ways for statistical analyses and data mining. To do this, well-designed data standards[1] are required

Metabolomics (2007) 3:211–221
DOI 10.1007/s11306-007-0082-2

ORIGINAL ARTICLE

Proposed minimum reporting standards for chemical analysis

Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)

Lloyd W. Sumner · Alexander Amberg · Dave Barrett · Michael H. Beale · Richard Beger · Clare A. Daykin · Teresa W.-M. Fan · Oliver Fiehn · Royston Goodacre · Julian L. Griffin · Thomas Hankemeier · Nigel Hardy · James Harnly · Richard Higashi · Joachim Kopka · Andrew N. Lane · John C. Lindon · Philip Marriott · Andrew W. Nicholls · Michael D. Reily · John J. Thaden · Mark R. Viant

Received: 9 January 2007 / Accepted: 27 July 2007 / Published online: 12 September 2007

Abstract There is a general consensus that supports the need for standardized reporting of metadata or information describing large-scale metabolomics and other functional genomics data sets. Reporting of standard metadata provides a biological and empirical context for the data, facilitates experimental replication, and enables the re-interrogation and comparison of data by others. Accordingly, the Metabolomics Standards Initiative is building a general consensus concerning the minimum reporting standards for metabolomics experiments of which the Chemical Analysis Working Group (CAWG) is a member of this community effort. This article proposes the minimum reporting standards related to the chemical analysis aspects of metabolomics experiments including: sample preparation, experimental analysis, quality control, metabolite identification, and data pre-processing. These minimum standards currently focus mostly upon mass spectrometry and nuclear magnetic resonance spectroscopy due to the popularity of these techniques in metabolomics. However, additional input concerning other techniques is welcomed and can be provided via the CAWG on-line discussion forum at http://msi-workgroups.sourceforge.net/

The Metabolomics Standards Initiative

To the editor:

The standards papers that *Nature Biotechnology* hosted online as part of a community consultation (http://www.nature.com/nbt/consult/index.html), in particular those by the Human Proteome Organization Proteomics Standardization Initiative (HUPO-PSI)[1,2] and the Functional Genomics Experiment (FuGE)[3] working groups, represent an important first step toward permitting the sharing of high-quality, structured data. We particularly applaud the open consultation solicited by *Nature Biotechnology* and advocate the early-community-involvement approach taken by HUPO-PSI, FuGE and the other working groups in the development of such guidelines and standards. These are the most effective ways to ensure that the output generated is pragmatic and the standards are both useful and widely accepted by the community.

As representatives of the nascent Metabolomics Standards Initiative (MSI)[4], we are following closely the work of the FuGE and the PSI working groups, leveraging on their work where commonality exists, such as the mass

