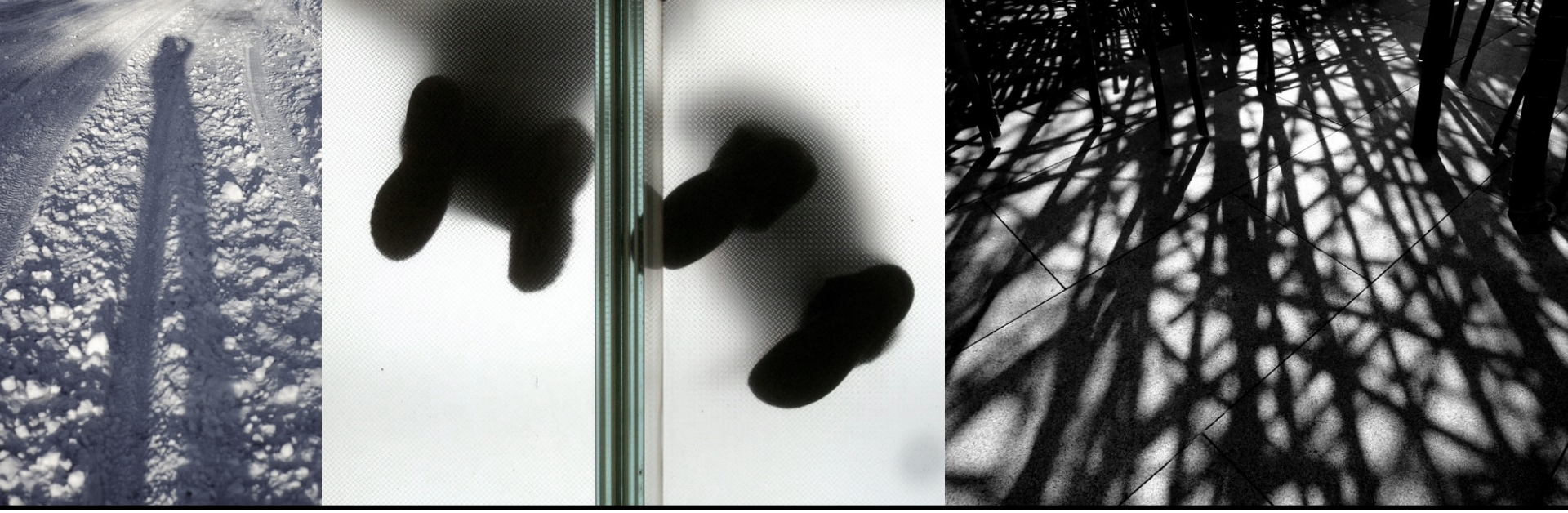


Large-scale Transcriptome Mining:
Building Integrative Regulatory Models, while Protecting Individual Privacy

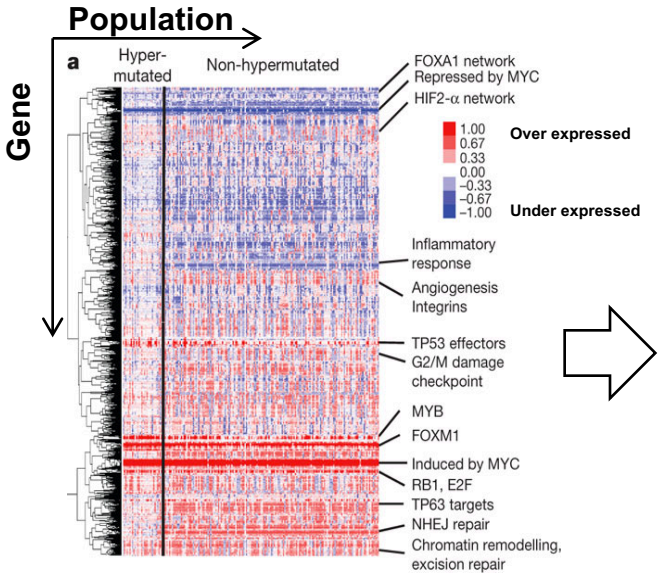
Mark Gerstein, Yale

Slides freely downloadable from Lectures.GersteinLab.org & “tweetable” (via @markgerstein)

See last slide for more info.

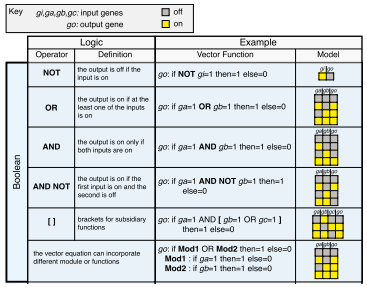
Modeling for RNA-seq data across many samples & individuals… while still protecting individual privacy

* Recent advent of much large scale RNA-seq (& other functional genomics data) following on DNA sequencing
* Often this is of human subjects & produced by large consortia (eg TCGA, PCAWG, GTEx) and needs to be protected
* Useful to build tools & approaches that interact with these data

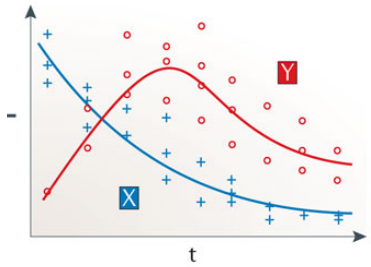


The Cancer Genome Atlas Network Nature 487, 330-337 (2012) doi:10.1038/nature11252

- Logical model



- Continuous model



$$\frac{dx_i}{dt} = \sum_{j=1}^{n} a_{i,j} x_j$$

- Probabilistic model

- Gene Regulatory Mechanisms

Istrail & Davidson, PNAS, '04 Nicolas Le Novère, Nature Reviews Genetics, '15

The Human Genome Project

ENCODE Pilot

ENCODE Production

Comparative ENCODE

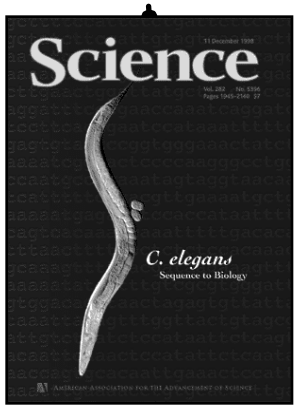
Epigenome Roadmap

2000

2005

2010

2015



Worm Genome

modENCODE

1000 Genomes Pilot

1000 Genomes Phase 3

GTEx

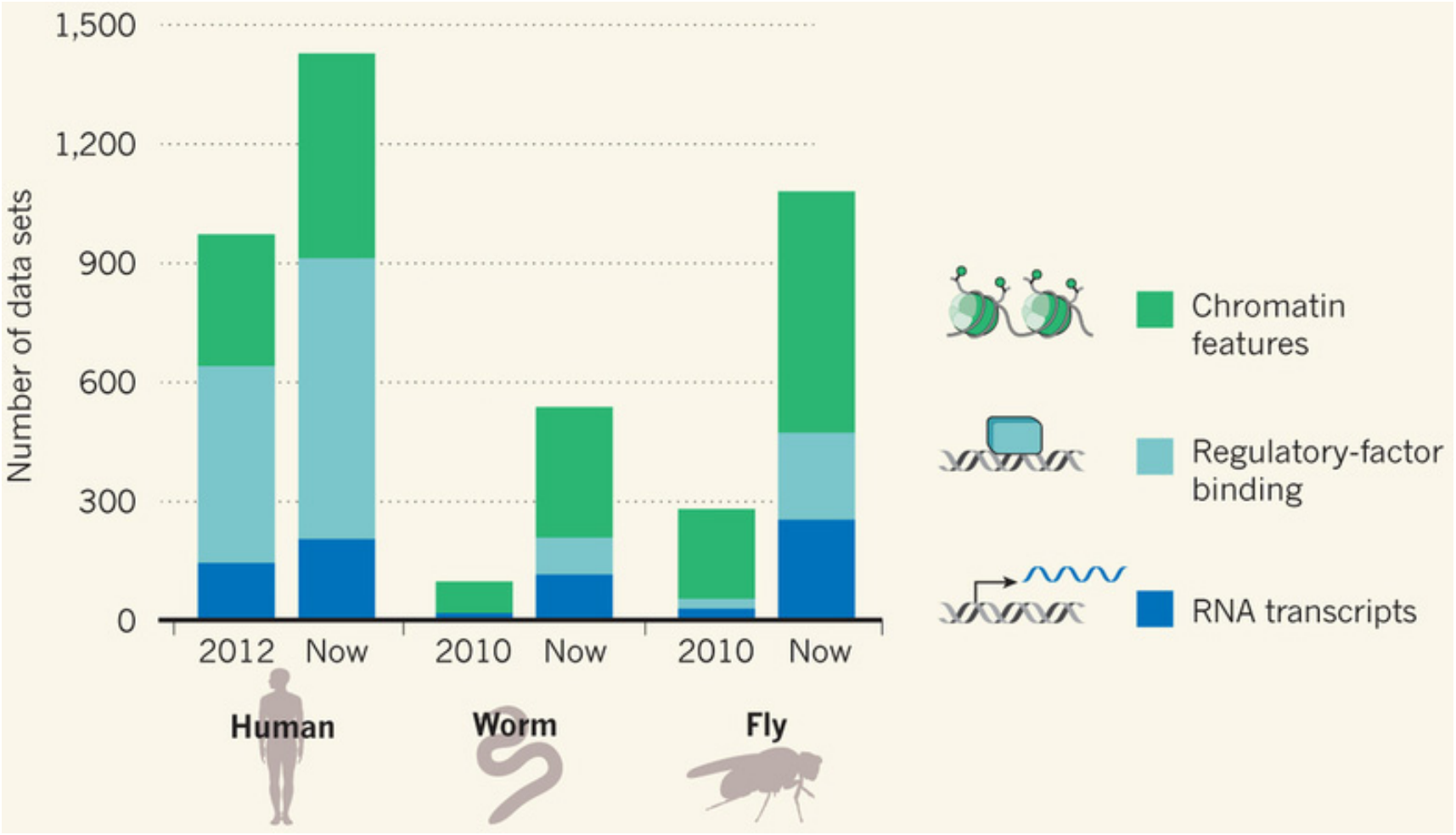
2-sided nature of functional genomics data: Analysis can be very General/Public or Individual/Private

- General quantifications related to overall aspects of a condition & are not tied to an individual's genotype - ie what genes go up in cancer
  - However, data is derived from an individual & tagged with an individual's genotype
- Other calculations aim to use genotype & specific aspects of the quantification to derive general relations related to sequence variation & gene expression
- Some calculations and data derive finding very specific to the variants in a particular individual

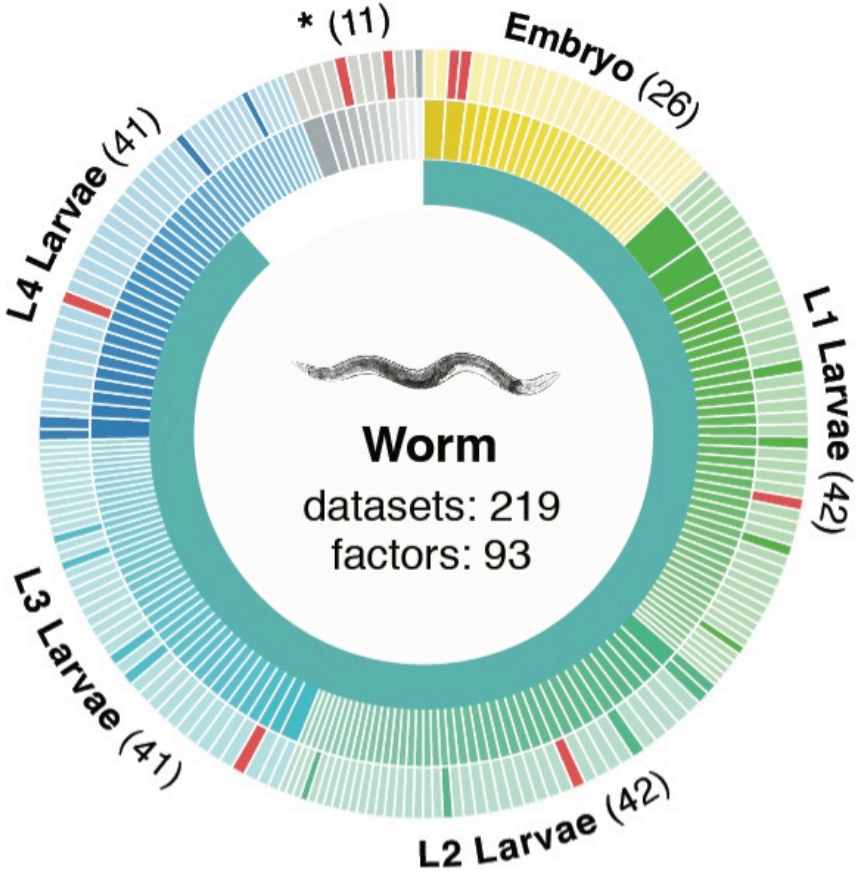
Comparative ENCODE Functional Genomics Resource

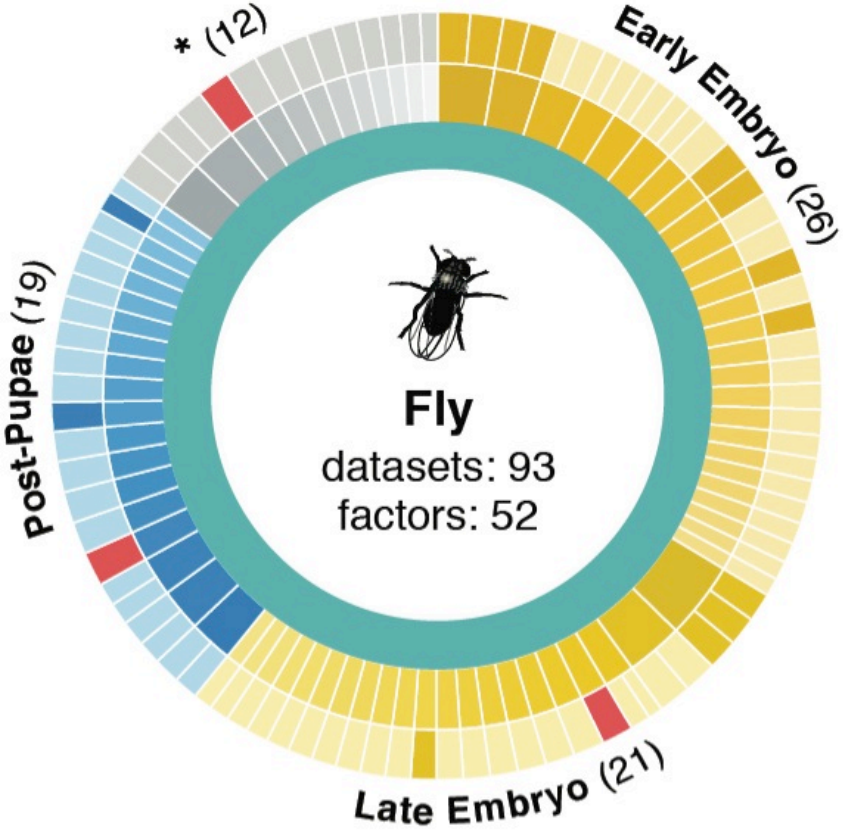
(EncodeProject.org/comparative)

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



Time-course gene expression data of worm & fly development



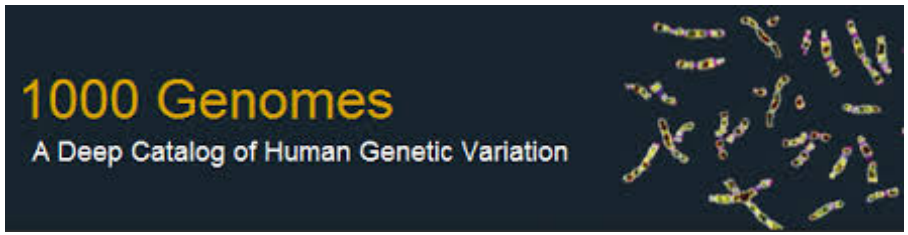


Organism	Major developmental stages
worm (*C. elegans*)	33 stages: 0, 0.5, 1, …, 12 hours, L1, L2, L3, L4, …, Young Adults, Adults
fly (*D. mel.*)	30 stages: 0, 2, 4, 6, 8,…, 20, 22 hours, L1-L4, Pupaes, Adults

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

Representative Expression, Genotype, eQTL Datasets

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



Large-scale Transcriptome Mining: Building Integrative Regulatory Models, while Protecting Individual Privacy

- Transcriptome analysis **data**
  - Comparative ENCODE – Lots of Matched Data
  - 1000G+Geuvadis for privacy
- **Expression Clustering**, Cross-species
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones.
  - Also, conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

- The General **Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies & burdensome security
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
  - Details on Relevant Hacks: Genomic, Computer Security, & Netfix

- **RNA-seq: How to Publicly Share Some of it**
  - Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels +
    eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
  - Quantifying accuracy of prediction, via gap between best & 2nd best match

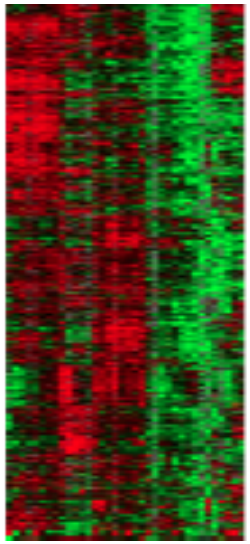
Large-scale Transcriptome Mining: Building Integrative Regulatory Models, while Protecting Individual Privacy

- Transcriptome analysis **data**
  - Comparative ENCODE – Lots of Matched Data
  - 1000G+Geuvadis for privacy
- **Expression Clustering**, Cross-species
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones.
  - Also, conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

- The General **Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies & burdensome security
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
  - Details on Relevant Hacks: Genomic, Computer Security, & Netfix

