What I did in the summer?

KKY
Group Meeting (summer, 2015)

Somethings I did

- How the spatial organization of genes shapes their expression patterns?

- A Bayesian framework for samples deconvolution

- Update/Introduction of the modERN (worm/fly) project

Somethings I did

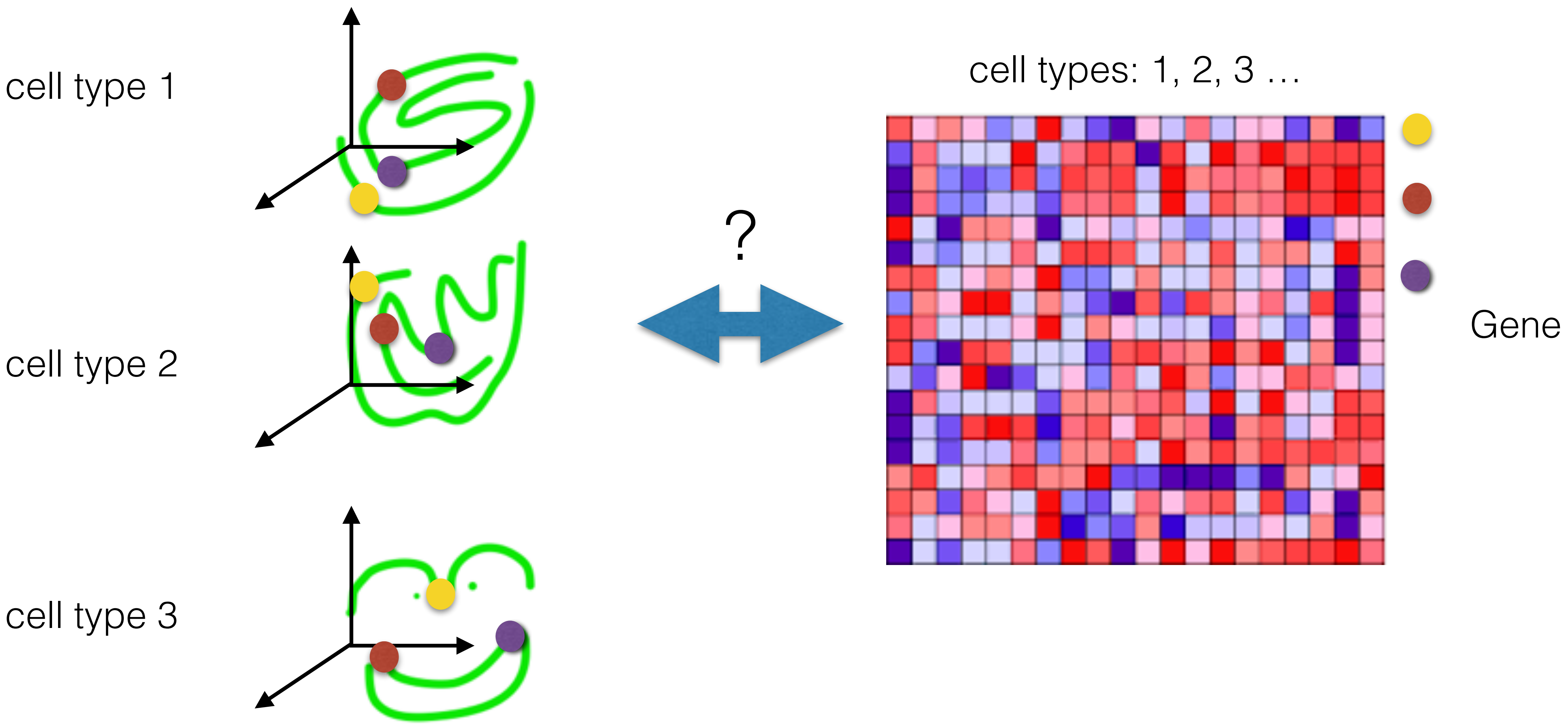
- **How the spatial organization of genes shapes their expression patterns?**
- A Bayesian framework for samples deconvolution
- Update/Introduction of the modERN (worm/fly) project

A fundamental question on gene regulation

- How come different kinds of cells or tissues have the same genome, but different expression profiles?
  - binding of transcription factors
  - nucleosomes positioning, histone marks
  - enhancers, networks …
  - spatial organization

A mapping between 2 spaces

real physical space abstract expression space



ENCODE3 Hi-C data

- Dekker Lab

- 12 completed cell lines: A549, Caki2, G401, LnCAP, NCI-H460, Panc1, PRMI-7951, SJCRH30, SK-MEL-5, SK-N-DZ, SK-NM-C, T470. 2 replicates per cell lines

- Contact maps binned in different sizes: 10mb, 2.5mb,1mb, 500kb, 250kb, 100kb, 40kb

- Raw counts and “ICED”

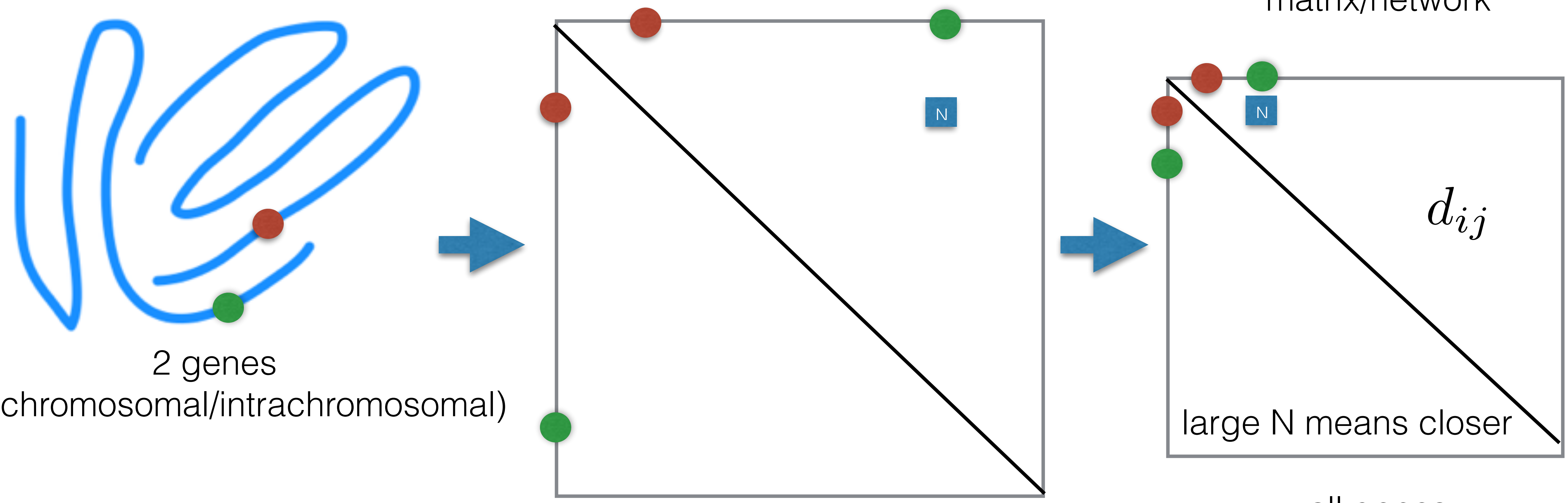
- In progress: HAc (AdrenocorPcal carcinoma) HA-s (Astrocytes spinal cord) HBVP (Brain vascular pericytes) DLD1 (Colon epithelial), ACHN (Kidney epithelial), HHSEC (HepaPc sinusoidal endothelial), HBMEC (BrainMicrovascularendothelial), HCMEC (Immortalized HBMEC)

Thoughts on Hi-C data

- Go into details: Identify the statistical significant contacts. Enhancer-target prediction. Interplay with other chromatin features.

- System-wide perspective: To understand the contacts as a whole

A simple construction: Gene-Gene Proximity Network



2 genes (interchromosomal/intrachromosomal)

Hi-C contact matrix

N

Genomics coordinates
100kb resolution, ICED

Gene-Gene proximity matrix/network

N

d_{ij}

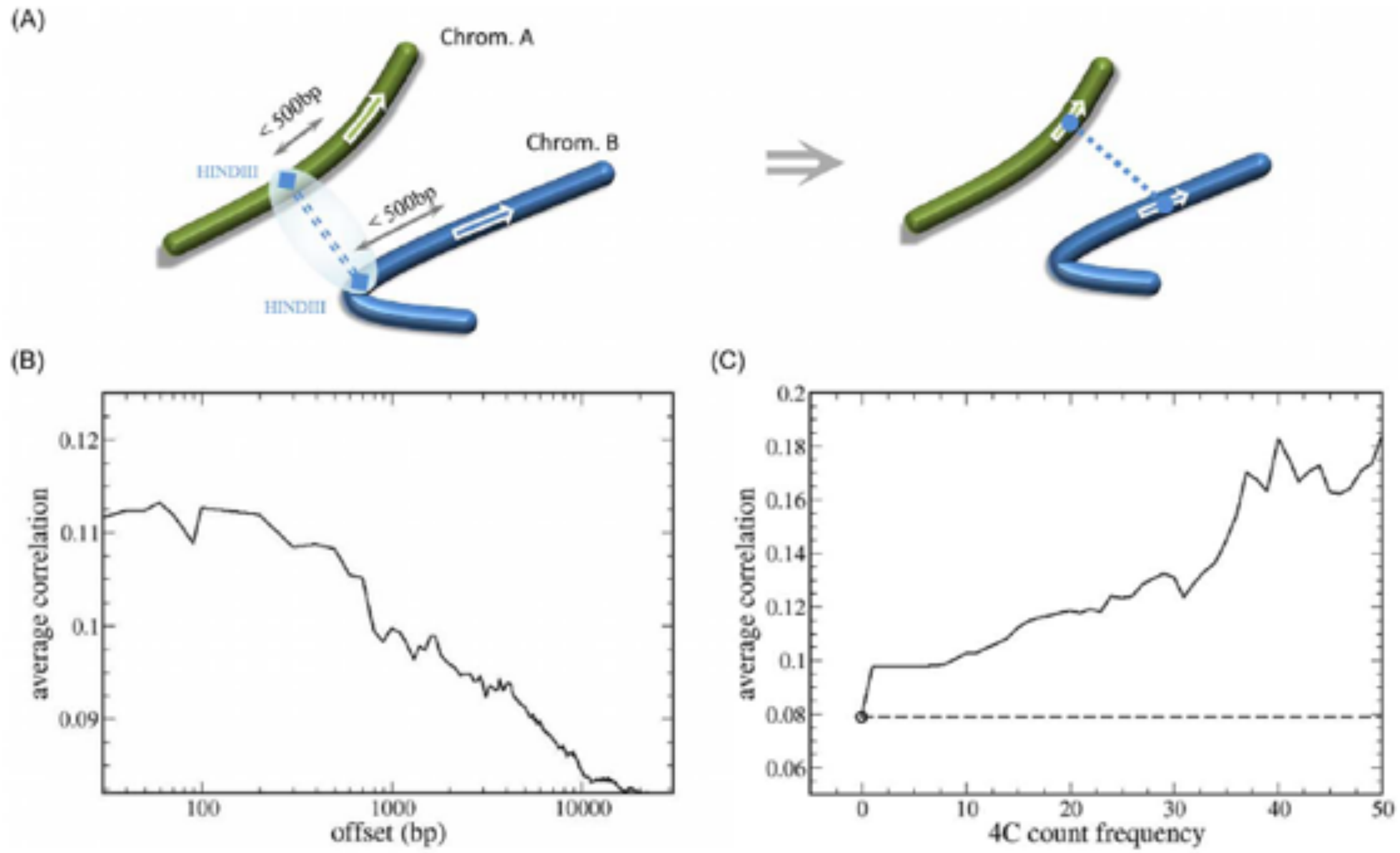
large N means closer

all genes

Example: A549
19100 genes
14% of gene pairs
are connected

Gene-Gene proximity versus Gene-Gene expression

- Many evidences showing that co-expressed genes tend to be sit next to each other in the genome (1D) as well as spatially close together (3D).

