Transcription regulation (in *YEAST*): a genomic network

Nicholas Luscombe
Laboratory of ***Mark Gerstein***
Department of Molecular Biophysics and Biochemistry
Yale University

***Gerstein lab*: Haiyuan Yu**
***Teichmann lab*: Madan Babu**

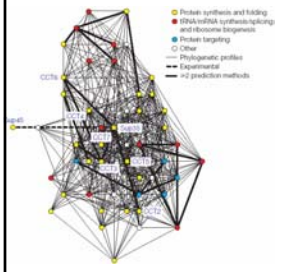
Comprehensive regulatory dataset in *YEAST*

	Dataset	Authors	URL	# of genes	# of regulations
Manual collection	TRANSFAC	Wingender, E., *et al.* 2001	http://transfac.gbf.de/TRANSFAC/	288	356
	Kepes' dataset	Guelzim, N., *et al.* 2002	http://www.nature.com/ng/journal/v31/n1/suppinfo/ng873_S1.html	477	906
ChIp-chip experiments	ChIp-chip data by Snyder's lab	Horak, C. E., *et al.* 2002	http://array.mbb.yale.edu/yeast/transcription/download.html	1560	2124
	ChIp-chip data by Young's lab	Lee, T. I., *et al.* 2002	http://web.wi.mit.edu/young/regulator_network/	2416	4358

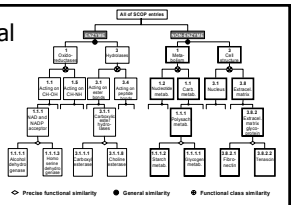
142 transcription factors
3,420 target genes
7,074 regulatory interactions

[Yu, Luscombe et al (2003), *Trends Genet*, 19: 422]

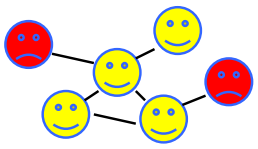
Networks provide a universal language to describe disparate systems



Protein interactions

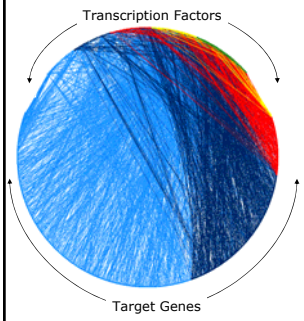


Hierarchies & DAGs



Social interactions

Comprehensive Yeast TF network

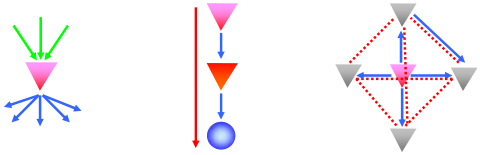


- Very complex network
- But we can simplify with standard graph-theoretic statistics:
  - Global topological measures
  - Local network motifs

[Barabasi, Alon]

1. Global topological measures

Indicate the gross topological structure of the network

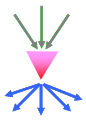


Degree | Path length | Clustering coefficient

[Barabasi]

1. Global topological measures

Number of incoming and outgoing connections



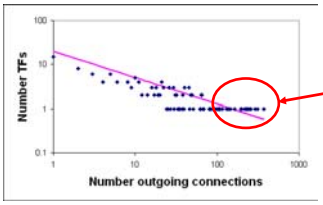
Degree

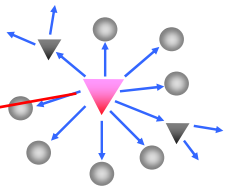
Incoming degree = 2.1
→each gene is regulated by ~2 TFs

Outgoing degree = 49.8
→each TF targets ~50 genes

[Barabasi]

Scale-free distribution of outgoing degree





Regulatory hubs

>100 target genes

Dictate structure of network

- Most TFs have few target genes
- Few TFs have many target genes

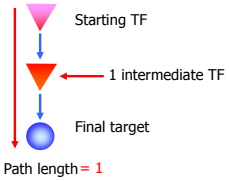
[Barabasi]

1. Global topological measures

Number of intermediate TFs until final target

Indicate how immediate a regulatory response is

Average path length = 4.7



Path length = 1

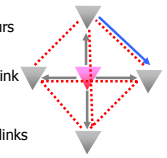
[Barabasi]

