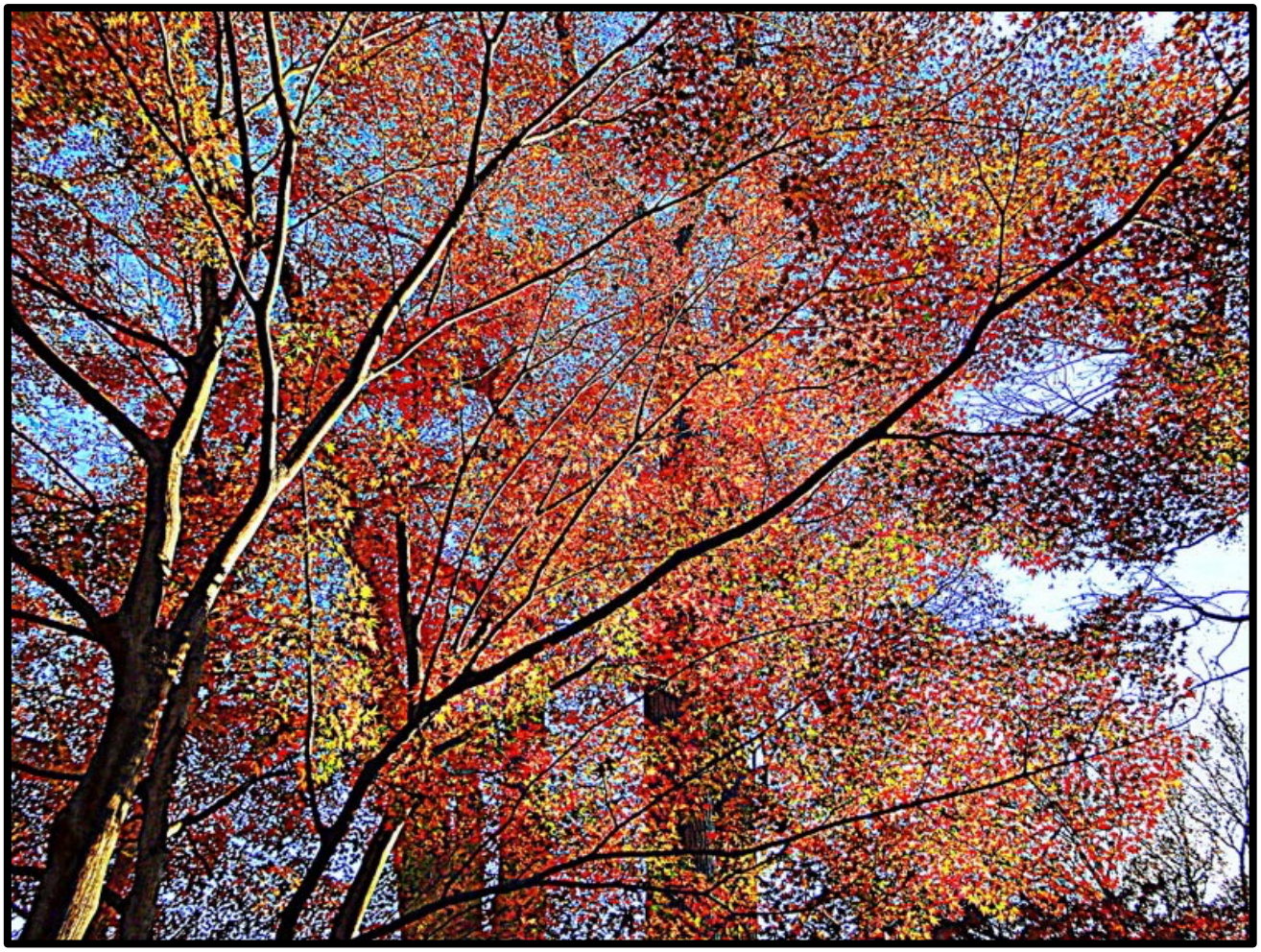
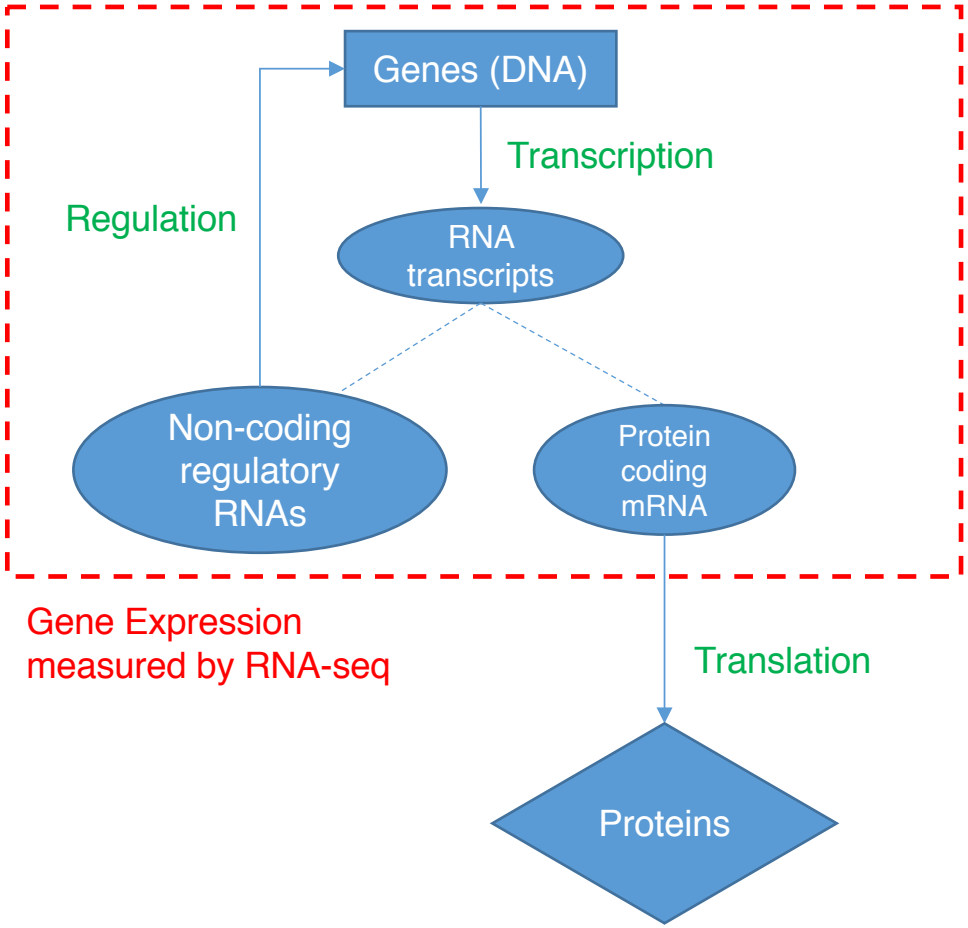
Transcriptome Mining:

Tackling core issues related to gene regulation
& also analyzing the "data exhaust" associated with this activity

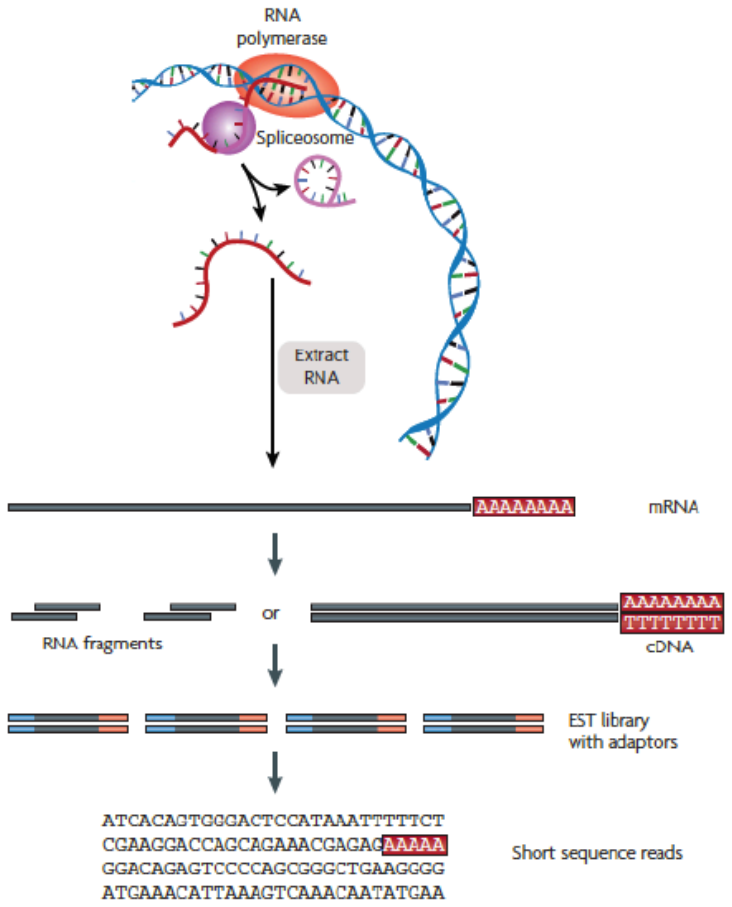


Mark Gerstein, Yale. Slides freely downloadable from **Lectures.GersteinLab.org** & “tweetable” (via **@markgerstein**). See last slide for more info.

Transcriptome = Gene Activity of All Genes in the Genome, usually quantified by RNA-seq



Gene Expression measured by RNA-seq



[*NATURE 459: 927; NAT. REV. GEN. 10: 57*]

Expression of genes is quantified by transcription:
RNA-Seq measures mRNA transcript amounts

RNA-Seq Overview

Successive steps of Data Reduction

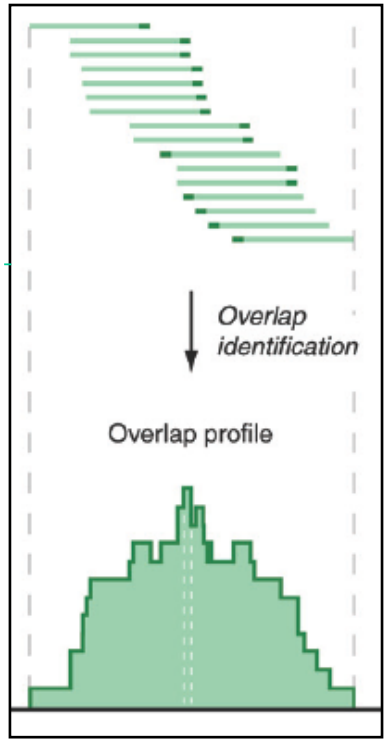
Fastq sequence files
~5-10 GB

```
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
TACCGATCGAGTCGACTGTAAACGTAGGCATA
ATTCTGACTGGTGTCATGCTGATGTACTTAAA
```

Index-building + Alignment to reference genome

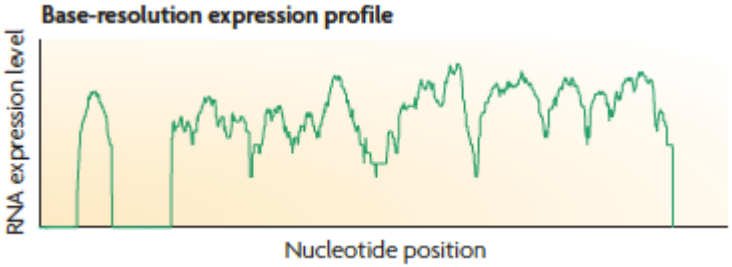
BAM files
~1-2-fold reduction

Conversion to signal track by overlapping reads



Reads => Signal

BigWig files
~25-fold reduction



Quantitative information from RNA-seq signal:
average signals at exon level (RPKMs)

Mapping to genes

Gene/Transcript expression matrix
~20-fold reduction

[*NAT. REV.* 10: 57; *PLOS CB* 4:e1000158; *PNAS* 4:107: 5254]

Activity Patterns

- RNA Seq. gives rise to activity patterns of genes & regions in the genome

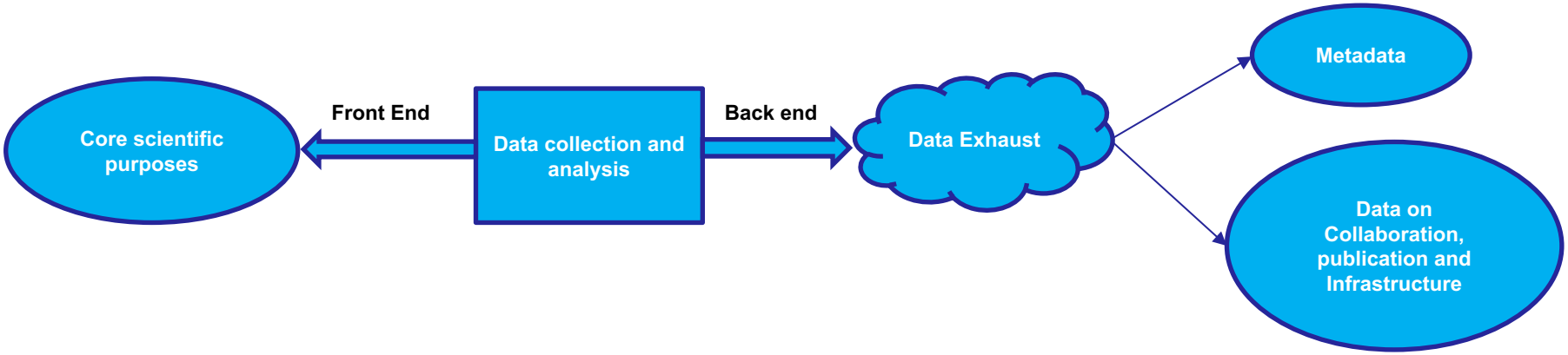
Some Core Science Qs Addressed by RNA-seq

- Gene activity as a function of:
  - **Developmental** stage: basic patterns of co-active genes across development
  - **Cell-type** & Tissue: relationship to specialized functions
  - **Evolutionary** relationships: behavior preserved across a wide range of organisms; patterns in model organisms in relation to those in humans
  - **Disease** phenotypes: disruption of patterns in disease

- Our overarching Qs:
  Are there core, ancient patterns of gene expression?
  Are they associated with development?
  Are they disrupted by disease?

Studying large-scale transcriptome data also produces

Data Exhaust



- Data Exhaust = Exploitable byproducts of big data collection and analysis
- Creative use of Data is key to Data Science !

[PHOTO: RELAXNEWS; from http://www.lapresse.ca]

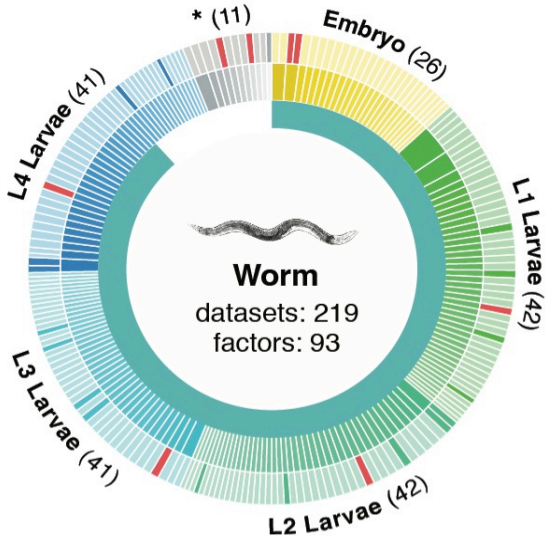
Transcriptome Mining: Tackling core issues related to gene regulation & also analyzing the "data exhaust" associated with this activity

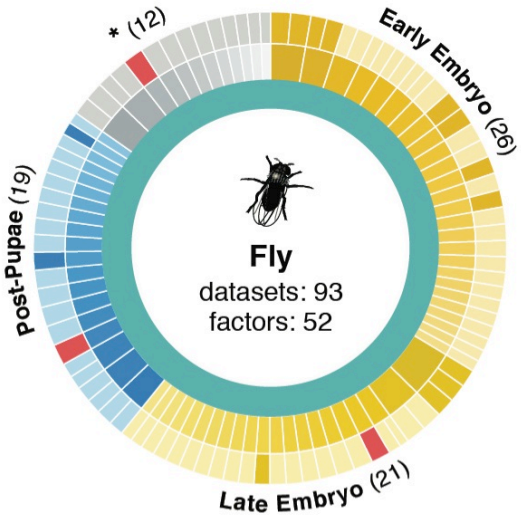
- *[Core-1]* **Expression Clustering**, Cross-species
  - Comparative ENCODE – Lots of worm-fly-human matched data & developmental timecourses
  - Optimization gives 16 conserved co-expression modules, 12 w/ hourglass
- *[Core-2]* **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers; Decoupling expression changes into those from conserved vs species-specific genes
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- *[Core-3]* **Logic Gates** Modeling
  - Preponderance of OR gates in cancer v. cell-cycle (esp. for MYC)
- *[Exhaust-1]* **Genomic Privacy** & RNA-seq
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
  - Using file formats to remove obvious variants
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- *[Exhaust-2]* **Publication Patterns** from data producing consortia
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

Transcriptome Mining: Tackling core issues related to gene regulation & also analyzing the "data exhaust" associated with this activity

- *[Core-1]* **Expression Clustering**, Cross-species
  - Comparative ENCODE – Lots of worm-fly-human matched data & developmental timecourses
  - Optimization gives 16 conserved co-expression modules, 12 w/ hourglass
- *[Core-2]* **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers; Decoupling expression changes into those from conserved vs species-specific genes
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- *[Core-3]* **Logic Gates** Modeling
  - Preponderance of OR gates in cancer v. cell-cycle (esp. for MYC)
- *[Exhaust-1]* **Genomic Privacy & RNA-seq**
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
  - Using file formats to remove obvious variants
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- *[Exhaust-2]* **Publication Patterns** from data producing consortia
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

ENCODE Time-course gene expression data of worm & fly development + human conditions





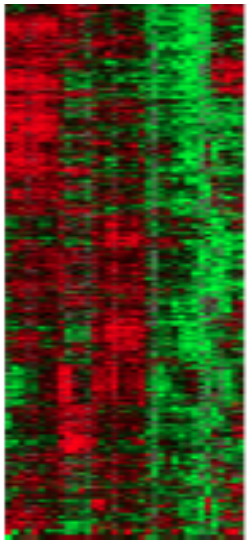
Comparative ENCODE Functional Genomics Resource (EncodeProject.org/comparative)

Organism	Major developmental stages
worm (*C. elegans*)	33 stages: 0, 0.5, 1, …, 12 hours, L1, L2, L3, L4, …, Young Adults, Adults
fly (*D. mel.*)	30 stages: 0, 2, 4, 6, 8,…, 20, 22 hours, L1-L4, Pupaes, Adults

- Broad sampling of conditions across transcriptomes for human, worm & fly
  - embryo & ES cells
  - developmental time course (worm-fly)
- In total: ~3000 datasets (~130B reads)

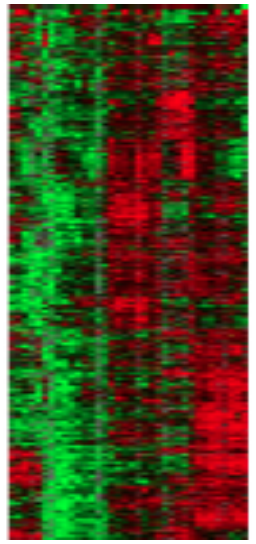
[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Expression clustering: revisiting an ancient problem



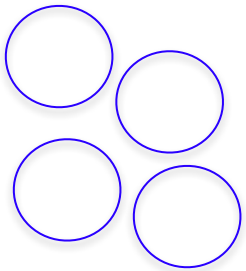
Species A

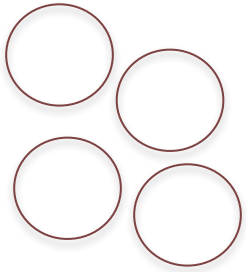
Species B



Clustering algorithm

Clustering algorithm



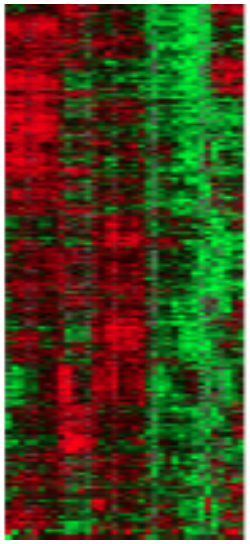


co-expressed genes responsible for the same function in a species

two independent sets of modules

Eisen MB et al. PNAS 1998
Langfelder P et al. BMC Bioinfo. 2008
Tamayo P et al. PNAS 1999
Kluger Y et al. Genome Res. 2003

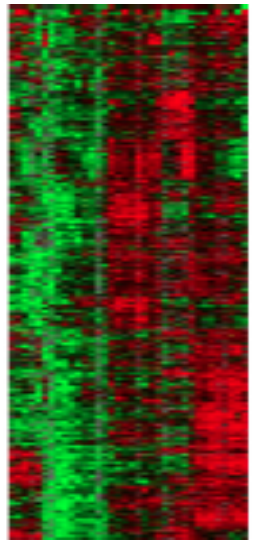
Expression clustering: revisiting an ancient problem