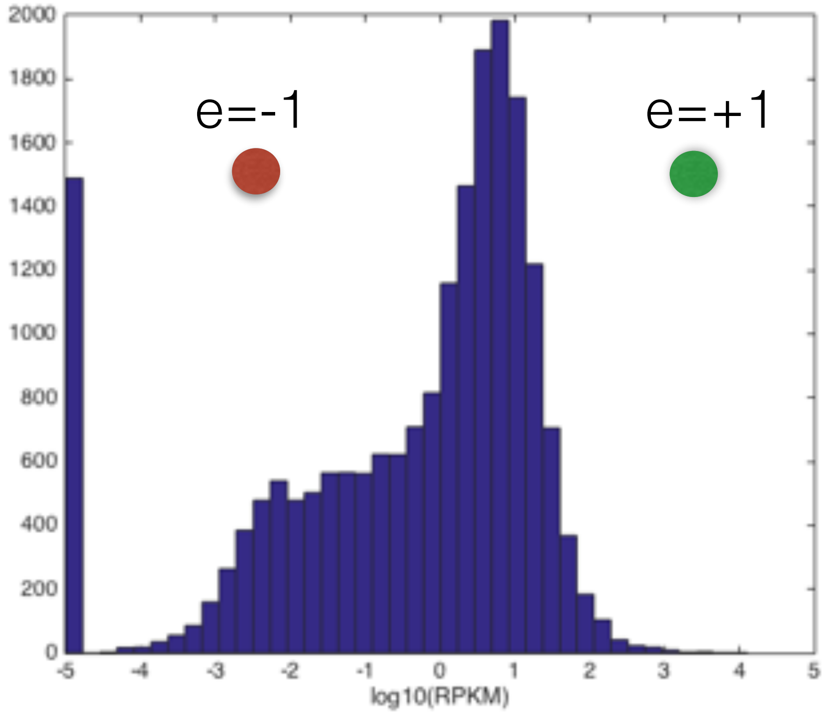


drawback:
spatial structure in one cell type correlates with expression profiles across many cell types

