Biochem 3BP3

Beyond the Gene – Networks, Ontologies

Week of Oct 4, 2021

This week...

WEEK 5 (OCTOBER 4 and 6) - BEYOND THE GENE

LIVE lecture in class Wednesday 12:30pm,

Recorded content

- Overview of Laboratory #4 - Ontology and Antimicrobial Resistance, https://web.microsoftstream.com/video/fa0c2a0b-f976-4c11-b3c6-f8b9adf5e4a9

Tutorial

- **LIVE** session with Teaching Assistants and Flash Updates
  - Monday,
  - Wednesday,
- Tutorial content can be found at GitHub, answers due on A2L

Flash Updates

- **Gene Ontology**. Introduce the Gene Ontology. See Nucleic Acids Res. 2019 Jan 8;47(D1):D330-D338 [PMID 30395331].
- **KEGG**. Introduce the Kyoto Encyclopedia of Genes and Genomes (KEGG). See Nucleic Acids Res. 2016 Jan 4;44(D1):D457-62 [PMID 26476454] and Nucleic Acids Res. 2019 Jan 8;47(D1):D590-D595 [PMID 30321428].
- **CARD**. Introduce the Comprehensive Antibiotic Resistance Database. See Nucleic Acids Res. 2017 45(Database issue):D566-D573 [PMID 27789705] and Nucleic Acids Res. 2020 48(Database issue):D517-D525 [PMID 31665441].

Lab 4 problems due Oct 22nd

Need to submit:

- answers to the problems
- FASTA file
- PHYLIP alignment
- RaxML file

Coming up ...

WEEK 6 (OCTOBER 11 and 13) - MID-TERM RECESS

WEEK 7 (OCTOBER 18 and 20) - LINUX & SEQUENCING INFORMATICS

No problems due

Coming up ...

Critical Review – Due October 27, 2021

This is a critical review exercise, worth 25% of the total course grade. Please follow the guidelines provided in the grading rubric and use the template WORD file provided.

Excluding references, the Critical Review cannot exceed 2 pages in length.

You are being asked to review a pre-publication manuscript submitted to www.biorxiv.org, an open access preprint repository for the biological sciences. Papers in the bioRxiv have generally not undergone peer review and thus are not considered formal publications. A pre-print at bioRxiv may latter appear as a publication in a scientific journal after peer-review.

You will be acting as a manuscript reviewer. You will be assigned a bioRxiv pre-print that includes genomics and/or bioinformatics as one of its primary methodologies. The overall topic of the pre-print will be in any aspect of the biological sciences, not just those covered by Biochem 3BP3. Critically read the pre-print and write a 2 page critique of the work, identifying any areas for improvement for the research described but also evaluating why you think the pre-print is valuable contribution to the field or not.

Notes:

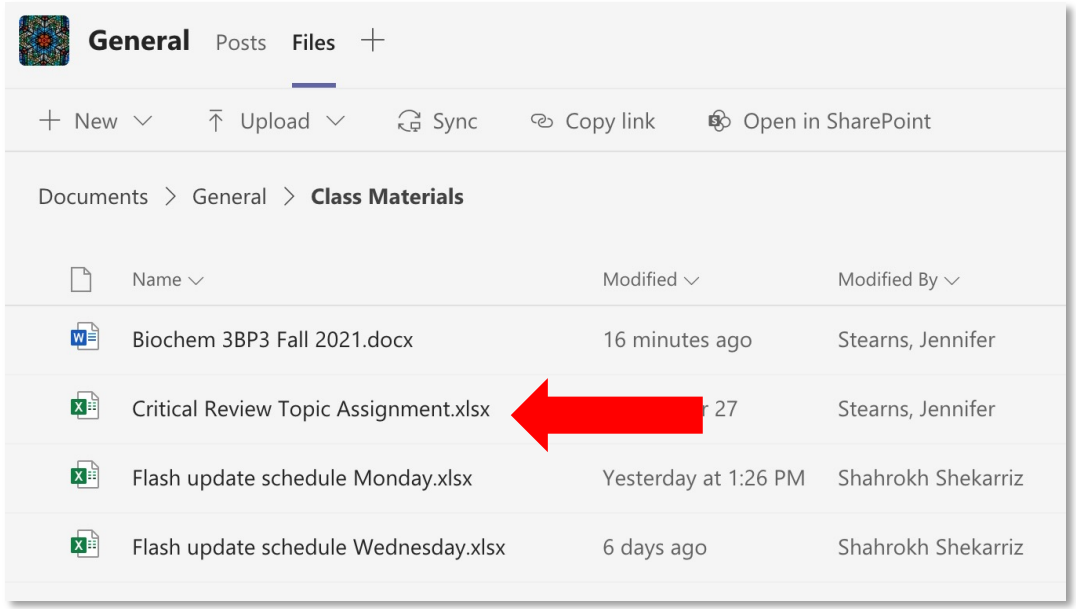
- Many pre-prints include online comments. You can use these to help guide your review, but TAs will be checking for plagerism.
- The assignments are posted in the course Teams under files

Coming up ...

Topic papers

1. **Metagenomic identification of viral sequences in laboratory reagents** https://www.biorxiv.org/content/10.1101/2021.09.10.459871v1
2. **Improving long-read consensus sequencing accuracy with deep learning** https://www.biorxiv.org/content/10.1101/2021.06.28.450238v3
3. **Strong experimental support for the hologenome hypothesis revealed from *Drosophila melanogaster* selection lines** https://www.biorxiv.org/content/10.1101/2021.09.09.459587v1
4. **Active growth signalling promotes cancer cell sensitivity to the CDK7 inhibitor ICEC0942** https://www.biorxiv.org/content/10.1101/2021.09.10.459733v1
5. **Computational drug repurposing against SARS-CoV-2 reveals plasma membrane cholesterol depletion as key factor of antiviral drug activity** https://www.biorxiv.org/content/10.1101/2021.09.10.459786v1
6. **Trypanosomal variant surface glycoprotein expression in human African trypanosomiasis patients** https://www.biorxiv.org/content/10.1101/2021.09.09.459620v2
7. **In silico design of bioactive chimeric peptide from archaeal antimicrobial peptides** https://www.biorxiv.org/content/10.1101/2021.08.14.456327v2
8. **Using BERT to identify drug-target interactions from whole PubMed** https://www.biorxiv.org/content/10.1101/2021.09.10.459845v1
9. **The population frequency of human mitochondrial DNA variants is highly dependent upon mutational bias** https://www.biorxiv.org/content/10.1101/2021.05.12.443844v3
10. **Implications of taxonomic and numerical resolution on DNA metabarcoding-based inference of benthic macroinvertebrate responses to river restoration** https://www.biorxiv.org/content/10.1101/2021.09.11.459893v1

Coming up ...