Species A

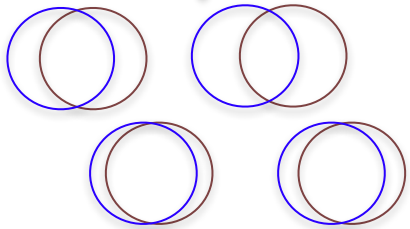
Orthologous pairs between species

Species B



A novel unified framework to integrate co-expression data across species

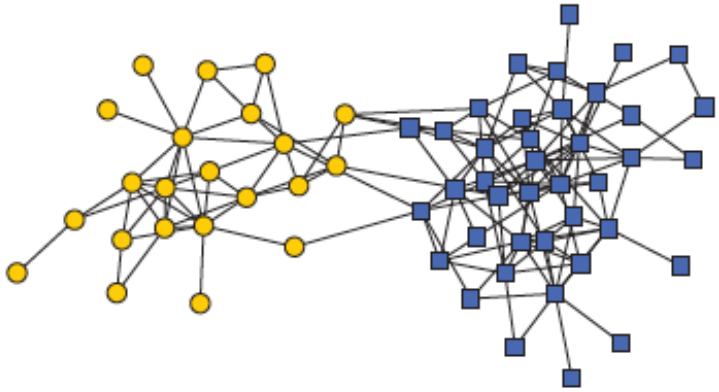
OrthoClust



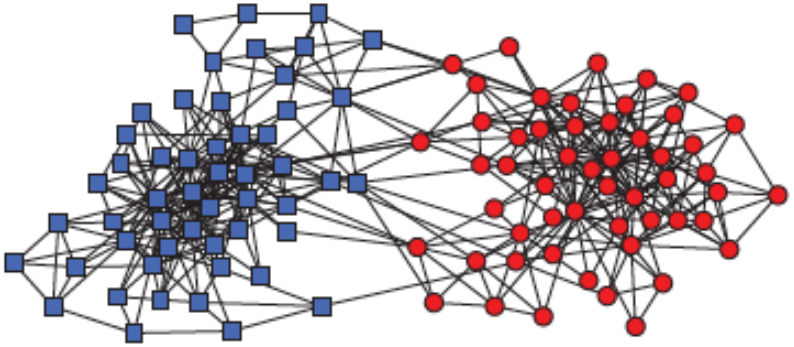
cross species modules

Yan et al. Genome Biol. 2014

Network modularity



Dolphin social network



Political books

Newman Phy. Rev. E 2013

$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

adjacency matrix

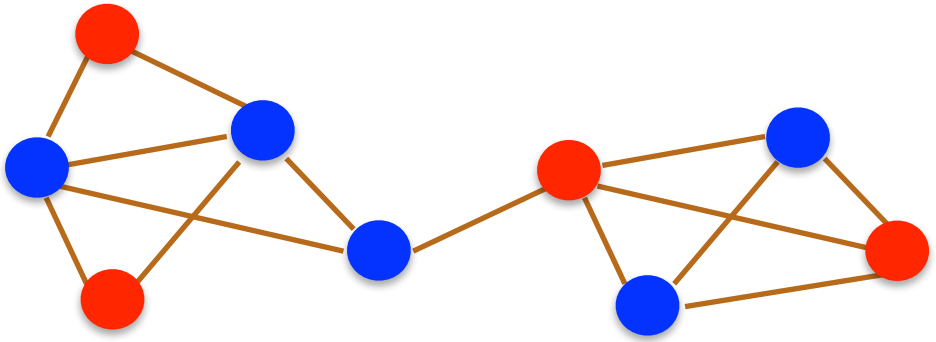
degree of node i

number of edges

expected number of edges between i and j

whether or not i, j are in the same module

Network modularity



$$Q \approx 0$$

$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

adjacency matrix

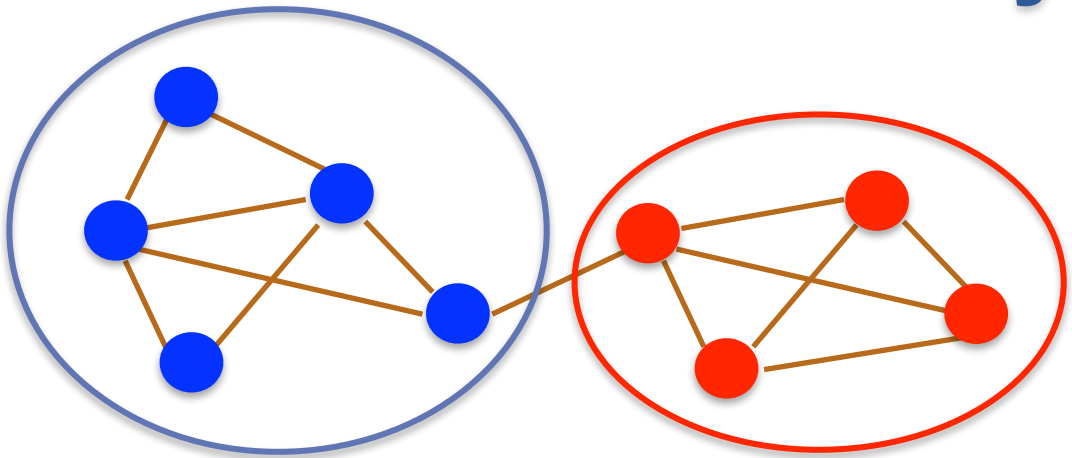
degree of node i

number of edges

expected number of edges between i and j

whether or not i, j are in the same module

Network modularity



$$Q = Q_{max}$$

Optimization problem for sim. annealing

$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

adjacency matrix

degree of node i

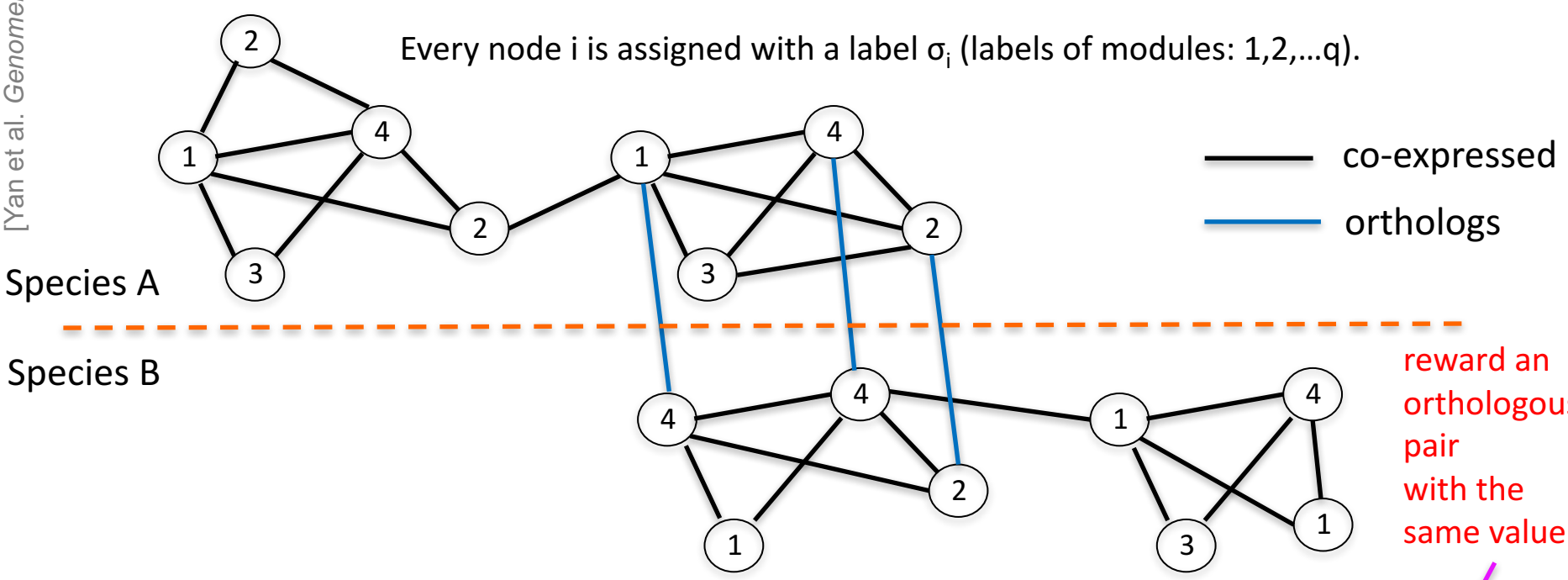
number of edges

expected number of edges between i and j

whether or not i, j are in the same module

A toy example [orthoclust]

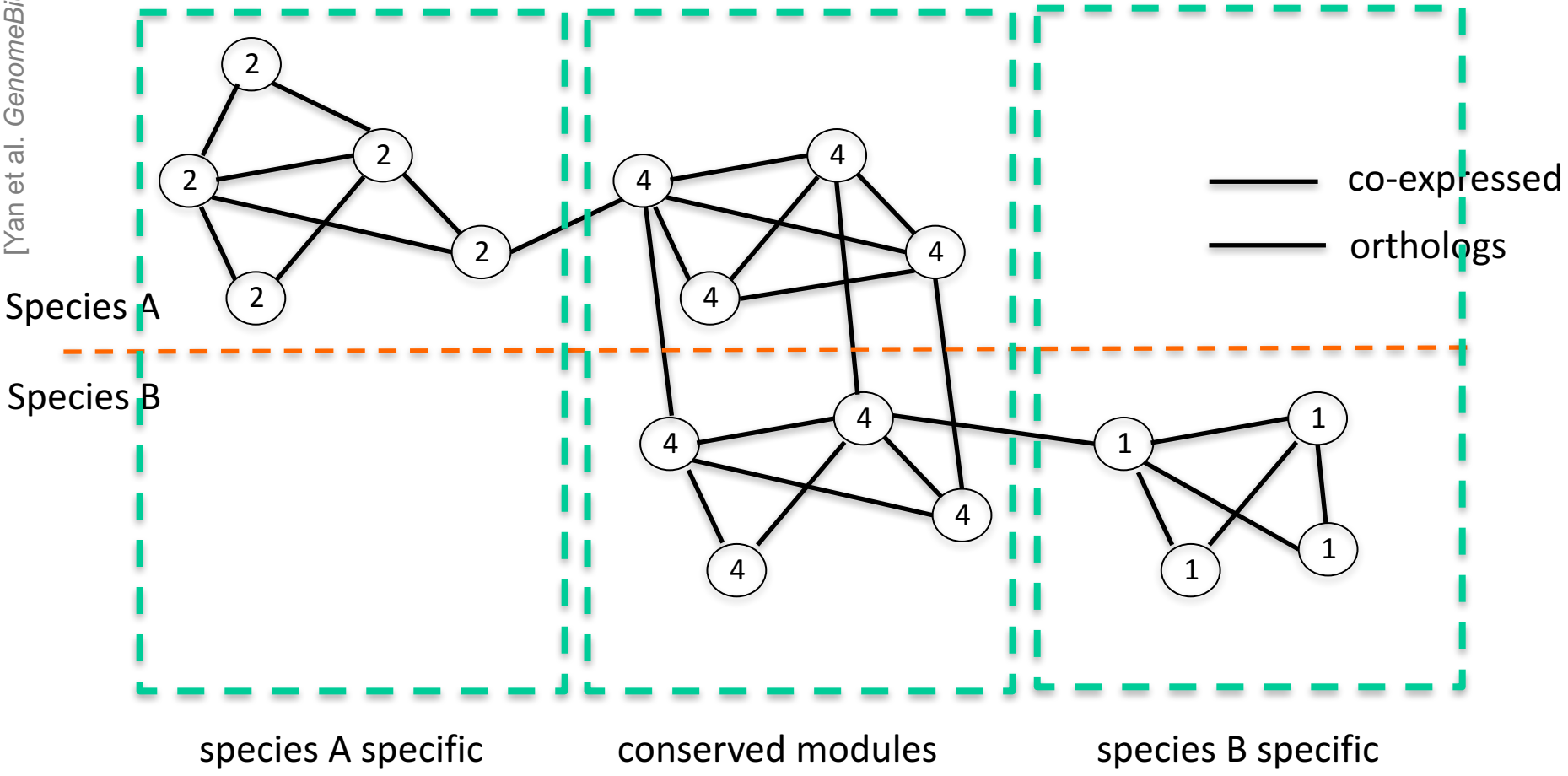
Every node i is assigned with a label σ_i (labels of modules: 1,2,…q).



$$H = \boxed{Q(\text{for all } \sigma_i \text{ in A}) \quad + \quad Q(\text{for all } \sigma_i \text{ in B})} + \kappa \boxed{\sum_{(i,j')\in Ortho} \delta_{\sigma_i \sigma_{j'}}}$$

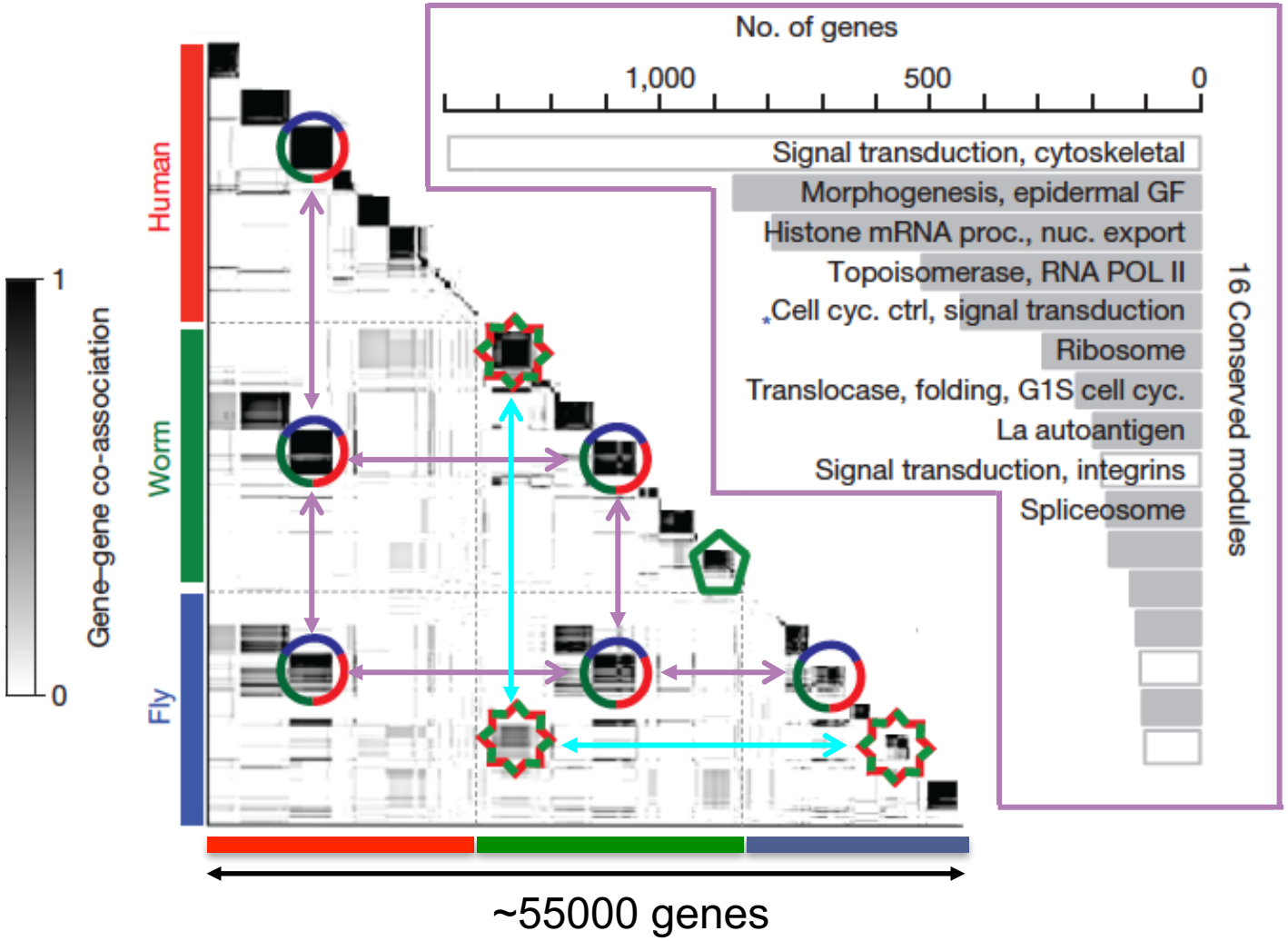
Favorableness = "Modularity" in species A + "Modularity" in species B + consistency betw. A & B

A toy example [orthoclust]



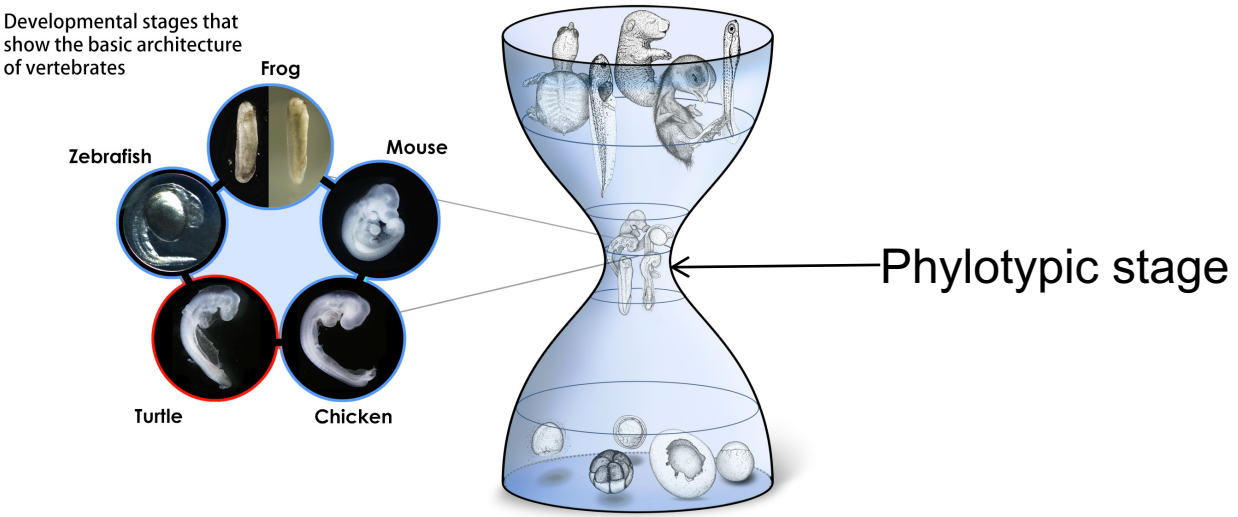
Use Potts model (generalized Ising model) to simultaneously cluster co-expressed genes within an organism as well as orthologs shared between organisms. Here, the ground state configuration correspond to three modules: 1, 2, 4.

Application for more than 2



[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Conserved modules exhibit canonical hourglass behavior



Illustrations courtesy Naoki Irie

Canonical Inter-organism Behavior

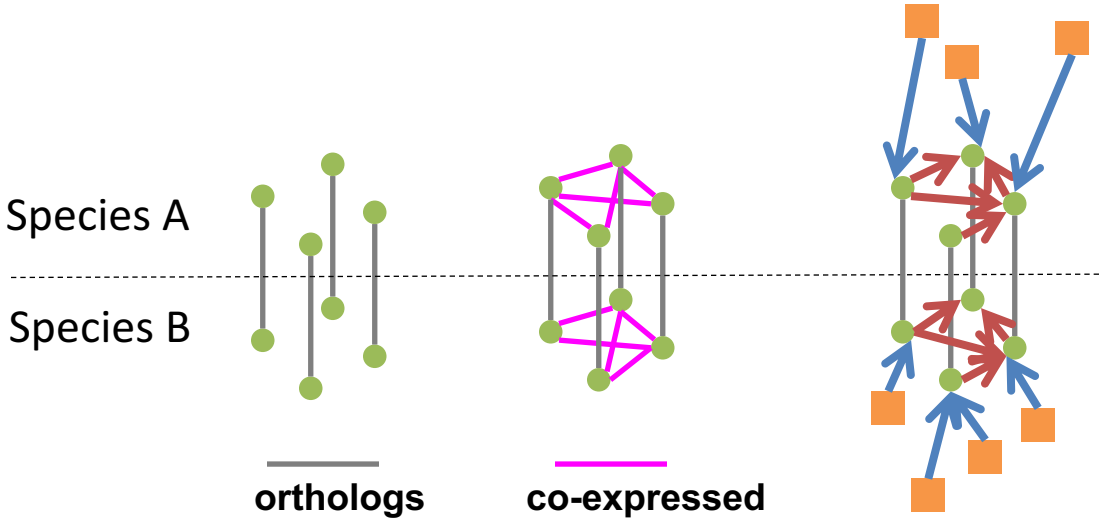
- "Hourglass hypothesis": all organisms go through a particular stage in embryonic development ("phylotypic" stage) where inter-organism expression differences of orthologous genes are smallest.
- **12 out of our 16 modules have this behavior**

[*Nature* 512:445 ('14); doi: 10.1038/nature13424]

Transcriptome Mining: Tackling core issues related to gene regulation & also analyzing the "data exhaust" associated with this activity

- *[Core-1]* **Expression Clustering**, Cross-species
  - Comparative ENCODE – Lots of worm-fly-human matched data & developmental timecourses
  - Optimization gives 16 conserved co-expression modules, 12 w/ hourglass
- *[Core-2]* **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers; Decoupling expression changes into those from conserved vs species-specific genes
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- *[Core-3]* **Logic Gates** Modeling
  - Preponderance of OR gates in cancer v. cell-cycle (esp. for MYC)
- *[Exhaust-1]* **Genomic Privacy & RNA-seq**
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
  - Using file formats to remove obvious variants
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- *[Exhaust-2]* **Publication Patterns** from data producing consortia
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

Is gene regulation among orthologs conserved?



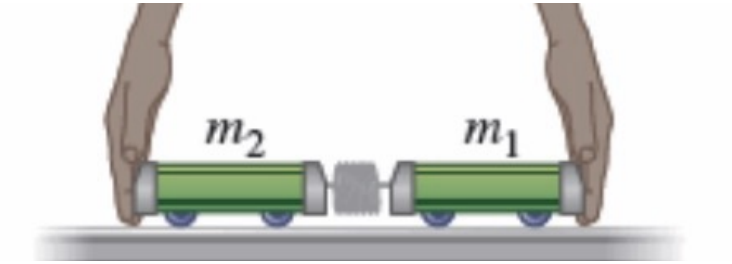
To what degree can't ortholog expression levels be predicted due to species-specific regulation?

