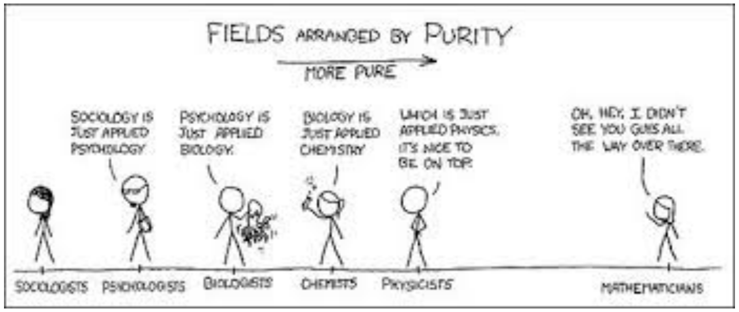
Physicist Meets Biology


FIELDS ARRANGED BY PURITY
MORE PURE
SOCIOLOGY IS JUST APPLIED PSYCHOLOGY
PSYCHOLOGY IS JUST APPLIED BIOLOGY.
BIOLOGY IS JUST APPLIED CHEMISTRY
WHICH IS JUST APPLIED PHYSICS. IT'S NICE TO BE ON TOP.
OH, HEY, I DIDN'T SEE YOU GUYS ALL THE WAY OVER THERE.
SOCIOLOGISTS
PSYCHOLOGISTS
BIOLOGISTS
CHEMISTS
PHYSICISTS
MATHEMATICIANS

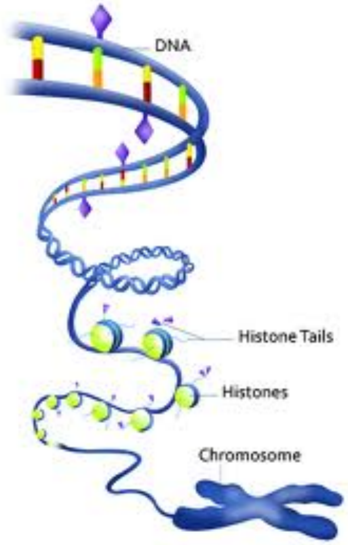
http://xkcd.com

Sara Knaack
Urbana Champaign

4/8/2016

Outline

- Introduction
- Work on the MuCap Experiment
  - Analysis of complex data in a high-throughput computing environment.
- A brief primer to biology and introduction to the work I do now
- Questions



Thanks to Sushmita Roy and Alireza Siahpirani for contributing slides on gene regulatory networks.

My own background

- B.S. from the University of Wisconsin at Madison - Math and Physics
  - Also took an introductory biology curriculum and organic chemistry.

- M.S. from the University of Illinois at Urbana
  - Course work in mathematical methods, quantum mechanics, field theory and statistical mechanics
  - Did beam line simulation work for the g-2 experiment.

- Ph.D. from the University of Illinois at Urbana
  - Work on the MuCap experiment, muon capture on the proton

- **2012 - Present** – Postdoctoral Trainee in Computational Biology at the Wisconsin Institute for Discovery.
  - Research in the regulation of gene expression in the context of evolution and cancer.
  - Funded by the CIBM program – more about that in a moment.
  - Capstone certificate in Bioinformatics – course work in computational biology, statistics and graphical models

More about the Wisconsin Institute for Discovery.

An inter disciplinary research environment, with many themes focusing on biomedical medical research, but also many other initiatives

WARF
Wisconsin Alumni Research Foundation

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

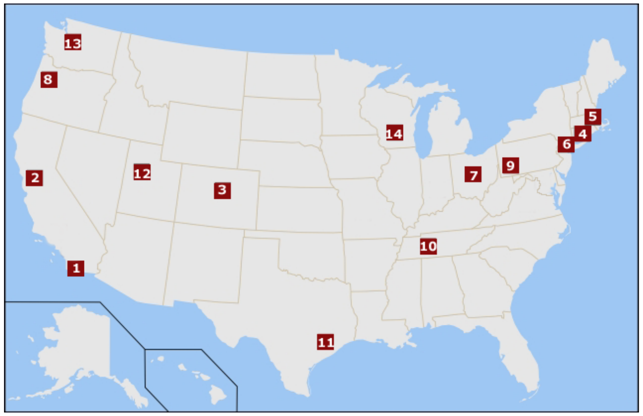
https://discovery.wisc.edu

The Computational Informatics in Biology and Medicine program.

- I am a funded by the CIBM program on the campus of the University of Wisconsin, which is a training program through the National Library of Medicine.
- http://www.cibm.wisc.edu
- We are one of multiple training programs run at campuses across the country, including Stanford and Harvard.
- https://www.nlm.nih.gov/ep/GrantTrainInstitute.html
- Provided a core community of fellow trainees and PI's to interact with
- Activities include the annual NLM conferences, and a weekly seminar during the academic year.

WISCONSIN

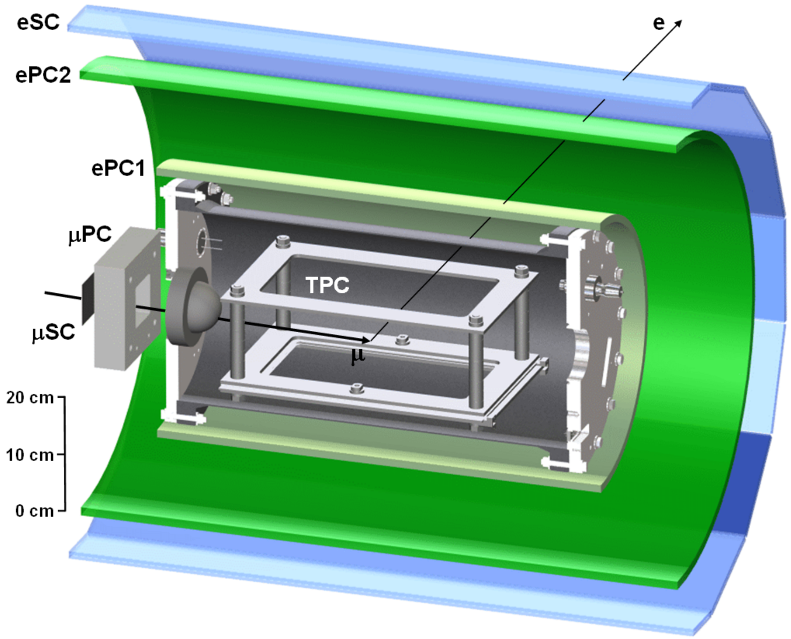
UNIVERSITY OF WISCONSIN–MADISON



Just what kind of training did I come from?

Indulge me three slides on my thesis work…

The MuCap Experiment



- Muon capture on the proton (MuCap)
- Grew out of the study of hydrogen fusion
  - at the level of fundamental particle interactions.
  - *Physics motivation: quark-gluon substructure of the proton,* g_p
- My work was to measure the rate of molecular state formation.

Description of the time distribution

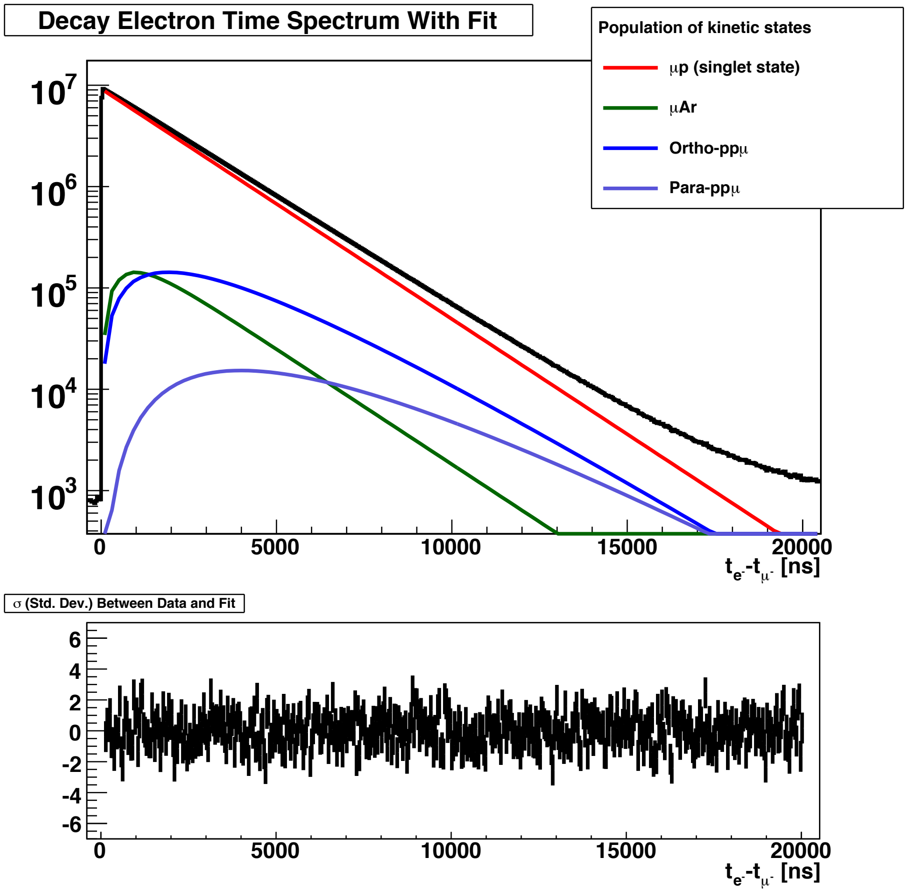
$$\begin{aligned}
n'_{\mu p}(t) &= -(\lambda_\mu + \Lambda_{pp\mu} + \Lambda_{\mathrm{pAr}} + \Lambda_{\mathrm{S}} + \Lambda_{\mathrm{pf}})n_{\mu p}(t), \\
n'_{\mu \mathrm{Ar}}(t) &= \Lambda_{\mathrm{pAr}} n_{\mu p}(t) - (h\lambda_\mu + \Lambda_{\mathrm{Ar}})n_{\mu \mathrm{Ar}}(t), \\
n'_{\mathrm{Ortho}}(t) &= \Lambda_{pp\mu} n_{\mu p}(t) - (\lambda_\mu + \lambda_{\mathrm{op}} + \Lambda_{\mathrm{O}})n_{\mathrm{Ortho}}(t), \\
n'_{\mathrm{Para}}(t) &= \Lambda_{\mathrm{pf}} n_{\mu p}(t) + \lambda_{\mathrm{op}} n_{\mathrm{Ortho}}(t) - (\lambda_\mu + \Lambda_{\mathrm{P}})n_{\mathrm{Para}}(t).
\end{aligned}$$

