Scalable, Dynamic Analysis and Visualization for Genomic Datasets

Grant Wallace, Matthew Hibbs, Maitreya Dunham, Rachel Sealfon, Kai Li, and Olga Troyanskaya

Olga Troyanskaya
Assistant Professor
Department of Computer Science &
Lewis-Sigler Institute for Integrative Genomics
Princeton University

Laboratory for BioInformatics and Functional Genomics

Science has become **data-centric**

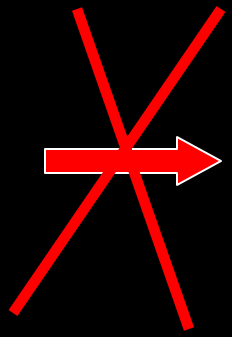
- Enormous amounts of data being observed
- Data then analyzed by experts in the field or computer scientists/statisticians

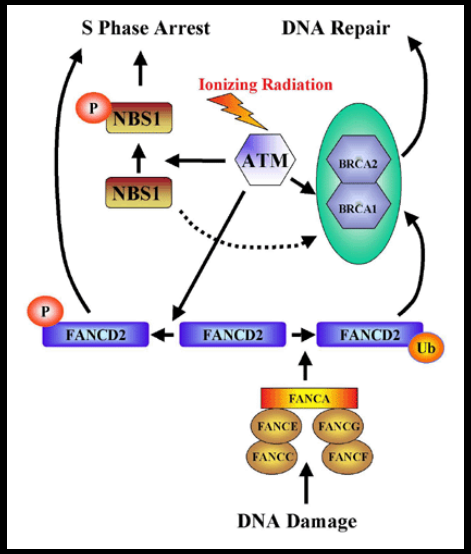
- Most disciplines – critical to have human insight during data analysis => visualization

Data growing in genomics: systems-level studies

- Cells & organisms are complex systems
- Many biological processes & diseases result from multiple changes on molecular level
- To understand and cure cancer & other diseases, need to observe and model cellular processes on a systems level

Data-Knowledge gap in biology





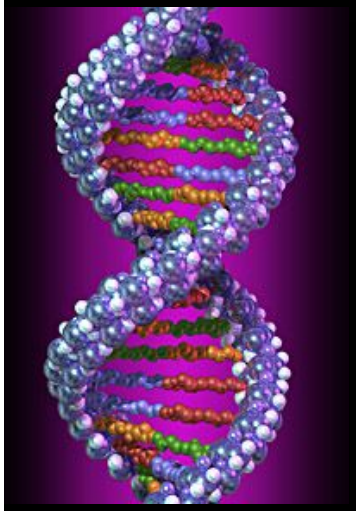
Explosion of genomic **DATA**

KNOWLEDGE of components and inter-relationships that lead to function

Outline

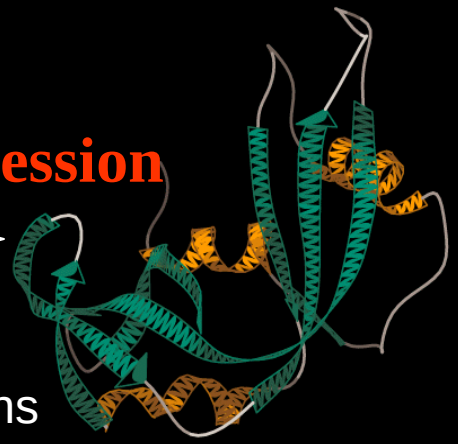
- Background
- Visualization and Analysis Solutions
  - ForestView
  - SPELL
  - GOLEM
- Integration of Solutions

Isn't genomics "done" with the human genome?