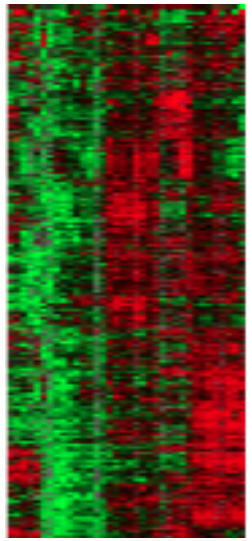
- **RNA-seq: How to Publicly Share Some of it**
  - Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
  - Quantifying accuracy of prediction, via gap between best & 2nd best match

Expression clustering: revisiting an ancient problem



Species A

Species B

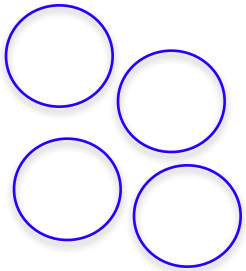


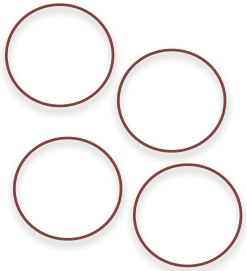
Clustering algorithm

Clustering algorithm

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

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

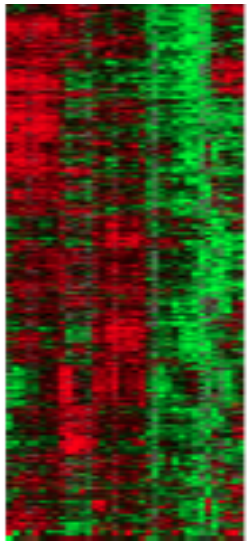




two independent sets of modules

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

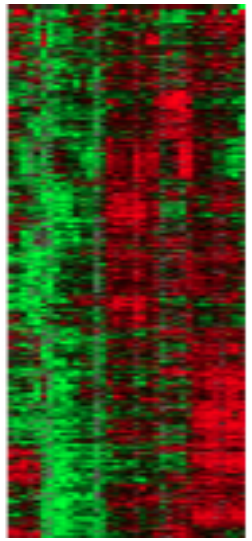
Expression clustering: revisiting an ancient problem



Species A

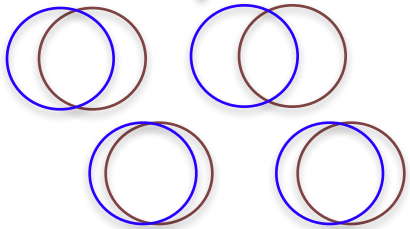
Orthologous pairs between species

Species B



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

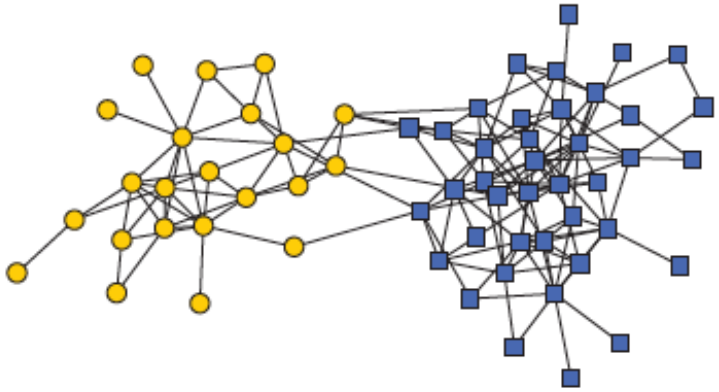
OrthoClust



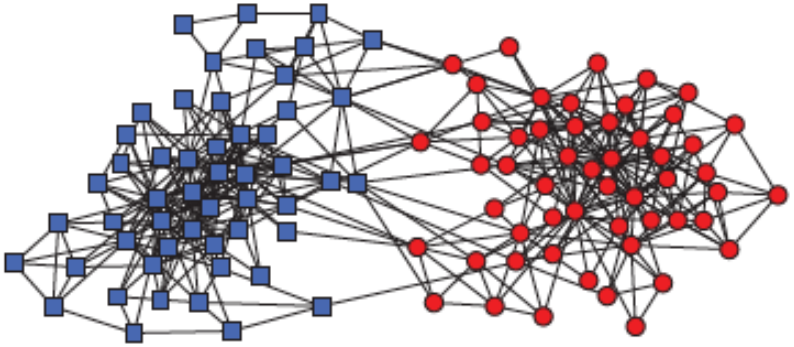
cross species modules

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

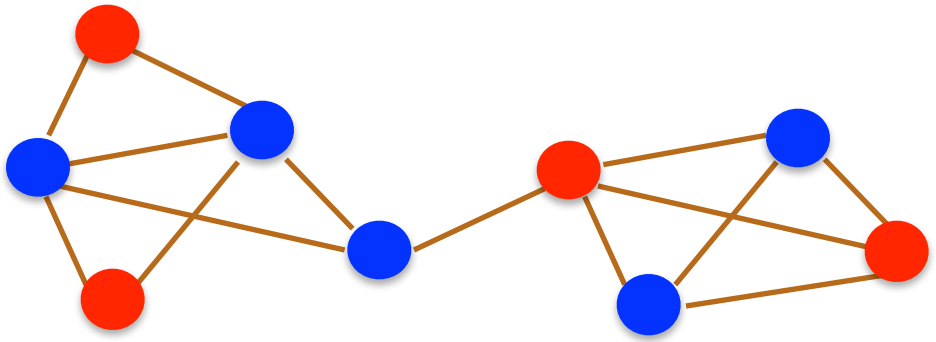
degree of node i

number of edges

expected number of edges between i and j

whether or not i, j are in the same module

Network modularity



$$Q \approx 0$$

degree of node i

adjacency matrix

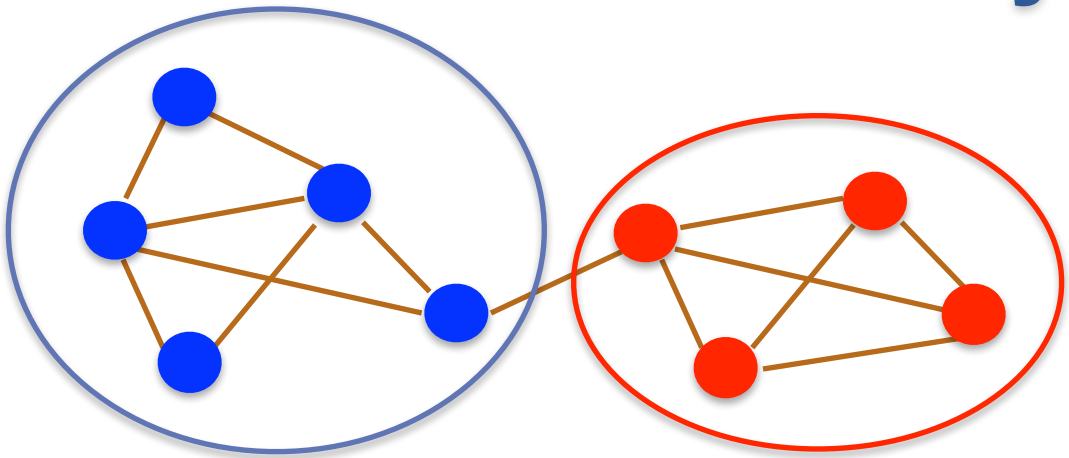
$$Q = \frac{1}{2m} \sum_{i,j} \left(W_{ij} - \frac{k_i k_j}{2m} \right) \delta_{\sigma_i \sigma_j}$$

whether or not i, j are in the same module

number of edges

expected number of edges between i and j

Network modularity



$$Q = Q_{max}$$

Optimization problem

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

adjacency matrix

degree of node i

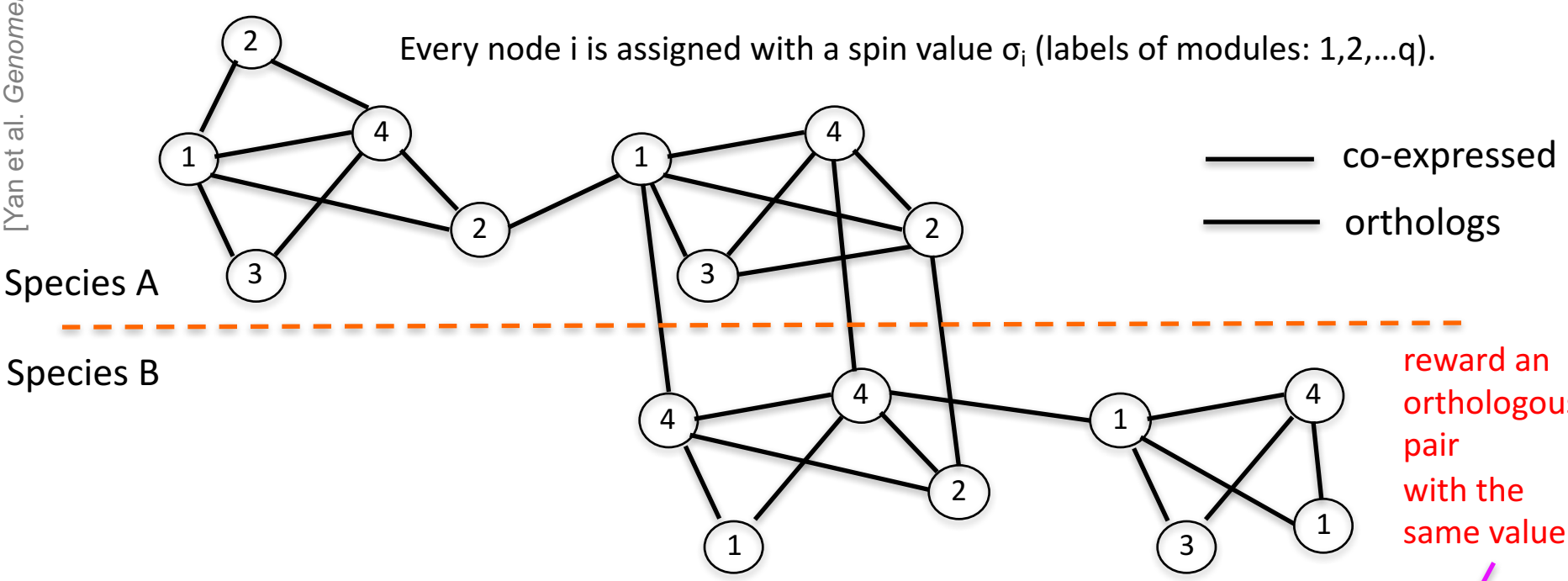
number of edges

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 spin value σ_i (labels of modules: 1,2,…q).



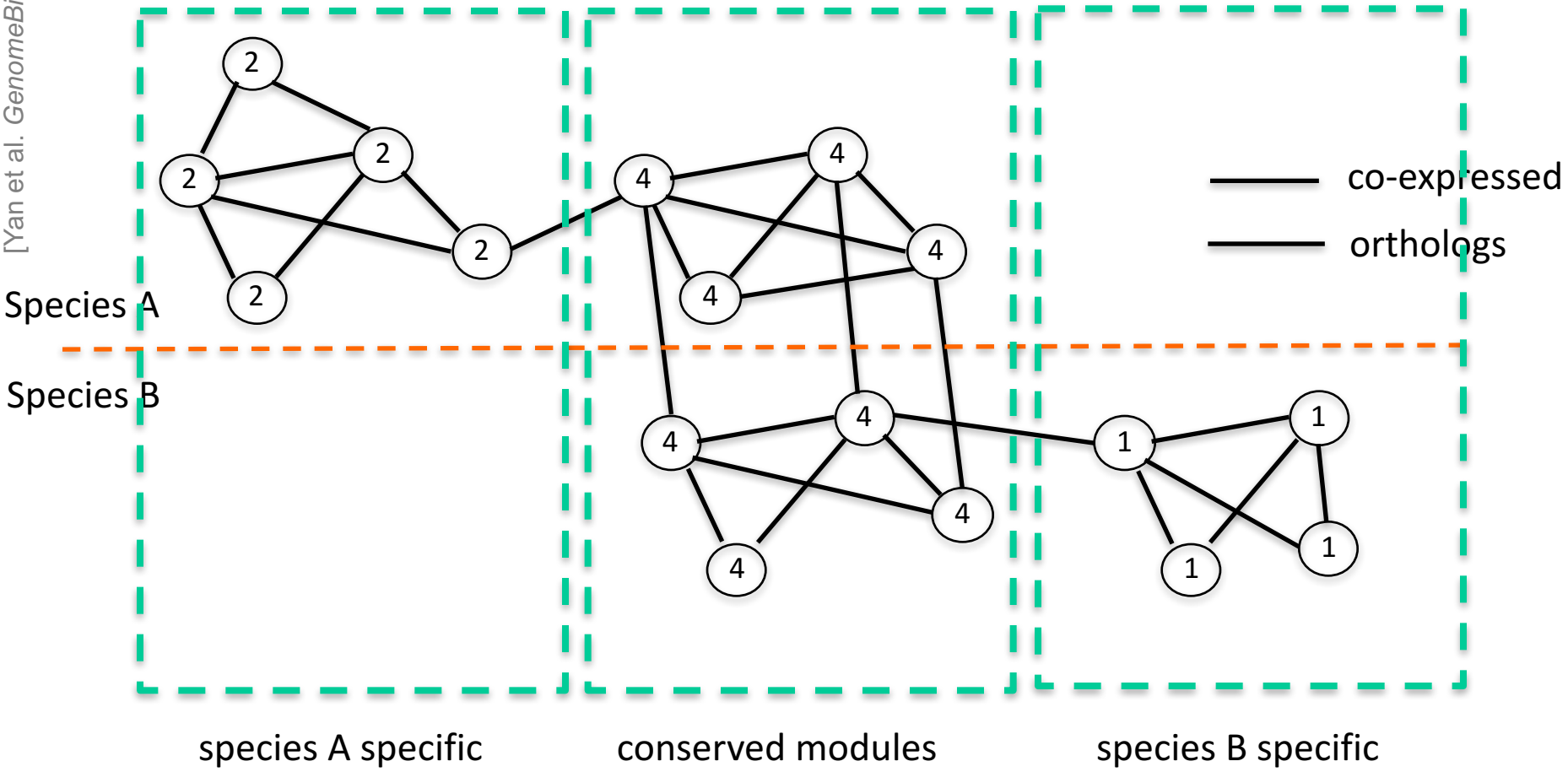
$$H = \sum_{i,j}\left(-W_{ij}^{(A)} + p_{ij}^{(A)}\right)\delta_{\sigma_i\sigma_j} + \sum_{i',j'}\left(-W_{i'j'}^{(B)} + p_{i'j'}^{(B)}\right)\delta_{\sigma_{i'}\sigma_{j'}} - \kappa \sum_{(i,j')\in Ortho}\delta_{\sigma_i\sigma_{j'}}$$

reward a co-expressed pair with the same value

punish a non co-expressed pair with the same value

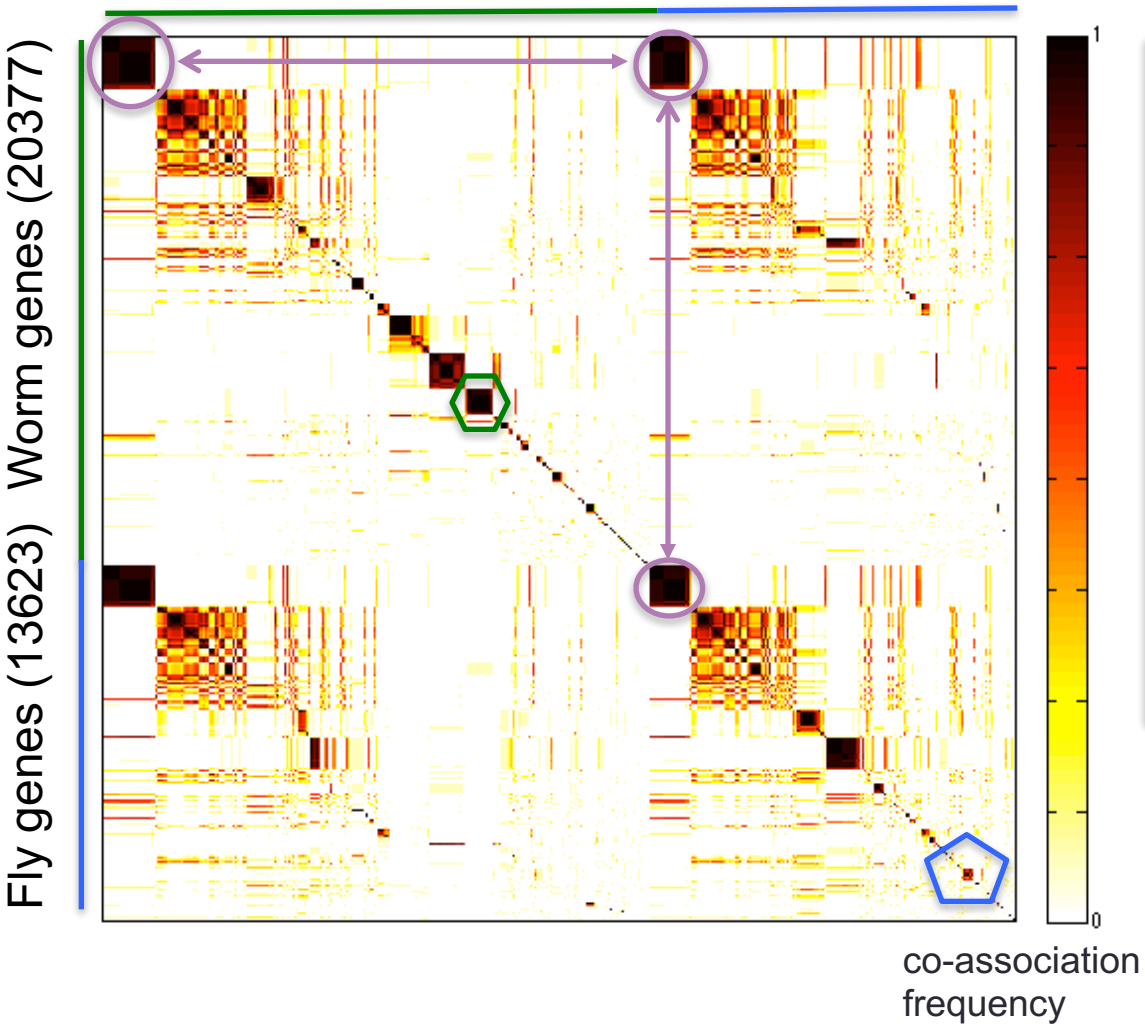
Favorableness = "Modularity" in species A + "Modularity" in species B + consistency betw. A & B

A toy example [orthoclust]

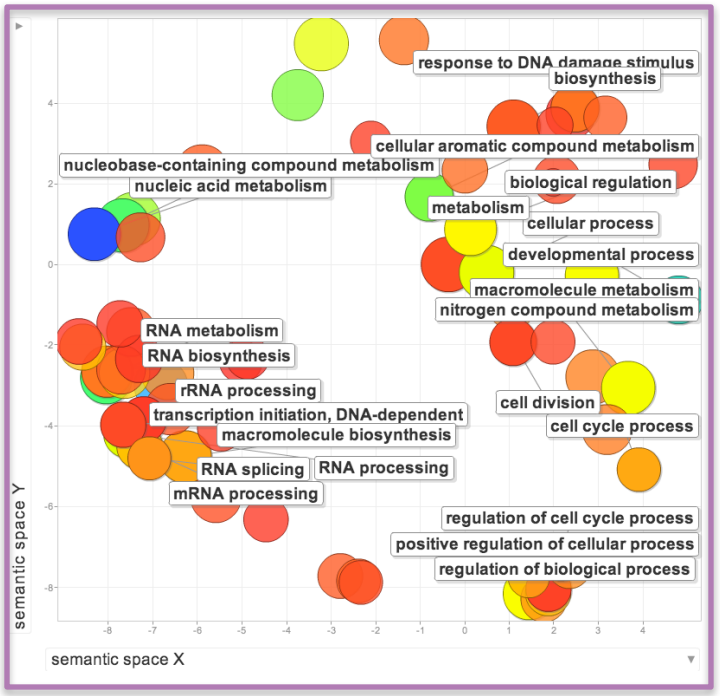


Use Potts model (generalized Ising model) to simultaneously cluster co-expressed genes within an organism as well as orthologs shared between organisms. Here, the ground state configuration correspond to three modules: 1, 2, 4.