No extensions will be granted

Please get in touch if you don't think you'll be able to catch up on all labs by Oct 22nd

BioCuration & Databases

- One of the most important tools in bioinformatics today is the database
  - Primary – raw data – DNA sequence, protein structure, chemical structure, gene expression measurements, etc
  - Secondary – integration of many sources of primary data, often organism-specific (e.g. *Drosophila* DB) or area specific (e.g. drug-drug interaction DB)
  - Literature curation (human or automated) and higher-level conceptual organization of knowledge (e.g. ontologies) increasingly important

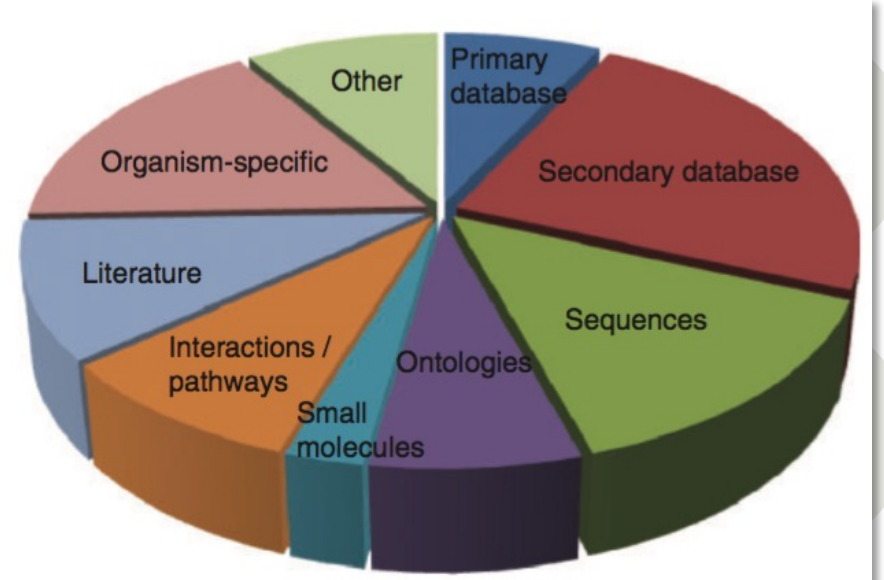


Figure 2. The types of data annotated by biocurators.

Burge et al. 2012. Biocurators and biocuration: surveying the 21st century challenges. Database Mar 20:bar059

Integrating Knowledge

- We have sequenced a large number of diverse genomes, each with thousands of predicted genes
- We have performed genome-wide gene expression or gene knock-down experiments to understand the role of these genes
- High-throughput experimental methods are producing increasingly complex data sets about the interactions between genes
- We have a century of knowledge about molecular & cell biology, physiology, biochemistry, etc. Yet publications rates are increasing!
- How do we integrate these data to obtain a deeper understanding of biology?
- How do we interpret experiments that generate thousands or millions of data points?

Integrating Knowledge

"Literature is informative but is not information"
– Dr. Suzanna Lewis, Lawrence Berkeley National Laboratory

- The scientific literature and the peer-reviewed publication system that drives it is at the heart of scientific advancement
- Transparency, repeatability, and readability are important aspects of scientific publishing
- A good scientific publication is highly informative – to a human reader – but very hard for a computer to work with
- To a computer scientist, information is "bits" organized via strict rules – good digital representations of knowledge

Integrating Knowledge

- Knowledge integration in the biological sciences has two “consumers”
  - Humans – databases and data sets that provide a ‘one stop shop’ for integrated knowledge since no single person can keep up with the literature
  - Computers – databases and data sets that provide integration of knowledge in a form that computer algorithms can use to make sense of new data
- Key to knowledge integration is the task of ‘Biocuration’ – translation of scientific knowledge into a digital format
- Two key tools of Biocuration are **Ontologies** and **Networks**
- Raw uncurated data equally valuable in the form of **Interactomes**

Ontologies

- Also known as “Controlled Vocabularies” – standardized names and relationships to be used in the description of biological concepts
- Ontologies do not contain data but instead are conceptual and map the relationships between concepts
- Data can be tagged with ontological terms, thus placing the data within a larger conceptual framework plus allowing comparison among data sets
- Computer algorithms see ontologies as graphs they can traverse to answer questions about data
- The most heavily adopted ontology is the Gene Ontology
  - represents gene and gene product information among all species
  - focused upon Molecular Function, Cellular Component, and Biological Process

CYP1A1 Annotated by the Gene Ontology

Gene Product Information

Symbol	CYP1A1
Name(s)	Cytochrome P450
Type	protein
Taxon	Homo sapiens
Synonyms	Q5J9B1_HUMAN
Database	UniProtKB, Q5J9B1
Related	Link to all direct and indirect **annotations** to CYP1A1.
	Link to all direct and indirect **annotations download** (limited to first 10,000) for CYP1A1.
Feedback	Contact the GO Helpdesk if you find mistakes or have concerns about the data you find here.

Gene/product	Gene/product name	Qualifier	Direct annotation	Annotation extension	Assigned by	Taxon	Evidence	Evidence with
CYP1A1	Cytochrome P450		iron ion binding		InterPro	Homo sapiens	IEA	InterPro:IPR001128 InterPro:IPR002401 InterPro:IPR008066
CYP1A1	Cytochrome P450		endoplasmic reticulum membrane		UniProt	Homo sapiens	IEA	UniProtKB-SubCell:SL-0097
CYP1A1	Cytochrome P450		heme binding		InterPro	Homo sapiens	IEA	InterPro:IPR001128 InterPro:IPR002401 InterPro:IPR008066
CYP1A1	Cytochrome P450		oxidation-reduction process		UniProt	Homo sapiens	IEA	UniProtKB-KW:KW-0560
CYP1A1	Cytochrome P450		aromatase activity		UniProt	Homo sapiens	IEA	EC:1.14.14.1

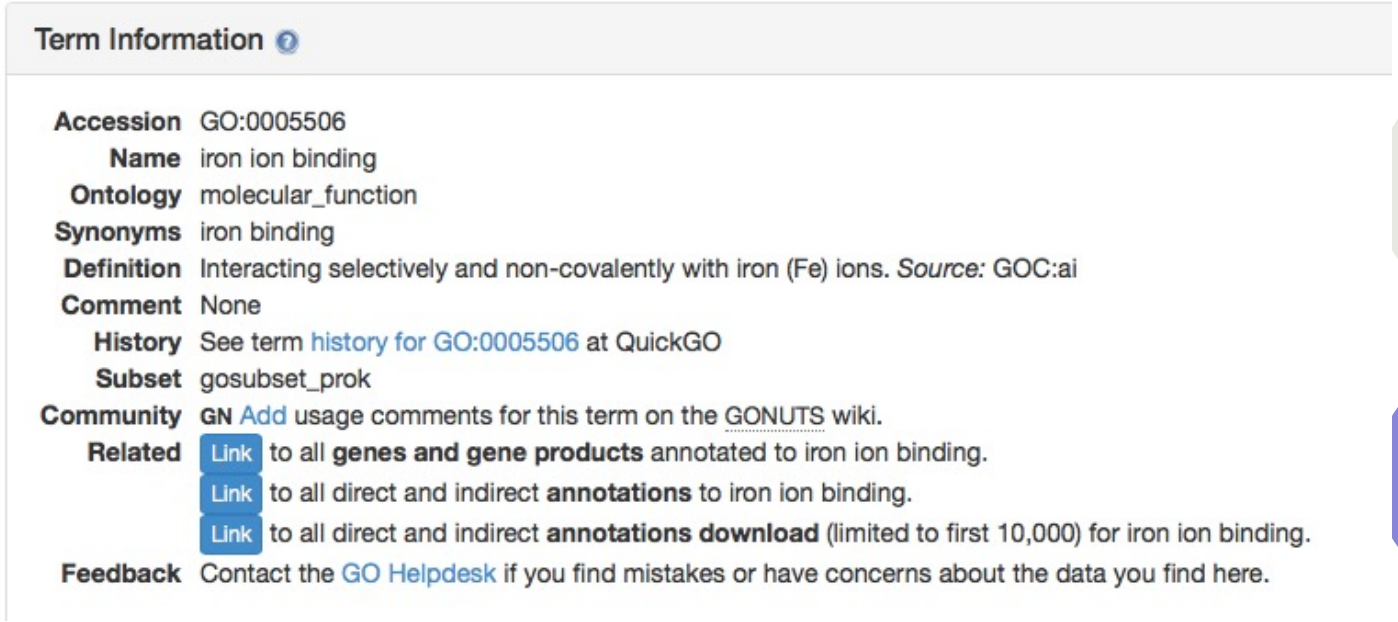
CYP1A1 Annotated by the Gene Ontology

Gene Product Information

Symbol	CYP1A1
Name(s)	Cytochrome P450
Type	protein
Taxon	Homo sapiens
Synonyms	Q5J9B1_HUMAN
Database	UniProtKB, Q5J9B1
Related	Link to all direct and indirect **annotations** to CYP1A1.
	Link to all direct and indirect **annotations download** (limited to first 10,000) for CYP1A1.
Feedback	Contact the GO Helpdesk if you find mistakes or have concerns about the data you find here.