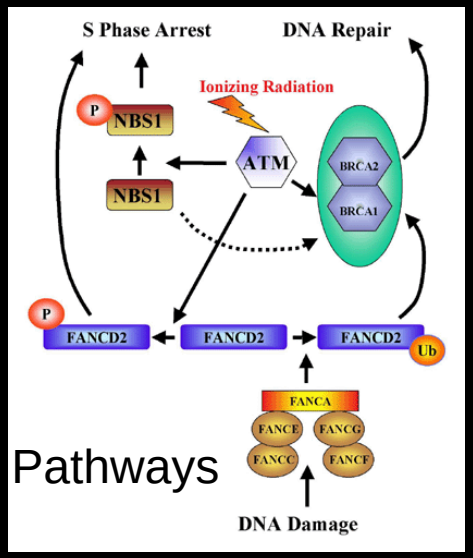


DNA

Gene Expression



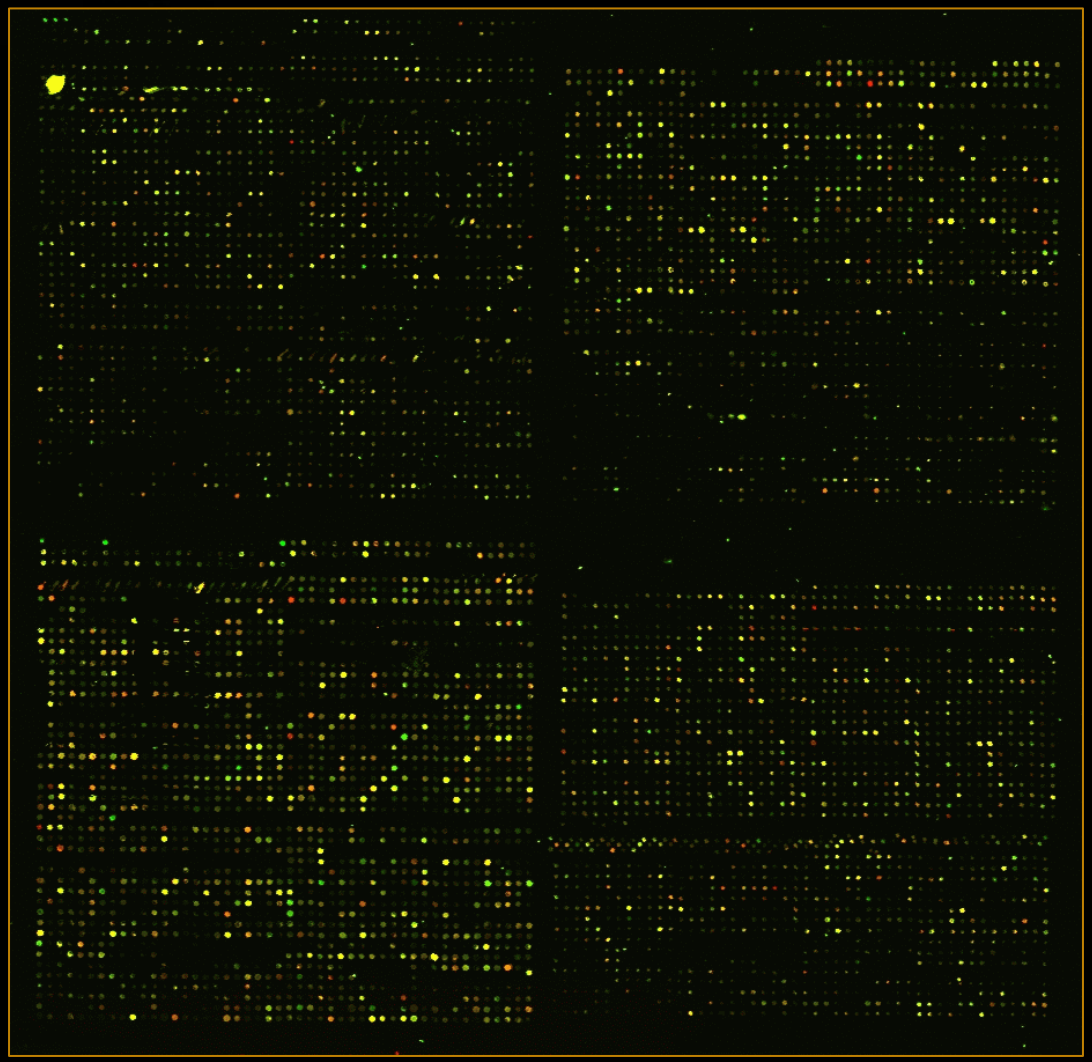
Proteins

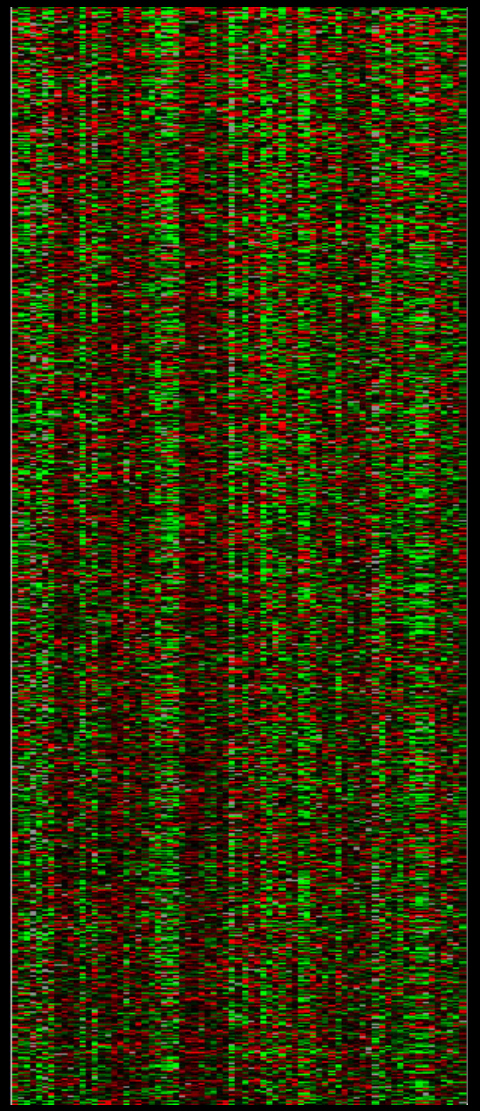


Pathways

Phenotypes

Microarray Results





Raw Image from Spellman et al., 98

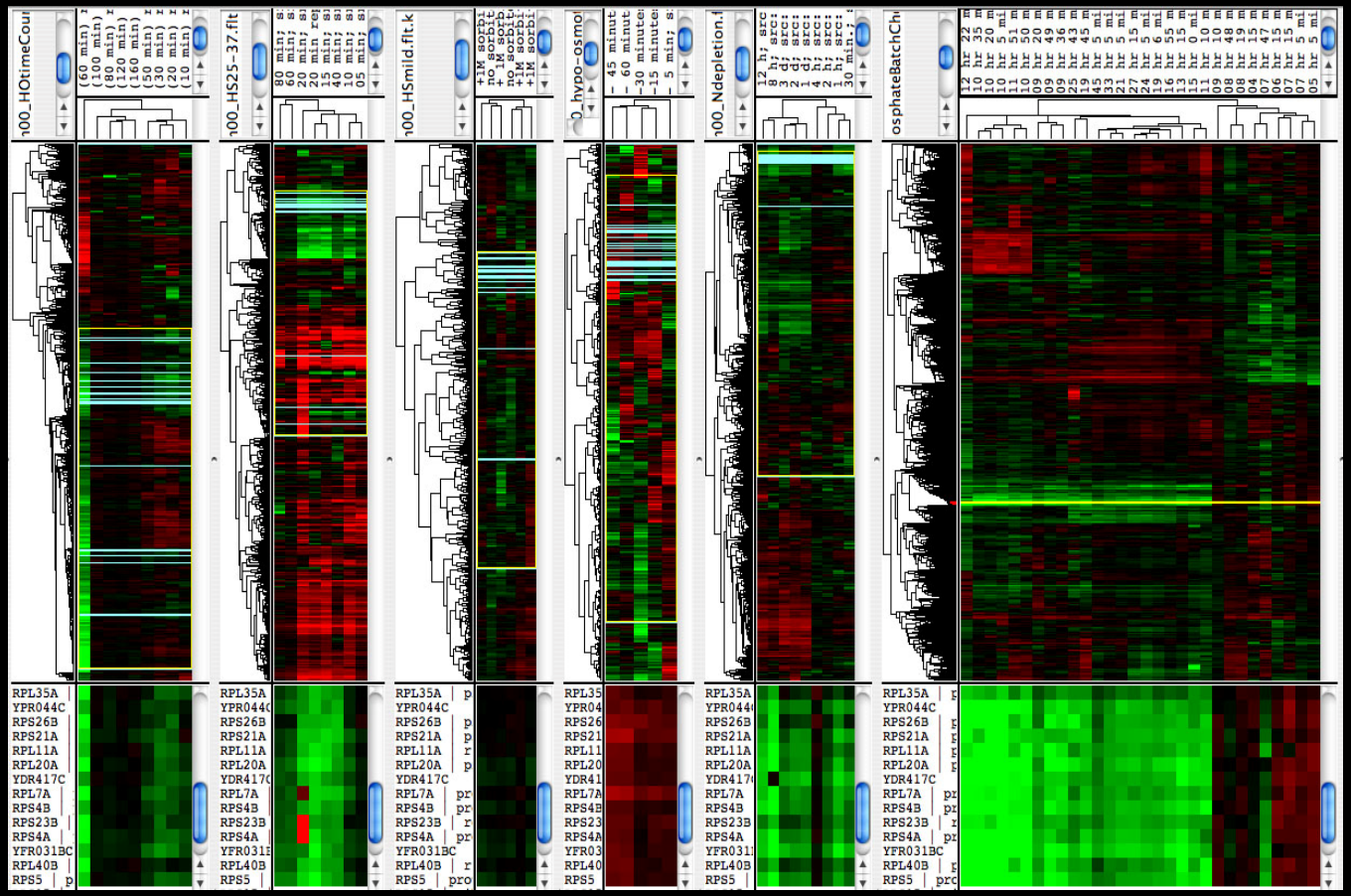
Challenges of Microarray Analysis

- LOTS of data
  - Hundreds of millions of measurements exploring cancer
- Strong need for collaboration
- Biologist's insight critical analysis => visualization-based analysis is critical
- Limited by effective visualization capacities
  - Screen space at a premium
  - Need to identify relevant data subsets
- Most analysis small-scale due to these issues