Cross-species clusters for worm and fly



GO terms of conserved modules



GO terms of specific modules

worm specific dauer entry

fly specific chitin activities

Yan KK et al. Genome Biology. 2014

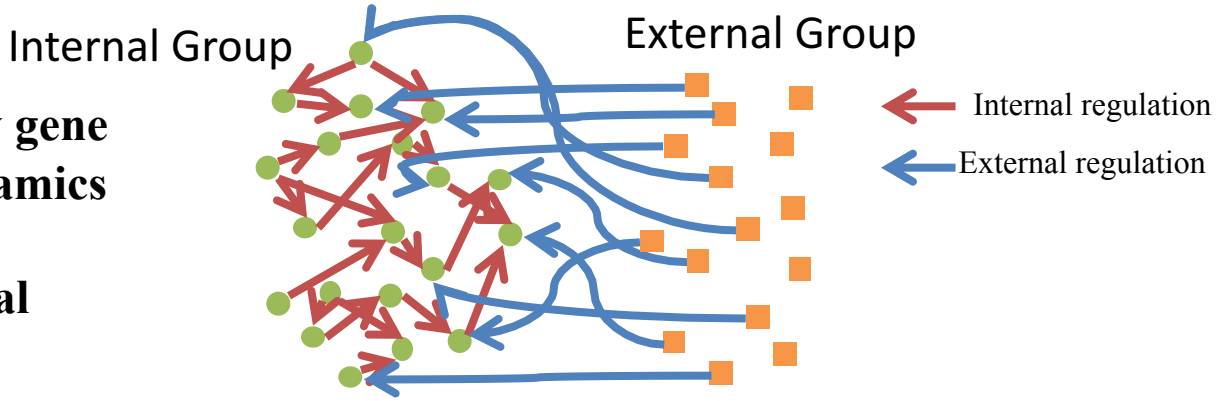
Large-scale Transcriptome Mining: Building Integrative Regulatory Models, while Protecting Individual Privacy

- Transcriptome analysis **data**
  - Comparative ENCODE – Lots of Matched Data
  - 1000G+Geuvadis for privacy
- **Expression Clustering**, Cross-species
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones.
  - Also, conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

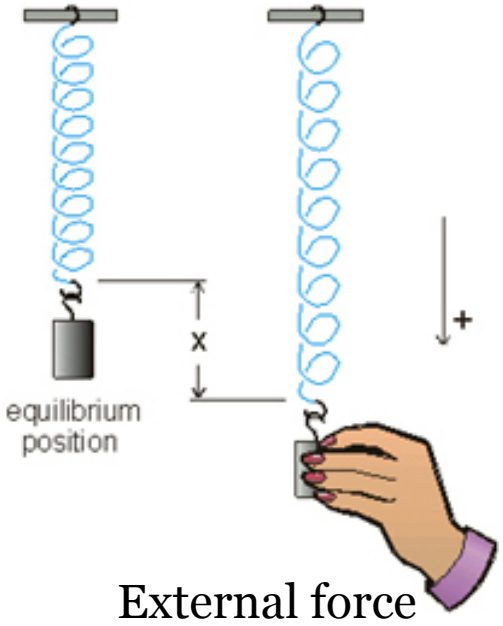
- The General **Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies & burdensome security
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
  - Details on Relevant Hacks: Genomic, Computer Security, & Netfix

- **RNA-seq: How to Publicly Share Some of it**
  - Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
  - Quantifying accuracy of prediction, via gap between best & 2nd best match

Internal & external gene regulatory networks



How to identify gene expression dynamics driven by internal/external regulation?

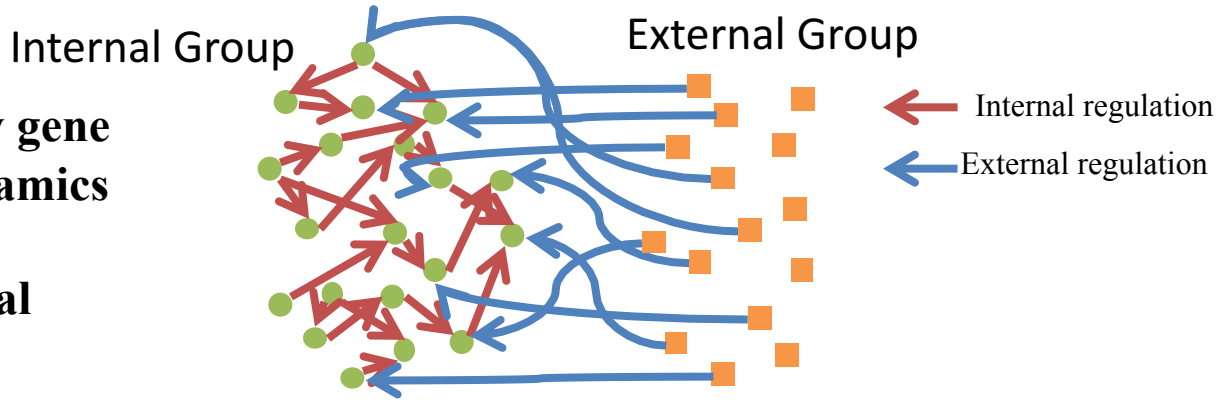


Interested system	Internal regulatory network	External regulatory network
Cross-species conserved genes	Conserved transcriptional factors (TFs)	Non-conserved TFs
Protein-coding genes	TFs	micro-RNAs
Individual's protein coding genes	Wild-type TFs	Somatic mutated TFs
Protein-coding genes in brain	Commonly expressed TFs	Brain-specific expressed TFs
Protein-coding genes in development	House-keeping TFs	Developmental TFs

[Wang et al. *PLOS CB* (in revision, '15)]

State-space model for internal and external gene regulatory networks

How to identify gene expression dynamics driven by internal/external regulation?



State space model

$$X_{t+1} = A X_t + B U_t$$

State: Gene expression vector of Group X at time t+1

A_{ij} captures temporal casual influence from Gene i to Gene j in internal group

State: Gene expression vector of internal group at time t

Control: Gene expression vector of external factors at time t

B_{kl} captures temporal casual influence from external factor k to Gene l in internal group

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

Decomposition of internal and external-related dynamic components

$$
\begin{aligned}
X_t &= AX_{t-1} + BU_{t-1} \\
&= A(AX_{t-2} + BU_{t-2}) + BU_{t-1} \\
&= A^2X_{t-2} + ABU_{t-2} + BU_{t-1} \\
&= A^3X_{t-3} + A^2BU_{t-3} + ABU_{t-2} + BU_{t-1} \\
&= \cdots \\
&= A^{t-1}X_1 + A^{t-2}BU_1 + A^{t-3}BU_2 + \cdots + ABU_{t-2} + BU_{t-1} \\
&= \underbrace{A^{t-1}X_1}_{x_t^{INT}} + \underbrace{\sum_{k=1}^{t-2} A^k BU_{t-1-k}}_{x_t^{INTER}} + \underbrace{BU_{t-1}}_{x_t^{EXT}}
\end{aligned}
$$

X_t^{EXT}: externally driven dynamic component

X_t^{INT}: Internally driven dynamic component

X_t^{INTER}: dynamic components driven by interactions between internal and external terms

* Subdivision of the rest of the terms $\sum_{k=1}^{t-2} A^k BU_{t-1-k} + BU_{t-1}$ is completely arbitrary

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

Effective state space model for meta-genes

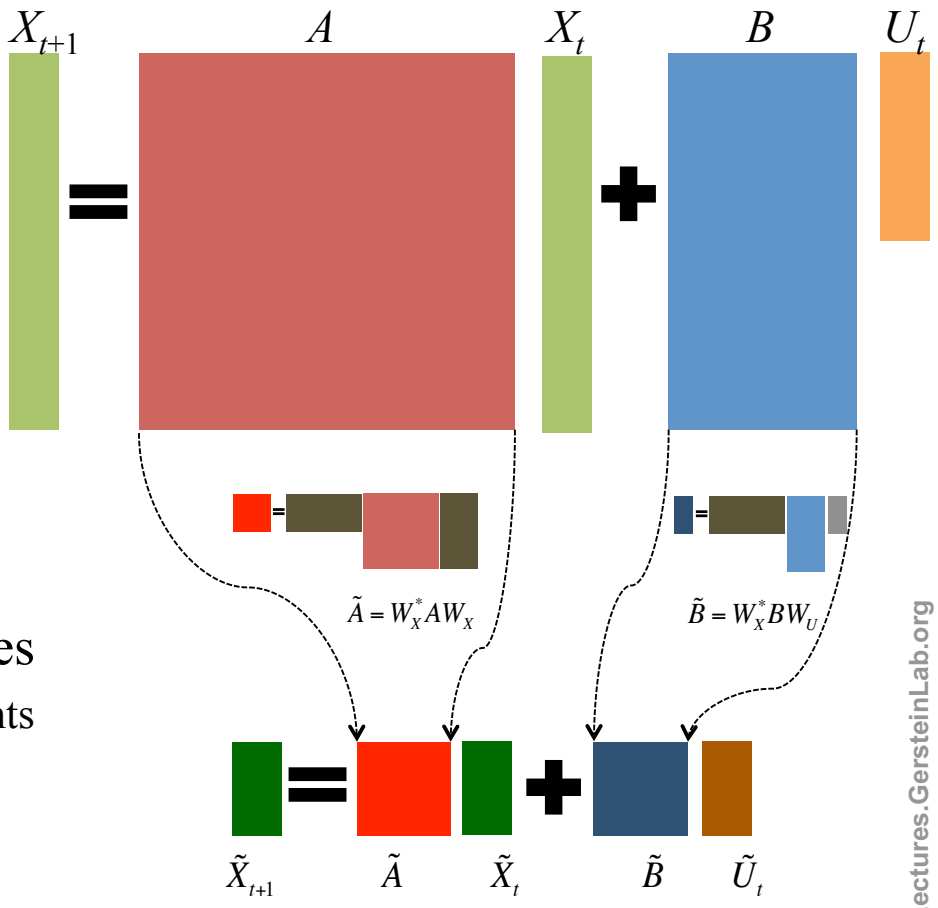
Not enough data to estimate state space model for genes
(e.g., 25 time points per gene to estimate 4 million elements of A or B for 2000 genes)

$$X_{t+1} = AX_t + BU_t$$

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)

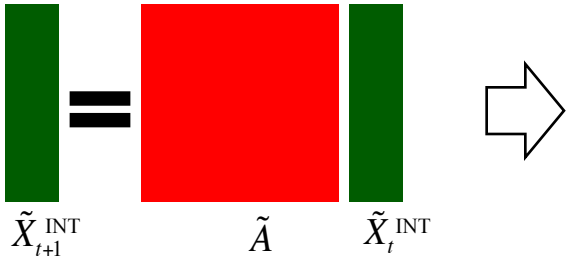
$$\tilde{X}_{t+1} = \tilde{A}\tilde{X}_t + \tilde{B}\tilde{U}_t$$



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

Eigenvalues of $\tilde{A}$ determine internal dynamics

First-order linear matrix difference equation

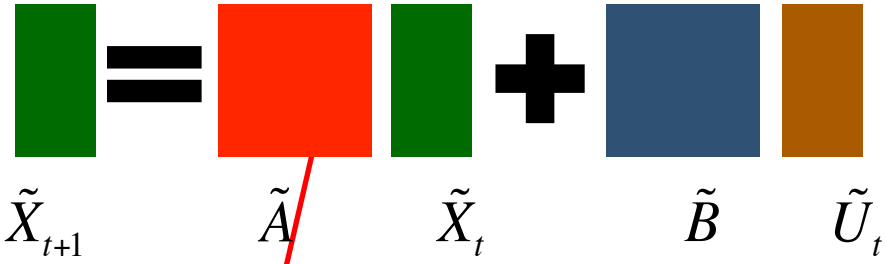


Analytic solution

A general first-order linear matrix difference equation, $Q_{t+1}=CQ_t$ is $Q_t=C^tQ_0=(HEH^{-1})^tQ_0=HE^tH^{-1}Q_0=HE^tS$, where the columns of the matrix H are eigenvectors of C, the diagonal elements of the diagonal matrix E are eigenvalues of C such that $CH=HE$, and the vector $S=H^{-1}Q_0$. Then, if we rewrite Q_t by a linear combination of the time exponential of eigenvalues of C, we have that $Q_t = HE^tS = \sum_{i=1}^{m_c} \alpha_i^t s_i H_i = \sum_{i=1}^{m_c} \alpha_i^t K_i$, where m_c is the total number of eigenvalues of C, α_i is the i^{th} eigenvalue of C, s_i is the i^{th} element of S, H_i is the i^{th} eigenvector of C (i.e., the i^{th} column of H), and $K_i=s_iH_i$ is the coefficient vector of Q_t over the t^{th} time exponential of α_i.

$\tilde{X}_t^{\text{INT}} = \sum_{p=1}^{M_1} \lambda_p^t \tilde{K}_p$; i.e., the internally driven component of i^{th} meta-gene's expression across all time points, $[\tilde{X}_1^{\text{INT}}(i) \quad \tilde{X}_2^{\text{INT}}(i) \quad \dots \quad \tilde{X}_T^{\text{INT}}(i)] = \sum_{p=1}^{M_1} \tilde{K}_p(i) \underbrace{[\lambda_p^1 \quad \lambda_p^2 \quad \dots \quad \lambda_p^T]}_{p^{\text{th}}\text{ iPDP}}$, a linear summation of the time exponential of eigenvalues of $\tilde{A}$

Canonical temporal expression trajectories from effective state space model

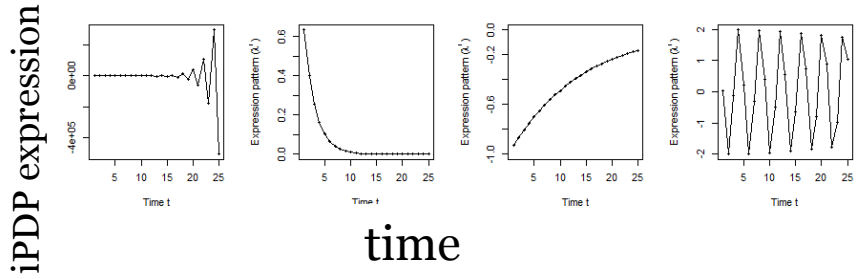


$\tilde{X}_{t+1}$ $\tilde{A}$ $\tilde{X}_t$ $\tilde{B}$ $\tilde{U}_t$

Internal driven dynamics

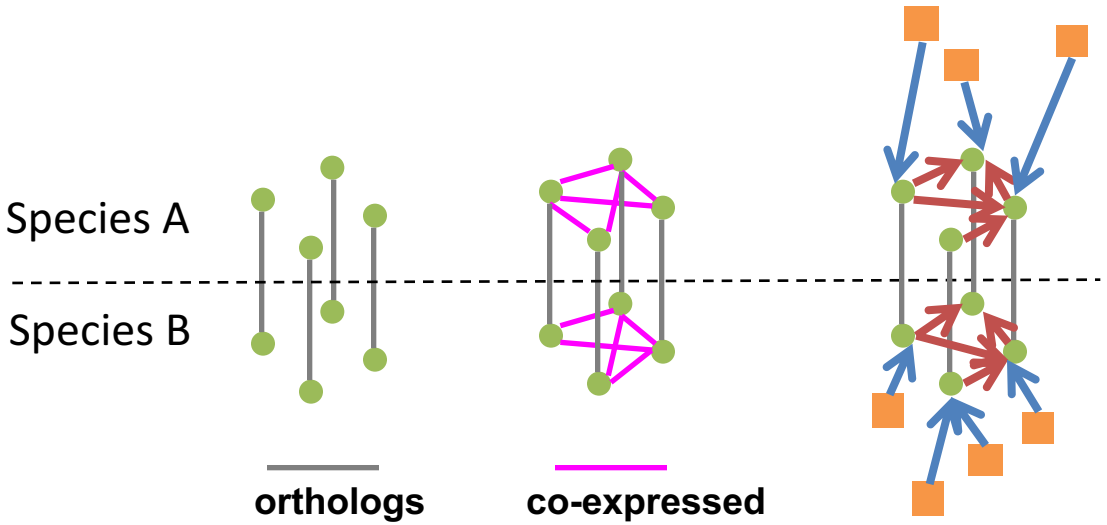
p^{th} internal principal dynamic pattern (iPDP): $[\lambda_p{}^1, \lambda_p{}^2, \ldots, \lambda_p{}^T]$, where λ_p is p^{th} eigenvalue of $\tilde{A}$.