1. Global topological measures

Ratio of existing links to maximum number of links for neighbouring nodes

Measure how inter-connected the network is

Average coefficient = 0.11

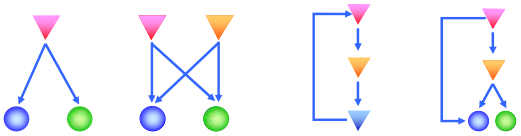


Clustering coefficient = 1/6 = 0.17

[Barabasi]

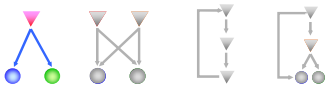
2. Local network motifs

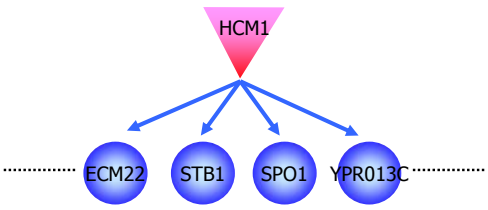
Regulatory modules within the network



[Alon]

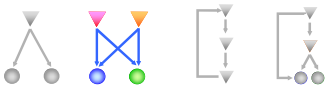
SIM = Single input motifs

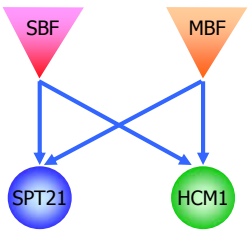




[Alon; Horak, Luscombe et al (2002), *Genes & Dev*, 16: 3017]

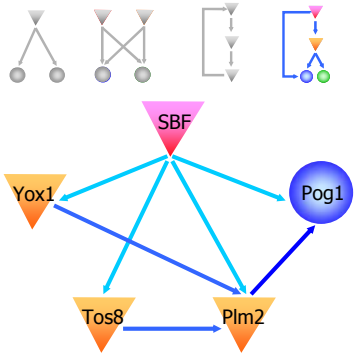
MIM = Multiple input motifs





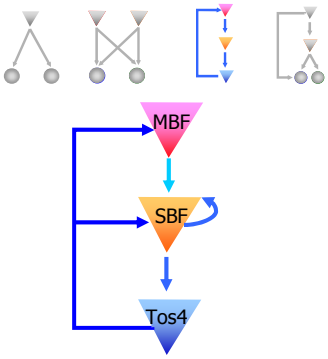
[Alon; Horak, Luscombe et al (2002), *Genes & Dev*, 16: 3017]

FFL = Feed-forward loops



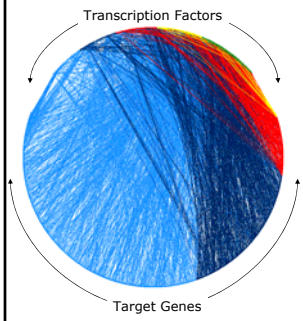
[Alon; Horak, Luscombe et al (2002), *Genes & Dev*, 16: 3017]

FBL = Feed-back loops



[Alon; Horak, Luscombe et al (2002), *Genes & Dev*, 16: 3017]

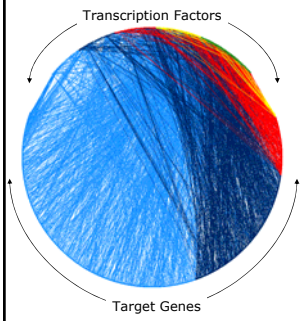
Comprehensive Yeast TF network



- Very complex network
- But we can simplify with graph-theoretic statistics:
  - Global topological measures
  - Local network motifs

[Barabasi, Alon]

Dynamic Yeast TF network



- Analyzed network as a static entity
- But network is *dynamic*
  - Different sections of the network are active under different cellular conditions
- Integrate gene expression data

[Luscombe et al, *Nature* (In press)]

Gene expression data

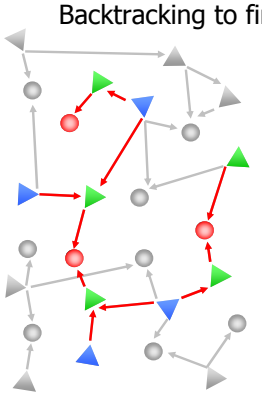
- Genes that are differentially expressed under five cellular conditions

Cellular condition	No. genes
Cell cycle	437
Sporulation	876
Diauxic shift	1,876
DNA damage	1,715
Stress response	1,385