- Future of discovery will require larger-scale, integrated analysis and visualization

ForestView

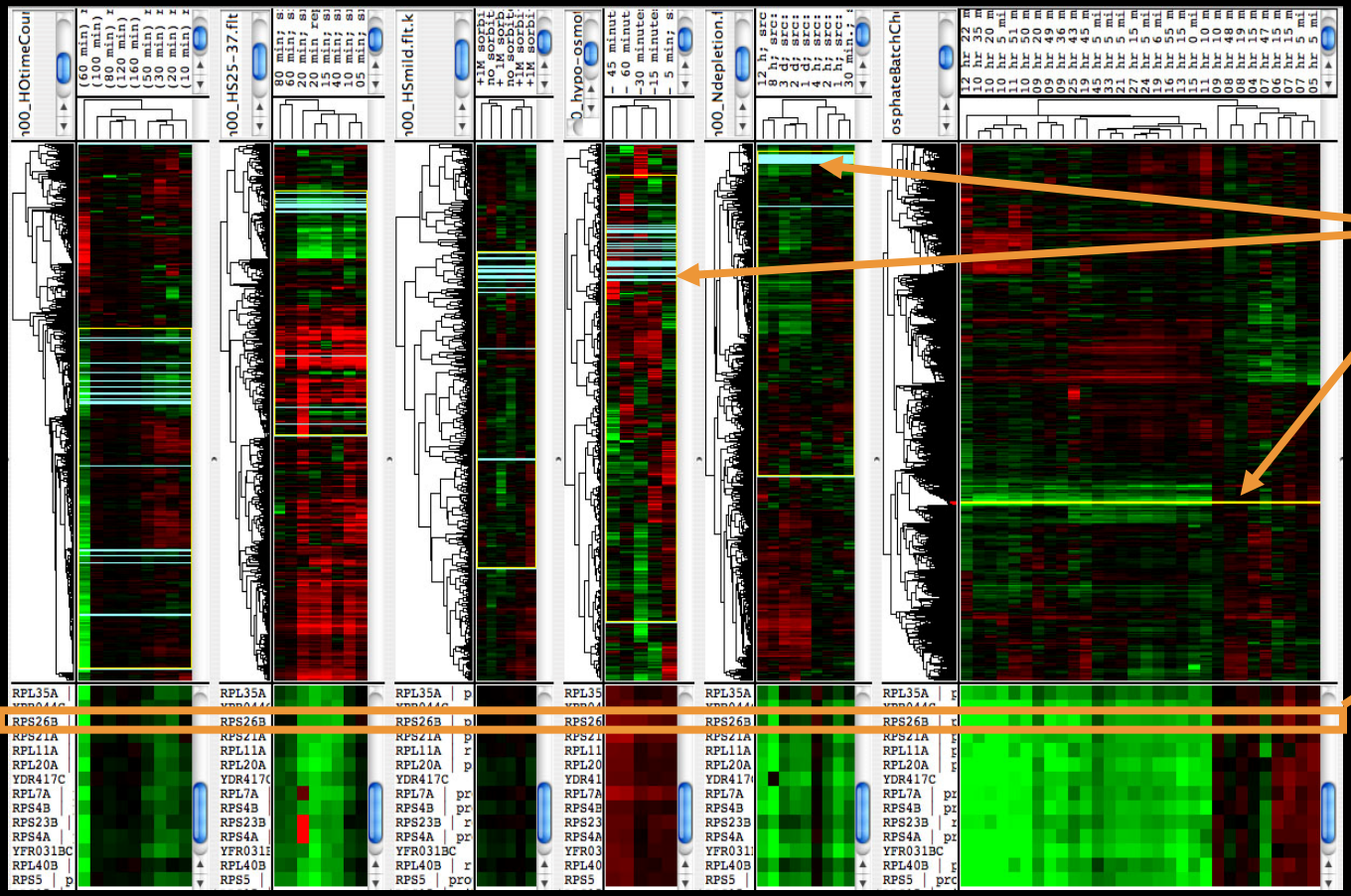
- Overview+Detail+Setting through Horizontal Integration
  - Places data in the larger context of other available data



Six diverse datasets linked together

ForestView

- Overview+Detail+Setting through Horizontal Integration
  - Places data in the larger context of other available data