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



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

Are gene regulations among orthologs conserved across species?



Are gene regulatory networks among orthologs conserved across species?

Regulation among orthologs (**internal**)

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

Orthologous genes (orthologs)

Species-specific transcription factors

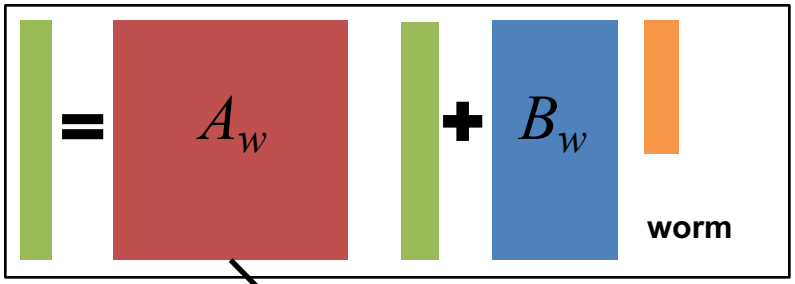
To what degree can't ortholog expression levels be predicted due to species-specific regulation

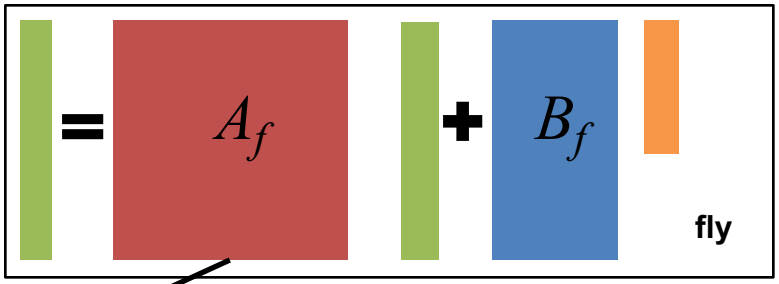
[Wang et al. *PLOS CB*, '16]

Are there any conserved regulatory networks between worm and fly during embryonic development?

Not enough time samples!

Dataset	Internal Group	External Group	Developmental stages	# of unknown parameters in A and B	# of available time samples
worm (*C. elegans*)	N_1=3147 worm-fly orthologs (incl. ortholog TFs)	N_2=509 worm-specific transcription factors	T=25 time points: 0, 0.5, 1, …, 12 hours	3147*3147+3147*509=11.5M	3147*25+509*25=91400
fly (*D. mel.*)		N_2=442 fly-specific transcription factors	T=12 time points: 0, 2, 4, 6, 8,…, 20, 22 hours	3147*3147+3147*442=11.3M	3147*25+442*25=89725



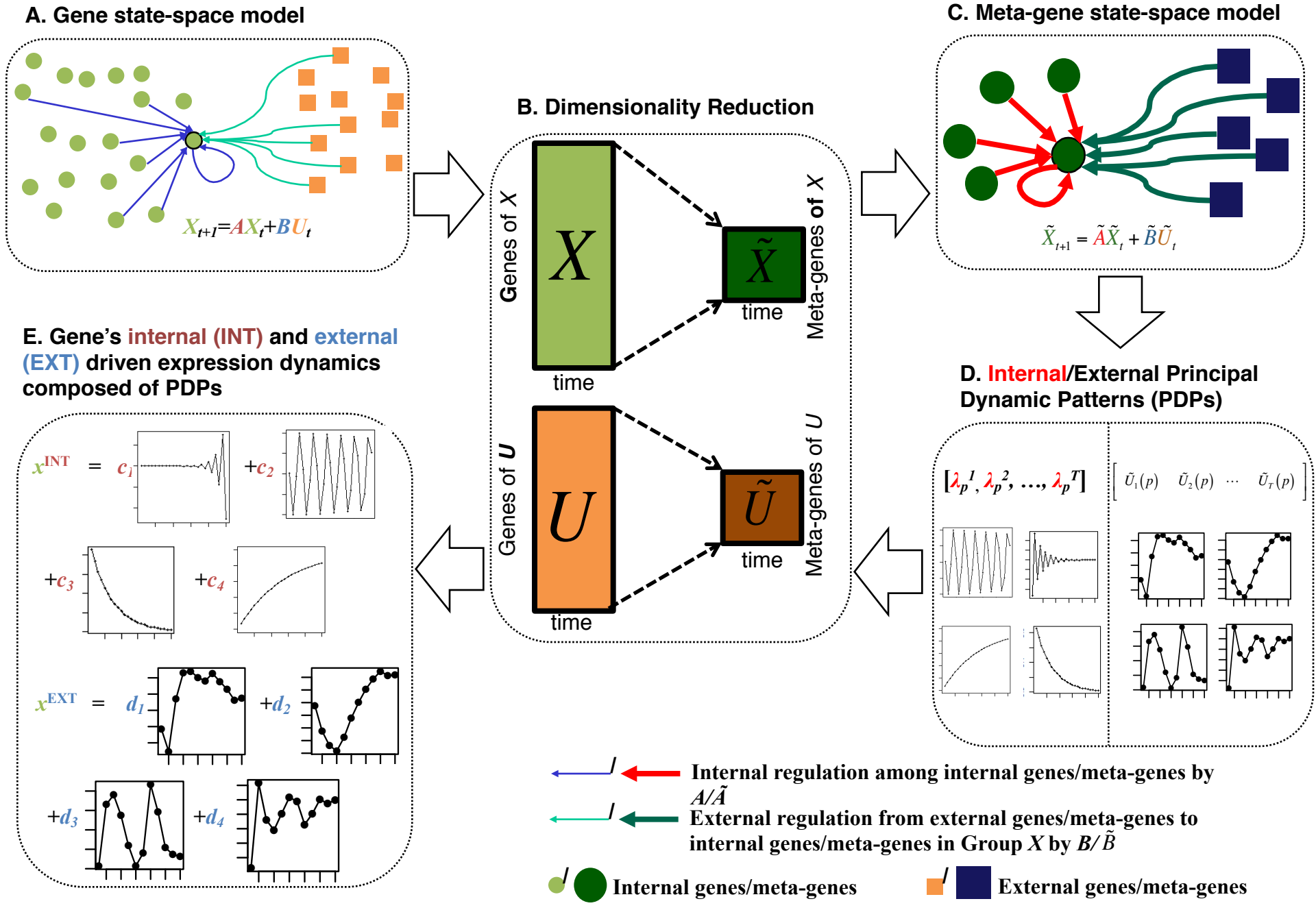


If A_w and A_f have similarities, cross-species conserved regulatory networks in embryonic development

Embryonic stem cells (ESCs)

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

Flowchart



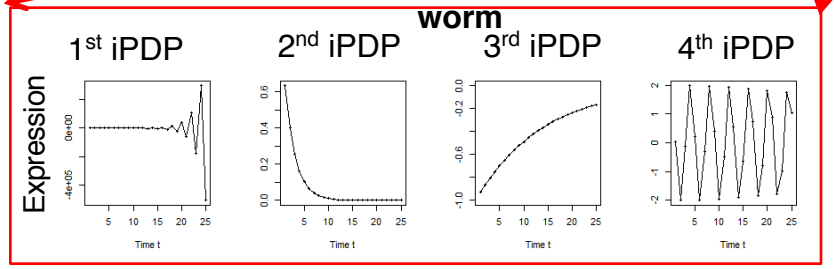
Orthologs have similar internal but different external dynamic patterns during embryonic development

Worm's effective state space model

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

A

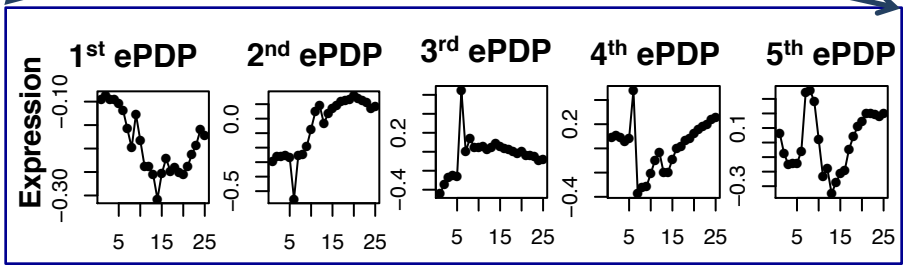
iPDPs: time exponentials of $\tilde{A}$ eigenvalues in worm



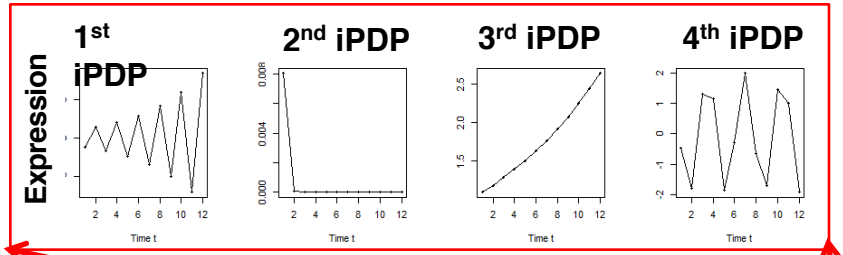
Similar iPDPs

B

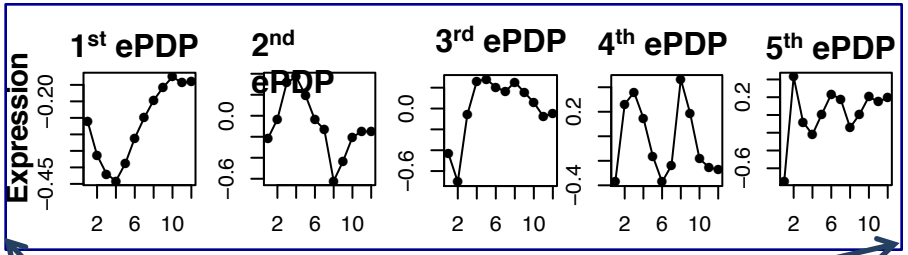
ePDPs: $\tilde{U}$ in worm



Different ePDPs



iPDPs: time exponentials of $\tilde{A}$ eigenvalues in fly



ePDPs: $\tilde{U}$ in fly

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

Fly's effective state space model

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

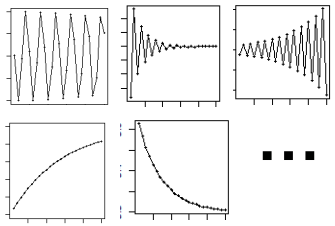
Projection back to gene space to get gene coefficients on iPDPs

Internal component of meta-genes: $\tilde{X}_{t+1}^{\mathrm{INT}} = \tilde{A}\tilde{X}_t^{\mathrm{INT}}$

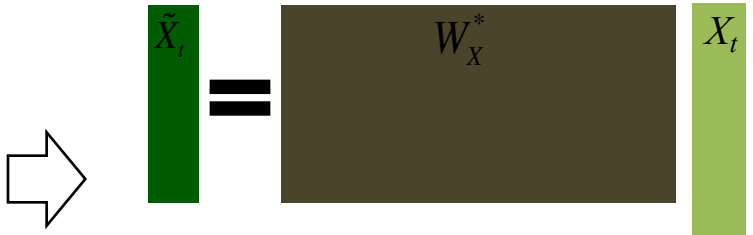
=>

$\tilde{X}_t^{\mathrm{INT}} = \sum_{p=1}^{M_1} \lambda_p^t \tilde{K}_p$; i.e., the internally driven component of i^{th} meta-gene's expression across all time points, $[\tilde{X}_1^{\mathrm{INT}}(i) \quad \tilde{X}_2^{\mathrm{INT}}(i) \quad \ldots \quad \tilde{X}_T^{\mathrm{INT}}(i)] =$

$$\sum_{p=1}^{M_1} \tilde{K}_p(i) \underbrace{\left[\lambda_p^1 \quad \lambda_p^2 \quad \ldots \quad \lambda_p^T\right]}_{p^{\mathrm{th}}\ \mathrm{iPDP}}$$



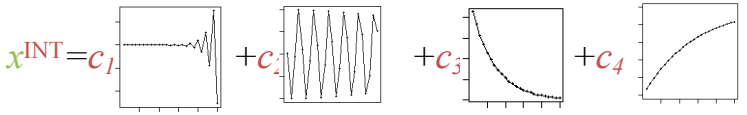
* V(i) represents i^{th} element of vector V



Linear transformation between genes and meta-genes

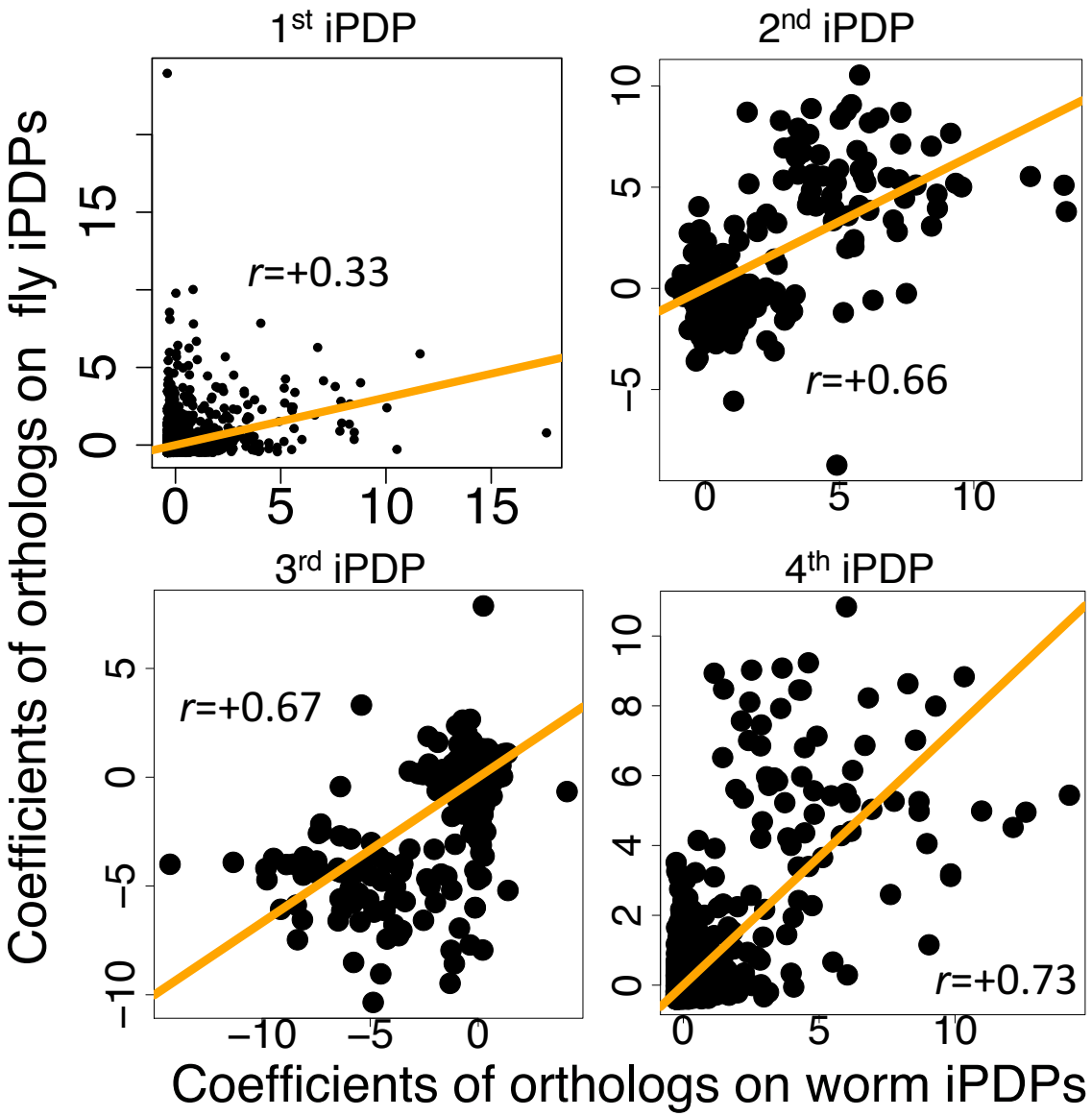
$$X_t^{\mathrm{INT}} = W_X \tilde{X}_t^{\mathrm{INT}} = \sum_{p=1}^{M_1} \lambda_p^t \underbrace{W_X \tilde{K}_p}_{C_p} = \sum_{p=1}^{M_1} \lambda_p^t C_p$$; i.e., the internally driven component of i^{th} gene's expression across all time points, $[X_1^{\mathrm{INT}}(i) \quad X_2^{\mathrm{INT}}(i) \quad \ldots \quad X_T^{\mathrm{INT}}(i)] =$

$$\sum_{p=1}^{M_1} C_p(i) \underbrace{\left[\lambda_p^1 \quad \lambda_p^2 \quad \ldots \quad \lambda_p^T\right]}_{p^{\mathrm{th}}\ \mathrm{iPDP}}$$



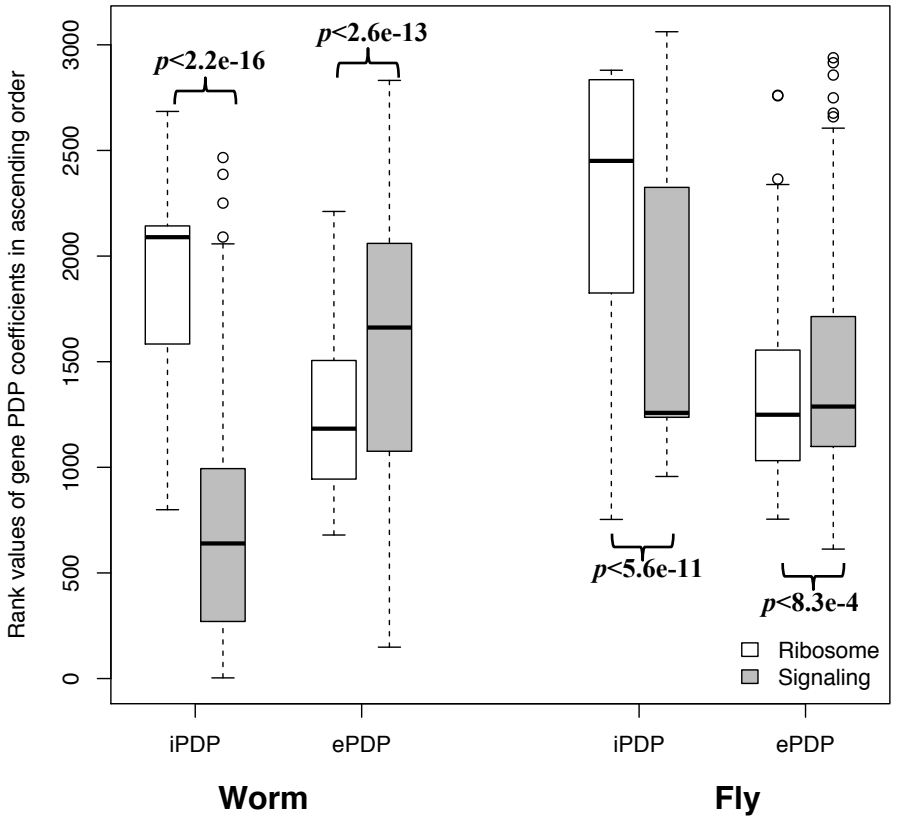
Individual gene x's coefficients on iPDPs

Orthologs have correlated iPDP coefficients



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

Evolutionarily conserved and younger genes exhibit the opposite internal and 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]

Large-scale Transcriptome Mining: Building Integrative Regulatory Models, while Protecting Individual Privacy

- Transcriptome analysis **data**
  - Comparative ENCODE – Lots of Matched Data
  - 1000G+Geuvadis for privacy
- **Expression Clustering**, Cross-species
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones.
  - Also, conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

- The General **Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies & burdensome security
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
  - Details on Relevant Hacks: Genomic, Computer Security, & Netfix

- **RNA-seq: How to Publicly Share Some of it**
  - Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
  - Quantifying accuracy of prediction, via gap between best & 2nd best match

The Conundrum of Genomic Privacy: Is it a Problem?