- Assume these genes undergo transcription regulation

[Luscombe et al, *Nature* (In press)]

Backtracking to find active sub-network



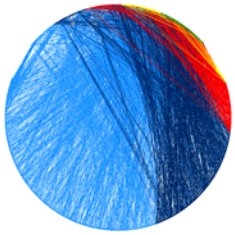
- Define differentially expressed genes
- Identify TFs that regulate these genes
- Identify further TFs that regulate these TFs

Active regulatory sub-network

[Luscombe et al, *Nature* (In press)]

Network usage under different conditions

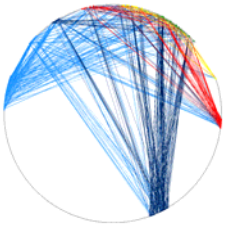
static



[Luscombe et al, *Nature* (In press)]

Network usage under different conditions

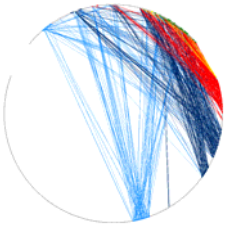
cell cycle



[Luscombe et al, *Nature* (In press)]

Network usage under different conditions

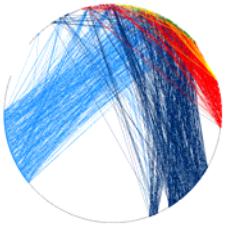
sporulation



[Luscombe et al, *Nature* (In press)]

Network usage under different conditions

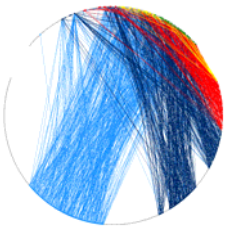
diauxic shift



[Luscombe et al, *Nature* (In press)]

Network usage under different conditions

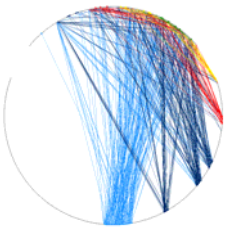
DNA damage



[Luscombe et al, *Nature* (In press)]

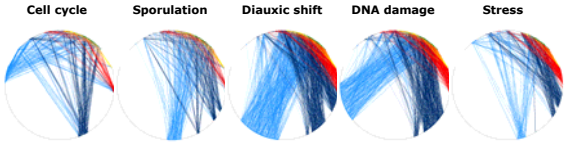
Network usage under different conditions

stress response



[Luscombe et al, *Nature* (In press)]

Network usage under different conditions



How do the networks change?

[Luscombe et al, *Nature* (In press)]

Methodology for analyzing network dynamics

Need a name!

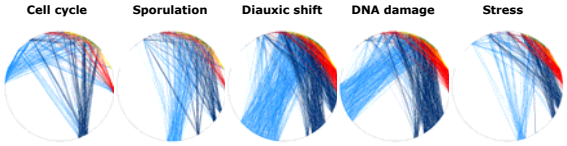
Dynamic **N**etwork **A**nalysis → ~~DNA~~

G(h)enomic **A**nalysis of **N**etwork **D(i)**namics → ~~GHANDI~~

Regulatory **A**nalysis of **N**etwork **Dy**namics → ~~RANDY~~

Statistical **A**nalysis of **N**etwork **Dy**namics → **SANDY**

Network usage under different conditions



SANDY:

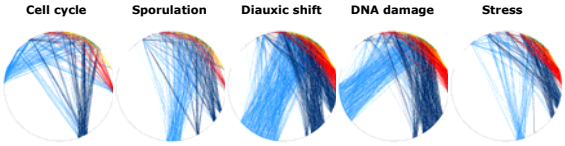
1. Standard graph-theoretic statistics:
- Global topological measures
- Local network motifs

2. Newly derived follow-on statistics:
- Hub usage
- Interaction rewiring

3. Statistical validation of results

[Luscombe et al, *Nature* (In press)]

Network usage under different conditions



SANDY:

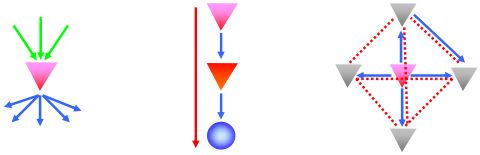
1. Standard graph-theoretic statistics:
- Global topological measures
- Local network motifs