Selections linked across datasets

Overview

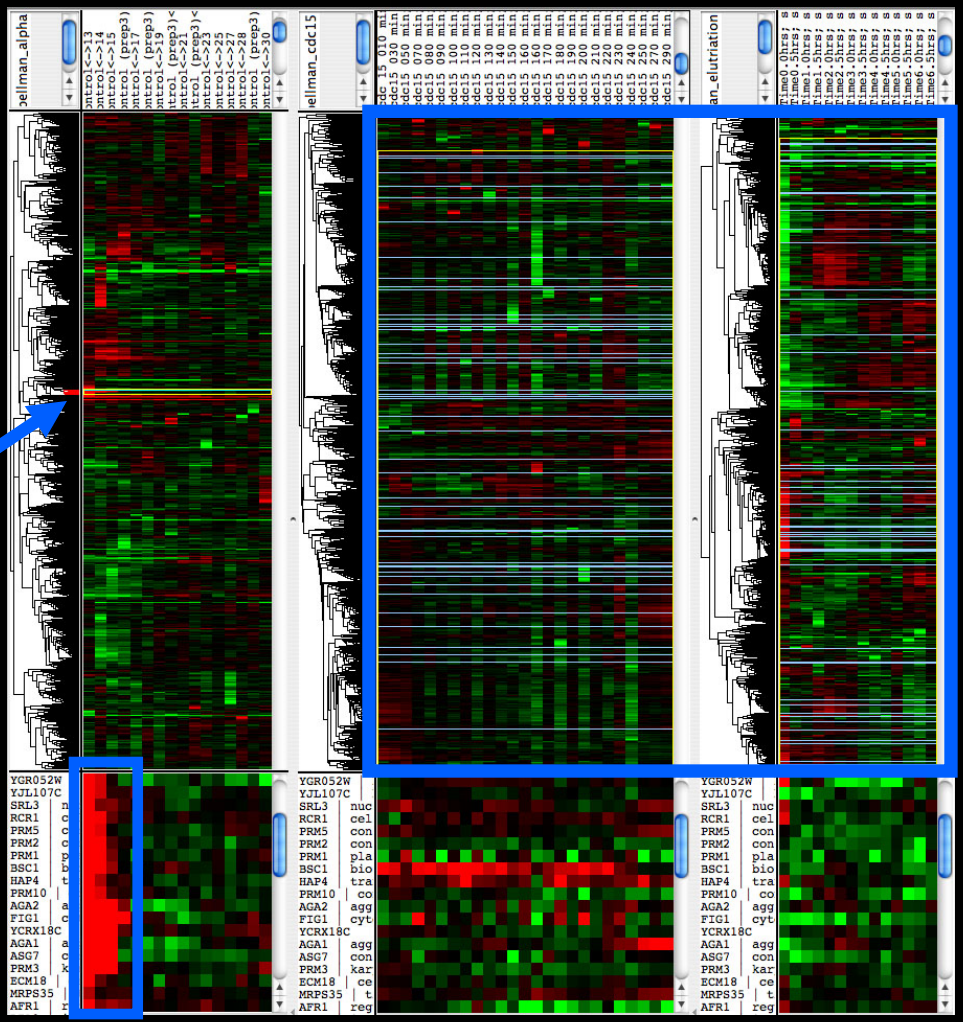
Details aligned horizontally

Detail/Zoom

Six diverse datasets linked together

ForestView Observation

alpha-factor cdc15 elutriation



Spellman cell cycle datasets

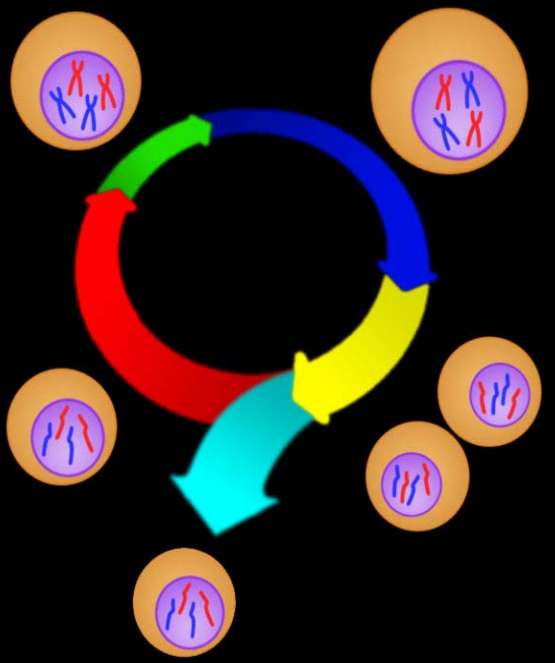
Three datasets studying cell cycle phases -- cancer can be related to cell cycle defects