$$\begin{aligned}
n_{\mu p}(t=0) &= 1 - f \\
\text{and } n_{\mu \mathrm{Ar}}(t=0) &= f, \\
\text{where } n_{\mathrm{Ortho}}(t=0) &= 0 \\
\text{and } n_{\mathrm{Para}}(t=0) &= 0.
\end{aligned}$$

$$n_e^{\mathrm{Obs.,Ar}}(t) = \lambda_\mu \left(n_{\mu p}(t) + n_{\mathrm{Para}}(t) + n_{\mathrm{Ortho}}(t)\right) + e_{\mathrm{Ar}} h \lambda_\mu n_{\mu \mathrm{Ar}}(t).$$

- Differential equations, initial conditions, full time distribution.
- Atomic physics parameters f, h, and e_{Ar}
  - relative contribution of μAr state decays
- The hydrogen kinetic rates, λ_μ, Λ_S, λ_{op}, Λ_{pf}, Λ_O, and Λ_P
  - Directly affect the time distribution of events
- The fit function is A n_e(t)+B

Fit to the decay electron time distribution



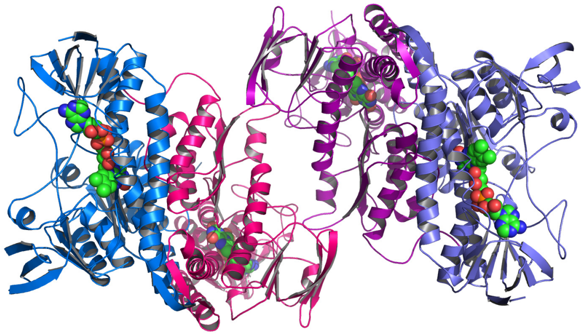
- 4.25 x 10^8 events
- Basic fit results
  - $\Lambda_{pp\mu}$=2.208(65) x 10^4 s^{-1}
  - Λ_{pAr}=4.529(15) x 10^4 s^{-1}
  - Λ_{Ar}=1.302(14) x 10^6 s^{-1}
- *χ2/Ndf*=0.983(64)

Carry-overs in Computational Biology

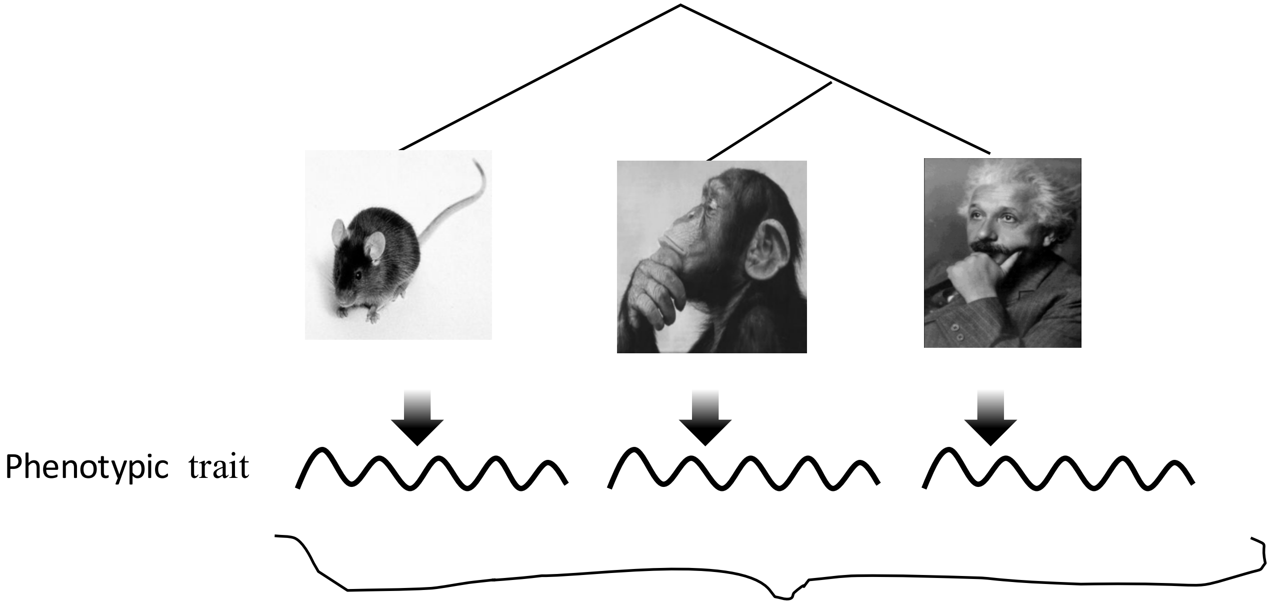
- Analysis of complex data from a high-throughput computational environment.
  - C++ code development
  - Statistical analysis

- Integrative study of processes on multiple spatial and temporal scales.

- The extraction and interpretation of results from complex systems.

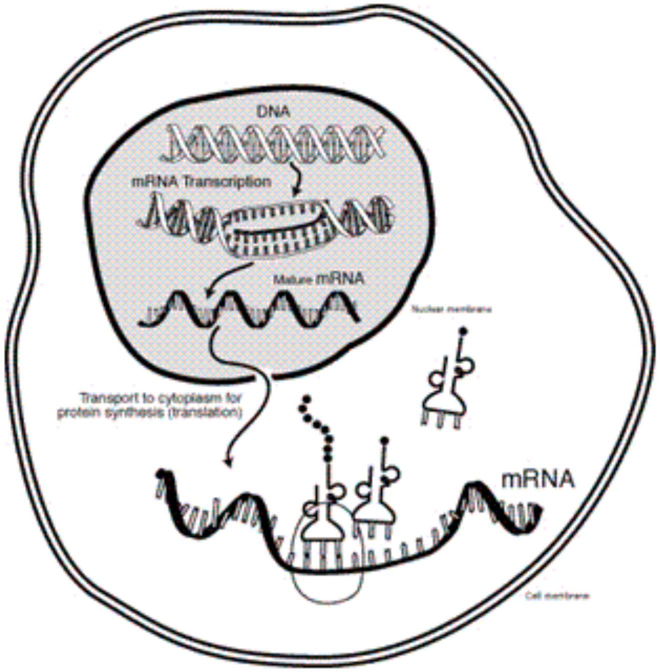


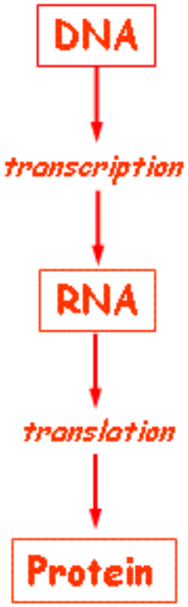
What controls phenotypic diversity?



Changes in gene expression & regulation play a major role in diversifying phenotype.

How are gene expressed? The central dogma.





DNA

transcription

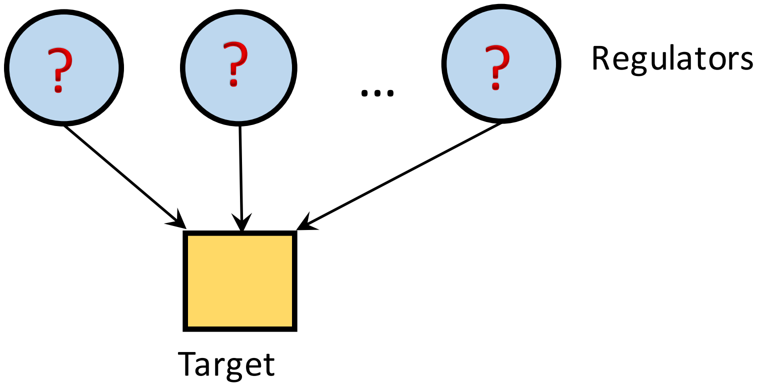
RNA

translation

Protein

http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/central_dogma.html

Who regulates whom?