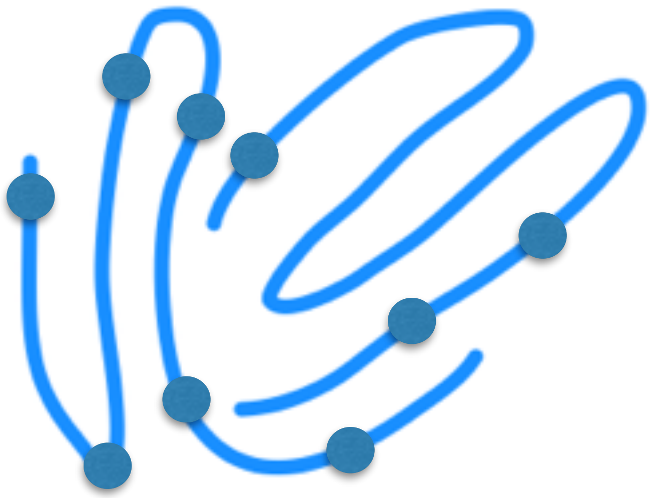
Homouz and Kudlicki, PLoS One, 2013

Gene-Gene proximity versus Gene-Gene expression

expression pattern of A549

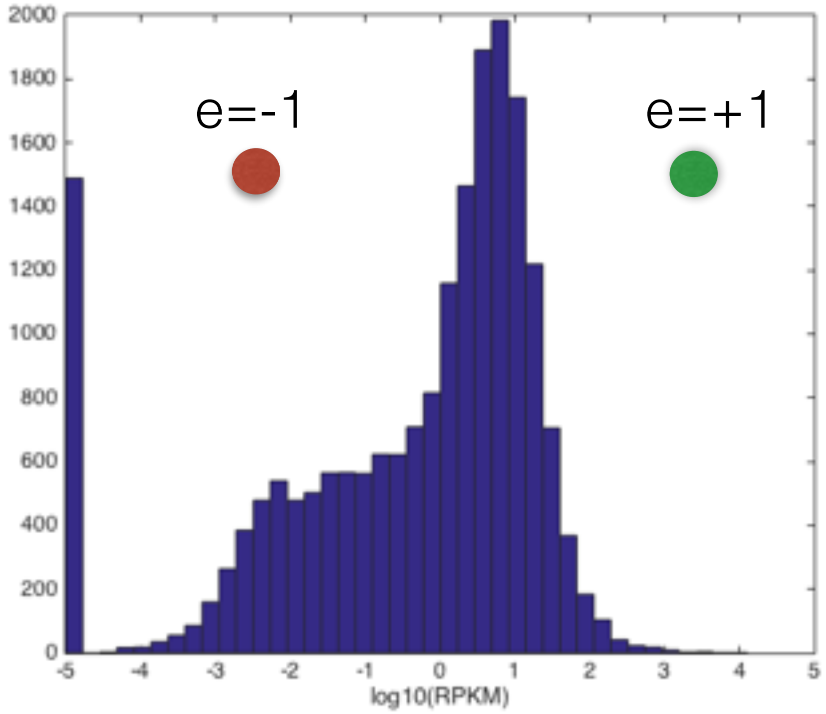


spatial structure of A549

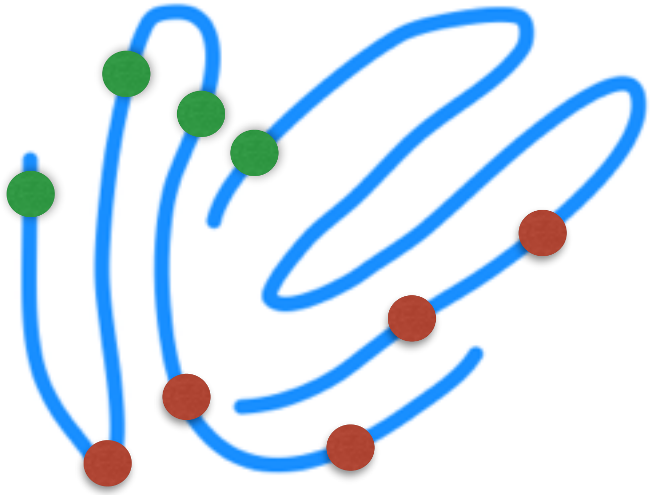


Gene-Gene proximity versus Gene-Gene expression

expression pattern of A549

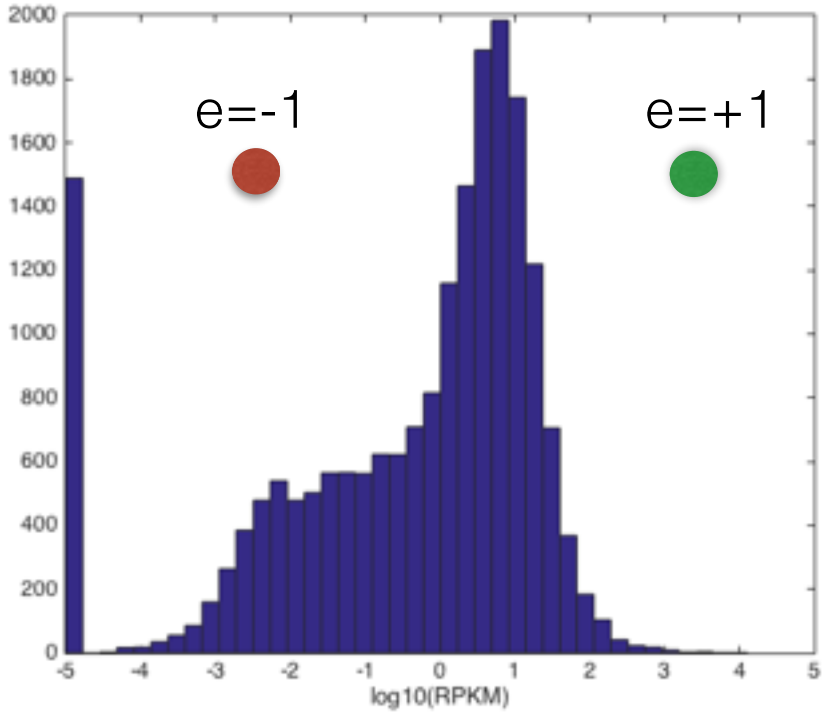


spatial structure of A549

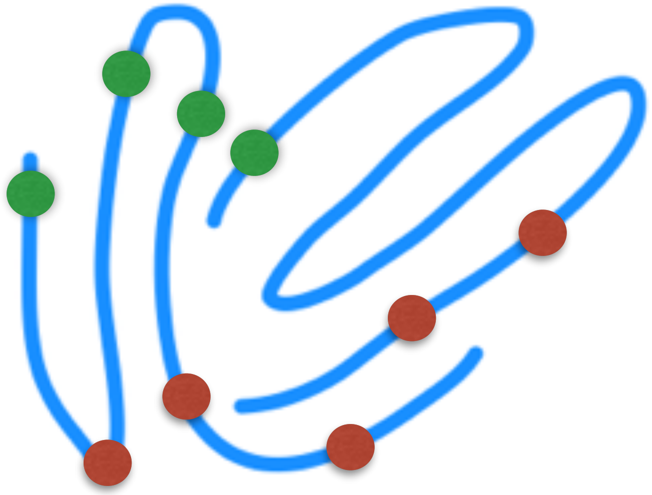


Gene-Gene proximity versus Gene-Gene expression

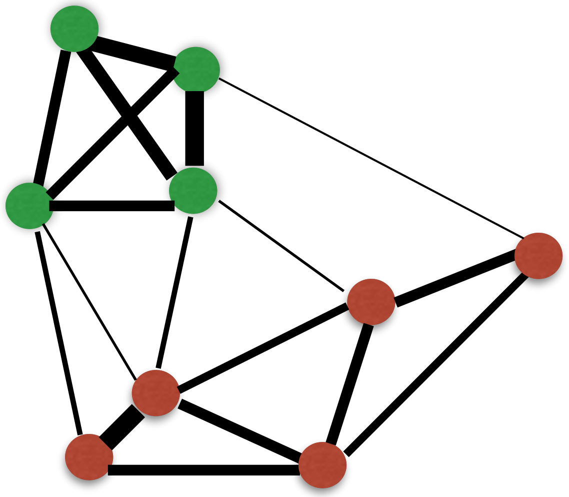
expression pattern of A549



spatial structure of A549



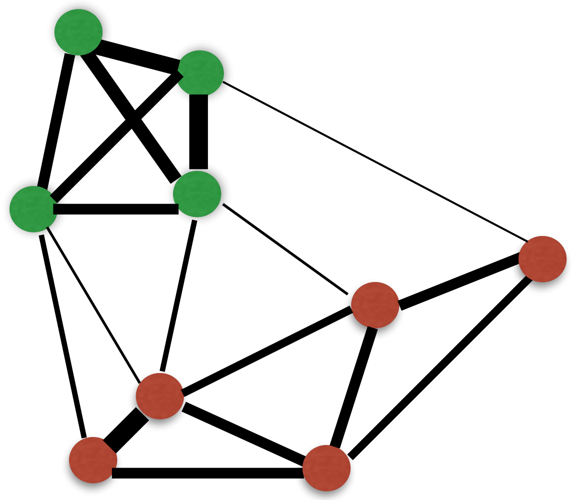
proximity network of A549



Graph partition (bisection) problem

Consider a graph G = (V, E), where V denotes the set of n vertices and E the set of edges. The objective is to partition G into k (k=2) components while minimizing the weights of the edges between separate components.

proximity network of A549



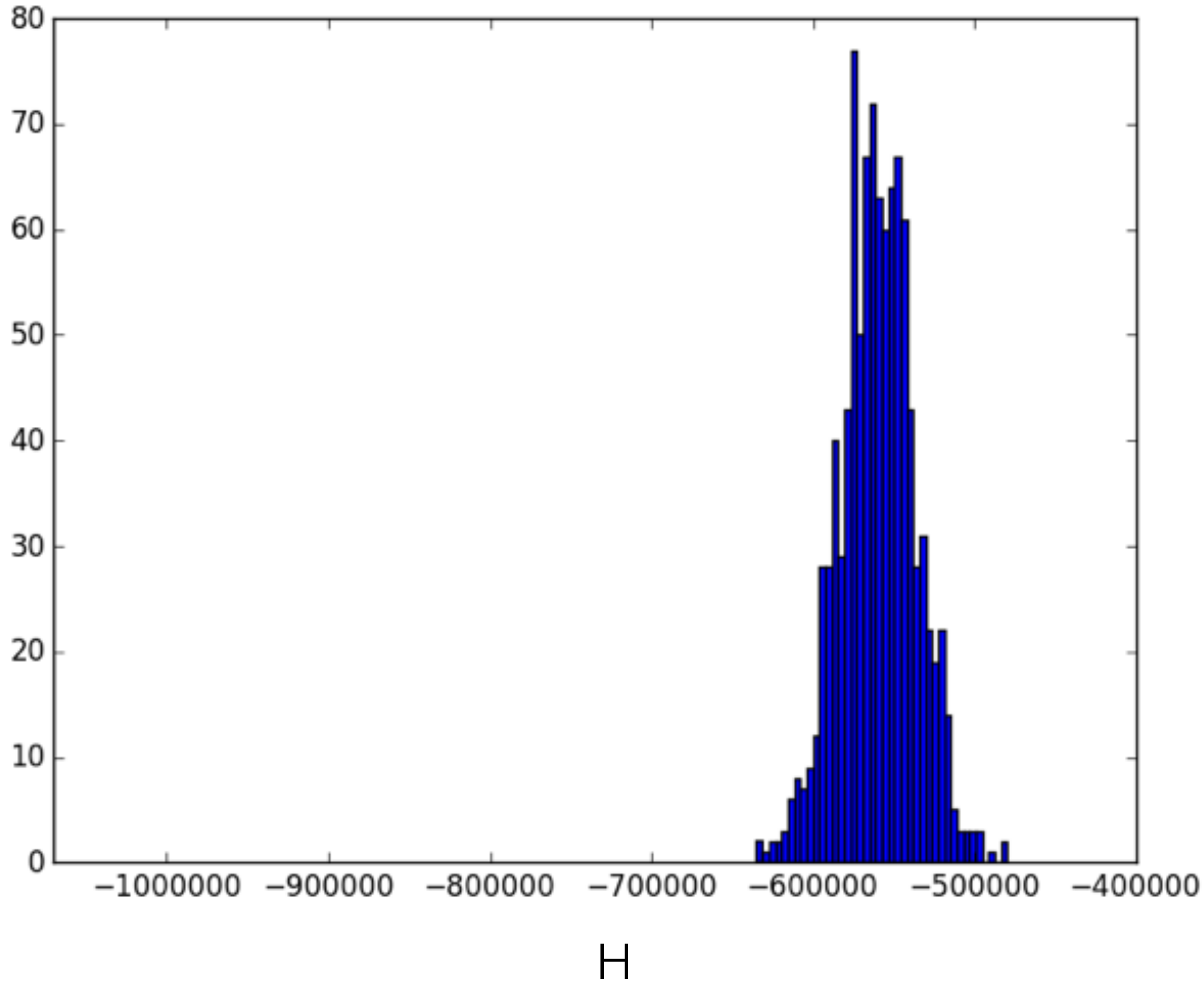
$$H = -\sum_{ij} d_{ij} e_i e_j$$

d is the weighted adjacency matrix and e=+1 or -1

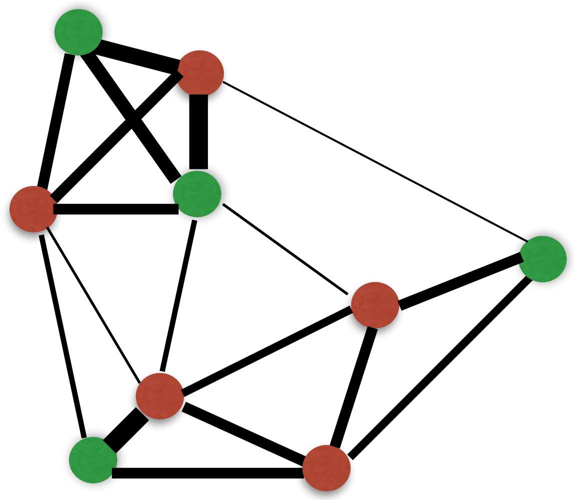
a low energy state means co-expressed genes are co localized