Cl set
wi
tin

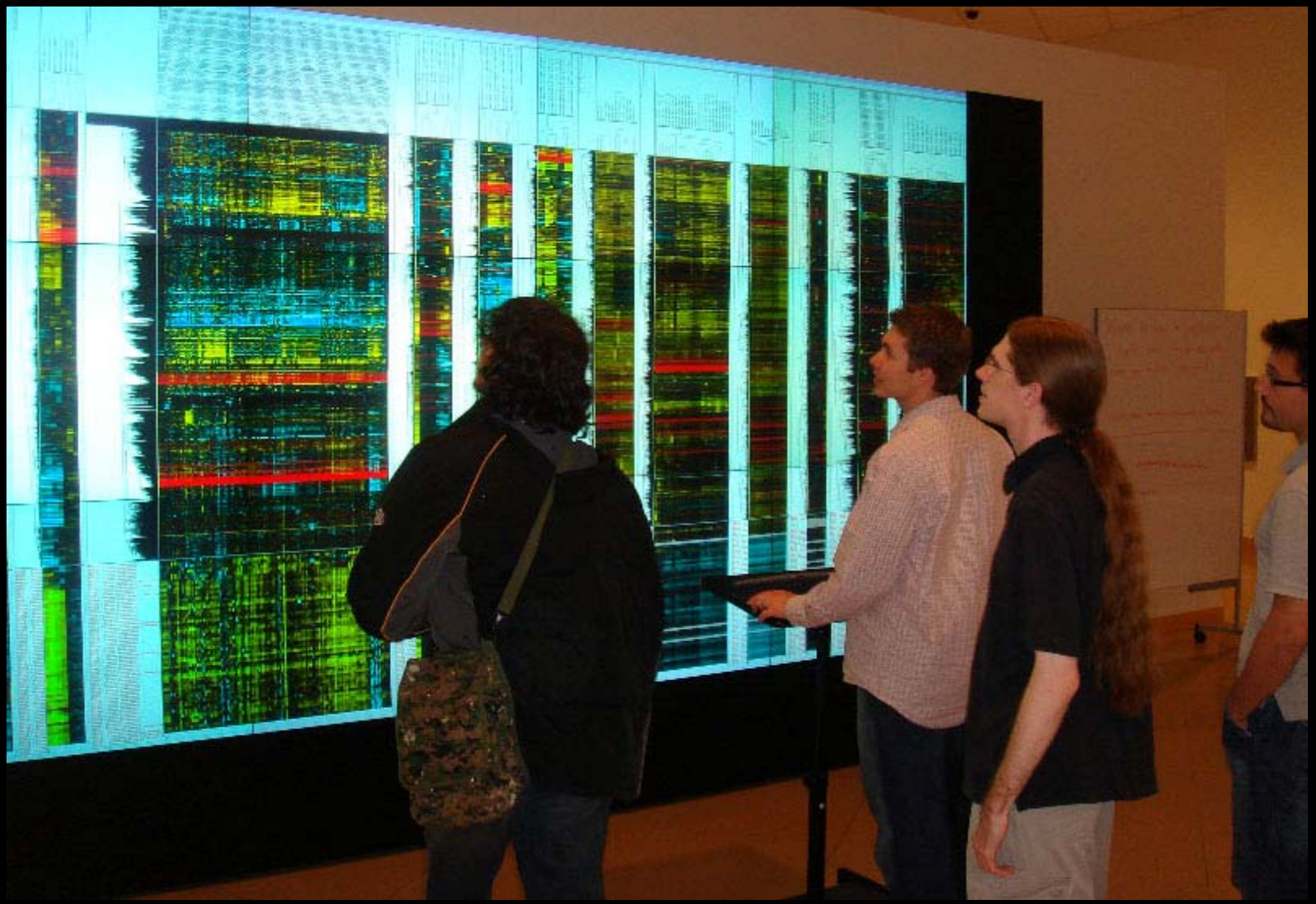
Th ed
in

G
ph ell

cycle -- artifact of the experiment



ForestView Scalability



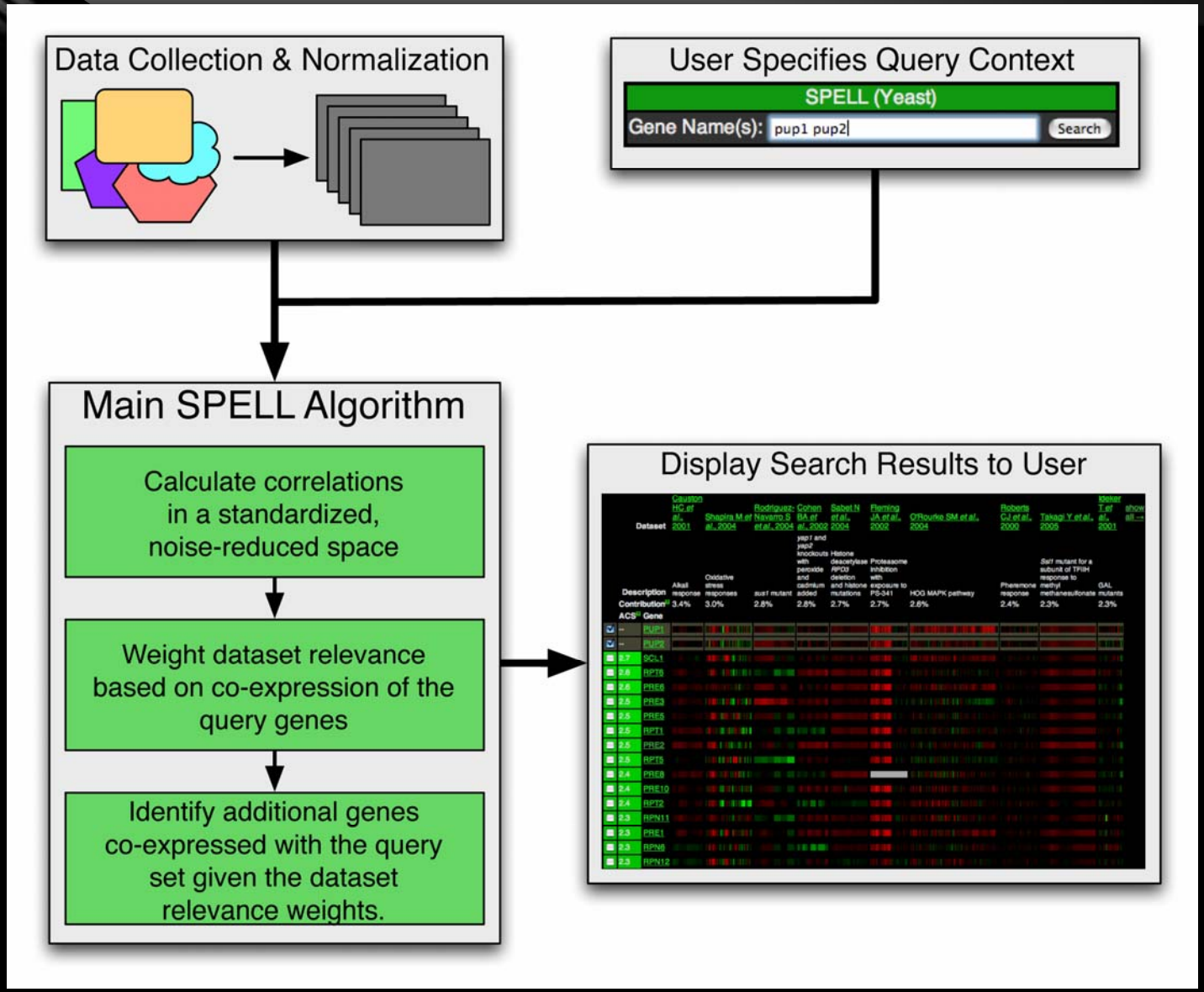
Analysis Across Multiple Datasets

- ForestView increases amount of data simultaneously viewed…
- But still much more data available

- Need to identify relevant datasets, and identify novel genes related to functions

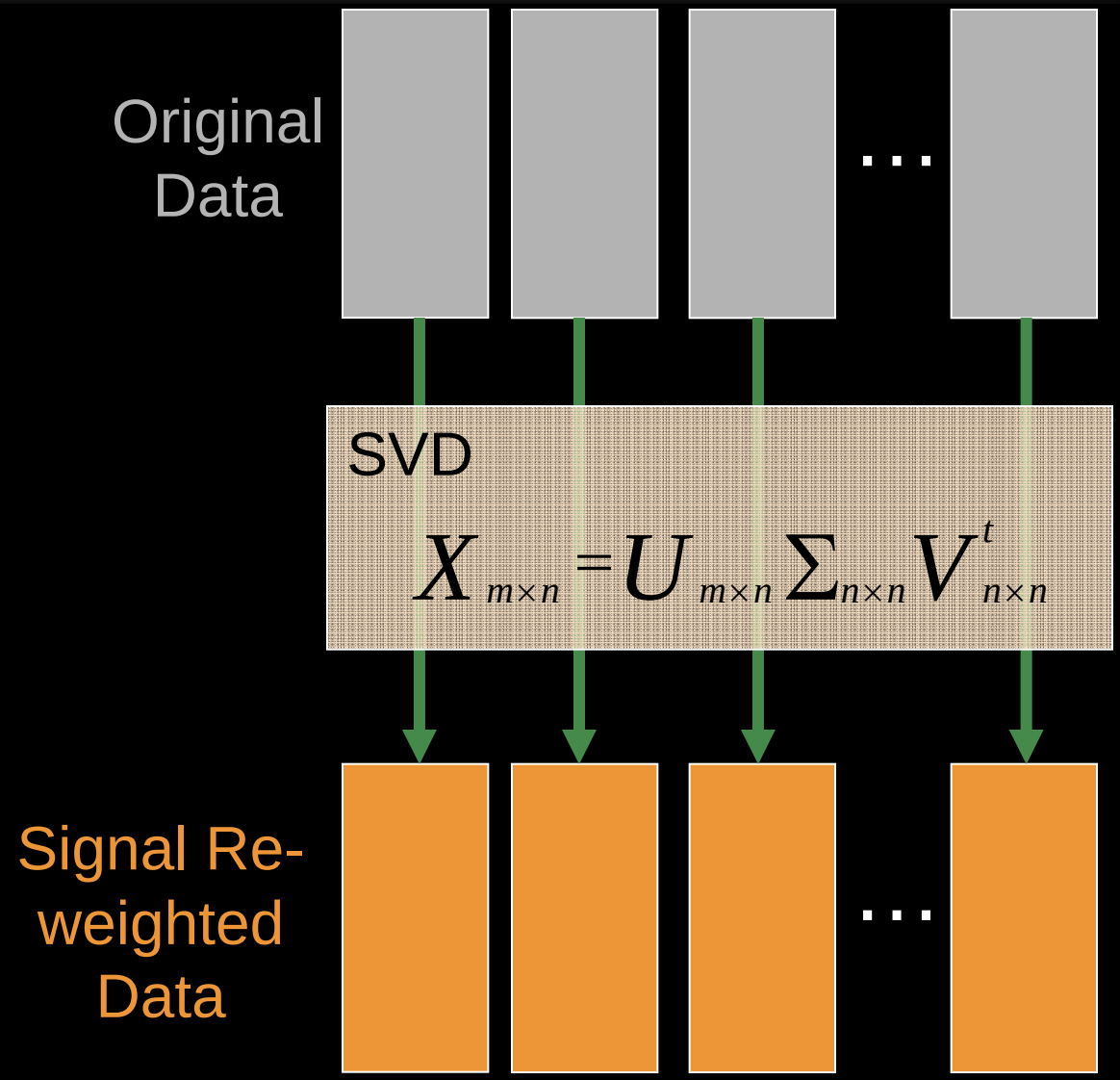
- Our solution:
- SPELL (Serial Patterns of Expression Levels Locator) – a similarity search and visualization engine