2. Newly derived follow-on statistics:
- Hub usage
- Interaction rewiring

3. Statistical validation of results

[Luscombe et al, *Nature* (In press)]

1. Standard statistics - global topological measures

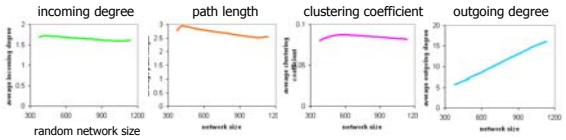


Degree | Path length | Clustering coefficient

[Barabasi]

Our expectation

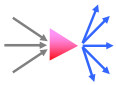
- Literature: Network topologies are perceived to be invariant
  - [Barabasi]
  - Scale-free, small-world, and clustered
  - Different molecular biological networks
  - Different genomes
- Random expectation: Sample different size sub-networks from complete network and calculate topological measures

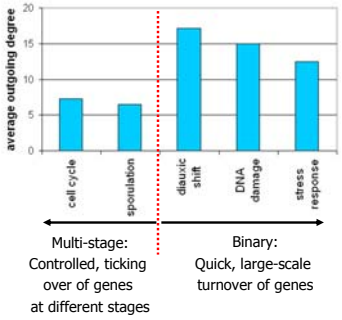


Measures should remain constant

[Luscombe et al, *Nature* (In press)]

Outgoing degree

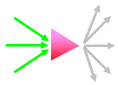


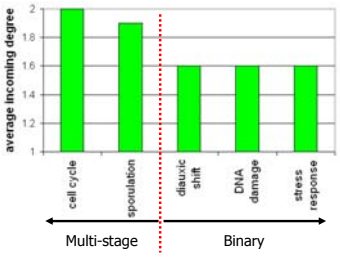


- "Binary conditions"
  → greater connectivity
- "Multi-stage conditions"
  → lower connectivity

[Luscombe et al, *Nature* (In press)]

Incoming degree

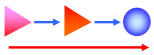


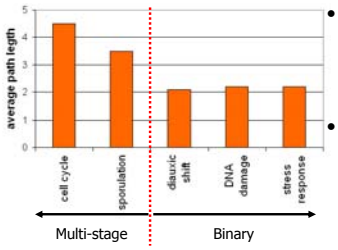


- "Binary conditions"
  → smaller connectivity
  → less complex TF combinations
- "Multi-stage conditions"
  → larger connectivity
  → more complex TF combinations

[Luscombe et al, *Nature* (In press)]

Path length

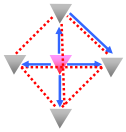


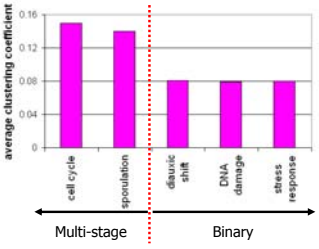


- "Binary conditions"
  → shorter path-length
  → "faster", direct action
- "Multi-stage" conditions
  → longer path-length
  → "slower", indirect action

[Luscombe et al, *Nature* (In press)]

Clustering coefficient

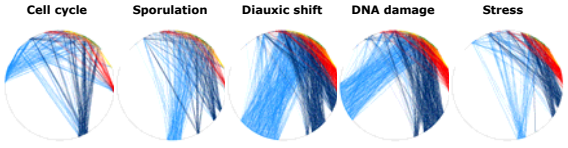




- "Binary conditions"
  → smaller coefficients
  → less TF-TF inter-regulation
- "Multi-stage conditions"
  → larger coefficients
  → more TF-TF inter-regulation

[Luscombe et al, *Nature* (In press)]

Network usage under different conditions



SANDY:

1. Standard graph-theoretic statistics:
- Global topological measures
- Local network motifs

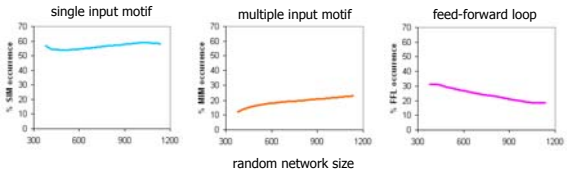
2. Newly derived follow-on statistics:
- Hub usage
- Interaction rewiring

3. Statistical validation of results

[Luscombe et al, *Nature* (In press)]