Gene-Gene proximity versus Gene-Gene expression

Distribution of H by shuffling the expression profile of A549

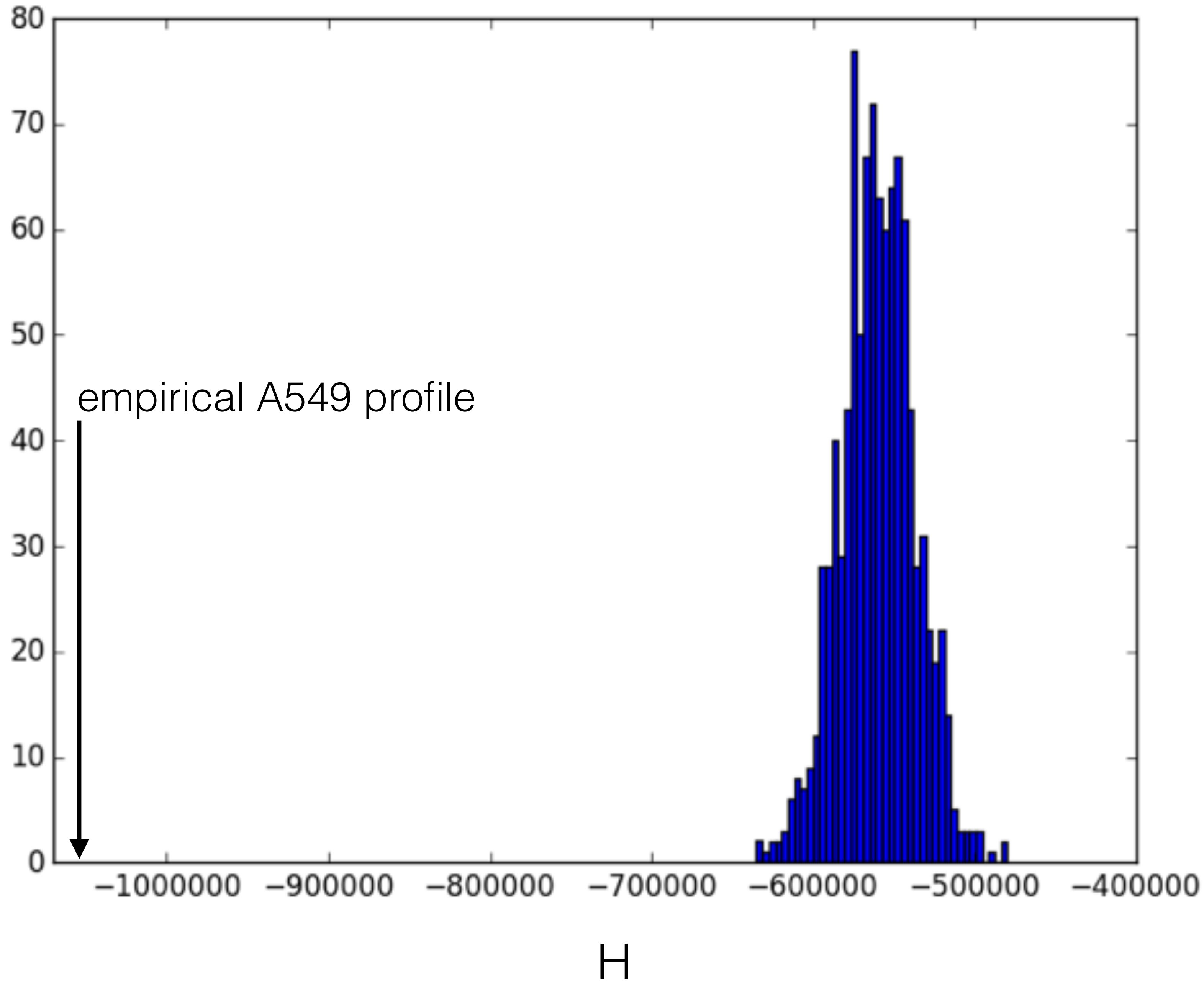


N nodes:
m is expressed, n is not

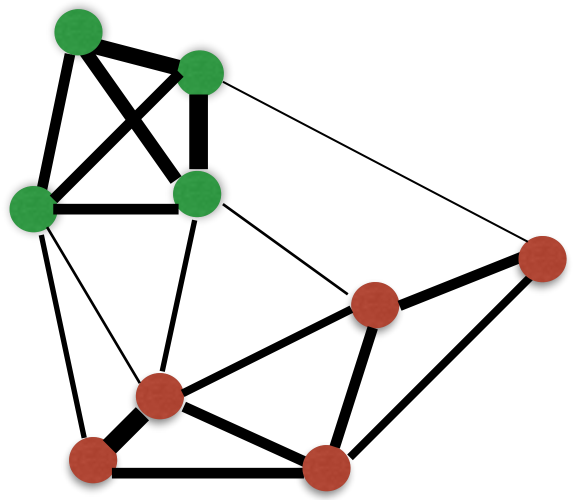


Gene-Gene proximity versus Gene-Gene expression

Distribution of H by shuffling the expression profile of A549

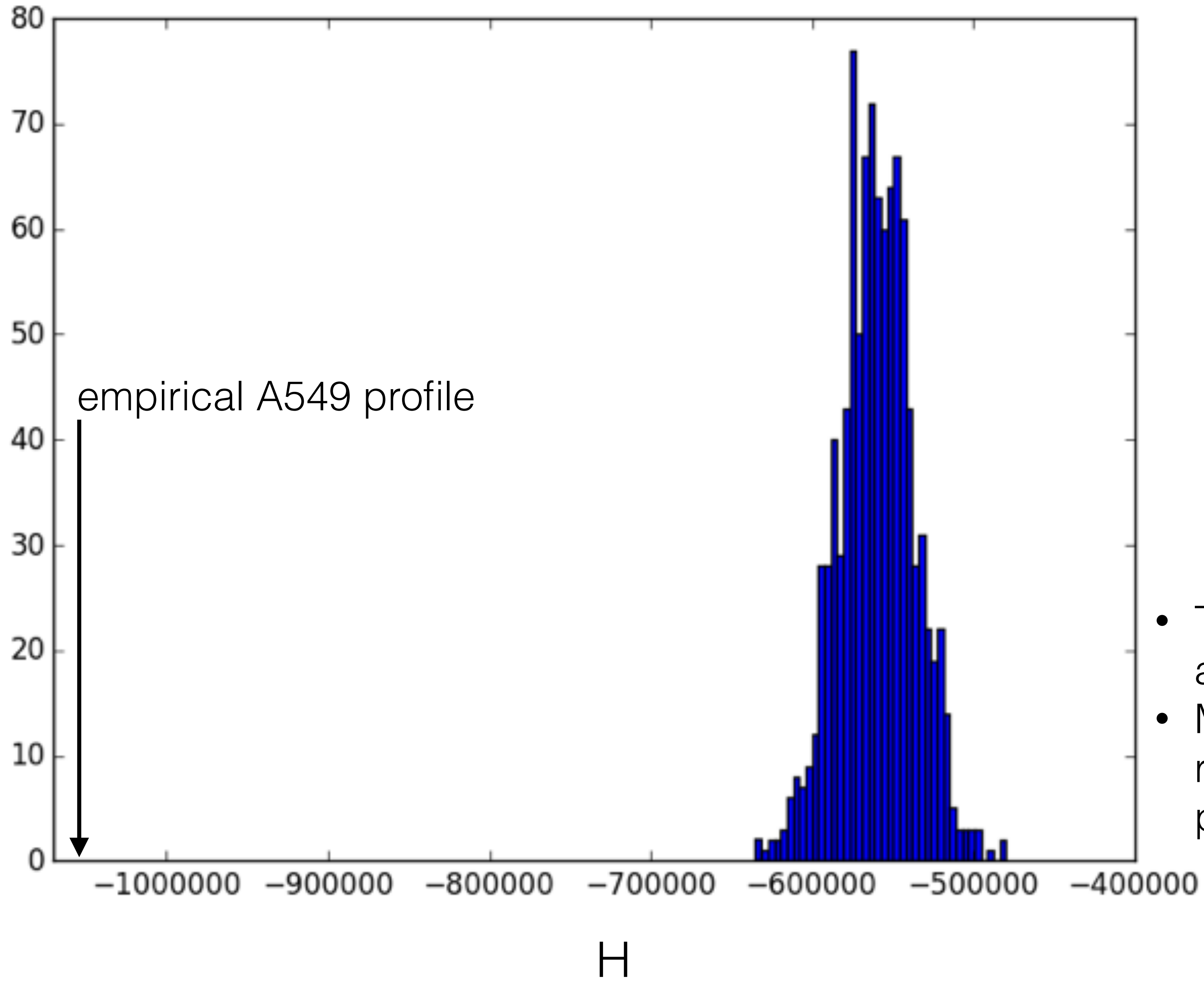


N nodes:
m is expressed, n is not



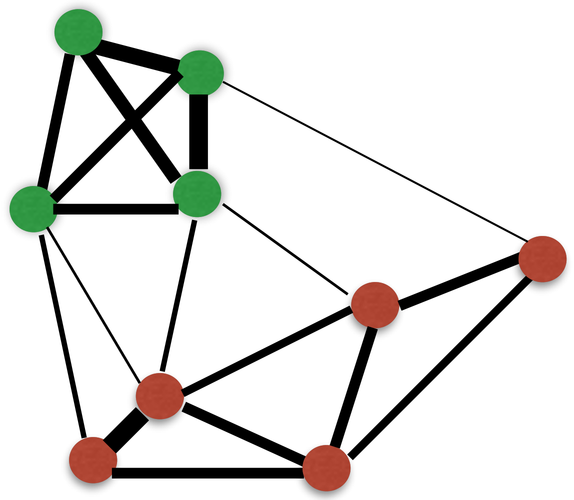
Gene-Gene proximity versus Gene-Gene expression

Distribution of H by shuffling the expression profile of A549



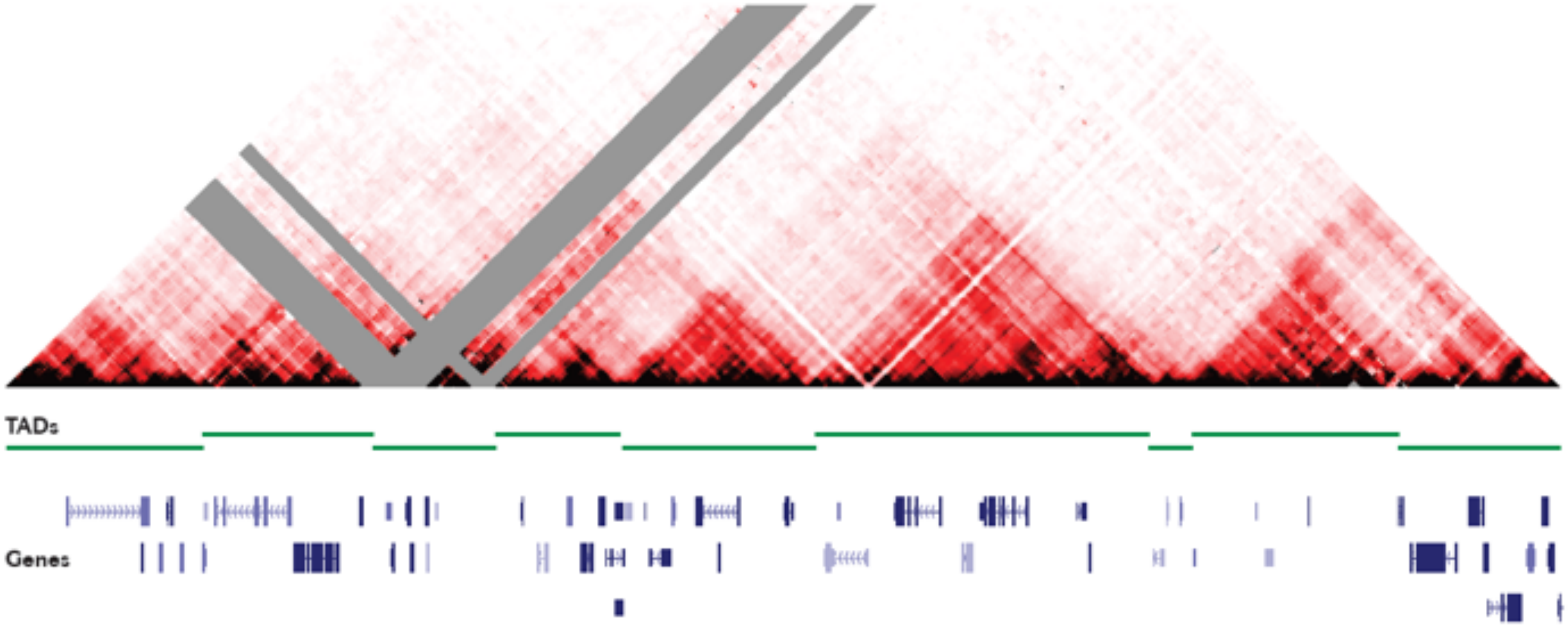
empirical A549 profile

N nodes:
m is expressed, n is not



- The spatial location of expressed genes are highly non-random.
- May be it's too naive to compare with random - perform shuffling while preserving other genomics features

In relationship with Topologically Associating Domains (TADs)