http://tinyurl.com/jgi2016-bowen

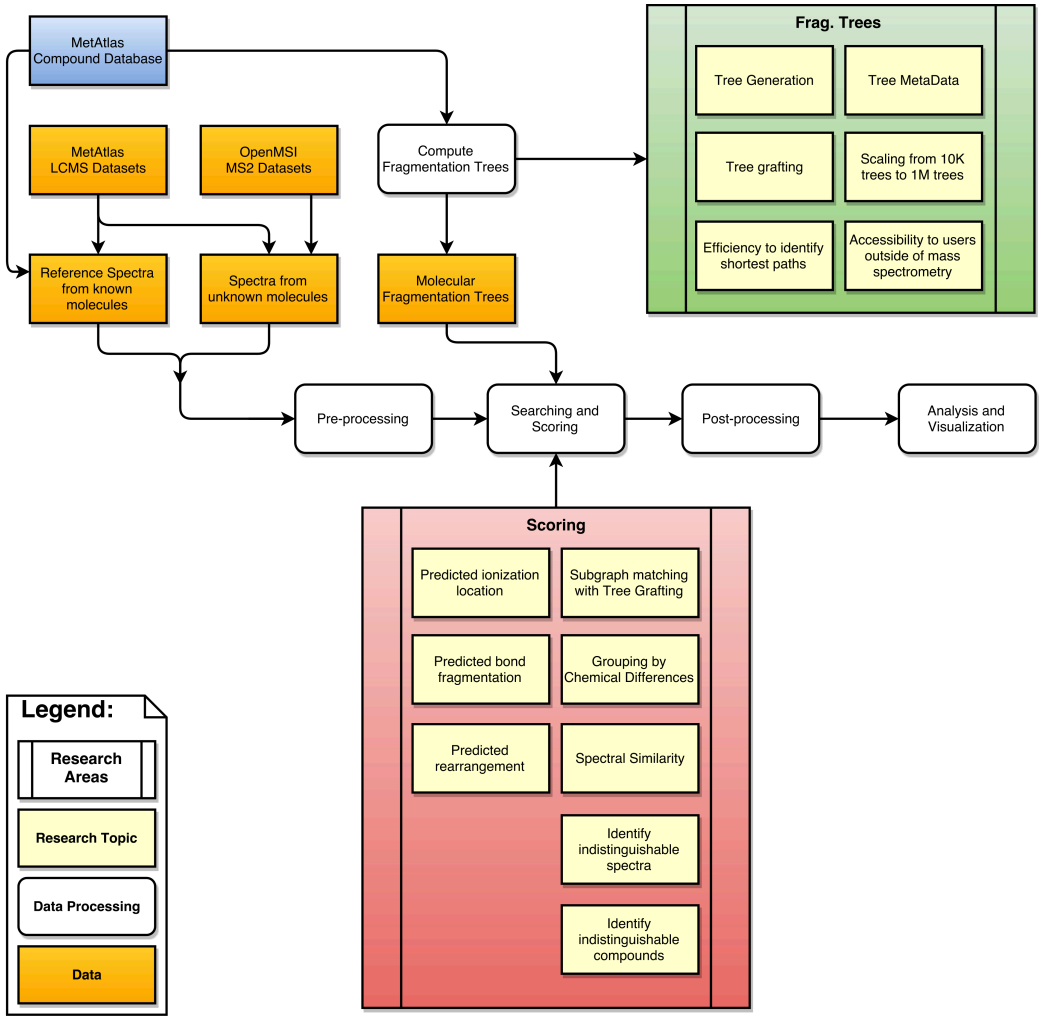
Dealing with the unknown

Compounds you observe: authentic standard is not available

1. First choice is always third party spectral libraries:
   - Not downloadable
   - Low quality
   - Compound not available

2. First principles calculation
   - Working for EI
   - ESI still has a way to go

3. Hybrid methods
   - Dependent on large graphs
   - Optimization of scoring algorithms needed



http://tinyurl.com/jgi2016-bowen

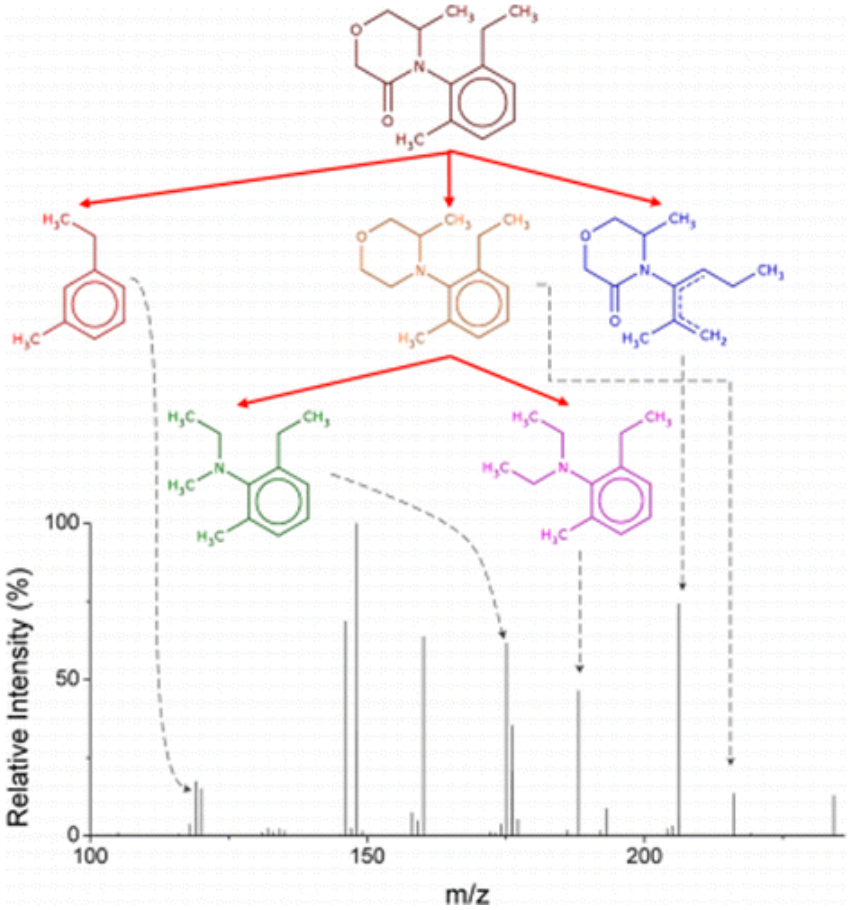
Dealing with the unknown

Compounds you observe: authentic standard is not available

- **Complete enumeration**
- **Scoring is higher when branch of tree is observed consistently**