Our expectation

- Literature: motif usage is well conserved for regulatory networks across different organisms [Alon]
- Random expectation: sample sub-networks and calculate motif occurrence



Motif usage should remain constant

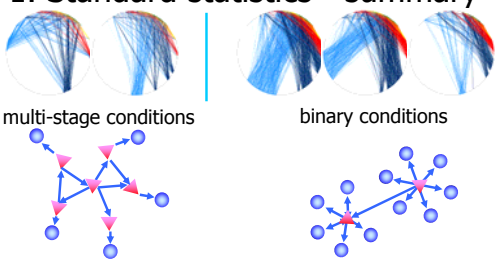
[Luscombe et al, *Nature* (In press)]

1. Standard statistics – local network motifs

Motifs		Cell cycle	Sporulation	Diauxic shift	DNA damage	Stress response
SIM		32.0%	38.9%	57.4%	55.7%	59.1%
MIM		23.7%	16.6%	23.6%	27.3%	20.2%
FFL		44.3%	44.5%	19.0%	17.0%	20.7%

[Luscombe et al, *Nature* (In press)]

1. Standard statistics - summary



multi-stage conditions

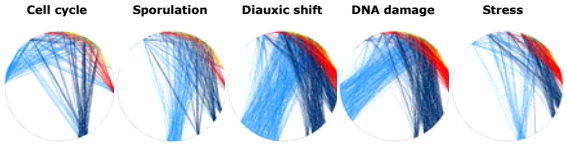
- fewer target genes
- longer path lengths
- more inter-regulation between TFs

binary conditions

- more target genes
- shorter path lengths
- less inter-regulation between TFs

[Luscombe et al, *Nature* (In press)]

Network usage under different conditions



SANDY:

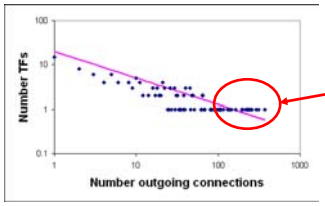
1. Standard graph-theoretic statistics:
- Global topological measures
- Local network motifs

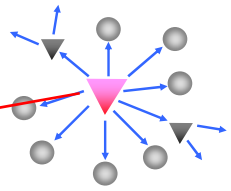
2. Newly derived follow-on statistics:
- Hub usage
- Interaction rewiring

3. Statistical validation of results

[Luscombe et al, *Nature* (In press)]

1. Follow-on statistics – network hubs





Regulatory hubs

>100 target genes

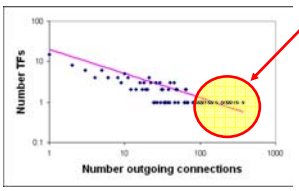
Dictate structure of network

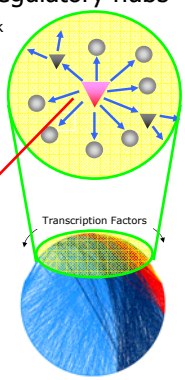
- Most TFs have few target genes
- Few TFs have many target genes

[Barabasi]

An aside: Essentiality of regulatory hubs

- Hubs dictate the overall structure of the network
- Represent vulnerable points
- **How important are hubs for viability?**
- Integrate gene essentiality data

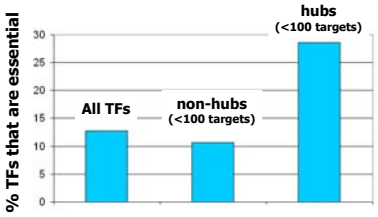




[Yu et al (2004), *Trends Genet*, 20: 227

Essentiality of regulatory hubs

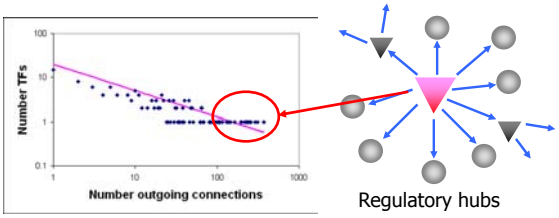
- Essential genes are *lethal* if deleted from genome
  - 1,105 yeast genes are essential
- Which TFs are essential?