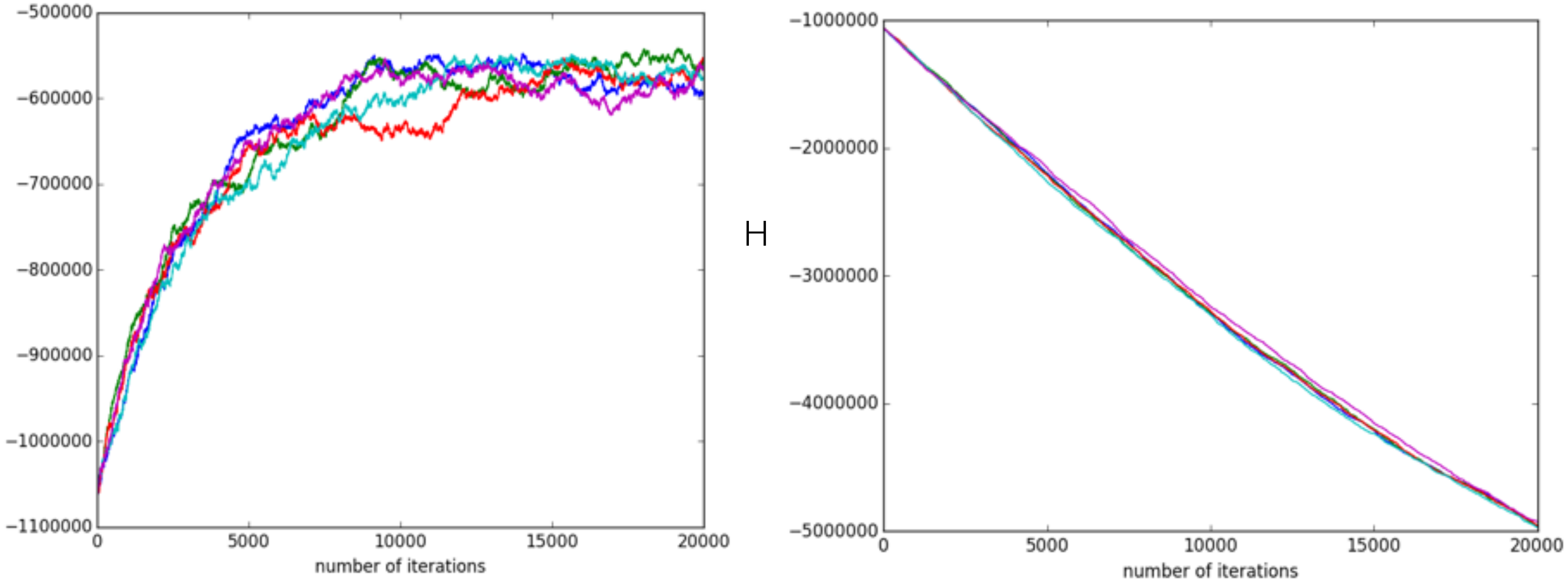


Dekker et al. Nat. Rev. Genetics 2013

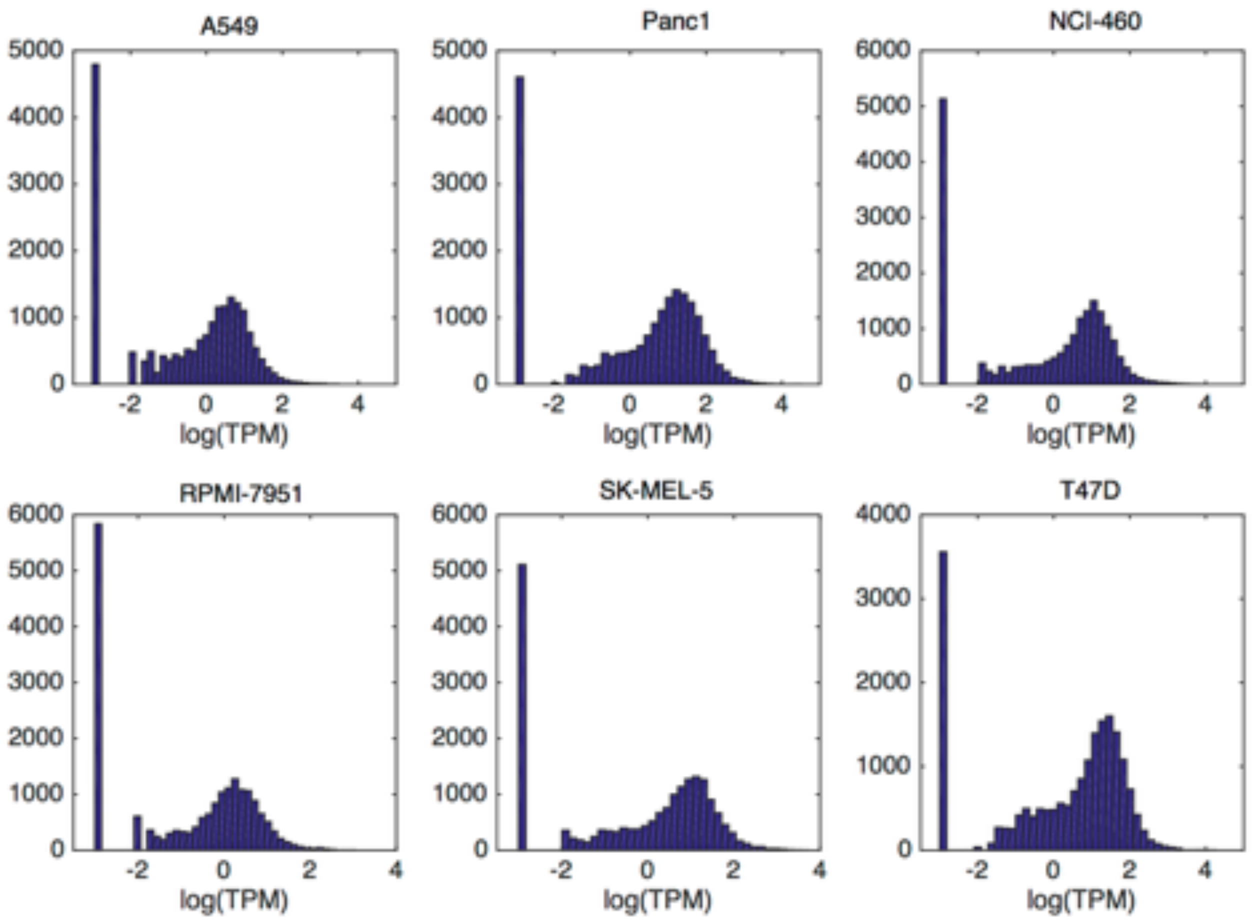
TADs are defined based on intra-chromosomal contacts.
Our approach takes into account of inter-chromosomal contacts.

Is the expression profile optimal?

Given a spatial configuration, the observed expression profile has a much lower energy than random, but is it optimal?

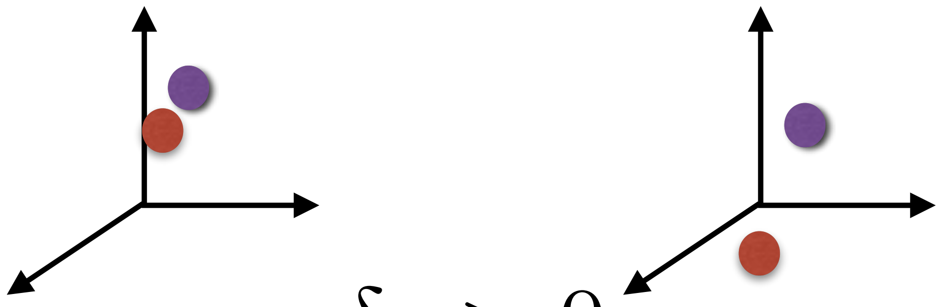


Gene-Gene proximity versus Gene-Gene expression



cell type I

cell type II

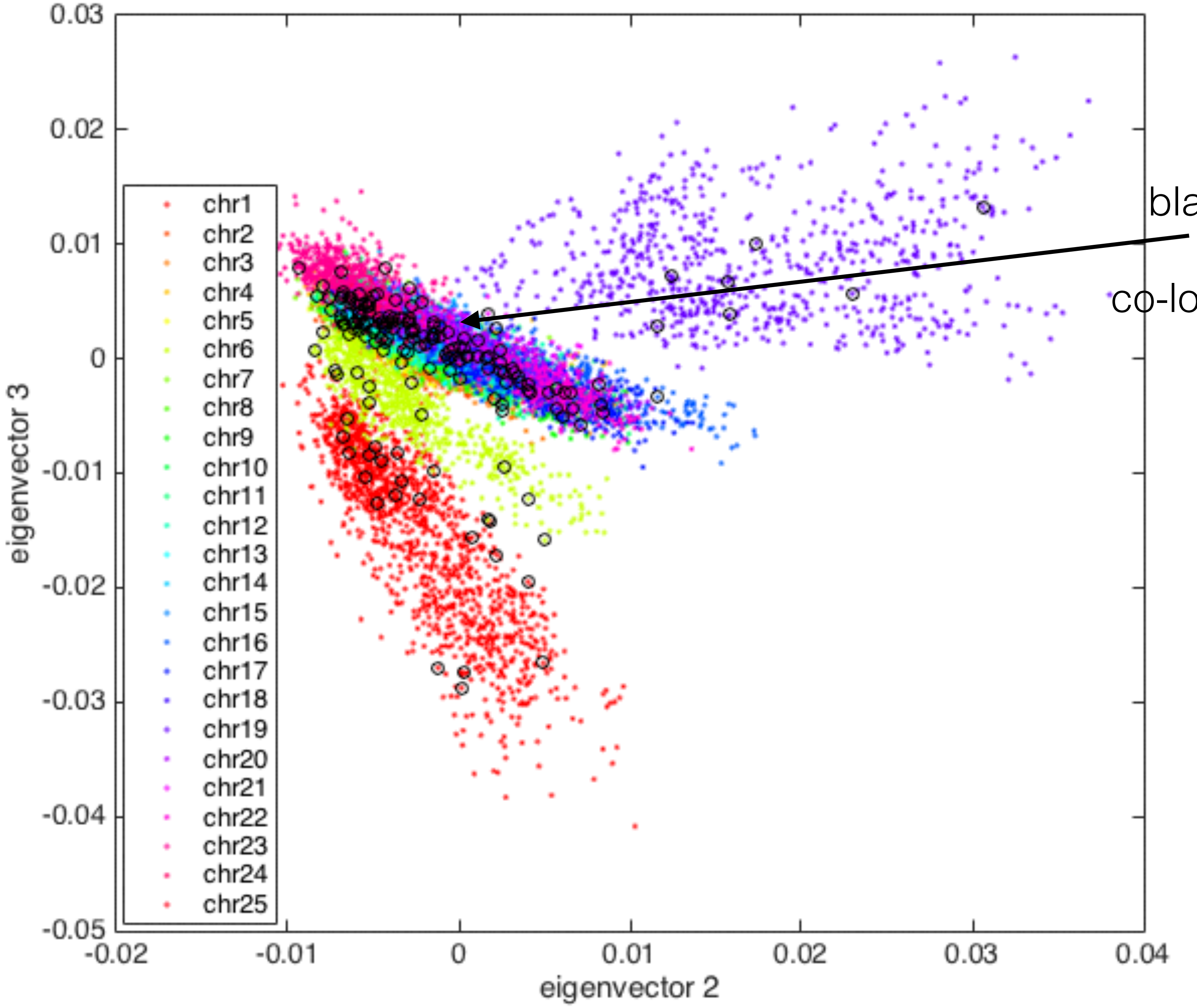


$\delta r > 0$

	I	II
🔴	ON	ON
🟣	ON	OFF

$\delta e > 0$

Targets of transcription factors in the Gene-Gene proximity network



black: targets of Jun, based on literature

co-localized targets - transcription factories?

Standard spectral clustering:
Project the network onto a few eigenvectors of the diffusion matrix.

Comparison of GGP networks between 12 cell types

- Gene-Gene proximity, conserved? specific?
  - what’s the proper distance metric?
- We have been working on the comparison of networks:
  - Network rewiring - addition/removal of nodes, edges
  - OrthoClust, multi-layers network clustering
  - Compare regulatory networks of worm, fly, human
  - BrainSpan, co-expression networks in different parts of the brain
  - Tissue specific PPI networks

On-going work

- Representing the spatial structure of a genome in a network offers a unified framework to integrate quite many existing data we have been working on.