Yes

Genetic Exceptionalism :
The Genome is very fundamental data, potentially very revealing about one's identity & characteristics

Identification Risk: Find that someone participated in a study [eg Craig, Erlich]

Characterization Risk: Finding that you have a particular trait from studying your identified genome [eg Watson ApoE status]

No

Shifting societal foci

No one really cares about *your* genes

You might not care

[Klitzman & Sweeney ('11), J Genet Couns 20:98l; Greenbaum & Gerstein ('09), New Sci. (Sep 23)]

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
- (...or is the genome more exceptional & fundamental?)

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

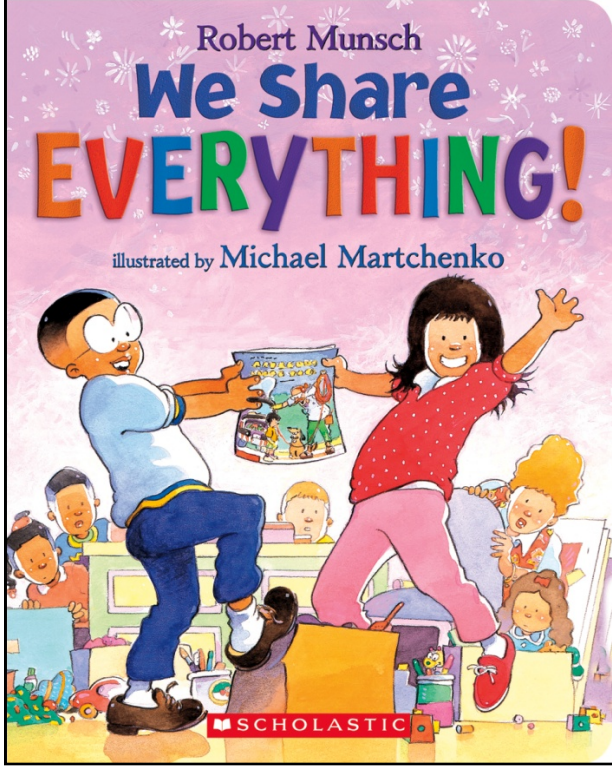
Tricky Privacy Considerations in Personal Genomics

- **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? What is acceptable data leakage? Can we quantify leakage?
    - Ex: photos of eye color
  - Cost Benefit Analysis: how helpful is identifiable data in genomic research v. potential harm from a breach?

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]

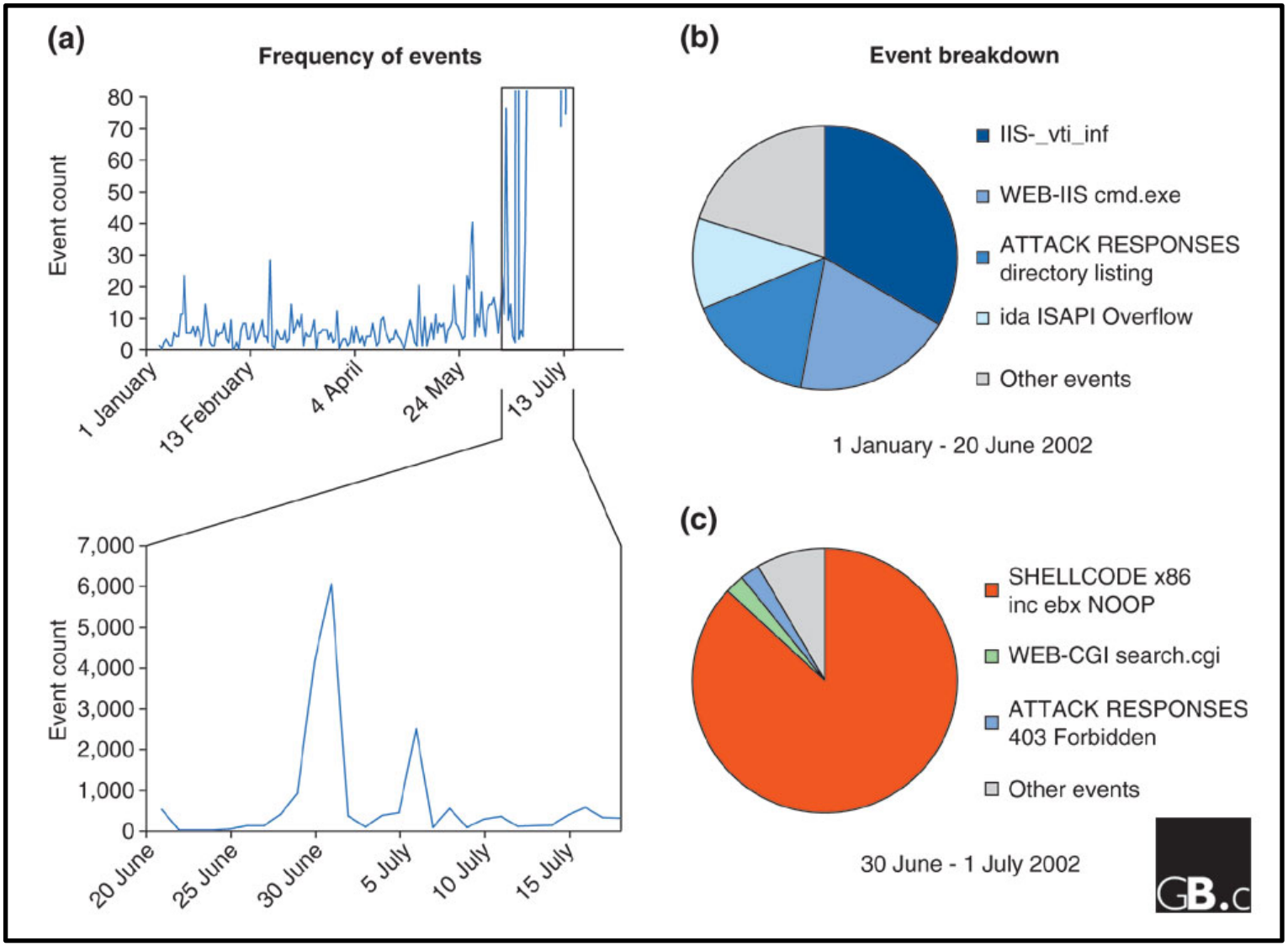
Strawman Hybrid Social & Tech Proposed Solution?

- 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)]

Difficulty in Securing Computers & Data



[Smith et al ('05), Genome Bio]

Genomic Privacy Hacks, Mostly Focusing on Identification

- Early genomic studies were based on small cohorts
  - Individuals give consent to participate but request anonymity
    - HAPMAP, PGP, 1000 Genomes…
  - Focus on hiding the participation of individuals
  - Attacks aimed at detecting whether an individual with known genotypes participated a study
    - "Detection of genomes in a mixture" (Homer et al 2008, Im et al 2012)
- As more people are genotyped, more individuals are in large private genomic databases
  - Detection of an individual is irrelevant, as their participation is already known
    - Current EX: "An individual's genomic/phenotypic data is most certainly stored in their hospital"
    - Future: >1M people's health information is part of a NIH/PMI or NHS databases
- Identification attacks now focus on pinpointing individuals by cross-referencing large seemingly independent datasets
  - Illustrates that a leaked/hacker/stolen dataset, even when anonymized, can leak information
  - Sweeney et al 2013, Gymrek et al 2013

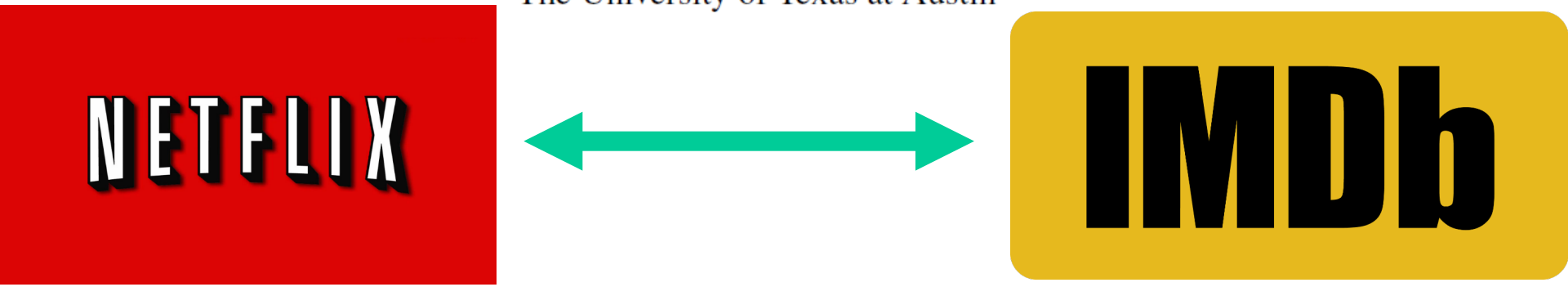
Gymrek et al, "Identifying Personal Genomes by Surname Inference" (2013)
Homer et al, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays." (2008)
Im et al, "On Sharing Quantitative Trait GWAS Results in an Era of Multiple-omics Data and the Limits of Genomic Privacy" (2012)
Sweeney et al, "Identifying Participants in the Personal Genome Project by Name" (2013)

What is a linking attack? Case of Netflix Prize

Robust De-anonymization of Large Datasets
(How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin



1. Very large datasets
2. A lot of users have a Netflix and an IMDb account
3. A user rates similar scores to a movie in Netflix and IMDb
4. A user rates a particular movie around the same date in Netflix and IMDb

What is a linking attack? Case of Netflix Prize

Movie ratings database

Anonymized Netflix Prize Training Dataset made available to contestants

100 million ratings
500,000 users

200 movie ratings/user
5,000 users/movie rating

User (ID)	Movie (ID)	Date of Rating	Rating [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
…	…	…	…
…	…	…	…

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**

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**

Large-scale Transcriptome Mining: Building Integrative Regulatory Models, while Protecting Individual Privacy

- Transcriptome analysis **data**
  - Comparative ENCODE – Lots of Matched Data
  - 1000G+Geuvadis for privacy
- **Expression Clustering**, Cross-species
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones.
  - Also, conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

- The General **Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies & burdensome security
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
  - Details on Relevant Hacks: Genomic, Computer Security, & Netfix

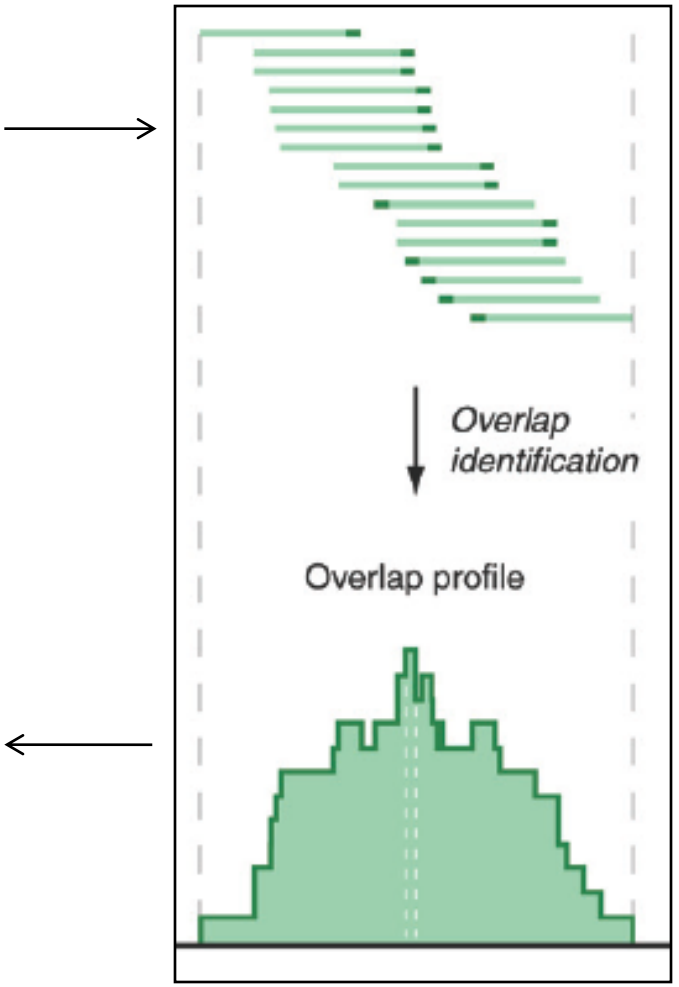
- **RNA-seq: How to Publicly Share Some of it**
  - Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
  - Quantifying accuracy of prediction, via gap between best & 2nd best match

RNA-seq

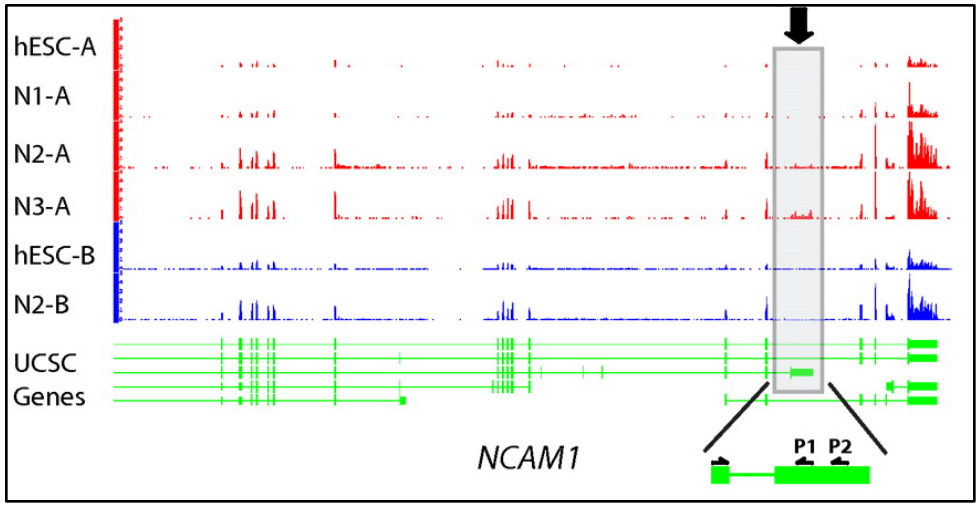
RNA-seq uses next-generation sequencing technologies to reveal RNA presence and quantity within a biological sample.

```
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
TACCGATCGAGTCGACTGTAAACGTAGGCATA
ATTCTGACTGGTGTCATGCTGATGTACTTAAA
```

Reads (fasta)
- Quality scores (fastq)
- Mapping (BAM)
- Contain variant information in transcribed regions