***Hubs* tend to be more *essential* than non-hubs**

[Yu et al (2004), *Trends Genet*, 20: 227

Derived statistics 1 – network hubs



Regulatory hubs

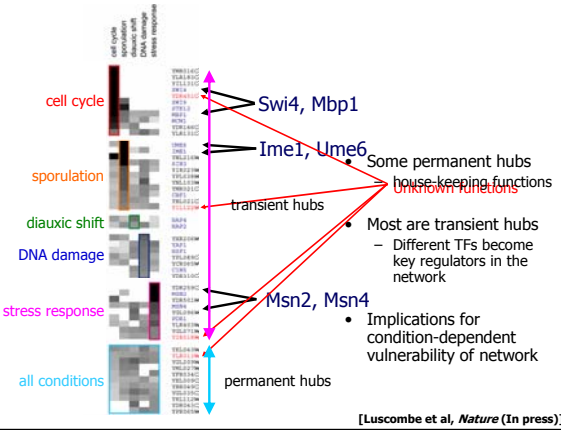
Do hubs stay the same or do they change over between conditions?
Do different TFs become important?

[Luscombe et al, *Nature* (In press)]

Our expectation

- Literature:
  - Hubs are permanent features of the network regardless of condition
- Random expectation: sample sub-networks from complete regulatory network
  - Random networks converge on same TFs
  - 76-97% overlap in TFs classified as hubs
  - *ie* hubs are permanent

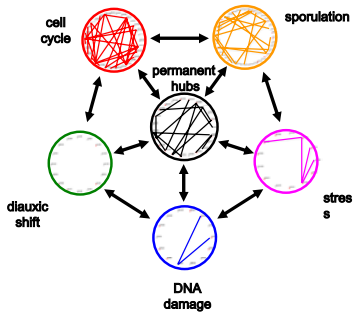
[Luscombe et al, *Nature* (In press)]



- Some permanent hubs
  - house-keeping functions
  - ~~Unknown functions~~
- Most are transient hubs
  - Different TFs become key regulators in the network
- Implications for condition-dependent vulnerability of network

[Luscombe et al, *Nature* (In press)]

Derived statistics 1 – network hubs

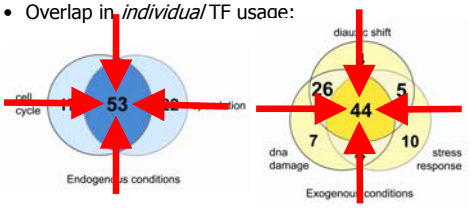


Network shifts its weight between centres

[Luscombe et al, *Nature* (In press)]

Derived statistics 2 – TF usage

- 95/142 TFs are used in > 1 condition
- 31 TFs are used in 5 conditions
- Overlap in *individual* TF usage:

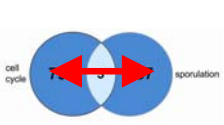


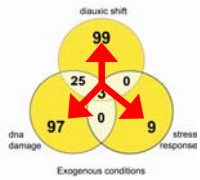
Great overlap in individual TF usage
Little regulatory specificity

[**]

Derived statistics 2 – TF usage

- BUT: Overlap in *combinatorial* TF usage:





Little overlap in combinatorial TF usage
Great regulatory specificity

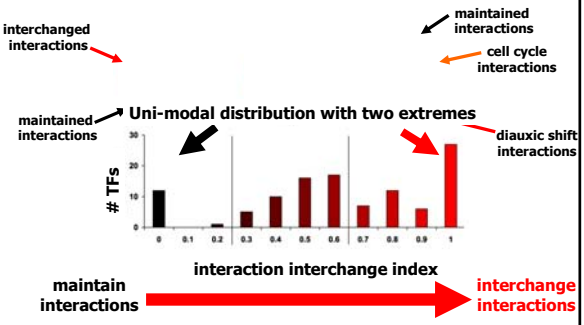
[**]

2. Follow-on statistics – interaction interchange

- Network undergoes substantial rewiring between conditions
- TFs must be replacing interactions with new ones between conditions
- Interchange Index = proportion of interactions that are maintained between conditions

[Luscombe et al, *Nature* (In press)]

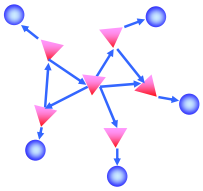
2. Follow-on statistics – interaction interchange



- **TFs maintain/interchange interactions by differing amounts**
- **Regulatory functions shift as new interactions are made**

[Luscombe et al, *Nature* (In press)]

Regulatory circuitry of cell cycle time-course



- Multi-stage conditions have:
  - longer path lengths
  - more inter-regulation between TFs
- How do these properties actually look?
  - examine TF inter-regulation during the cell cycle

[Luscombe et al, *Nature* (In press)]

Regulatory circuitry of cell cycle time-course

Phase-specific TFs show ***serial*** inter-regulation

→ drive cell cycle forward through time-course

[Luscombe et al, *Nature* (In press)]

Regulatory circuitry of cell cycle time-course

2. parallel inter-regulation

Ubiquitous and phase-specific TFs show ***parallel*** inter-regulation

Many ubiquitous TFs are permanent ***hubs***

→ Channel of communication between house-keeping functions and cell cycle progression

[Luscombe et al, *Nature* (In press)]

Network dynamics in FANTOM 3?

- SANDY framework is generally applicable to many types of networks
- Example applications:
  - Tissue-specific regulatory sub-networks
  - Sub-network time-courses during development
  - Conservation of networks and sub-networks in mouse and human
  - Prediction of TF hubs
  - Human disease loci, and vulnerability of network (eg hubs and lethality)
  - Integrated molecular network (protein-protein, protein-DNA, RNA-RNA etc)
- Current challenges:
  - How do we get to a network using current data?
  - Networks are obviously more complex than in yeast eg alternative TSS

"They say they built the train tracks over the Alps between Vienna and Venice before there was a train that could make the trip.

They built it anyway.

They knew one day a train would come."

Movie - Under the Tuscan Sun

Mark Gerstein
Haiyuan Yu

Sarah Teichmann
Madan Babu

Summary 1

- Incorporation of many data sources to build a comprehensive network of regulatory interactions
  - 142 TFs
  - 3,420 target genes
  - 7,074 regulatory interactions
- Very complex network of inter-regulation

Summary 2

- Dynamic usage of the regulatory network
- Different sections of the network are used in different cellular conditions
- How do we measure the changes?
  - SANDY
- Expectation is for networks to be constant

Summary 3

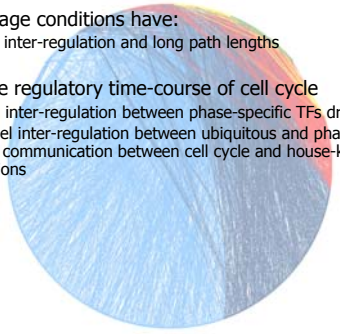
- Standard statistics:
- Topological measures indicate global network structure
  - Multi-stage: fewer targets, long path lengths, high clustering
  - Binary: many targets, short path lengths, low clustering
- Network motifs measure local network structure
  - Multi-stage: indirect-acting motifs - SIMS
  - Binary: direct-acting motifs – MIMS
- *Aside:* gene expression profiles
  - Targets are co-expressed
  - TF and target are often time-shifted

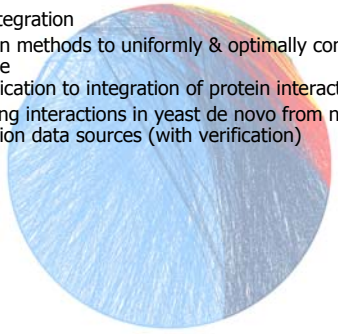
Summary 4

- Derived statistics:
- Hubs:
  - Hubs tend to be essential
  - Some hubs are permanent
  - Most hubs are transient
- TF usage:
  - Individual TF usage has large overlap
  - Combinatorial TF usage has small overlap -> regulatory specificity
- Interactions interchange:
  - TFs maintain or interchange interactions between conditions to different degrees
  - Highly interchanging TFs shift regulatory functions

Summary 5

- Multi-stage conditions have:
  - Much inter-regulation and long path lengths
- Examine regulatory time-course of cell cycle
  - Serial inter-regulation between phase-specific TFs drive cell cycle
  - Parallel inter-regulation between ubiquitous and phase-specific TFs allow communication between cell cycle and house-keeping functions



Summary 6

- Data integration
- Bayesian methods to uniformly & optimally combine evidence
  (in application to integration of protein interaction data)
- Predicting interactions in yeast de novo from non-interaction data sources (with verification)