Regulation of Gene Expression is Multilayered

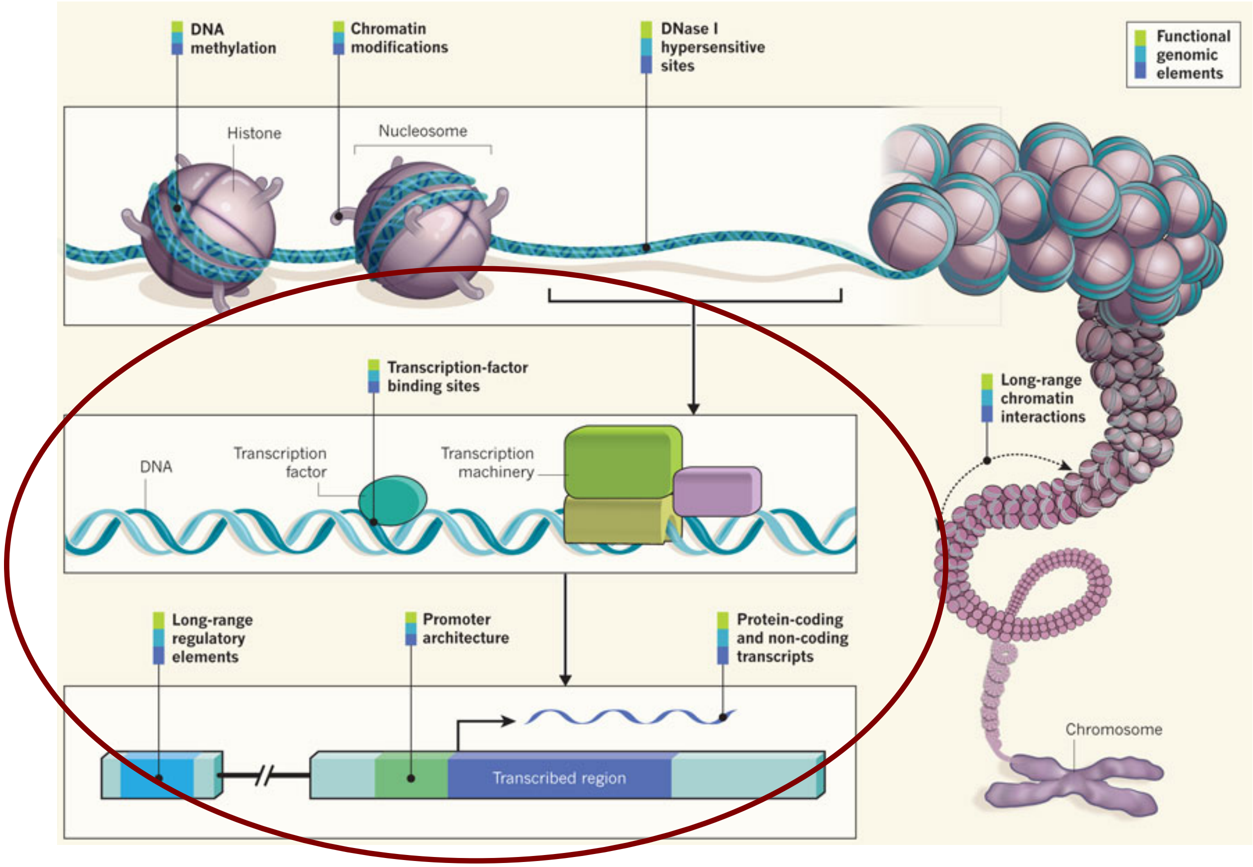
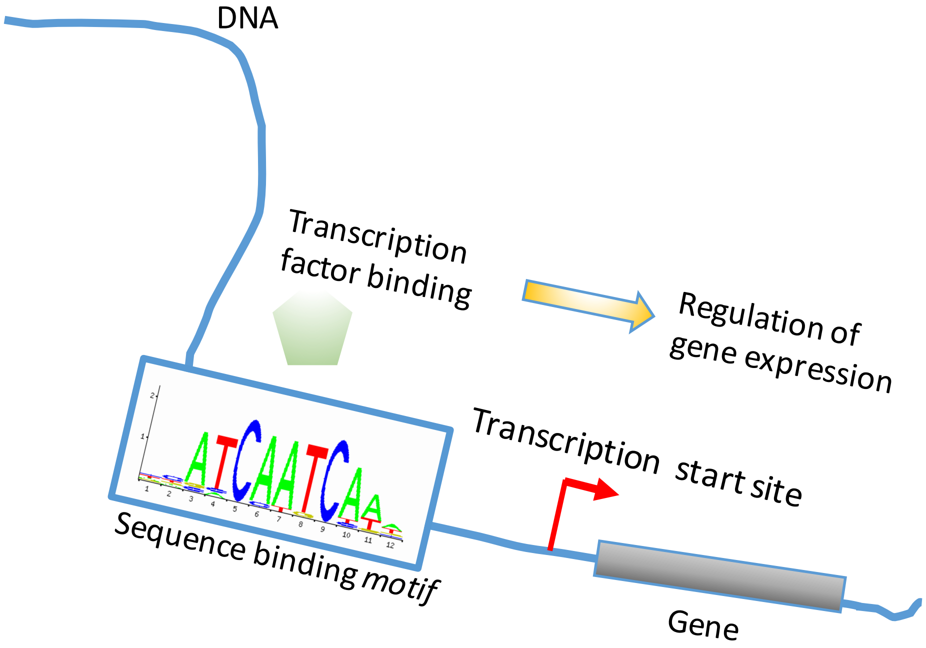
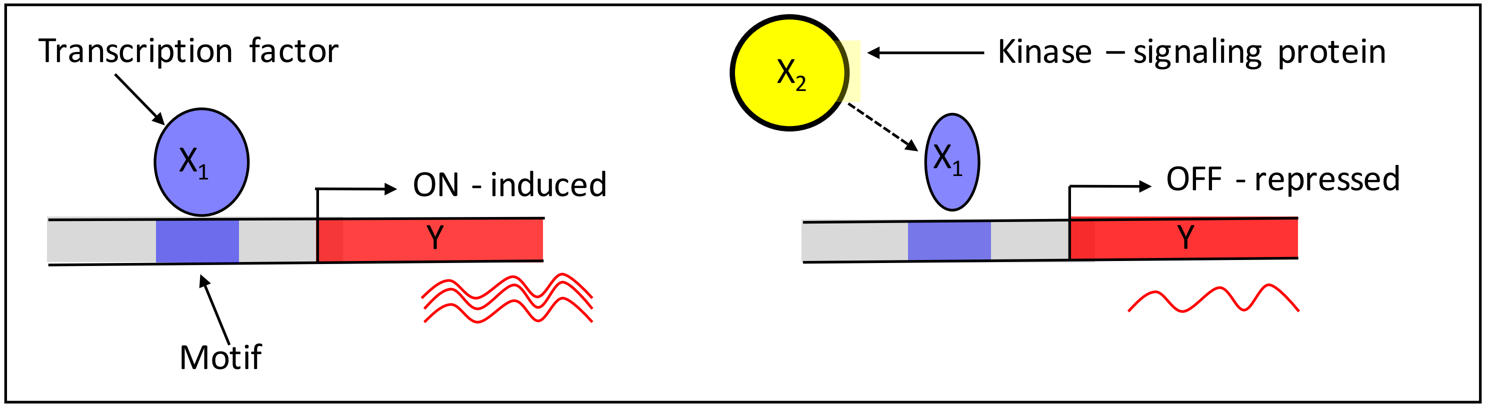

DNA methylation
Chromatin modifications
DNase I hypersensitive sites
Functional genomic elements
Histone
Nucleosome
Transcription-factor binding sites
Long-range chromatin interactions
DNA
Transcription factor
Transcription machinery
Long-range regulatory elements
Promoter architecture
Protein-coding and non-coding transcripts
Transcribed region
Chromosome

Image: ENCODE Consortium

One mode of gene regulation: transcription factor binding at cis-regulatory elements in the genome.