System Overview



Algorithm Details

1. Transform data through signal re-weighting



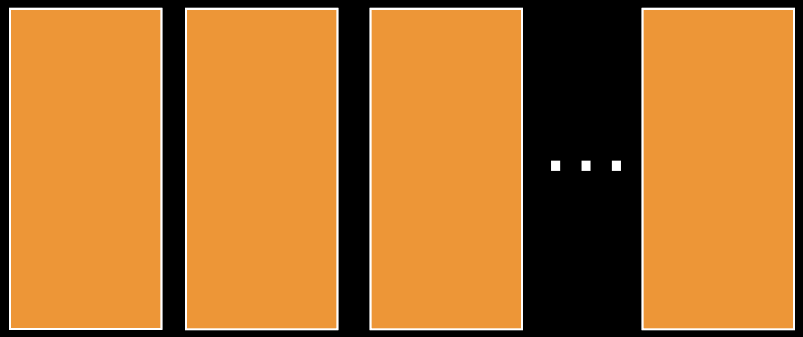
Algorithm Details

1. Transform data through signal re-weighting
2. Given a query, determine dataset weights

User specifies N query genes

Avg z-score between query pairs is dataset weight

$$w_d = \left(\frac{2}{N(N-1)}\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(z'_{q_i,q_j})$$

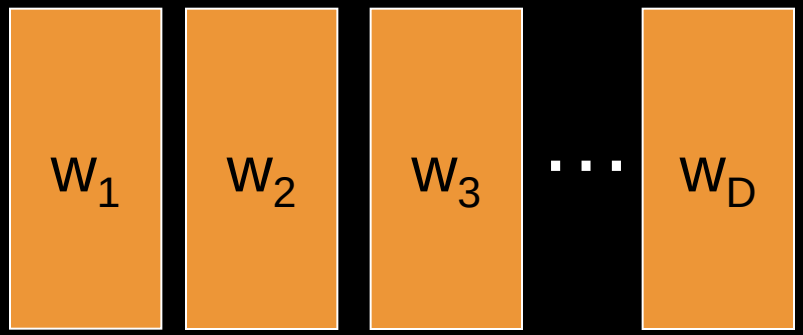


Algorithm Details

1. Transform data through signal re-weighting
2. Given a query, determine dataset weights
3. Identify additional co-expressed genes

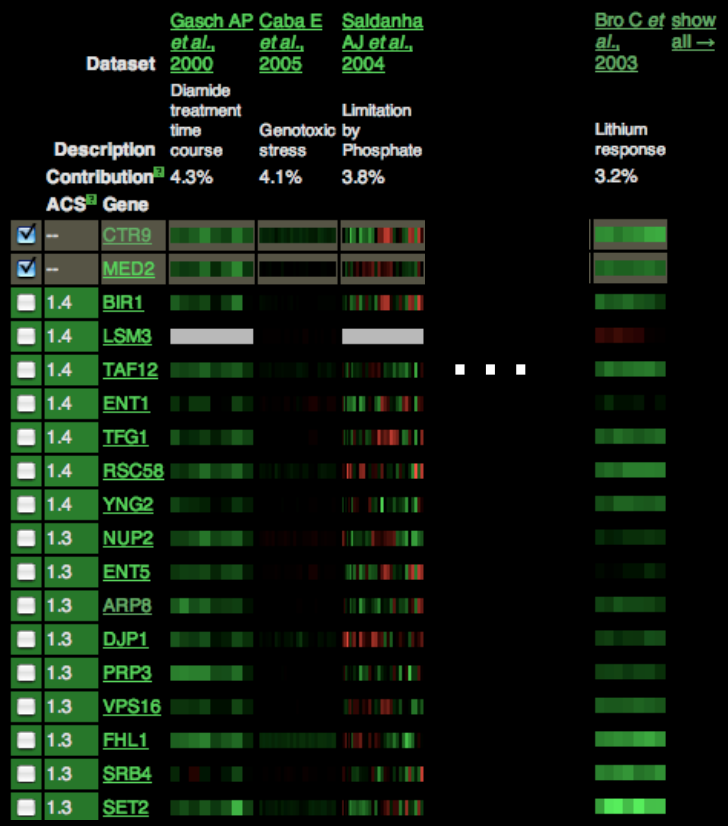
Scores for all other genes are weighted average z-scores

$$S_x = \left(\frac{1}{DN}\right)\sum_{d=1}^{D}\sum_{i=1}^{N} w_d z'_{x,q_i}$$



Algorithm Details

1. Transform data through signal re-weighting
2. Given a query, determine dataset weights
3. Identify additional co-expressed genes
4. Display results



Search from Cell Cycle Cluster

SPELL - *S. cerevisiae*

Search results for DEM1, CDC28

Refine Search

New Search

Dataset Listing

Show Expression Levels

	Dataset	Description	Contribution	ACS	Gene
	Spellman PT *et al.*, 1998	Cell cycle, cdc15 block-release	4.7%		
	Ideker T *et al.*, 2001	GAL mutants	3.5%		
	Spellman PT *et al.*, 1998	Cell cycle, alpha-factor block-release	3.3%		
	Pitkanen JP *et al.*, 2004	Phosphomannose isomerase *pmi40* deletion strain response to excess mannose	3.3%		
	Gasch AP *et al.*, 2000	Heat Shock from various temp to 37C	3.0%		
	Saldanha AJ *et al.*, 2004	Limitation by Leucine	2.9%		
	Gasch AP *et al.*, 2000	Dithiothreitol exposure time course (y13)	2.6%		
	Saldanha AJ *et al.*, 2004	Limitation by Uracil	2.5%		
	Yoshimoto H *et al.*, 2002	Ca(2+) exposure	2.5%		
	Eriksson PR *et al.*, 2005	*spt10* global transcription regulator null mutant	2.5%		
☑				--	CDC28
☑				--	DEM1
☐				1.9	POP7
☐				1.8	MSG5
☐				1.7	UBS1
☐				1.7	AME1
☐				1.6	MCM2
☐				1.6	ARL1
☐				1.6	PRP9
☐				1.5	PPS1
☐				1.5	FTH1
☐				1.5	TSC10
☐				1.5	ALG14
☐				1.5	YBR255W
☐				1.5	TAH1
☐				1.5	SWD3
☐				1.5	CSH1
☐				1.5	TOS1
☐				1.5	ERF2
☐				1.5	PBP2
☐				1.4	YBR259W
☐				1.4	HSL7