Regulation among orthologs (**internal**)

Regulation from species-specific factors (**external**)

Orthologous genes (orthologs)

Species-specific transcription factors



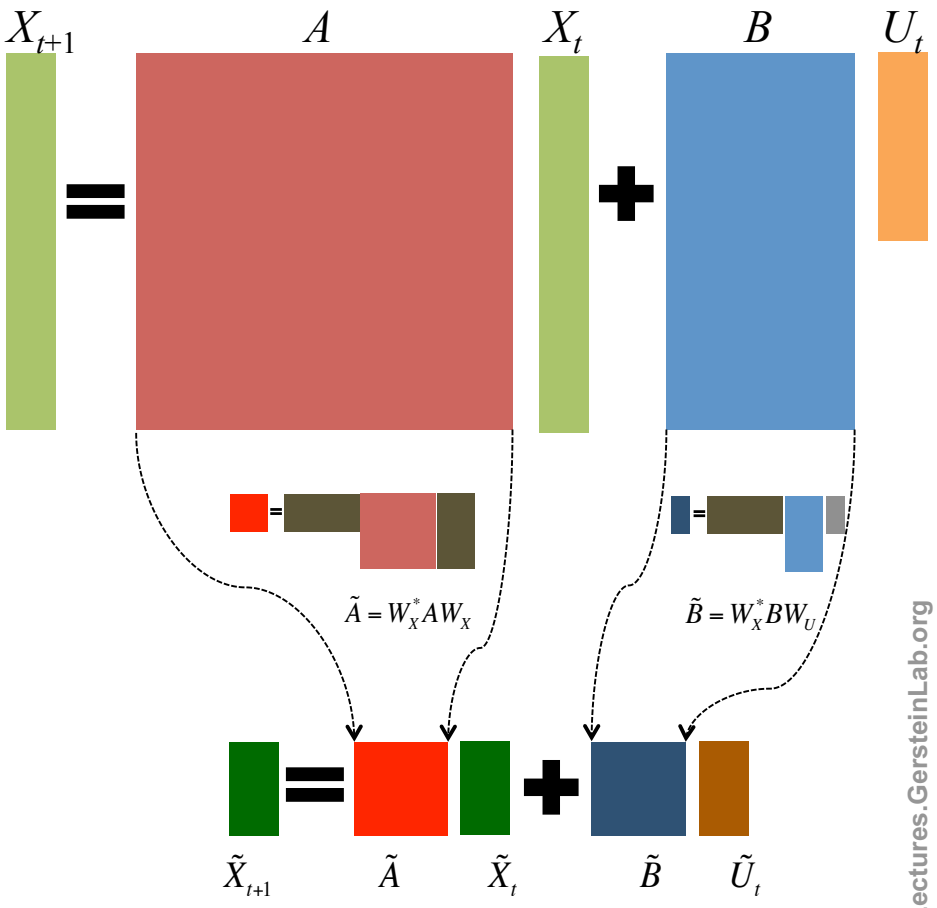
With External force

Purely Internal Dynamics

[Wang et al. *PLOS CB*, '16]

State-space model for internal and external gene regulatory networks

- **State X_t:** Gene expression vector of internal group at time t
- A_{ij} captures temporal casual influence from Gene i to Gene j in internal group
- B_{kl} captures temporal casual influence from external factor k to Gene l in internal group
- **Control U_t:** Gene expression vector of external factors at time t



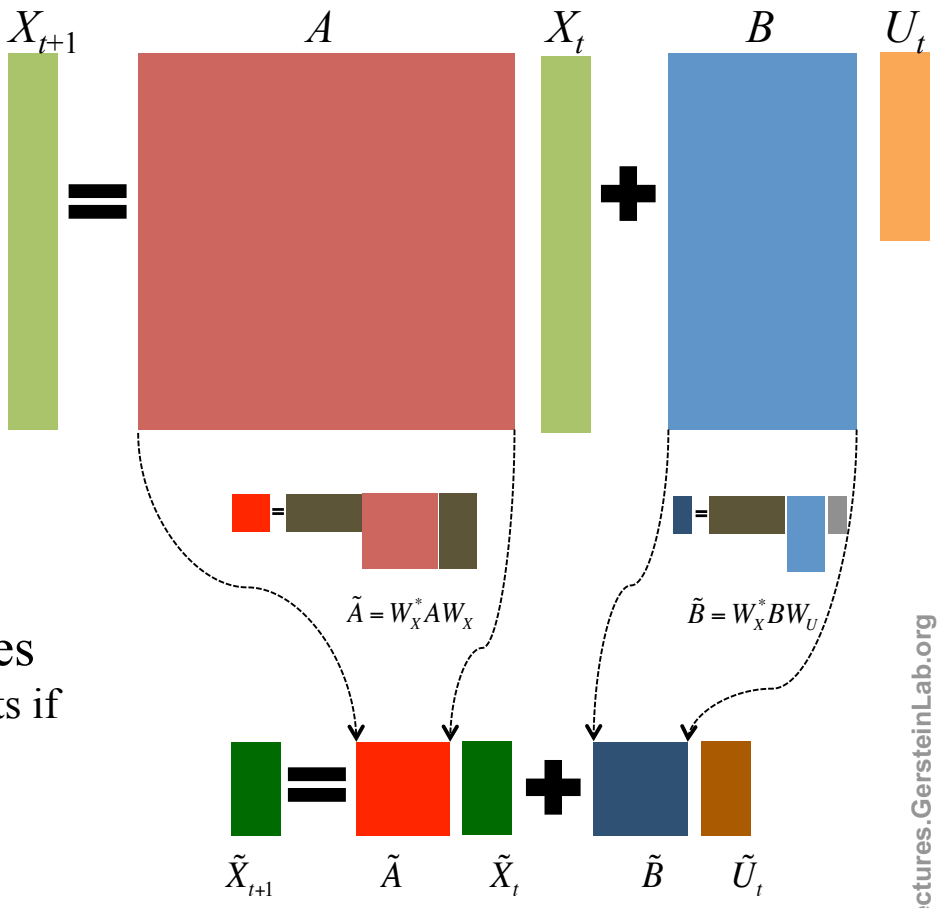
[Wang et al. *PLOS CB*, '16]

State-space model for internal and external gene regulatory networks

Not enough data to estimate state space model for genes
(e.g., 25 time points per gene to estimate 4 million elements of A or B for 2000 genes)

Dimensionality reduction from genes to meta-genes (e.g., SVD)

Effective state space model for meta-genes
(e.g., 250 time points to estimate 50 matrix elements if 5 meta-genes)



[Wang et al. *PLOS CB*, '16]

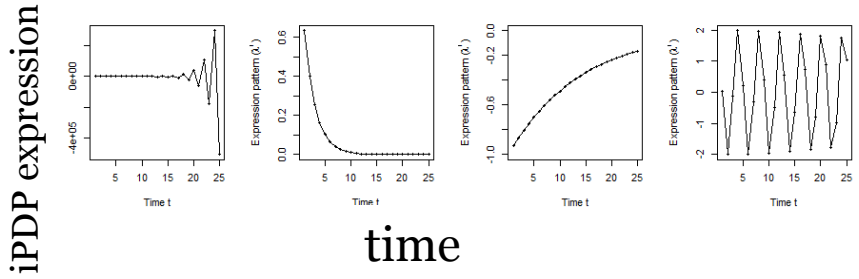
Canonical temporal expression trajectories from effective state space model

Is a std. 1st order homogeneous matrix difference equation. It can solved by diagonalizing A giving….

$$\tilde{X}_{t+1} = \tilde{A}\,\tilde{X}_t + \tilde{B}\,\tilde{U}_t$$

p^{th} internal principal dynamic pattern (iPDP): $[\lambda_p{}^1, \lambda_p{}^2, \ldots, \lambda_p{}^T]$,
where λ_p is p^{th} eigenvalue of $\tilde{A}$.

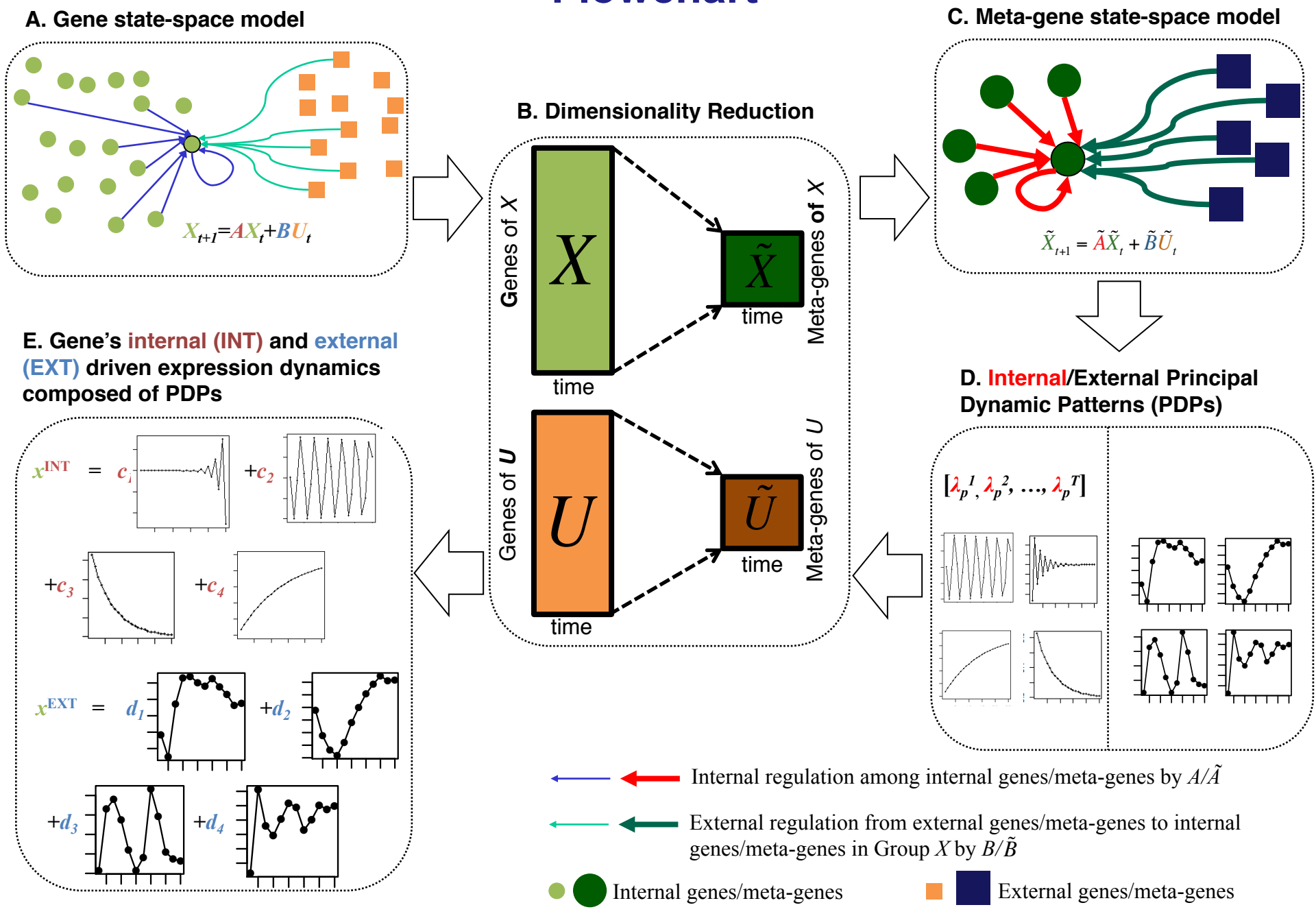
Canonical temporal expression trajectories
(e.g., degradation, growth, damped oscillation, etc.)



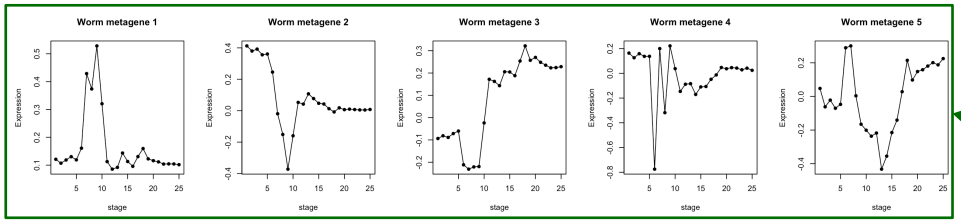
[Wang et al. *PLOS CB*, '16]

[Wang et al. *PLOS CB*, '16]

Flowchart



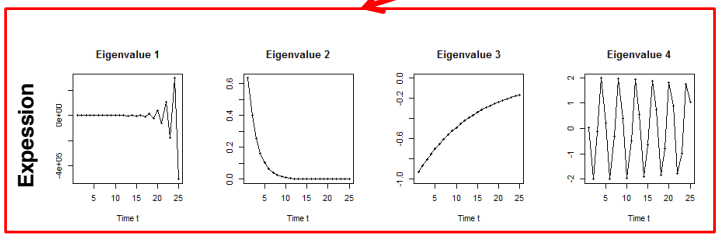
Orthologs have similar internal but different external dynamic patterns during embryonic development

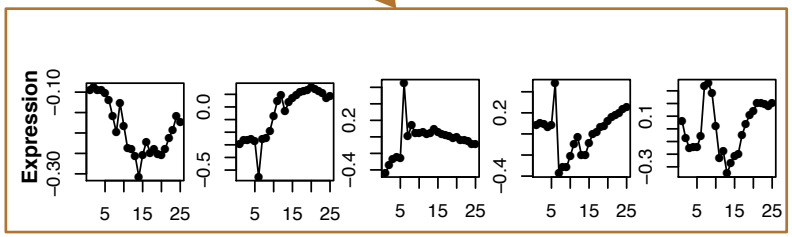


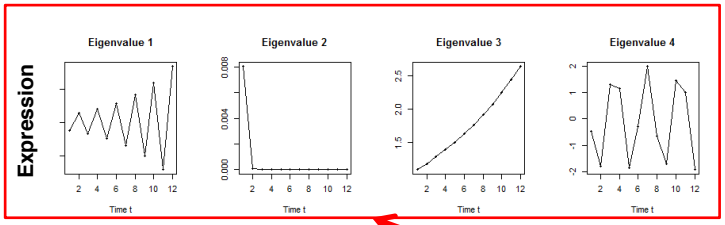
Worm

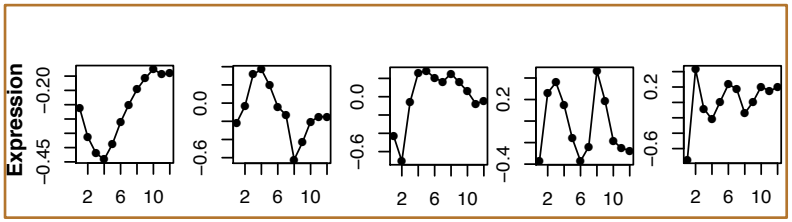
ePDPs

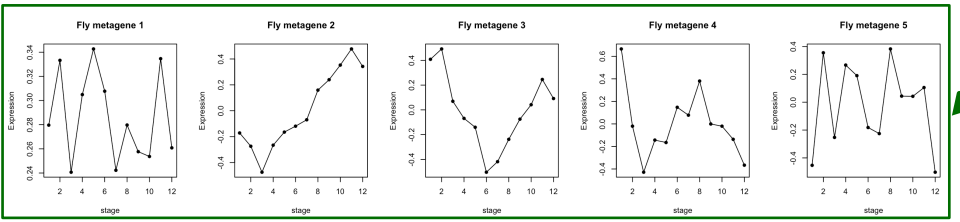
iPDPs











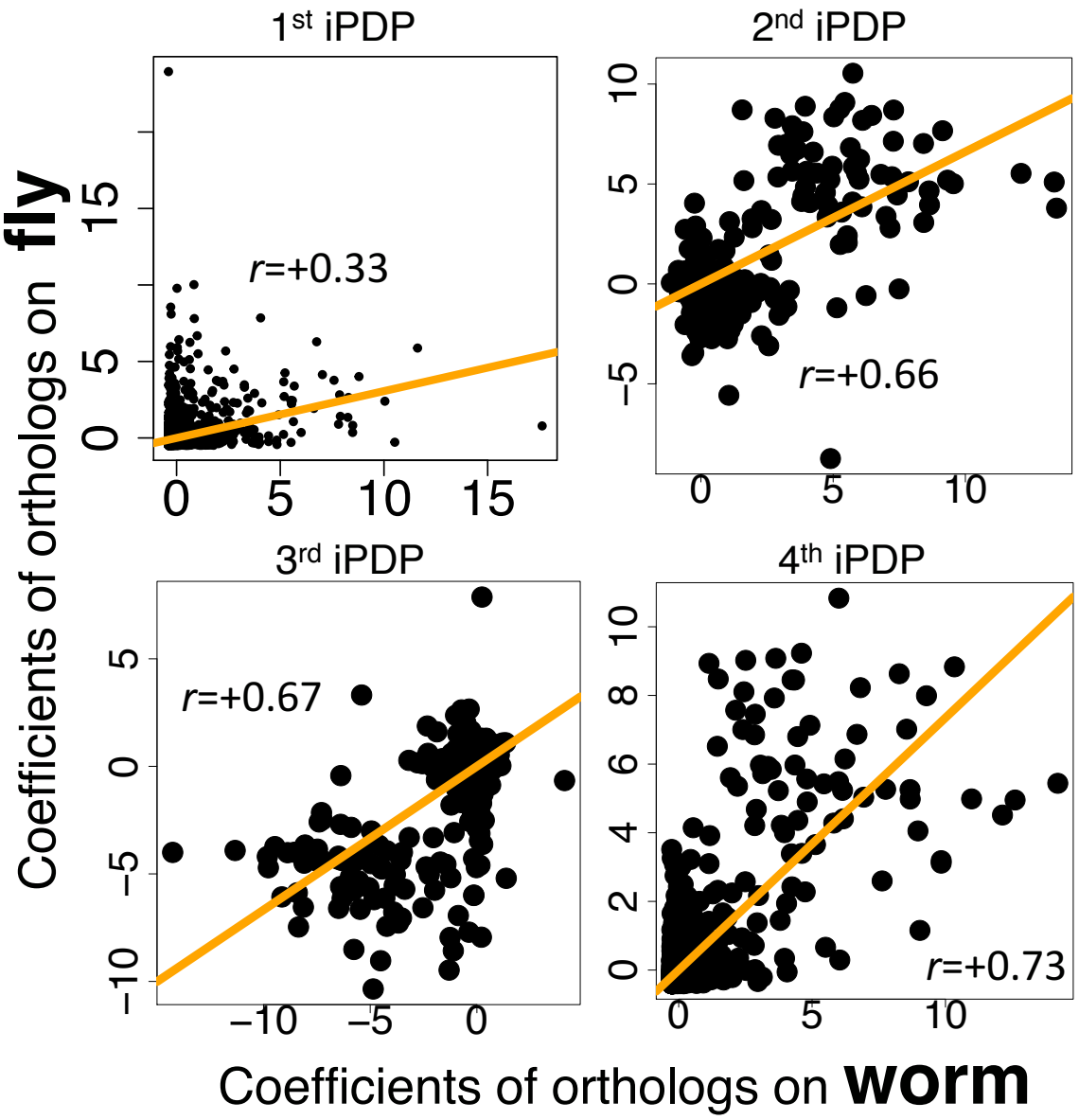
Fly's
effective state space model

Meta-genes

$$\tilde{X}_{t+1} = \tilde{A}\tilde{X}_t + \tilde{B}\tilde{U}_t$$

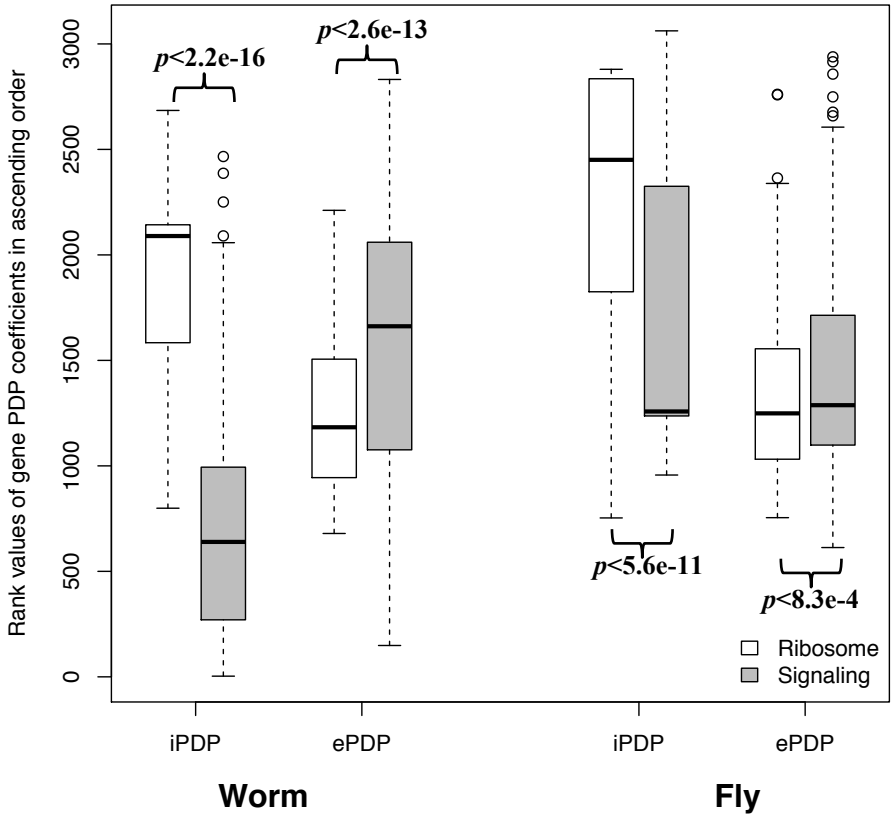
[Wang et al. *PLOS CB*, '16]

Orthologs have correlated iPDP coefficients



[Wang et al. *PLOS CB*, '16]

Evolutionarily conserved & younger genes exhibit the opposite internal & external PDP coefficients



