Challenges of integrating and analyzing different genomic data types

Jim Cavalcoli
University of Michigan
ISMB 2012 Bioinformatics Core Workshop

Outline

- Technological Challenges – Different platforms, different software and methods yields variation in results **(for Discussion)**
- Comparing and Integrating data from different methods and within and between samples
- Tools for adding biological relevance and integrating data

Method Comparisons

- Different platforms – same(?) biological measurement
  - RNA-Seq v. Microarray
  - Exome Variants v SNP arrays
- Same Samples – Different Biological Measurement
  - RNA expression & TF Binding
  - Genome Variants & Metabolite changes

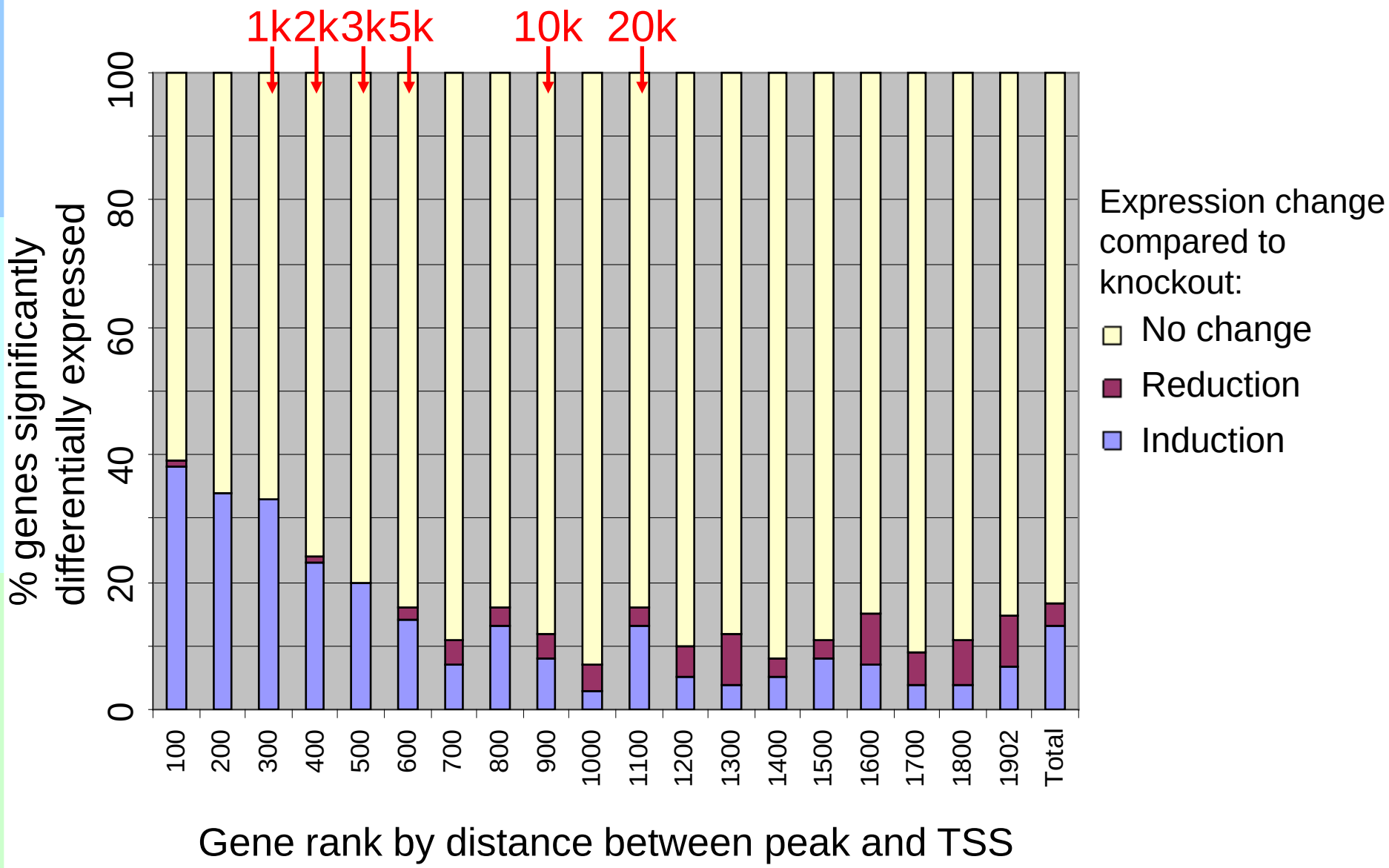
Challenges in Comparing Across Experiments

- Same Samples, but looking across different platforms
    - "From the same tissues/treatments we want to examine RNA-Seq and ChIP for TF SP1"
    - Or RNA-Seq and Methylation or Histone Binding profile
- **What RNA-seq comparison is best to integrate with the ChIP-seq data?** For the transcription factor studied, we had RNA-seq data for WT versus KO of the TF. We also had untreated WT versus WT with a treatment that stimulated the TF activity.
  - Although it might be expected that the WT versus KO for the TF would have served as the best comparison to integrated with the ChIP-seq data, it turned out that the WT treated versus untreated correlated better with the ChIP-Seq results.