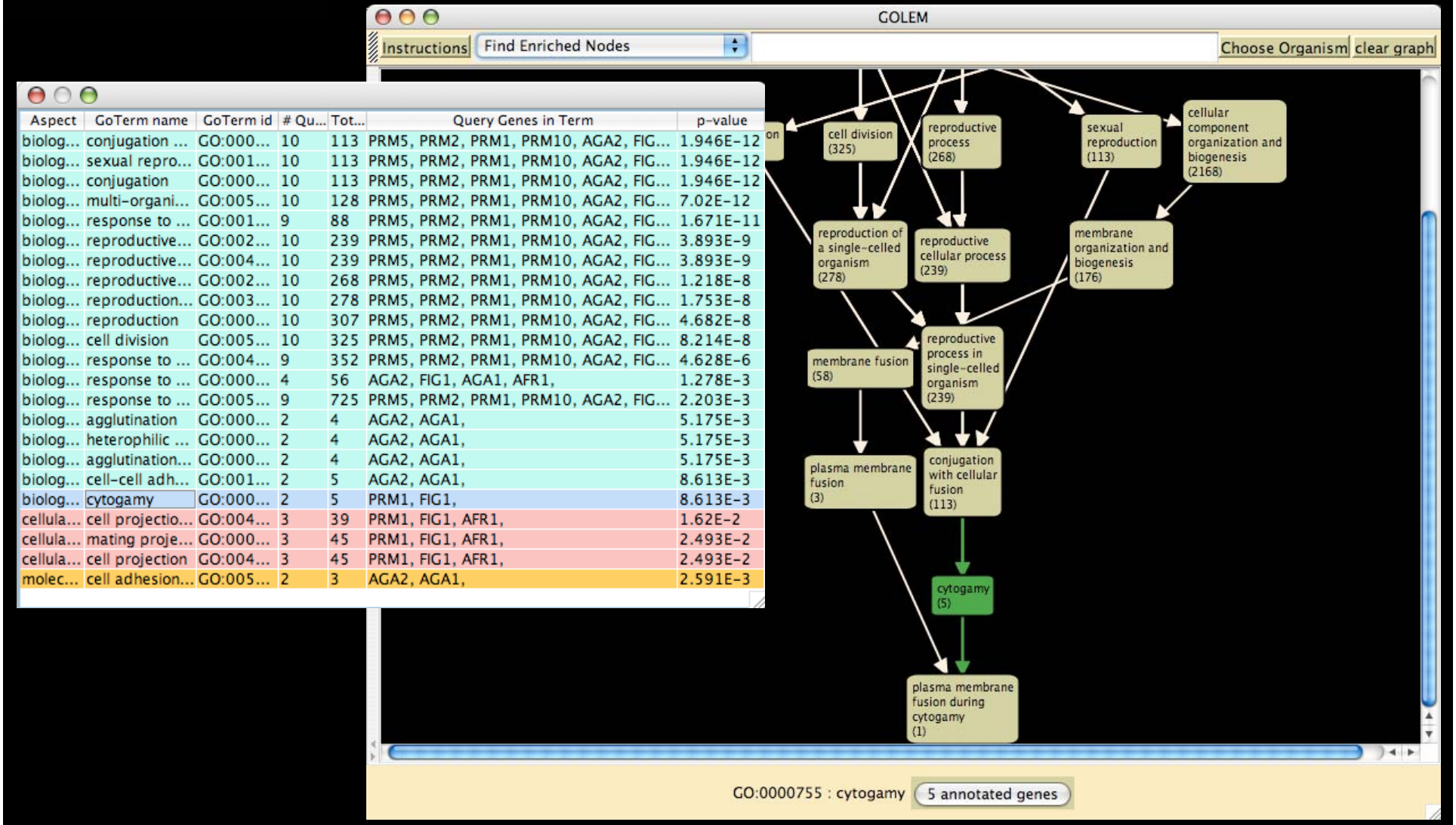
show all →

GOLEM: going beyond genomic data

- Also need analysis tools for verification and interpretation of search or visual analysis results
- Gene Ontology (GO) - curated hierarchical structure that represents known biology
- GOLEM - tool for finding GO enrichment, viewing ontology structure

GOLEM

- Using genes from the cell cycle example:



Aspect	GoTerm name	GoTerm id	# Qu...	Tot...	Query Genes in Term	p-value
biolog...	conjugation ...	GO:000...	10	113	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.946E-12
biolog...	sexual repro...	GO:001...	10	113	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.946E-12
biolog...	conjugation	GO:000...	10	113	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.946E-12
biolog...	multi-organi...	GO:005...	10	128	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	7.02E-12
biolog...	response to ...	GO:001...	9	88	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.671E-11
biolog...	reproductive...	GO:002...	10	239	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	3.893E-9
biolog...	reproductive...	GO:004...	10	239	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	3.893E-9
biolog...	reproductive...	GO:002...	10	268	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.218E-8
biolog...	reproduction...	GO:003...	10	278	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.753E-8
biolog...	reproduction	GO:000...	10	307	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	4.682E-8
biolog...	cell division	GO:005...	10	325	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	8.214E-8
biolog...	response to ...	GO:004...	9	352	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	4.628E-6
biolog...	response to ...	GO:000...	4	56	AGA2, FIG1, AGA1, AFR1,	1.278E-3
biolog...	response to ...	GO:005...	9	725	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	2.203E-3
biolog...	agglutination	GO:000...	2	4	AGA2, AGA1,	5.175E-3
biolog...	heterophilic ...	GO:000...	2	4	AGA2, AGA1,	5.175E-3
biolog...	agglutination...	GO:000...	2	4	AGA2, AGA1,	5.175E-3
biolog...	cell-cell adh...	GO:001...	2	5	AGA2, AGA1,	8.613E-3
biolog...	cytogamy	GO:000...	2	5	PRM1, FIG1,	8.613E-3
cellula...	cell projectio...	GO:004...	3	39	PRM1, FIG1, AFR1,	1.62E-2
cellula...	mating proje...	GO:000...	3	45	PRM1, FIG1, AFR1,	2.493E-2
cellula...	cell projection	GO:004...	3	45	PRM1, FIG1, AFR1,	2.493E-2
molec...	cell adhesion...	GO:005...	2	3	AGA2, AGA1,	2.591E-3