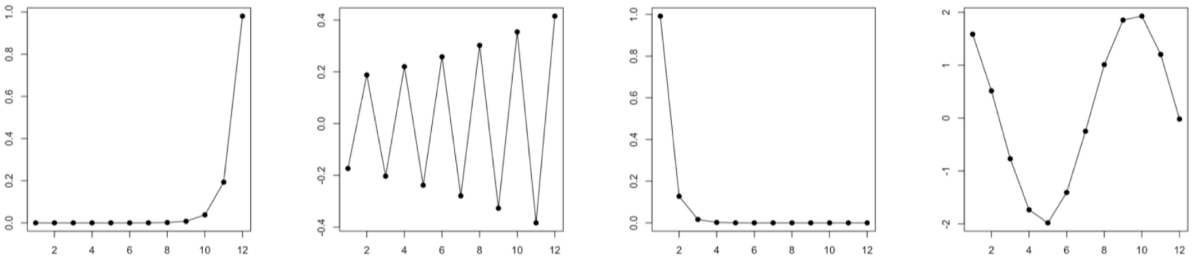
Ribosomal genes have significantly larger coefficients for the internal than external PDPs, but signaling genes exhibit the opposite trend

[Wang et al. *PLOS CB*, '16]

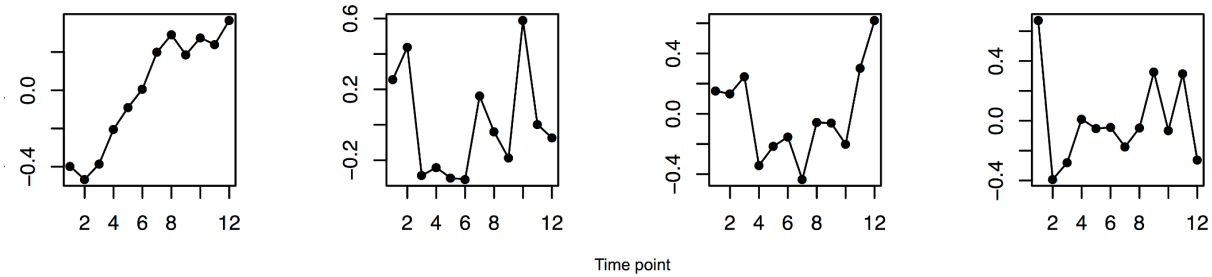
Human-specific TFs respond more strongly to hormonal stimulation during cell-cycle than conserved genes in breast cancer cell

- Applied to Breast Cancer Cell Cycle (2 periods) under hormonal stim.
  - **INT = conserved human genes**
    - ~1100 H-F-W orthologs
    - follow normal cell cycle
  - **EXT = human spec TFs**
    - diff from above
    - perhaps responding to stimulation

iPDPs



ePDPs

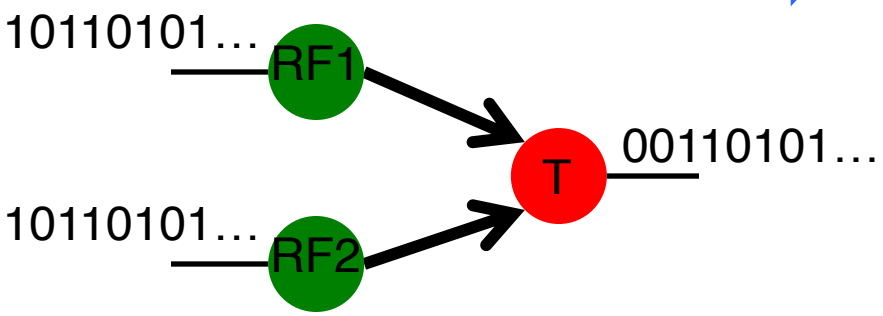


Transcriptome Mining: Tackling core issues related to gene regulation & also analyzing the "data exhaust" associated with this activity

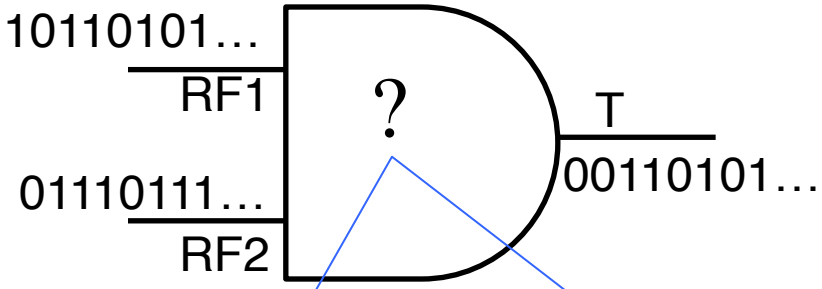
- *[Core-1]* **Expression Clustering**, Cross-species
  - Comparative ENCODE – Lots of worm-fly-human matched data & developmental timecourses
  - Optimization gives 16 conserved co-expression modules, 12 w/ hourglass
- *[Core-2]* **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers; Decoupling expression changes into those from conserved vs species-specific genes
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- *[Core-3]* **Logic Gates** Modeling
  - Preponderance of OR gates in cancer v. cell-cycle (esp. for MYC)
- *[Exhaust-1]* **Genomic Privacy & RNA-seq**
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
  - Using file formats to remove obvious variants
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- *[Exhaust-2]* **Publication Patterns** from data producing consortia
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

Modeling cooperativity between TFs to target gene using logic gates

A regulatory triplet



2-input-1-output logic gate

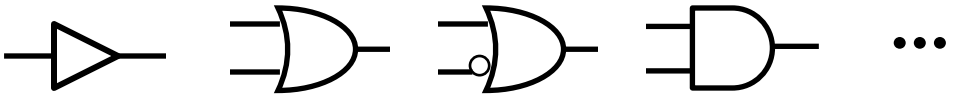


0 – gene off
1 – gene on
after binarizing gene expression data*

Input type (RF1, RF2)	RF1	0	0	1	1
	RF2	0	1	0	1
Output	T	*X*	*X*	*X*	*X*

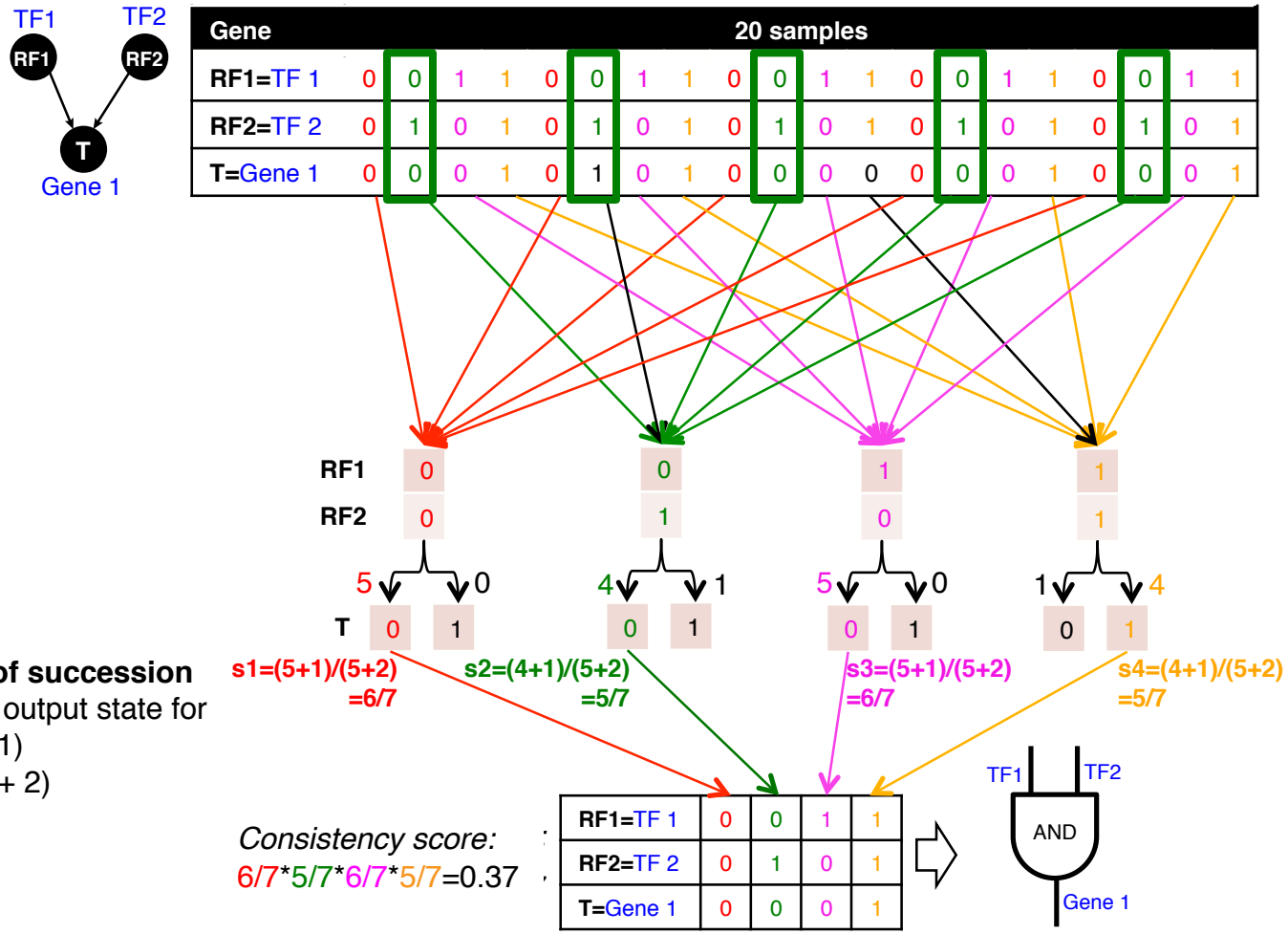
Binarized expression

X can be 0 or 1, so there are 2^4=16 possible output combinations, each of which corresponds to a unique 2-input-1-output logic gate



*BoolNet, R package

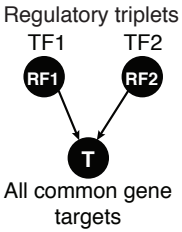
An example: selection of the best-matched logic gate



Laplace's rule of succession
s=(# of selected output state for the input type + 1)
/(# of input type + 2)

Wang, et al., *PLoS Computational Biology*, 2015

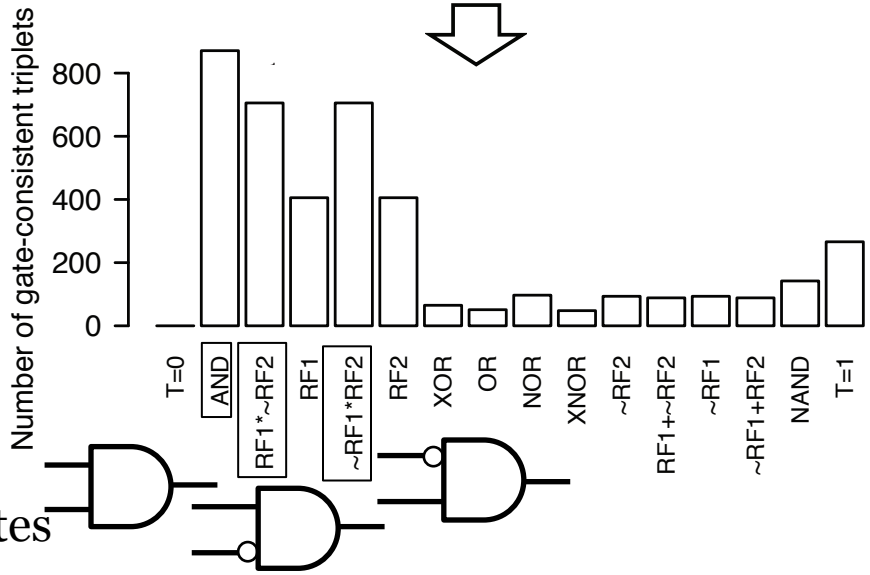
App. 1 – TF cooperativity in the cell cycle



Target gene	2464
TF	176
Triplet	39,011
Time point	59

Yeast Cell Cycle

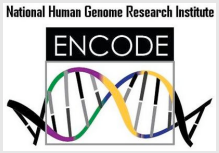
Triplet ID	RF1	RF2	Common Target Gene (T)	Matched logic gate
1	YHR084W	YBR083W	YBR082C	AND
2	YKL112W	YIL131C	YMR198W	OR
…	…	…	…	…
39011	YOR113W	YBL103C	YDR042C	XOR



AND-like gates

Wang, et al., *PLoS Computational Biology*, 2015

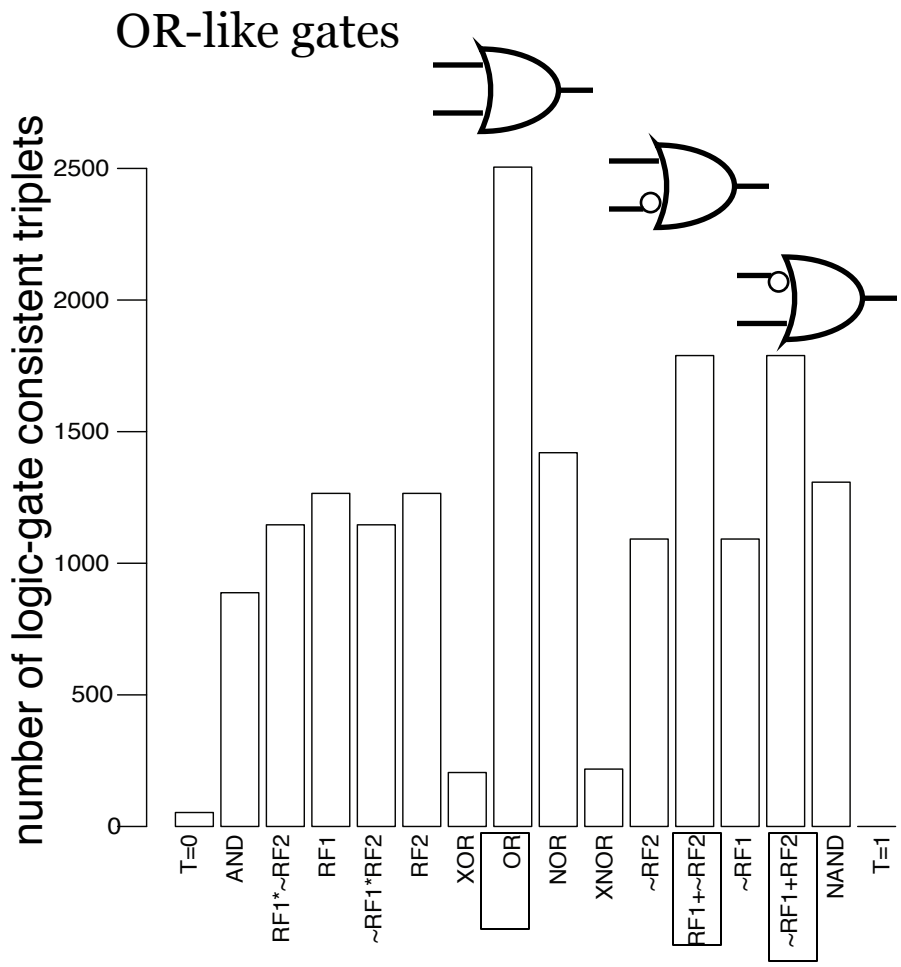
Acute Myeloid Leukemia (AML)

Target gene	1824	ENCODE Data (K562, ChIP-seq)
TF	70	
Regulatory triplet	50,865	TCGA Data (AML, level 3, RNA-seq) https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp
Patient sample	197	

Wang, et al., *PLoS Computational Biology*, 2015

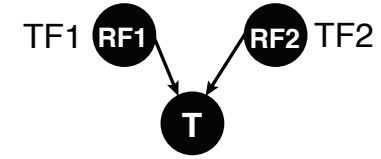
App. 2 – TF cooperativity in AML

Regulatory triplet from ENCODE	50,865
Patient sample for TCGA AML expression data	197



Human TF-TF-target

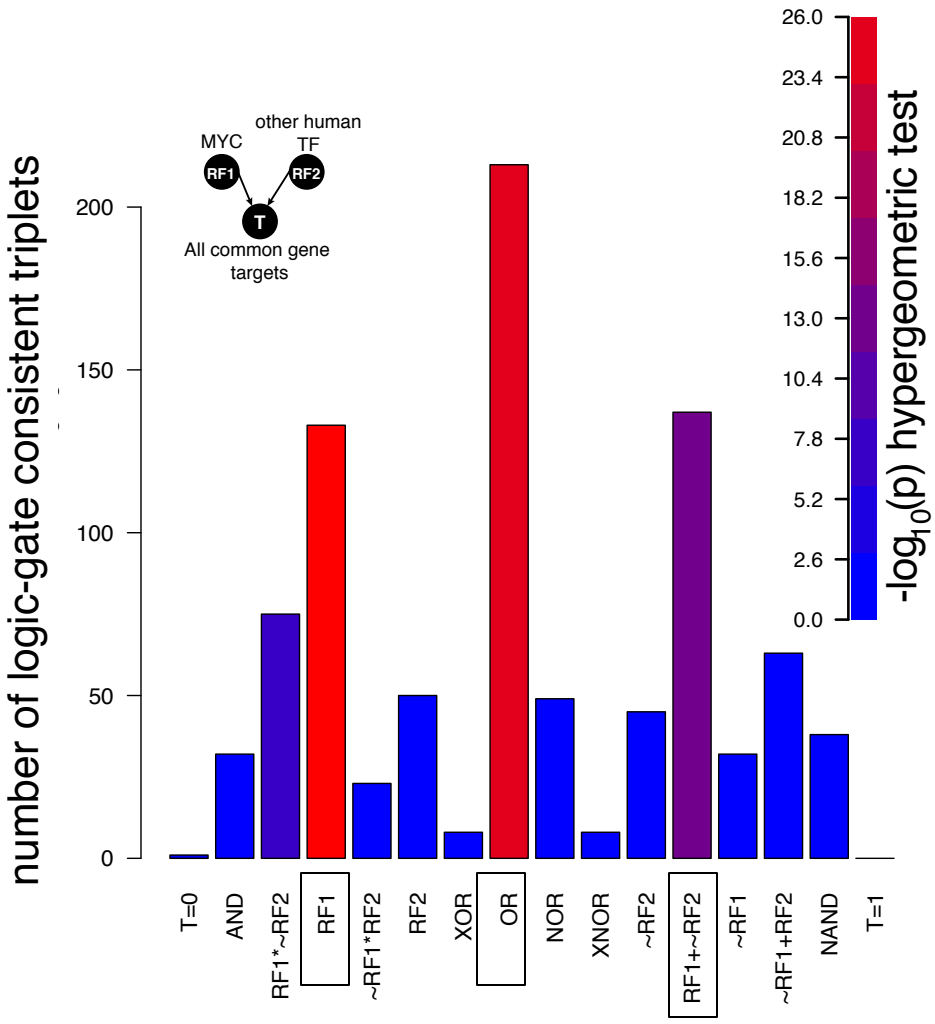
RF1	RF2	Common Target Gene (T)	Matched logic gate
ATF3	BDP1	YPEL1	AND
MYC	BCL3	BCR	T=RF1
ATF3	BRF2	AIF1L	AND
…	…	…	…



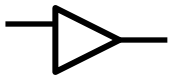
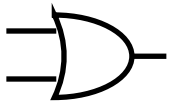
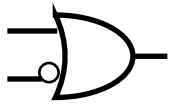
All common gene targets

Wang, et al., *PLoS Computational Biology*, 2015

Cancer-related TF, MYC, universally amplifies target expression



Restrict to RF1=MYC, giving 2,153 triplets

- RF1 
- **OR**(RF1, RF2) 
- **OR**(RF1, **NOT** RF2) 

High expression of MYC is sufficient for high target gene expression

c-Myc Is a Universal Amplifier of Expressed Genes in Lymphocytes and Embryonic Stem Cells

Cell

Zuqin Nie,[1,6] Gangqing Hu,[2,6] Gang Wei,[2] Kairong Cui,[2] Arito Yamane,[3] Wolfgang Resch,[3] Ruoning Wang,[4] Douglas R. Green,[4] Lino Tessarollo,[5] Rafael Casellas,[3] Keji Zhao,[2,*] and David Levens[1,*]

Wang, et al., *PLoS Computational Biology*, 2015

Transcriptome Mining: Tackling core issues related to gene regulation & also analyzing the "data exhaust" associated with this activity

- *[Core-1]* **Expression Clustering**, Cross-species
  - Comparative ENCODE – Lots of worm-fly-human matched data & developmental timecourses
  - Optimization gives 16 conserved co-expression modules, 12 w/ hourglass
- *[Core-2]* **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers; Decoupling expression changes into those from conserved vs species-specific genes
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- *[Core-3]* **Logic Gates** Modeling
  - Preponderance of OR gates in cancer v. cell-cycle (esp. for MYC)
- *[Exhaust-1]* **Genomic Privacy & RNA-seq**
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
  - Using file formats to remove obvious variants
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- *[Exhaust-2]* **Publication Patterns** from data producing consortia
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

2-sided nature of functional genomics data: Analysis can be very General/Public or Individual/Private

- **General quantifications** related to overall aspects of a condition – ie gene activity as a function of:
  - Developmental stage, Evolutionary relationships, Cell-type, Disease

- **Above are not tied to an individual's genotype. However, data is derived from individuals & tagged with their genotypes**

- (Note, a few calculations aim to use explicitly genotype to derive general relations related to sequence variation & gene expression - eg allelic activity)

Genomics has similar "Big Data" Dilemma in the Rest of Society

- Sharing & "peer-production" is central to success of many new ventures, with the same risks as in genomics
  - **EG web search**: Large-scale mining essential

- We confront privacy risks every day we access the internet

[Seringhaus & Gerstein ('09), *Hart. Courant* (Jun 5); Greenbaum & Gerstein ('11), *NY Times* (6 Oct)]

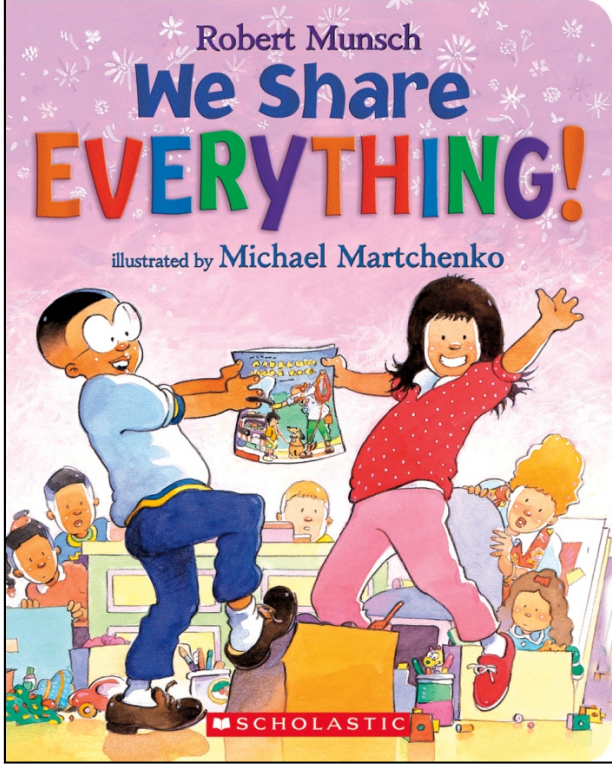
Tricky Privacy Considerations in Personal Genomics

- **Genetic Exceptionalism :** The Genome is very fundamental data, potentially very revealing about one's identity & characteristics
- **Personal Genomic info. essentially meaningless currently but will it be in 20 yrs? 50 yrs?**
  - Genomic sequence very revealing about one's children. Is true consent possible?
  - Once put on the web it can't be taken back
- **Culture Clash:** Genomics historically has been a proponent of "open data" but not clear personal genomics fits this.
  - Clinical Medline has a very different culture.
- **Ethically challenged** history of genetics
  - Ownership of the data & what consent means (Hela)
    - Could your genetic data give rise to a product line?