Gene/product	Gene/product name	Qualifier	Direct annotation	Annotation extension	Assigned by	Taxon	Evidence	Evidence with
CYP1A1	Cytochrome P450		iron ion binding		InterPro	Homo sapiens	IEA	InterPro:IPR001128 InterPro:IPR002401 InterPro:IPR008066
CYP1A1	Cytochrome P450		endoplasmic reticulum membrane		UniProt	Homo sapiens	IEA	UniProtKB-SubCell:SL-0097
CYP1A1	Cytochrome P450		heme binding		InterPro	Homo sapiens	IEA	InterPro:IPR001128 InterPro:IPR002401 InterPro:IPR008066
CYP1A1	Cytochrome P450		oxidation-reduction process		UniProt	Homo sapiens	IEA	UniProtKB-KW:KW-0560
CYP1A1	Cytochrome P450		aromatase activity		UniProt	Homo sapiens	IEA	EC:1.14.14.1

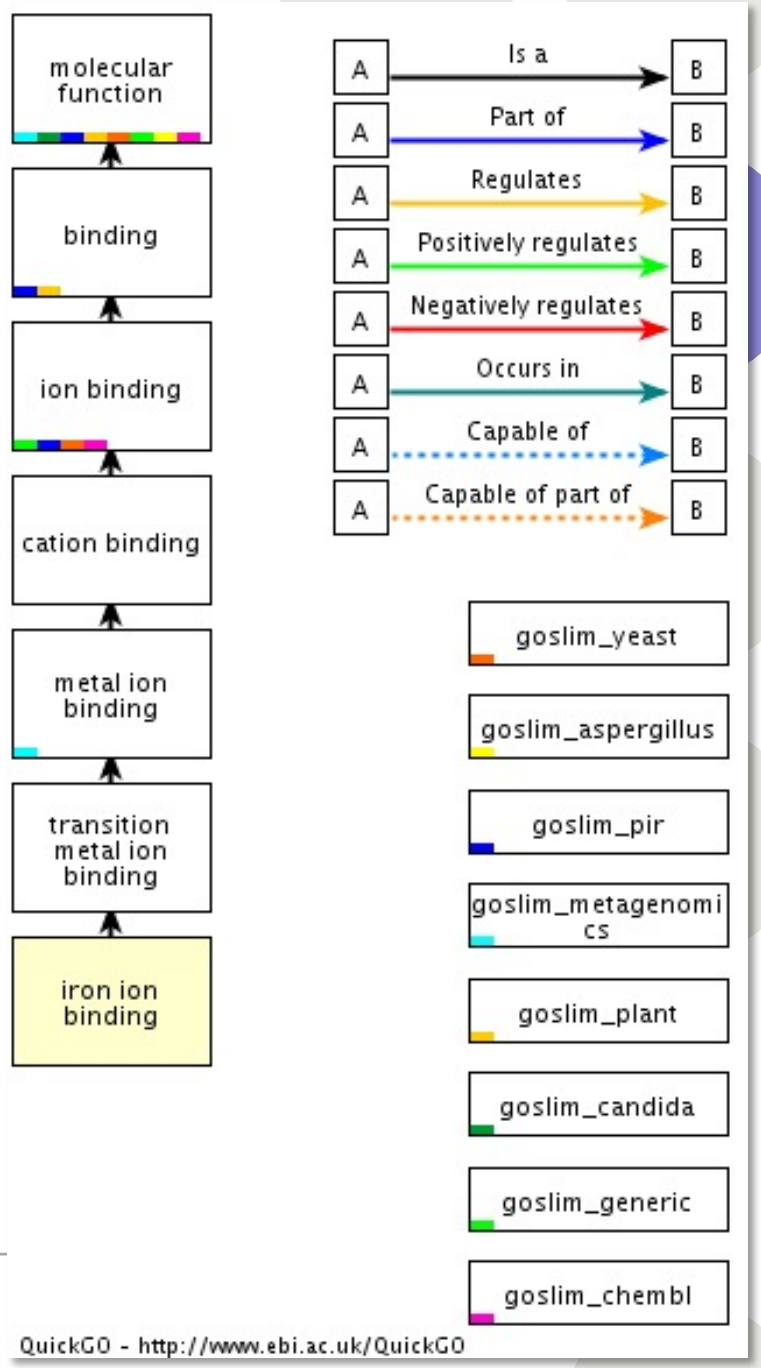
“Iron Ion Binding” Gene Ontology term



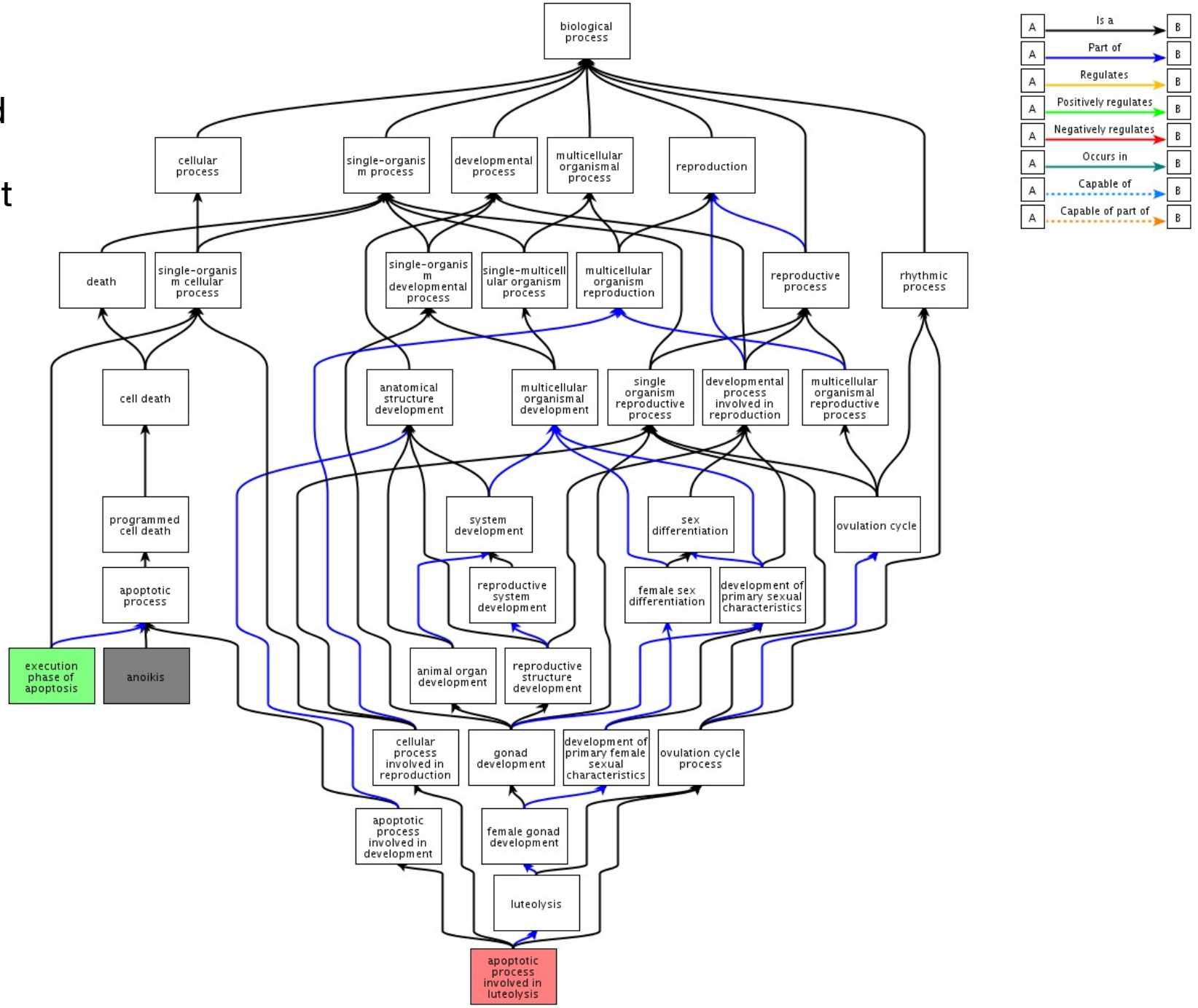
- An ontology term has an Accession, Name, and Definition as it’s minimum information
- Terms often include references to external information, such as a citation from the scientific literature
- External data can be linked to the term by analysis (e.g. gene annotation algorithms, Pfam analysis, etc) <- not part of the ontology itself