The future: dynamic, integrated search, analysis and visualization

SPELL query used to identify datasets relevant to a gene set

Datasets explored in ForestView, related genes selected

Results used to further explore current data, or refine SPELL search

Aspect	GoTerm name	GoTerm id	# Qu...	Tot...	Query Genes in Term	p-value
biolog...	conjugation ...	GO:000...	10	113	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.946E-12
biolog...	sexual repro...	GO:001...	10	113	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.946E-12
biolog...	conjugation	GO:000...	10	113	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.946E-12
biolog...	multi-organi...	GO:005...	10	128	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	7.02E-12
biolog...	response to ...	GO:001...	9	88	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.671E-11
biolog...	reproductive...	GO:002...	10	239	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	3.893E-9
biolog...	reproductive...	GO:004...	10	239	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	3.893E-9
biolog...	reproductive...	GO:002...	10	268	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.218E-8
biolog...	reproduction...	GO:003...	10	278	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	1.753E-8
biolog...	reproduction	GO:000...	10	307	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	4.682E-8
biolog...	cell division	GO:005...	10	325	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	8.214E-8
biolog...	response to ...	GO:004...	9	352	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	4.628E-6
biolog...	response to ...	GO:000...	4	56	AGA2, FIG1, AGA1, AFR1,	1.278E-3
biolog...	response to ...	GO:005...	9	725	PRM5, PRM2, PRM1, PRM10, AGA2, FIG...	2.203E-3
biolog...	agglutination	GO:000...	2	4	AGA2, AGA1,	5.175E-3
biolog...	heterophilic ...	GO:000...	2	4	AGA2, AGA1,	5.175E-3
biolog...	agglutination...	GO:000...	2	4	AGA2, AGA1,	5.175E-3
biolog...	cell-cell adh...	GO:001...	2	5	AGA2, AGA1,	8.613E-3
biolog...	cytogamy	GO:000...	2	5	PRM1, FIG1,	8.613E-3
cellula...	cell projectio...	GO:004...	3	39	PRM1, FIG1, AFR1,	1.62E-2
cellula...	mating proje...	GO:000...	3	45	PRM1, FIG1, AFR1,	2.493E-2
cellula...	cell projection	GO:004...	3	45	PRM1, FIG1, AFR1,	2.493E-2
molec...	cell adhesion...	GO:005...	2	3	AGA2, AGA1,	2.591E-3

GOLEM used to identify functional enrichments of clusters

Conclusions

- Scalable, dynamic visualization-based analysis enables novel biological discoveries
- Input from biology researchers critical in analysis
- Integration of multiple data sets is important in modeling and analyzing biological data
- Integration of visualization and analysis/search is important in genomics

Acknowledgements

Research Staff
* Camelia Chiriac

Postdoctoral Fellows
* Florian Markowetz
* Edo Airoldi
* David Hess

Graduate Students
* Chad Myers
* **Matthew Hibbs**
* Curtis Huttenhower
* Patrick Bradley
* Maria Chikina
* Yuanfang Guan
* Zafer Barutcuoglu
* Lars Bongo

Undergraduate Students
* Drew Robson
* Daniel Barrett
* Adam Wible
* **Rachel Sealfon**

Collaborator:

Kai Li

Laboratory for BioInformatics and Functional Genomics

http://function.princeton.edu

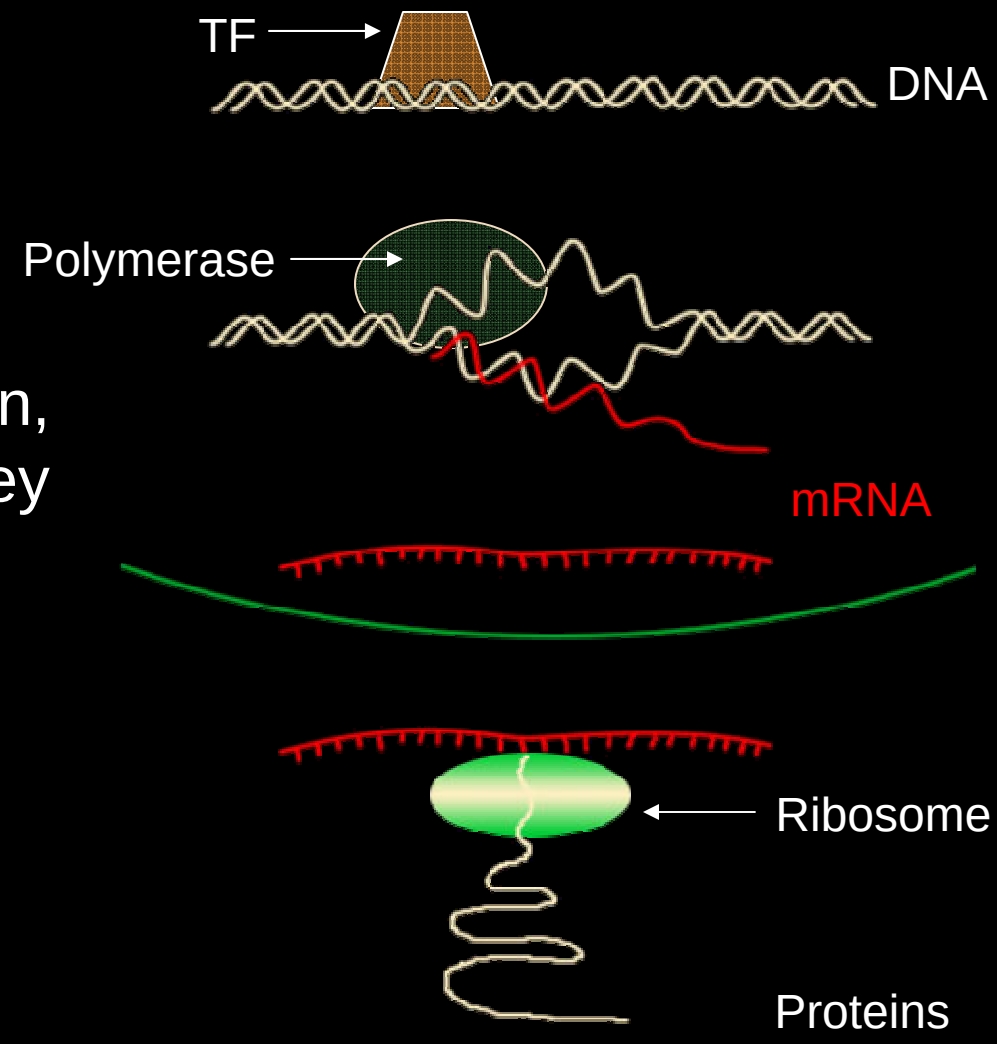
SPELL

bioPIXIE

GRIFn

Central Dogma

- Common cellular mechanisms to create proteins
- Understanding the function, coordination of proteins key to understand disease
- Gene expression microarrays give us a picture of cellular states



Basic Microarray Methodology