[D Greenbaum & M Gerstein ('08). Am J. Bioethics; D Greenbaum & M Gerstein, Hartford Courant, 10 Jul. '08 ; SF Chronicle, 2 Nov. '08; Greenbaum et al. *PLOS CB* ('11) ; Greenbaum & Gerstein ('13), The Scientist; Photo from NY Times]

The Other Side of the Coin: Why we should share

- Sharing helps **speed research**
  - Large-scale mining of this information is important for medical research
  - Privacy is cumbersome, particularly for big data
- Sharing is important for **reproducible research**
- Sharing is useful for **education**
  - More fun to study a known person's genome
    - Eg Zimmer's Game of Genomes in STAT



[Yale Law Roundtable ('10). Comp. in Sci. & Eng. 12:8; D Greenbaum & M Gerstein ('09). Am. J. Bioethics; D Greenbaum & M Gerstein ('10). SF Chronicle, May 2, Page E-4; Greenbaum et al. *PLOS CB* ('11)]

[Economist, 15 Aug '15]

The Dilemma

- The individual (harmed?) v the collective (benefits)
  - But do sick patients care about their privacy?
- How to balance risks v rewards - Quantification
  - What is acceptable risk?
    Can we quantify leakage?
    - Ex: photos of eye color
  - Cost Benefit Analysis

Current Social & Technical Solutions

- **Closed Data** Approach
  - Consents
  - "Protected" distribution via dbGAP
  - Local computes on secure computer
- Issues with Closed Data
  - Non-uniformity of consents & paperwork
    - Different international norms, leading to confusion
  - Encryption & computer security creates burdensome requirements on data sharing & large scale analysis
  - Many schemes get "hacked"

- **Open Data**
  - Genomic "test pilots" (ala PGP)?
    - Sports stars & celebrities?
  - Some public data & data donation is helpful but is this a realistic solution for an unbiased sample of ~1M

[Greenbuam et al ('04), Nat. Biotech; Greenbaum & Gerstein ('13), The Scientist]

Strawman Hybrid Social & Tech Proposed Solution?

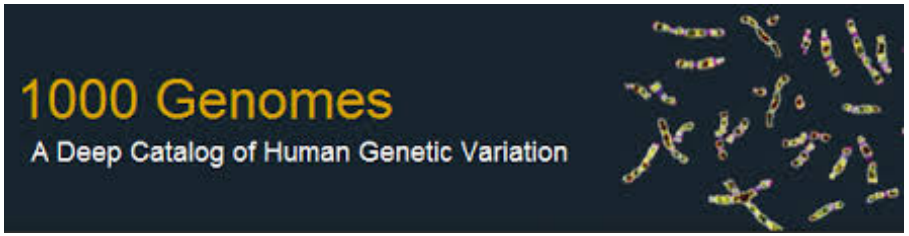
- Fundamentally, researchers have to keep genetic secrets.
  - **Need for an (international) legal framework**
  - Genetic Licensure & training for individuals (similar to medical license, drivers license)
- Technology to make things easier
  - Cloud computing & enclaves (eg solution of Genomics England)
- Technological barriers shouldn't create a social incentive for “hacking”

- **Quantifying Leakage & allowing a small amounts of it**
- **Careful separation & coupling of private & public data**
  - **Lightweight, freely accessible secondary datasets coupled to underlying variants**
  - Selection of stub & "test pilot" datasets for benchmarking
  - Develop programs on public stubs on your laptop, then move the program to the cloud for private production run

[D Greenbaum, M Gerstein (‘11). Am J Bioeth 11:39. Greenbaum & Gerstein, The Scientist ('13)]

Representative Expression, Genotype, eQTL Datasets

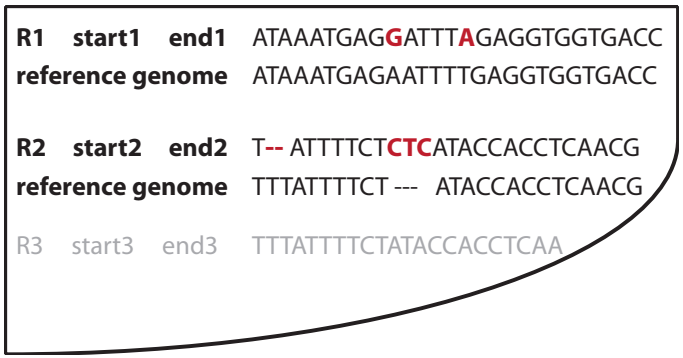
- Genotypes are available from the 1000 Genomes Project
- mRNA sequencing for 462 individuals from gEUVADIS and ENCODE
  - Publicly available quantification for protein coding genes
- Approximately 3,000 cis-eQTL (FDR<0.05)

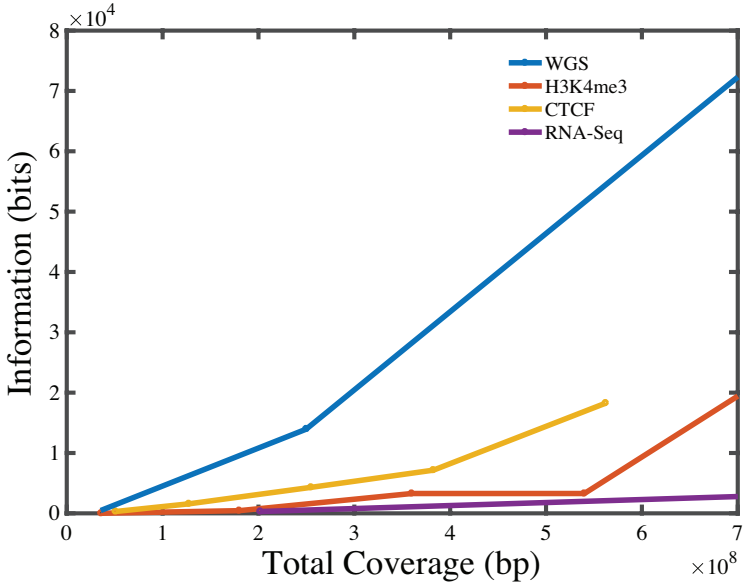


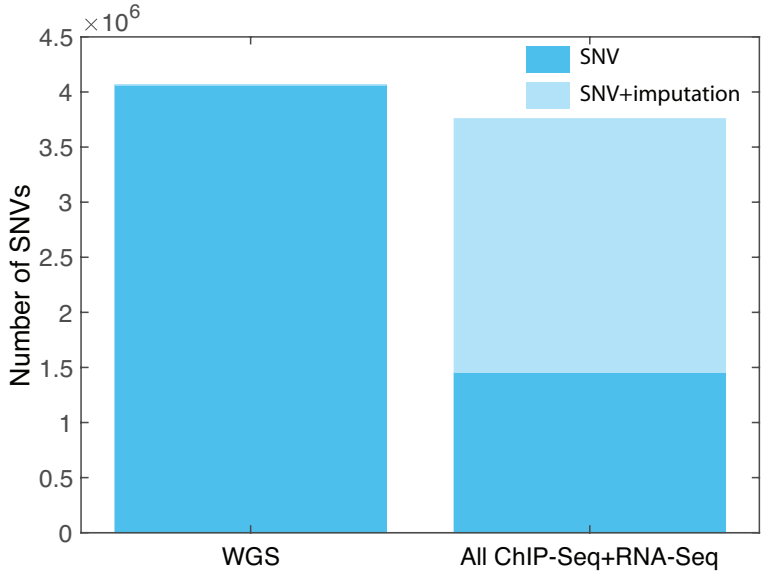
Variants directly in the reads

- Functional genomics data comes with a great deal of sequencing
  - NA12878 as case study - 1000 genomes variants are used as gold standard
- How much information, for example, do RNA-Seq reads (or ChIP-Seq) reads contain? Does that information enough to identify individuals?

Variants from RNA-Seq reads



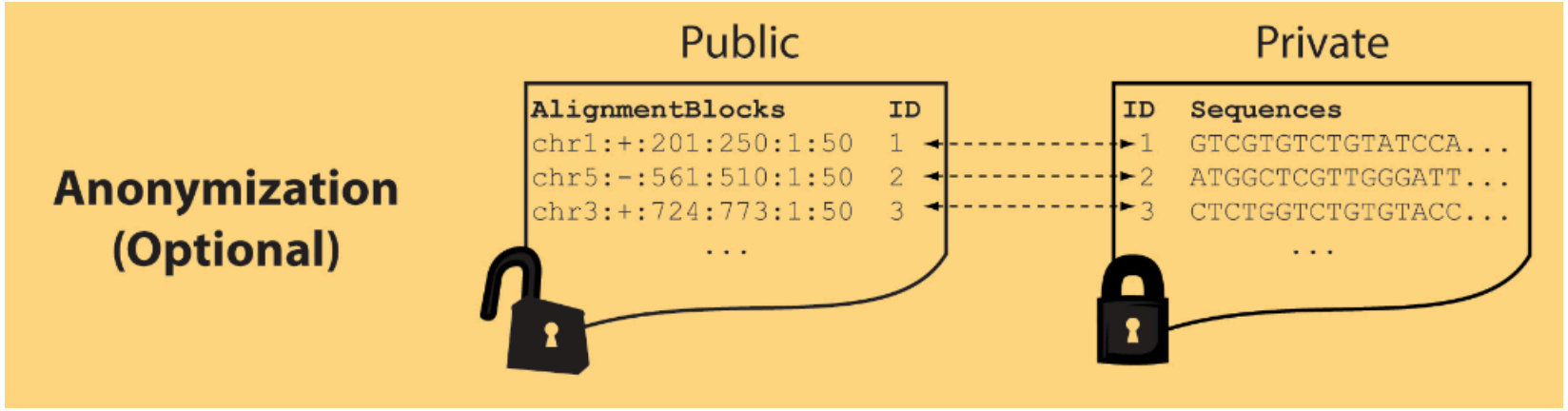


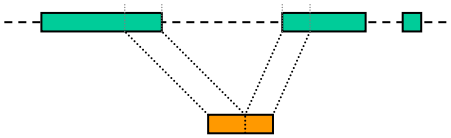


- It might seem like we don't infer much information from single ChIP-Seq and RNA-Seq experiments compared to WGS
  - However putting 10 different ChIP-Seq experiments and RNA-Seq together with imputation provides a great deal of information about the individual

Light-weight formats to Hide Most of the Read Data (Signal Tracks)

- Some lightweight format clearly separate public & private info., aiding exchange
- Files become much smaller
- Distinction between formats to compute on and those to archive with – become sharper with big data

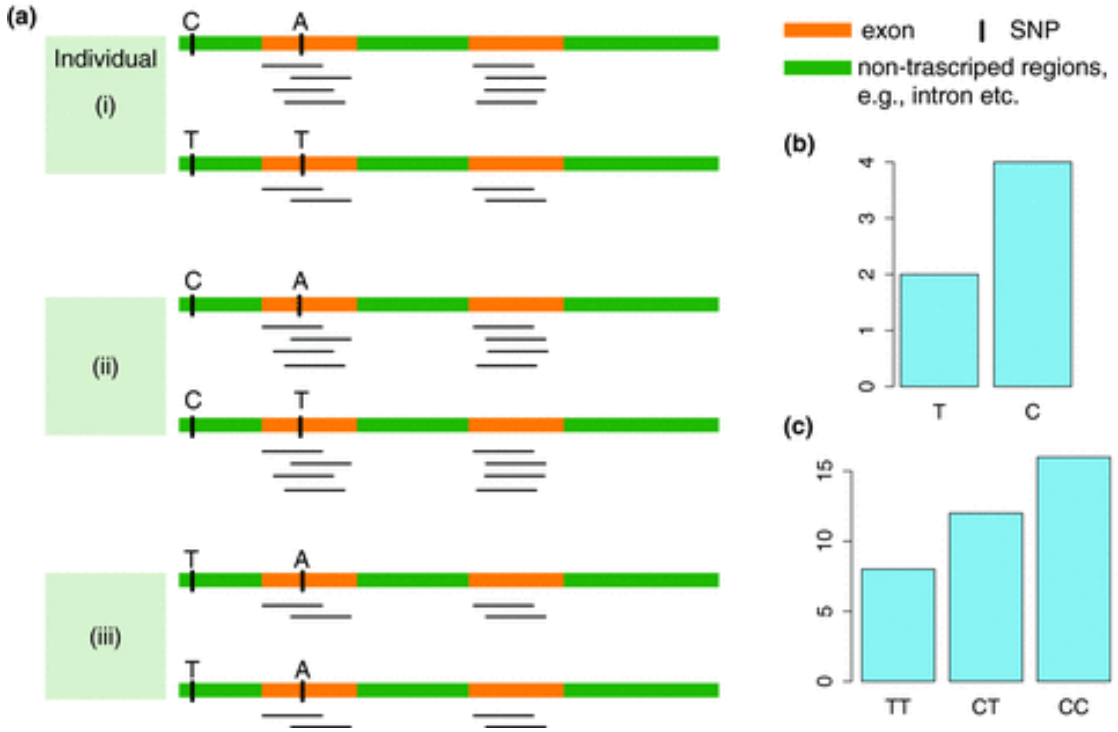




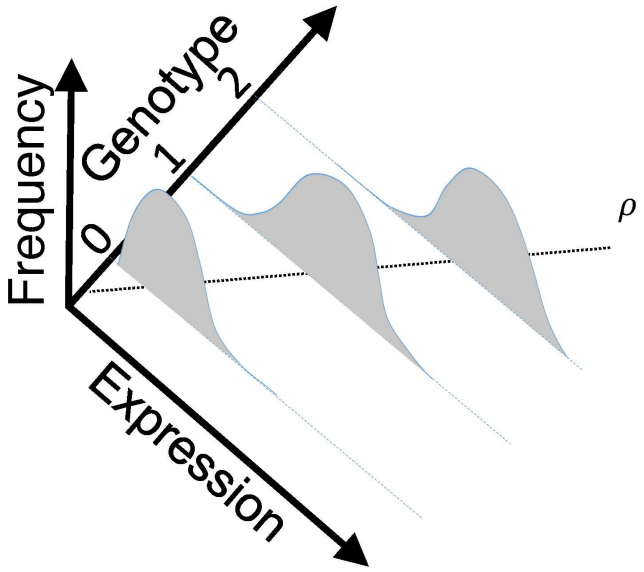
Mapping coordinates without variants (MRF)

Reads (linked via ID, 10X larger than mapping coord.)

[*Bioinformatics* 27: 281]



[*Biometrics 68(1) 1–11*]

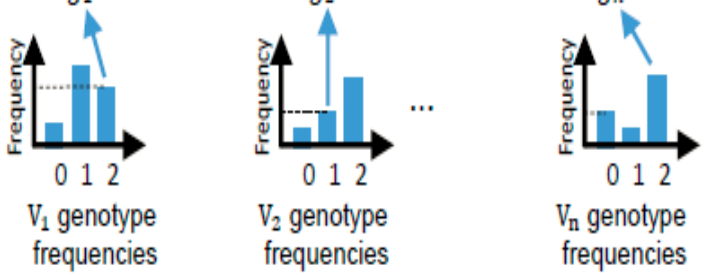


eQTL Mapping Using RNA-Seq Data

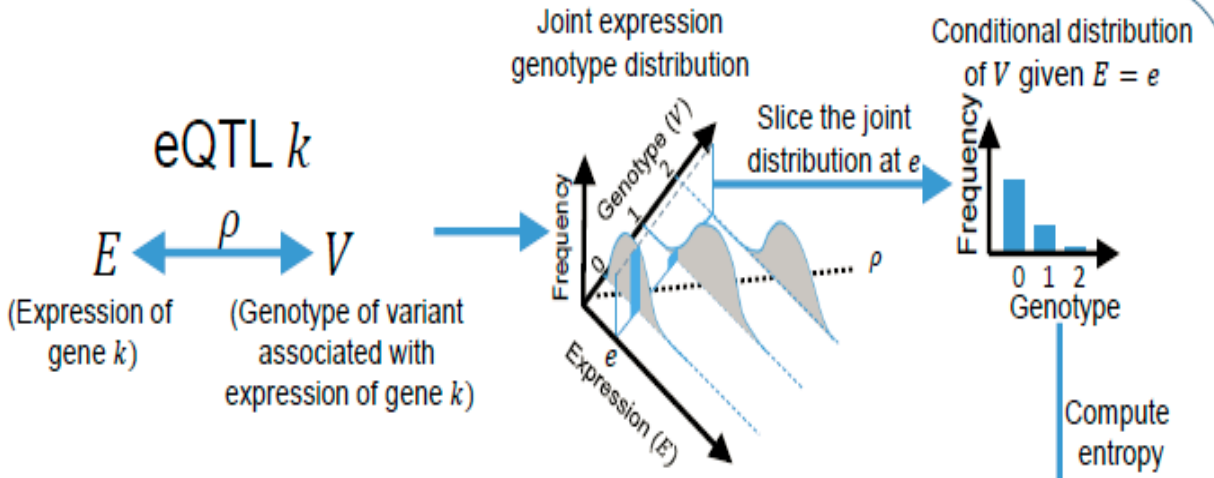
- eQTLs are genomic loci that contribute to variation in mRNA expression levels
- eQTLs provide insights on transcription regulation, and the molecular basis of phenotypic outcomes
- eQTL mapping can be done with RNA-Seq data

Information Content and Predictability

$$ICI\left(\begin{array}{c}\text{Individual has variant}\\ \text{genotypes } g_1, g_2, \ldots, g_n\\ \text{for variants } V_1, V_1, \ldots, V_n\end{array}\right) = \log\left(\frac{1}{\text{Frequency of } V_1 \text{ genotype } g_1 = 2}\right) + \log\left(\frac{1}{\text{Frequency of } V_2 \text{ genotype } g_2 = 1}\right) + \ldots + \log\left(\frac{1}{\text{Frequency of } V_n \text{ genotype } g_n = 2}\right)$$



- Naive measure of information (no LD, distant correlations, pop. struc., &c)
- Higher frequency: Lower ICI
- Additive for multiple variants