“Iron Ion Binding” Gene Ontology term

- Each ontology term is connected to other ontology terms via carefully defined relationship types
- “is_a” and “part_of” are the most common relationship types but each ontology may have its own specific types
- Tagging data with ontology terms allows researchers to search or classify their data using specific or general terms
- Ontologies can help researchers discover unrealized relationships among results



Example – Gene annotation highlighted three GO Terms in a data set (in colour) but the Gene Ontology shows how they are linked by common biological processes



OBO & OWL formats

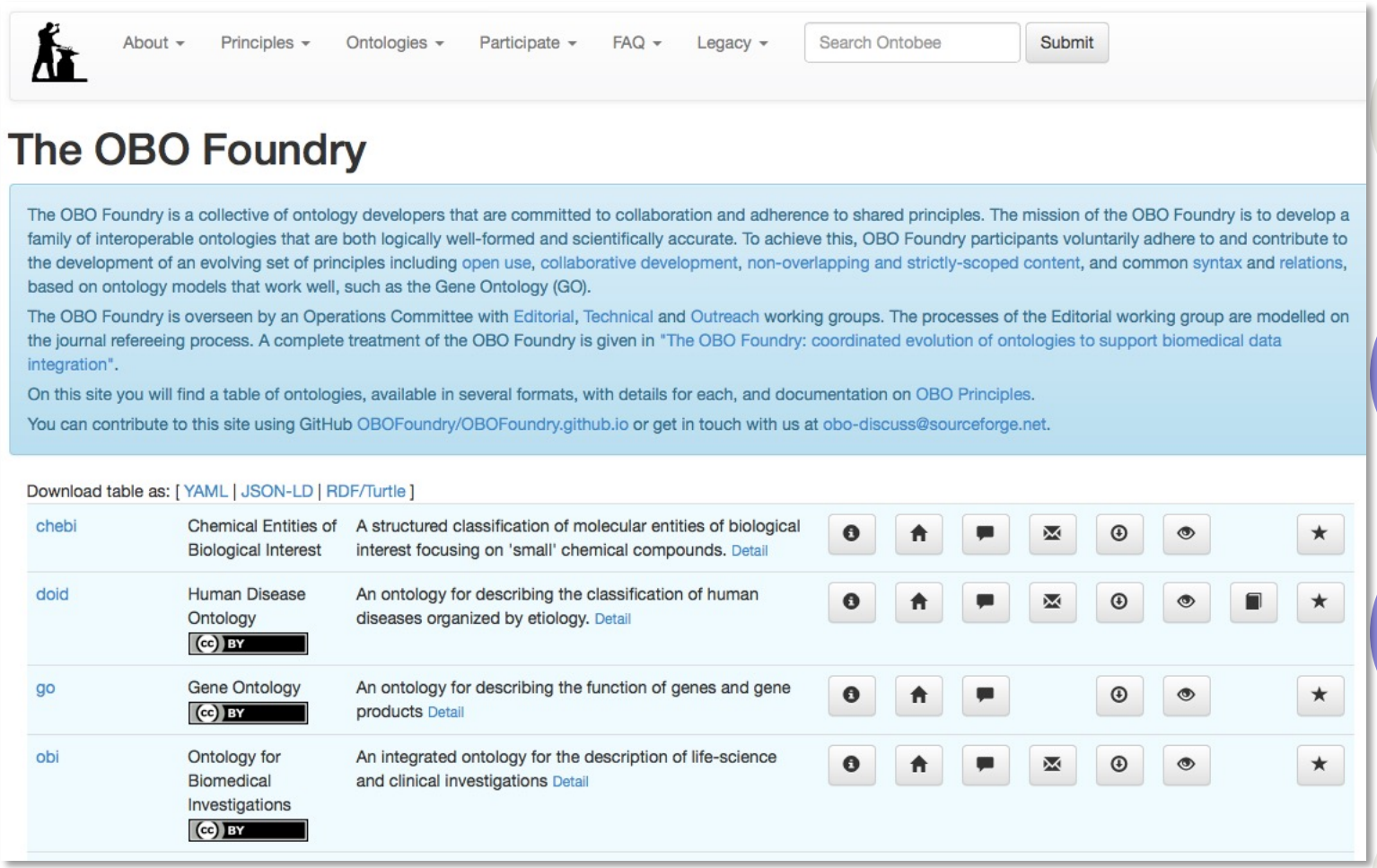
Computer readable descriptions of an ontology allow development of algorithms to perform analysis in an ontological context

```
[Term]
id: GO:0006248
name: CMP catabolic process
namespace: biological_process
def: "The chemical reactions and pathways resulting in the breakdown of CMP, cytidine monophosphate." [ISBN:0198506732]
subset: gosubset_prok
synonym: "CMP breakdown" EXACT []
synonym: "CMP catabolism" EXACT []
synonym: "CMP degradation" EXACT []
is_a: GO:0009175 ! pyrimidine ribonucleoside monophosphate catabolic process
is_a: GO:0009222 ! pyrimidine ribonucleotide catabolic process
is_a: GO:0046035 ! CMP metabolic process
is_a: GO:0046133 ! pyrimidine ribonucleoside catabolic process

[Term]
id: GO:0006249
name: dCMP catabolic process
namespace: biological_process
def: "The chemical reactions and pathways resulting in the breakdown of dCMP, deoxycytidine monophosphate." [ISBN:0198506732]
subset: gosubset_prok
synonym: "dCMP breakdown" EXACT []
synonym: "dCMP catabolism" EXACT []
synonym: "dCMP degradation" EXACT []
is_a: GO:0009178 ! pyrimidine deoxyribonucleoside monophosphate catabolic process
is_a: GO:0009223 ! pyrimidine deoxyribonucleotide catabolic process
is_a: GO:0046063 ! dCMP metabolic process

[Term]
id: GO:0006250
name: obsolete CDP reduction
namespace: biological_process
def: "OBSOLETE (was not defined before being made obsolete)." [GOC:ai]
comment: This term was made obsolete because it represents a molecular function rather than a biological process.
synonym: "CDP reduction" EXACT []
is_obsolete: true
consider: GO:0051063

[Term]
id: GO:0006251
name: dCDP catabolic process
namespace: biological_process
def: "The chemical reactions and pathways resulting in the breakdown of dCDP, deoxycytidine 5'-diphosphate." [ISBN:0198506732]
subset: gosubset_prok
synonym: "dCDP breakdown" EXACT []
synonym: "dCDP catabolism" EXACT []
synonym: "dCDP degradation" EXACT []
is_a: GO:0009198 ! pyrimidine deoxyribonucleoside diphosphate catabolic process
is_a: GO:0009223 ! pyrimidine deoxyribonucleotide catabolic process
is_a: GO:0046062 ! dCDP metabolic process
```