Reads => Signal

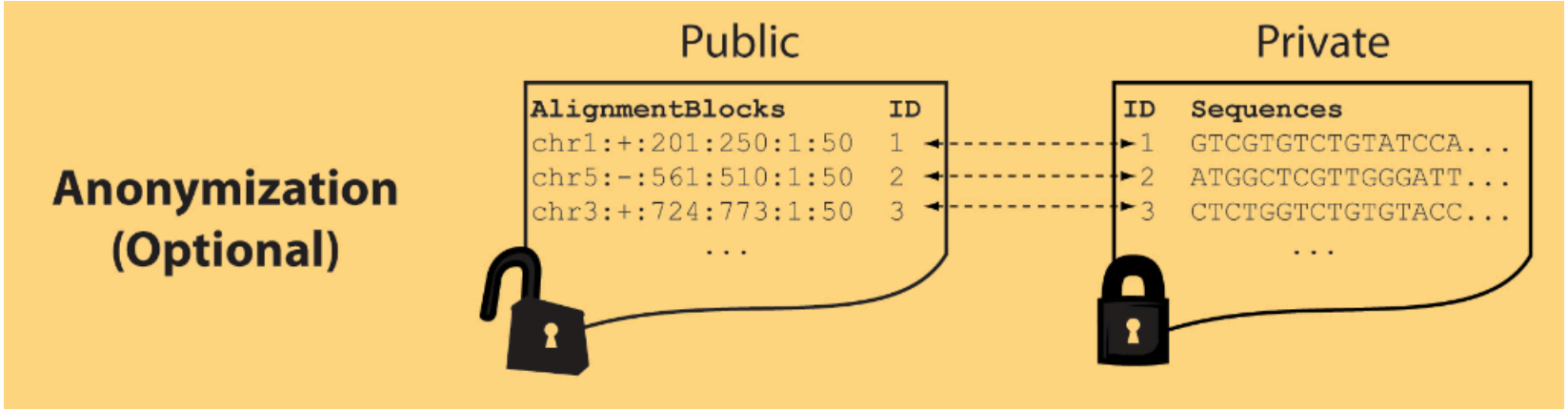


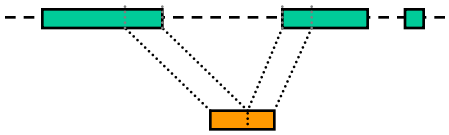
Quantitative information from RNA-seq signal: average signals at exon level (RPKMs)

[*PLOS CB* 4:e1000158; *PNAS* 4:107: 5254 ; *IJC* 123:569]

Light-weight formats

- 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)

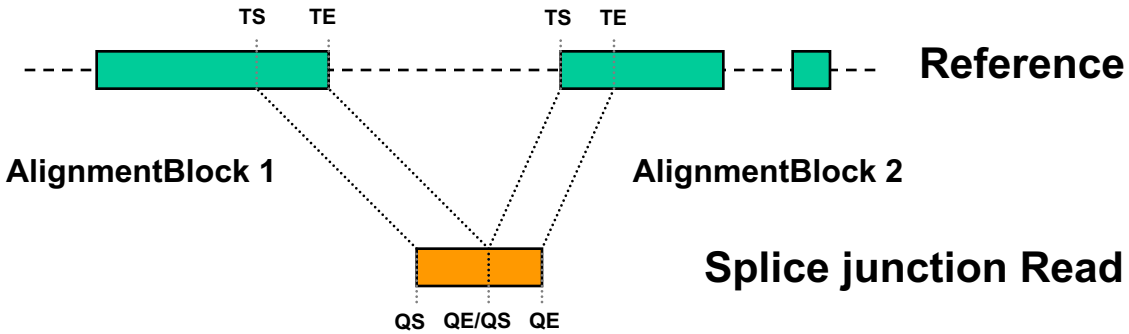
Reads (linked via ID, 10X larger than mapping coord.)

[*Bioinformatics* 27: 281]

MRF Examples

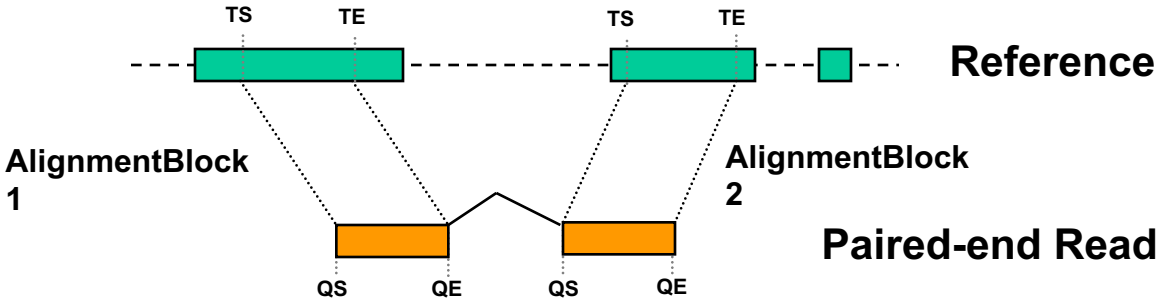
Reference based compression **(ie CRAM)** is similar but it stores actual variant beyond just position of alignment block

`chr2:+:601:630:1:30,chr2:+:921:940:31:50`



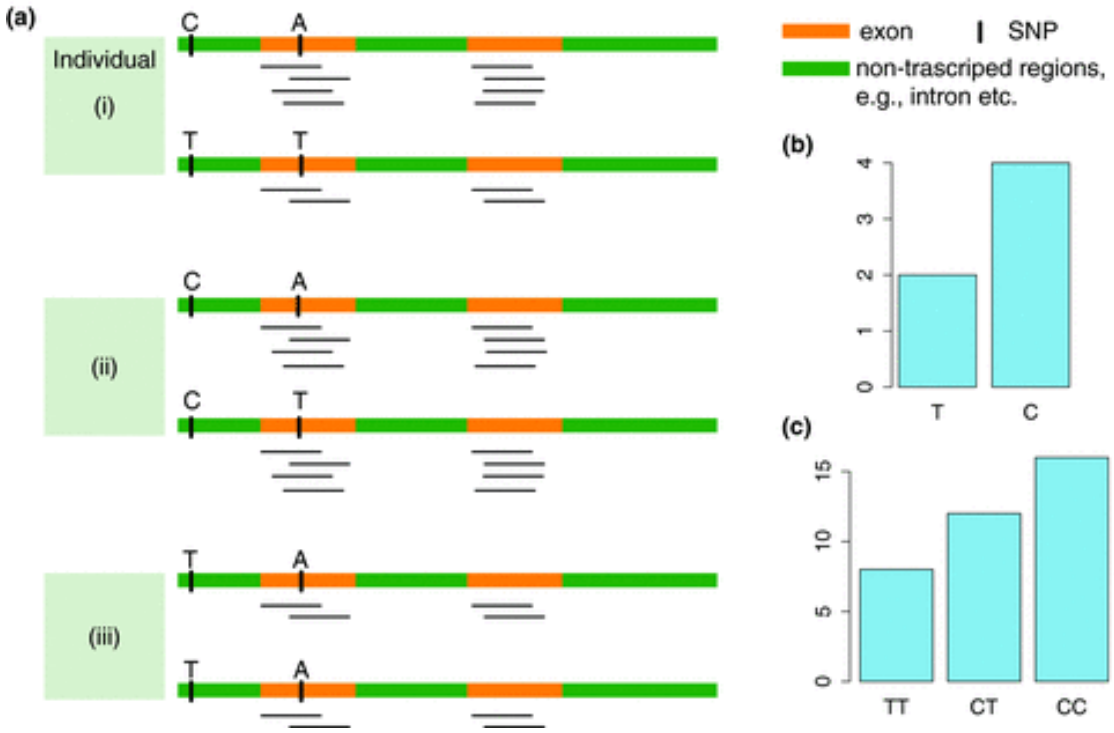
Legend: TS = TargetStart, TE = TargetEnd, QS = QueryStart, QE = QueryEnd

`chr9:+:431:480:1:50|chr9:+:945:994:1:50`

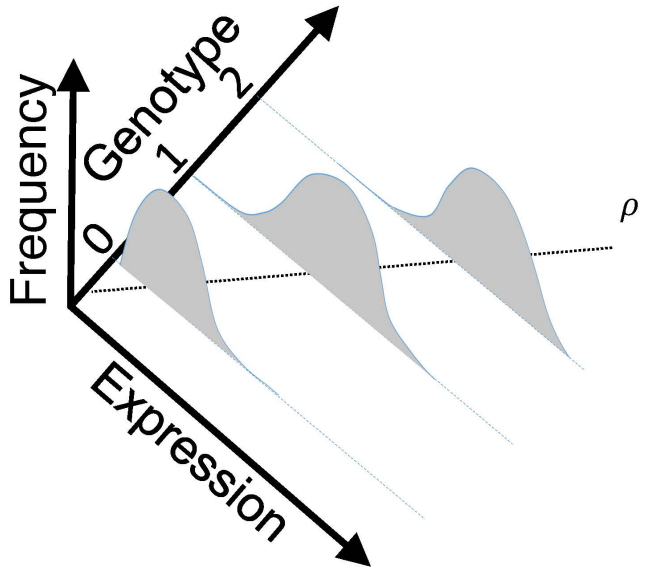


Legend: TS = TargetStart, TE = TargetEnd, QS = QueryStart, QE = QueryEnd

[Habegger et al., Bioinformatics ('11)]



[*Biometrics 68(1) 1–11*]

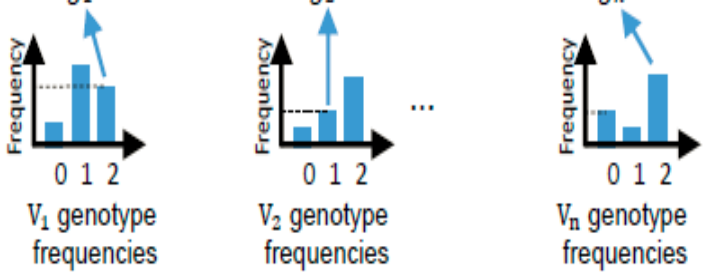


eQTL Mapping Using RNA-Seq Data

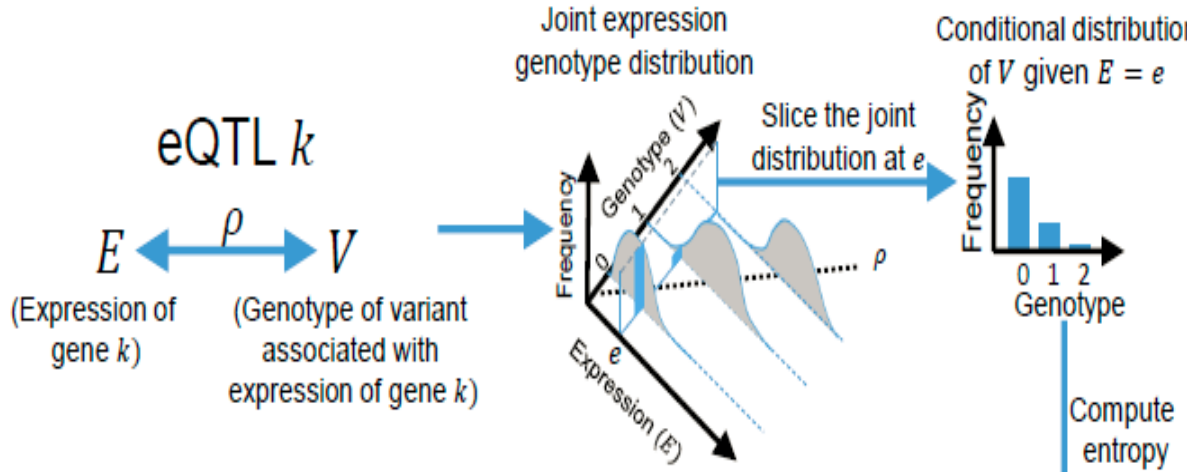
- eQTLs are genomic loci that contribute to variation in mRNA expression levels
- eQTLs provide insights on transcription regulation, and the molecular basis of phenotypic outcomes
- eQTL mapping can be done with RNA-Seq data

Information Content and Predictability

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



- **Higher frequency: Lower ICI**
- **Lower frequency: Higher ICI**
- **Additive for multiple variants**



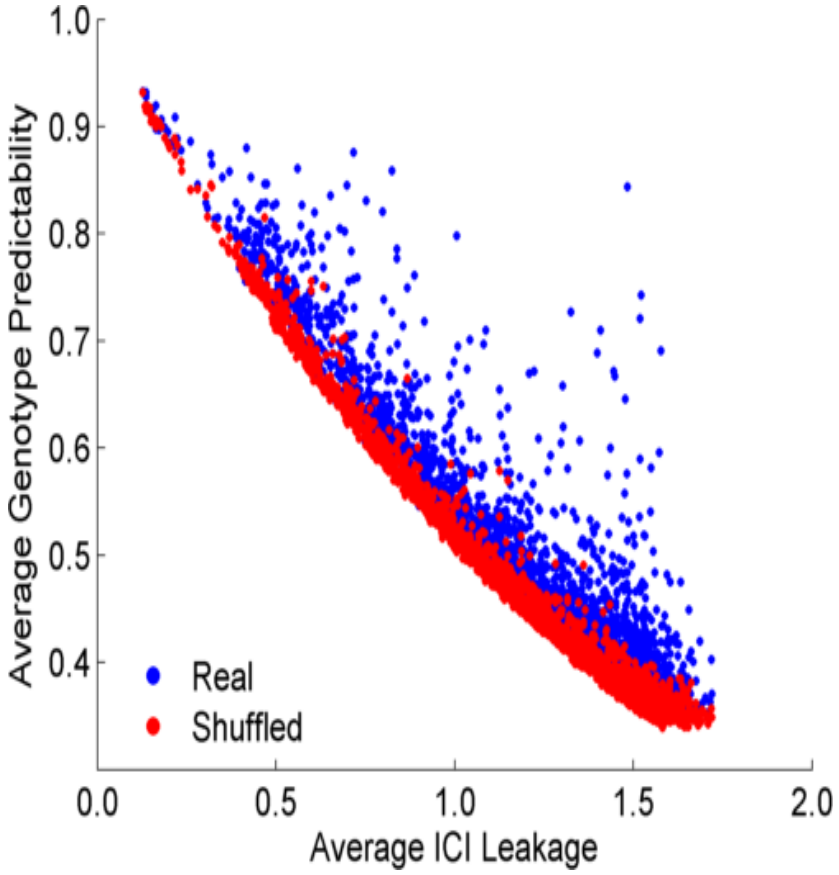
$$\pi(V \mid E = e) = e^{-H(V \mid E = e)}$$

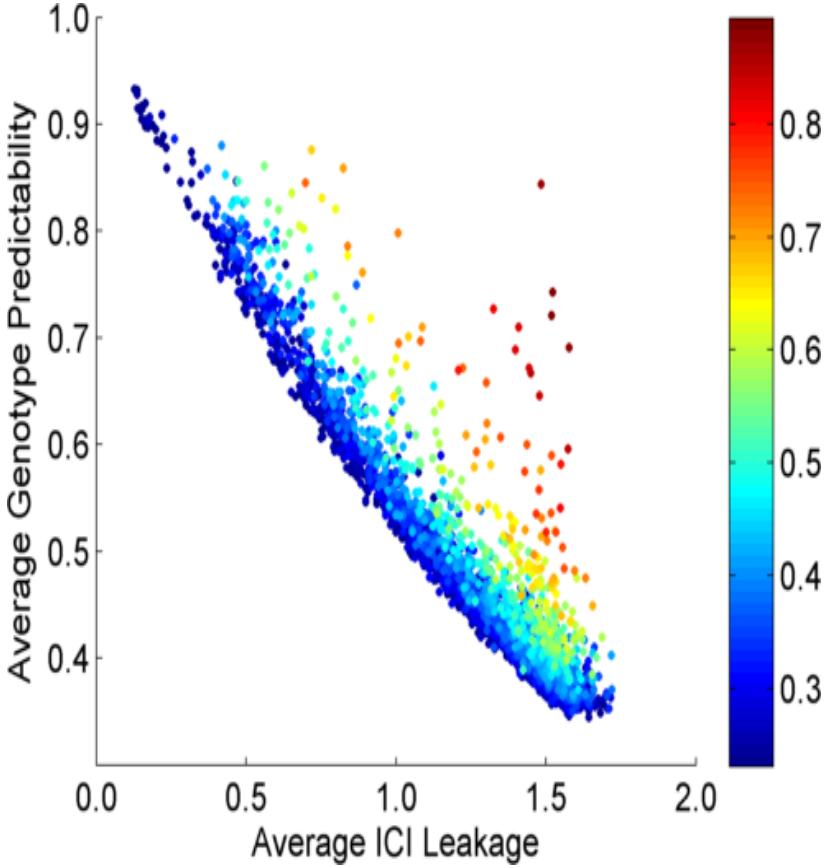
- **Higher cond. entropy: Lower predictability**
- **Lower cond. entropy: Higher predictability**
- **Additive for multiple eQTLs**

[Harmanci et al. Nat. Meth. 2016]