Challenges Cont'd

- **What peak regions regulate which gene(s)?** A lot of ChIP-Seq bindings do not apparently regulate any genes. They may serve a different purpose, no purpose in the context under study, or they might not even be functional.
- Some genes share promoters, and some peaks are far from any gene.
- Given all these complexities, it's difficult to know which gene(s), if any, a peak should be paired with.
- Typically, the further you get from a gene's TSS, the less likely the TF is to regulate the corresponding gene. However, the fall off appears to be different for different TFs.

TF binding sites near TSSs correspond to differential expression



Integrating Biological Relevance

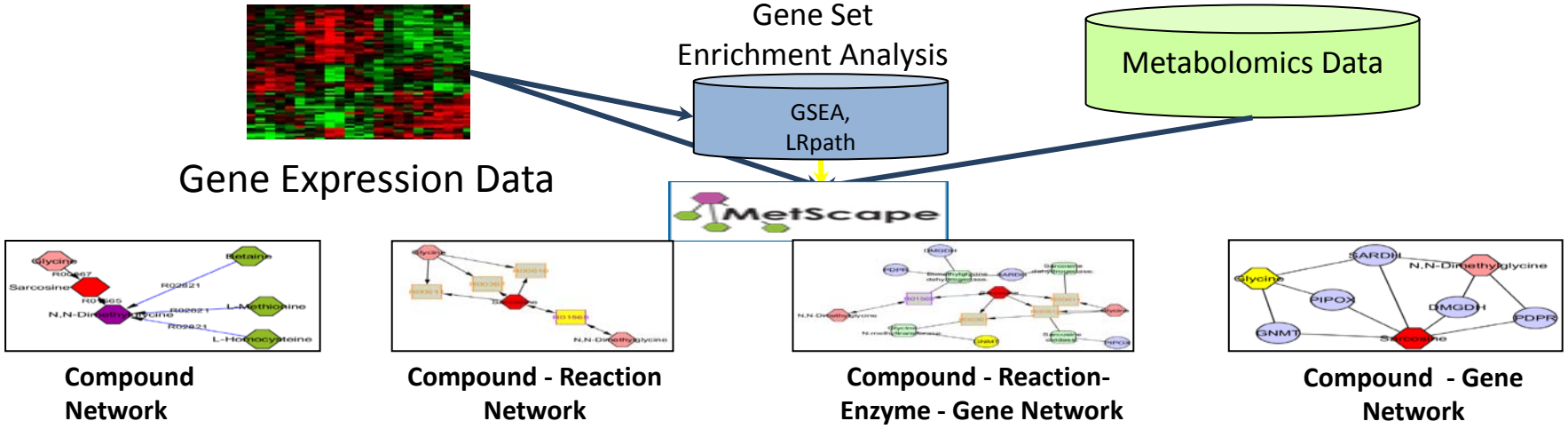
- Geneset enrichment –consenus by crowd
  - What sets are meaningful?
- Categorization and clustering –guilt by association
  - What level of binning will be helpful?
- Linking into Literature reference information
  - Accuracy of what is published? And NLP derrived?
- Workflows to maintain consistency and provenance **(outside the scope of today's talk)**

Gene Set Enrichment

- GSEA (Broad Institute)
  - http://www.broadinstitute.org/gsea/index.jsp
- LRPath (NCIBI)
  - http://lrpath.ncibi.org/
- DAVID
  - http://david.abcc.ncifcrf.gov/home.jsp
- ConceptGen
  - http://Conceptgen.ncibi.org

Cytoscape Plug-in: Metscape

- Provide the context for experimental data
- Utilize prior knowledge of metabolic networks
- Display multiple measurements across observations, time points, experimental conditions etc.
- Integrate multidimensional data
  - Can be gene expression and metabolomics data
- Provide broader view of metabolic networks
- Link to diseases



Karnovsky et al, Bioinformatics. 2012;28:373-80

http://metscape.ncibi.org/

http://metab2mesh.ncibi.org

Sartor et al, Bioinformatics. 2012;28(10):1408-10.