Modeling a regulatory network



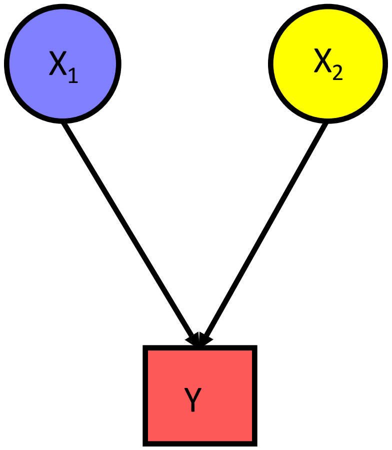
Transcription factor

Motif

ON - induced

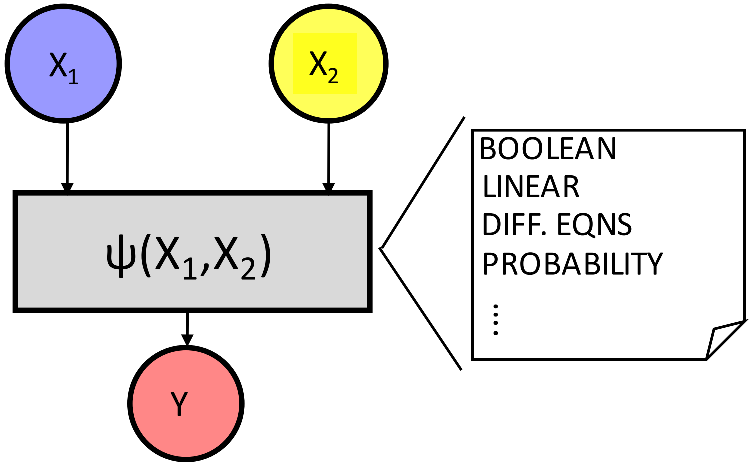
Kinase – signaling protein

OFF - repressed



Structure

Who are the regulators?



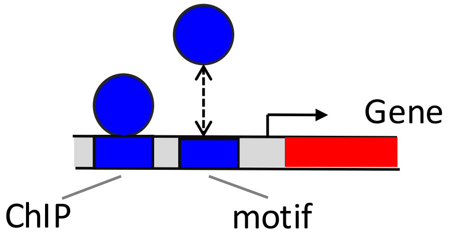
BOOLEAN
LINEAR
DIFF. EQNS
PROBABILITY
⋮

Function

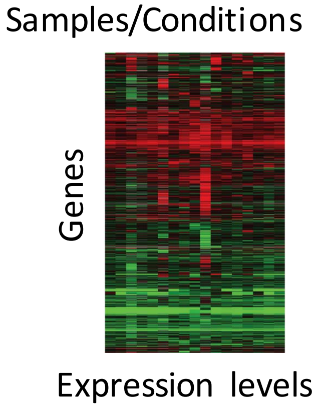
How do they determine expression levels?

Types of data for reconstructing networks

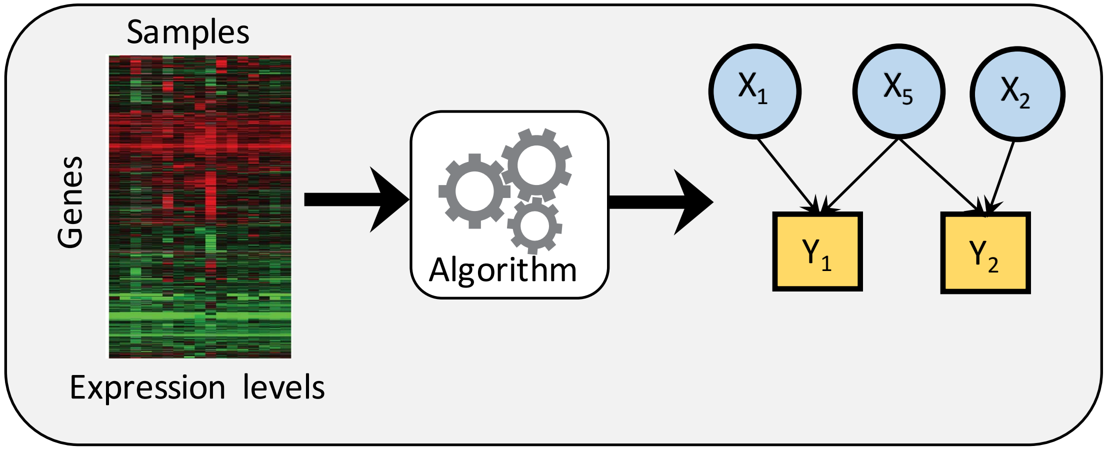
- Physical
  - ChIP-chip and ChIP-seq
  - Sequence-specific binding - motifs
  - Regulator centric



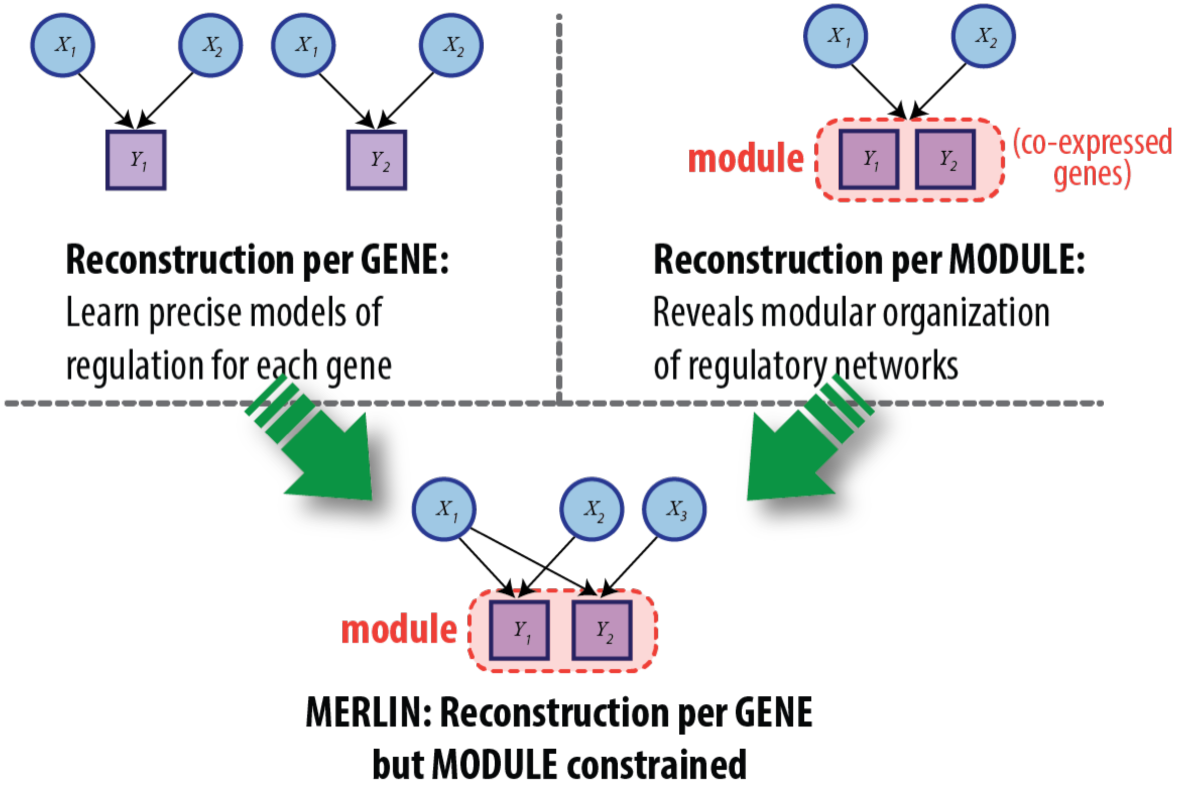
- Functional
  - Gene expression
  - Measure dynamic information
  - Can potentially recover genome-wide regulatory networks



Expression-based network inference



MERLIN: A network reconstruction method to predict regulators of genes and modules



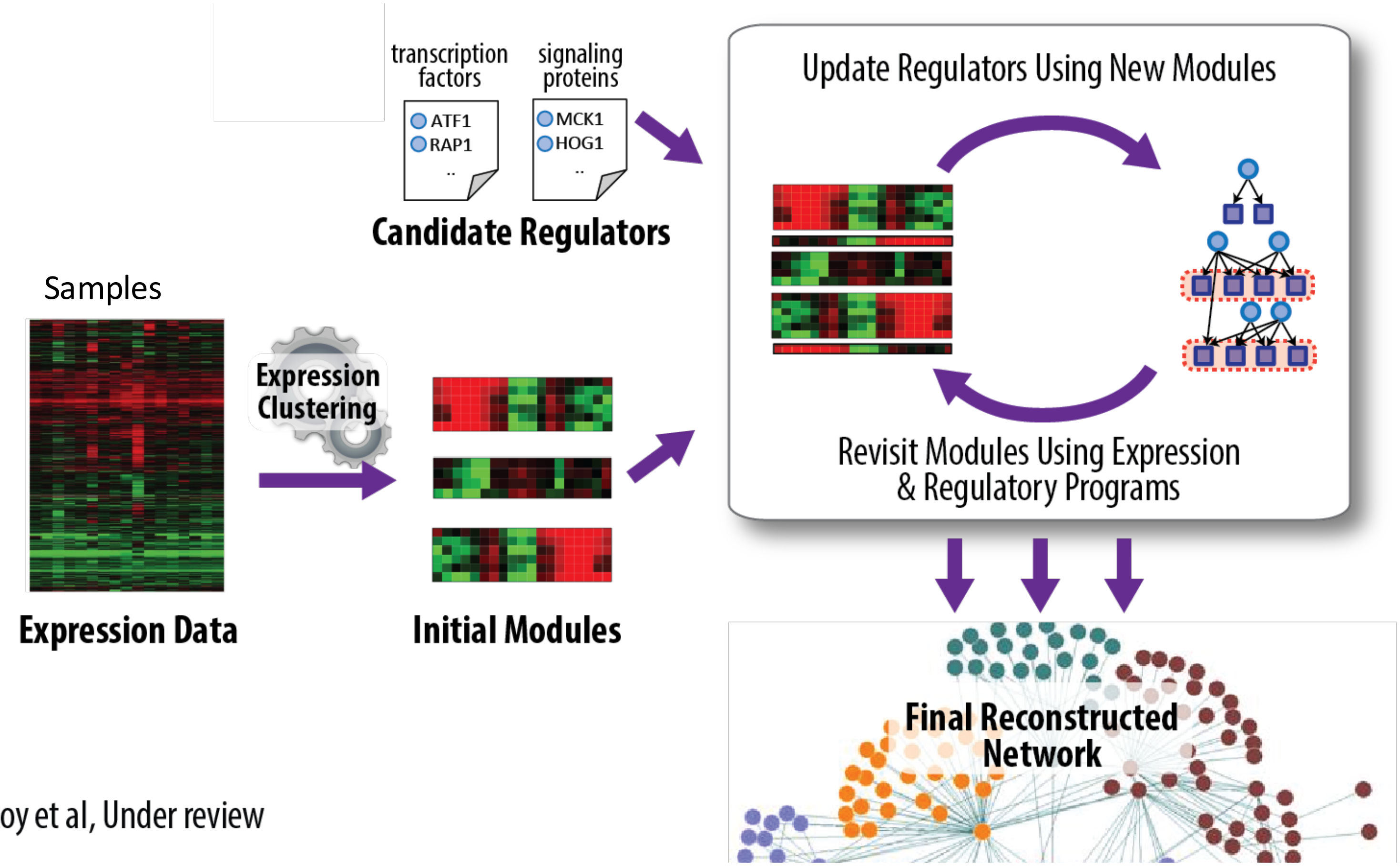
Reconstruction per GENE:
Learn precise models of regulation for each gene

Reconstruction per MODULE:
Reveals modular organization of regulatory networks

MERLIN: Reconstruction per GENE but MODULE constrained

Roy et al, 2013 Plos comp bio

MERLIN learning algorithm

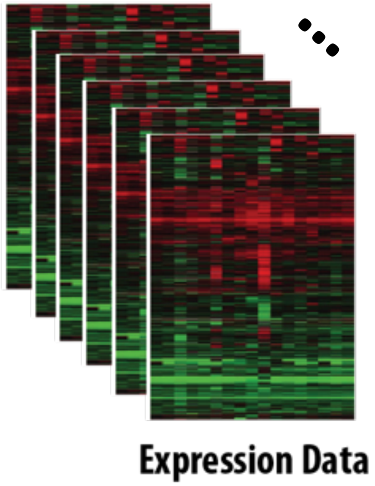


Roy et al, Under review

Data from The Cancer Genome Atlas

- Cancer Genome Atlas Research Network:
  - Weinstein et al. *Nat Genet*. 2013

- Microarray gene expression data for 6 cancers:
  (1) Breast (BRCA)
  (2) Colon (COAD)
  (3) Kidney ma (KIRC)
  (4) Lung (LUSC)
  (5) Ovarian (OV)
  (6) Uterine (UCEC).

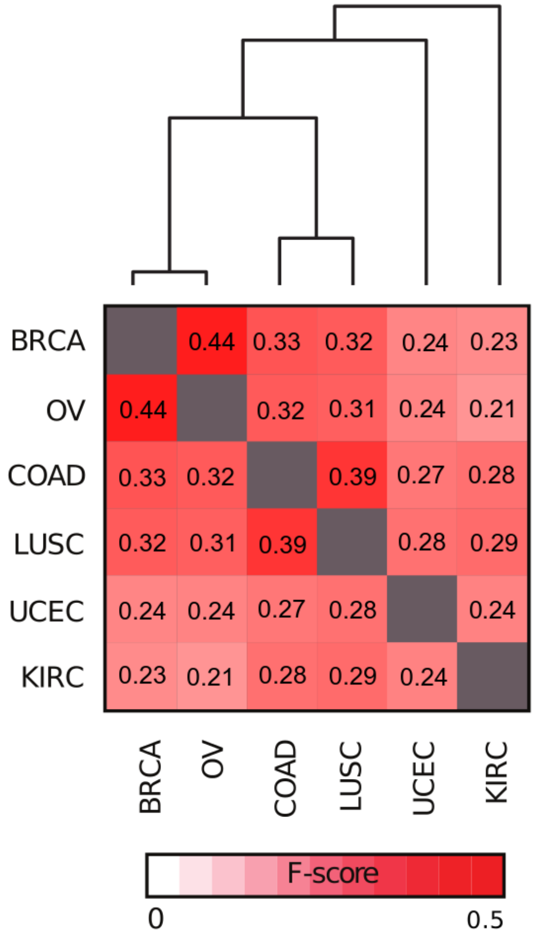


- 54 (UCEC) to 598 (OV) patient samples

- 8499 genes were selected
  - Variation in expression across patient samples in each data set
  - Any gene annotated in curated NCI cancer pathways

- 1050 were known transcription factors TFs and kinases – regulators

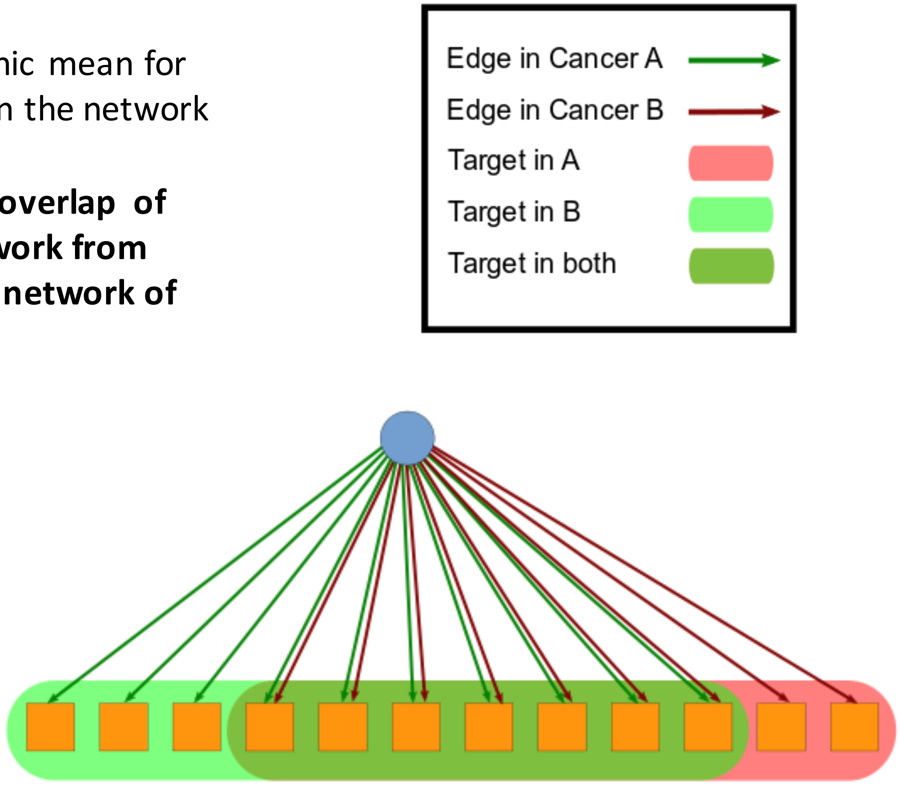
The inferred networks are distinctly different



	BRCA	OV	COAD	LUSC	UCEC	KIRC
BRCA		0.44	0.33	0.32	0.24	0.23
OV	0.44		0.32	0.31	0.24	0.21
COAD	0.33	0.32		0.39	0.27	0.28
LUSC	0.32	0.31	0.39		0.28	0.29
UCEC	0.24	0.24	0.27	0.28		0.24
KIRC	0.23	0.21	0.28	0.29	0.24	

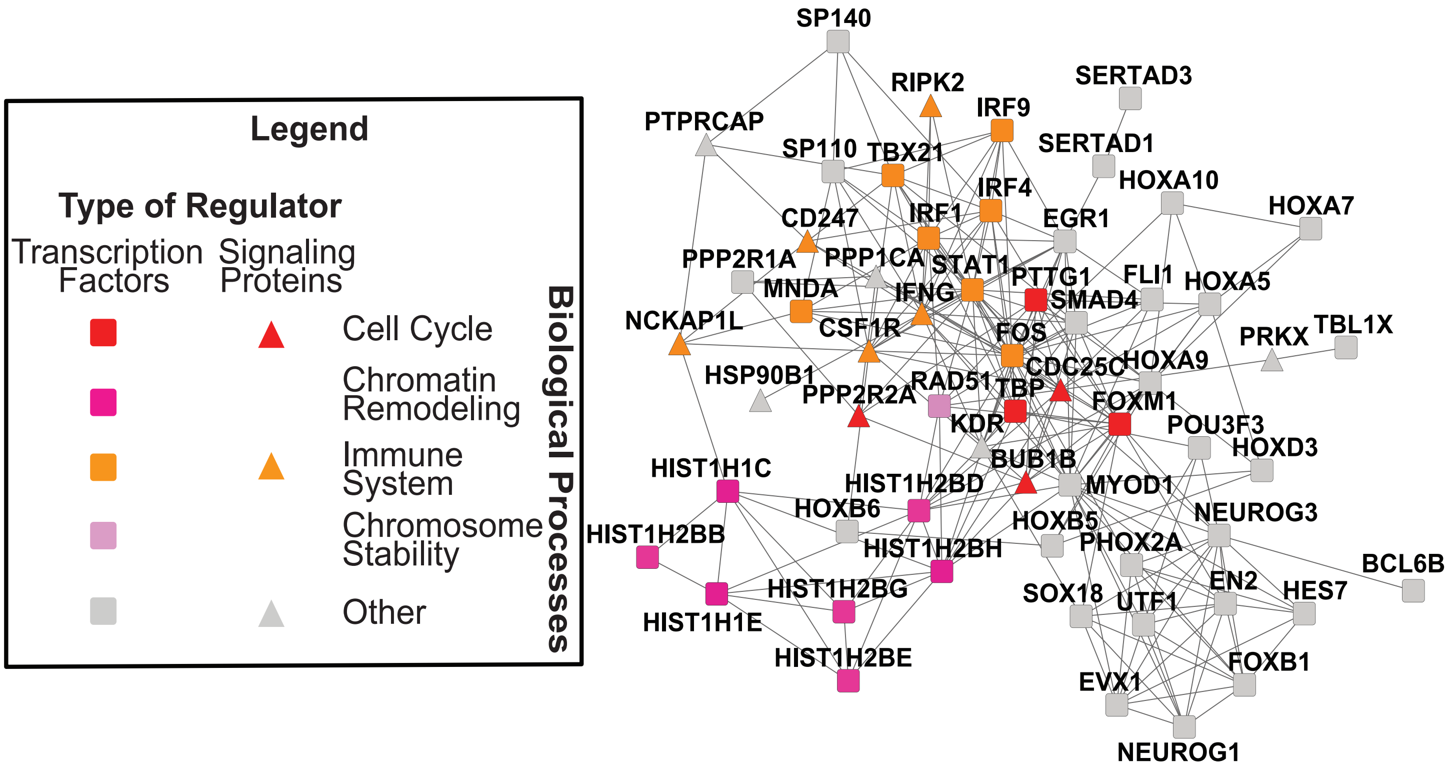
F-score or harmonic mean for the set of edges in the network

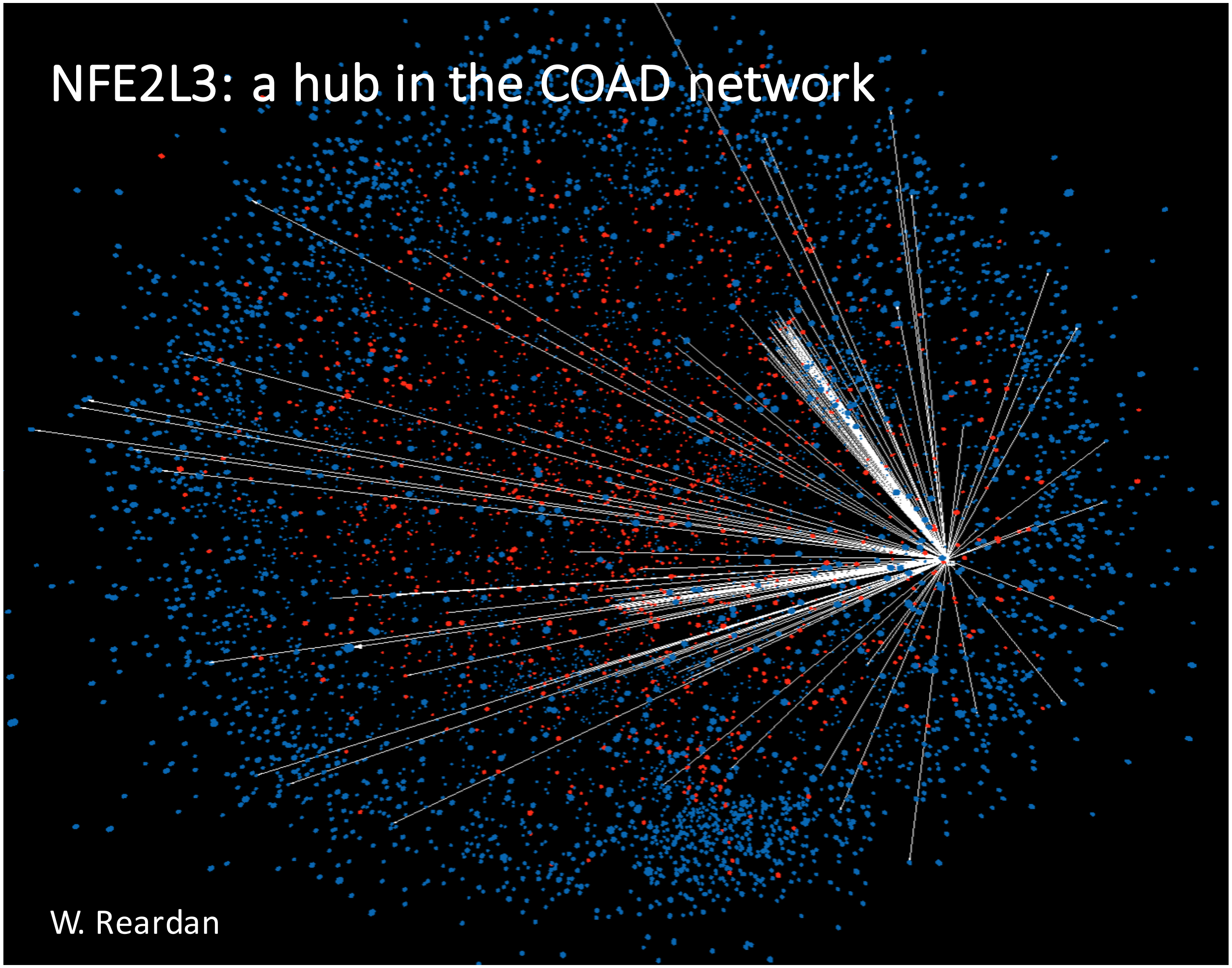
How great is the overlap of edges in the network from cancer A and the network of cancer B?



Common regulators are associated with chromatin, cell cycle and immune response

Consists of edges between 75 regulatory proteins and 156 target genes

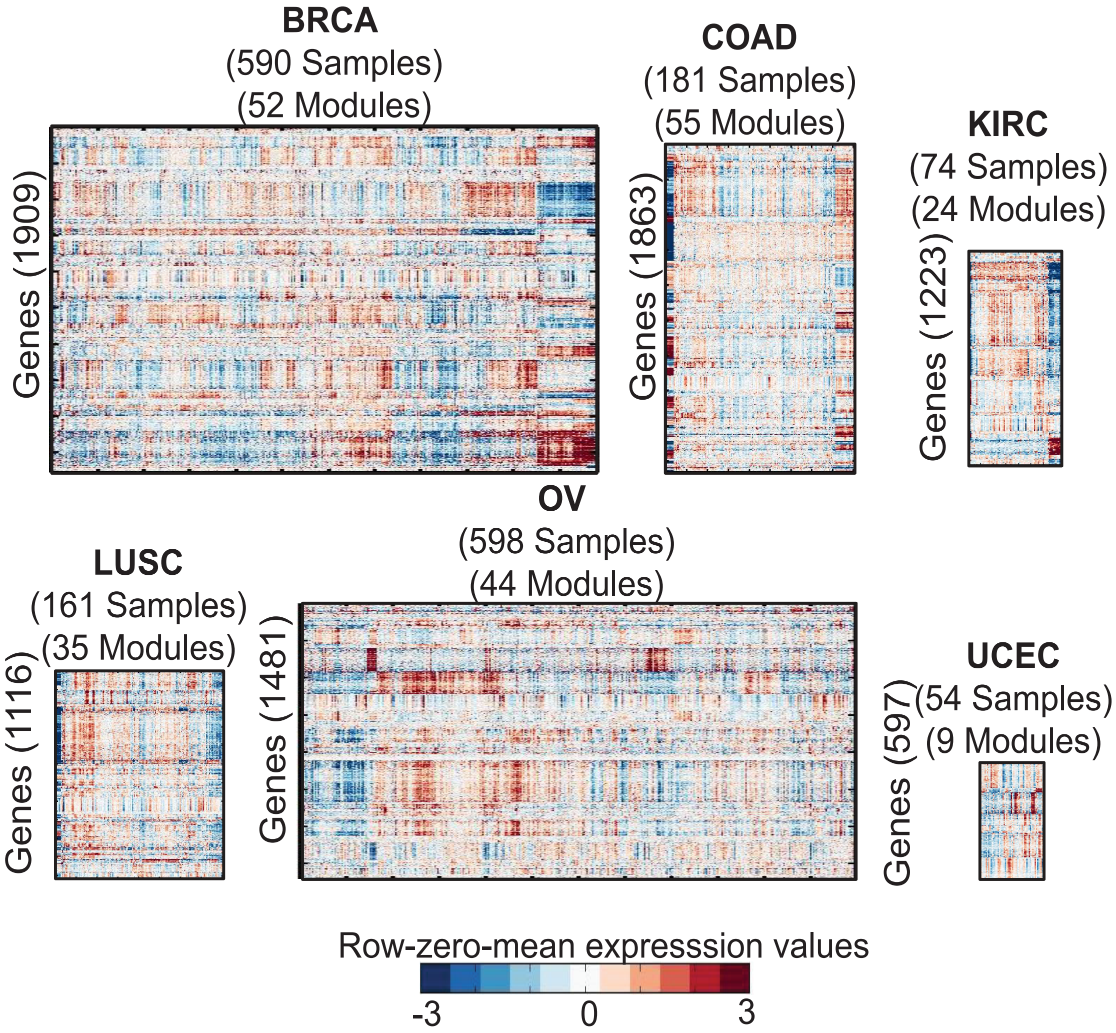




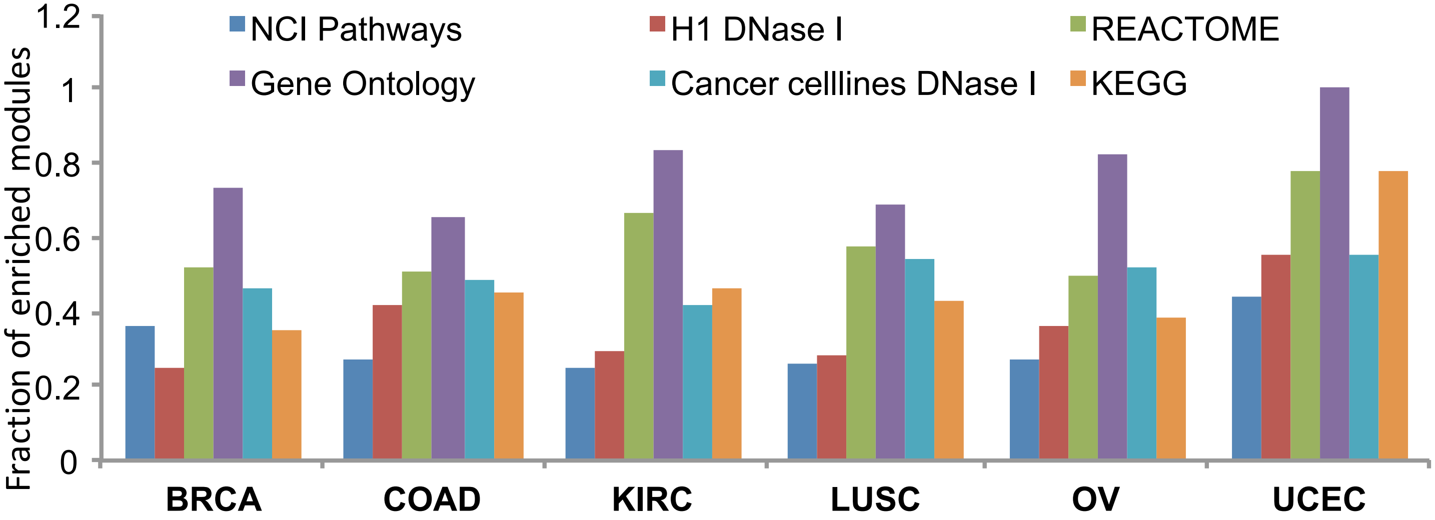
NFE2L3: a hub in the COAD network

W. Reardan

Expression in consensus modules



Validating that our modules are biologically coherent

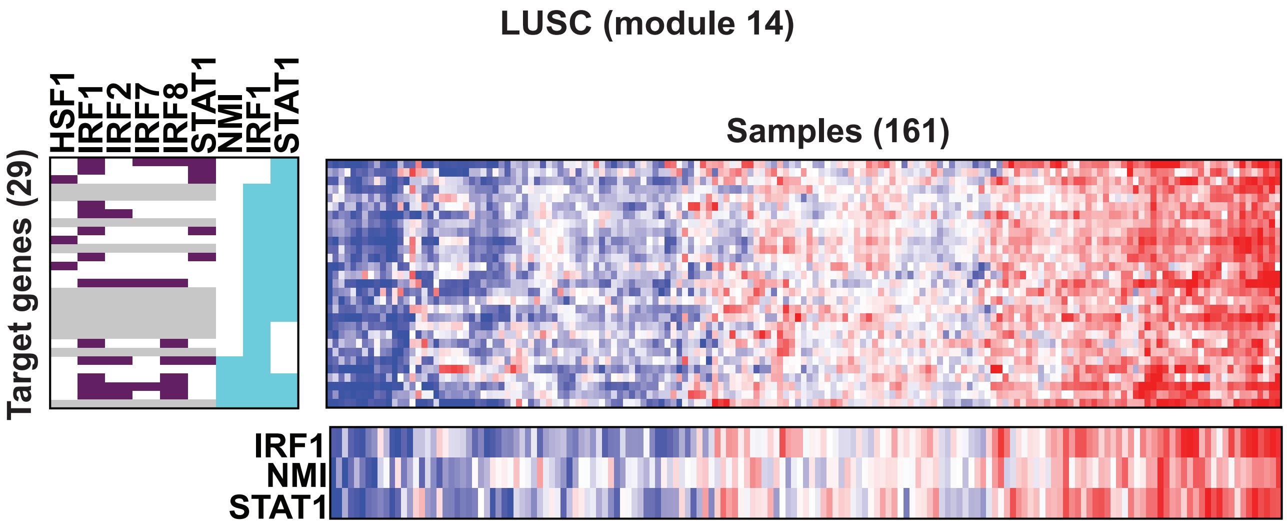


Here we have several sets of annotated genes, and each set provides us lists of gens with a certain biological significance.

We look to find if our modules are significantly enriched in genes from any of these annotated sets using a Hypergeometric test

Above we count the fraction of modules that have an enrichment with 0.05 significance in the results from each data set.

An example of a module



🟦 Targets of regulators predicted by MERLIN

🟦 Targets of regulators predicted by MERLIN

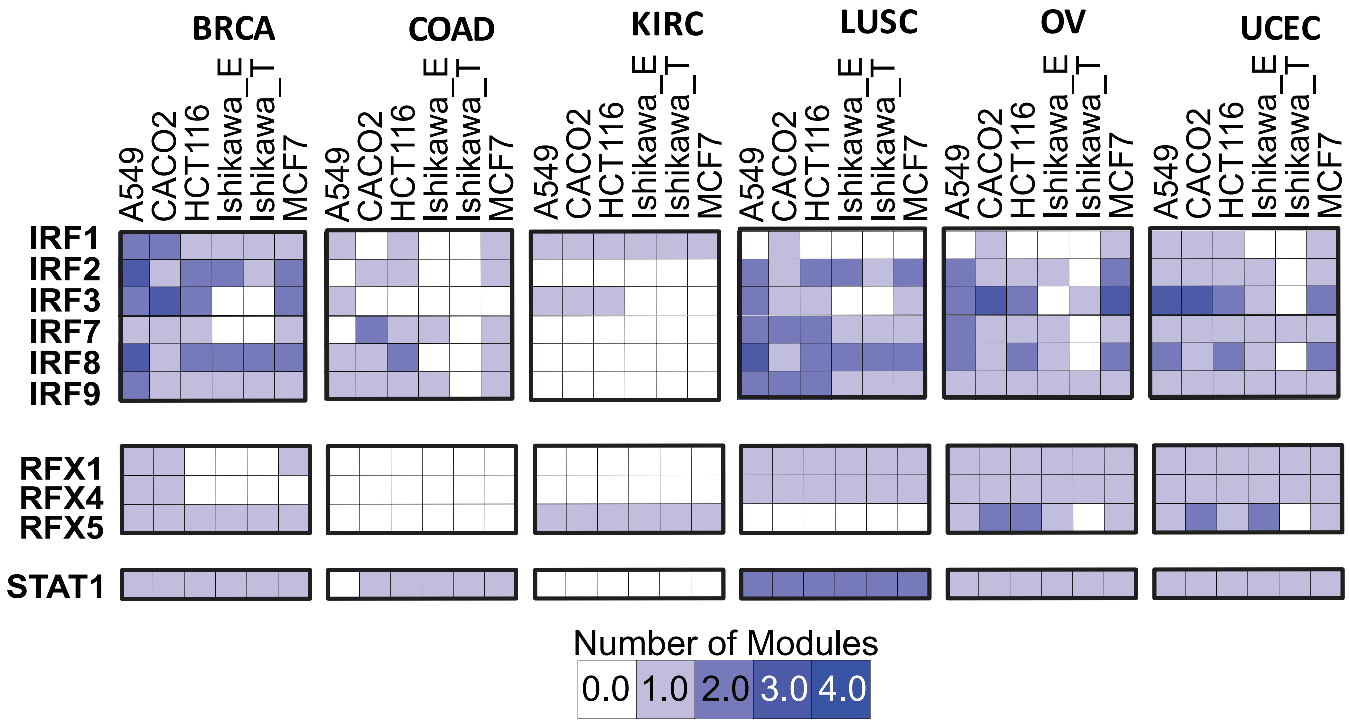
🟪 Targets of regulators f MSigDB motifs

⬜ Not annotated

Row-zero-meaned expression values

-2.0	-1.6	-1.2	-0.8	-0.4	0.0	0.4	0.8	1.2	1.6	2.0

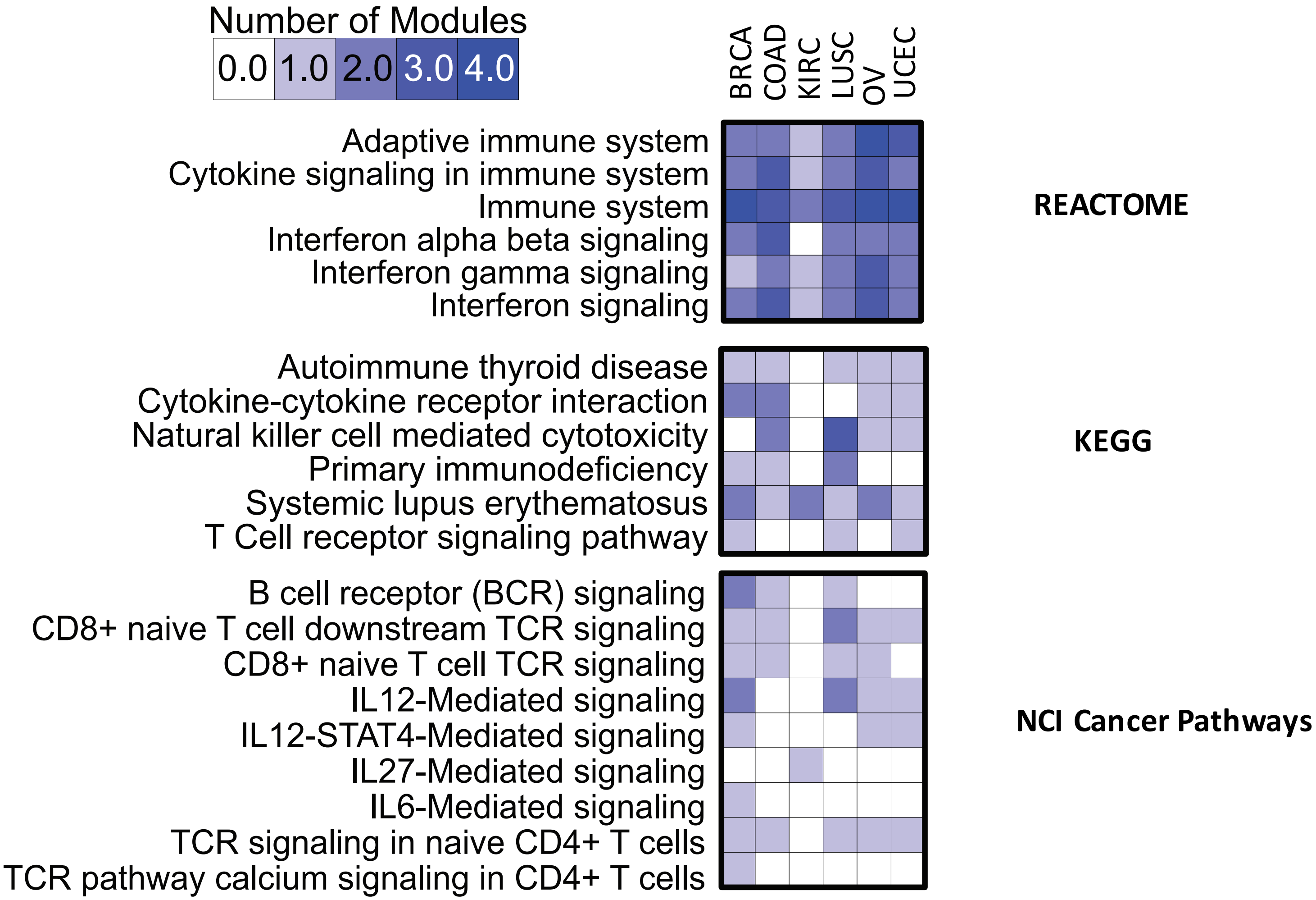
Enrichments for motifs of immune system regulators



- interferon regulatory factor (IRF) family – all cancers
- regulatory factor X (RFX) – five cancers
- signal transducer and activator of transcription (STAT1) – five cancers

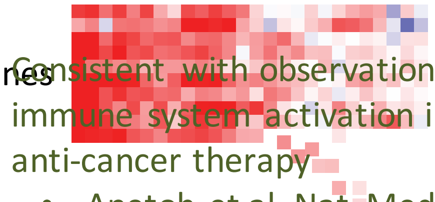
- All regulators of the immune system

Immune system function is over-represented



Is the immune system induced or repressed?

Modules associated with the immune system:

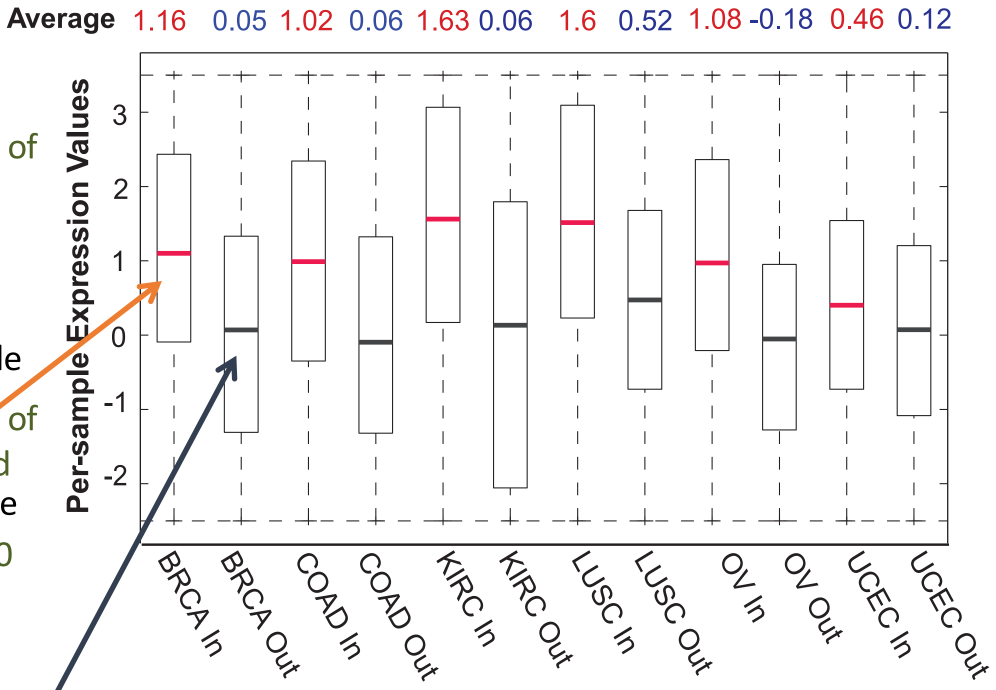


- Consistent with observations of immune system activation in anti-cancer therapy
  - Apetoh et al. Nat. Med. 2007
- Known cross-talk mechanism of activation between STAT3 and interleukin-6 signaling
  - Lee et al. Nat. Med. 2010

Per-gene, per-sample expression values

Modules not associated with the immune system:

Samples
Genes



Summary

- We have introduced stability-selection into our MERLIN-based approach to infer regulatory networks across different conditions.
- Our approach builds on the idea that both module- and network-based characterization of transcriptional programs are important.
- Our methods can be extended with additional data types.

Knaack SA, Siahpirani AF, Roy S. A pan-cancer modular regulatory network analysis to identify common and cancer-specific network components. *Cancer Inform*. 2014 Oct 28;13(Suppl. 5):69-84.

doi: 10.4137/CIN.S14058
PMID: 25374456 [PubMed]

Conclusions

Work with beautiful complex systems, rich for exploration and discovery.

Computational biology is a fast paced field, with many emerging technologies and methods.

The mindset towards measurements is different than what you are used to from physics.

It’s a field that will have an increasing impact on medicine and human health as we learn more.

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Acknowledgements

WISCONSIN
INSTITUTE FOR DISCOVERY
AT THE UNIVERSITY OF WISCONSIN-MADISON

- My mentor, Sushmita Roy
- Members of the Roy Group
- Funding support from
  - National Science Foundation, CAREER grant (SR)
  - National Library of Medicine training grant NLM5T15LM007359 (SK)
- You!

Thanks to Sushmita Roy and Alireza Siahpirani for contributing slides on gene regulatory networks.