Per eQTL and ICI Cumulative Leakage versus Genotype Predictability

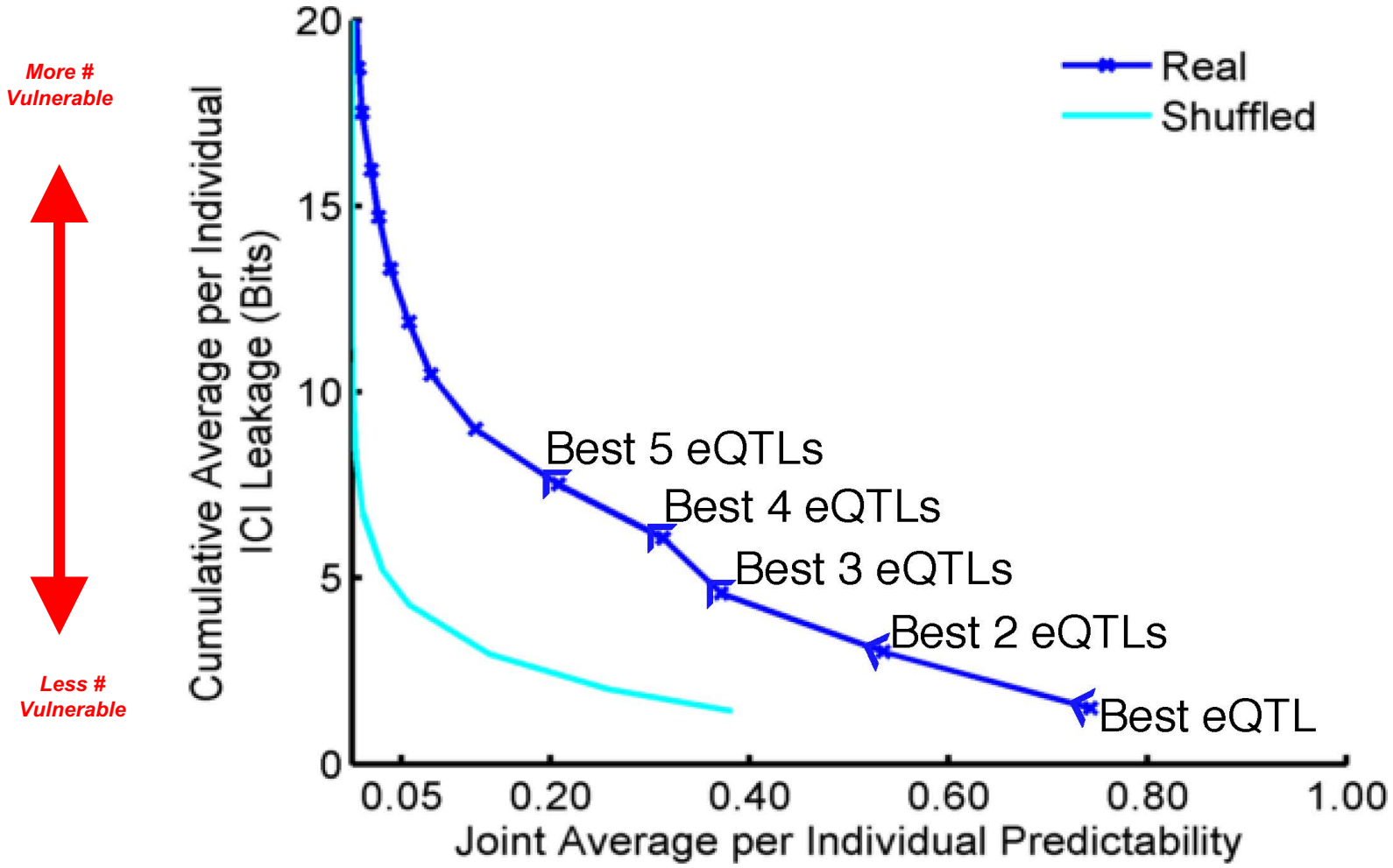
Colors by absolute correlation





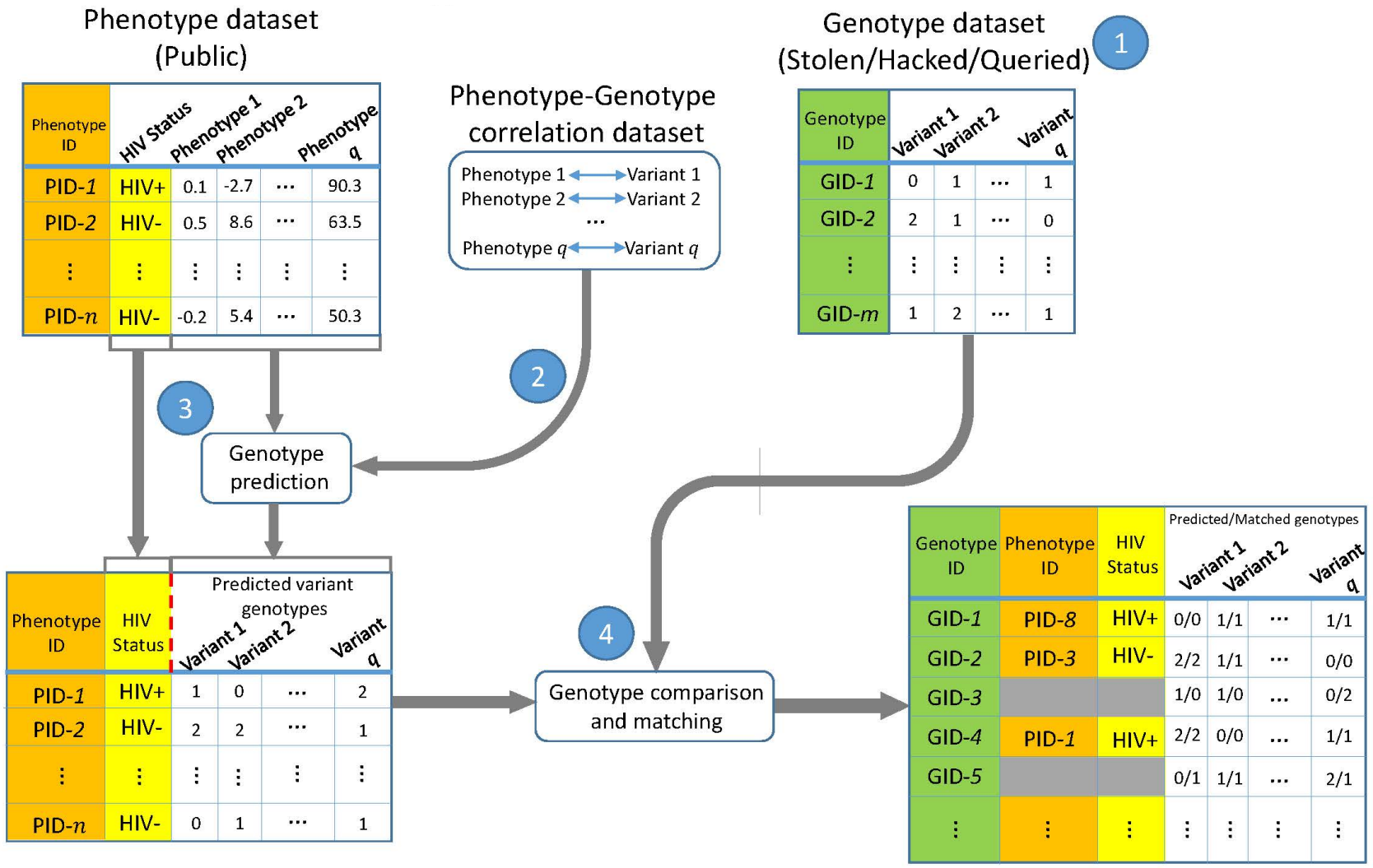
[Harmanciet al. Nat. Meth. (in revision)]

Cumulative Leakage versus Joint Predictability



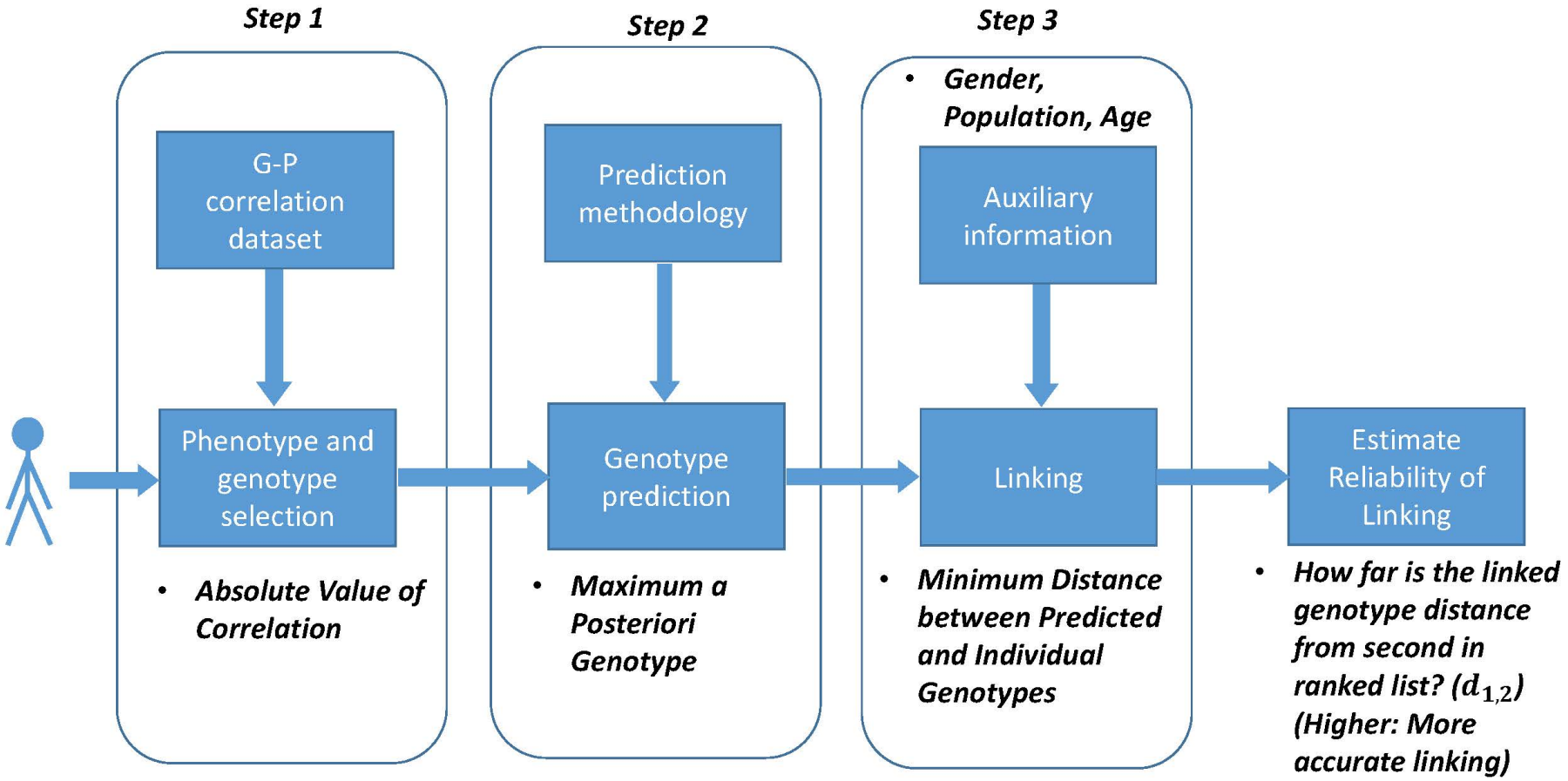
[Harmanciet al. Nat. Meth. (in revision)]

Linking Attack Scenario



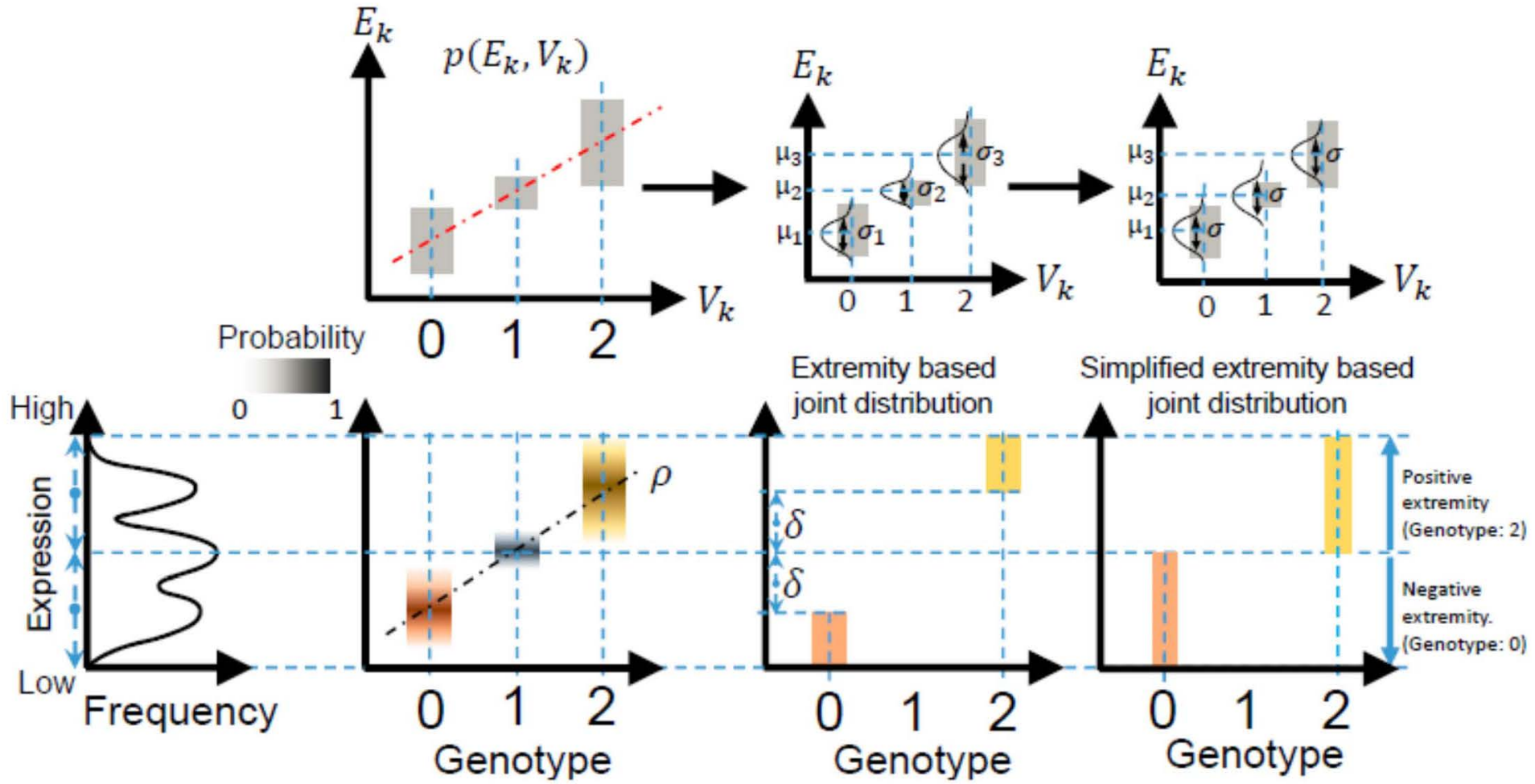
[Harmanciet al. Nat. Meth. (in revision)]

Steps in Instantiation of a (Mock) Linking Attack



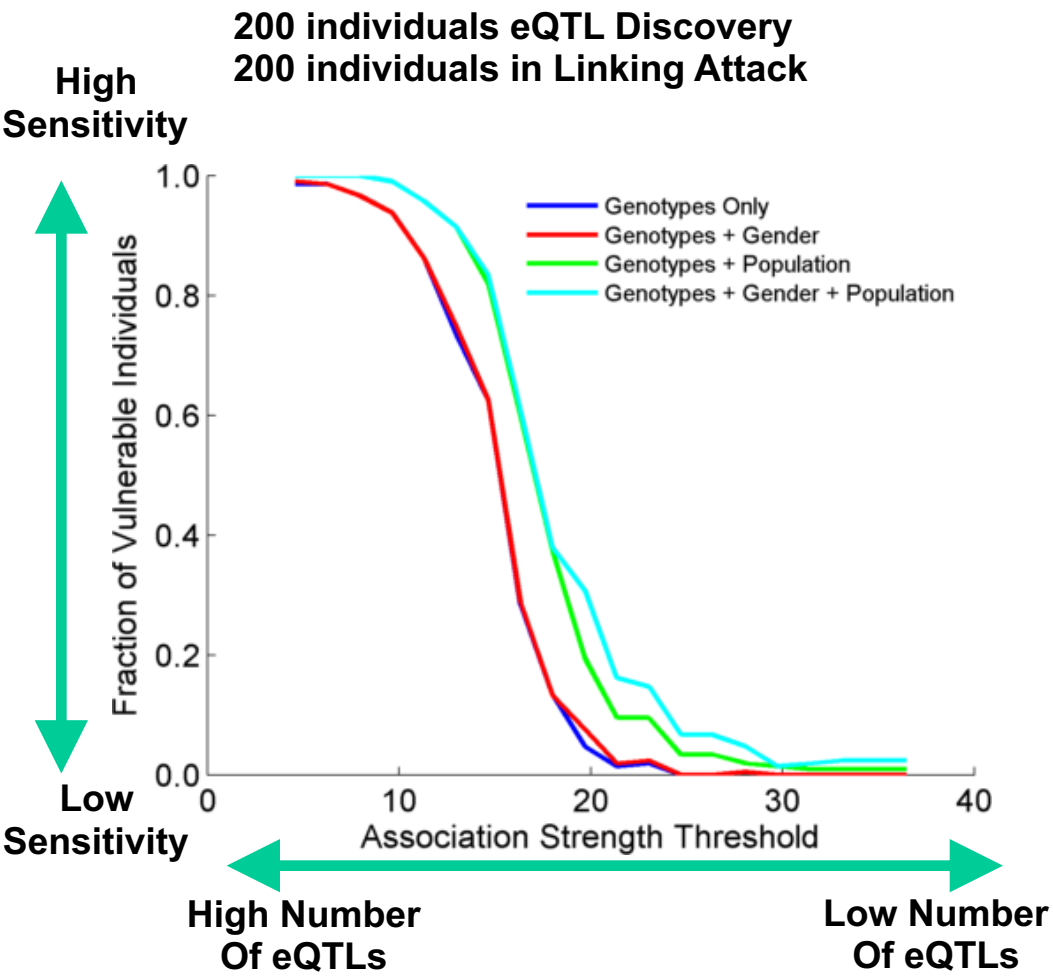
[Harmanci et al. *Nat. Meth.* (16)]