Much higher FDR than spectral libraries

Enables exploration of hypothetical molecular structures which are not in databases

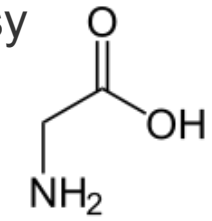


http://tinyurl.com/jgi2016-bowen

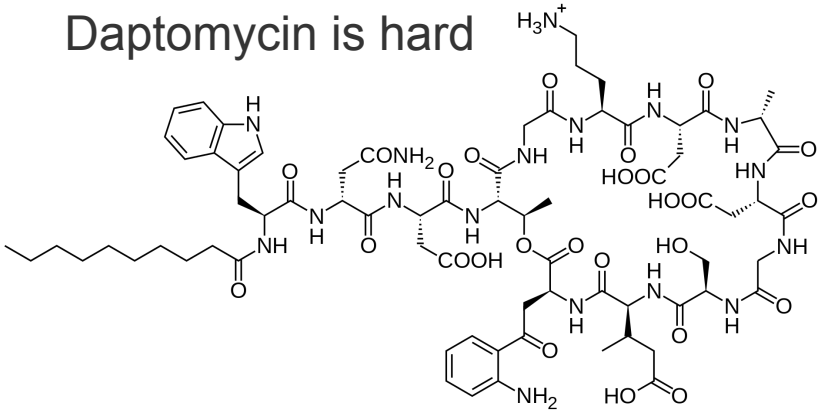
Dealing with the unknown

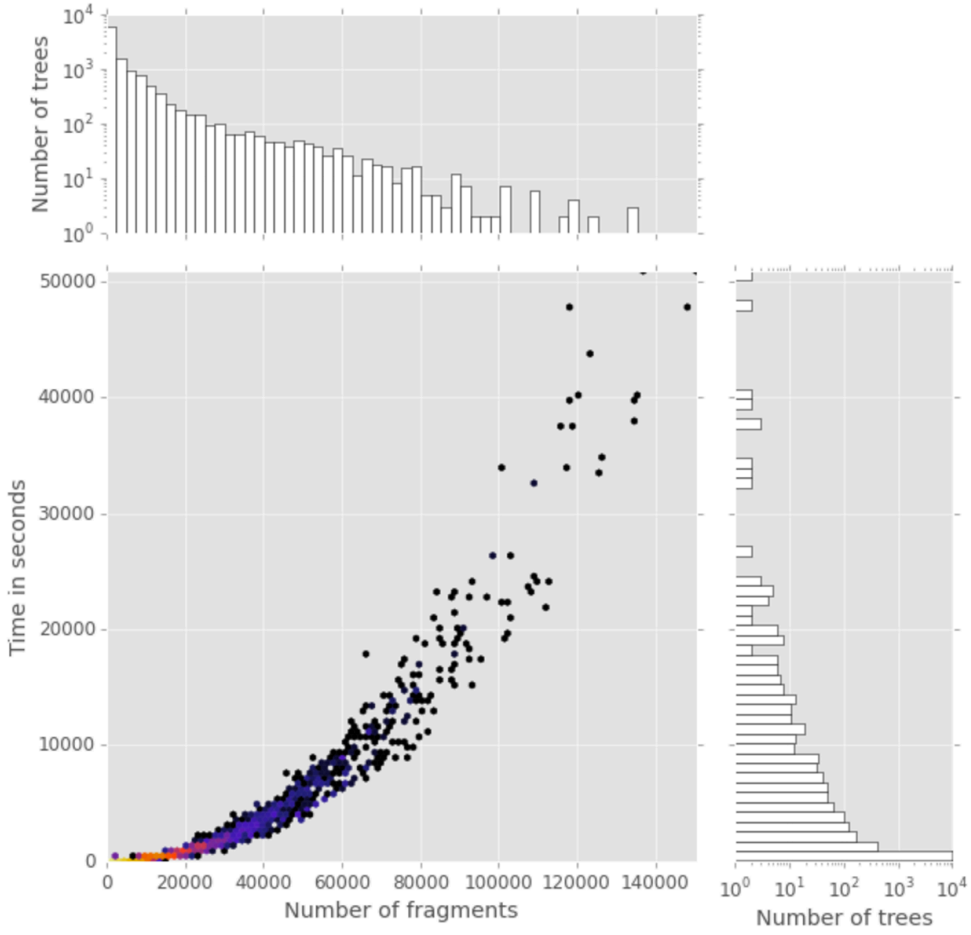
Without supercomputing this is not possible