OBO Foundry

- www.obofoundry.org - a repository of thousands of ontologies, each developed by ‘domain experts’, i.e. Biocurators with expert knowledge

- Dr. McArthur’s lab are ‘domain experts’ in Antibiotic Resistance and have developed the Antibiotic Resistance Ontology



OBO Foundry

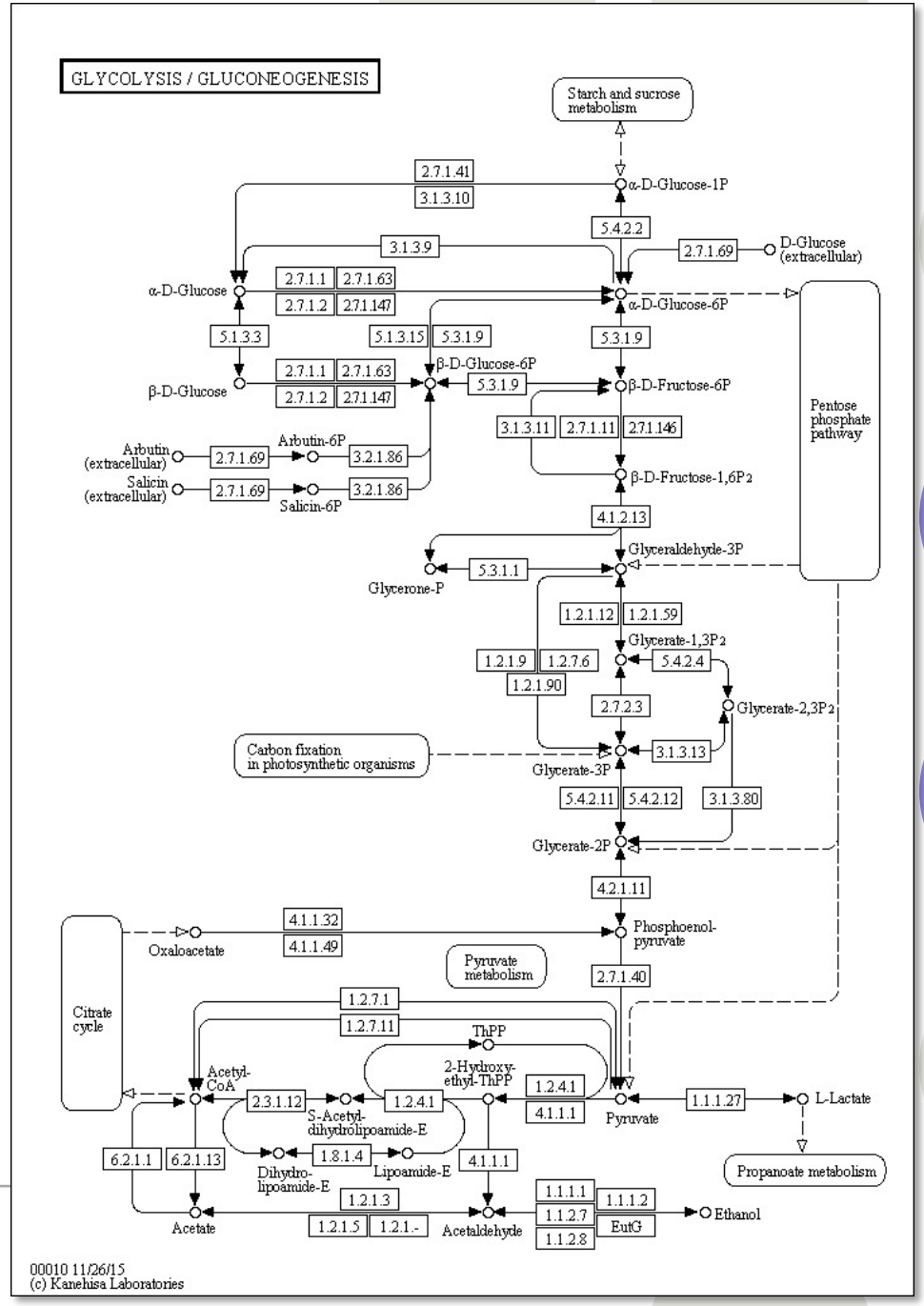
- Enforces standards – particularly for syntax and relationship types – defined by Principles (three shown below)
- Principle #1 – OPEN - Available to all without constraint, preferably using a Creative Commons license.
- Principle #2 – FORMAT – Ontologies must be available in a common format language using standard syntax – OWL or OBO.
- Principle #3 – ORTHOGONAL – terms must be unique – no duplicate terms among ontologies (i.e. a beta-lactamase GO term and a beta-lactamase ARO term). Where ontologies overlap, they use the exact same term (i.e. the ARO and GO both cite the GO term for beta-lactamase). This creates interconnectedness among ontologies!
- There is often conflict in ontology development between ‘correctness’ and ‘usefulness’ - see Goble & Wroe. 2004. The Montagues and the Capulets. *Comp Funct Genomics* 5: 623-32.

Networks

- Whereas ontologies map relationships between concepts, networks generally map relationships between real entities that can be measured
- Unlike ontologies, networks are generally not hierarchical in nature
- The best examples of a familiar network are a biochemical reaction or a regulatory pathway
- The most commonly used network resource is the KEGG: Kyoto Encyclopedia of Genes and Genomes
- Networks reflect a synthesis of knowledge, particularly around gene regulation and biochemical reactions