Metab2MeSH – Compound Annotation with MeSH Terms

NCIBI

Search Metab2MeSH | About Metab2MeSH

"cardiovascular diseases"

Search by: MeSH term Compound | Exact Match?

Metab2MeSH Search

Compound search examples:
methylmalonic acid
glucose-6-phosphate
MeSH term search examples:
diabetes mellitus
metabolism, inborn errors

Trouble finding the right MeSH term? Check the MeSH browser.

history : "brain injuries" -> cardiovascular diseases -> "cardiovascular diseases"

884 compounds found matching MeSH heading ""**cardiovascular diseases**""
= requery Metab2MeSH with compound or MeSH descriptor.
Disclaimer: Compound name matched Compound ID (from PubChem) at the time of computations. Due to PubChem updates, the list of synonyms may have changed.

download tab-delimited results

Filter by top level MeSH Heading: --Show All--

Compound Name	Compound ID(s)	MeSH Heading(s)	MeSH Descriptor	MeSH Qualifier	PubMed Articles*	P-Value	Q-Value	Fold Change	ChiSq
L-Homocysteine	91552	Diseases	Cardiovascular Diseases	-	1031	0.00e-1	0.00e-1	25.5	24022.8
DL-Homocysteine									
2-ammonio-4-sulfanylb									
(2S)-2-azaniumyl-4-su									
Spectrum_001666									
Lycopene, all-trans-									
LYCOPENE									
Prasugrel									
Norethindrone acetate									
Ambap51-98-9									
AC1LEXP8									

Compound Name	Compound ID(s)	MeSH Heading(s)	MeSH Descriptor	MeSH Qualifier	PubMed Articles*	P-Value	Q-Value	Fold Change	ChiSq
LYCOPENE	446925	Chemicals and Drugs	Carotenoids	-	1432	0.00e-1	0.00e-1	645.8	910388.2
LYCOPENE	446925	Chemicals and Drugs	beta Carotene	-	506	0.00e-1	0.00e-1	390.9	196022.3
LYCOPENE	446925	Organisms Technology, Industry, Agriculture	Lycopersicon esculentum	-	334	0.00e-1	0.00e-1	541.7	178117
LYCOPENE	446925	Chemicals and Drugs	Lutein	-	212	0.00e-1	0.00e-1	728.4	154005
LYCOPENE	446925	Chemicals and Drugs	Xanthophylls	-	200	0.00e-1	0.00e-1	482.2	96048.9
LYCOPENE	446925	Chemicals and Drugs	Anticarcinogenic Agents	-	231	0.00e-1	0.00e-1	166.3	37464.5
LYCOPENE	446925	Chemicals and Drugs	Antioxidants	-	569	0.00e-1	0.00e-1	49.7	26838.3
LYCOPENE	446925	Chemicals and Drugs	Vitamin E	-	188	1.00e-221	8.77e-218	36.8	6515.5
LYCOPENE	446925	Diseases	Prostatic Neoplasms	-	168	2.95e-187	2.21e-183	32.8	5091.4
LYCOPENE	446925	Anatomy Technology, Industry, Agriculture	Fruit	-	137	1.51e-186	1.13e-182	60.2	7800.5
LYCOPENE	446925	Chemicals and Drugs	Vitamin A	-	144	5.24e-163	3.47e-159	33.3	4486.1
LYCOPENE	446925	Technology, Industry, Agriculture	Dietary Supplements	-	120	2.91e-136	1.64e-132	33.8	3790.7
LYCOPENE	446925	Organisms Technology, Industry, Agriculture	Vegetables	-	84	1.09e-114	5.26e-111	58.7	4704.7

Web Services for NCIBI Tools

http://ws.ncibi.org/

- Data Services
  - Natural Language Processing Pipeline for PubMed and PMCOA
  - Gene2MeSH
  - Metab2MeSH
  - Michigan Molecular Interactions Database (MiMI)
  - Metabolomics
- Computational Analysis Services
  - Natural Language Processing
    - Sentence Segmentation
    - Phrase Structure Parsing
  - Gene Set Enrichment Analysis
    - LRPath
    - ThinkBACK

In Summary

- There are many remaining challenges in Technical areas yet the potential benefits to science and medicine are huge.

- It’s very important to harness all the knowledge that’s out there and this comes from integrating multiple data and annotation streams.