Levels of Expression-Genotype Model Simplifications



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

Linking Attack with Extremity based Genotype Prediction



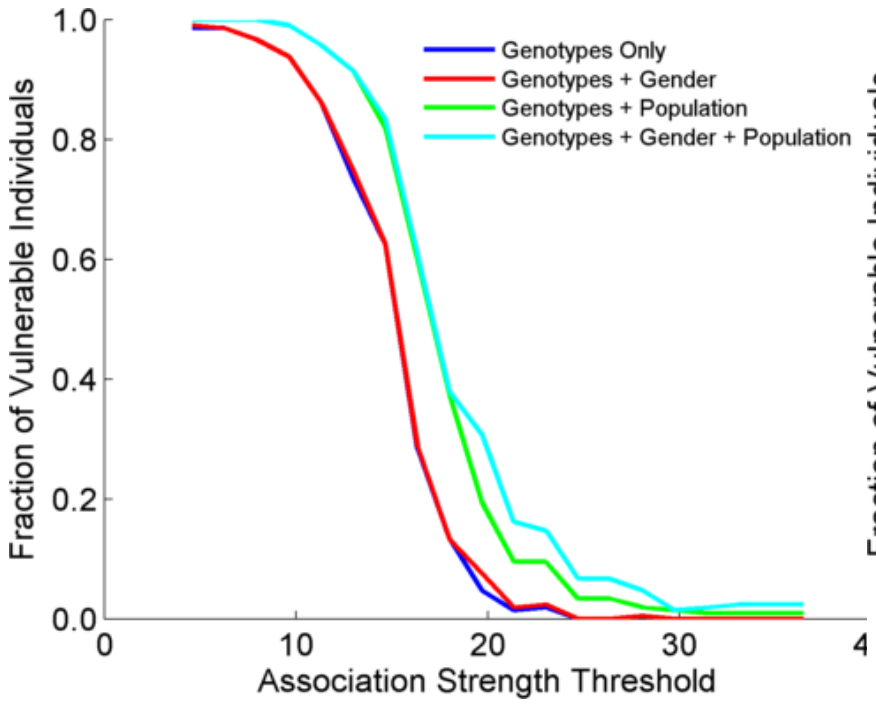
- X-axis: The threshold of association for selecting the eQTLs
  - Higher threshold: Smaller number of eQTLs
- Y-axis: Fraction of correctly linked individuals
  - Measures the ***Sensitivity of the attack***

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

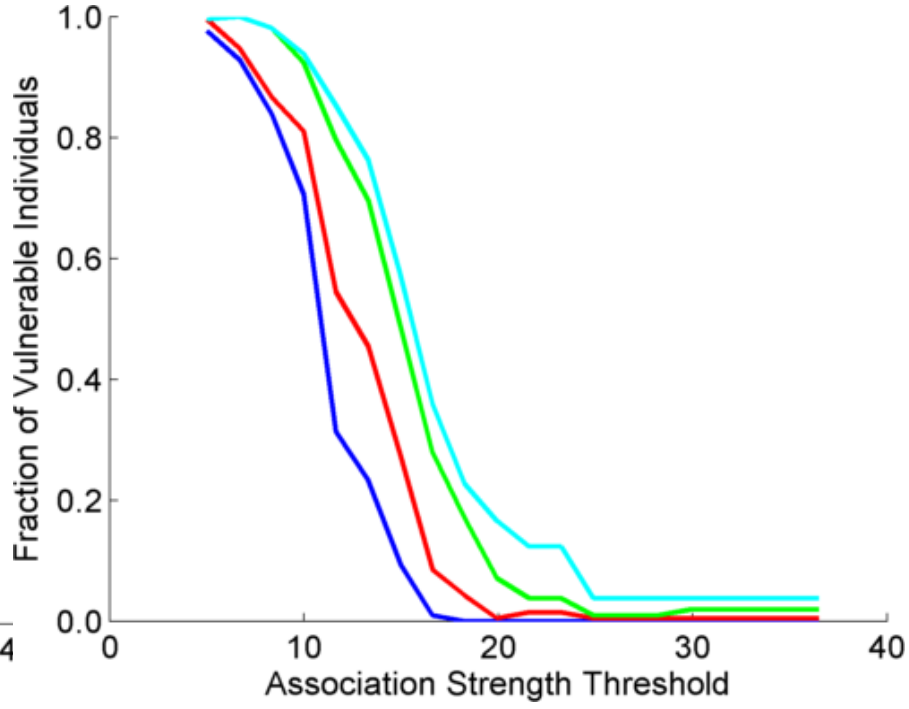
58 - Lectures.GersteinLab.org

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)]

59 - Lectures.GersteinLab.org

Which 70%?

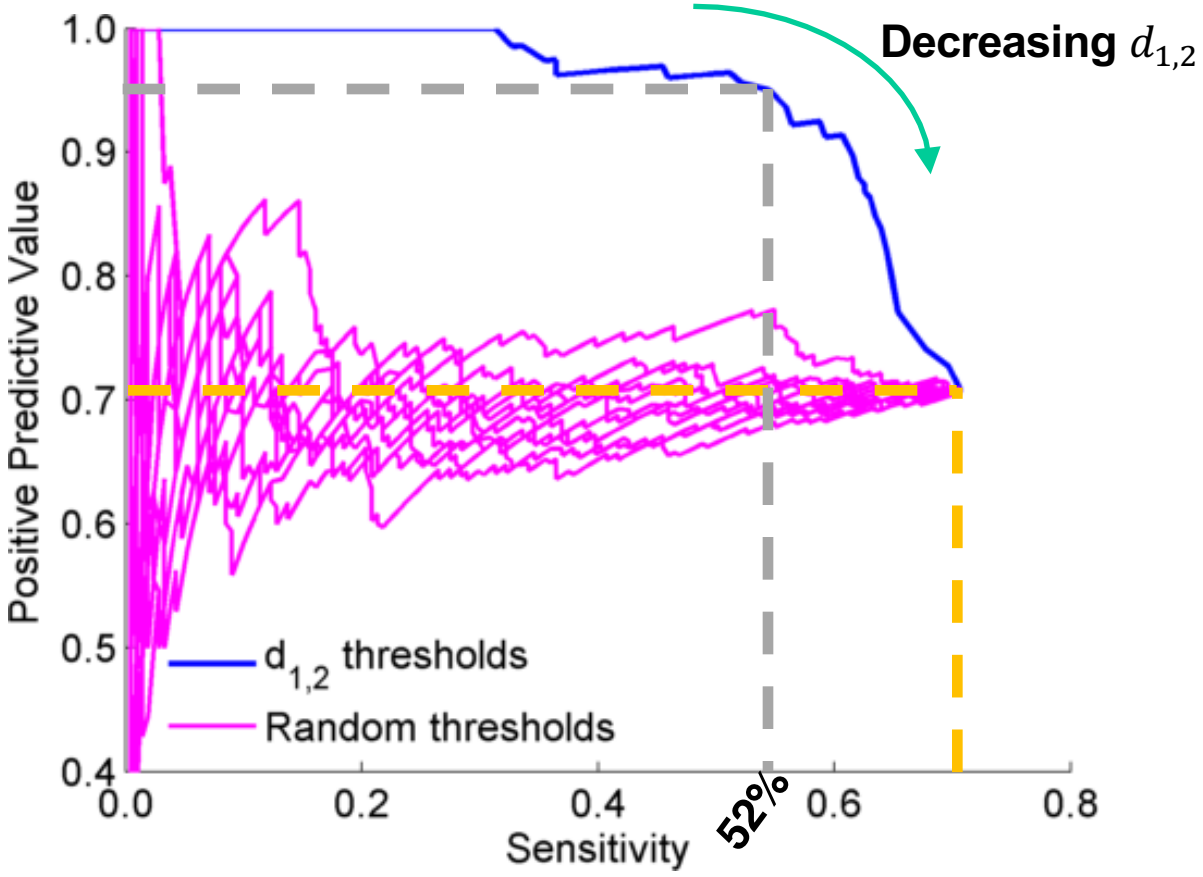
- Attacker arbitrarily selects eQTLs with association strength above 10
- 70% of the individuals are linked correctly
- But which 70%?
- Is there a way to differentiate between linkings to distinguish their reliability?
- First Distance Gap:
  - Difference between the genotype distance of second best matching and best matching individuals
  - $d_{1,2} = d_{second} - d_{first}$

- Attacker arbitrarily selects eQTLs with association strength above 10
- 70% of the individuals are linked correctly
- But which 70%?
- Is there a way to differentiate between linkings to distinguish their reliability?
- First Distance Gap:
  - Difference between the genotype distance of second best matching and best matching individuals
  - $d_{1,2} = d_{second} - d_{first}$

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

60 ▪ Lectures.GersteinLab.org

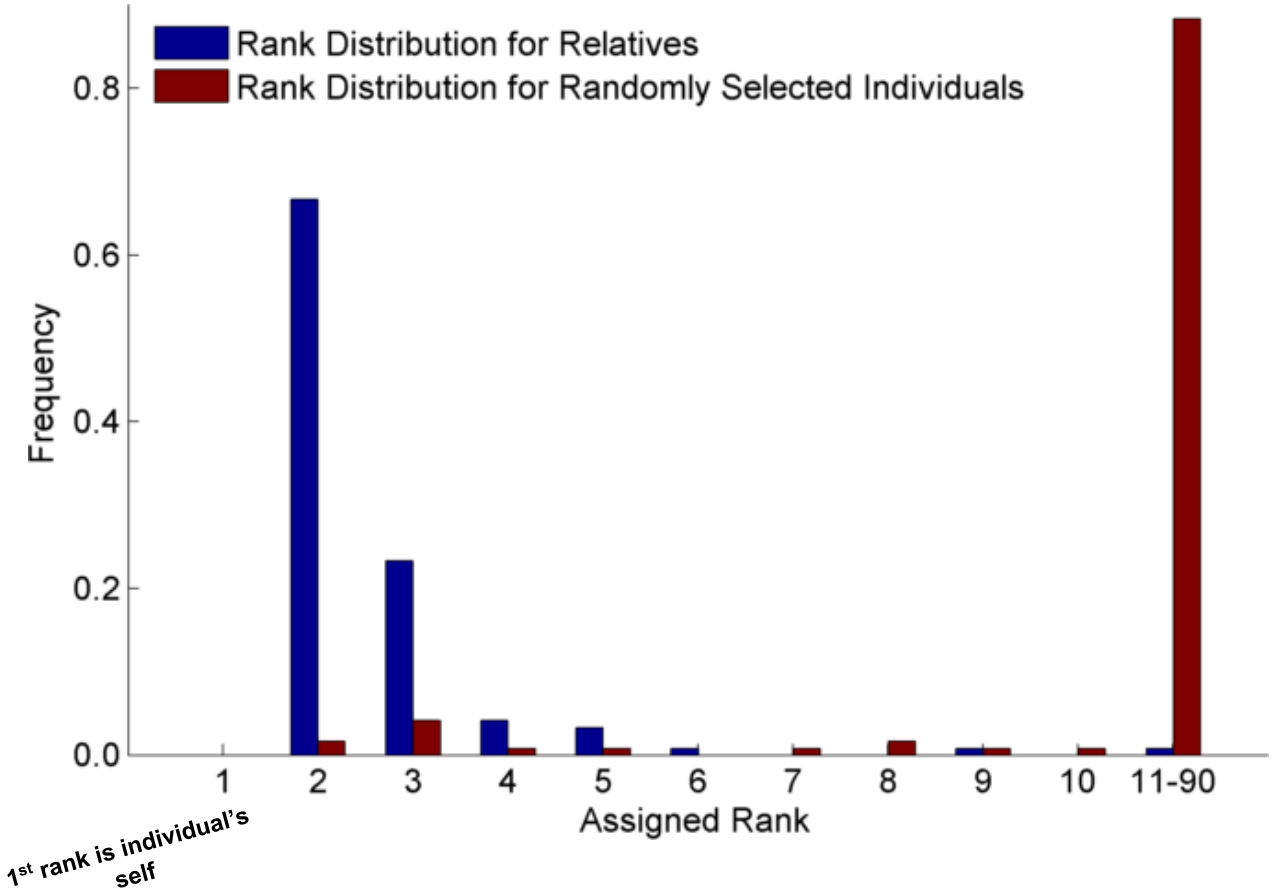
Sensitivity vs PPV for Linkings selected per f*irst distance gap,* $d_{1,2}$



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

Lectures.GersteinLab.org

Relatives are also vulnerable (30 CEU Trios)



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

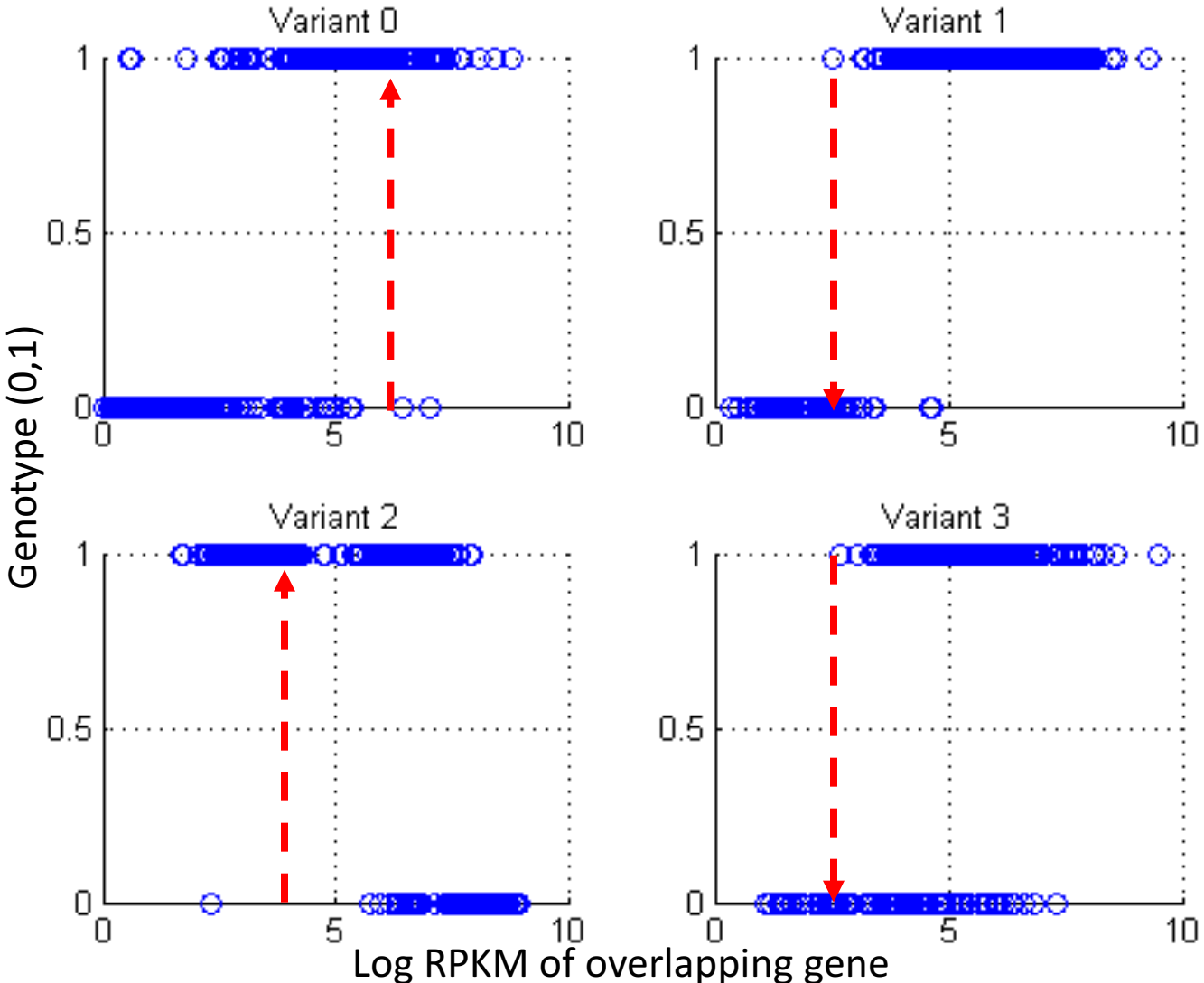
62 - Lectures.GersteinLab.org

Small Data Leakage from just Gene Expression Data: 4 eQTL-SNP genotypes

Example: Vulnerable sample variants, expressions

- Variant 0 (1, 6)
- Variant 1 (0, 2)
- Variant 2 (1, 3)
- Variant 3 (0, 2)

Expression levels are outliers and are predictive of the genotype!



Large-scale Transcriptome Mining: Building Integrative Regulatory Models, while Protecting Individual Privacy

- Transcriptome analysis **data**
  - Comparative ENCODE – Lots of Matched Data
  - 1000G+Geuvadis for privacy
- **Expression Clustering**, Cross-species
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones.
  - Also, conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

- The General **Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies & burdensome security
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
  - Details on Relevant Hacks: Genomic, Computer Security, & Netfix

- **RNA-seq: How to Publicly Share Some of it**
  - Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels + eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
  - Quantifying accuracy of prediction, via gap between best & 2nd best match

Large-scale Transcriptome Mining: Building Integrative Regulatory Models, while Protecting Individual Privacy

- Transcriptome analysis **data**
  - Comparative ENCODE – Lots of Matched Data
  - 1000G+Geuvadis for privacy
- **Expression Clustering**, Cross-species
  - Potts-model optimization gives 16 conserved co-expression modules (which can potentially annotate ncRNAs/TARs)
- **State Space Models** of Gene Expression
  - Using dimensionality reduction to help determine internal & external drivers
  - Decoupling expression changes into those driven by worm-fly conserved genes vs species-specific ones.
  - Also, conserved genes have similar canonical patterns (iPDPs) in contrast to species specific ones (Ex of ribosomal v signaling genes)

- The General **Dilemma of Genomic Privacy**
  - Fundamental, inherited info that's very private v need for large-scale mining for med. research
  - Issues w/ current social & tech approaches: inconsistencies & burdensome security
  - Strawman Hybrid Soc-Tech Proposal (Cloud Enclaves. Quantifying Leaks & Closely Coupled priv.-public datasets)
  - Details on Relevant Hacks: Genomic, Computer Security, & Netfix

- **RNA-seq: How to Publicly Share Some of it**
  - Presents a tricky privacy issue since much of the sequencing is for general, non-individual specific results yet it's tagged with individual information
  - Removing SNVs in reads w/ MRF
  - Quantifying & removing variant info from expression levels +
    eQTLs using ICI & predictability
  - Instantiating a practical linking attack using extreme expression levels
  - Quantifying accuracy of prediction, via gap between best & 2nd best match



DREISS.gersteinlab.org - D **Wang**, F He, S Maslov

papers.gersteinlab.org/subject/privacy - D **Greenbaum**

PrivaSeq.gersteinlab.org - A **Harmanci**

github.com/gersteinlab/**OrthoClust** - K **Yan**, D Wang, J Rozowsky, H Zheng, C Cheng

Acknowledgements

Hiring Postdocs.
See gersteinlab.org/jobs !

Extra

Info about content in this slide pack

- General PERMISSIONS
  - This Presentation is copyright Mark Gerstein, Yale University, 2016.
  - 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