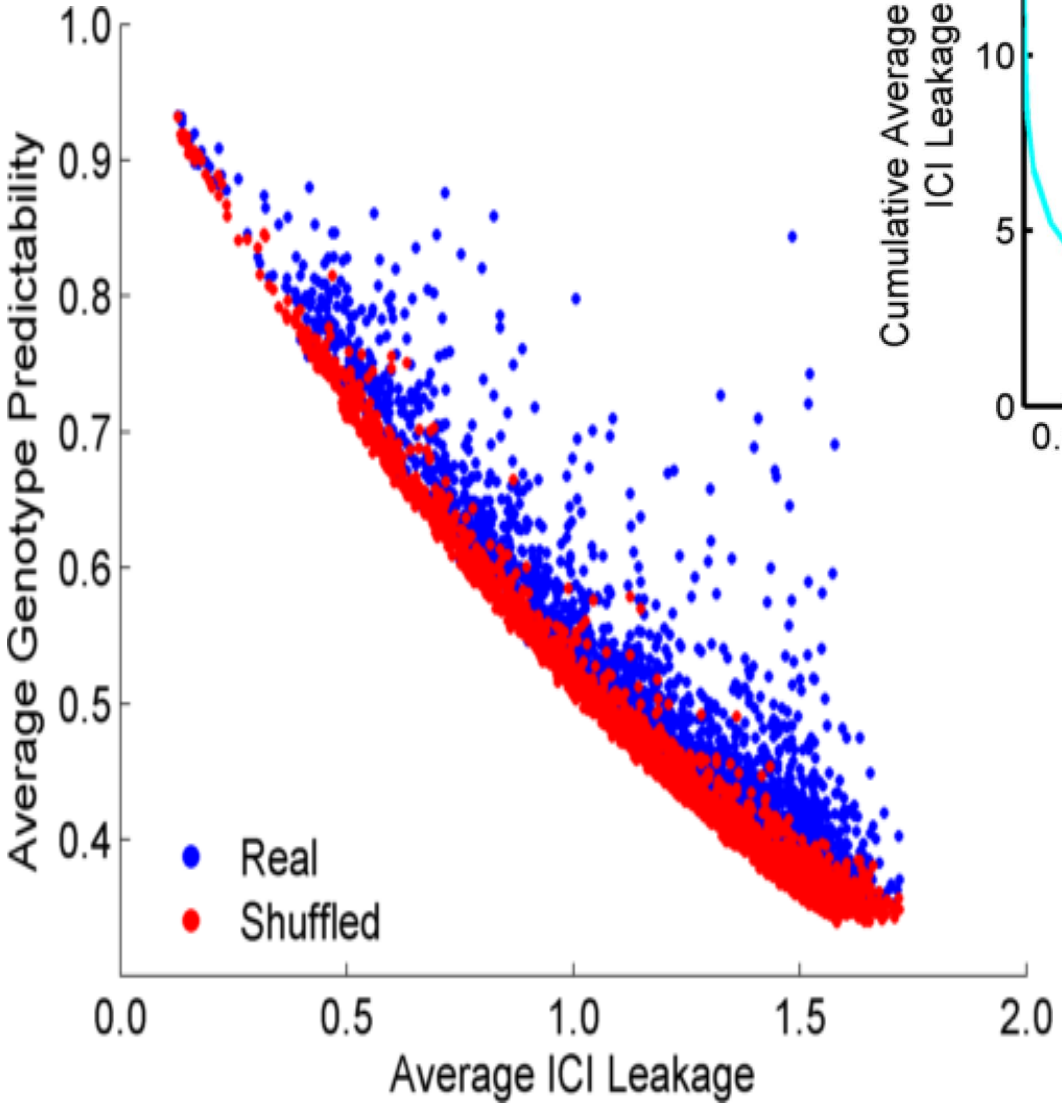


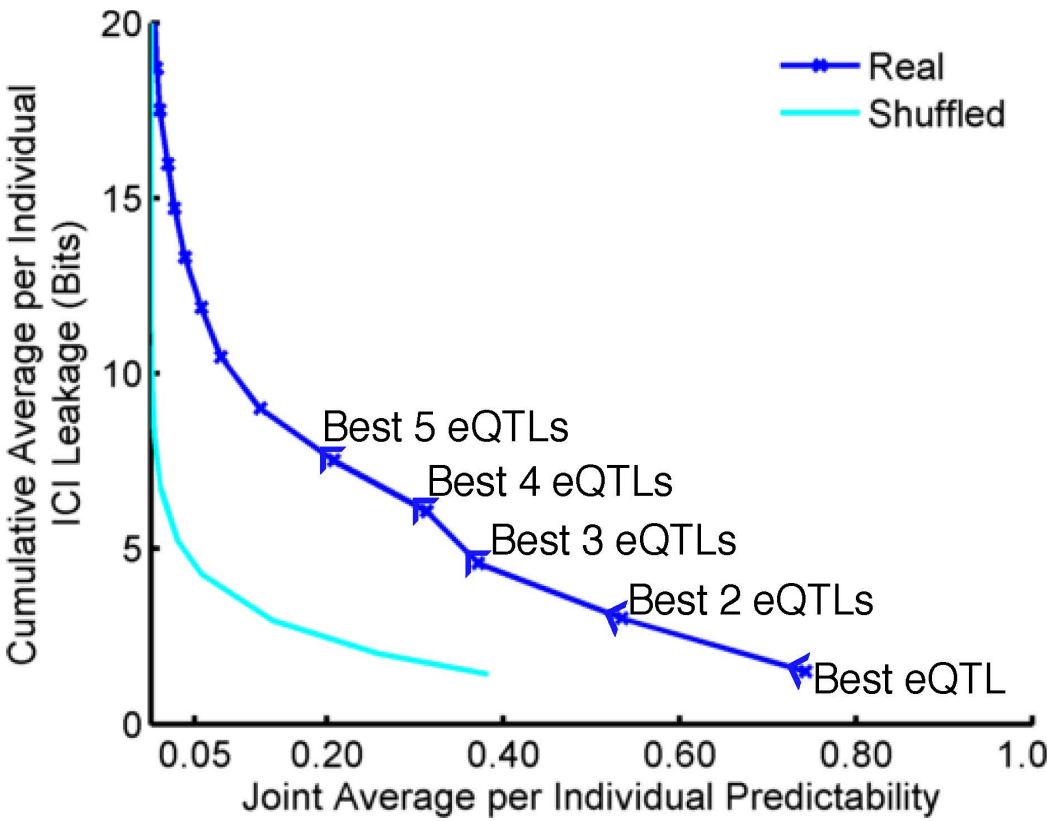
$$\pi\left(V \mid E = e\right) = e^{-H\left(V \mid E = e\right)}$$

- Condition specific entropy
- Higher cond. entropy: Lower predictability
- Additive for multiple eQTLs

[Harmanci et al. Nat. Meth. 2016]

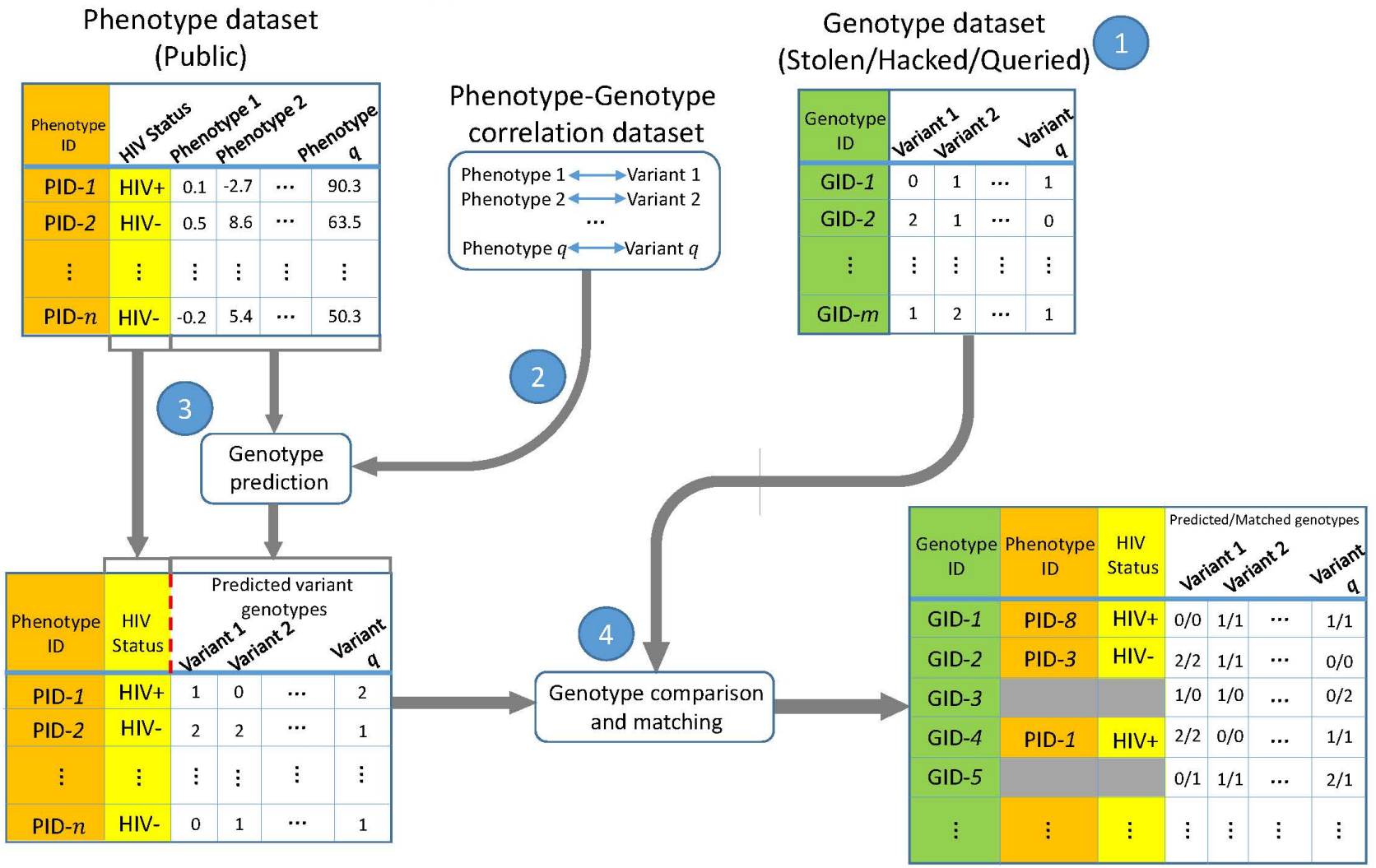
[Harmanciet al. Nat. Meth. (‘16]





ICI Leakage versus Genotype Predictability

Linking Attack Scenario



[Harmanciet al. Nat. Meth. (in revision)]

Linking Attacks: Case of Netflix Prize

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
…	…	…	…
…	…	…	…

Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
…	…	…	…
…	…	…	…
…	…	…	…

- **Many users are shared**
- **The grades of same users are correlated**
- **A user grades one movie around the same date in two databases**

Anonymized Netflix Prize Training Dataset made available to contestants

Linking Attacks: Case of Netflix Prize

Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
…	…	…	…
…	…	…	…

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
…	…	…	…
…	…	…	…
…	…	…	…

- **Many users are shared**
- **The grades of same users are correlated**
- **A user grades one movie around the same date in two databases**

- **IMDB users are public**

- **NetFLIX and IMdB moves are public**

Linking Attacks: Case of Netflix Prize

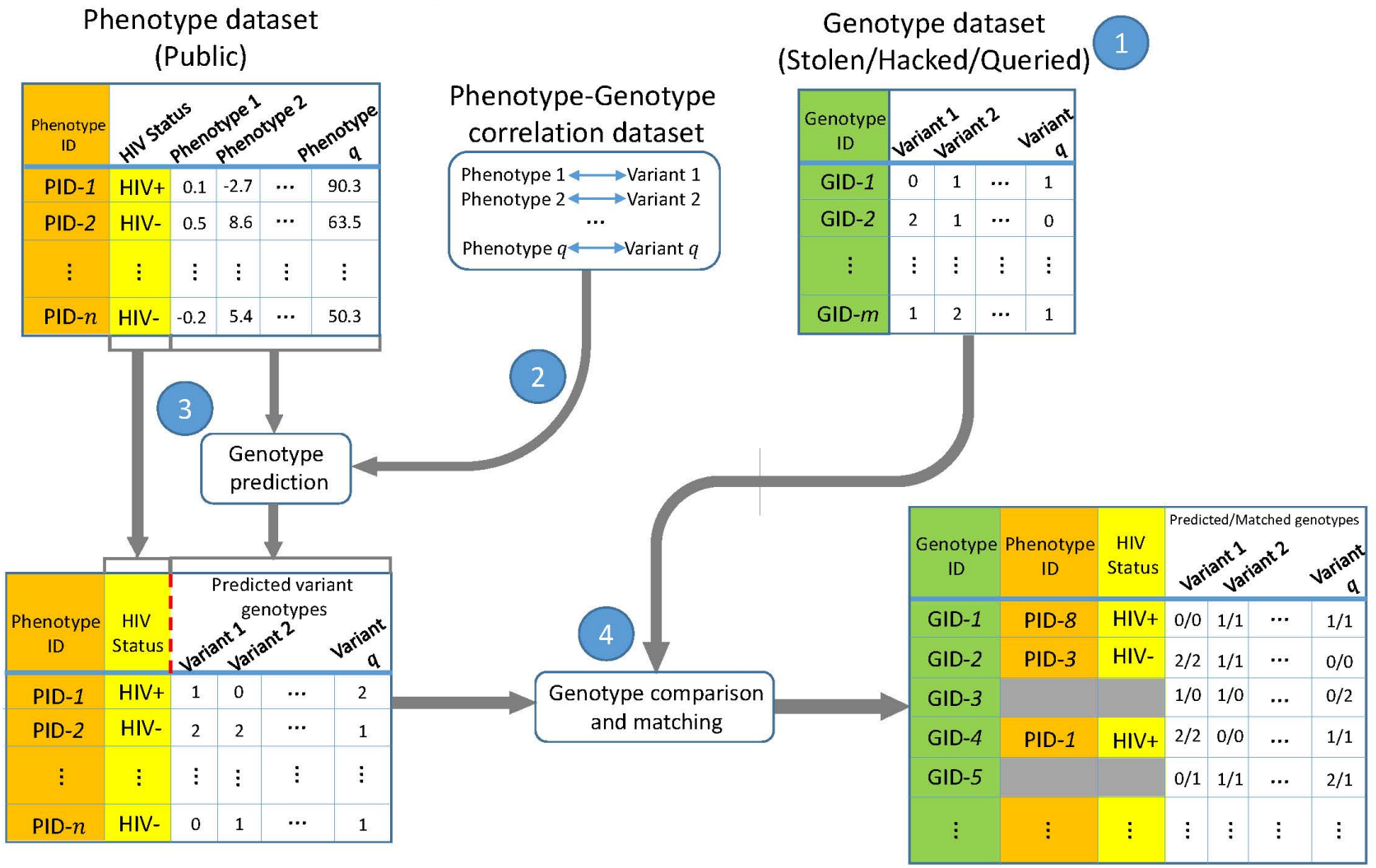
Names available for many users!

User (ID)	Movie (ID)	Date of Grade	Grade [1,2,3,4,5]
NTFLX-0	NTFLX-19	10/12/2008	1
NTFLX-1	NTFLX-116	4/23/2009	3
NTFLX-2	NTFLX-92	5/27/2010	2
NTFLX-1	NTFLX-666	6/6/2016	5
…	…	…	…
…	…	…	…

User (ID)	Movie (ID)	Date of Grade	Grade [0-10]
IMDB-0	IMDB-173	4/20/2009	5
IMDB-1	IMDB-18	10/18/2008	0
IMDB-2	IMDB-341	5/27/2010	-
…	…	…	…
…	…	…	…
…	…	…	…

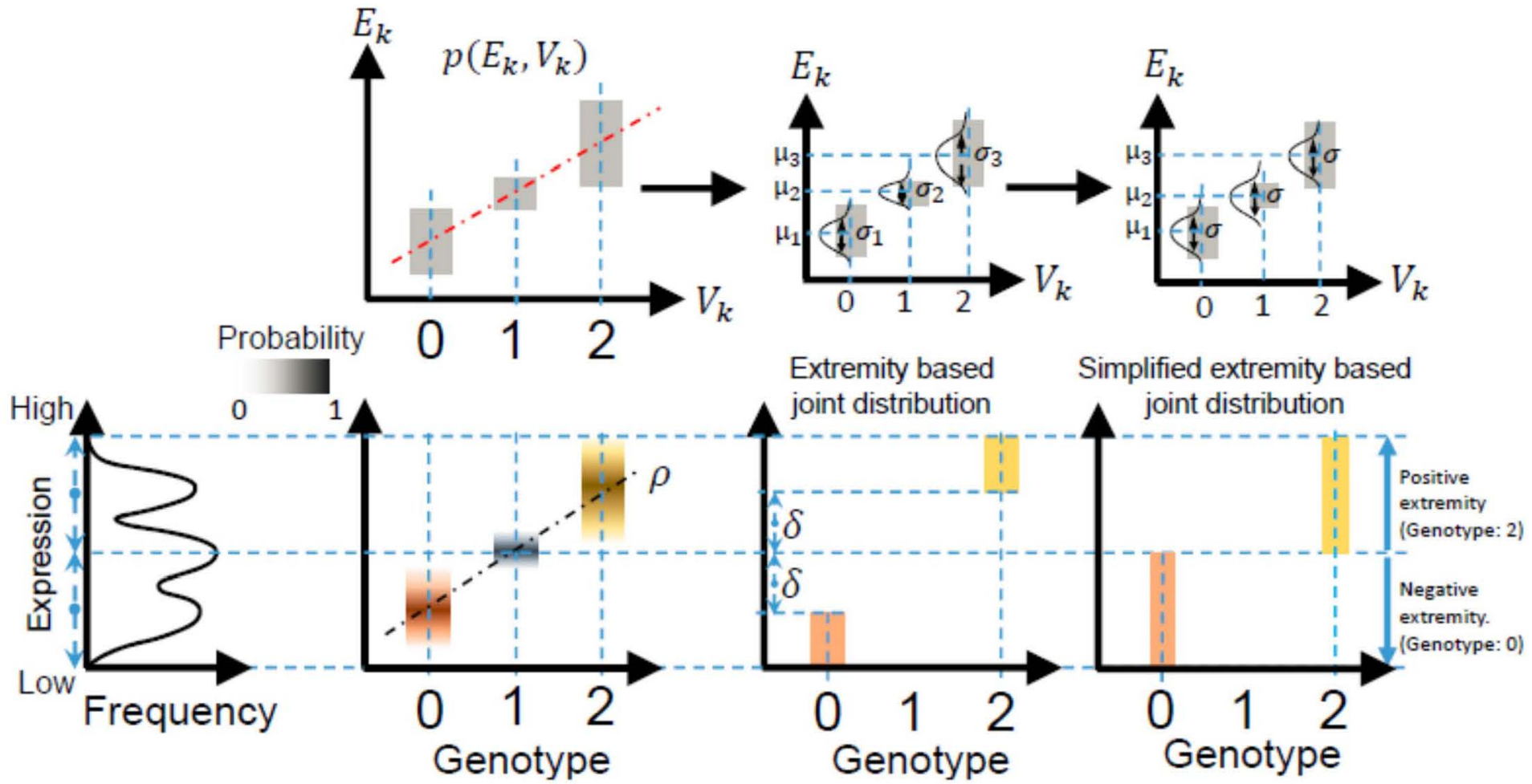
- **Many users are shared**
- **The grades of same users are correlated**
- **A user grades one movie around the same date in two databases**

Linking Attack Scenario



[Harmanciet al. Nat. Meth. (in revision)]

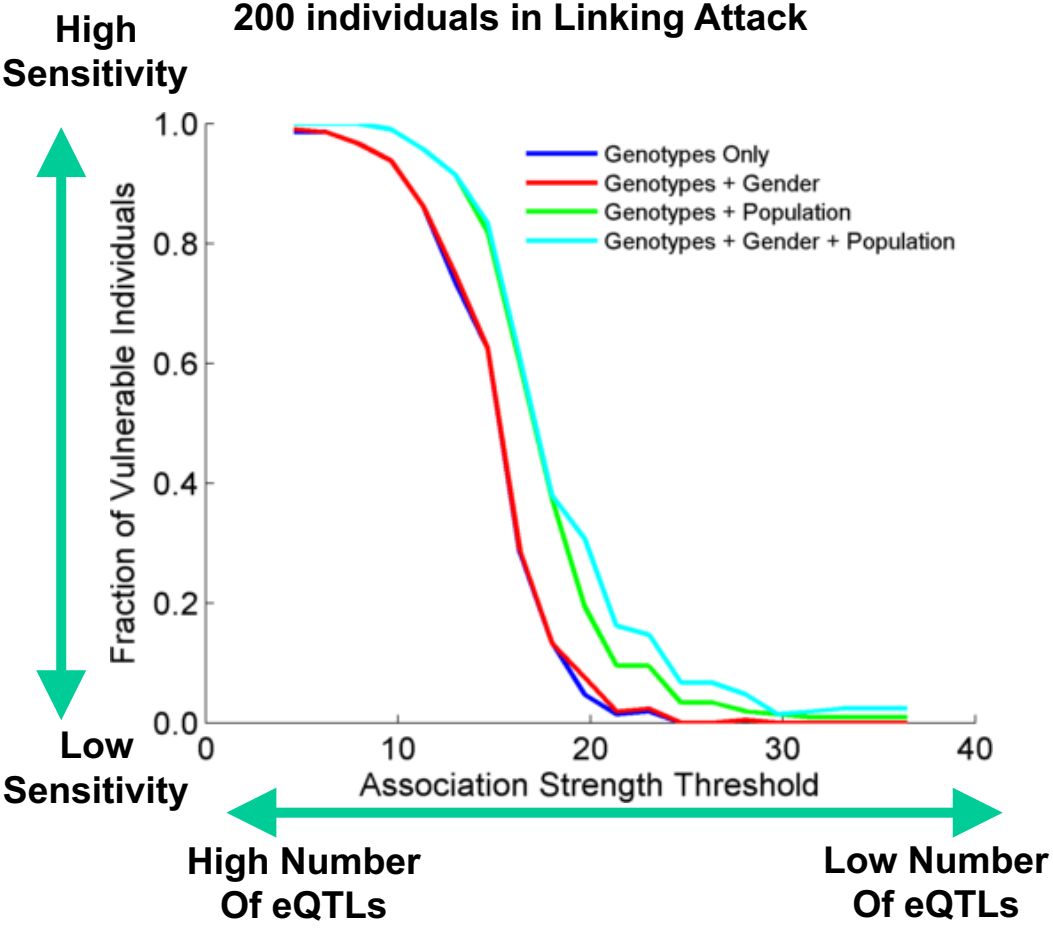
Levels of Expression-Genotype Model Simplifications for Genotype Prediction



[Harmanci et al. *Nat. Meth.* (16)]

Success in Linking Attack with Extremity based Genotype Prediction

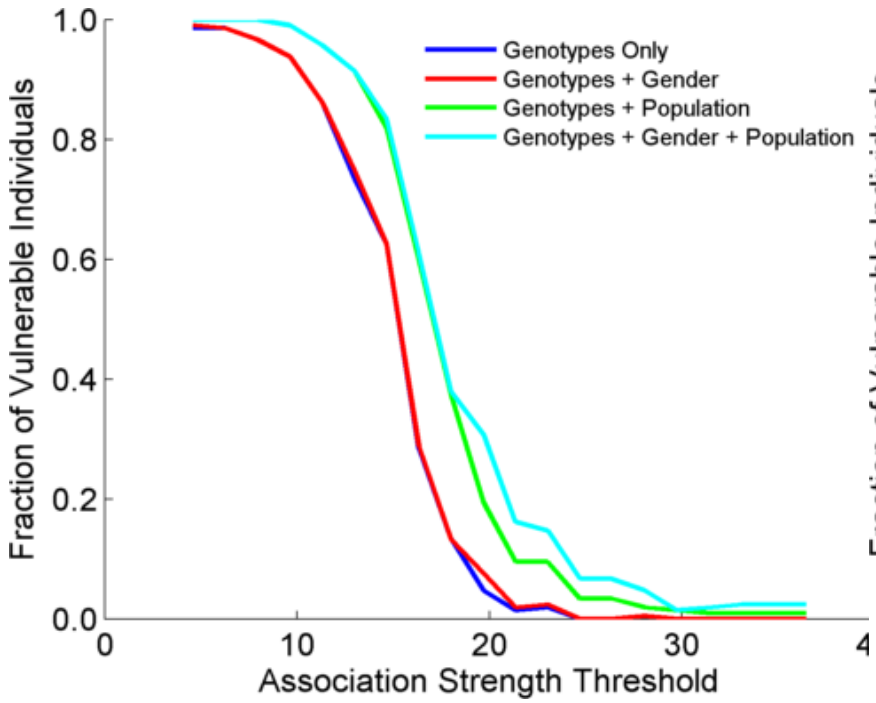
200 individuals eQTL Discovery
200 individuals in Linking Attack