Glycine is easy



Daptomycin is hard





http://tinyurl.com/jgi2016-bowen

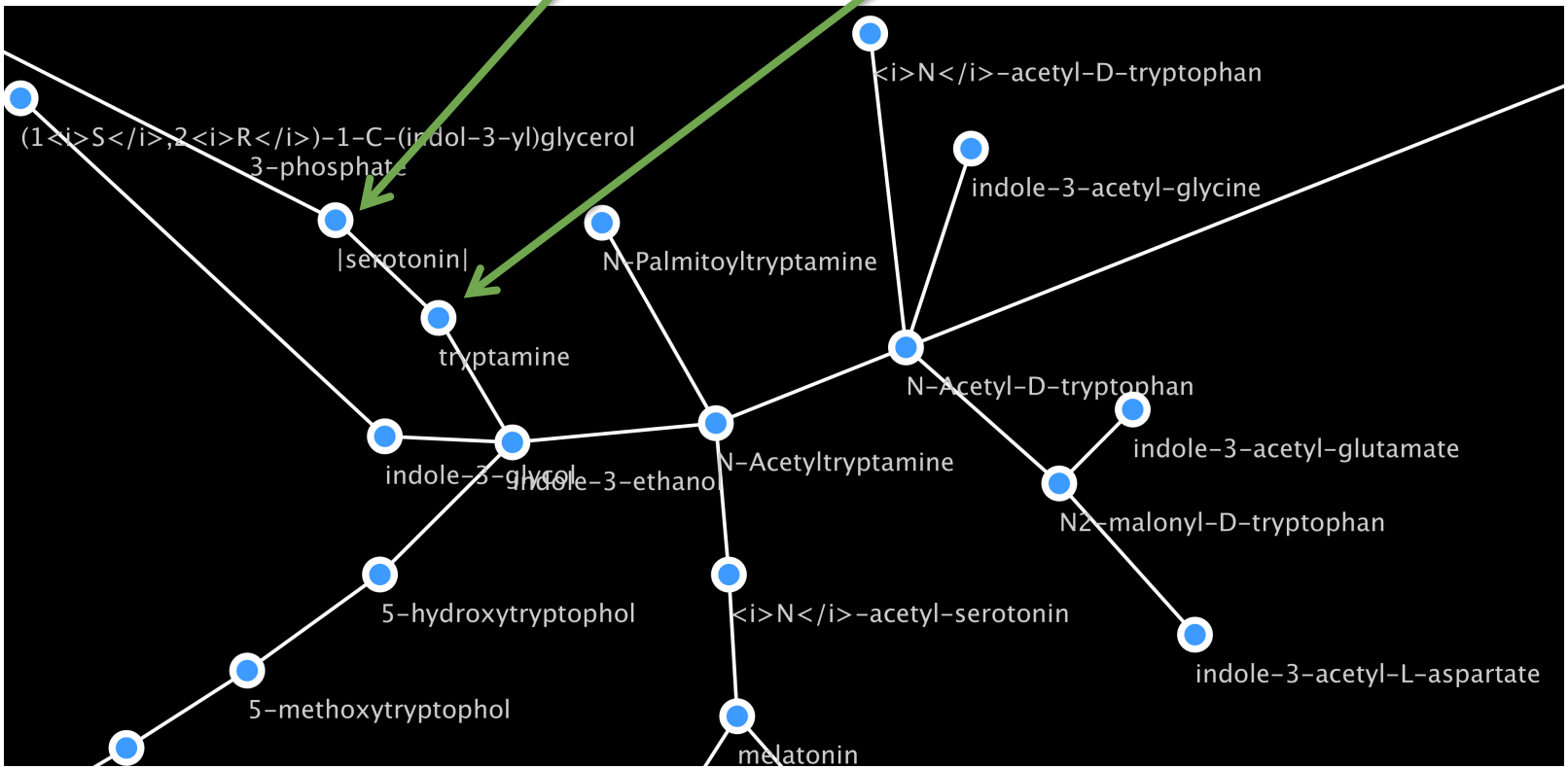
Data integration

Exometabolomics of primary metabolite utilization

- **Consider a study where**
  - Consumption of medium #1 induces sleep
  - Consumption of control media has no effect
- **Identify the patterns of metabolite utilization**
  - Find co-varying transcripts
  - Find genes responsible

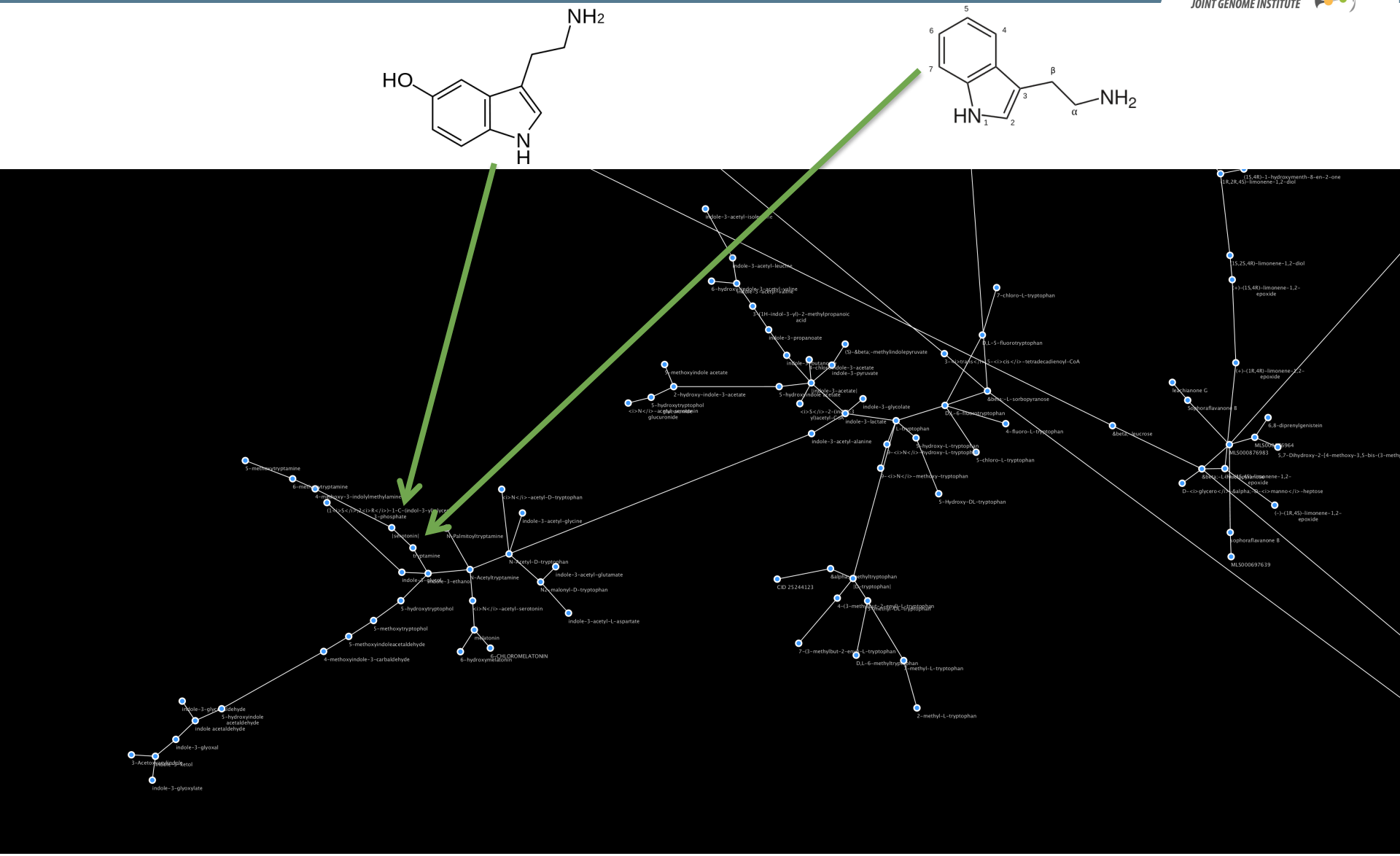
Chemical networks

- **Relationships of new molecules**
- **Data integration**
  - Transcriptomics
  - Gene predictions



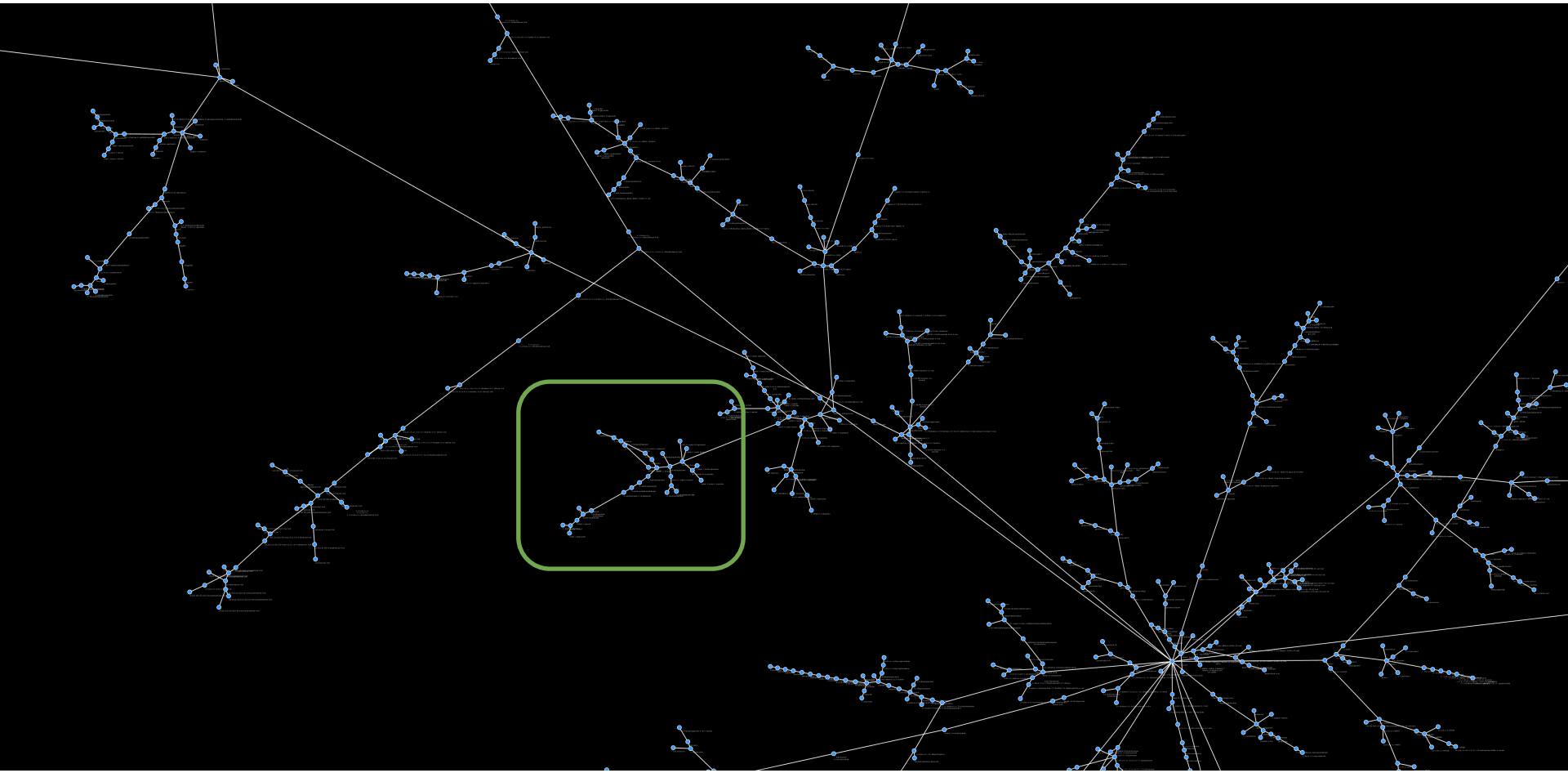
http://tinyurl.com/jgi2016-bowen

Chemical networks



http://tinyurl.com/jgi2016-bowen

Chemical networks



http://tinyurl.com/jgi2016-bowen

Chemical networks