- Expression data (graph partition), TF targets, histone marks, (may be other network properties)

- Network may help us to compare contact maps

Somethings I did

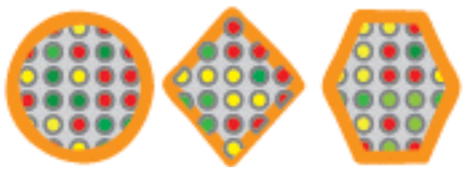
- How the spatial organization of genes shapes their expression patterns, or vice versa?

- **A Bayesian framework for samples deconvolution**

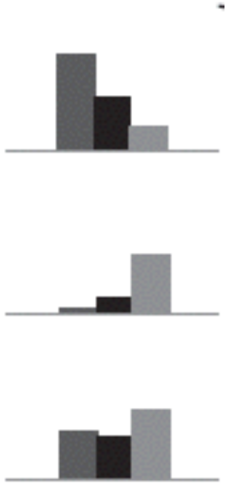
- Update/Introduction of the modERN (worm/fly) project

Samples deconvolution

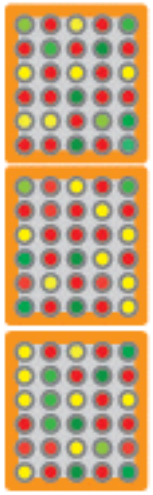
cell types specific expression profiles

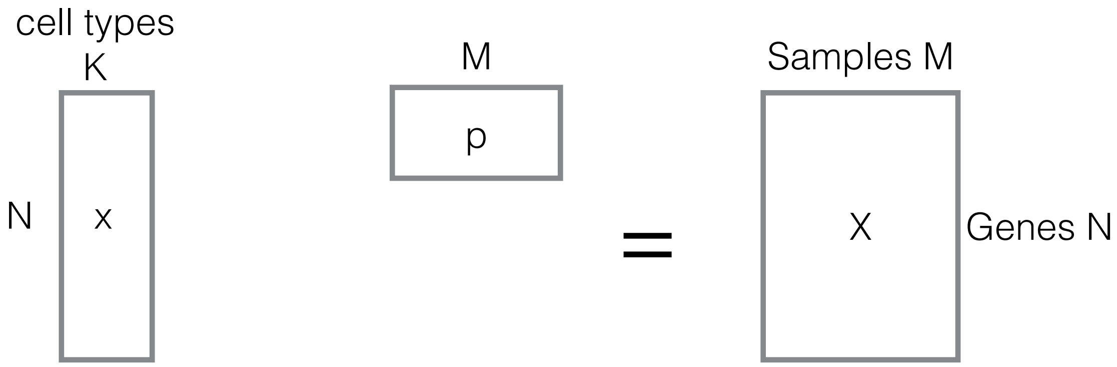


cell-type proportion



samples with convoluted expression profiles





cell types K

N

x

M

p

=

Samples M

X

Genes N

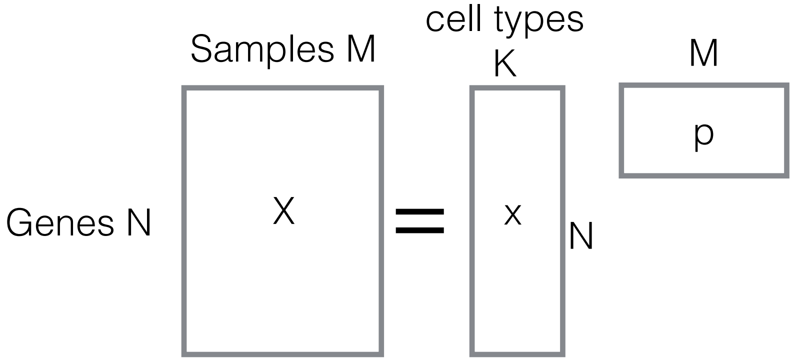
$$X_i^\mu = \sum_{k=1}^{K} p_k^\mu x_i^k + \zeta$$

If either one of x or p is known, inferring the other is essentially a quadratic programming problem by minimizing the function:

$$\sum_{\mu=1}^{M} \sum_{i=1}^{N} (X_i^\mu - \sum_{k=1}^{K} p_k^\mu x_i^k)^2$$

non-negative matrix factorization

$$X_i^\mu = \sum_{k=1}^{K} p_k^\mu x_i^k + \zeta$$



$$NM > NK + (K-1)M$$

decompose X into x and p but minimize $\sum_{\mu=1}^{M}\sum_{i=1}^{N}(X_i^\mu - \sum_{k=1}^{K} p_k^\mu x_i^k)^2$

subjected to constraints x_i, p_i >0 and $\sum p_i = 1$

existing algorithm: deconf: Repsilber et al. 2010
algorithms based on standard NMF alone, do not take into account of the prior information

a Bayesian framework

$$P(x,p|X) = \frac{P(X|x,p)}{P(X)} P(x,p)$$

prior determined by incorporating knowledge of cell types

$$P(X|x,p) \sim \exp(-H)$$

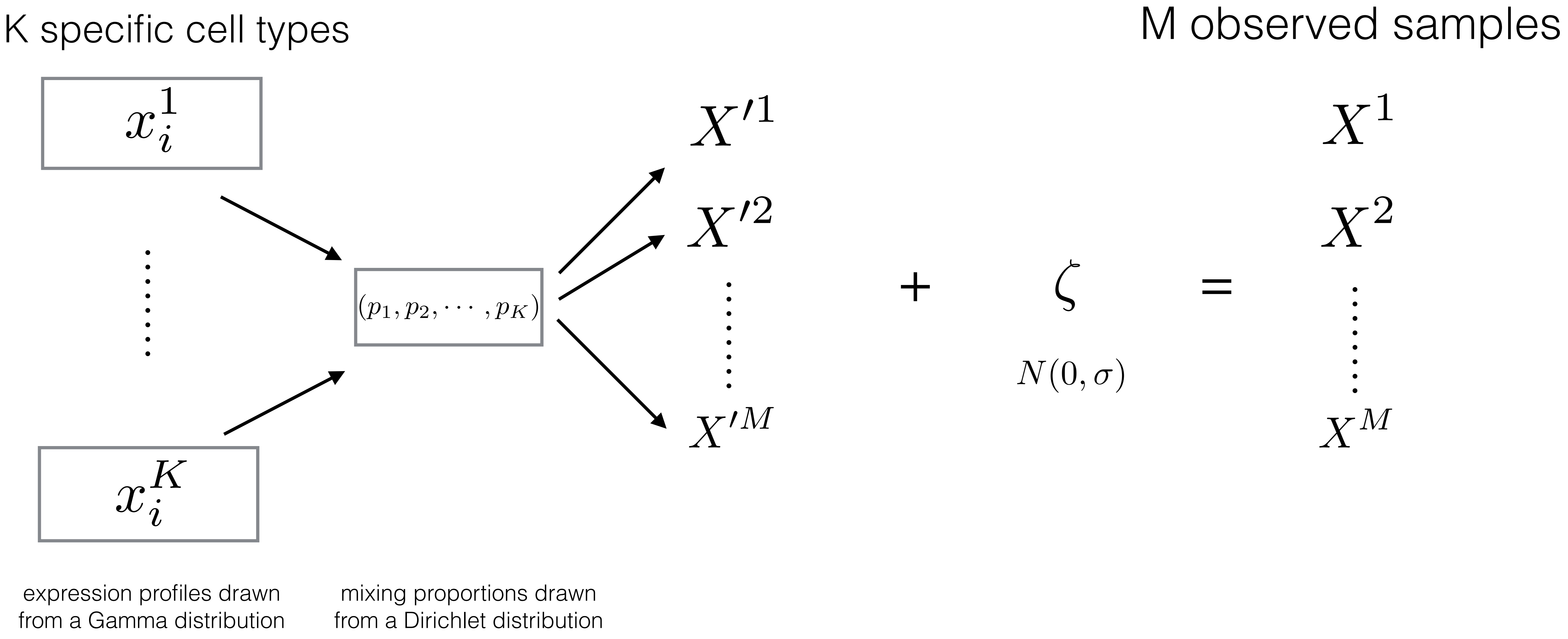
$$H = \sum_{\mu} \sum_{i} \frac{(X_i^{\mu} - \sum_k p_k^{\mu} x_i^k)^2}{2\sigma^2}$$

$$P(x) \sim Gamma()$$

$$P(p) \sim Dirichlet()$$

sample the posteri by MCMC, obtaining many (x,p) configurations, use the means $\hat{x}, \hat{p}$ as estimates of gold standards

A simulation



K specific cell types

M observed samples

x_i^1

x_i^K

$(p_1, p_2, \cdots, p_K)$

X'^1

X'^2

X'^M

$+$

ζ

$N(0, \sigma)$

$=$

X^1

X^2

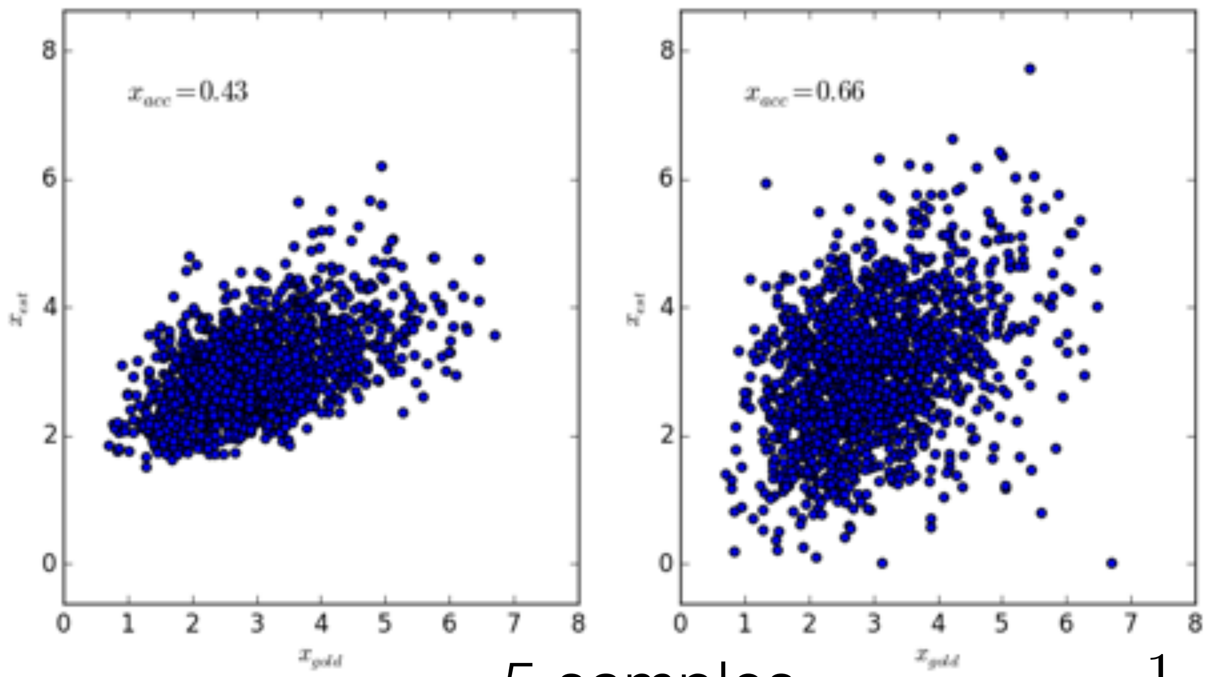
X^M

expression profiles drawn from a Gamma distribution

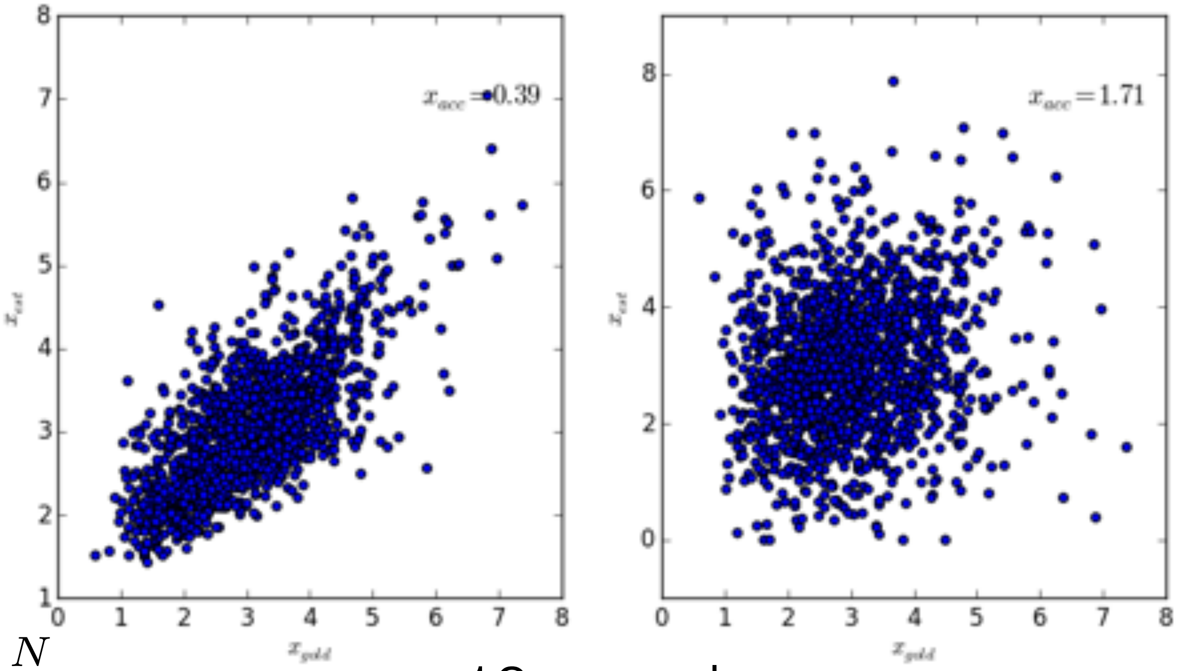
mixing proportions drawn from a Dirichlet distribution

Given the observable X, we want to infer x and p, and then compare with the original gold standards.

reconstruction of cell-type specific expression profiles: Bayesian versus deconf



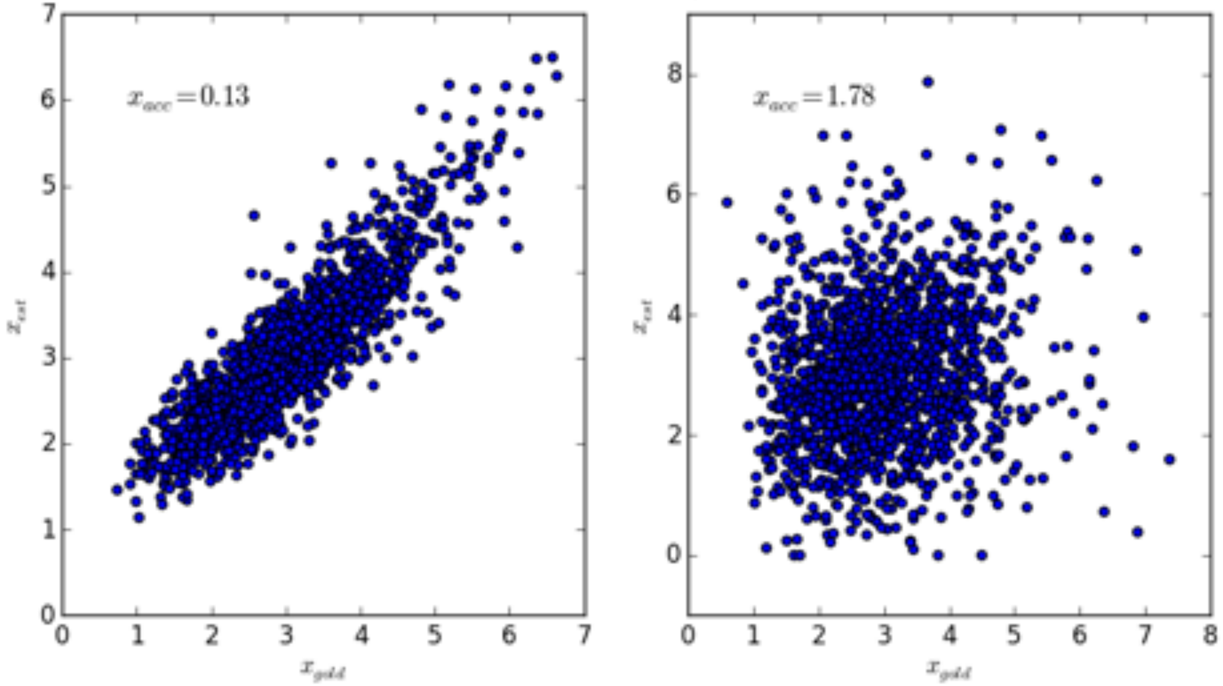
5 samples



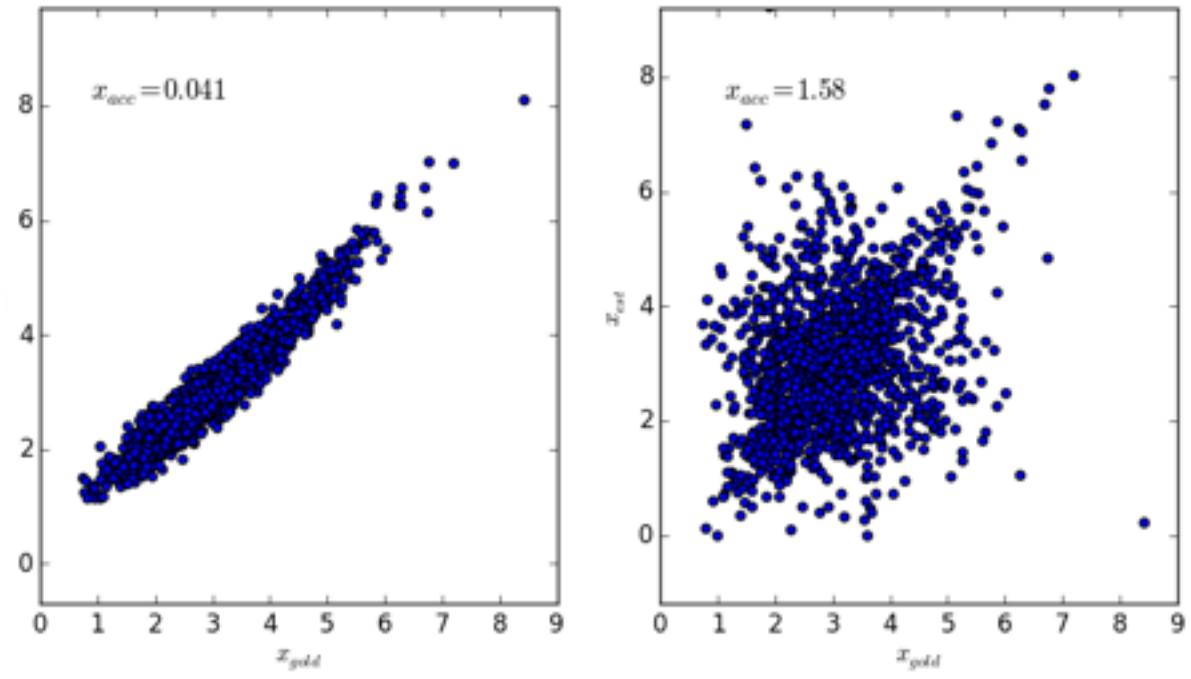
10 samples

$$x_{acc} = \frac{1}{NK} \sum_{k=1}^{K} \sum_{i=1}^{N} (x_i^k - \hat{x}_i^k)^2$$

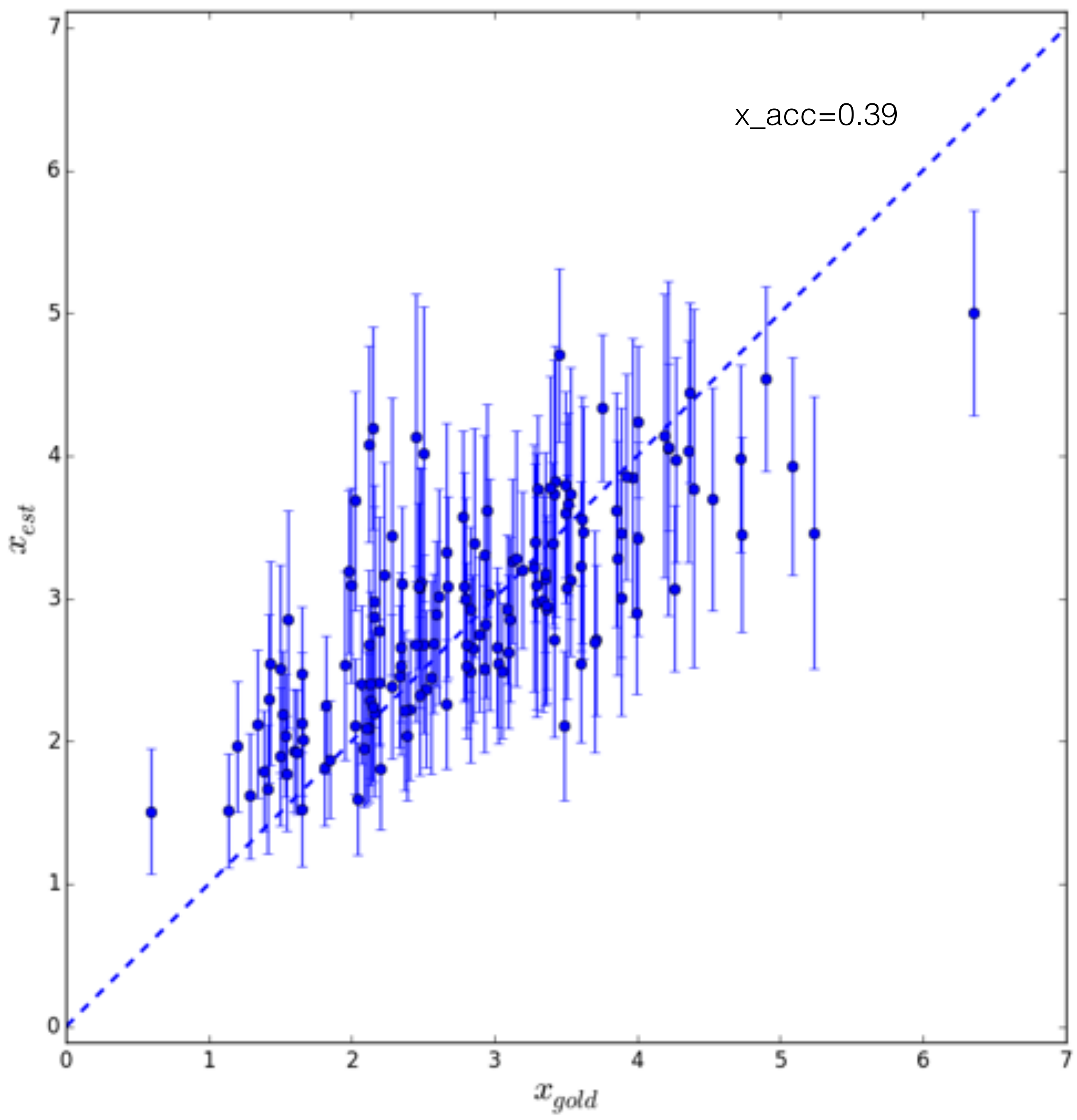
20 samples



30 samples



reconstruction of cell-type specific expression profiles: error estimate



10 samples

On-going work

- in principle, prior knowledge could improve deconvolution. but,

- for practical problems, which prior distributions should be used?

    - make sense in modeling gene expression, i.e. could well fit the data

    - some distributions are easier for MCMC, like certain conjugate priors

    - currently struggling with MCMC

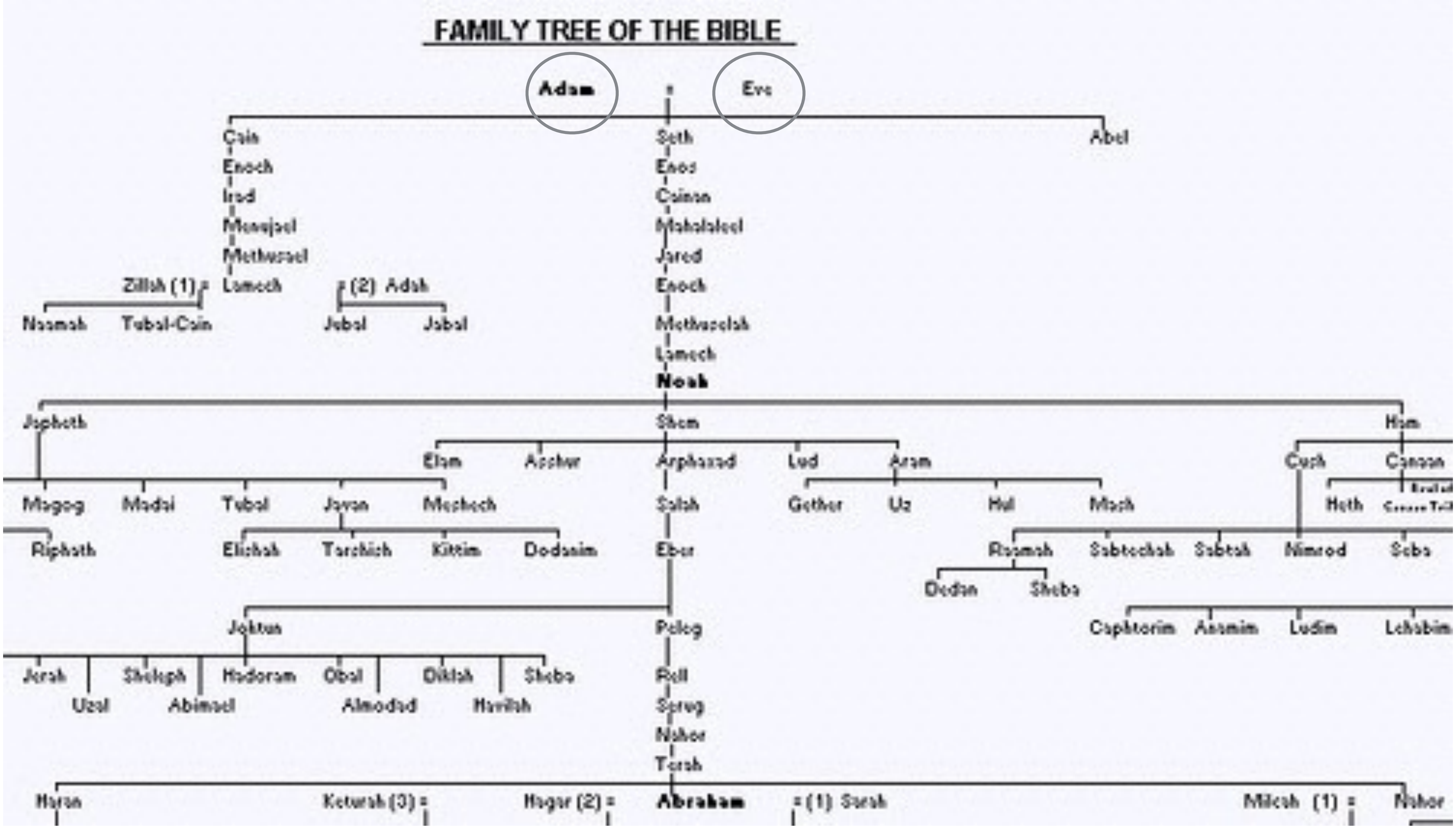
Somethings I did

- How the spatial organization of genes shapes their expression patterns, or vice versa?
- A Bayesian framework for samples deconvolution
- **Update/Introduction of the modERN (worm/fly) project**

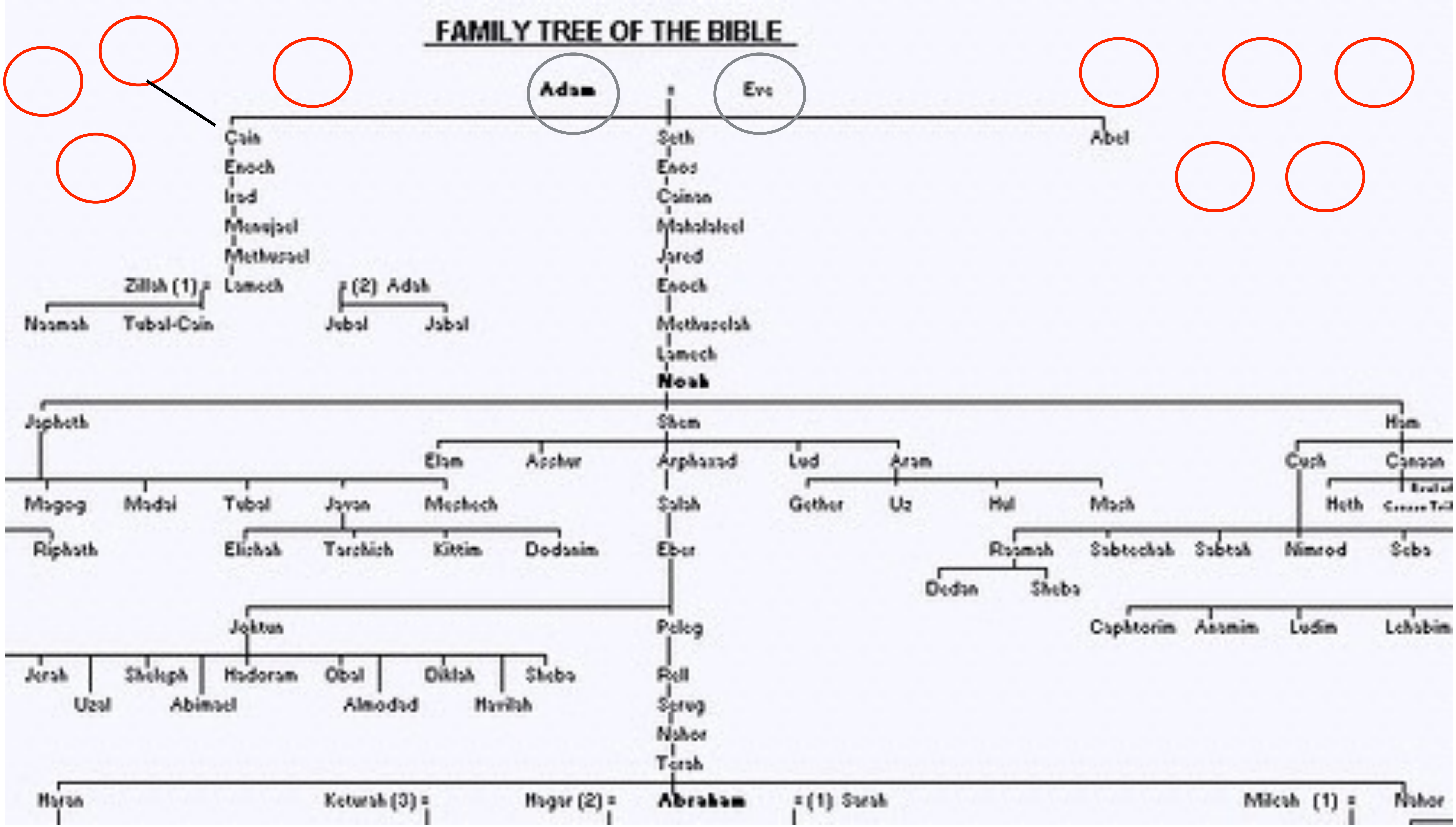
modERN (model organism Encyclopedia of Regulatory Networks)

- Currently,
  - worm: ~270 ChIP-Seq experiments in various stages, with a few stages have 40-70 TFs. Total 113 unique TFs (aim: 687).
  - fly: ~240 ChIP-Seq experiments in various stages. Total 170 unique TFs (aim: 703).
  - look at orthologs ~10 pairs
- In the future, ChIP-Seq profiles of more TFs, and RNA-Seq of ~100 TF-knockout mutants
- Compare regulatory networks

One more thing I did



One more thing I did



Acknowledgement

- How the spatial organization of genes shapes their expression patterns, or vice versa?
    - **ANS**
- A Bayesian framework for samples deconvolution
    - **DW, SKL**
- Update/Introduction of the modERN (worm/fly) project
    - **TG**