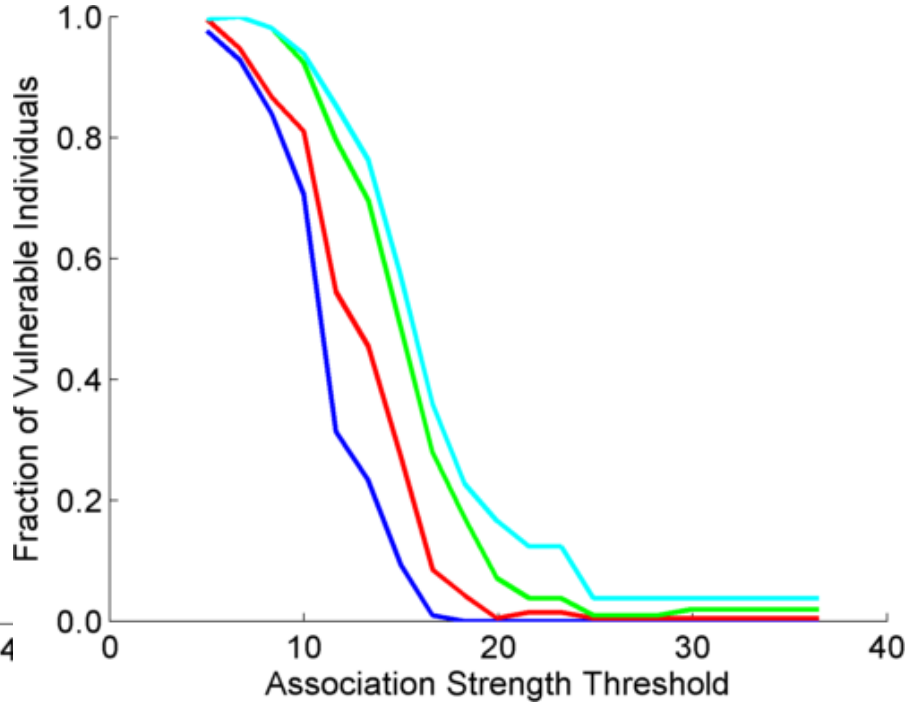
[Harmanci et al. *Nat. Meth.* (16)]

Success in Linking Attack with Extremity based Genotype Prediction

200 individuals eQTL Discovery
200 individuals in Linking Attack



200 individuals eQTL Discovery
100,200 individuals in Linking Attack

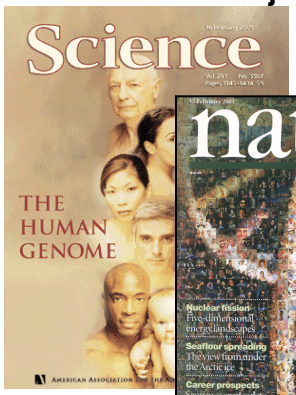


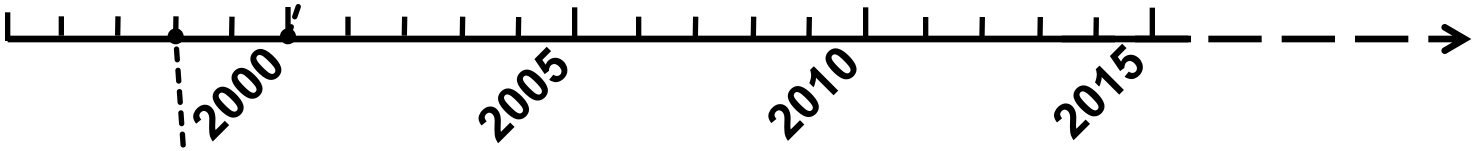
[Harmanci et al. *Nat. Meth.* (16)]

Transcriptome Mining: Tackling core issues related to gene regulation & also analyzing the "data exhaust" associated with this activity

- *[Core-1]* **Expression Clustering**, Cross-species
  - Comparative ENCODE – Lots of worm-fly-human matched data & developmental timecourses
  - Optimization gives 16 conserved co-expression modules, 12 w/ hourglass
- *[Core-2]* **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers; Decoupling expression changes into those from conserved vs species-specific genes
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- *[Core-3]* **Logic Gates** Modeling
  - Preponderance of OR gates in cancer v. cell-cycle (esp. for MYC)
- *[Exhaust-1]* **Genomic Privacy & RNA-seq**
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
  - Using file formats to remove obvious variants
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- *[Exhaust-2]* **Publication Patterns** from data producing consortia
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

The Human Genome Project



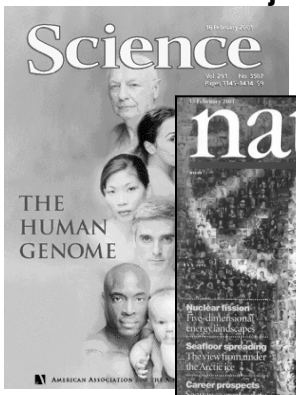


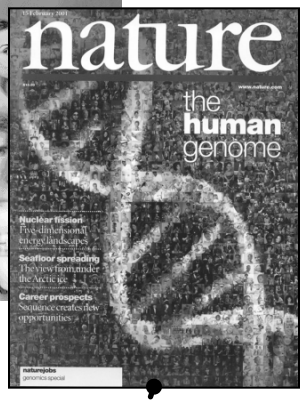
Worm Genome

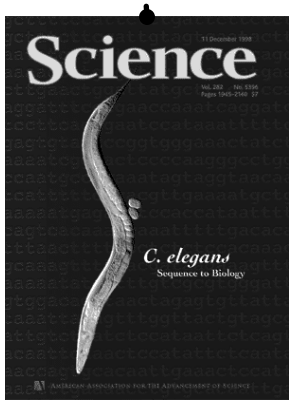
The Human Genome Project

ENCODE Pilot

ENCODE Production







Worm Genome

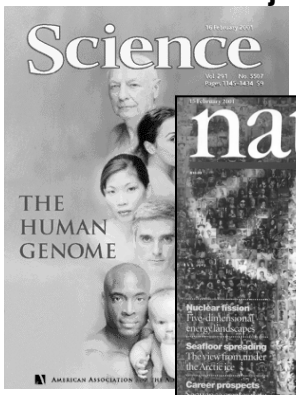
modENCODE

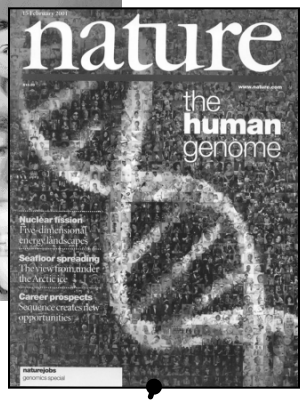
The Human Genome Project

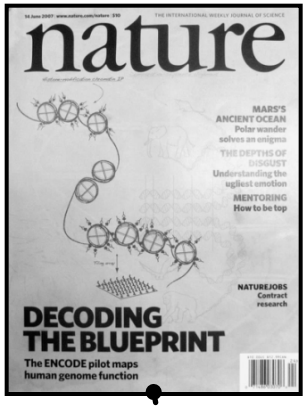
ENCODE Pilot

ENCODE Production

Comparative ENCODE





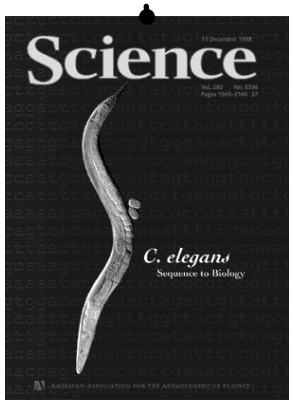


2000

2005

2010

2015



Worm Genome

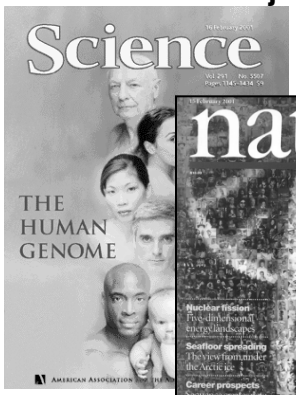
modENCODE

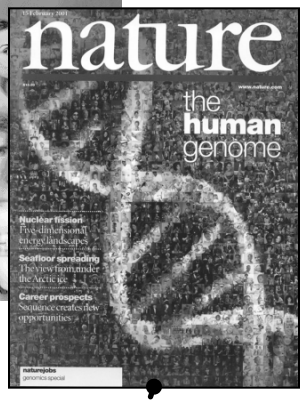
The Human Genome Project

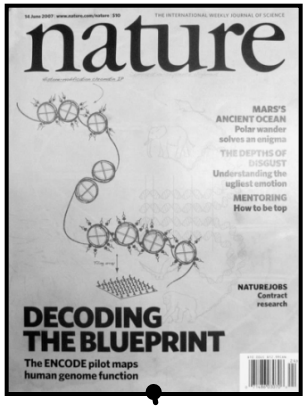
ENCODE Pilot

ENCODE Production

Comparative ENCODE





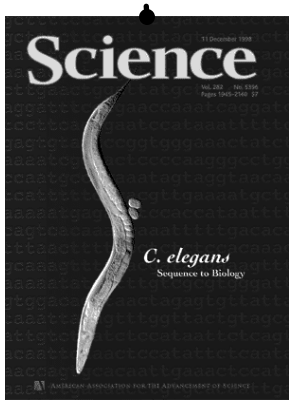


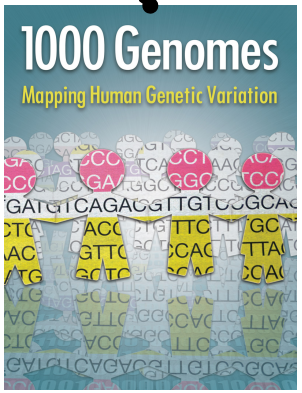
2000

2005

2010

2015





Worm Genome

modENCODE

1000 Genomes Pilot

1000 Genomes Production

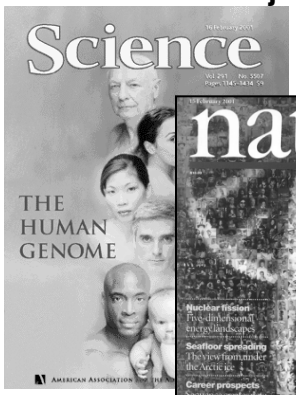
The Human Genome Project

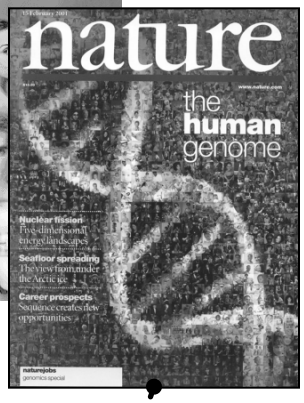
ENCODE Pilot

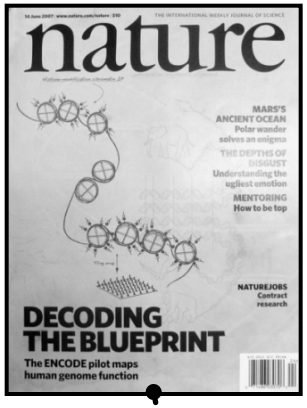
ENCODE Production

Comparative ENCODE

Epigenome Roadmap





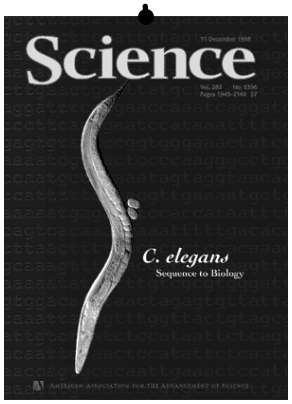


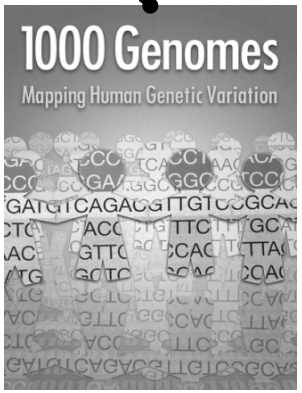
2000

2005

2010

2015





Worm Genome

modENCODE

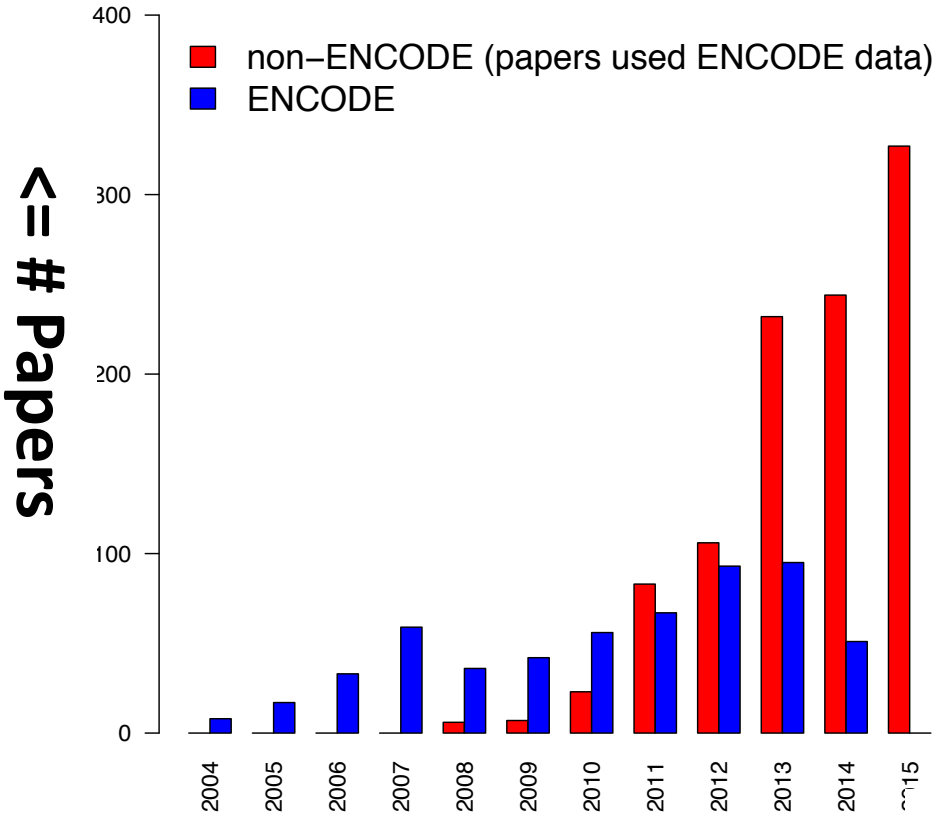
1000 Genomes Pilot

1000 Genomes Production

GTEx

With help of M Pazin at NHGRI, identified: 702 community papers that used ENCODE data but were not supported by ENCODE funding &
558 consortium papers supported by ENCODE funding
(https://www.encodeproject.org/search/?type=Publication for up-to-date query)
Then identified 1,786 ENCODE members & 8,263 non-members .

non–ENCODE (papers used ENCODE data) ENCODE



 # Authors

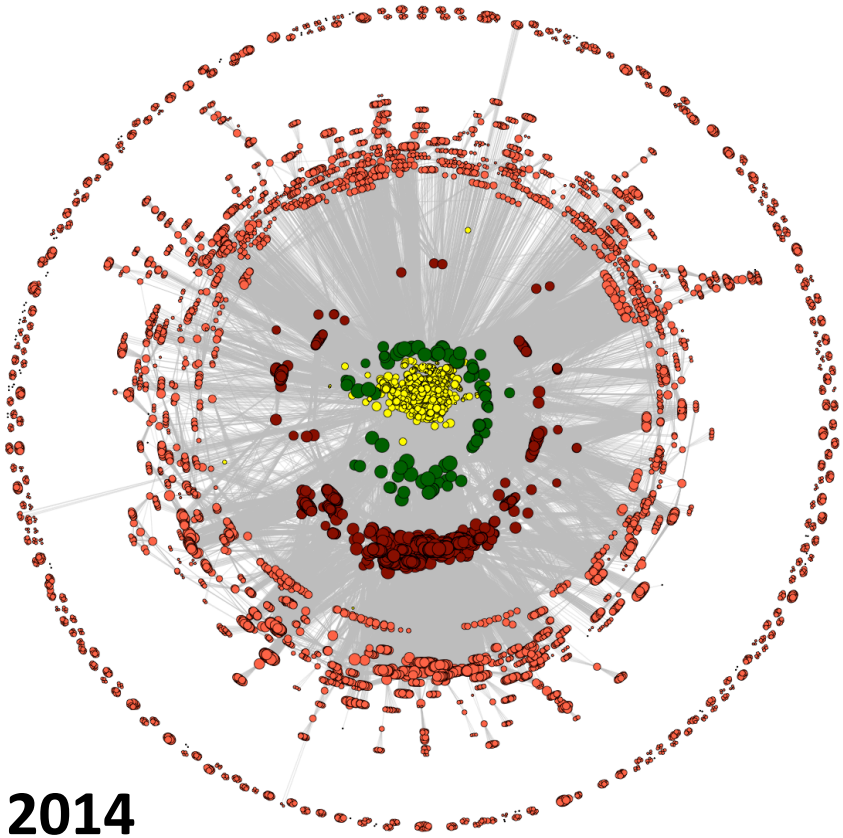
[Wang et al., TIG ('16)]

Yr. ('04 to '15)

Co-authorship Network of ENCODE members & Data Users

- ENCODE member
- non-member
- ENCODE member broker
- non-member broker

co-authorship



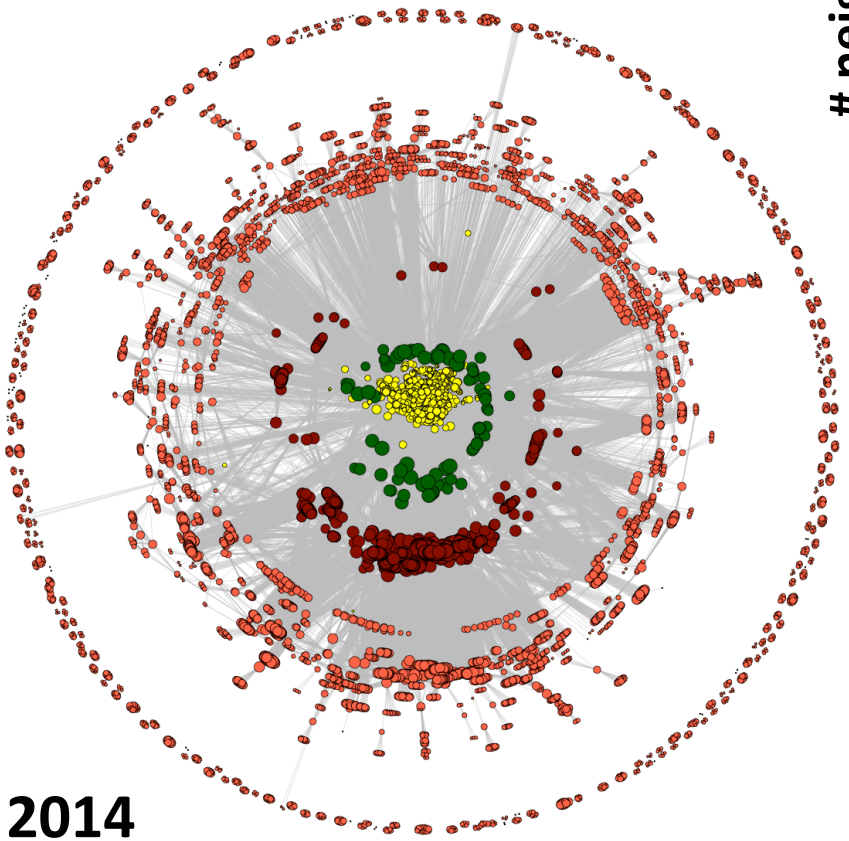
2014

[Wang et al., TIG ('16)]