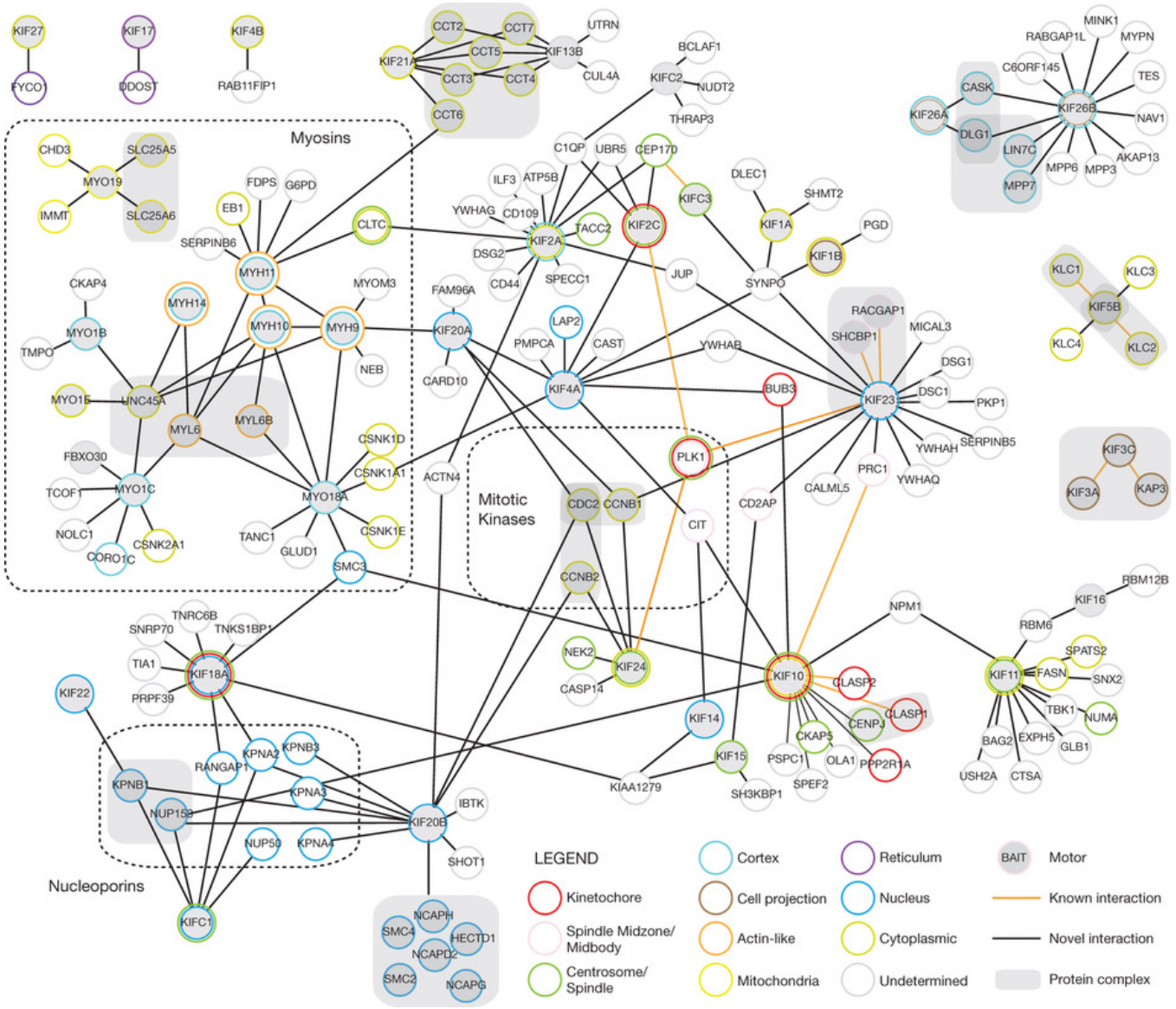
Glycolysis in KEGG

- The glycolysis network in KEGG represents understood relationships among enzymes and compounds
- Each enzyme or compound “term” is associated with a large volume of curated information on its classification, association with disease, role in the cell, etc.
- Networks like KEGG are useful for either learning about a pathway or annotating / analyzing data in the context of a pathway
- KEGG includes computer-readable data formats



Interactomes

- Whereas networks & ontologies reflect a synthesis of knowledge, they do not integrate raw experimental data, i.e. data that can lead to new knowledge
- Interactomes reflect raw data from single experiments or integration of raw data from multiple experiments – they are not curated knowledge, they are experimental observations!
- Interactomes track “interactions” between two entities. Examples:
  - compound – protein interaction
  - protein – protein interaction
  - compound – gene expression interaction
- While the interaction has been observed, often the mechanism or the biological importance is not understood – a focus of Systems Biology



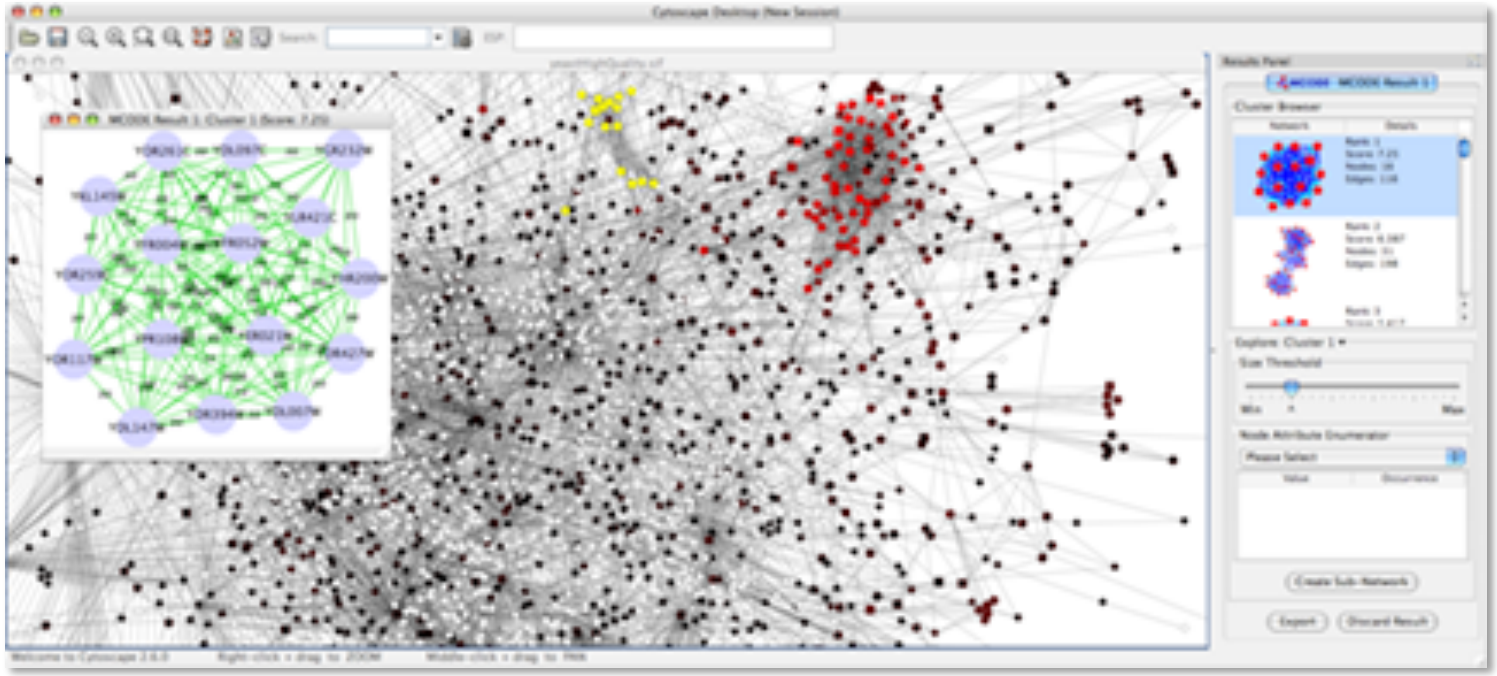
- 191 candidate protein–protein interactions measured by affinity-purification mass spectrometry
- Nodes (circles) = proteins; shaded gray for known motor proteins
- Edges (line) = interactions; shaded black if previously unknown
- Nodes are coloured by localization

A genomic toolkit to investigate kinesin and myosin motor function in cells. Maliga *et al*. 2013. *Nature Cell Biology* 15: 325-334

Data Formats & Software

- Interactome data can be stored in a number of formats (SIF, NNF, XGMML, SBML, etc.)
- There are a number of domain specific repositories but not one centralized repository – data files often provided as supplementary files with a scientific publication
- Data repositories and the role of publishers in ensuring scientific data is accessible and not lost over time is a very active discussion in the scientific community, e.g. *GigaScience Journal*, www.gigasciencejournal.com

Data Formats & Software



- The most popular software for analysis of interactome data is Cytoscape, www.cytoscape.org
- Cytoscape allows interactome mining and visualization but also analysis of outside data (e.g. RNA-Seq) in the context of observed interactomes
  - e.g. are the genes upregulated in my RNA-Seq experiment reflective of a specific sub-set of protein-protein interactions?

Connections & Cross-References

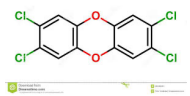
- Connections & Cross-References lead to standardization & normalization
  - Standardization – a common language is used to describe biological phenomena
  - Normalization – data, models, citations, and algorithms are annotated using the same common language, allowing easier sharing and comparison of data
- Connections & Cross-Reference combined with computer-readable representations leads to powerful analytical tools
- Examples
  - A protein-protein interactome dataset uses GenBank accessions for all the proteins observed
  - Gene Ontology annotates GenBank so all protein GenBank accessions are associated with GO terms
  - InterPro curates connections between Pfam and PROSITE models and Gene Ontology terms

Enrichment

- A powerful tool for high-level interpretation of complex experimental data, allowing investigator to detect biological processes otherwise not apparent
- Computational and statistical in nature – uses ontologies, networks, or interactomes to test if a specific biological process, pathway, or interaction is highlighted by the data
- Null hypothesis – no ontological term, biochemical or regulatory network, or known interactions are over-represented in my data (relative to controls or background signal)

Enrichment Example:

Danio rerio



dioxin

CONTROL

EXPOSED

RNA-Seq

1300 genes up-regulated in response to dioxin

VS

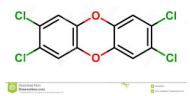
All genes in the genome

Gene Ontology enrichment analysis

GO terms for oxidative stress response, DNA damage repair statistically enriched in the up-regulated genes

Enrichment Example:

Danio rerio



dioxin

CONTROL

EXPOSED

RNA-Seq

1300 genes up-regulated in response to dioxin

VS

All genes in the genome

Gene Ontology enrichment analysis

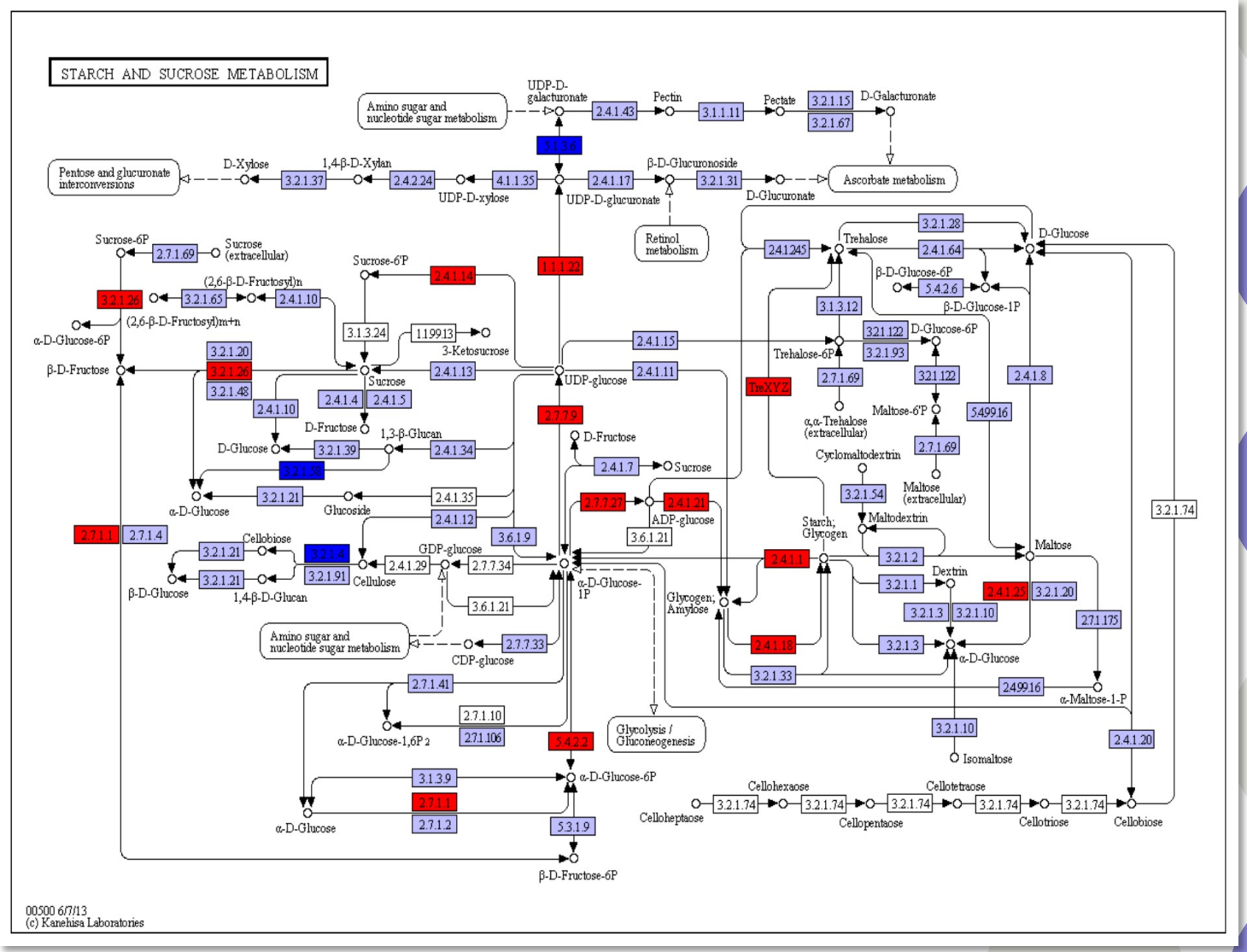
GO terms for oxidative stress response, DNA damage repair statistically enriched in the up-regulated genes

Each gene tagged with a different set of ontology terms in the genome annotation

Algorithm traverses the entire ontology to determine which higher-level ontology terms are involved in enrichment

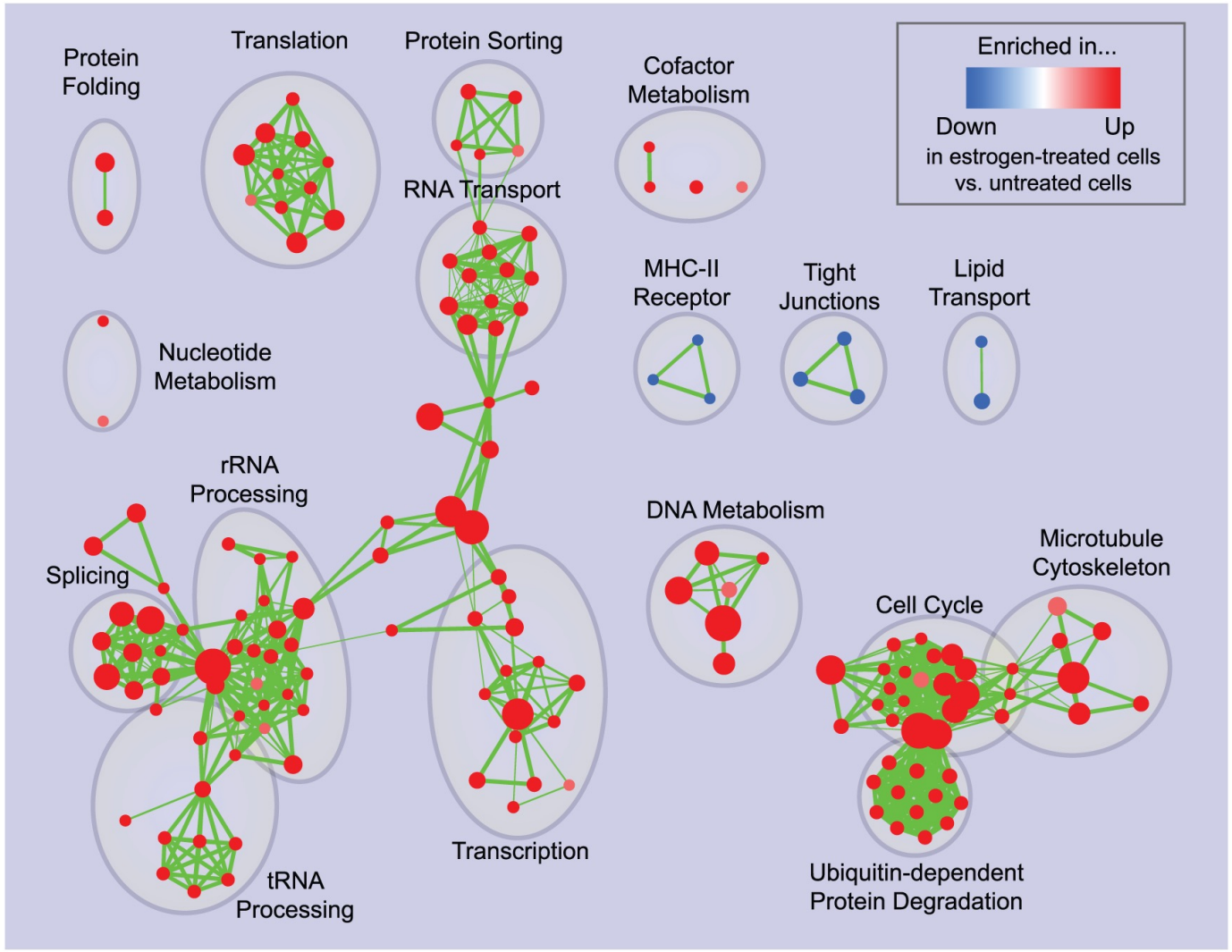
Enrichment Example - KEGG

- Purple – enzyme in cotton but not differentially regulated
- Blue – downregulated
- Red – upregulated
- Bowman *et al*. 2013. RNA-Seq Transcriptome Profiling of Upland Cotton (Gossypium hirsutum L.) Root Tissue under Water-Deficit Stress. *PLoS ONE* 8(12):e82634



Enrichment Example - Interactomes

- Gene expression with estrogen treatment examined
- Down- and upregulated genes examined in the context of known protein and other interactions

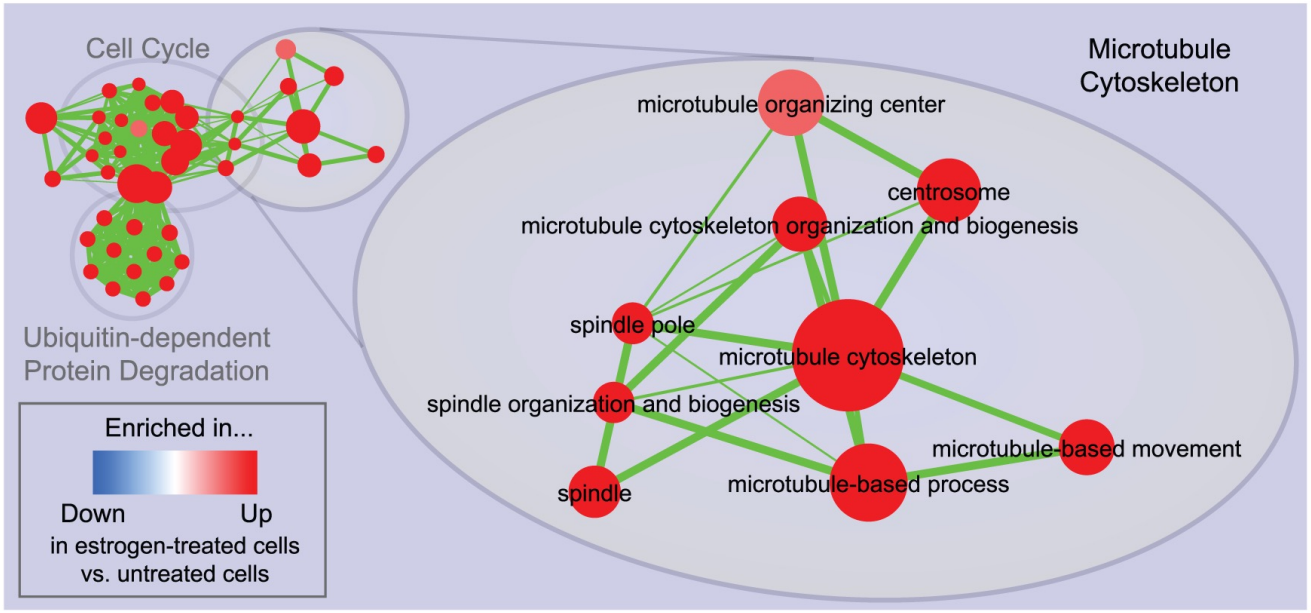


Enrichment map for estrogen treatment of breast cancer cells at 24 hours of culture.

Merico *et al*. 2010. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS One* 5(11):e13984.

Enrichment Example - Interactomes

- Cell cycle & microtubule cytoskeleton genes and processes are up-regulated at the transcript level – can this help guide drug discovery?



Enrichment map for estrogen treatment of breast cancer cells at 24 hours of culture.

Merico *et al*. 2010. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS One* 5(11):e13984.

Enrichment – Fisher's Exact Test

- Example: In human genome background (30,000 gene total), 40 genes are involved in p53 signaling pathway. A given gene list has found that 3 out of 300 belong to p53 signaling pathway. We ask the question if 3/300 is more than random chance comparing to the human background of 40/30000.

- Null Hypothesis: the gene list is specifically associated (enriched) in the p53 signaling pathway no more than random chance.

	Gene List	Genome
p53 pathway	3	40
not p53 pathway	297	29960

- Fisher Exact p-value = 0.008. Since $p<0.05$ the gene list is specifically associated (enriched) in the p53 signaling pathway more than random chance.

Enrichment – EASE

- The Fisher Exact test is sensitive to the inter-relationship of ontology terms and thus tends to over-predict.

- DAVID's EASE statistic is a modified Fisher Exact test that requires more genes as evidence of an enriched ontology term:

	Gene List	Genome
p53 pathway	3-1	40
not p53 pathway	297	29960

- EASE p-value = 0.06. Since p>0.05 the gene list is not associated (enriched) in the p53 signaling pathway more than random chance.

Comprehensive Antibiotic Resistance Database

- Integrates ontologies, sequence similarity models, and genome annotation algorithms to predict resistome and antibiogram for pathogens
  - Resistome – the complement of resistance genes in a pathogen
  - Antibiogram – the range of drug resistance and susceptibility in a pathogen
- All data and algorithms organized by the Antibiotic Resistance Ontology
- We will be working with the latest version in the lab