

Integration of Expression Data and Transcriptional Control Network: Significant Regulators Driving Expression Changes

Andrey Sivachenko, Anton Yuryev, Nikolai Daraselia, Ilya Mazo
Ariadne Genomics Inc.

What we get:

- Whole-genome snapshot of transcription level changes
- High levels of noise
- No structure: “One gene at a time”

What we want:

- System-level information
- Molecular mechanisms, pathways
- Quantifiable hypotheses ranked by objective criteria



Data integration

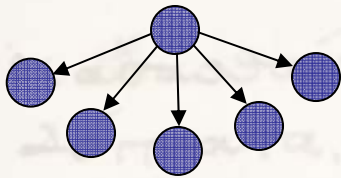
What we do:

- Integrate expression data with known transcription regulation network
- Consider set of targets $T(r)$ of every regulator r
- Search for $T(r)$ exhibiting consistent changes
- Significant expression changes in $T(r) \Leftrightarrow$ “significant regulator”

Transcription regulation network:

- Transcription control relations $A \rightarrow B$ mined from PubMed
- Natural language processing full-sentence parsing algorithm
- ~12,000 direct and indirect transcription control relations among ~3,800 genes; ~90% reliability.

Model



- *Null hypothesis*: activity of the regulator is *not* associated with the studied changes in cellular state.
- Under the *null*, targets' expression is a random sample

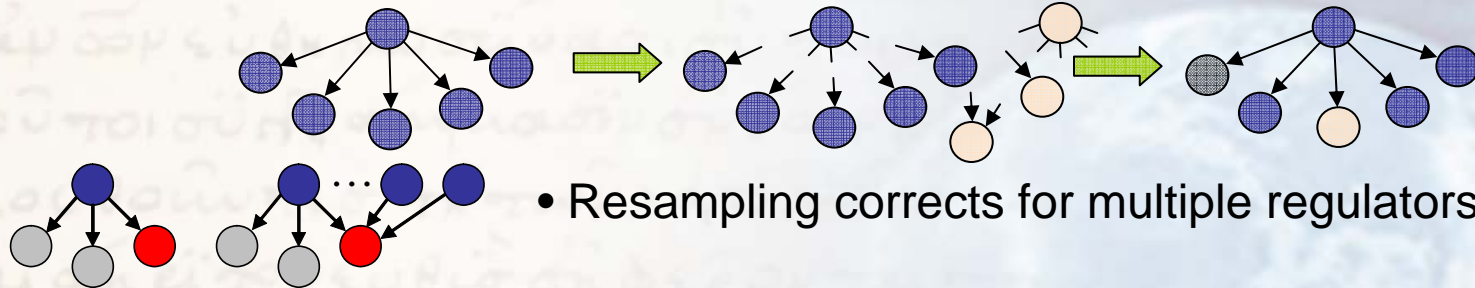
Sampling distribution:

- The structure (regulator – set of targets) is imposed by the network
- *Network resampling* should be used to calculate significance

To get the expected distribution

Break all the links

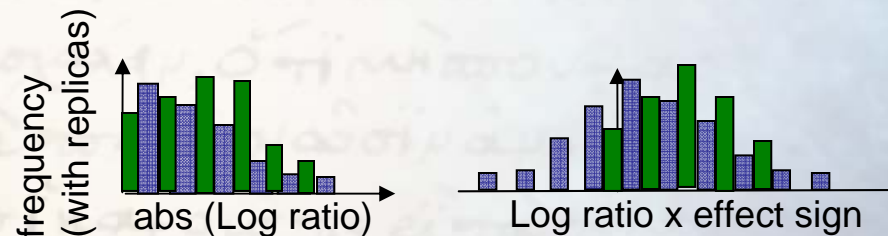
Reconnect randomly



- Resampling corrects for multiple regulators

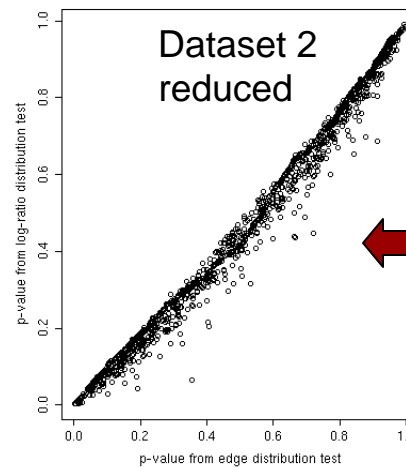
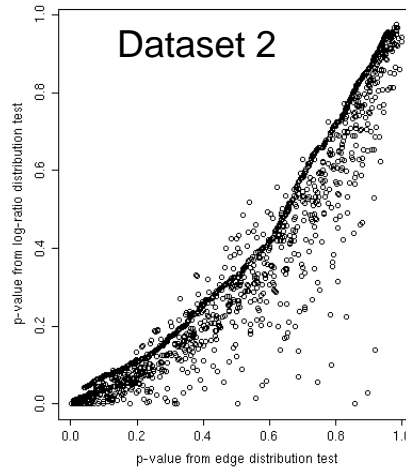
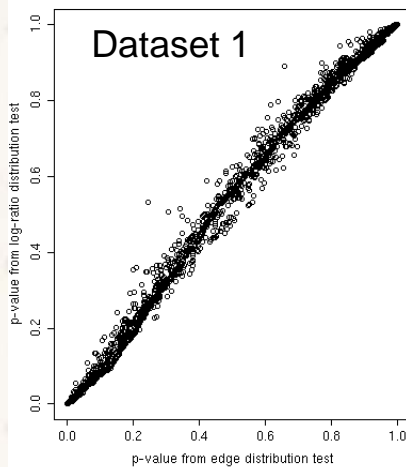
Sampling approximation:

- Replicate each log-ratio ℓ in-degree times to correct for connectivity.
- Take $\text{abs}(\ell)$ for unsigned test; $\ell \times (\text{effect sign})$ for signed test.



Results: Assessments

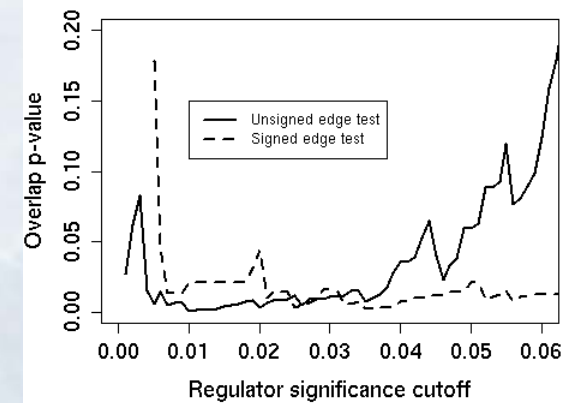
Do replicas matter? Test against distribution of measured log-ratios vs. test against distribution with replicas (scatterplots)



Remove:
In-degree > 12
AND
Log-ratio > 0.9
36 targets total

Overlap with known facts:

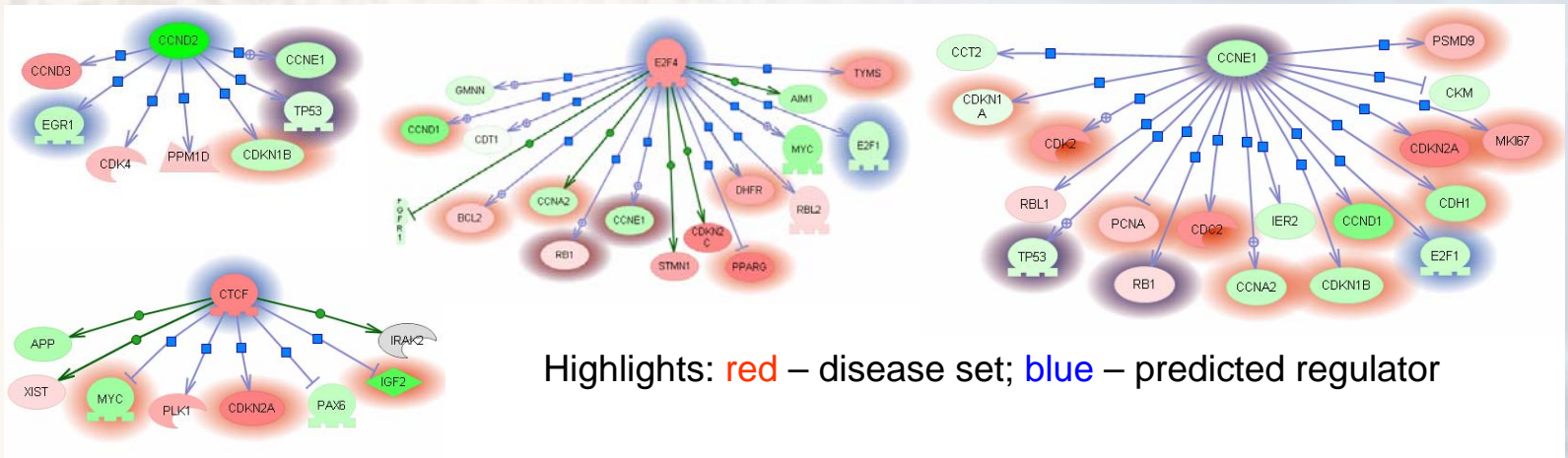
- Protein-disease associations mined from the literature (315 total, 155 regulators in the network)
- Set of predicted regulators is defined by the p-value cutoff $S(p)$
- Evaluate overlap between $S(p)$ (predicted “significant regulators”) and the “disease set”



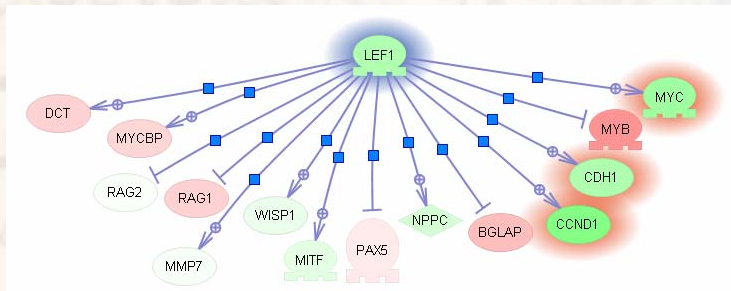
Robustness: Predicted sets of regulators are *stable* against relation removal

Results: Regulators

- Unsigned test (log-ratio absolute values): 21 regulators with $p < 0.01$



- Signed test (log-ratio x effect sign): 7 regulators with $p < 0.01$



- Expression data are reduced to biologically meaningful hypotheses
- Hypotheses are ranked by an objective measure
- Further investigation and validation is required