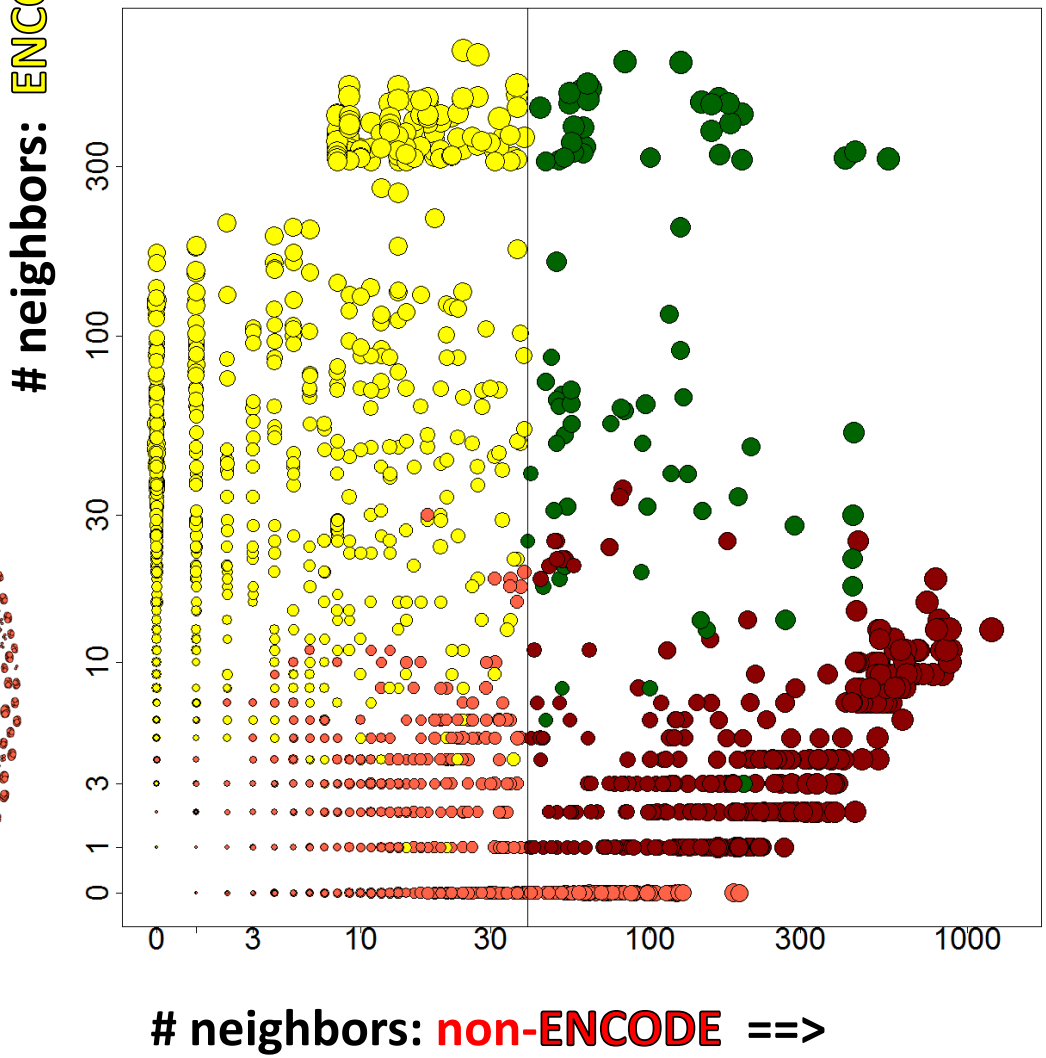
Co-authorship Network of ENCODE members & Data Users

- ENCODE member
- non-member
- ENCODE member broker
- non-member broker

co-authorship



2014

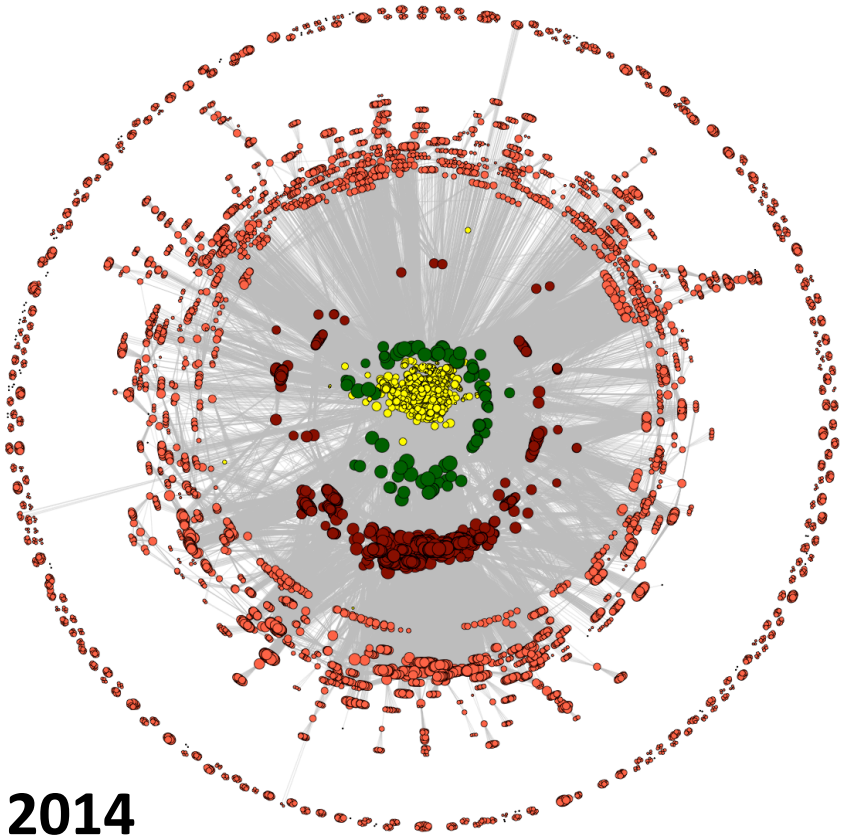


[Wang et al., TIG ('16)]

Co-authorship Network of ENCODE members & Data Users

- ENCODE member
- non-member
- ENCODE member broker
- non-member broker

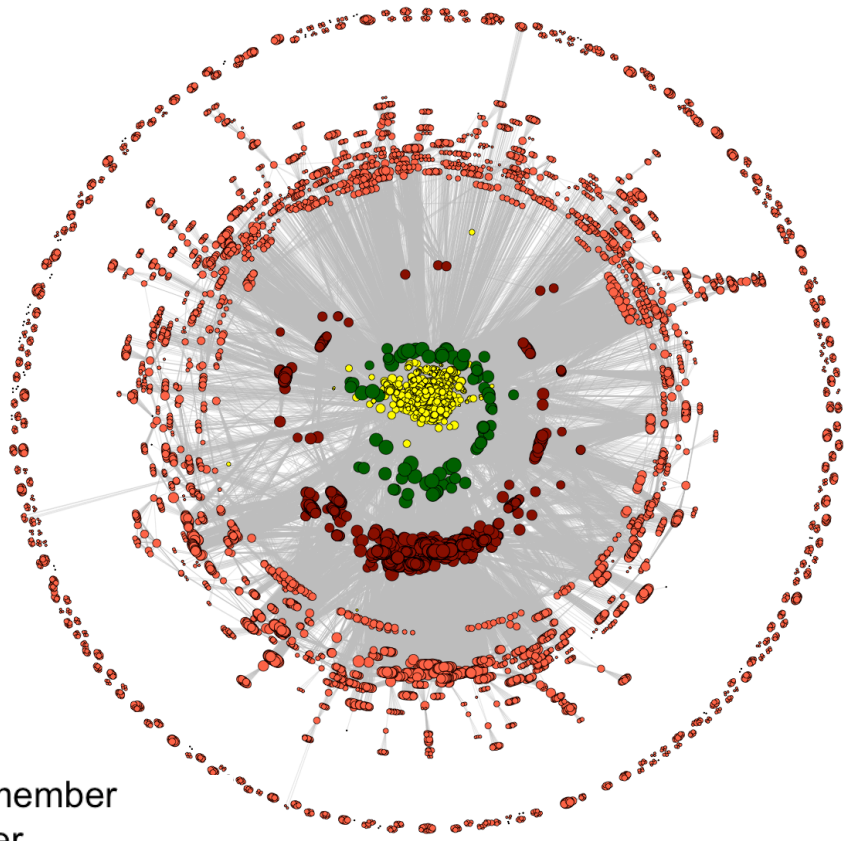
co-authorship



2014

[Wang et al., TIG ('16)]

Dynamics of co-authorship network

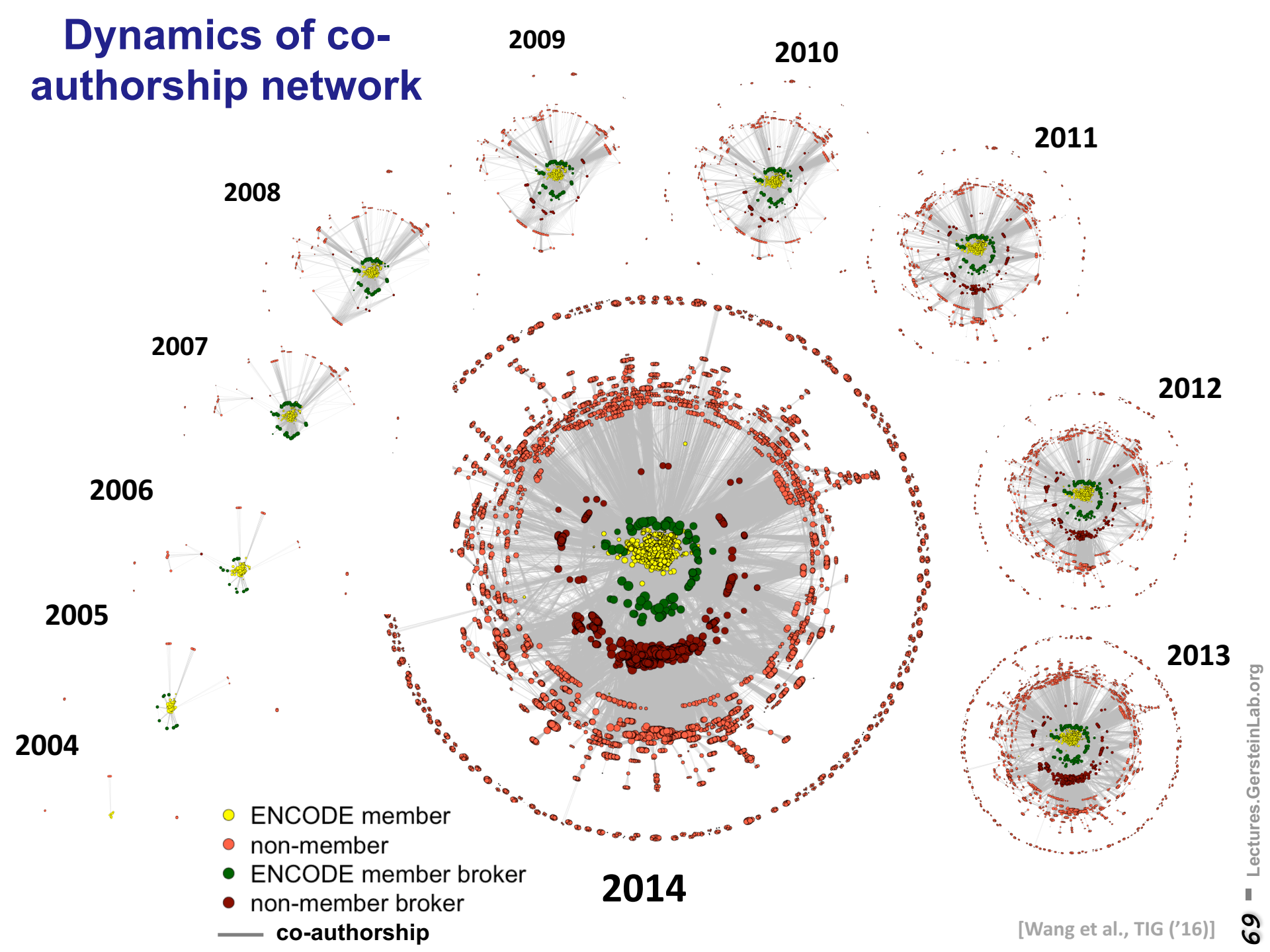


ENCODE member
non-member
ENCODE member broker
non-member broker
co-authorship

2014

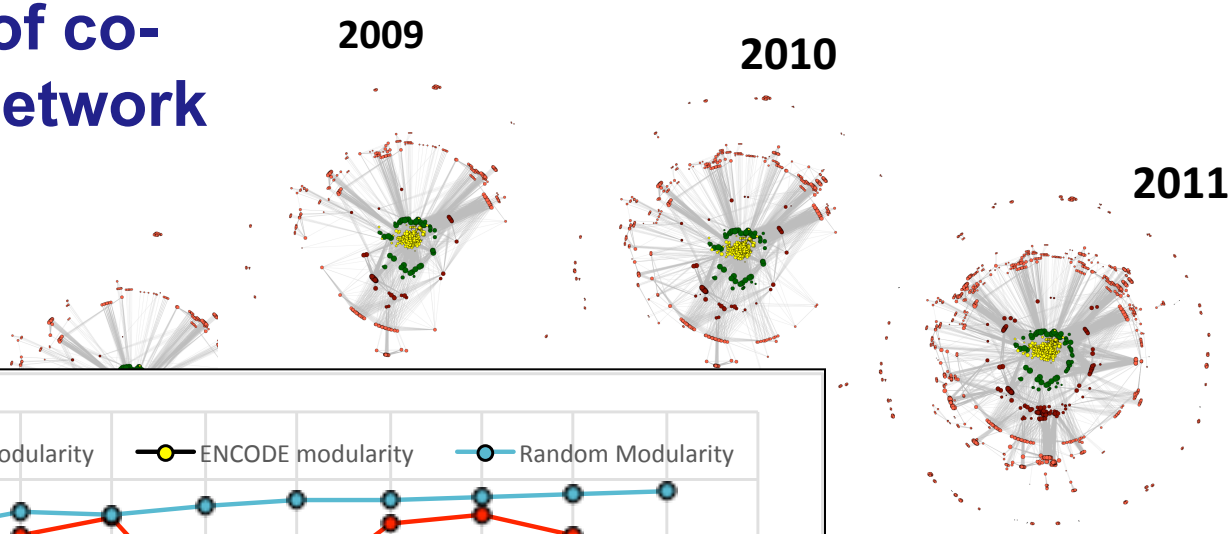
[Wang et al., TIG ('16)]

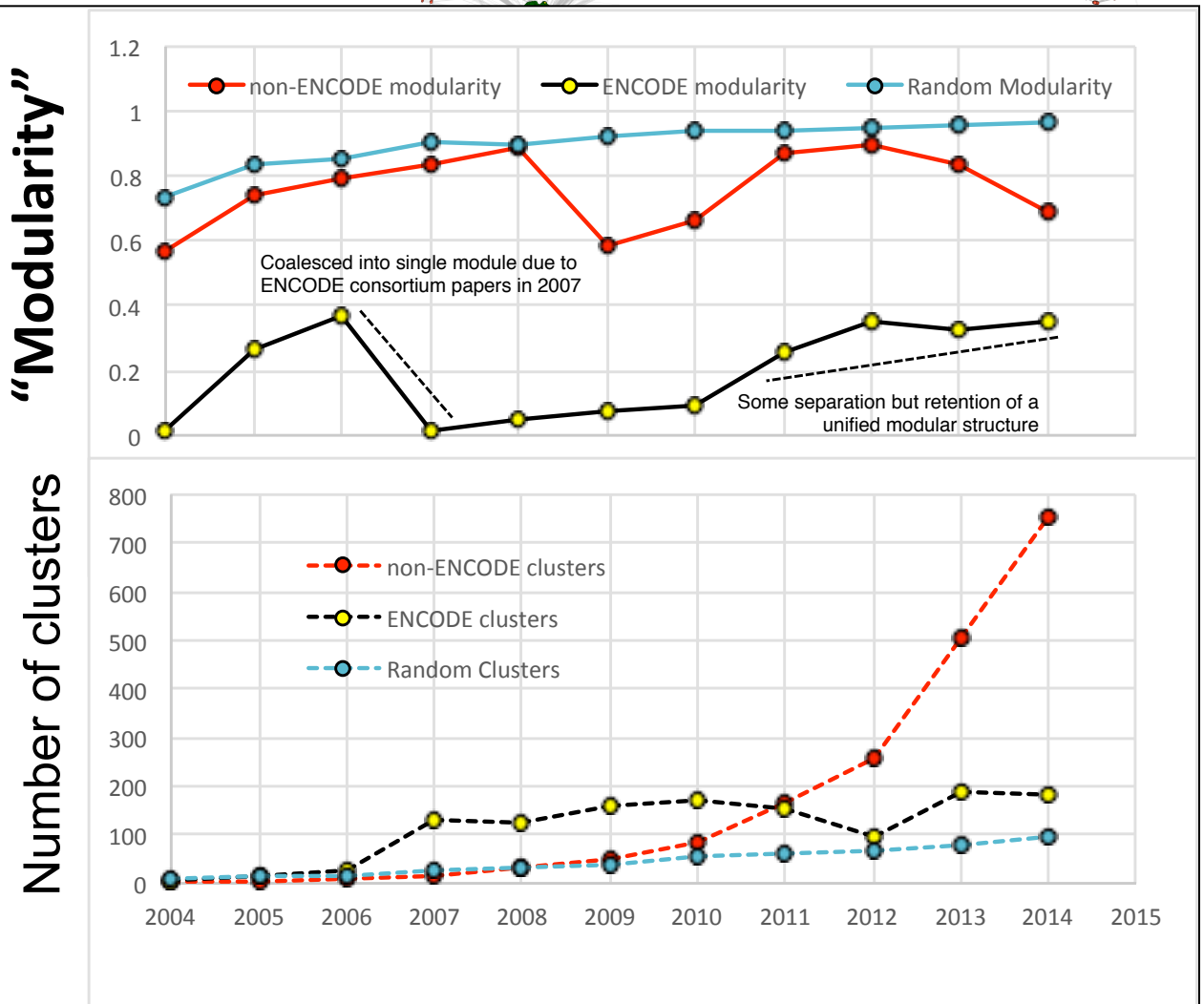
Dynamics of co-authorship network



[Wang et al., TIG ('16)]

Dynamics of co-authorship network





[Wang et al., TIG ('16)]

Transcriptome Mining: Tackling core issues related to gene regulation & also analyzing the "data exhaust" associated with this activity

- *[Core-1]* **Expression Clustering**, Cross-species
  - Comparative ENCODE – Lots of worm-fly-human matched data & developmental timecourses
  - Optimization gives 16 conserved co-expression modules, 12 w/ hourglass
- *[Core-2]* **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers; Decoupling expression changes into those from conserved vs species-specific genes
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- *[Core-3]* **Logic Gates** Modeling
  - Preponderance of OR gates in cancer v. cell-cycle (esp. for MYC)
- *[Exhaust-1]* **Genomic Privacy & RNA-seq**
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
  - Using file formats to remove obvious variants
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- *[Exhaust-2]* **Publication Patterns** from data producing consortia
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

Transcriptome Mining: Tackling core issues related to gene regulation & also analyzing the "data exhaust" associated with this activity

- *[Core-1]* **Expression Clustering**, Cross-species
  - Comparative ENCODE – Lots of worm-fly-human matched data & developmental timecourses
  - Optimization gives 16 conserved co-expression modules, 12 w/ hourglass
- *[Core-2]* **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers; Decoupling expression changes into those from conserved vs species-specific genes
  - Conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)
- *[Core-3]* **Logic Gates** Modeling
  - Preponderance of OR gates in cancer v. cell-cycle (esp. for MYC)
- *[Exhaust-1]* **Genomic Privacy** & RNA-seq
  - The dilemma: The genome as fundamental, inherited info that's very private v need for large-scale mining for med. research
  - 2-sided nature of RNA-seq presents a particularly tricky privacy issue
  - Using file formats to remove obvious variants
  - Quantifying & removing further variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
- *[Exhaust-2]* **Publication Patterns** from data producing consortia
  - Co-authorship network statistics relate to publication rollouts & show gradual adoption by a diverse community
  - Key role of brokers in data dissemination

Acknowledgements: ENCODE/modENCODE Transcriptome Group

Joel Rozowsky, Koon-Kiu Yan, Daifeng Wang, Chao Cheng, James B. Brown, Carrie A. Davis, LaDeana Hillier, Cristina Sisu, **Jingyi Jessica Li,** Baikang Pei, Arif O. Harmanci, Michael O. Duff, Sarah Djebali, Roger P. Alexander, Burak H. Alver, Raymond K. Auerbach, Kimberly Bell, Peter J. Bickel, Max E. Boeck, Nathan P. Boley, Benjamin W. Booth, Lucy Cherbas, Peter Cherbas, Chao Di, Alex Dobin, Jorg Drenkow, Brent Ewing, Gang Fang, Megan Fastuca, Elise A. Feingold, Adam Frankish, Guanjun Gao, Peter J. Good, Phil Green, Roderic Guigó, Ann Hammonds, Jen Harrow, Roger A. Hoskins, Cédric Howald, Long Hu, Haiyan Huang, Tim J. P. Hubbard, Chau Huynh, Sonali Jha, Dionna Kasper, Masaomi Kato, Thomas C. Kaufman, Rob Kitchen, Erik Ladewig, Julien Lagarde, Eric Lai, Jing Leng, **Zhi Lu,** Michael MacCoss, Gemma May, Rebecca McWhirter, Gennifer Merrihew, David M. Miller, Ali Mortazavi, Rabi Murad, Brian Oliver, Sara Olson, Peter Park, Michael J. Pazin, Norbert Perrimon, Dmitri Pervouchine, Valerie Reinke, Alexandre Reymond, Garrett Robinson, Anastasia Samsonova, Gary I. Saunders, Felix Schlesinger, Anurag Sethi, Frank J. Slack, William C. Spencer, Marcus H. Stoiber, Pnina Strasbourger, Andrea Tanzer, Owen A. Thompson, Kenneth H. Wan, Guilin Wang, Huaien Wang, Kathie L. Watkins, Jiayu Wen, Kejia Wen, Chenghai Xue, Li Yang, Kevin Yip, Chris Zaleski, Yan Zhang, Henry Zheng,

Steven E. Brenner, Brenton R. Graveley, Susan E. Celniker, Thomas R Gingeras, Robert Waterston

Hiring Postdocs. See gersteinlab.org/jobs !

EncodeProject.org/comparative/transcriptome

Acknowledgements

DREISS.gersteinlab.org

D **Wang**, F He, S Maslov

papers.gersteinlab.org/subject/privacy –

D **Greenbaum**

Loregic.gersteinlab.org

D **Wang**, KK Yan, C Sisu, C Cheng, J Rozowsky, W Meyerson

PrivaSeq.gersteinlab.org

A **Harmanci**, G Gürsoy, F Navarro

github.com/gersteinlab/**OrthoClust**

K **Yan**, D Wang, J Rozowsky, H Zheng, C Cheng

Publication patterns [“encode authors”]

D **Wang**, KK Yan, J Rozowsky, E Pan

Hiring Postdocs. See JOBS.gersteinlab.org !

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2017.
  - Please read permissions statement at www.**gersteinlab.org/misc/permissions.html** .
  - Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
  - Paper references in the talk were mostly from Papers.GersteinLab.org.

- PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info .
  - In particular, many of the images have particular EXIF tags, such as kwpotppt , that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt