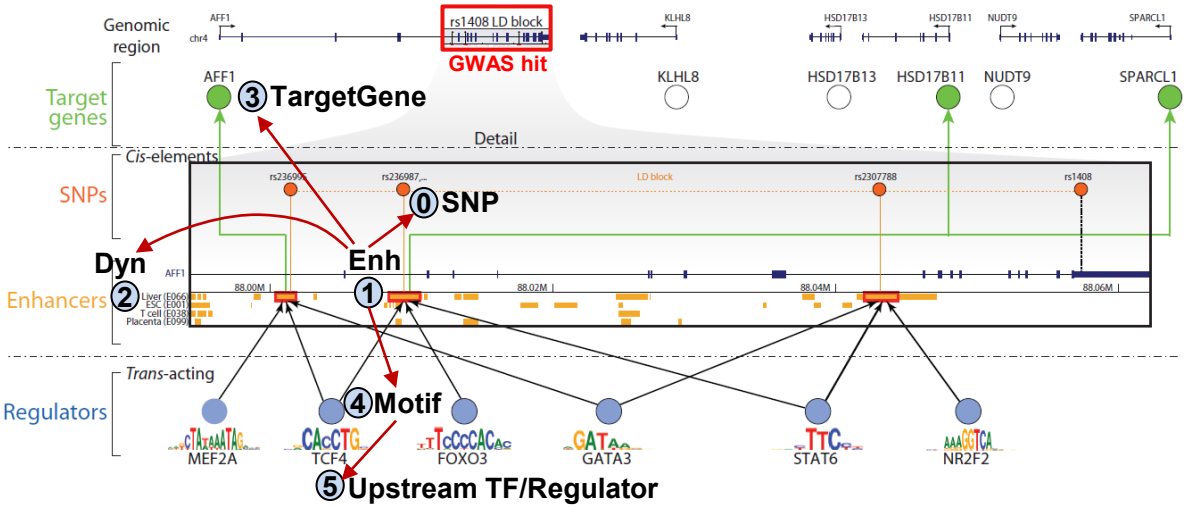
# MIT 18.700/18.701/6.047/6.878/HST.507
# Computational Biology: Genomes, Networks, Evolution




# Lecture 7: Regulatory circuitry

## Epigenome Dynamics: Joint Chromatin State Learning
## Enhancer-gene linking: Correlation, Hi-C, eQTLs
## TF motif discovery: Enrichment, EM, Gibbs Sampling
## Deep learning convolution CNNs for motif discovery
## Global motif discovery: Comparative Genomics
## Motif Instance Identification: Branch Length Score
## Regulatory region dissection: MPRA, HiDRA

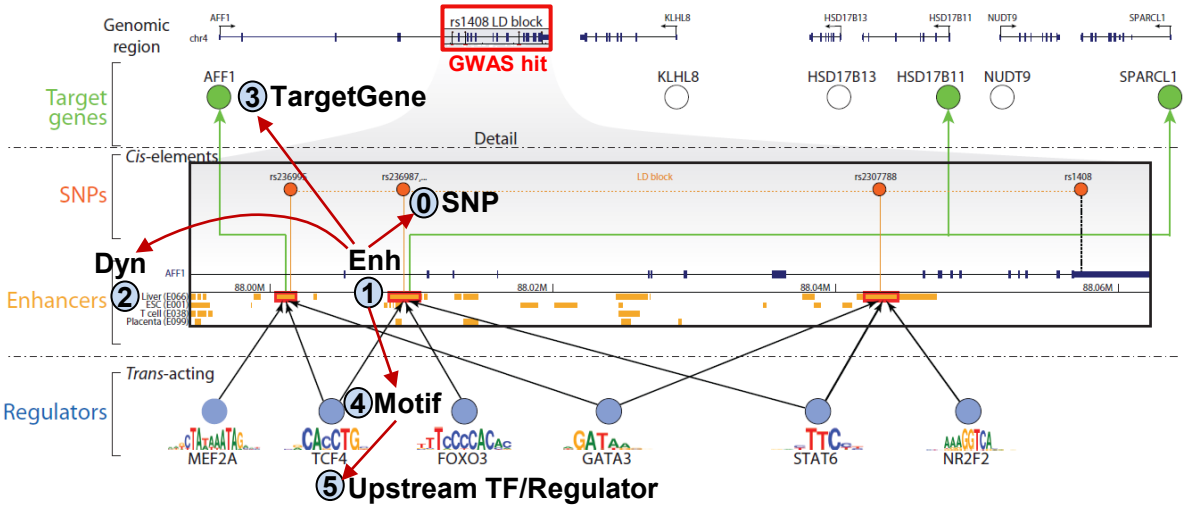
Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

**MLCB - Machine Learning in Computational Biology - Fall 2024 - Profs. Manolis Kellis + Eric Alm - 6.8700/6.8701/20.s900/20.s948/HST.507**

| Homeworks | Project / Mentoring (Fri) | Wk | Date | Lec | Topic |
|---|---|---|---|---|---|
| | | | **Introduction: Machine Learning, Deep Learning, Generative AI, and the Unification of Biology** | | |
| HW0 out Thu 9/5 | | 1 | Thu-Sep-05 | L1 | Course Overview, Machine Learning, Deep Learning, Inference, Genome, Proteins, Chemistry, Imaging |
| | | | **Module 1: Genomics, Epigenomics, Single-Cell, Networks, Circuitry** | | |
| HW0 due Wed 9/11 | | 2 | Tue-Sep-10 | L2 | Expression Analysis, Clustering/Classification, Gaussian Mixture Models, K-means, Bayesian Inf, Gen-vs-DiscrML |
| HW1 out Thu 9/12 | 0=Self Introductions | 2 | Thu-Sep-12 | L3 | Single-cell genomics, sc-mutli-omics, non-linear embeddings, spatial transcriptomics, next-gen technologies |
| | | 3 | Tue-Sep-17 | L4 | Sequential Data, Alignment, DynProg, Hidden Markov Models, Parsing, Posterior Decoding, HMM architectures |
| | | 3 | Thu-Sep-19 | L5 | Epigenomics: Signal Modeling, Peak calling, Chromatin states, 3D structure, Hi-C, Genome Topology |
| | | 4 | Tue-Sep-24 | L6 | Regulatory Genomics: Motifs, Information, ChIP, Gibbs Sampling, EM, CNNs for Genome Parsing |
| HW1 due Mon 9/30 | 1=Select previous paper(s) | 4 | Thu-Sep-26 | L7 | Regulatory Networks: Graphs, Linear Algebra, PCA, SVD, Dimentionality Reduction, TF-enhancer-gene circuitry |
| | | | **Module 2: Protein Structure, Protein Language Models, Geometric Deep Learning** | | |
| HW2 out Thu 10/3 | | 5 | Tue-Oct-01 | L8 | Intro to structural biology |
| | | 5 | Thu-Oct-03 | L9 | Protein structure and folding: Diffusion models, Cryo-EM, Protein design |
| | | 6 | Tue-Oct-08 | L10 | Intro to transformers and Large Language Models LLMs |
| | 2=Proposal+Feasibility | 6 | Thu-Oct-10 | L11 | Protein Language Models PLMs and Transfer Learning |
| | | 7 | Tue-Oct-15 | - | -- No Class -- Student holiday |
| HW2 due Mon 10/21 | | 7 | Thu-Oct-17 | L12 | DNA language models: Chromatin Structure |
| | | | **Module 3: Chemistry, Therapeutics, Graph Neural Networks** | | |
| HW3 out Thu 10/24 | | 8 | Tue-Oct-22 | L13 | Overview of drug development |
| | 3=OffHrs\|Update\|Feedback | 8 | Thu-Oct-24 | L14 | Intro to small molecules |
| | | 9 | Tue-Oct-29 | L15 | Representation of small molecules: Graphs, GNNs, Transformers, RDKit |
| | | 9 | Thu-Oct-31 | L16 | Docking: Small molecule - proteins docking |
| | | 10 | Tue-Nov-05 | L17 | Disease Association Mapping, genetics, GWAS, linkage analysis, disease circuitry, variant-to-function |
| HW3 due Tue 11/12 | 4=OffHrs\|Update\|Feedback | 10 | Thu-Nov-07 | L18 | Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS, multi-modal QTLs |
| | | | **Module 4: Electronic Health Records, Imaging, Evolution, Metabolism** | | |
| No HW | | 11 | Tue-Nov-12 | L19 | Electronic Health Records, AllOfUs, UKBioBank, Medical Genomics, Pop-Scale Cohorts, Multi-Ancestry [not quiz'd] |
| | | 11 | Thu-Nov-14 | L20 | **-- In-class Quiz** |
| | | 12 | Tue-Nov-19 | L21 | Imaging methods for biological applications |
| | 5=Midcourse report | 12 | Thu-Nov-21 | L22 | Comparative genomics, Conservation, Evolutionary signatures, PhyloCSF, RNA structure, Motif BLS2conf |
| | | 13 | Tue-Nov-26 | L23 | Evolution, Phylogenetics, Phylogenomics, Duplication, RNA world, RNA folding, lincRNAs, RNA modifications, m6A |
| | | 13 | Thu-Nov-28 | - | -- No Class -- Thanksgiving Holiday |
| | | 14 | Tue-Dec-03 | L24 | Modeling metabolism: Flux balance analysis |
| | 6=WriteUp, Slides Due | 14 | Thu-Dec-05 | L25 | Measuring metabolism: Metabolomics and Deep Learning |
| | | | **Final Projects** | | |
| | 7=In Class Presentations | 15 | Tue-Dec-10 | L26 | **Project Presentations (6-8 mins/team). Report due Fri@11.59p, Slides due Mon@11.59p, Present Live Tue** |

MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution




# Lecture 7: Regulatory circuitry

## Epigenome Dynamics: Joint Chromatin State Learning

## Enhancer-gene linking: Correlation, Hi-C, eQTLs

## TF motif discovery: Enrichment, EM, Gibbs Sampling

## Deep learning convolution CNNs for motif discovery

## Global motif discovery: Comparative Genomics

## Motif Instance Identification: Branch Length Score

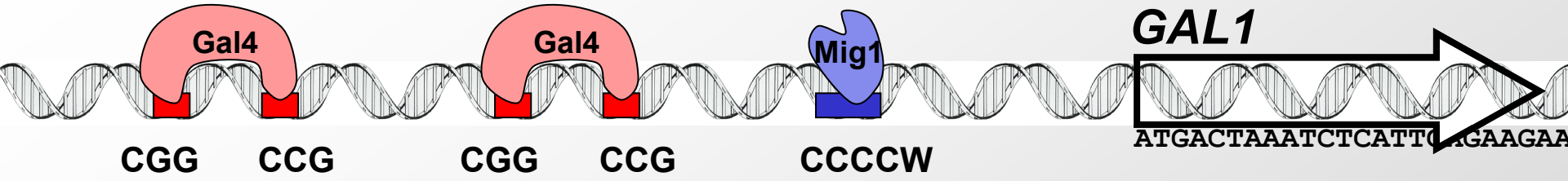
## Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

# Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint (M,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

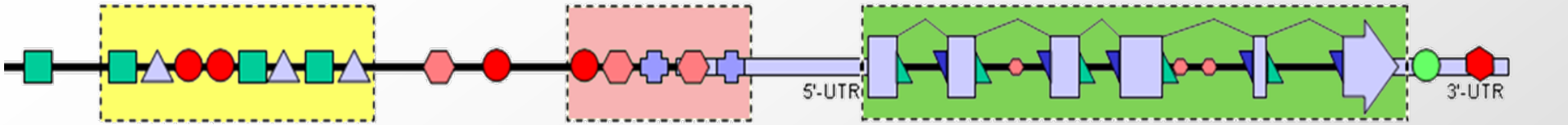
# Regulatory motif discovery




- Regulatory motifs
  - Genes are turned on / off in response to changing environments
  - No direct addressing:  subroutines (genes) contain sequence tags (motifs)
  - Specialized proteins (transcription factors) recognize these tags

- What makes motif discovery hard?
  - Motifs are short (6-8 bp), sometimes degenerate
  - Can contain any set of nucleotides (no ATG or other rules)
  - Act at variable distances upstream (or downstream) of target gene

# The regulatory code: All about regulatory motifs




**Enhancer regions**
**Where in the body?**

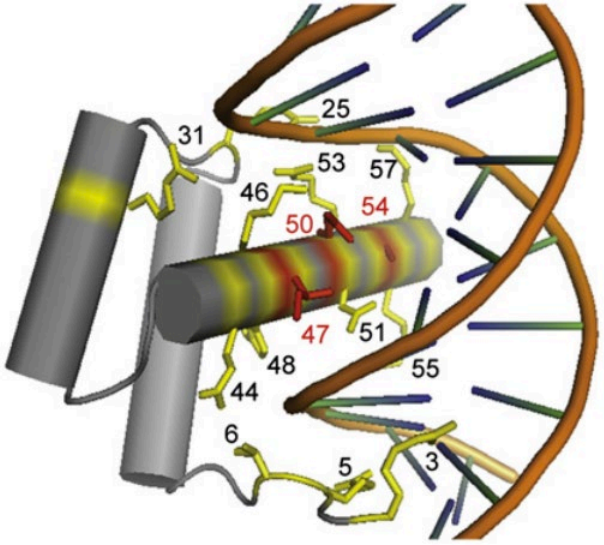
**Promoter motifs**
**When in time?**

**Splicing signals**
**Which variants?**

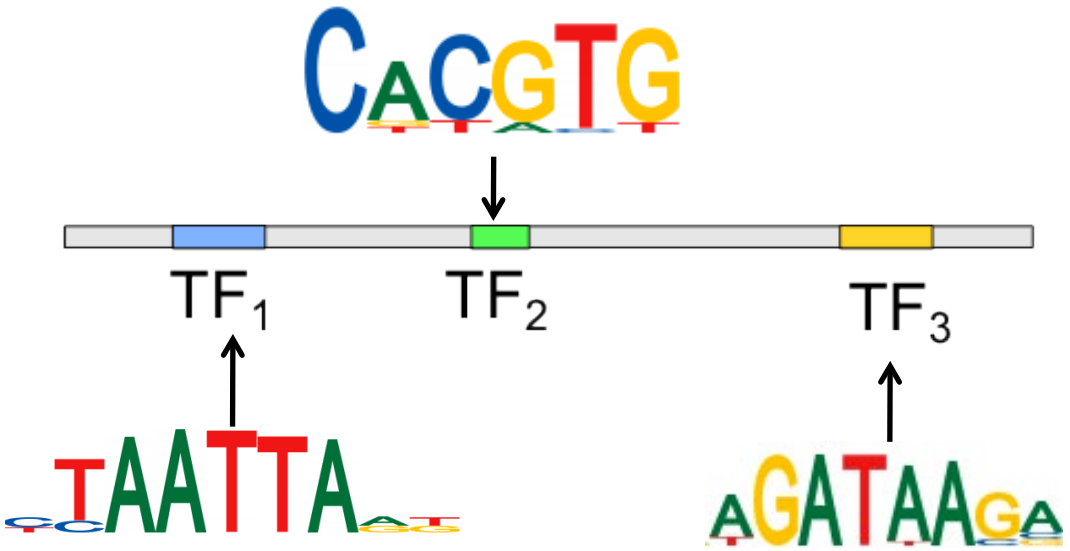
**Motifs at RNA level**
**Which subsets?**

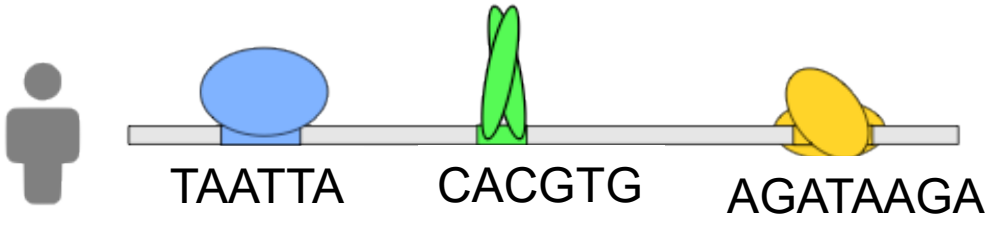
- The parts list: ~20-30k genes
  - Protein-coding genes, RNA genes (tRNA, microRNA, snRNA)
- The circuitry: constructs controlling gene usage
  - Enhancers, promoters, splicing, post-transcriptional motifs
- The regulatory code, complications:
  - Combinatorial coding of 'unique tags'
    - Data-centric encoding of addresses
  - Overlaid with 'memory' marks
    - Large-scale on/off states
  - Modulation of the large-scale coding
    - Post-transcriptional and post-translational information
- Today: discovering motifs in co-regulated promoters and *de novo* motif discovery & target identification

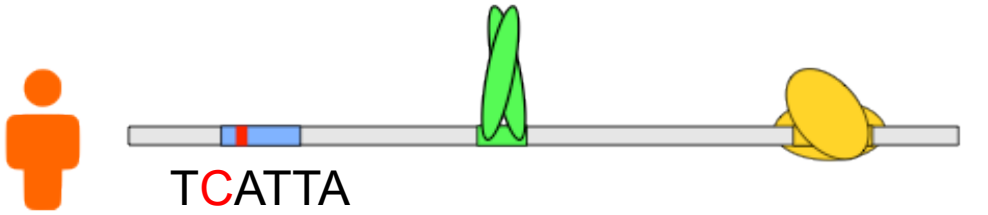
# TFs use DNA-binding domains to recognize specific DNA sequences in the genome



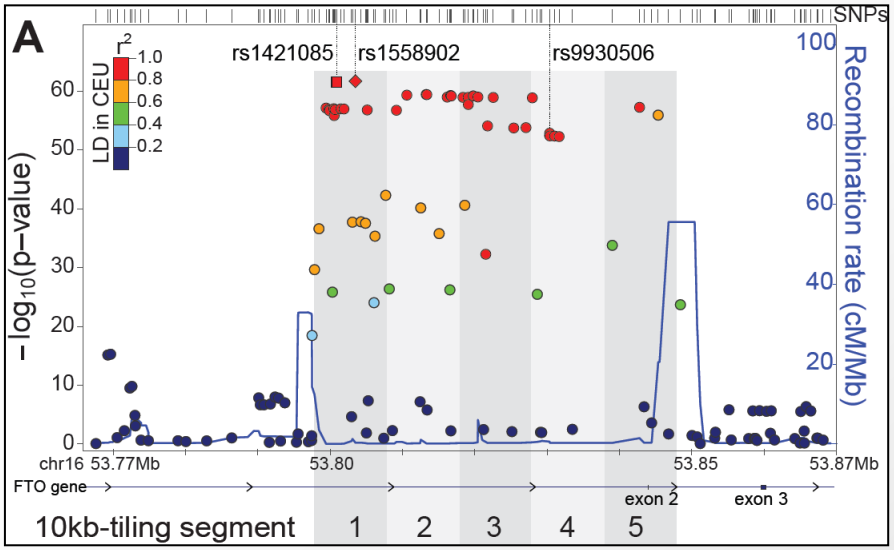

DNA-binding domain of
*Engrailed*



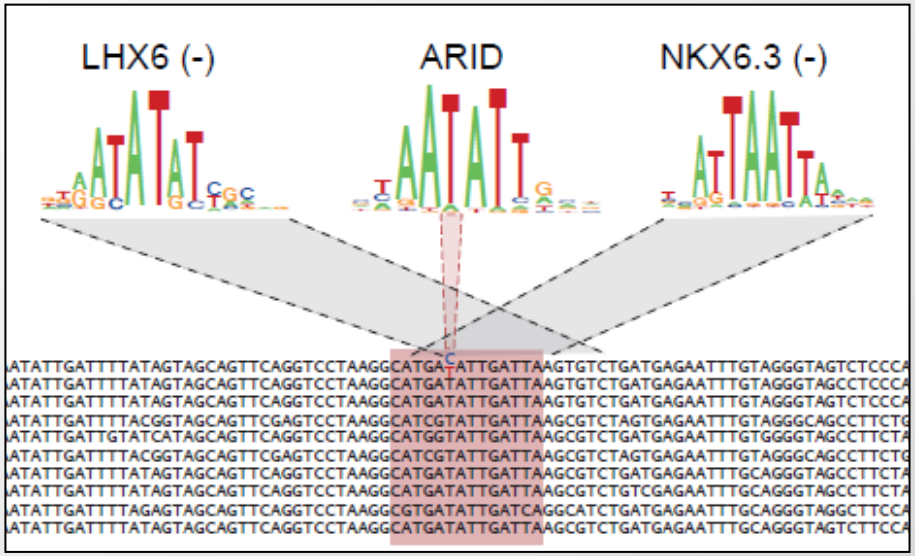



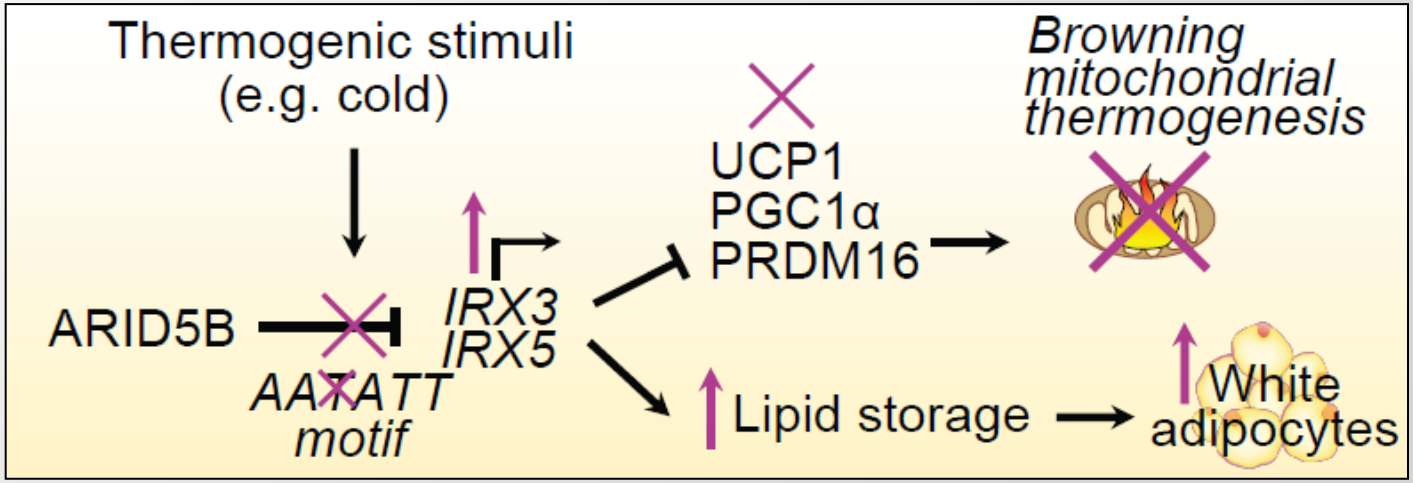
# Disrupted motif at the heart of FTO obesity locus




***Strongest association with obesity***



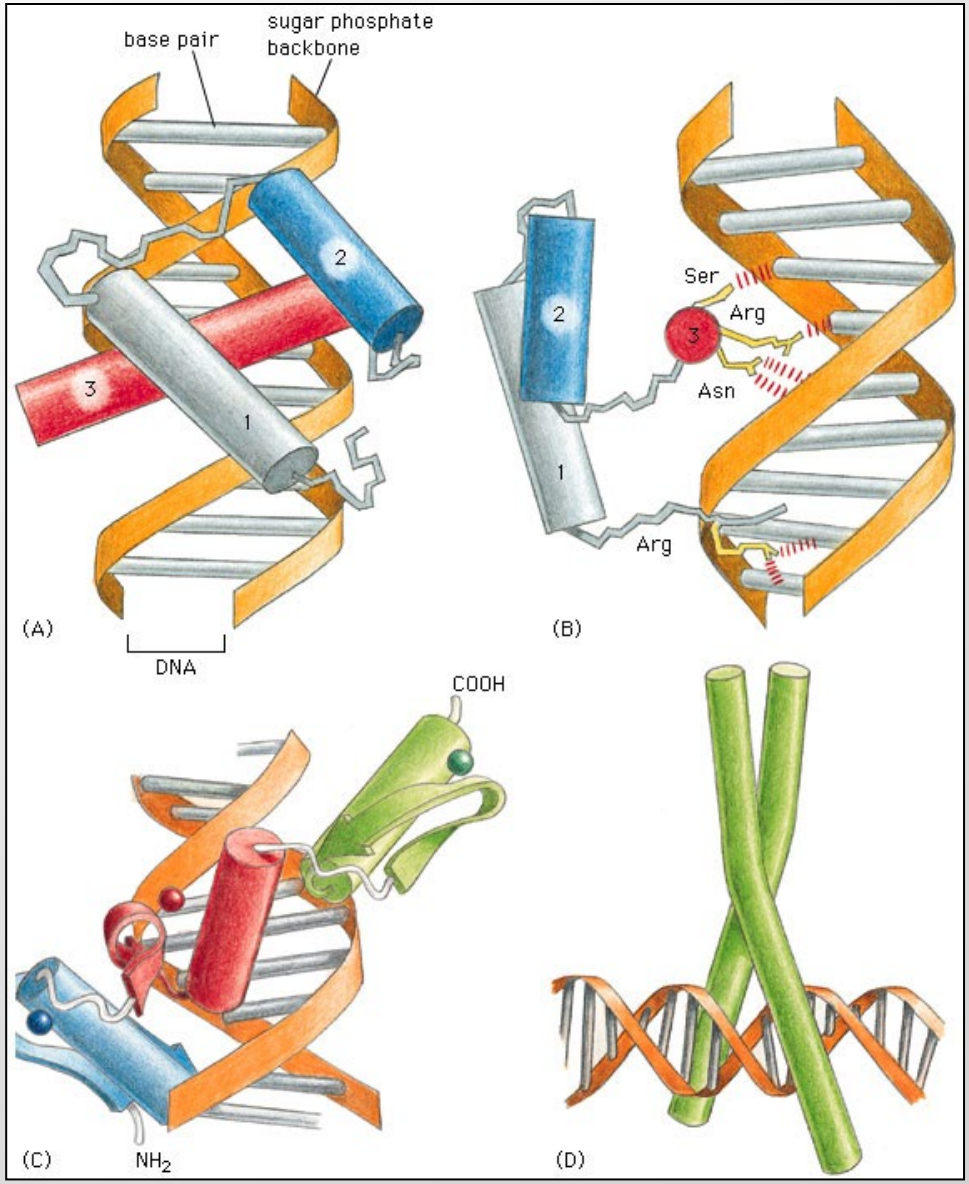

***C-to-T disruption of AT-rich regulatory motif***




***Restoring motif restores thermogenesis***

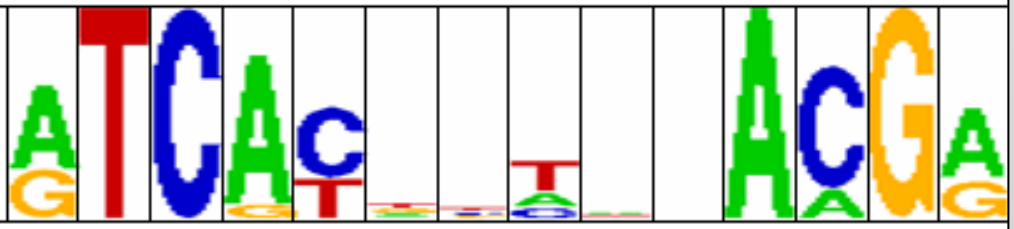
# Regulator structure ⇔ recognized motifs

- Proteins ‘feel’ DNA
  - Read chemical properties of bases
  - Do NOT open DNA (no base complementarity)
- 3D Topology dictates specificity
  - Fully constrained positions: ➔ every atom matters
  - “Ambiguous / degenerate” positions ➔ loosely contacted
- Other types of recognition
  - MicroRNAs: complementarity
  - Nucleosomes: GC content
  - RNAs: structure/seqn combination



# Motifs summarize TF sequence specificity

| Target genes bound by ABF1 regulator | | Coordinates | | Genome sequence at bound site |
|---|---|---|---|---|
| ACS1 | acetyl CoA synthetase | -491 | -479 | \|ATCATTCTGGACG\| |
| ACS1 | acetyl CoA synthetase | -433 | -421 | \|ATCATCTCGGACG\| |
| ACS1 | acetyl CoA synthetase | -311 | -299 | \|ATCATTTGCCACG\| |
| CHA1 | catabolic L-serine dehydratase | -280 | -254 | A\|ATCACCGCGAACG\|GA |
| ENO2 | Enolase | -470 | -461 | ggcgttat\|GTCACTAACGACG\|tgcacca |
| HMR | silencer | -256 | -283 | ATCAATAC\|ATCATAAAATACG\|AACGATC |
| LPD1 | lipoamide dehydrogenase | -288 | -300 | gat\|ATCAAAATTAACG\|tag |
| LPD1 | lipoamide dehydrogenase | -301 | -313 | gat\|ATCACCGTTGACG\|tca |
| PGK | phosphoglycerate kinase | -523 | -496 | CAAACAA\|ATCACGAGCGACG\|GTAATTTC |
| RPC160 | RNA pol III/C 160 kDa subunit | -385 | -349 | \|ATCACTATATACG\|TGAA |
| RPC40 | RNA pol III/C 40 kDa subunit | -137 | -116 | \|GTCACTATAAACG\| |
| rpL2 | ribosomal protein L2 | -185 | -167 | TAAT\|aTCAcgtcACACG\|AC |
| SPR3 | CDC3/10/11/12 family homolog | -315 | -303 | \|ATCACTAAATACG\| |
| YPT1 | TUB2 | -193 | -172 | CCTAG\|GTCACTGTACACG\|TATA |

| Position | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Position Weight Matrix (PWM) | A | **56** | 4 | 4 | **81** | 4 | 23 | 15 | 27 | 31 | 31 | **89** | 23 | 4 | **58** |
| | G | **32** | 4 | 4 | 12 | 4 | 31 | 23 | 4 | 19 | 23 | 4 | 4 | **89** | **35** |
| | C | 4 | 4 | **89** | 4 | **58** | 12 | 23 | 19 | 19 | 23 | 4 | **69** | 4 | 4 |
| | T | 4 | **89** | 4 | 4 | **35** | 35 | 39 | 50 | 31 | 23 | 4 | 4 | 4 | 4 |
| Motif Logo | |  | | | | | | | | | | | | | |
| Consensus | | **R** | **T** | **C** | **A** | **Y** | N | N | H | N | N | **A** | **C** | **G** | **R** |

- Summarize information
- Integrate many positions
- Measure of information
- Distinguish motif vs. motif instance
- Assumptions:
  - Independence
  - Fixed spacing

# Motifs are not limited to DNA sequences

- Splicing Signals at the RNA level
  - Splice junctions
  - Exonic Splicing Enhancers (ESE)
  - Exonic Splicing Surpressors (ESS)
- Domains and epitopes at the Protein level
  - Glycosylation sites
  - Kinase targets
  - Targetting signals
  - MHC binding specificities
- Recurring patterns at the physiological level
  - Expression patterns during the cell cycle
  - Heart beat patterns predicting cardiac arrest
    - Final project in previous year, now used in Boston hospitals!
  - Any probabilistic recurring pattern

# Approaches to regulatory motif discovery

**Region-based motif discovery**

- Expectation Maximization (e.g. MEME)
  - Iteratively refine positions / motif profile
- Gibbs Sampling (e.g. AlignACE)
  - Iteratively sample positions / motif profile
- Enumeration with wildcards (e.g. Weeder)
  - Allows global enrichment/background score
- Peak-height correlation (e.g. MatrixREDUCE)
  - Alternative to cutoff-based approach

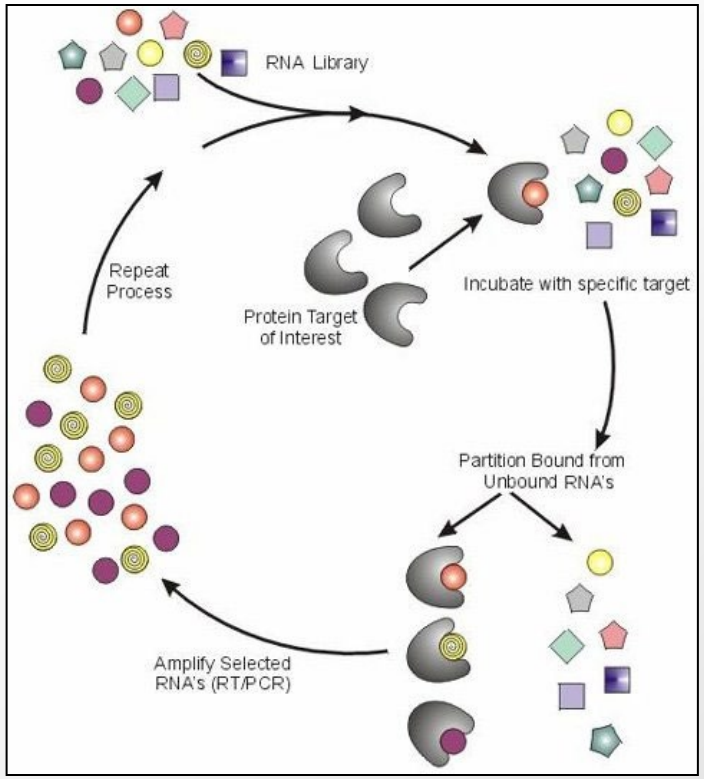
**Genome-wide**

- Conservation-based discovery (e.g. MCS)
  - Genome-wide score, up-/down-stream bias

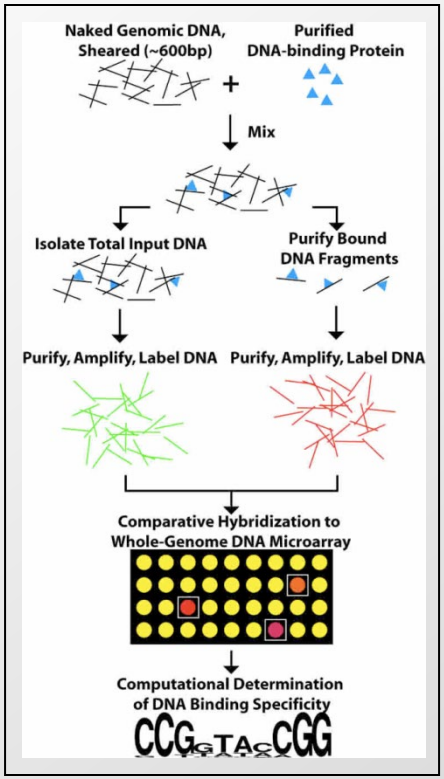
***In vitro / trans***

- Protein Domains (e.g. PBMs, SELEX)
  - In vitro motif identification, seq-/array-based

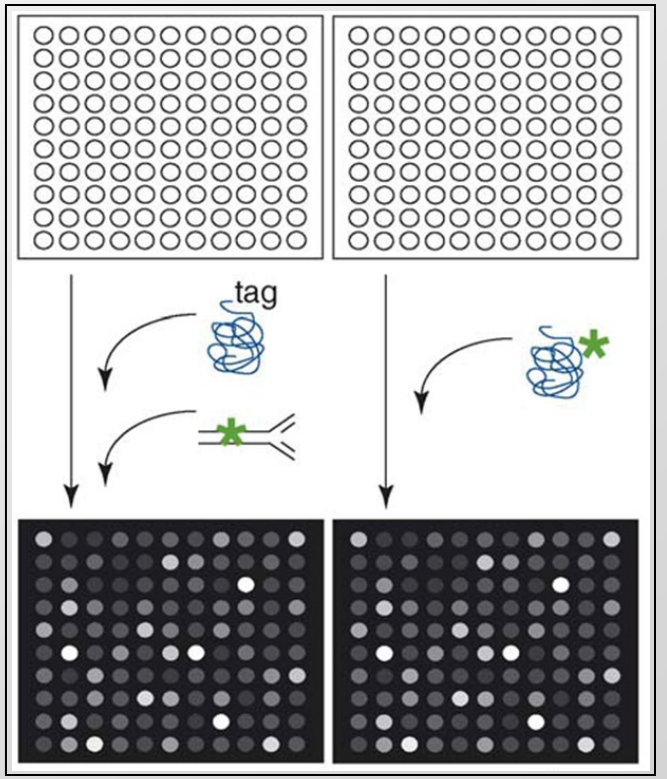
# Experimental factor-centric discovery of motifs




SELEX (Systematic Evolution of Ligands by Exponential Enrichment; Klug & Famulok, 1994).

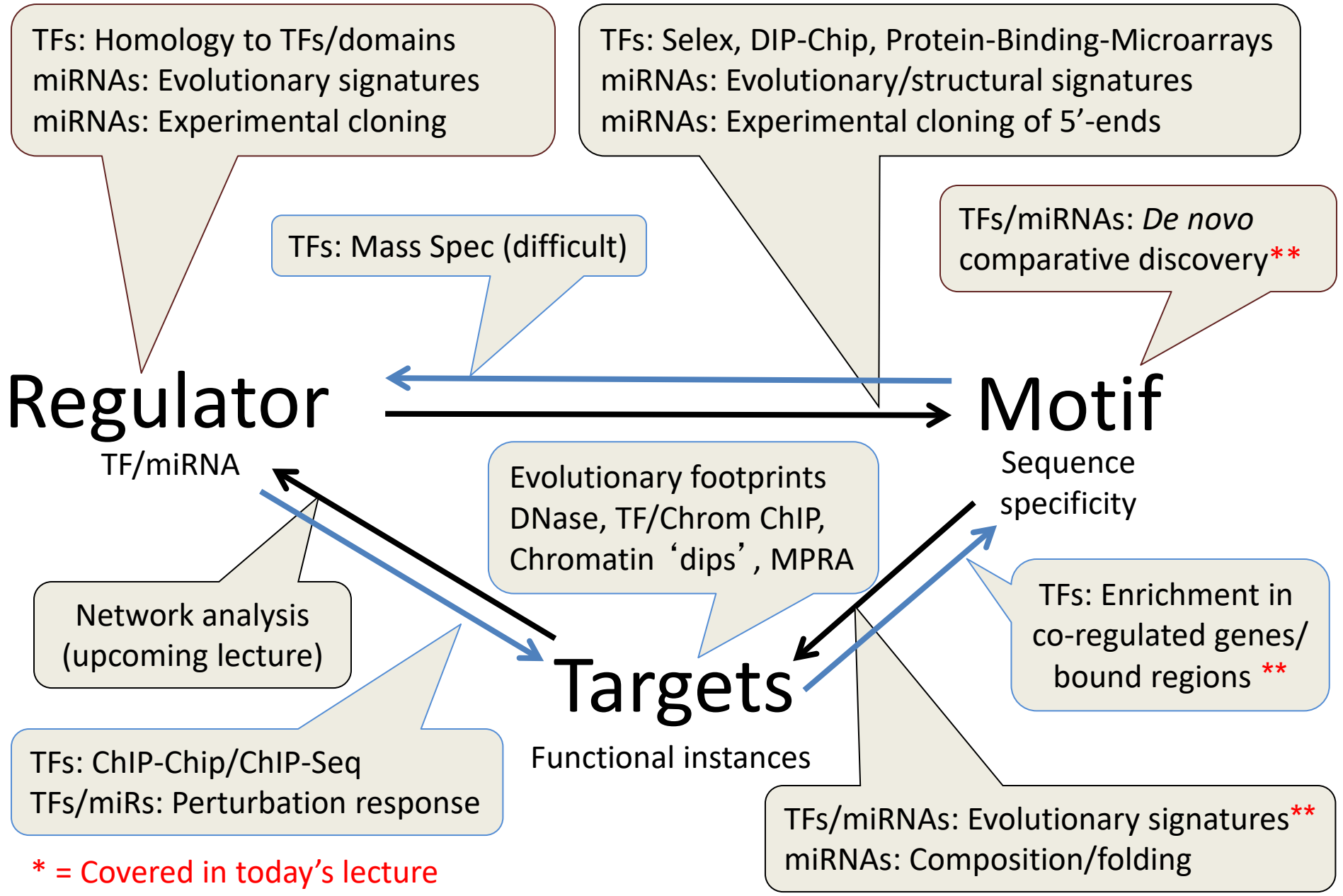



DIP-Chip (DNA-immunoprecipitation with microarray detection; Liu et al., 2005)




PBMs (Protein binding microarrays; Mukherjee, 2004)
Double stranded DNA arrays

# Challenges in regulatory genomics

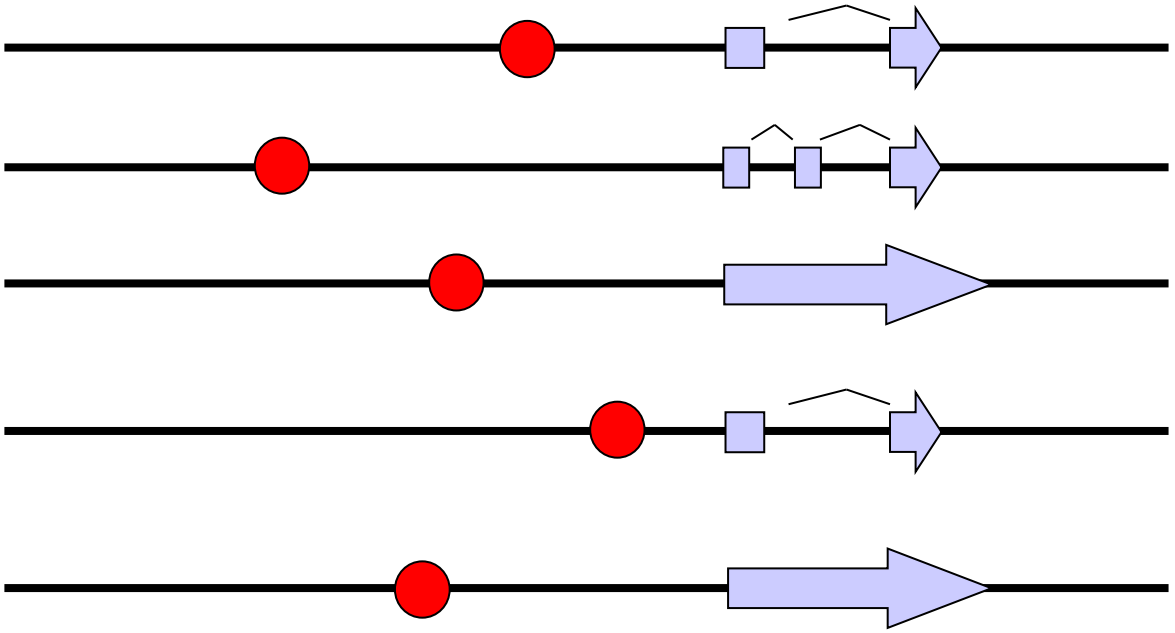


# Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint (M,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

# Enrichment-based discovery methods

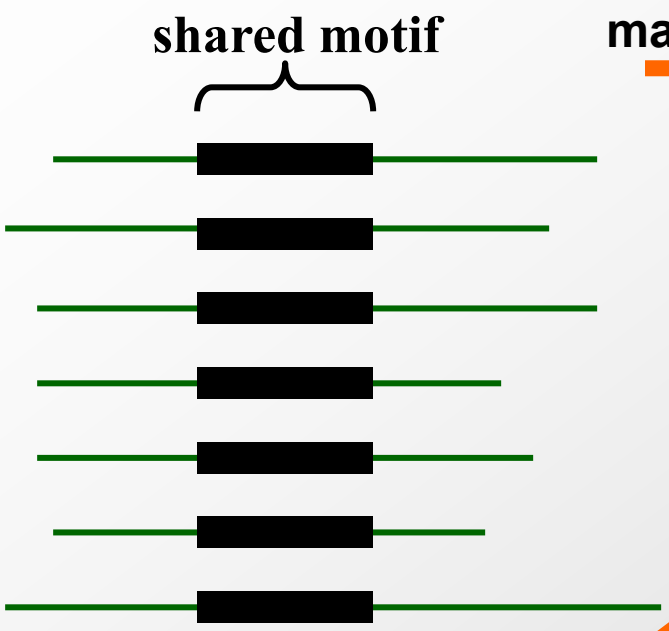
**Given a set of co-regulated/functionally related genes, find common motifs in their promoter regions**



- **Align the promoters to each other using local alignment**
- **Use expert knowledge for what motifs should look like**
- **Find ‘median’ string by enumeration (motif/sample driven)**
- **Start with conserved blocks in the upstream regions**

# Starting positions ⇔ Motif matrix

- given aligned sequences ➔ easy to compute profile matrix




maximization

sequence positions

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| A | 0.1 | 0.3 | 0.1 | 0.2 | 0.2 | 0.4 | 0.3 | 0.1 |
| C | 0.5 | 0.2 | 0.1 | 0.1 | 0.6 | 0.1 | 0.2 | 0.7 |
| G | 0.2 | 0.2 | 0.6 | 0.5 | 0.1 | 0.2 | 0.2 | 0.1 |
| T | 0.2 | 0.3 | 0.2 | 0.2 | 0.1 | 0.3 | 0.3 | 0.1 |

expectation

given profile matrix

- easy to find starting position probabilities

**Key idea: Iterative procedure for estimating both, given uncertainty (learning problem with hidden variables: the starting positions)**

# Three options for assigning points, and their parallels across K-means, HMMs, Motifs

| Update rule | Update assignments (E step) ➔ Estimate hidden labels | Algorithm implementing E step in each of the three settings | | | Update model parameters (M step) ➔ max likelihood |
|---|---|---|---|---|---|
| | | **Expression clustering** | **HMM learning** | **Motif discovery** | |
| The hidden label is: | | Cluster labels | State path π | Motif positions | |
| Pick a best | Assign each point to best label | **K-means:** Assign each point to nearest cluster | **Viterbi training:** label sequence with best path | **Greedy:** Find best motif match in each sequence | Average of those points assigned to label |
| Average all | Assign each point to all labels, probabilistically | **Fuzzy K-means:** Assign to all clusters, weighted by proximity | **Baum-Welch training:** label sequence w all paths (posterior decoding) | **MEME:** Use all positions as a motif occurrence weighed by motif match score | Average of all points, weighted by membership |
| Sample one | Pick one label at random, based on their relative probability | **N/A:** Assign to a random cluster, sample by proximity | **N/A:** Sample a single label for each position, according to posterior prob. | **Gibbs sampling:** Use one position for the motif, by sampling from the match scores | Average of those points assigned to label(a sample) |

# Basic Iterative Approach

Given: length parameter *W,* training set of sequences

set initial values for **motif**

do

➔ re-estimate ***starting-positions*** from ***motif***

➔ re-estimate ***motif*** from ***starting-positions***

until convergence (change < ε)

return: ***motif, starting-positions***

# Representing Motif M(k,c) and Background B(c)

- Assume motif has fixed width, *W*
- Motif represented by matrix of probabilities: ***M(k,c)***
  the probability of character ***c*** in column ***k***

$$M = \begin{array}{c|ccc} & 1 & 2 & 3 \\ A & 0.1 & 0.5 & 0.2 \\ C & 0.4 & 0.2 & 0.1 \\ G & 0.3 & 0.1 & 0.6 \\ T & 0.2 & 0.2 & 0.1 \end{array}$$

**(~CAG)**

- **Background represented by B(c), frequency of each base**

$$B = \begin{array}{c|c} A & 0.26 \\ C & 0.24 \\ G & 0.23 \\ T & 0.27 \end{array}$$

**(near uniform)**

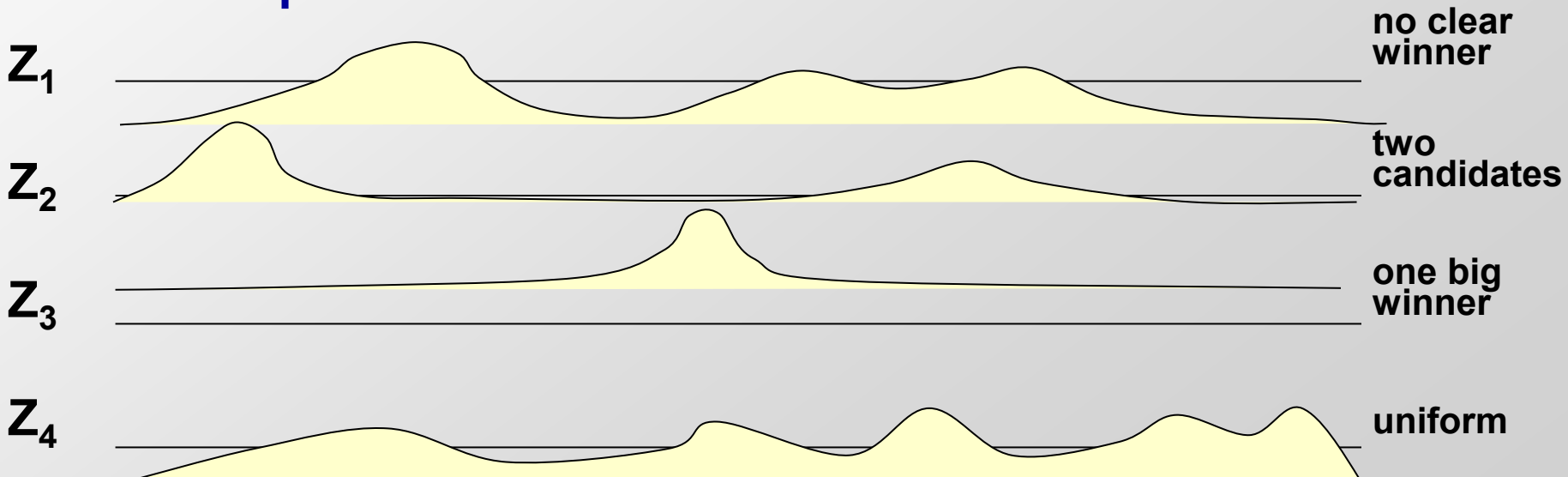
**(see also: di-nucleotide etc)**

# Representing the starting position probabilities ($Z_{ij}$)

- the element $Z_{ij}$ of the matrix $Z$ represents the probability that the motif starts in position *j* in sequence *i*

$$Z = \begin{array}{c} \\ \texttt{seq1} \\ \texttt{seq2} \\ \texttt{seq3} \\ \texttt{seq4} \end{array} \begin{array}{cccc} \texttt{1} & \texttt{2} & \texttt{3} & \texttt{4} \\ \texttt{0.1} & \texttt{0.1} & \texttt{0.2} & \texttt{0.6} \\ \texttt{0.4} & \texttt{0.2} & \texttt{0.1} & \texttt{0.3} \\ \texttt{0.3} & \texttt{0.1} & \texttt{0.5} & \texttt{0.1} \\ \texttt{0.1} & \texttt{0.5} & \texttt{0.1} & \texttt{0.3} \end{array}$$

**Some examples:**



# Starting positions ($Z_{ij}$) ⇔ Motif matrix M(k,c)


M-step

E-step

| | k=1 | k=2 | k=3 | k=4 | k=5 | k=6 | k=7 | k=8 |
|---|---|---|---|---|---|---|---|---|
| c=A | 0.1 | 0.3 | 0.1 | 0.2 | 0.2 | 0.4 | 0.3 | 0.1 |
| c=C | 0.5 | 0.2 | 0.1 | 0.1 | 0.6 | 0.1 | 0.2 | 0.7 |
| c=G | 0.2 | 0.2 | 0.6 | 0.5 | 0.1 | 0.2 | 0.2 | 0.1 |
| c=T | 0.2 | 0.3 | 0.2 | 0.2 | 0.1 | 0.3 | 0.3 | 0.1 |

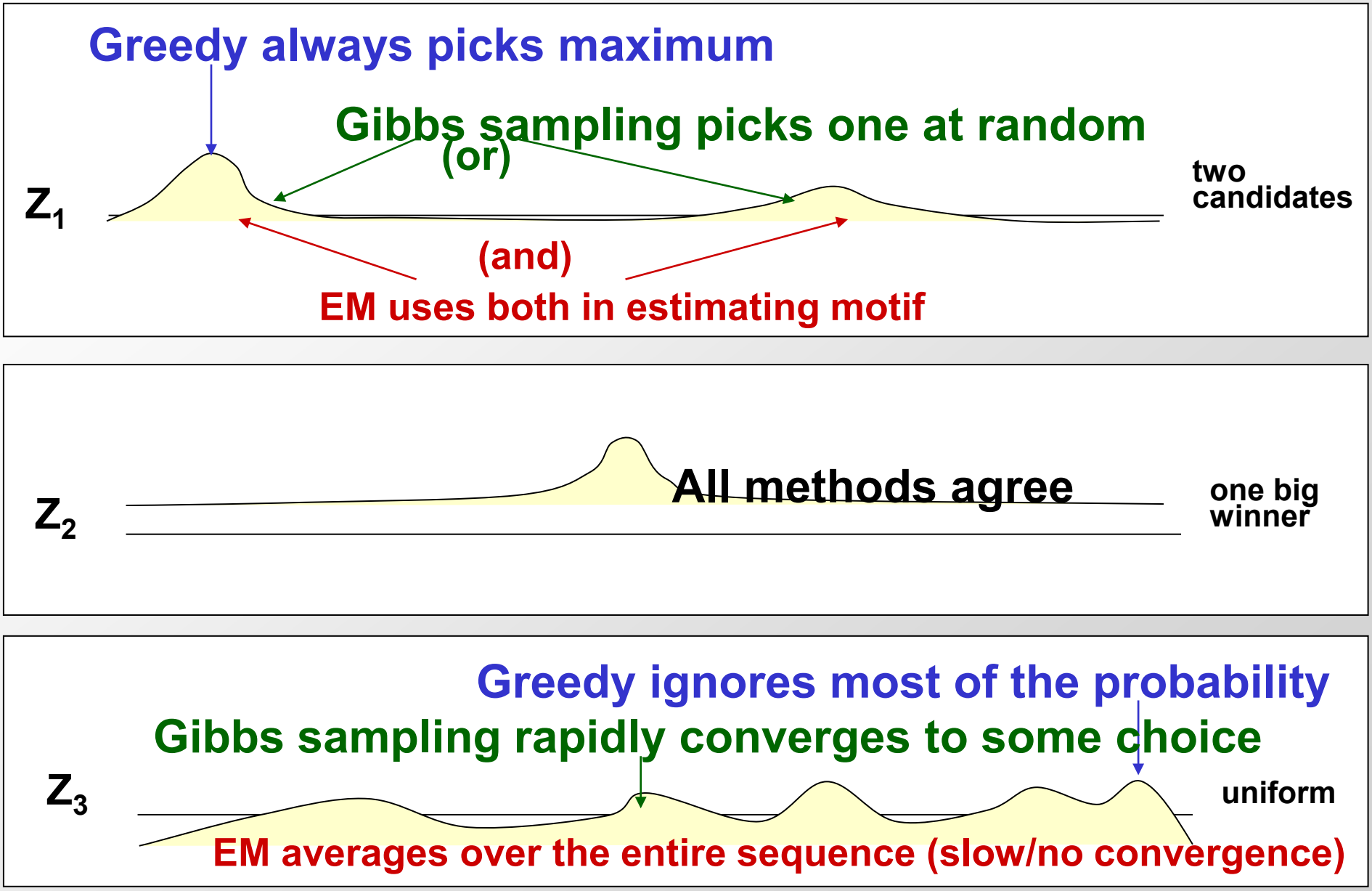
**Starting positions: $Z_{ij}$** **Motif: M(k,c)**

- $Z_{ij}$: Probability that on sequence i, motif start at position j
- M(k,c): Probability that $k^{th}$ character of motif is letter c

- **Computing $Z_{ij}$ matrix from M(k,c) is straightforward**
  - **At each position, evaluate start probability by multiplying across the matrix**

- **Three variations for re-computing motif M(k,c) from $Z_{ij}$ matrix**
  - **Expectation maximization** ➔ **All starts weighted by $Z_{ij}$ prob distribution**
  - **Gibbs sampling** ➔ **Single start for each seq $X_i$ by sampling $Z_{ij}$**
  - **Greedy approach** ➔ **Best start for each seq $X_i$ by maximum $Z_{ij}$**

# Three examples for Greedy, Gibbs Sampling, EM



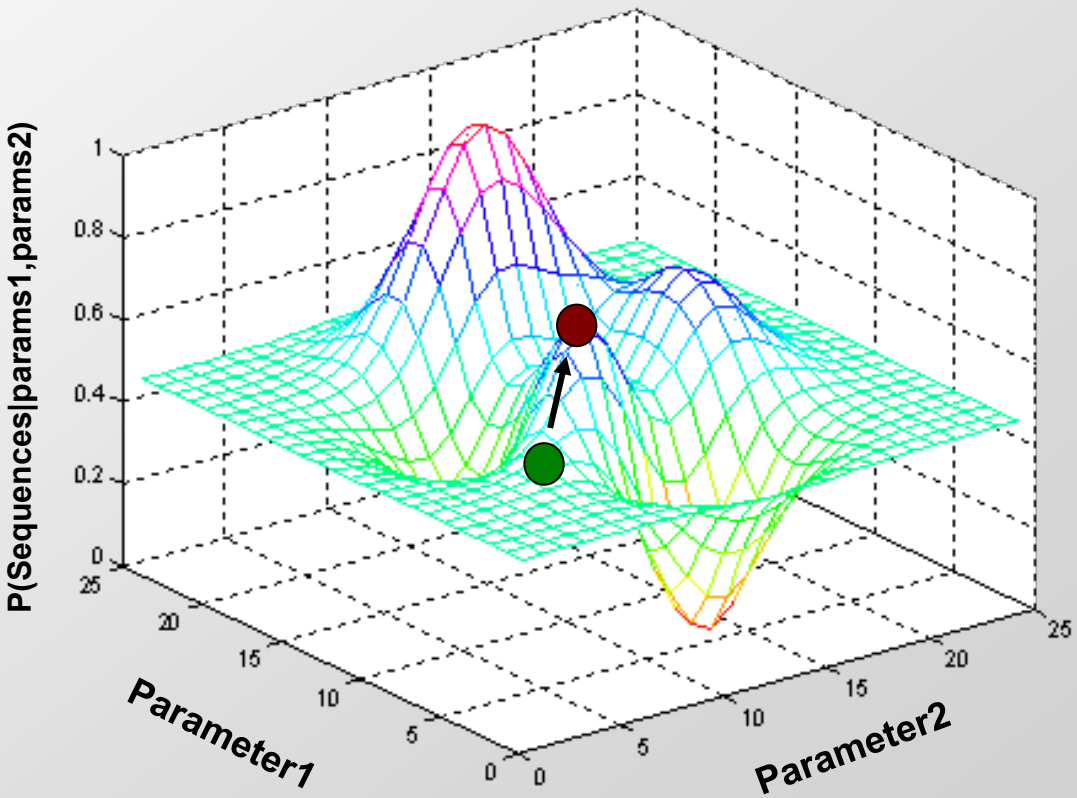
# P(Seq|Model) Landscape

**EM searches for parameters to increase P(seqs|parameters)**

Useful to think of P(seqs|parameters) as a function of parameters

EM starts at an initial set of parameters

And then "climbs uphill" until it reaches a local maximum




***Where EM starts can make a big difference***

# One solution: Search from Many Different Starts

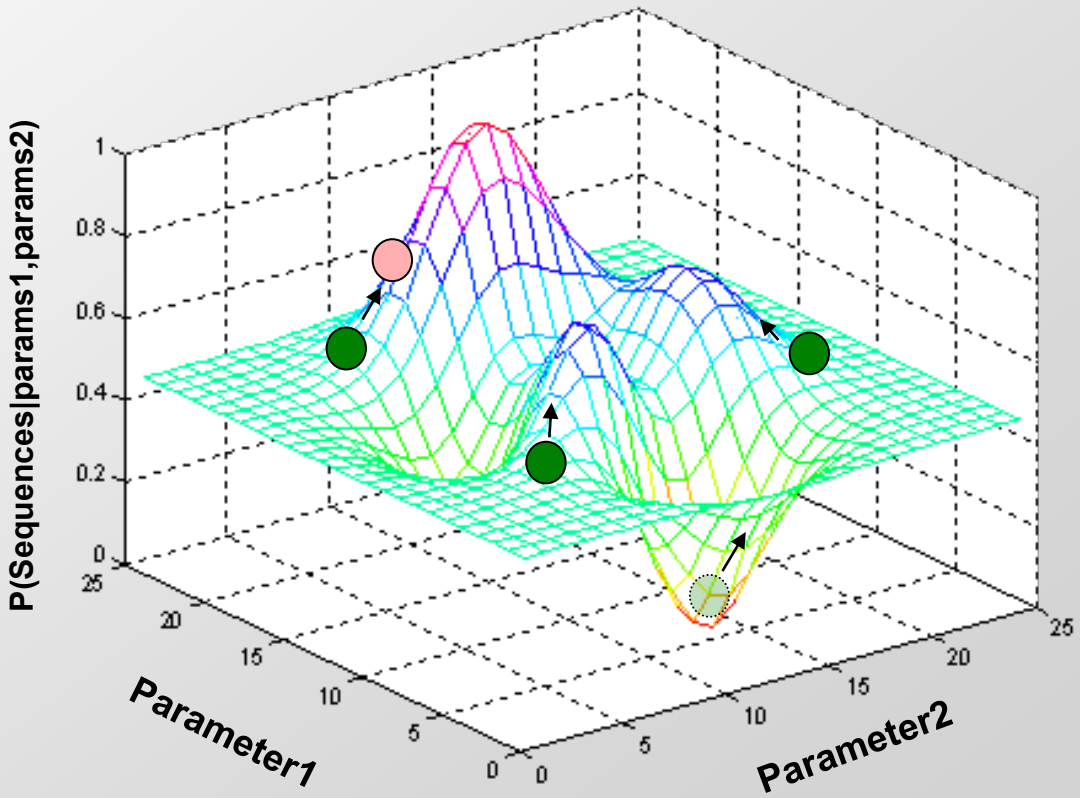
**To minimize the effects of local maxima, you should search multiple times from different starting points**

MEME uses this idea

Start at many points

Run for one iteration

Choose starting point that got the “highest” and continue



# Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint ($M$,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

# Three options for assigning points, and their parallels across K-means, HMMs, Motifs

| Update rule | Update assignments (E step) ➔ Estimate hidden labels | Algorithm implementing E step in each of the three settings | | | Update model parameters (M step) ➔ max likelihood |
|---|---|---|---|---|---|
| | | **Expression clustering** | **HMM learning** | **Motif discovery** | |
| The hidden label is: | | Cluster labels | State path π | Motif positions | |
| Pick a best | Assign each point to best label | **K-means:** Assign each point to nearest cluster | **Viterbi training:** label sequence with best path | **Greedy:** Find best motif match in each sequence | Average of those points assigned to label |
| Average all | Assign each point to all labels, probabilistically | **Fuzzy K-means:** Assign to all clusters, weighted by proximity | **Baum-Welch training:** label sequence w all paths (posterior decoding) | **MEME:** Use all positions as a motif occurrence weighed by motif match score | Average of all points, weighted by membership |
| Sample one | Pick one label at random, based on their relative probability | **N/A:** Assign to a random cluster, sample by proximity | **N/A:** Sample a single label for each position, according to posterior prob. | **Gibbs sampling:** Use one position for the motif, by sampling from the match scores | Average of those points assigned to label(a sample) |

# Gibbs Sampling

- A general procedure for sampling from the joint distribution of a set of random variables $\Pr(U_1 \ldots U_n)$ by iteratively sampling from for each $j$ $\Pr(U_j \mid U_1 \ldots U_{j-1}, U_{j+1} \ldots U_n)$
- Useful when it's hard to explicitly express means, stdevs, covariances across the multiple dimensions
- Useful for supervised, unsupervised, semi-supervised learning
  - Specify variables that are known, sample over all other variables
- Approximate:
  - Joint distribution: the samples drawn
  - Marginal distributions: examine samples for subset of variables
  - Expected value: average over samples
- Example of Markov-Chain Monte Carlo (MCMC)
  - The sample approximates an unknown distribution
  - Stationary distribution of sample (only start counting after burn-in)
  - Assume independence of samples (only consider every 100)
- Special case of Metropolis-Hastings
  - In its basic implementation of sampling step
  - But it's a more general sampling framework

# Gibbs Sampling for motif discovery

- First application to motif finding: Lawrence et al 1993
  - Can view as a stochastic analog of EM for motif discovery task
  - Less susceptible to local minima than EM
- EM maintains distribution $\boldsymbol{Z_i}$ over the starting points for each seq
- Gibbs sampling selects specific starting point $\mathbf{a_i}$ for each seq

  ➔ but keeps resampling these starting points

given: length parameter *W,* training set of sequences

choose random positions for *a*

do

pick a sequence $\boldsymbol{X_i}$

estimate ***p*** given current motif positions *a* (update step)

(using all sequences but $X_i$)

sample a new motif position $a_i$ for $X_i$ (sampling step)

until convergence

return: *p, a*

# Popular implementation: AlignACE, BioProspector

AlignACE: first statistical motif finder

BioProspector: improved version of AlignACE

Both use basic Gibbs Sampling algorithm:

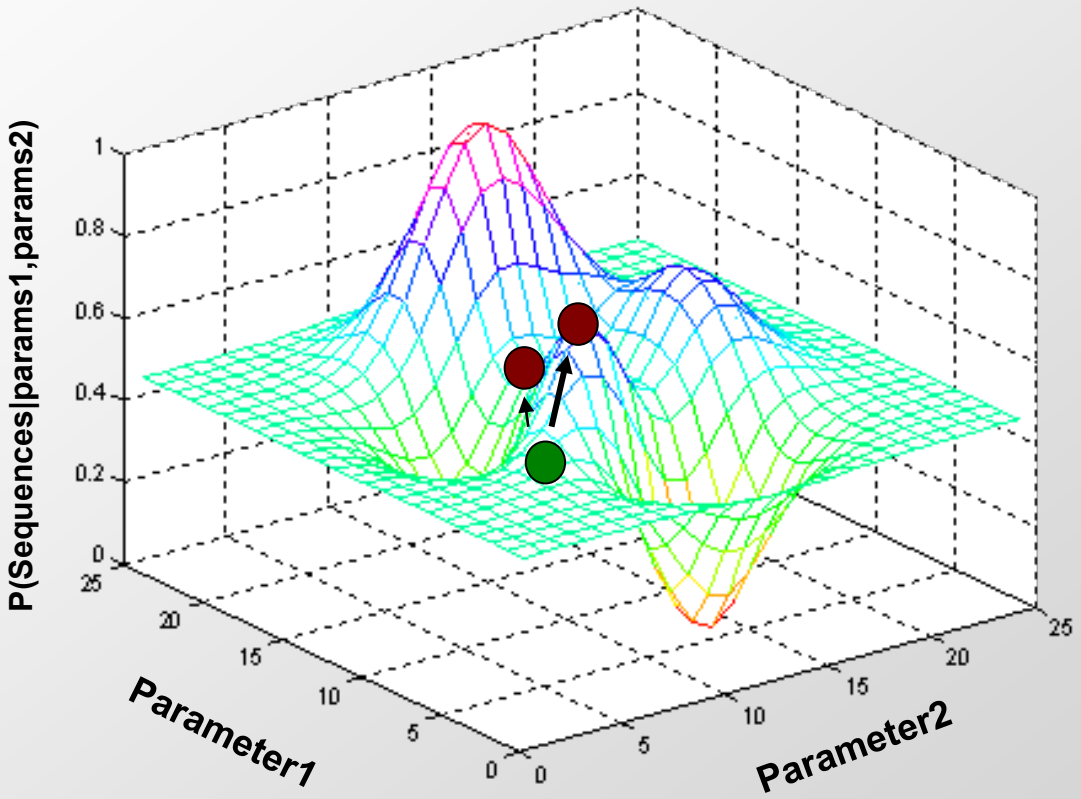
1. Initialization:
   a. Select random locations in sequences $X_1, \ldots, X_N$
   b. Compute an initial model M from these locations
2. Sampling Iterations:
   a. Remove one sequence $X_i$
   b. Recalculate model
   c. Pick a new location of motif in $X_i$ according to probability the location is a motif occurrence

In practice, run algorithm from multiple random initializations:

1. Initialize
2. Run until convergence
3. Repeat 1,2 several times, report common motifs

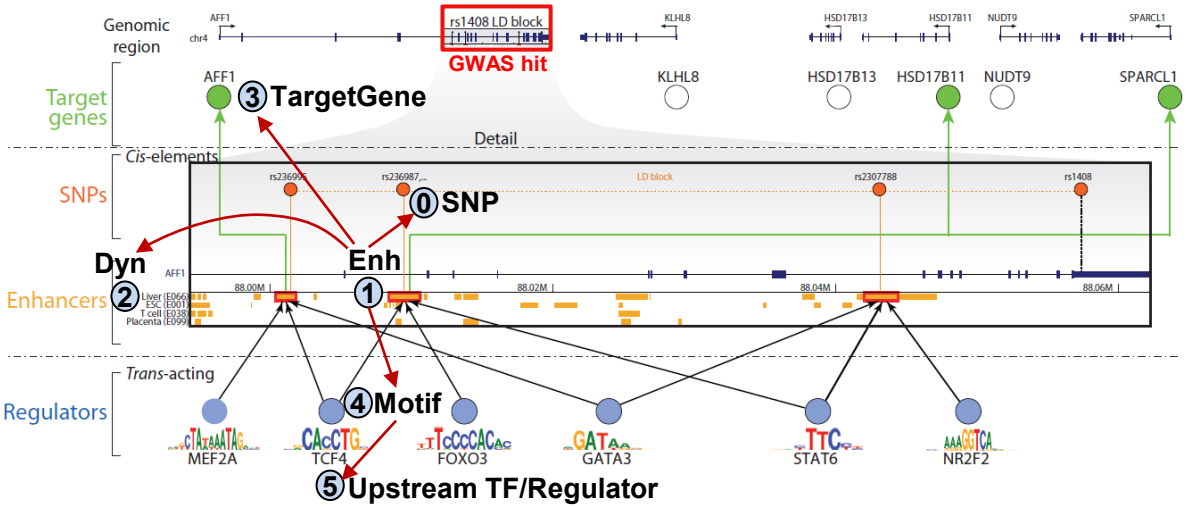
# Gibbs Sampling and Climbing

**Because gibbs sampling does always choose the best new location it can move to another place not directly uphill**




***In theory,*** **Gibbs Sampling less likely to get stuck a local maxima**

# MIT 18.700/18.701/6.047/6.878/HST.507
# Computational Biology: Genomes, Networks, Evolution

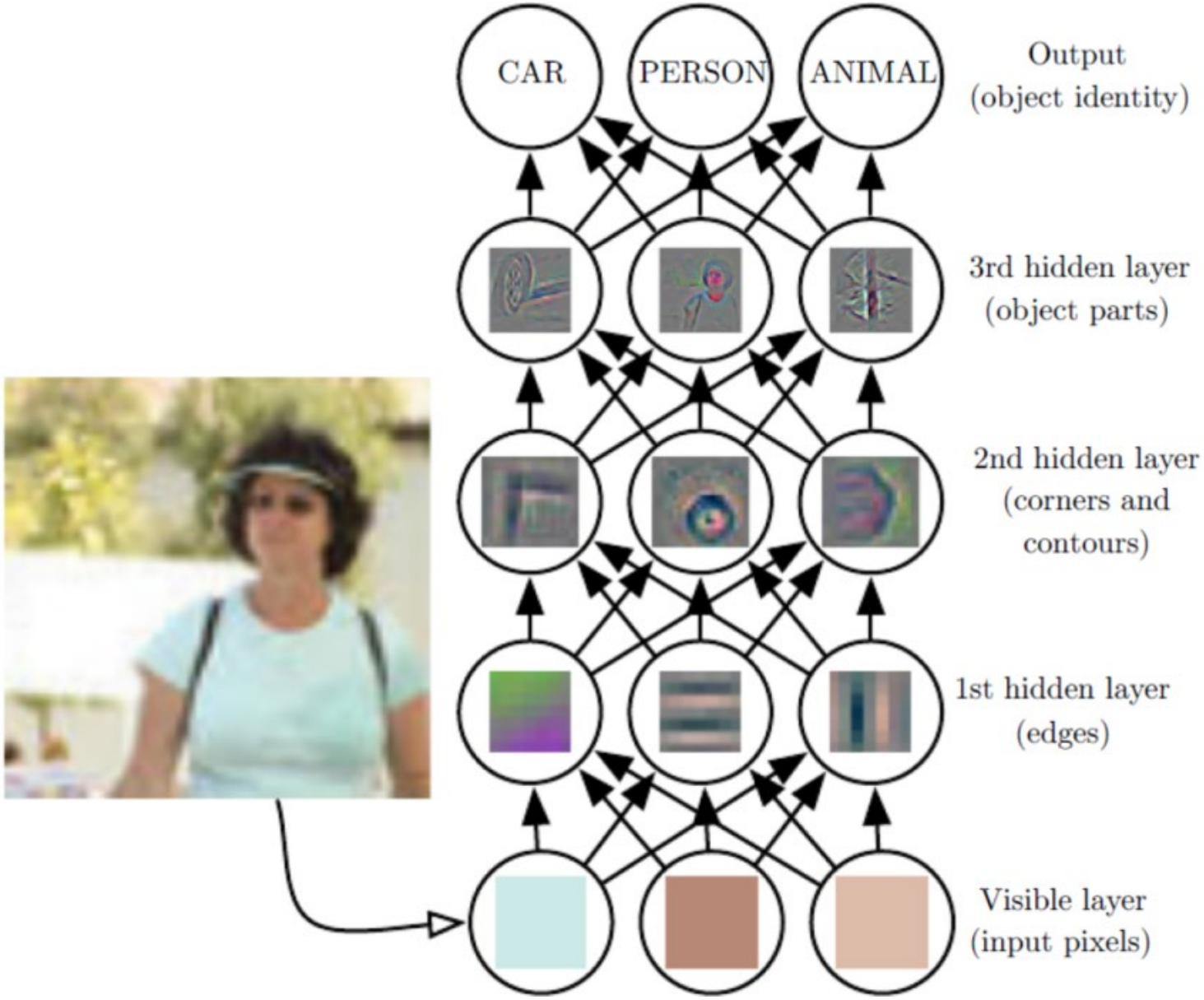



# Lecture 7: Regulatory circuitry

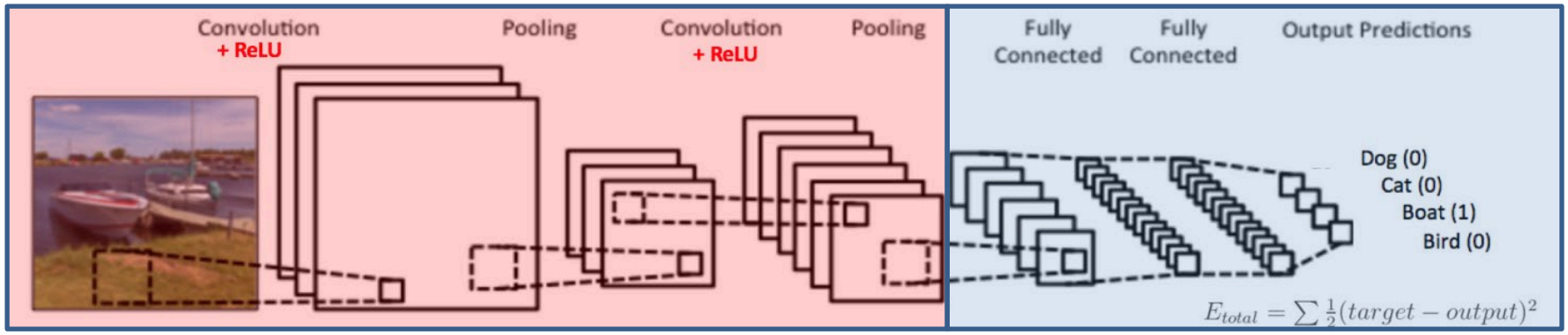
## Epigenome Dynamics: Joint Chromatin State Learning
## Enhancer-gene linking: Correlation, Hi-C, eQTLs
## TF motif discovery: Enrichment, EM, Gibbs Sampling
## Deep learning convolution CNNs for motif discovery
## Global motif discovery: Comparative Genomics
## Motif Instance Identification: Branch Length Score
## Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

# **Human Vision** ⇔ many **layers** of **abstraction** ⇔ **Deep** learning



# Key idea: **Representation** learning




**‘Modern’ Deep learning:**
**Hierarchical Representation Learning**
**Feature extraction**

**‘Classical’ Fully-connected**
**Neural Networks**
**Classification**

In deep learning, the two tasks are **coupled**:

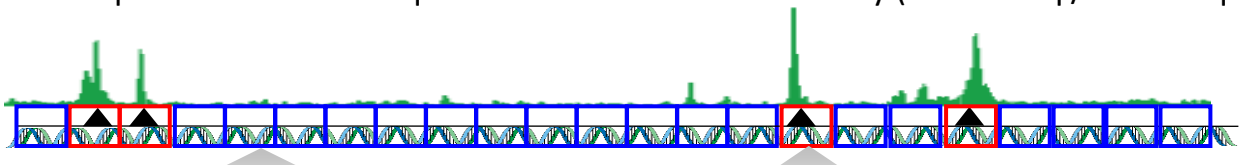
- the classification task “drives” the feature extraction
- Extremely powerful and general paradigm
  - ➔ **Be creative**! The field is still at its infancy!
  - ➔ New application domains (e.g. beyond images) can have **structure** that current architectures **do not capture/exploit**
  - ➔ Genomics/biology/neuroscience can help drive development of **new architectures**

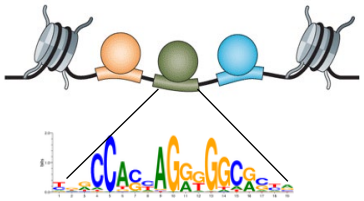
# Key design principles of CNNs (+brain counterparts)

| Property | Human Visual System Property | Deep Learning CNN Building Block |
|---|---|---|
| Locality | Low-level neurons respond to local patches (receptive field) | Local computation of convolutional filters (not a fully-connected network) |
| Filters | Specialized neurons carry out low-level detection operation | Low-level filters carry out the same operation throughout the network |
| Layers / abstraction | Layers of neurons learn increasingly abstract 'concepts' | Layers of hidden units, abstract concepts learned from simpler parts / building blocks |
| Threshold | Neurons fire after cross activation threshold ➔ non-linearity | Activation functions introduce non-linearities ➔ expand universe of functions |
| Pooling | Higher-level neurons invariant to exact position, sum/max of prev. | Max/Avg pooling layers: positional invarnce reduced # parameters, speed up compute |
| Multimodal | Different neurons extract different features of image | Multiple filters applied simultaneously, each captures different aspects of original image |
| Sampling Density | Central vision sampled densely by photoreceptors than periphery | Adjust stride of filter application to denser (slower) vs. sparser (faster) sampling |
| Saturation | Neurons 'tired' after activation, signal quiets down | Limiting weight of individual hidden units, dropout learning, regularization |
| Learn/Reinforcement | Useful connections strengthened over time | Back-propagation, adjusting weights across the hierarchy |
| Feed-fward edges | Neurons with long connections from lower levels to higher ones | Residual networks (ResNets) feed lower-level signal, avoid vanishing gradients |

# Predictive model of regulatory DNA

Transcription factor ChIP-seq data OR chromatin accessibility (DNase-seq / ATAC-seq data)





…GACTTGAAACGGCATTG
… Inactive (0) (0.3)

…GACAGATAATGCATTGA…
Active (+1) (20.2)

…GACAGATAATGCATTGA…

…ACTGTCATGGATATTCT…

…GATATTCTACTGTAAG…

DNA sequences ($S_i$)

…CAACCTTGAACGGCATTG…

…GACTTGAAACGGCATTG…

…CAGTATGCATACGTGAA…

Classification or Regression model $F(S_i)$

*Arvey et al. 2012*
*Ghandi et al. 2014*
*Setty et al. 2015*

Class = +1 (20.2)

Class = +1 (10.6)

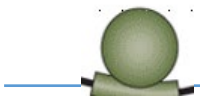
Class = +1 (15.8)

Measured Labels ($Y_i$)

Class = 0 (0.3)

Class = 0 (1.2)

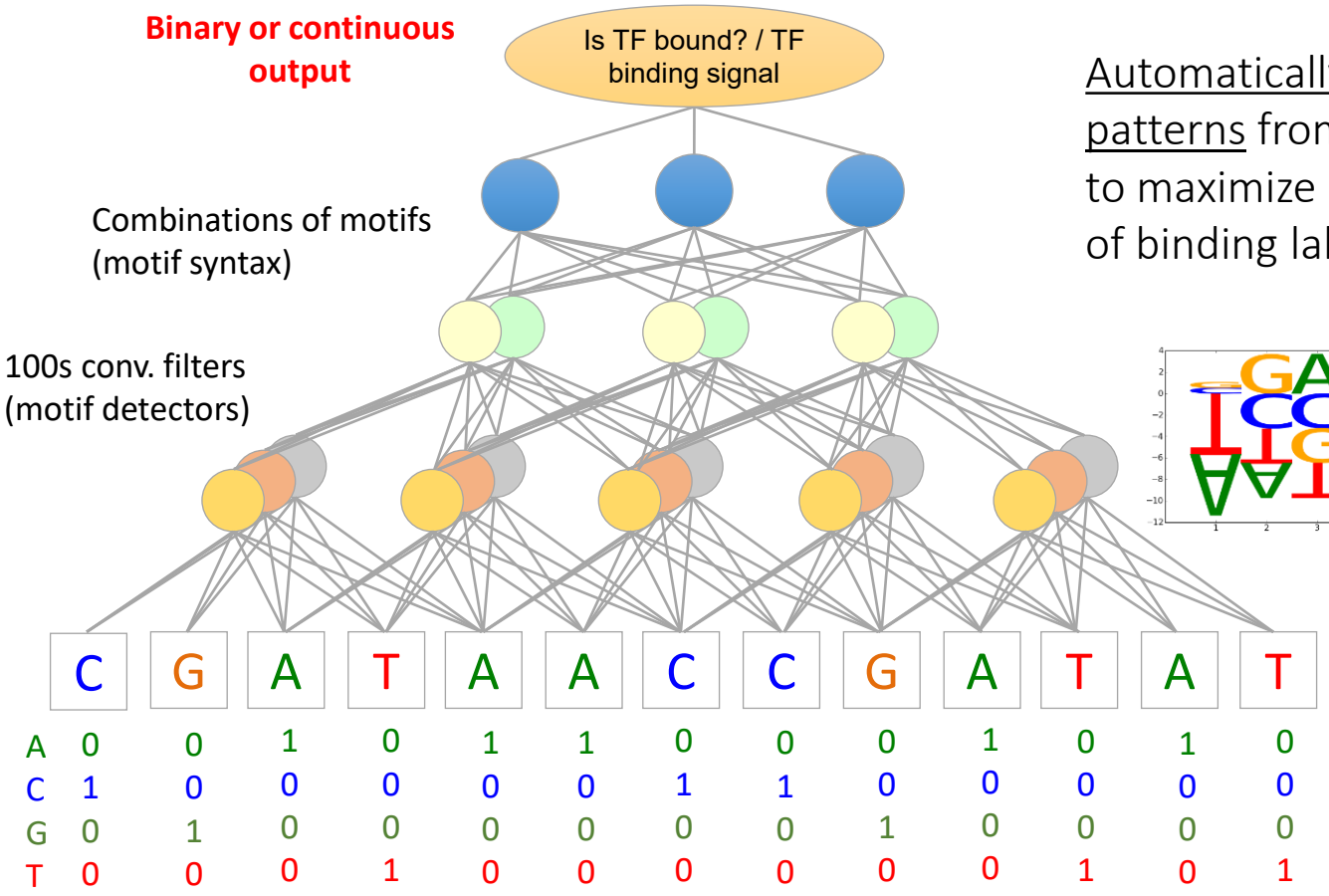
Class = 0 (3.5)



Bound

Unbound

# Convolutional neural network (CNN) with DNA sequence inputs



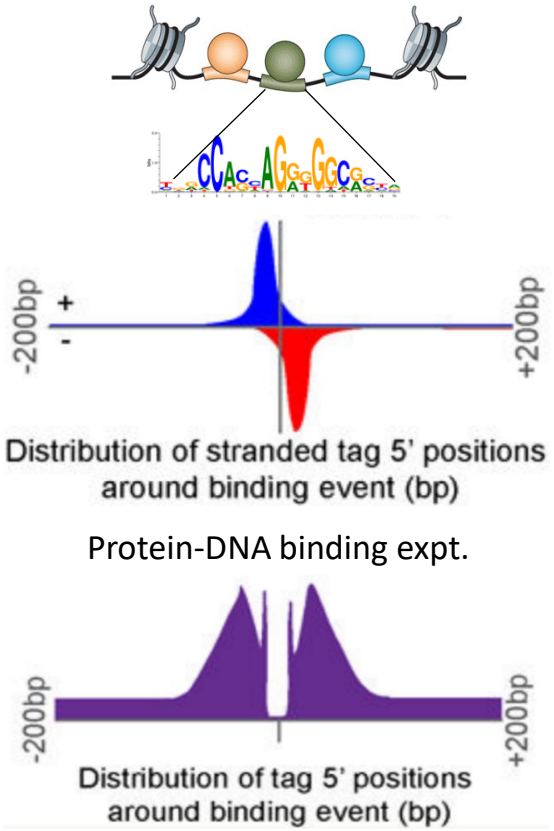



Automatically learns predictive patterns from raw DNA sequence to maximize prediction accuracy of binding labels

**One-hot encoded input:** DNA sequence represented as ones and zeros

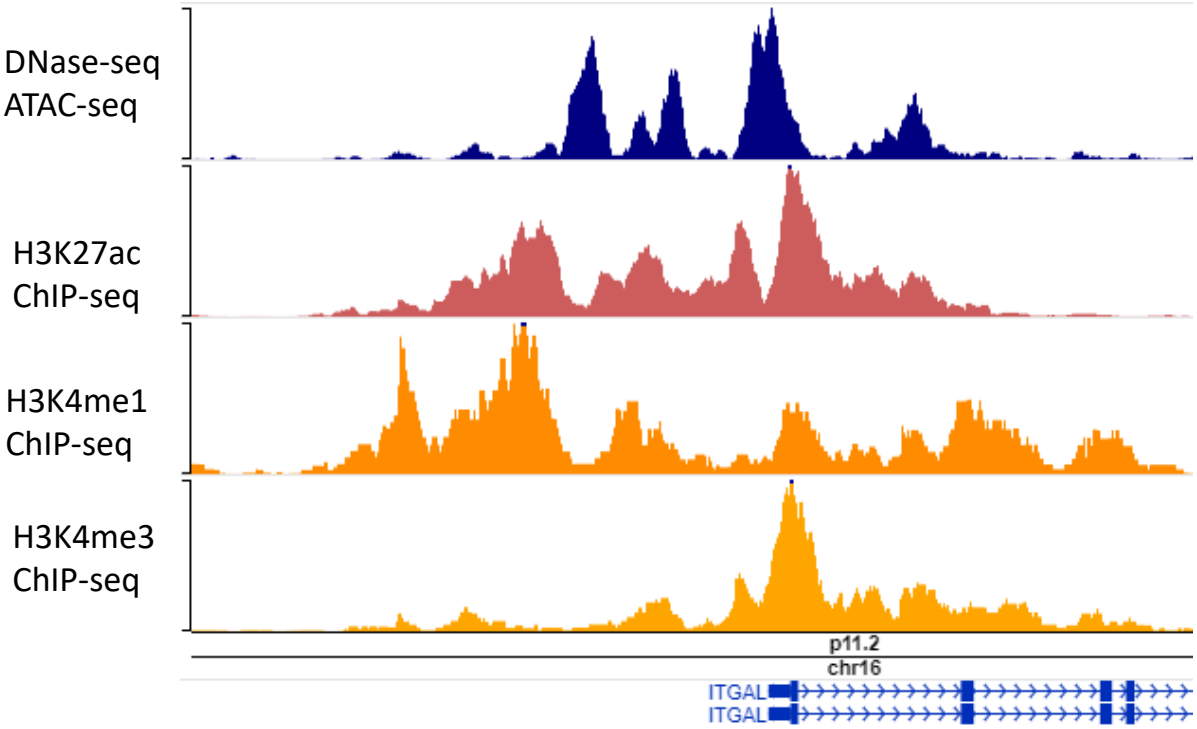
*Alipanahi et al. 2015 (DeepBind)*
*Kelley et al. 2016 (Basset)*
*Zhou et al. 2015 (DeepSEA)*

# High-resolution ‘shapes’ and ‘spans‘ of TF and chromatin profiles capture exquisite information about protein-DNA contacts




Protein-DNA binding expt.

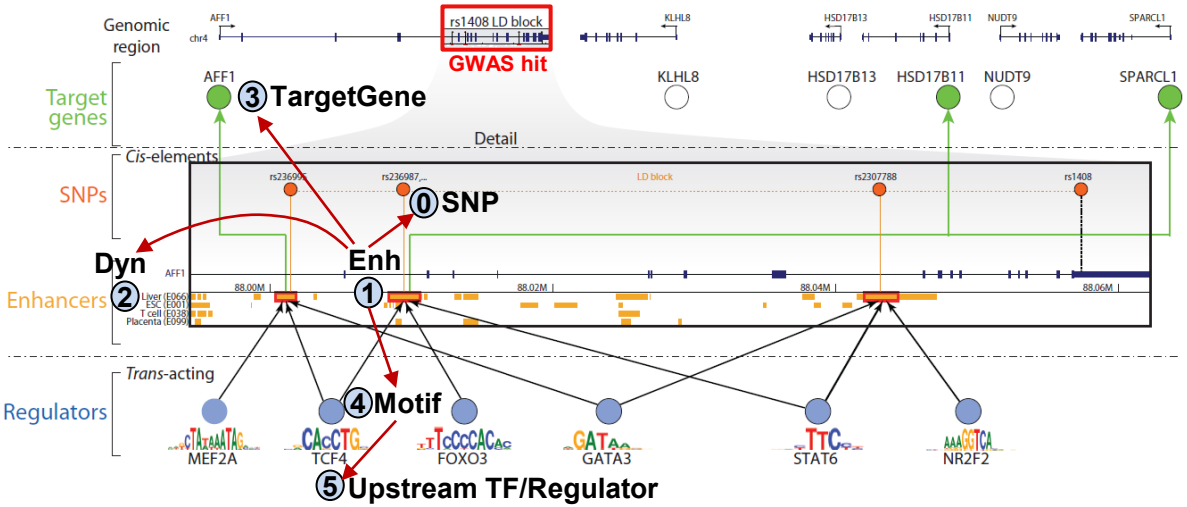
DNA accessibility experiments




https://doi.org/10.3109/10409238.2015.1051505

MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution




# Lecture 7: Regulatory circuitry

## Epigenome Dynamics: Joint Chromatin State Learning

## Enhancer-gene linking: Correlation, Hi-C, eQTLs

## TF motif discovery: Enrichment, EM, Gibbs Sampling

## Deep learning convolution CNNs for motif discovery

## Global motif discovery: Comparative Genomics

## Motif Instance Identification: Branch Length Score

## Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

# Regulatory genomics: motifs, instances, regions

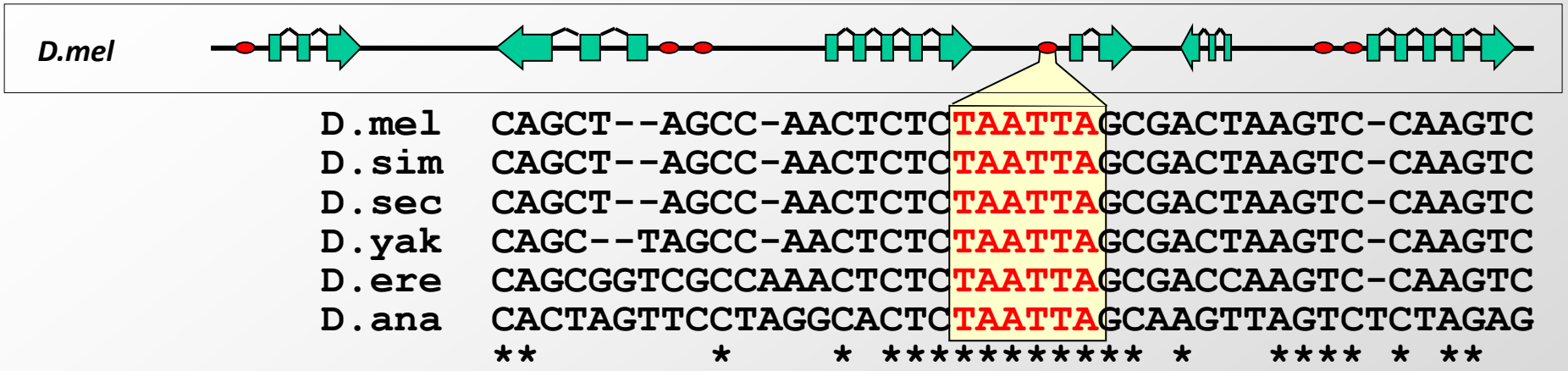
1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint (M,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

# Motivation for *de novo* genome-wide motif discovery

- Both TF and region centric approaches are not comprehensive and are biased
- TF centric approaches generally require transcription factor (or antibody to factor)
  - Lots of time and money
  - Also have computational challenges
- *De novo* discovery using conservation is unbiased but can't match motif to factor and require multiple genomes

# Evolutionary signatures for regulatory motifs

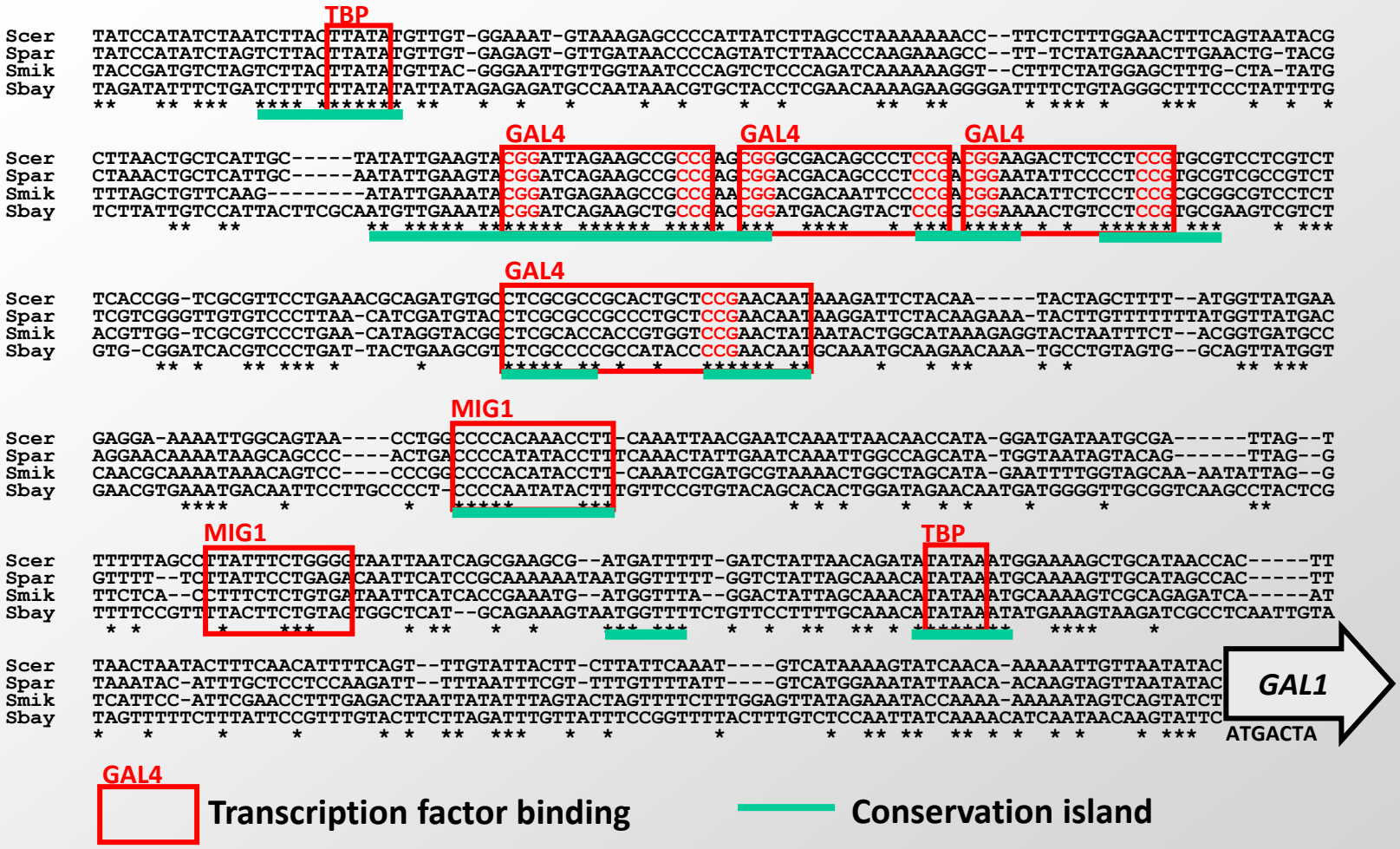
**Known engrailed binding site**




- Start by looking at known motif instances
- Individual motif instances are preferentially conserved
- Can we just take conservation islands and call them motifs?
  - No. Many conservation islands are due to chance or perhaps due to non-motif conservation

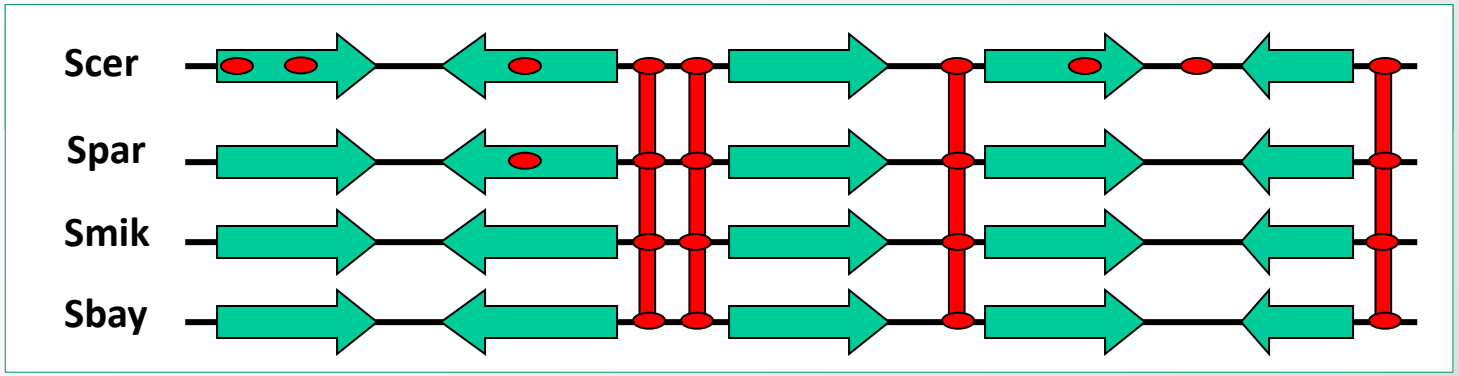
**Kellis *el al*, Nature 2003**
**Xie *et al.* Nature 2005**
**Stark *et al*, Nature 2007**

# Conservation islands overlap known motifs




**Increase power by testing conservation in many regions**

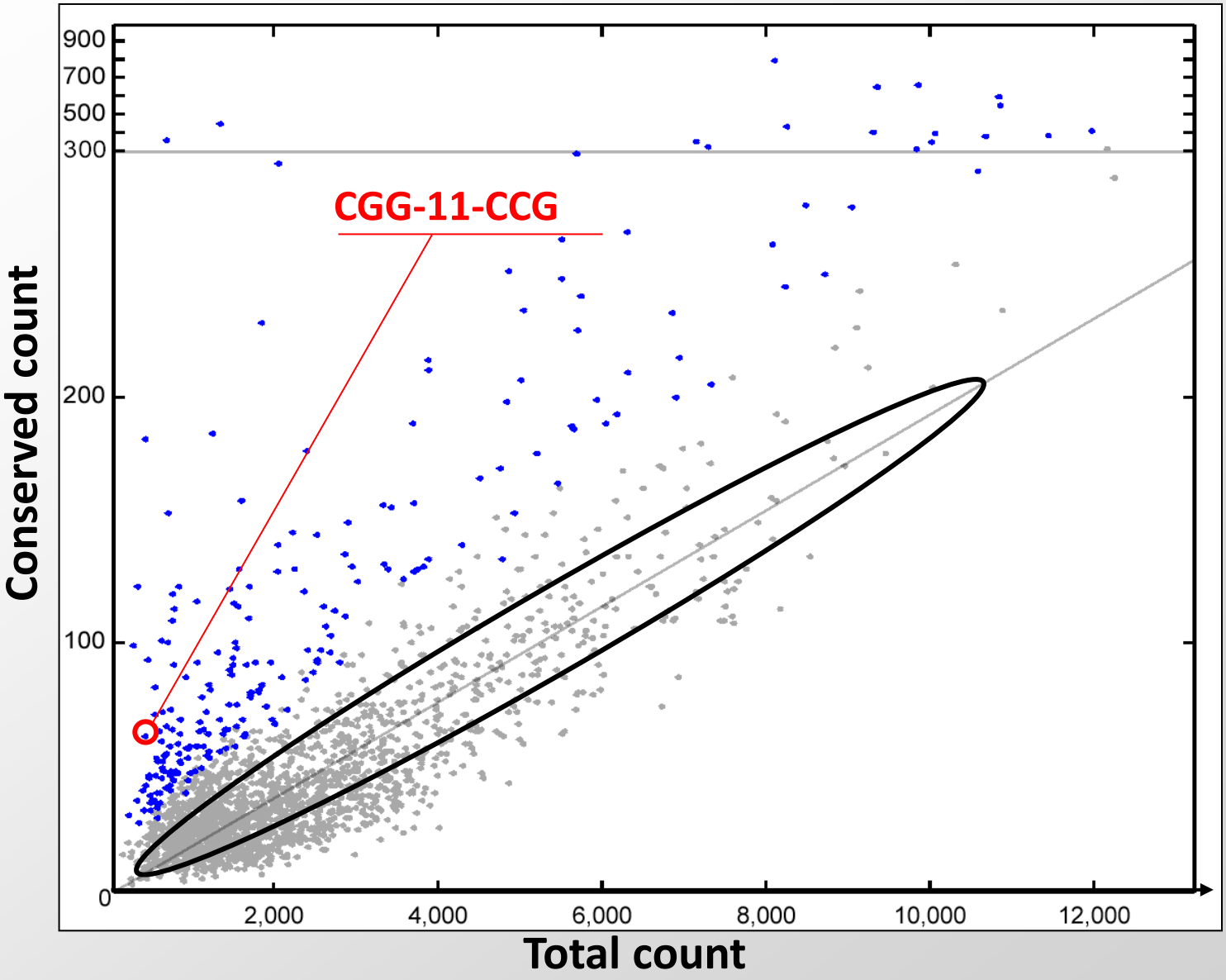
# Genome-wide conservation



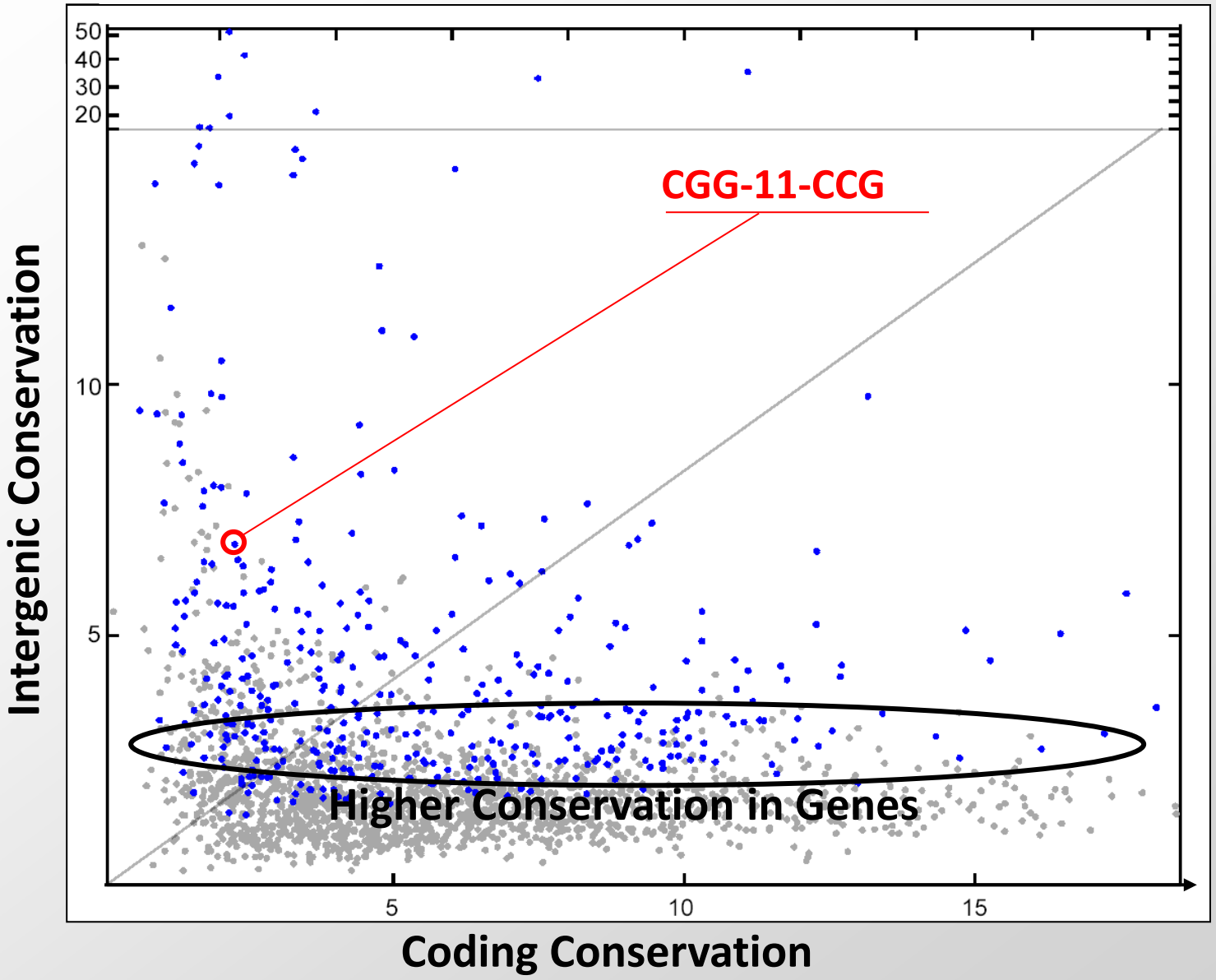

| Evaluate conservation within: | Gal4 | Controls |
|---|---|---|
| (1) All intergenic regions | 13% | 2% |
| (2) Intergenic : coding | 13% : 3% | 2% : 7% |
| (3) Upstream : downstream | 12:0 | 1:1 |

**A signature for regulatory motifs**

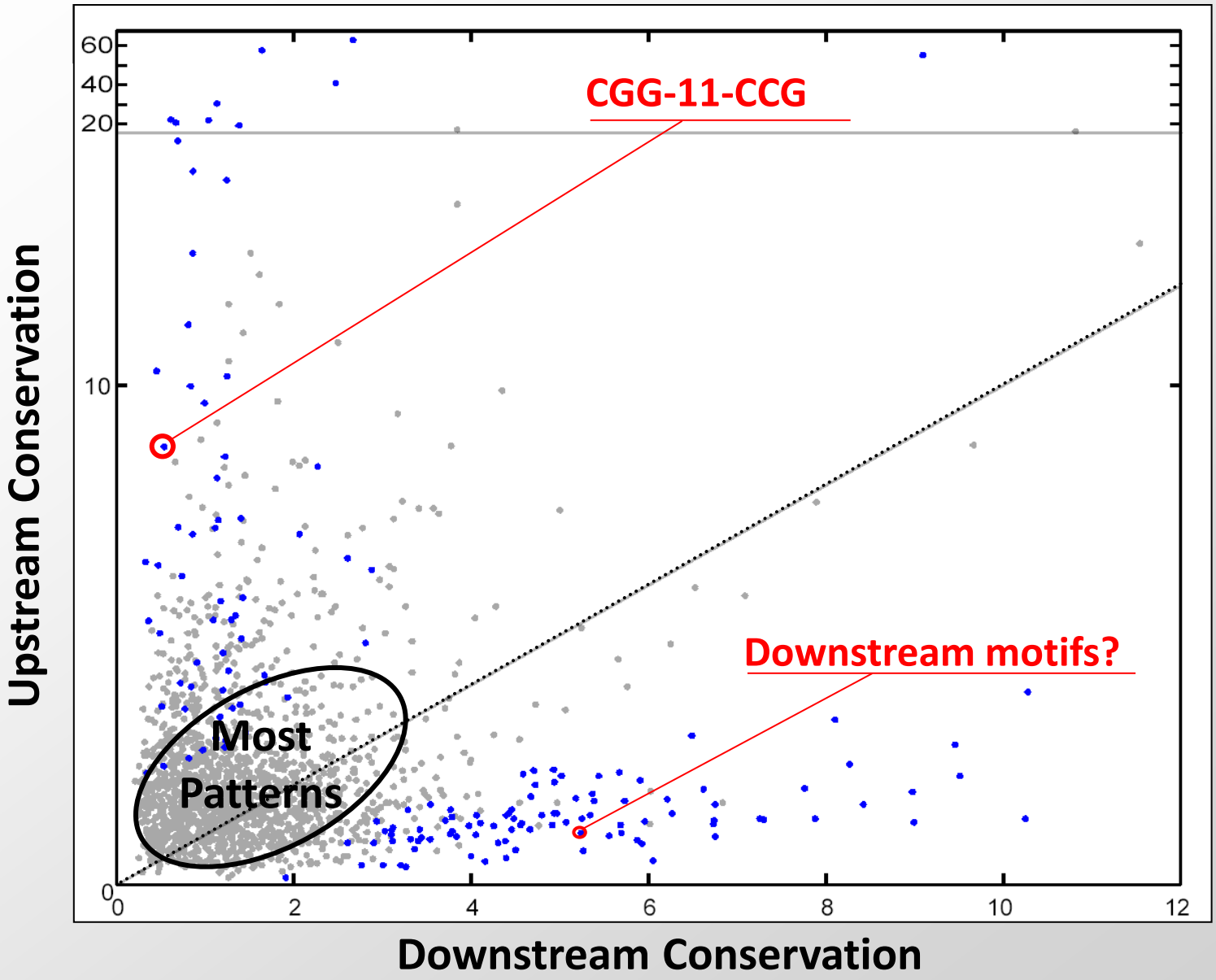
# Test 1: Intergenic conservation



# Test 2: Intergenic vs. Coding



# Test 3: Upstream vs. Downstream



# Conservation for TF motif discovery

1. **Enumerate motif seeds**

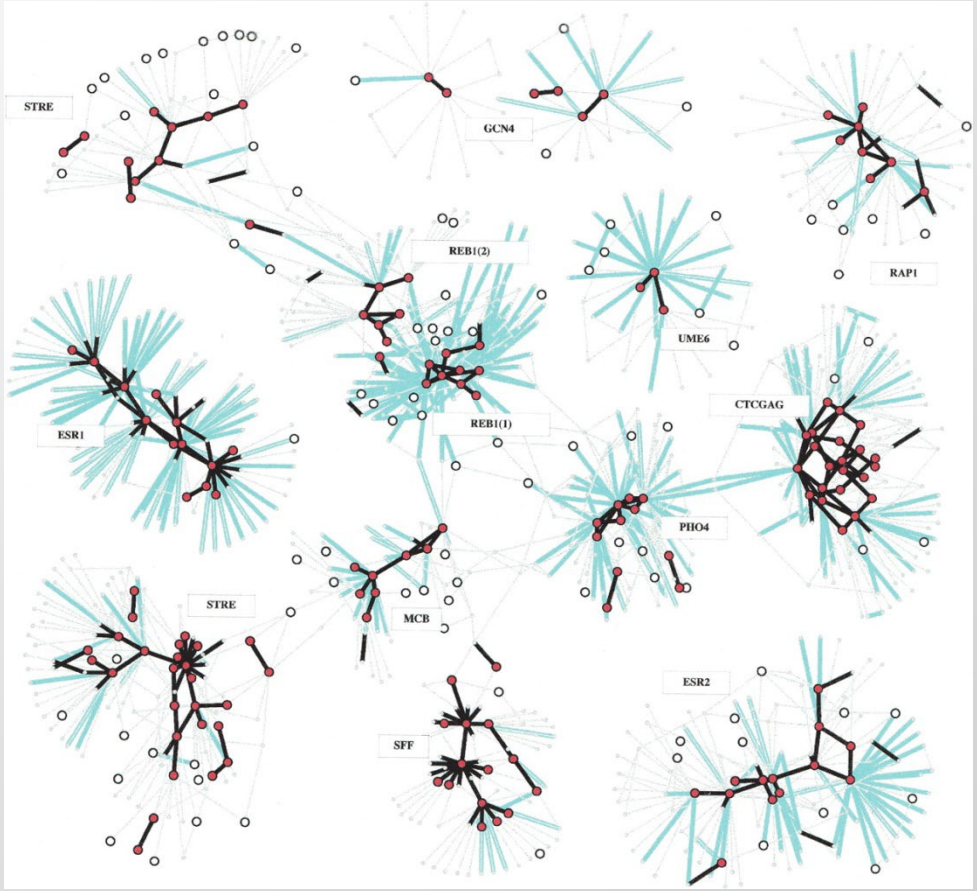

   - Six non-degenerate characters with variable size gap in the middle
2. **Score seed motifs**
   - Use a conservation ratio corrected for composition and small counts to rank seed motifs
3. **Expand seed motifs**


   - Use expanded nucleotide IUPAC alphabet to fill unspecified bases around seed using hill climbing
4. **Cluster to remove redundancy**
   - Using sequence similarity

**Kellis, Nature 2003**

# Learning motif degeneracy using evolution

- Record frequency with which one sequence is "replaced" by another in evolution
- Use this to find clusters of k-mers that correspond to a single motif




Tanay, Genome Research 2004

# Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint (M,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

# Validation of the discovered motifs

- Because genome-wide motif discovery is *de novo*, we can use functional datasets for validation
  - Enrichment in co-regulated genes
  - Overlap with TF binding experiments
  - Enrichment in genes from the same complex
  - Positional biases with respect to transcription start
  - Upstream vs. downstream / inter vs. intra-genic bias
  - Similarity to known transcription factor motifs
- Each of these metrics can also be used for discovery
  - In general, split metrics into discovery vs. validation
  - As long as they are *independent* !
  - Strategies that combine them all lose ability to validate
    - Directed experimental validation approaches are then needed

# Similarity to known motifs

- If discovered motifs are real, we expect them to match motifs in large databases of known motifs
- We find this (significantly higher than with random motifs)
- Why not perfect agreement?
  - Many known motifs are not conserved
  - Known motifs are biased; may have missed real motifs

| MCS | Discovered motif | Known Factor |
|---|---|---|
| 46.8 | **GGGCGG**R | SP-1 |
| 34.7 | **GCCATnTT**g | YY1 |
| 32.7 | **CACGTG** | MYC |
| 31.2 | G**ATTGGY** | NF-Y |
| 30.8 | **TGA**n**TCA** | AP-1 |
| 29.7 | **GGGAGG**RR | MAZ |
| 29.5 | **TGACGTM**R | CREB |
| 26.0 | **CGGCCAT**YK | NF-MUE1 |
| 25.0 | **TGACCTTG** | ERR□ |
| 22.6 | **CCGGAAR**Y | ELK-1 |
| 19.8 | S**CGGAAG**Y | GABP |
| 17.9 | **CA**T**TTCC**K | STAT1 |

Xie, Nature 2005

**70/174 mammalian motifs**

| MCS | Discovered motif | Known Factor |
|---|---|---|
| 65.6 | CTAATTAAA | en |
| 57.3 | TTKCAATTAA | repo |
| 54.9 | WATTRATTK | ara |
| 54.4 | AAATTKATGC | prd |
| 51 | GCAATAAA | vvl |
| 46.7 | DTAATTTRYN | Ubx |
| 45.7 | TGATTAAT | ap |
| 43.1 | YMATTAAAA | abd-A |
| 41.2 | AAACNNGTT | |
| 40 | RATTKAATT | |
| 39.5 | GCACGTGT | ftz |
| 38.8 | AACASCTG | br-Z3 |

Stark, Nature 2007

**35/145 fly motifs**

# Positional bias of motif matches

- Motifs are involved in initiation of transcription

→Motif matches biased versus TSS

  – 10% of fly motifs
  – 34% of mammalian motifs

→Depletion of TF motifs in coding sequence

  – 57% of fly motifs

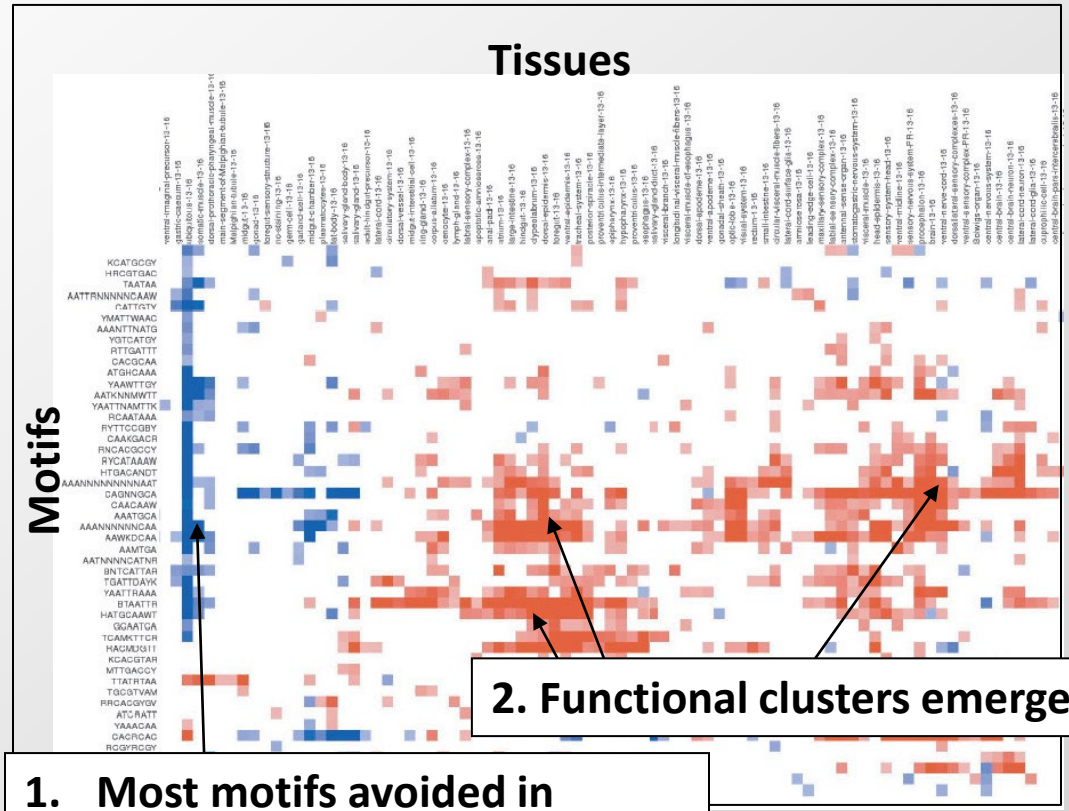
→Clustering of motif matches

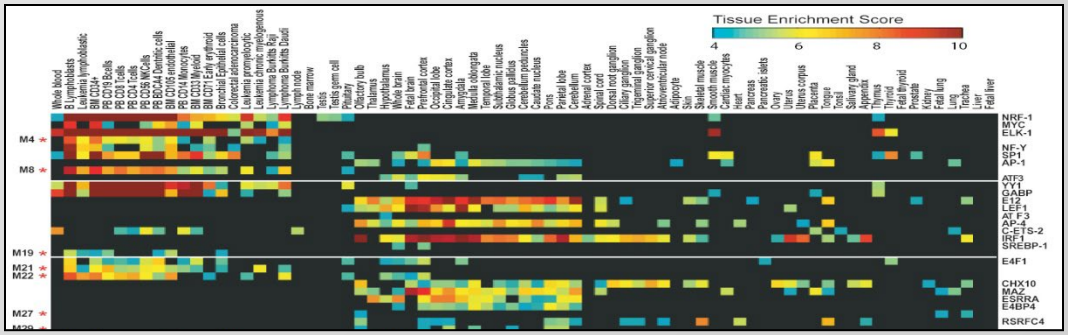
  – 19% of fly motifs

# Motifs have functional enrichments

For both fly (top) and mammals (bottom), motifs are enriched in genes expressed in specific tissues

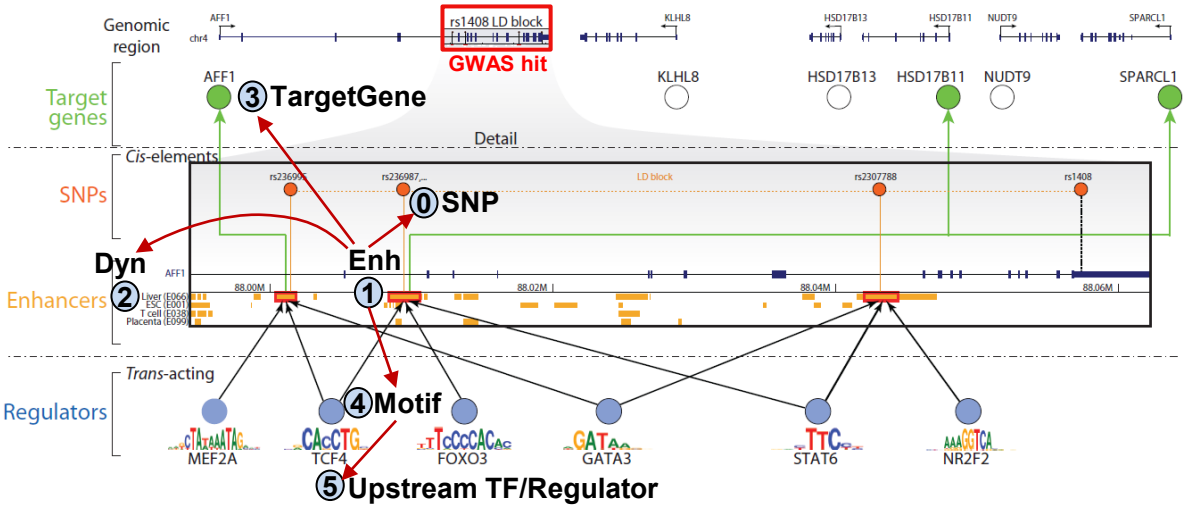
Reveals modules of cooperating motifs



MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution




# Lecture 7: Regulatory circuitry

## Epigenome Dynamics: Joint Chromatin State Learning

## Enhancer-gene linking: Correlation, Hi-C, eQTLs

## TF motif discovery: Enrichment, EM, Gibbs Sampling

## Deep learning convolution CNNs for motif discovery

## Global motif discovery: Comparative Genomics

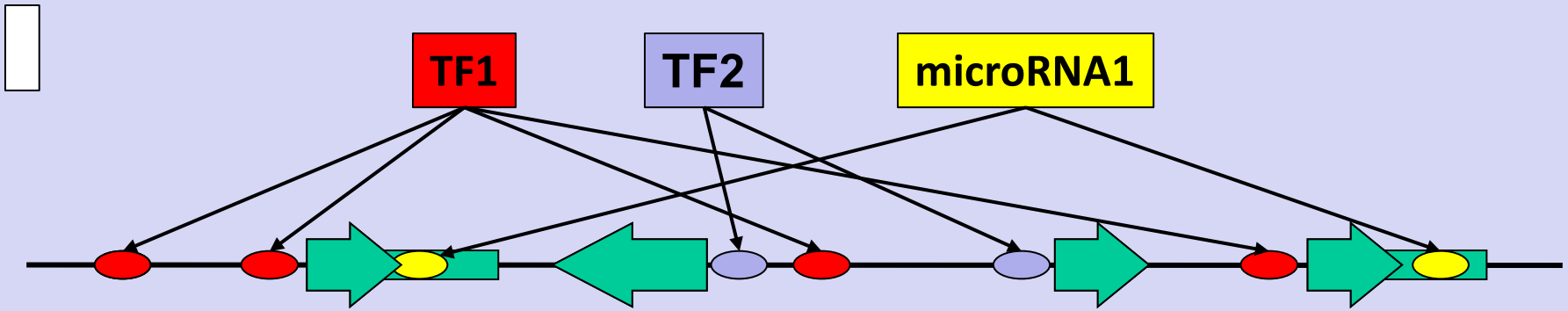
## Motif Instance Identification: Branch Length Score

## Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

# Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint (M,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays




# Motif instance identification

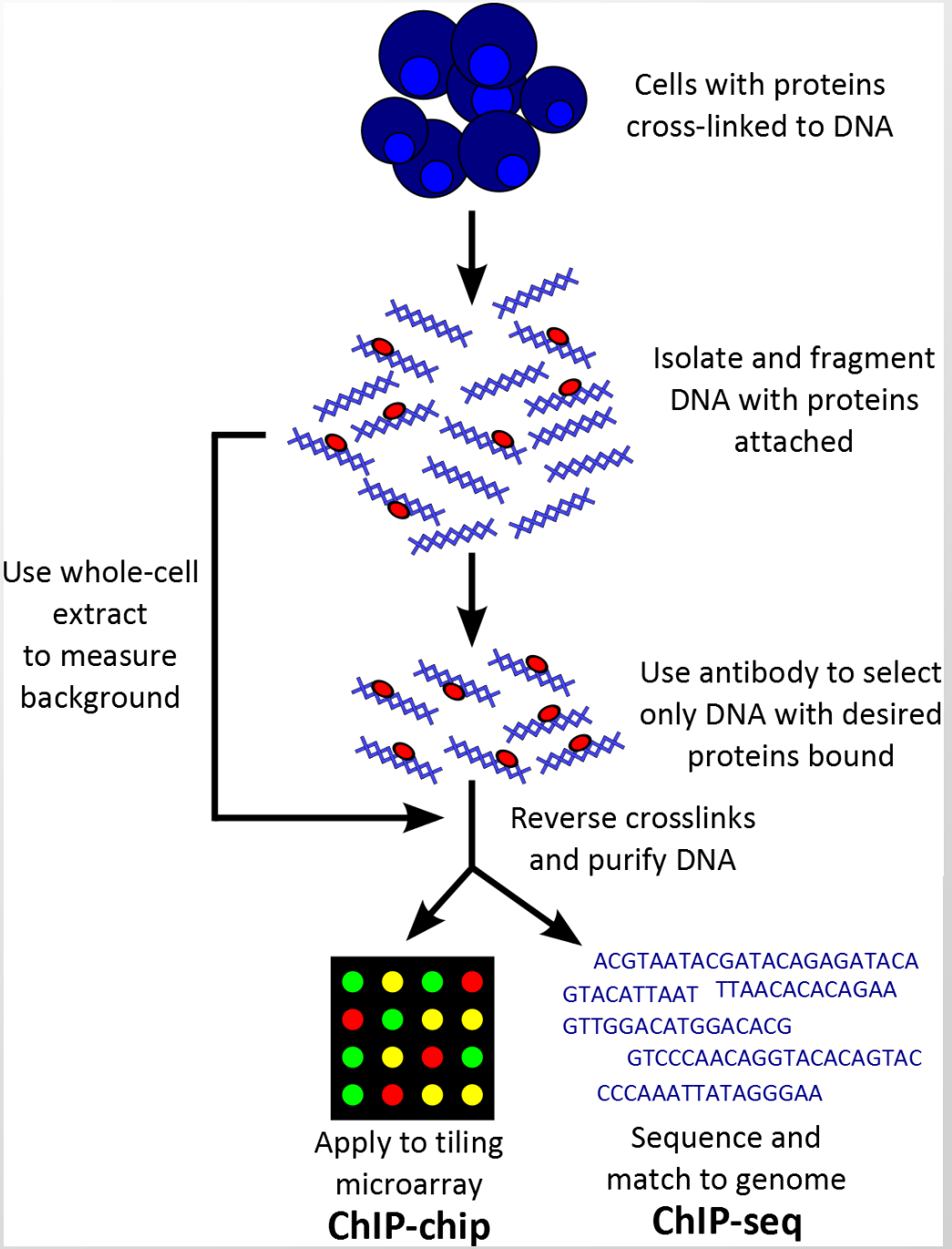
**How do we determine the functional binding sites of regulators?**

Kheradpour, Stark, Roy, Kellis, Genome Research 2007

# Experimental target identification: ChIP-chip/seq

## Limitations :

- Antibody availability
- Restricted to specific stages/tissues
- Biological functionality of most binding sites unknown
- Resolution can be limited (can't usually identify the precise base pairs)



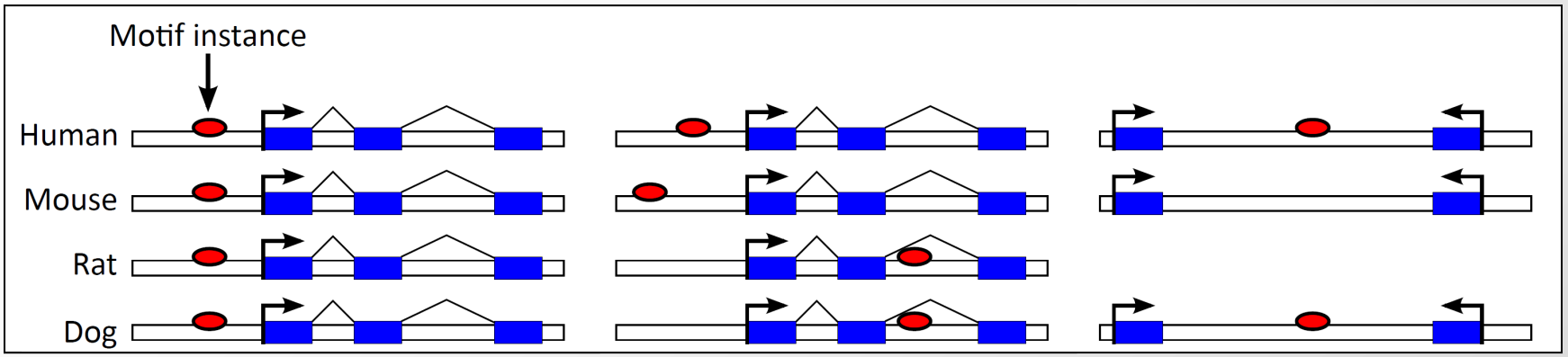

**Ren et al., 2000; Iyer et al., 2001 (ChIP-chip)**

**Robertson et al., 2007 (ChIP-seq)**

# Computational target identification

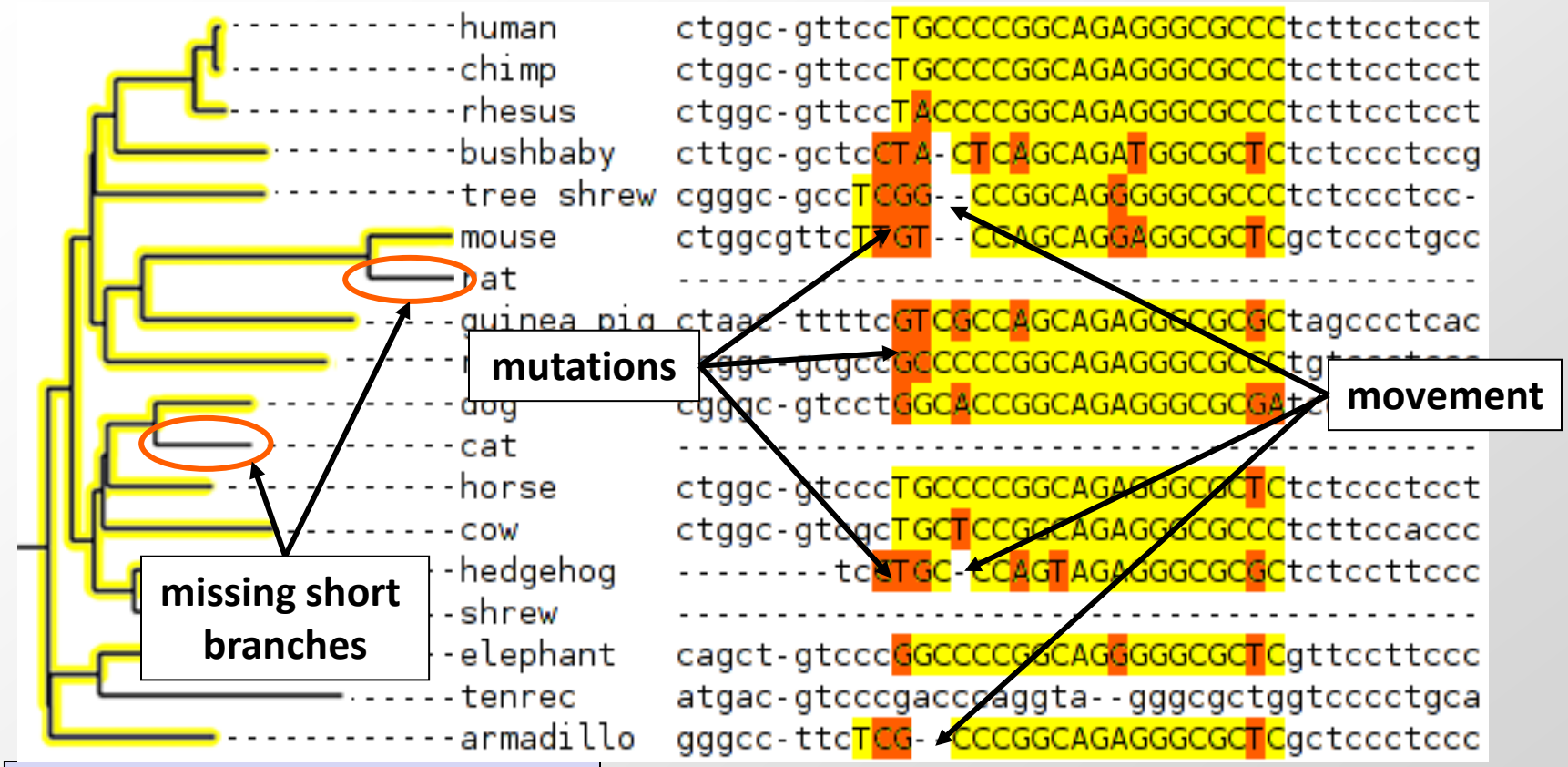
- Single genome approaches using motif clustering (e.g. Berman 2002; Schroeder 2004; Philippakis 2006)
  - Requires set of specific factors that act together
  - Miss instances of motifs that may occur alone
- Multi-genome approaches (phylogentic footprinting) (e.g. Moses 2004; Blanchette and Tompa 2002; Etwiller 2005; Lewis 2003)
  - Tend to either require absolute conservation or have a strict model of evolution

# Challenges in target identification

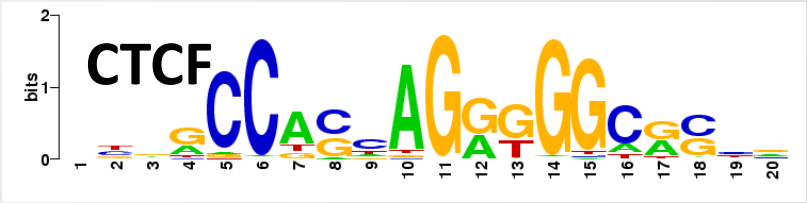



- Simple case
  - Instance fully conserved in orthologous position near genes
- Motif turn-around/movement
  - Motif instance is not found in orthologous place due to birth/death or alignment errors
- Distal/missing matches
  - Due to sequencing/assembly errors or turnover
  - Distal instances can be difficult to assign to gene

# Computing Branch Length Score (BLS)




**BLS = 2.23sps (78%)**



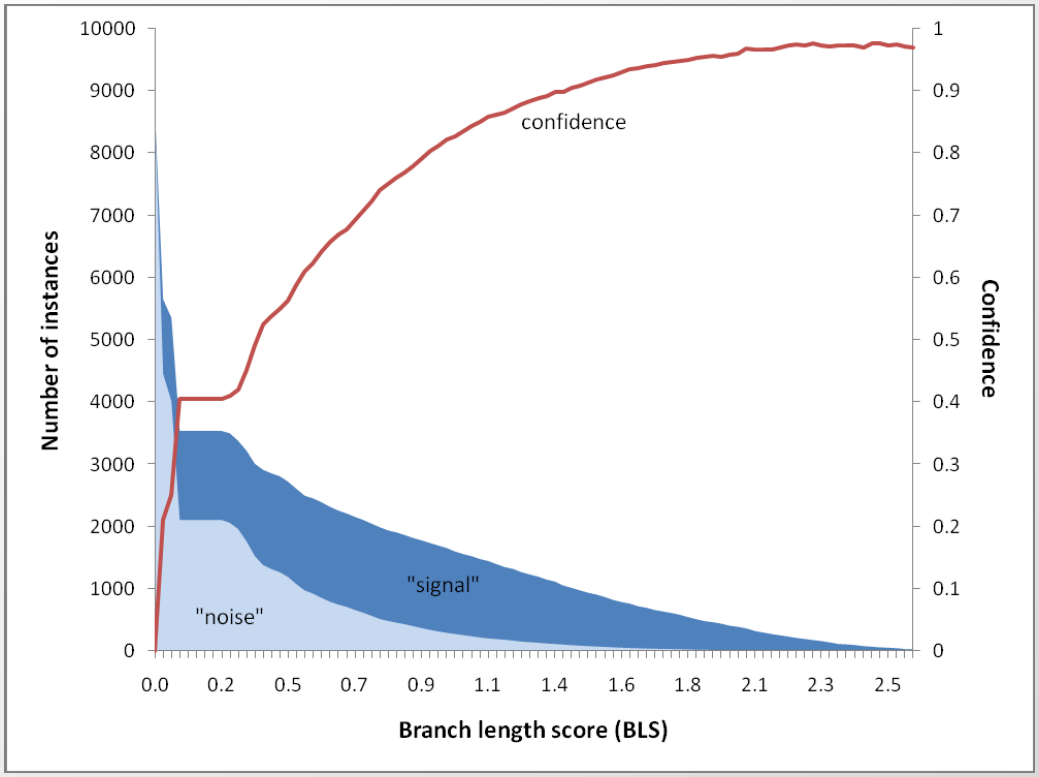

**Allows for:**

1. **Mutations permitted by motif degeneracy**
2. **Misalignment/movement of motifs within window (up to hundreds of nucleotides)**
3. **Missing motif in dense species tree**

# Branch Length Score → Confidence

1. Evaluate chance likelihood of a given score
   - Sequence could also be conserved due to overlap with un-annotated element (e.g. non-coding RNA)
2. Account for differences in motif composition and length
   - For example, short motif more likely to be conserved by chance

# Branch Length Score ➔ Confidence



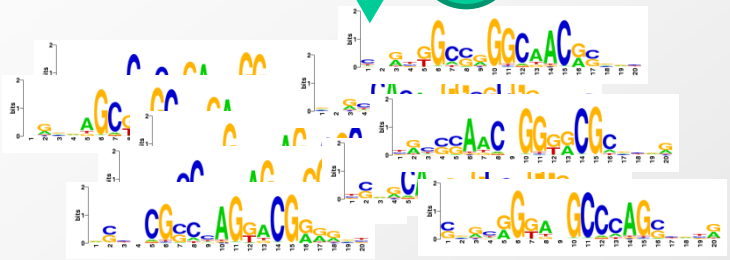

1. **Use motif-specific shuffled control motifs determine the expected number of instances at each BLS by chance alone or due to non-motif conservation**
2. **Compute Confidence Score as fraction of instances over noise at a given BLS (=1 – false discovery rate)**

# Producing control motifs

When evaluating the conservation, enrichment, etc, of motifs, it is useful to have a set of "control motifs"

Original motif

1 **Produce 100 shuffles of our original**



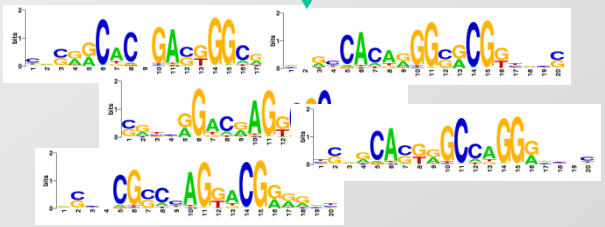
Genome sequence

Known motifs

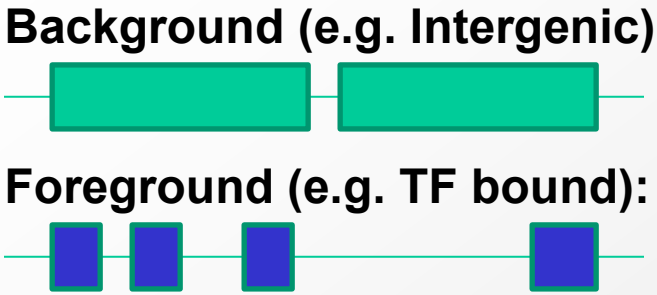
2 **Filter motifs, requiring they match the genome with about (+/- 20%) of our original motif**

3 **Sort potential control motifs based on their similarity to other known motifs**

4 **Cluster potential control motifs and take at most one from each cluster, in increasing order of similarity to known motifs**

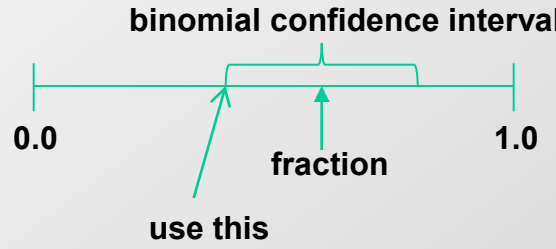


# Computing enrichments: background vs. foreground



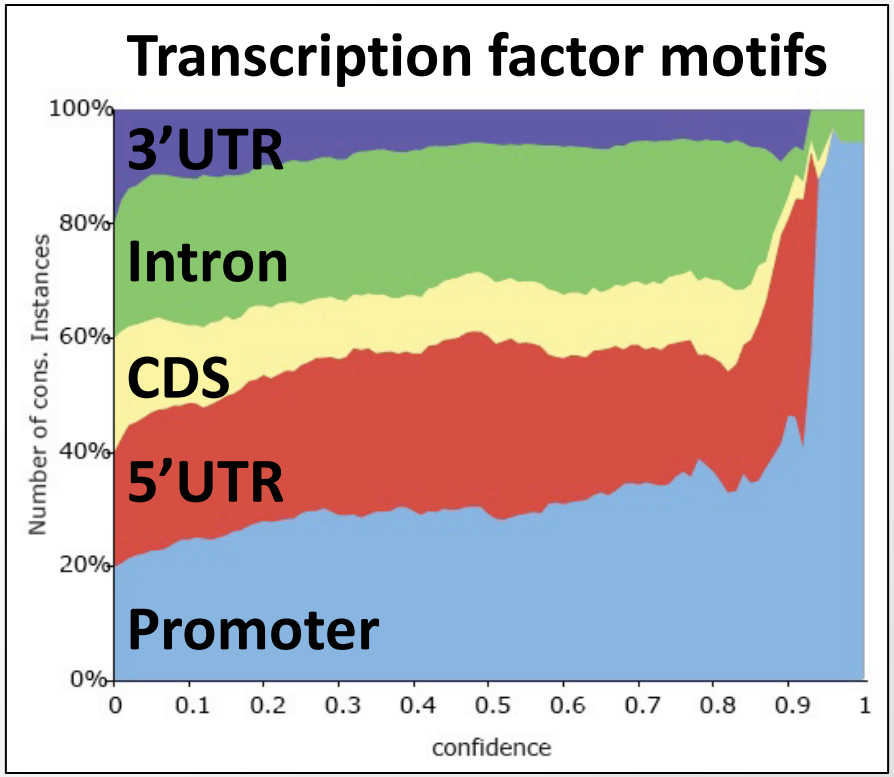

$$\frac{\#\text{ in foreground}}{\#\text{ in background}} \div \frac{\text{size of foreground}}{\text{size of background}}$$

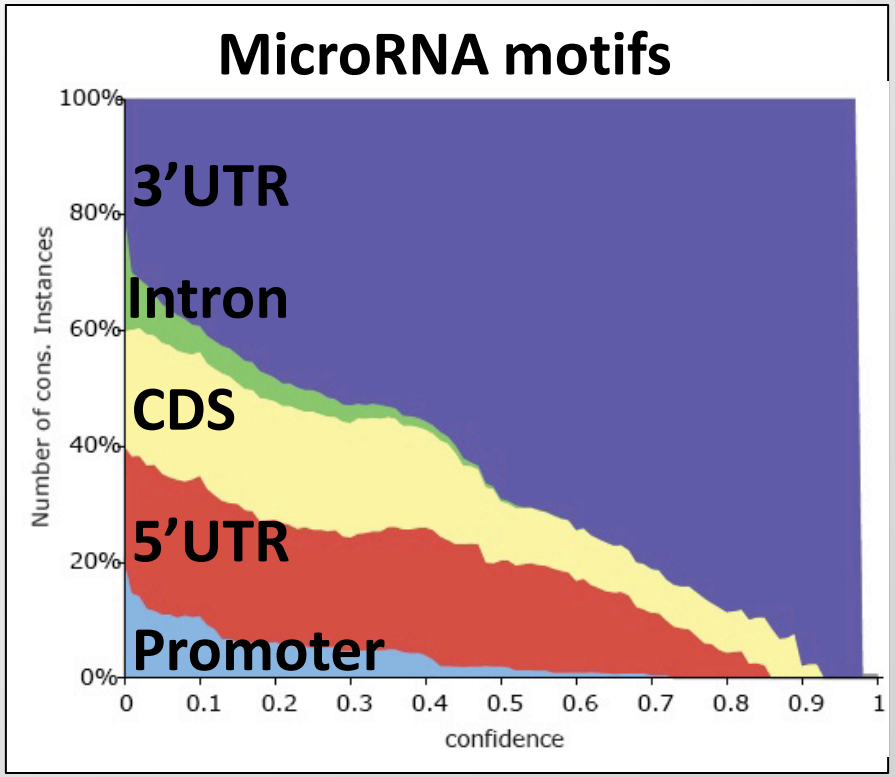
$$\frac{\#\text{ in foreground}}{\#\text{ in background}} \div \frac{\#\text{ control in foreground}}{\#\text{ control in background}}$$




- Background vs. forgeround
  - co-regulated promoters vs. all genes
  - Bound by TF vs. other intergenic regions
- Enrichment: ***fraction of motif instances in foreground*** vs. ***fraction of bases in foreground***
- Correct for composition/conservation level: compute enrichmt w/control motifs
  - Fraction of motif instances can be compared to **fraction of control motif instances in foreground**
  - A hypergeometric p-value can be computed (similar to $\chi^2$, but better for small numbers)
- Fractions can be made more conservative using a binomial confidence interval

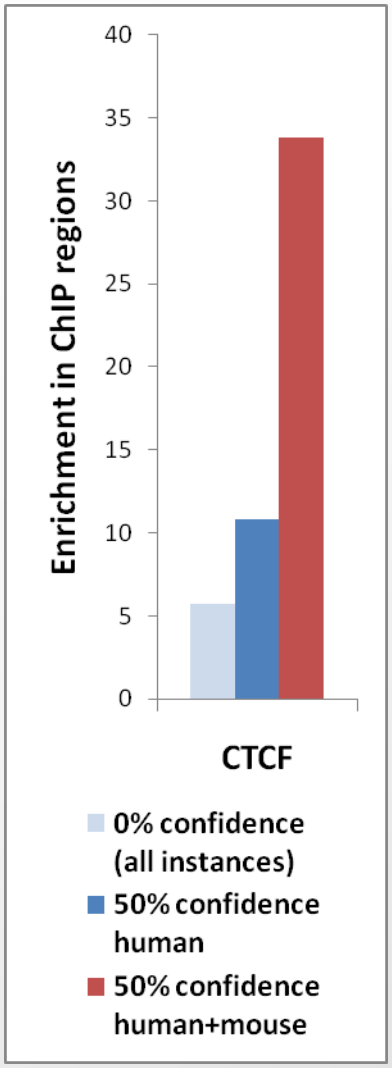
# Confidence selects for functional instances



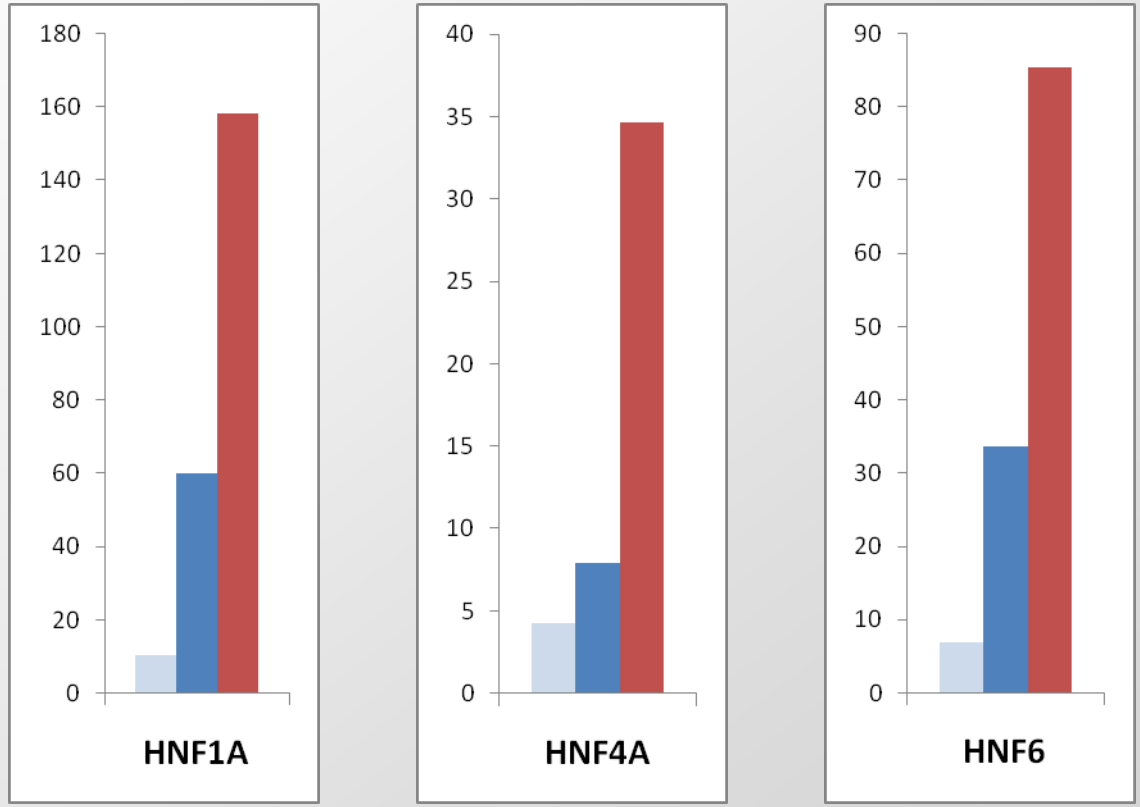




1. Confidence selects for transcription factor motif instances in promoters and miRNA motifs in 3' UTRs

# More enrichment when binding conserved

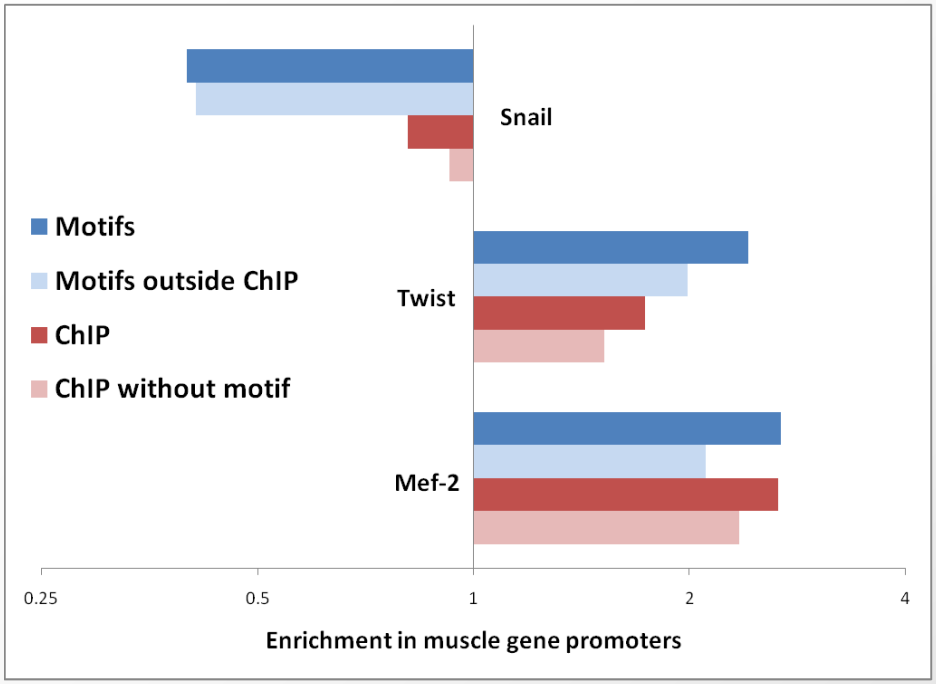


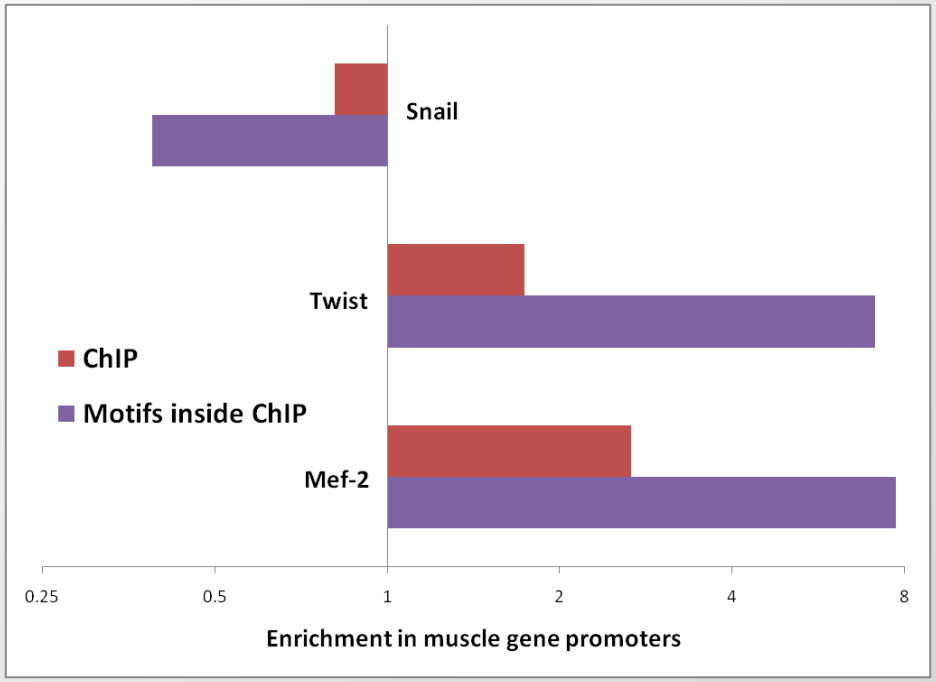

Human: Barski, *et al.*, *Cell* (2007)
Mouse: Bernstein, unpublished




1. ChIP bound regions may not be conserved
2. For CTCF we also have binding data in mouse
3. Enrichment in intersection is dramatically higher
4. Trend persists for other factors where we have multi-species ChIP data

# Comparing ChIP to Conservation



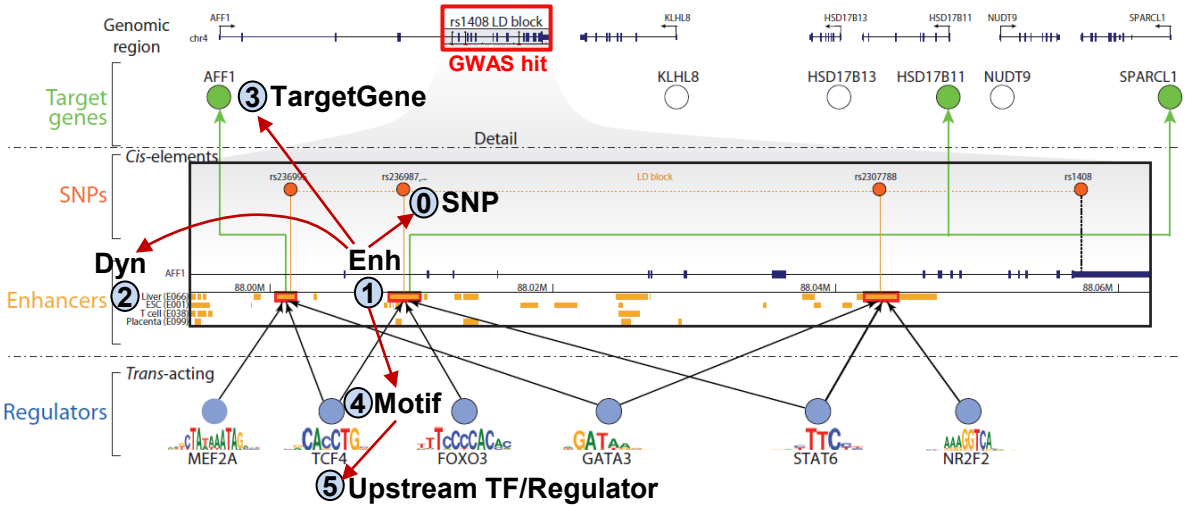




1. **Motifs at 60% confidence and ChIP have similar enrichments (depletion for the repressor Snail) in the functional promoters**
2. **Enrichments persist even when you look at non-overlapping subsets**
3. **Intersection of two regions has strongest signal**
4. **Evolutionary and experimental evidence is complementary**
   - **ChIP includes species specific regions and differentiate tissues**
   - **Conserved instances include binding sites not seen in tissues surveyed**

ChIP data from: Zeitlinger, *et al*., *G&D* (2007); Sandmann, *et al,*. *G&D* (2007); Sandmann, *et al., Dev Cell* (2006)

MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution




# Lecture 7: Regulatory circuitry

## Epigenome Dynamics: Joint Chromatin State Learning

## Enhancer-gene linking: Correlation, Hi-C, eQTLs

## TF motif discovery: Enrichment, EM, Gibbs Sampling

## Deep learning convolution CNNs for motif discovery

## Global motif discovery: Comparative Genomics

## Motif Instance Identification: Branch Length Score

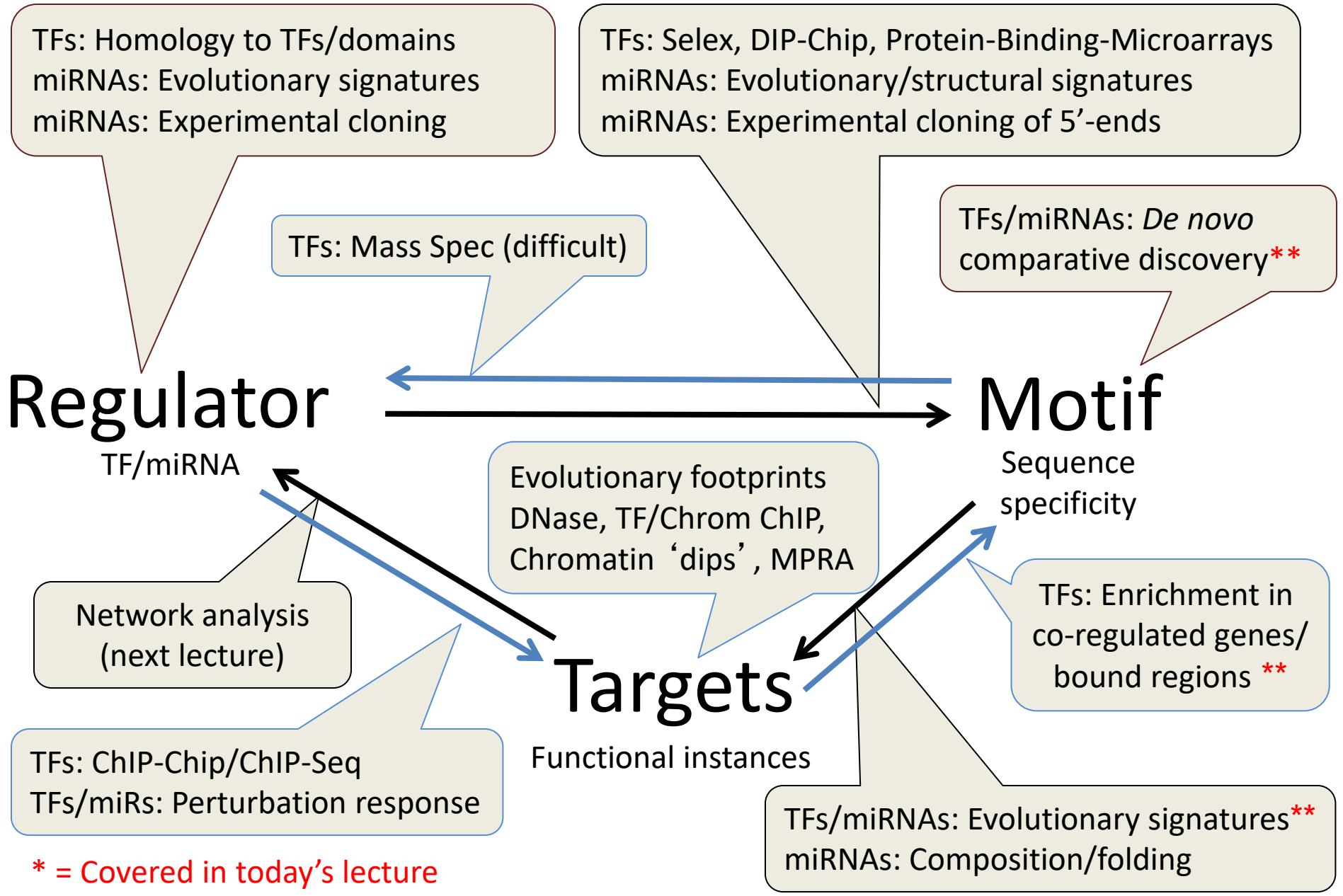
## Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

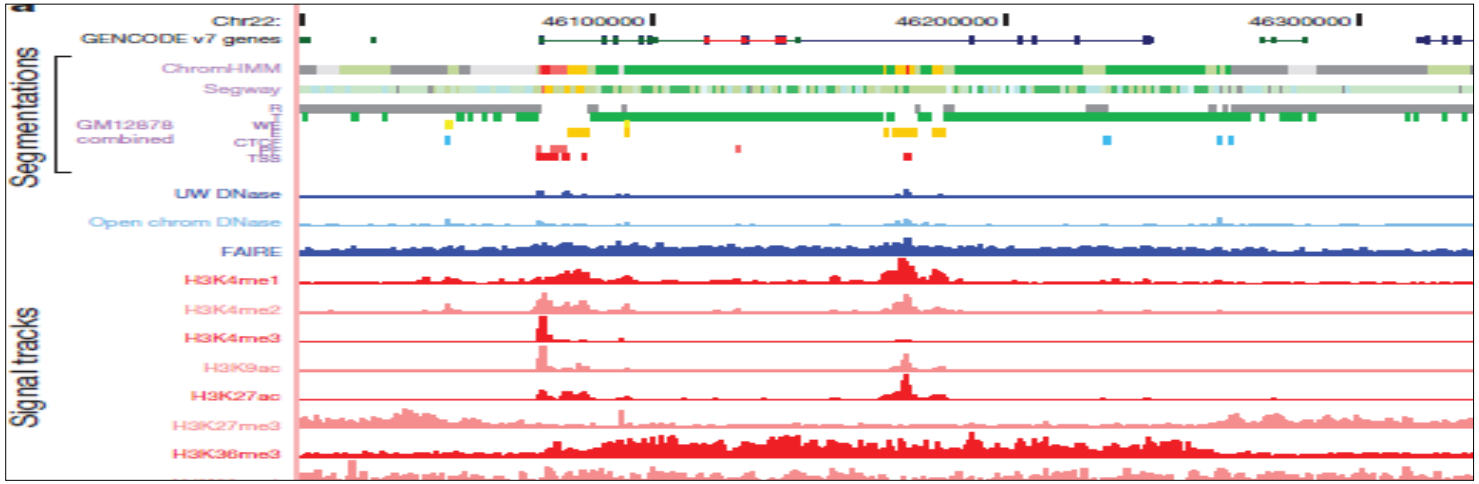
# Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint (M,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

# Challenges in regulatory genomics



# From Identification to Large-Scale Confirmation and Dissection of Candidate Regulatory Regions




ENCODE Project Consortium et al, 2012

ENCODE, Roadmap Epigenomics, *et al*: Histone marks, TF binding, DNase, FAIRE, …

➔ identification of **candidate** regulatory regions

**Next challenge: confirm/dissect 10,000s of regions!**

- Test **thousands** of candidate regulatory regions at once
- Identify regulatory positions at or near **nucleotide level** resolution independent of sequence motifs
- Distinguish **activating vs. repressive** nucleotides

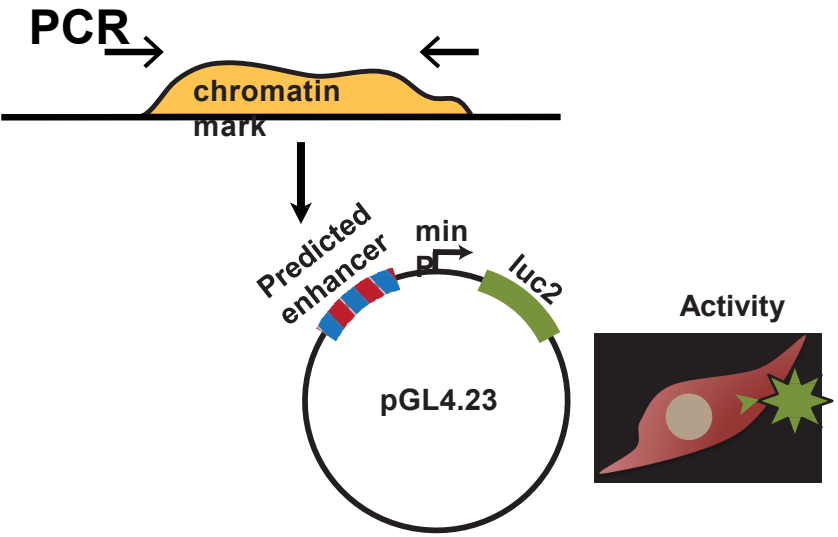
# Problem: Not all annotated enhancers are real

VISTA Enhancer Browser
whole genome enhancer browser

2659 *in vivo* tested elements
1444 elements with enhancer activity

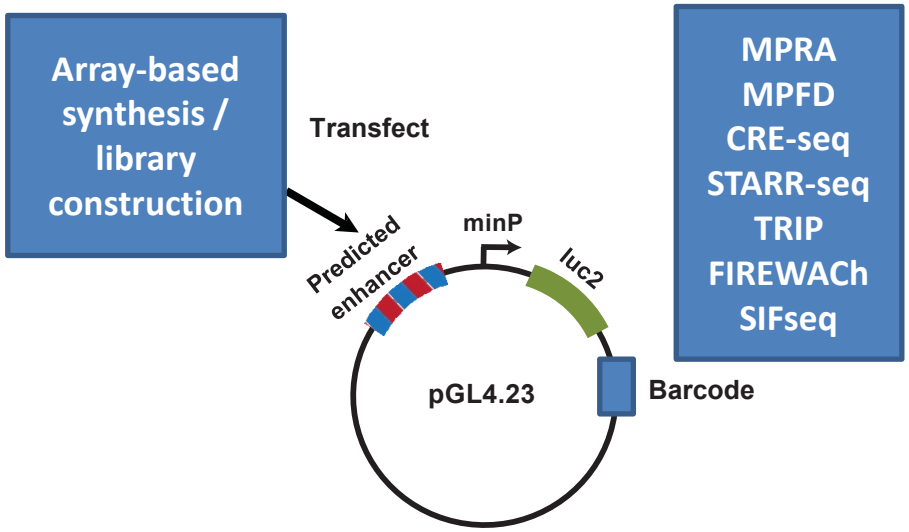
Visel *et al.* NAR 2007

## Luciferase assays




***Slow, tedious, time-consuming***

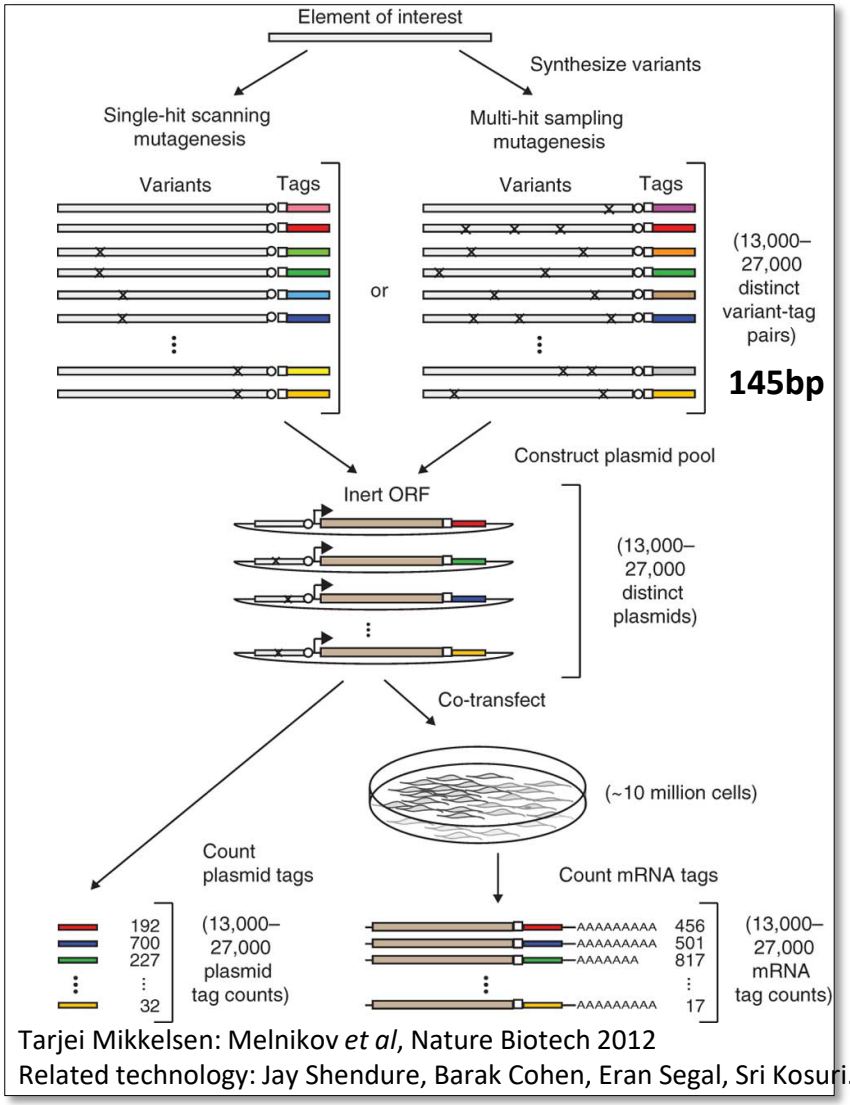
## Massively-parallel assays




***10,000+ elements at a time***

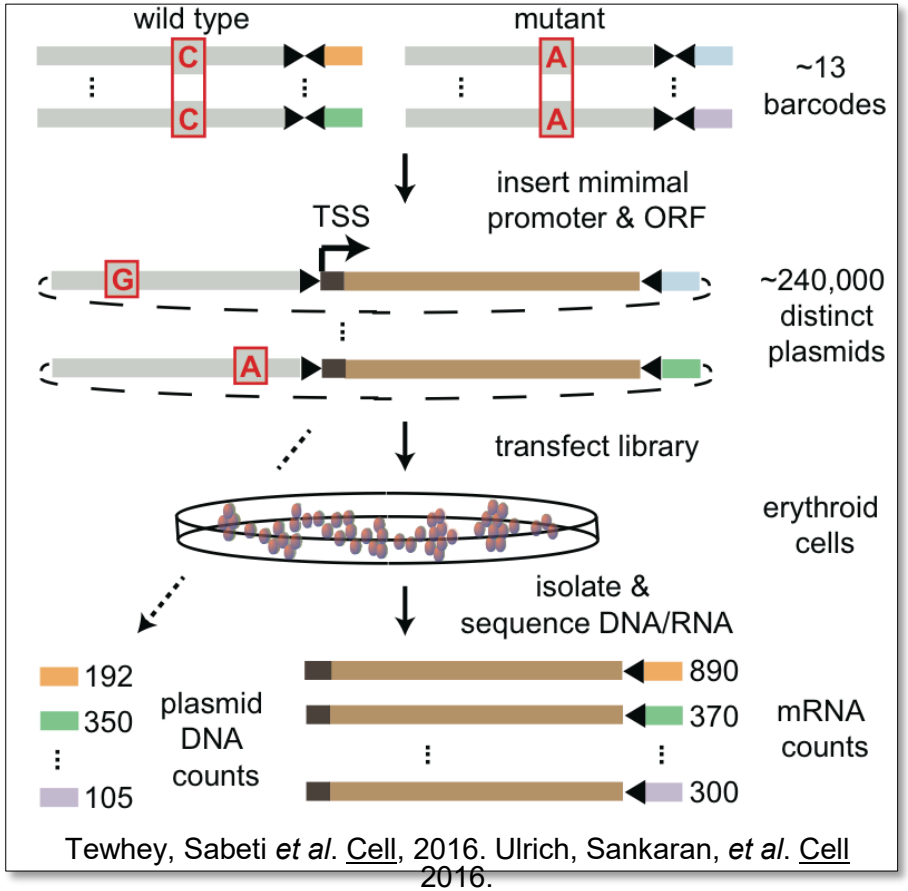
Difference between endogenous epigenomic signatures (e.g. H3K27ac)
vs. being able to actually drive expression of a reporter gene
(take DNA sequence segment out of context)

# Enabling Technology: Massively Parallel Reporter Assay (MPRA)




Tarjei Mikkelsen: Melnikov *et al*, Nature Biotech 2012
Related technology: Jay Shendure, Barak Cohen, Eran Segal, Sri Kosuri…

- Synthesize many enhancer versions ➔ insert upstream
- Couple each with a barcode ➔ insert downstream
- Make 10,000s of elements ➔ plasmids, transfection
- High-throughput test in diff. cell types ➔ 10k measurements

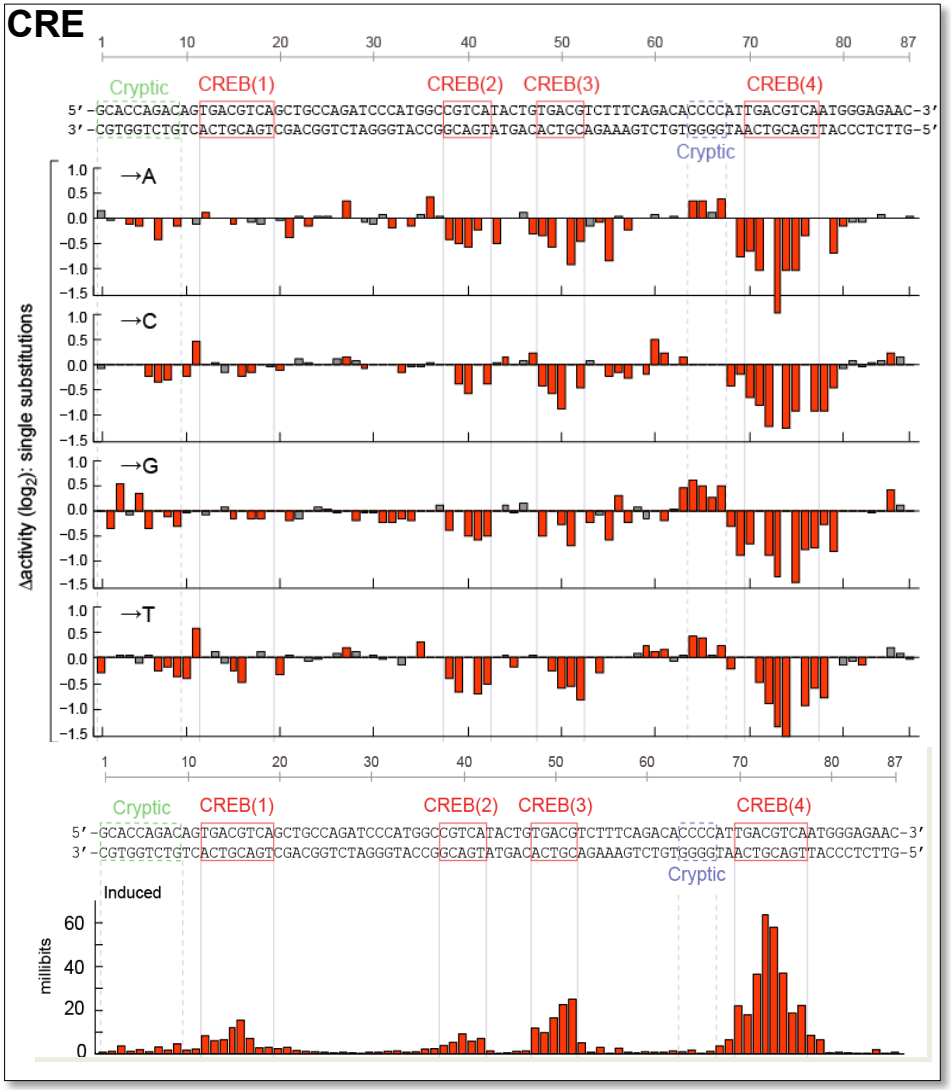


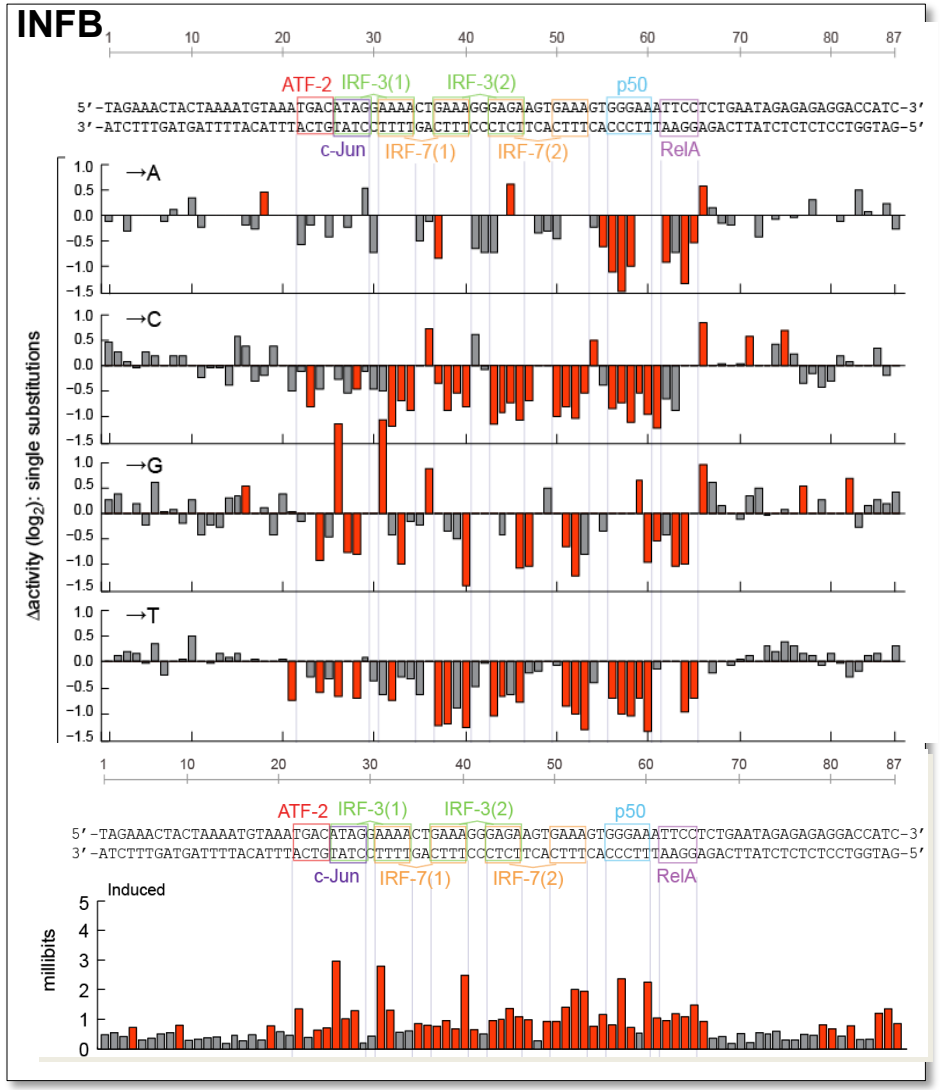

Tewhey, Sabeti *et al*. Cell, 2016. Ulrich, Sankaran, *et al*. Cell 2016.

***Application: Test 10,000 variants in 1 experiment***

Can we achieve (1) **large scale application** (2) **nucleotide level resolution**, and (3) **direction of effect**, all without knowing motifs or precise 145bp to test?

# High-resolution tiling dissection of individual regulatory regions

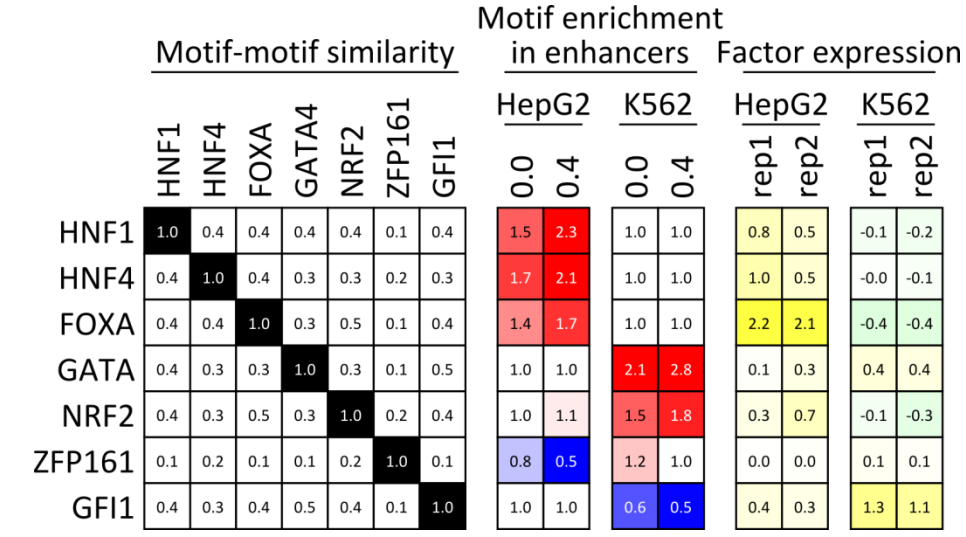


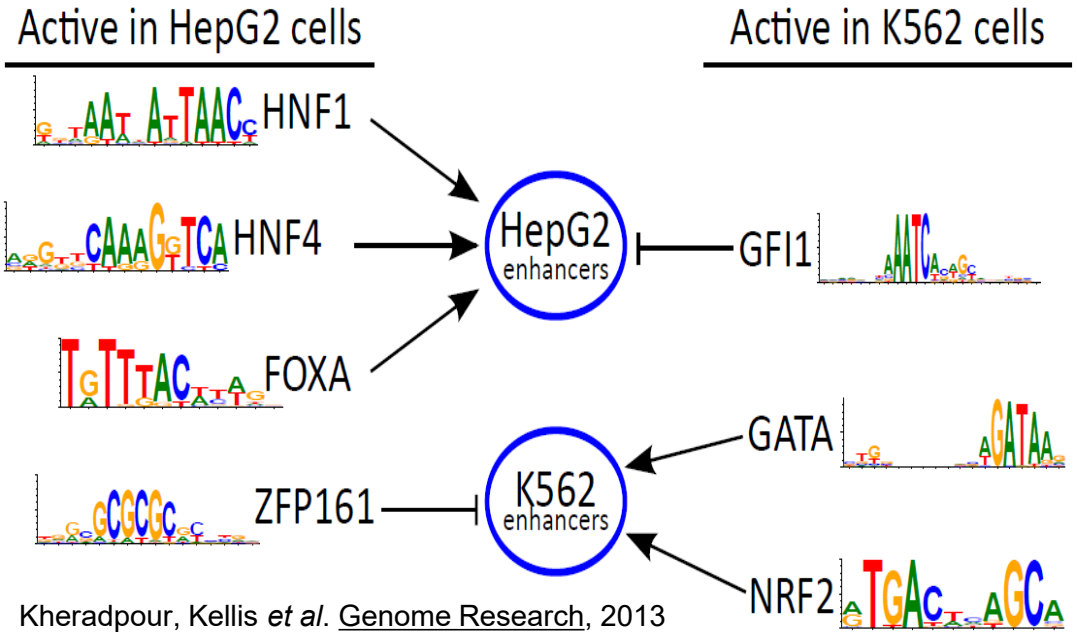



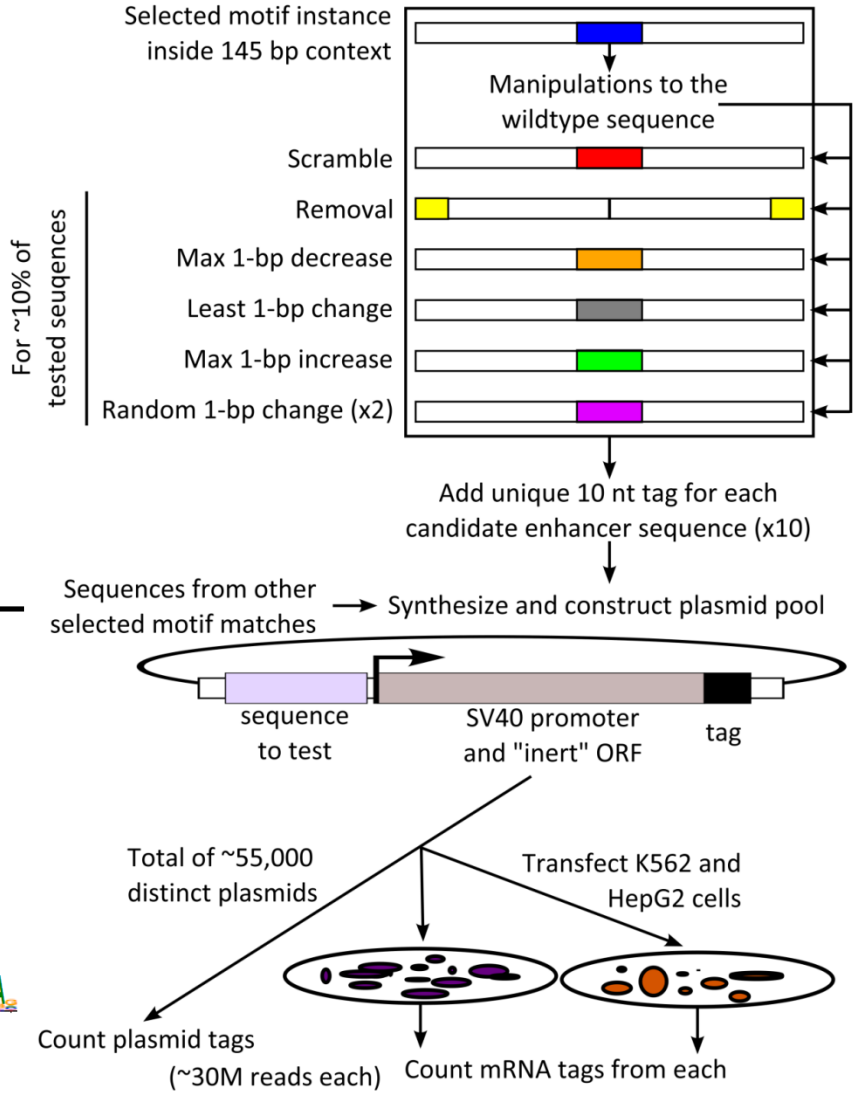

**Challenge:** hundreds of constructs needed for each region
Can test thousands of regions jointly?

# Systematic motif disruption in 2000 regions

## for 5 activators and 2 repressors in 2 human cell lines



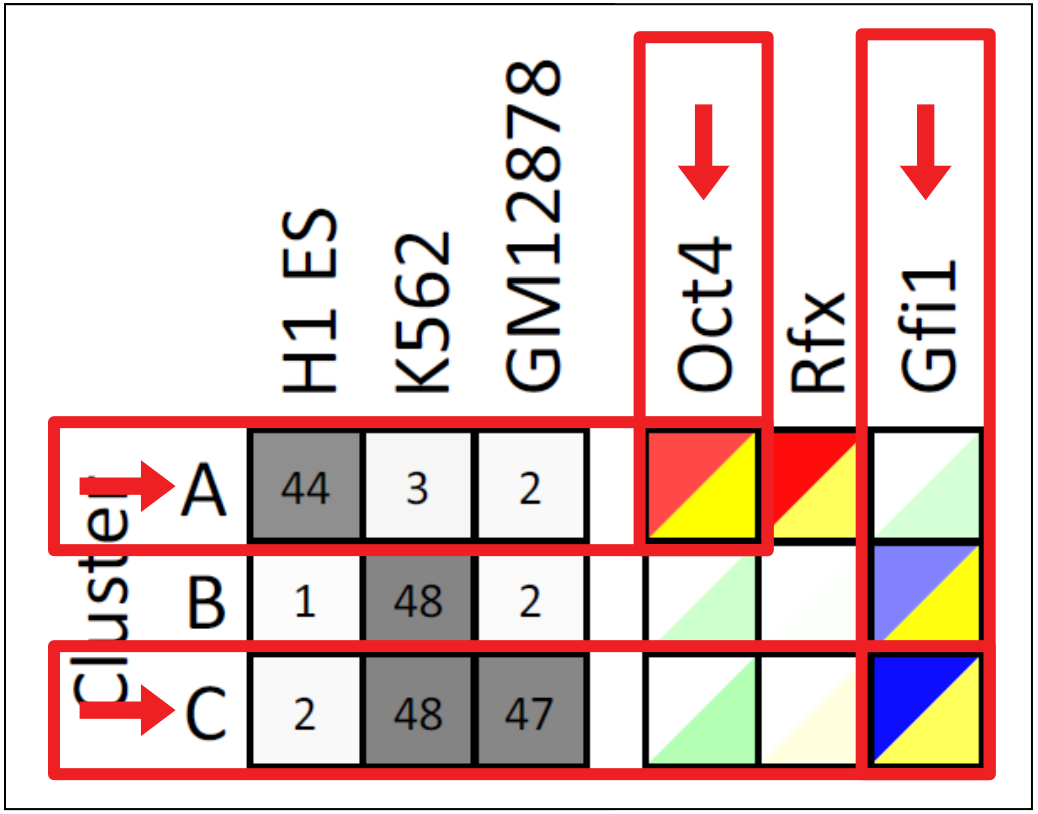



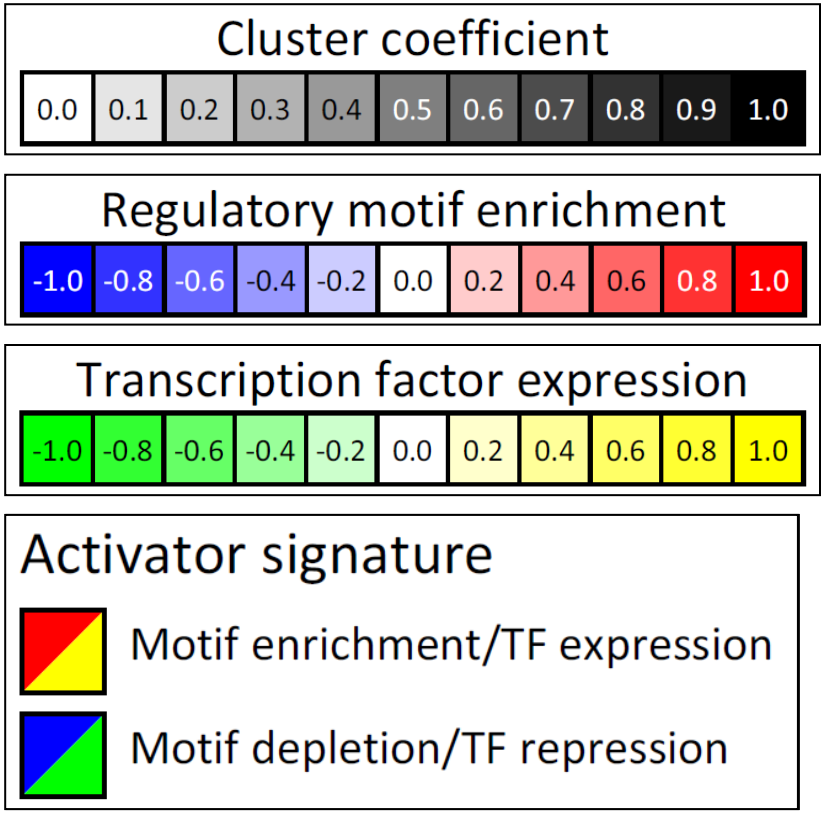




**54000+ measurements (x2 cells, 2x repl)**

Kheradpour, Kellis *et al*. Genome Research, 2013

# What to perturb: Guided by computational predictions

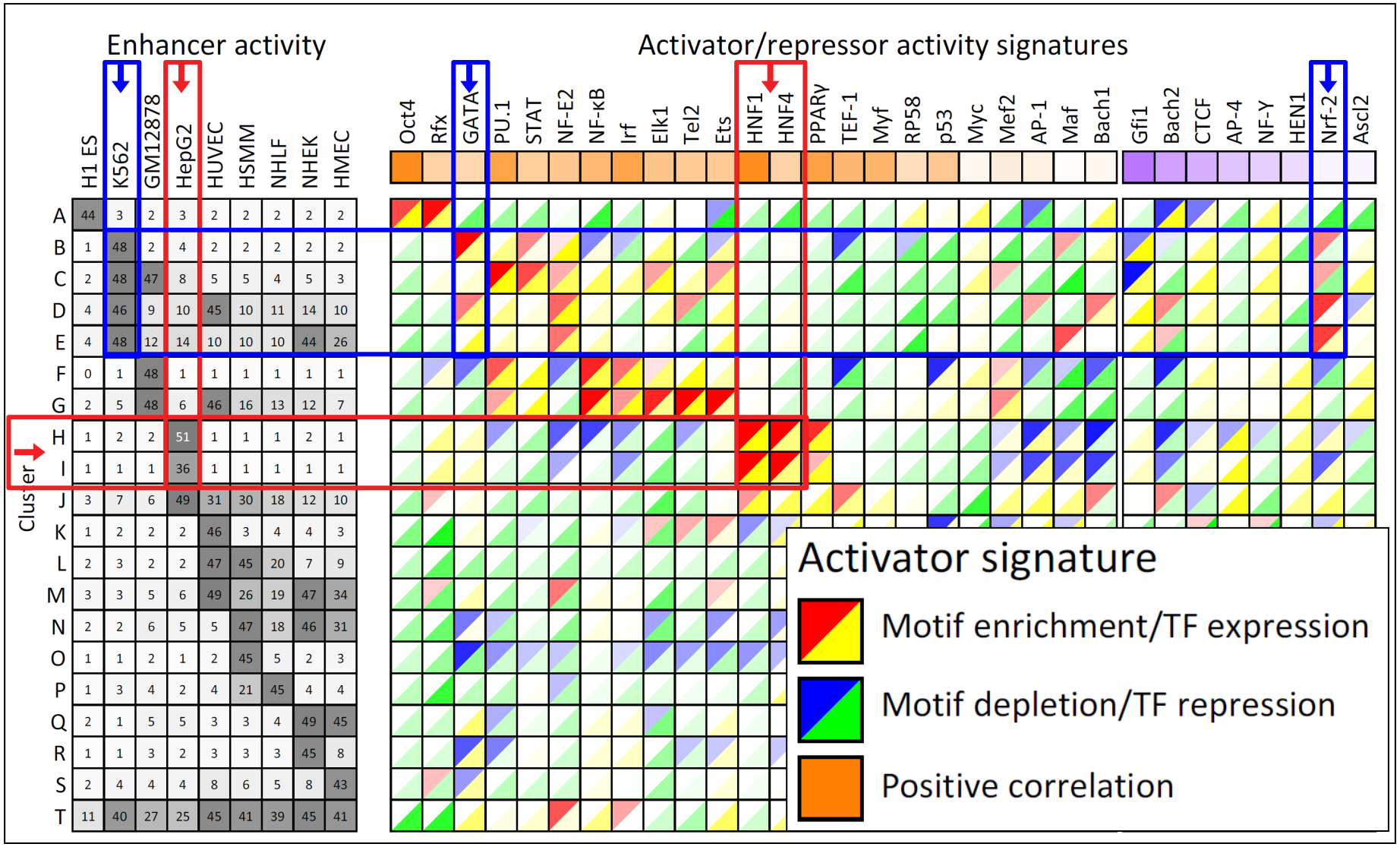



- Chromatin mark-based cell line specific enhancers
- Oct4 predicted activator of embryonic stem cells
- Gfi1 predicted repressor K562/GM12878 cells

Coordinated activity reveals activators/repressors

# HNF1 and HNF4 are predicted activators of HepG2 enhancers




- Model: Disruption of the motif site would abolish enhancer state

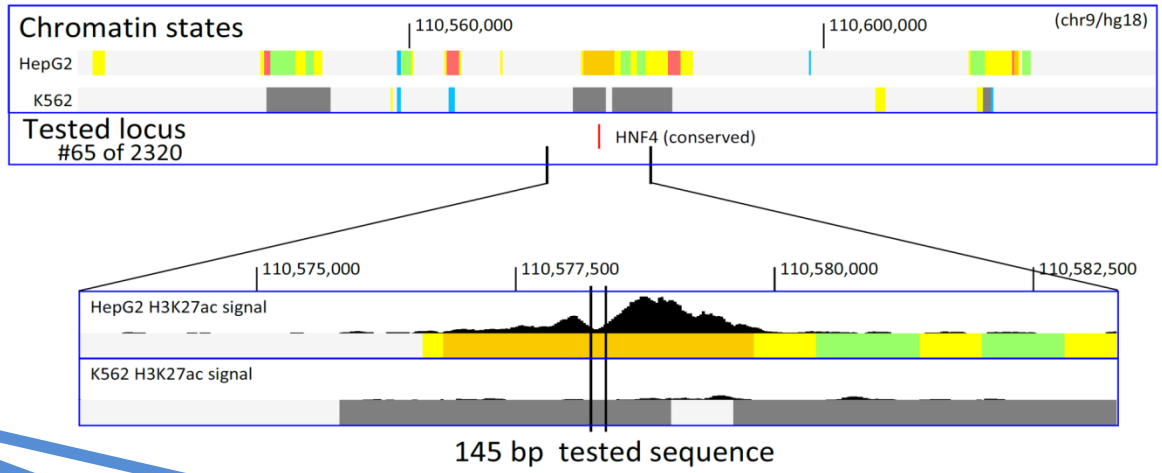
# Example activator: conserved HNF4 motif match

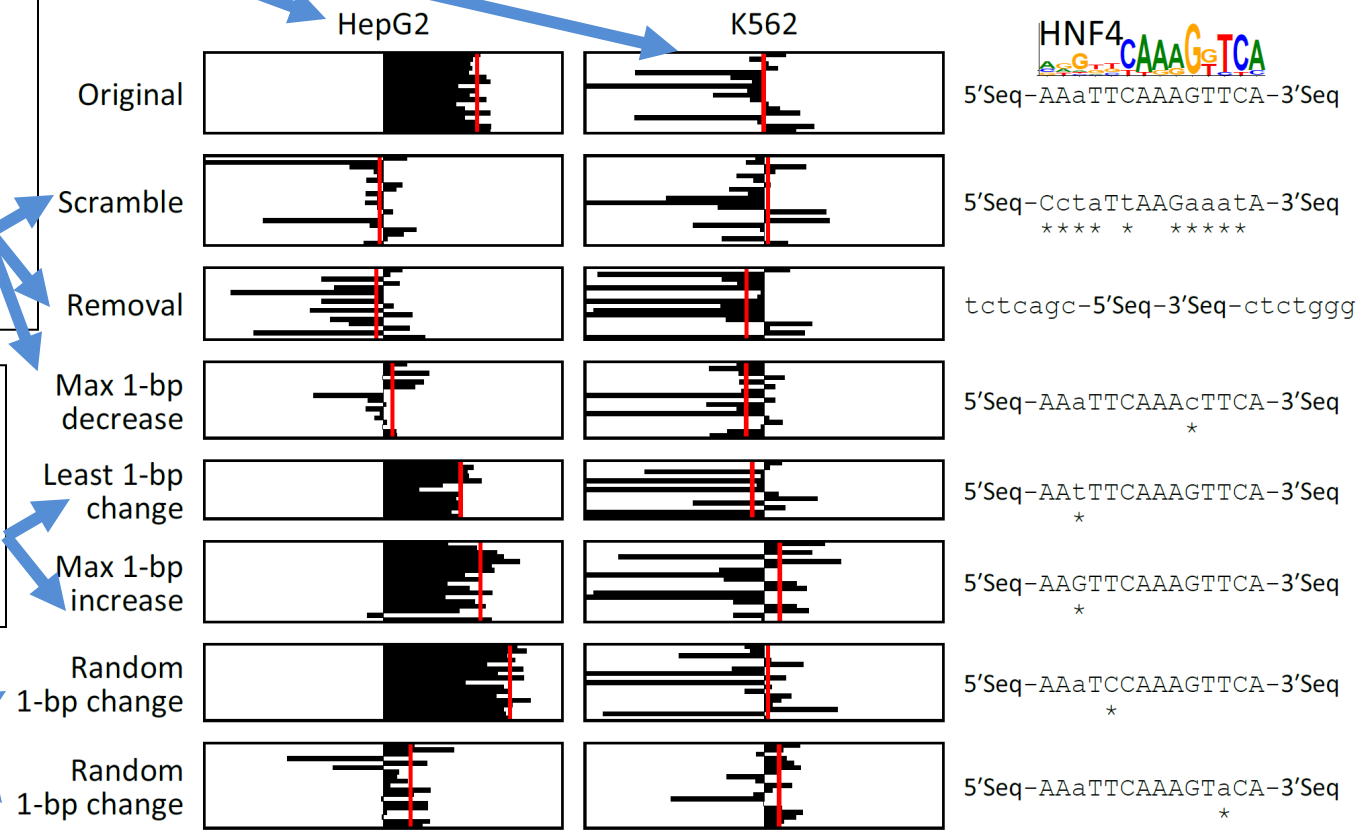
WT expression specific to HepG2

Motif match disruptions reduce expression to background

Non-disruptive changes maintain expression

Random changes depend on effect to motif match

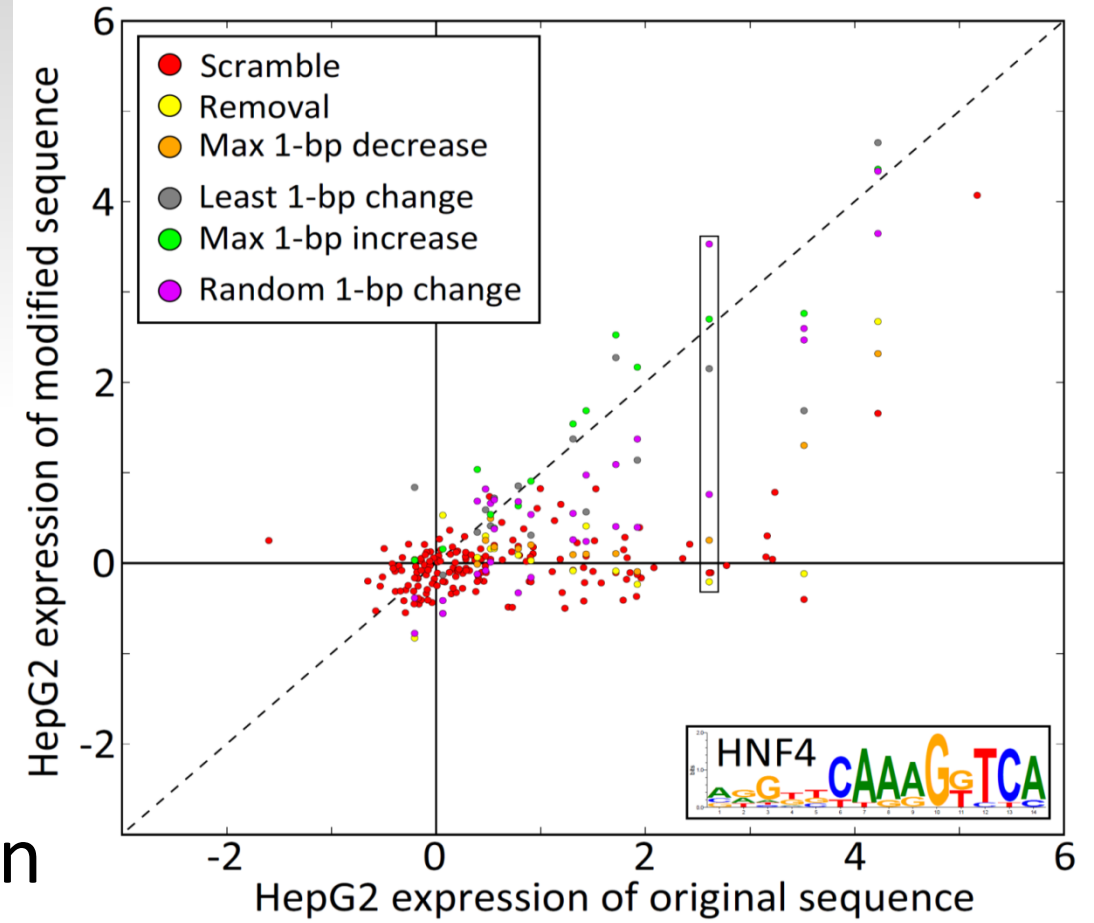


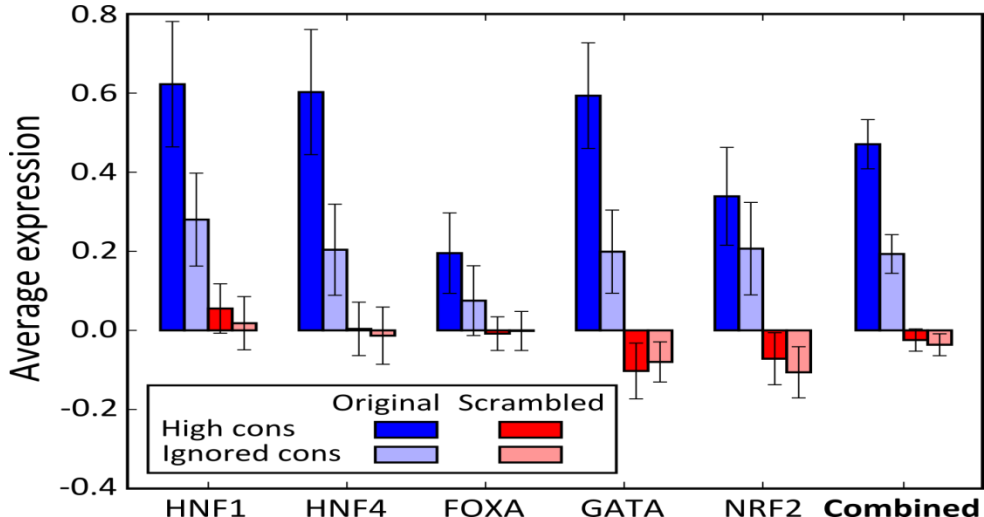



# Results hold across 2000+ enhancers

- Scramble abolishes reporter expression
- Neutral mutations show no change
- Increasing mutations show more expression

- Repressor mutations ➔ expression increase
- Motif context matters

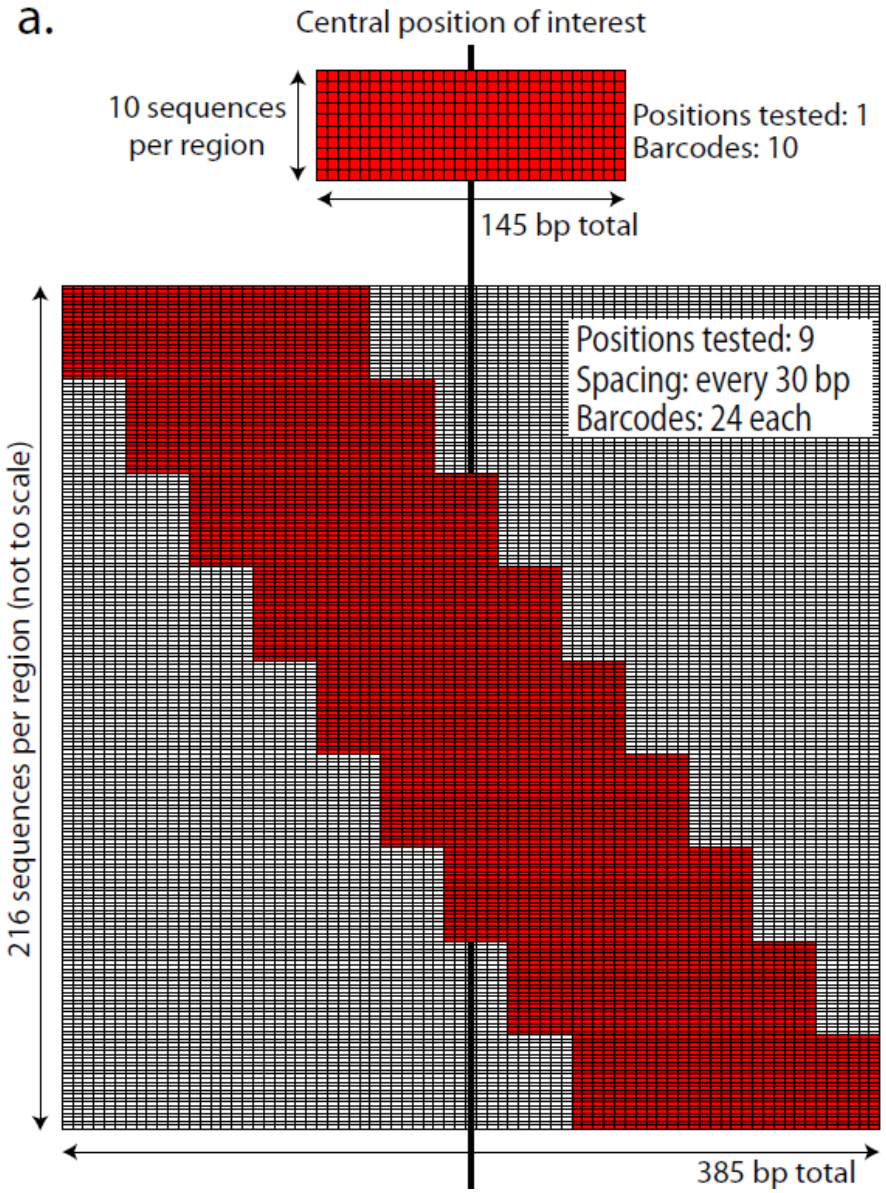


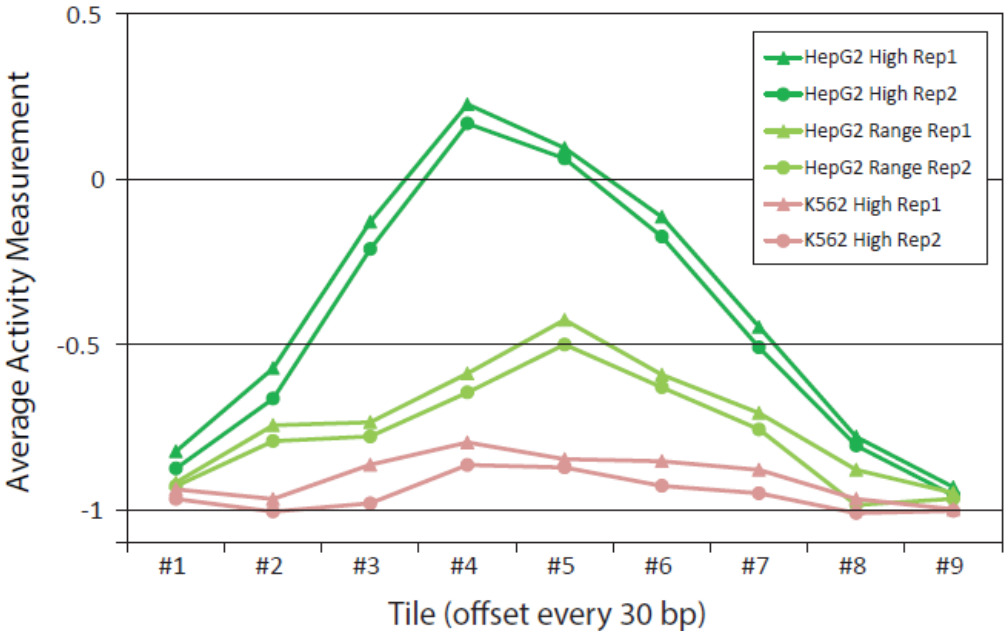
# Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint ($M$,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

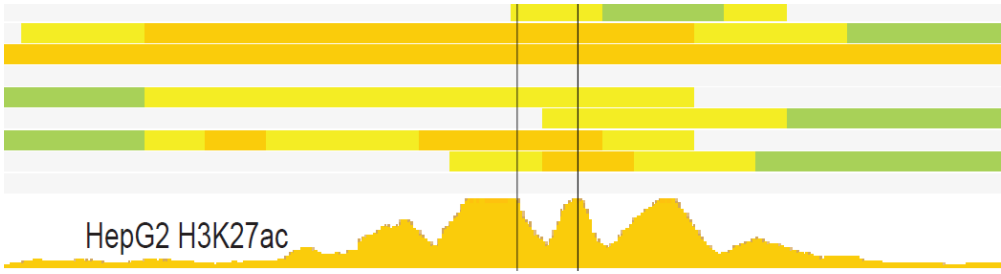
# Effect of enhancer position on reporter activity




***Tile regions at 30-bp offsets***

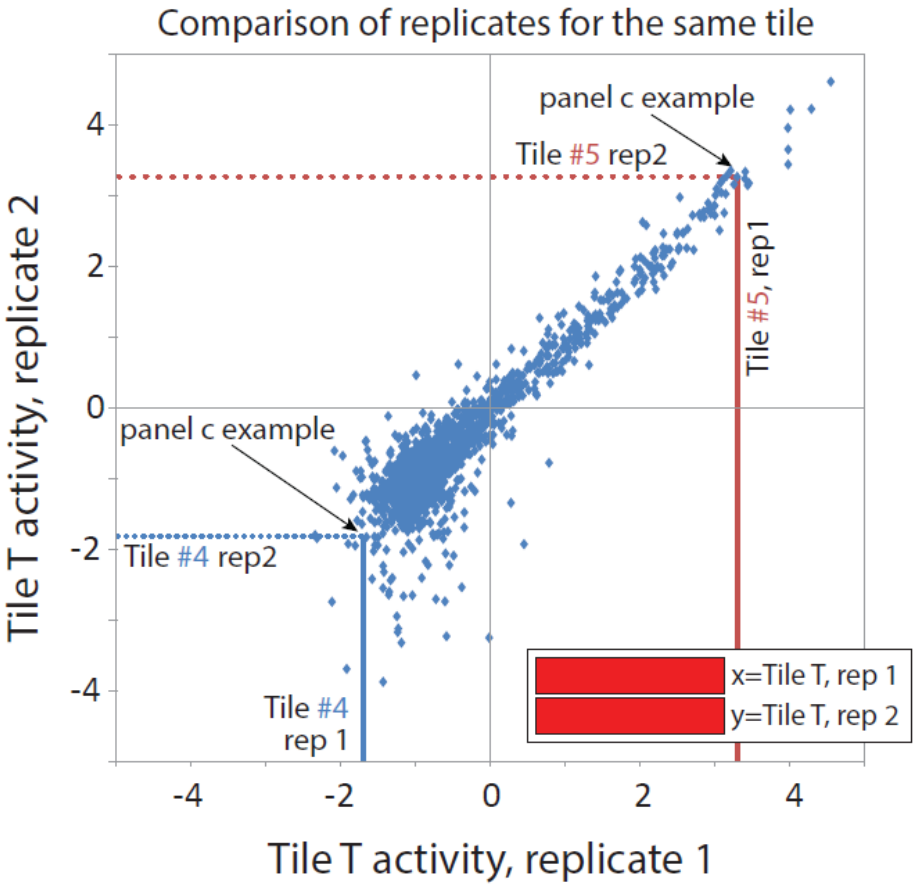


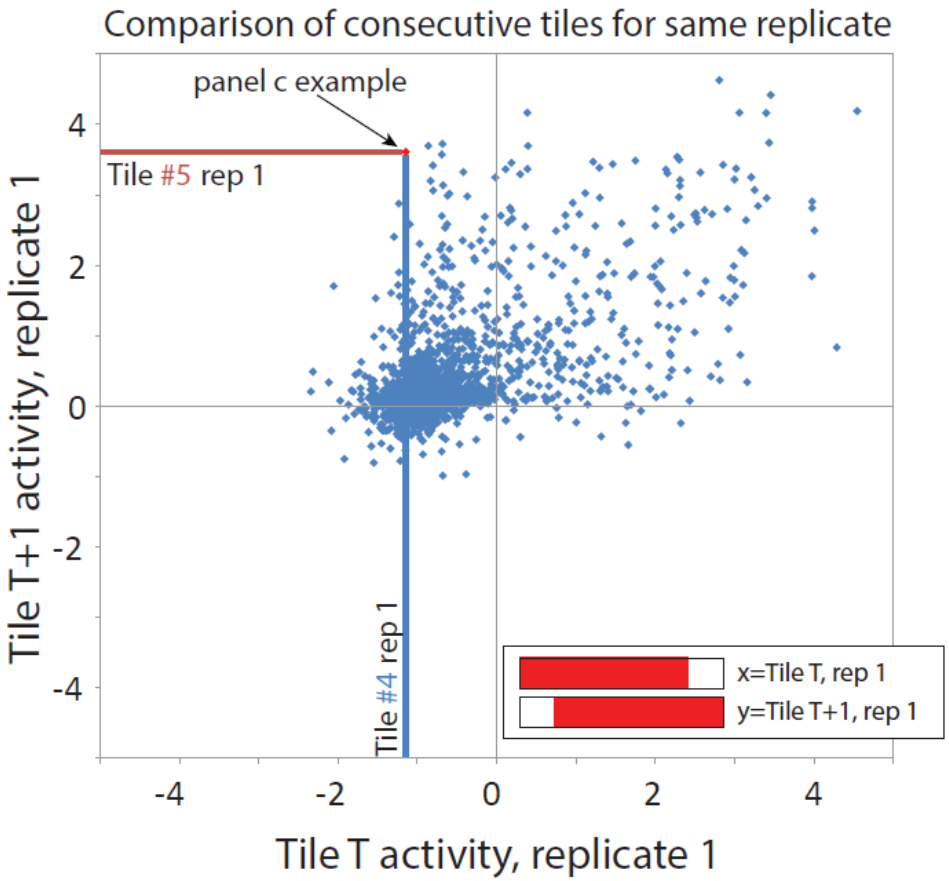

***Centers of selected regions show strongest activity***




***Chromatin dips in matched cell show strongest activity***

# An offset of 30-bp can make a big difference

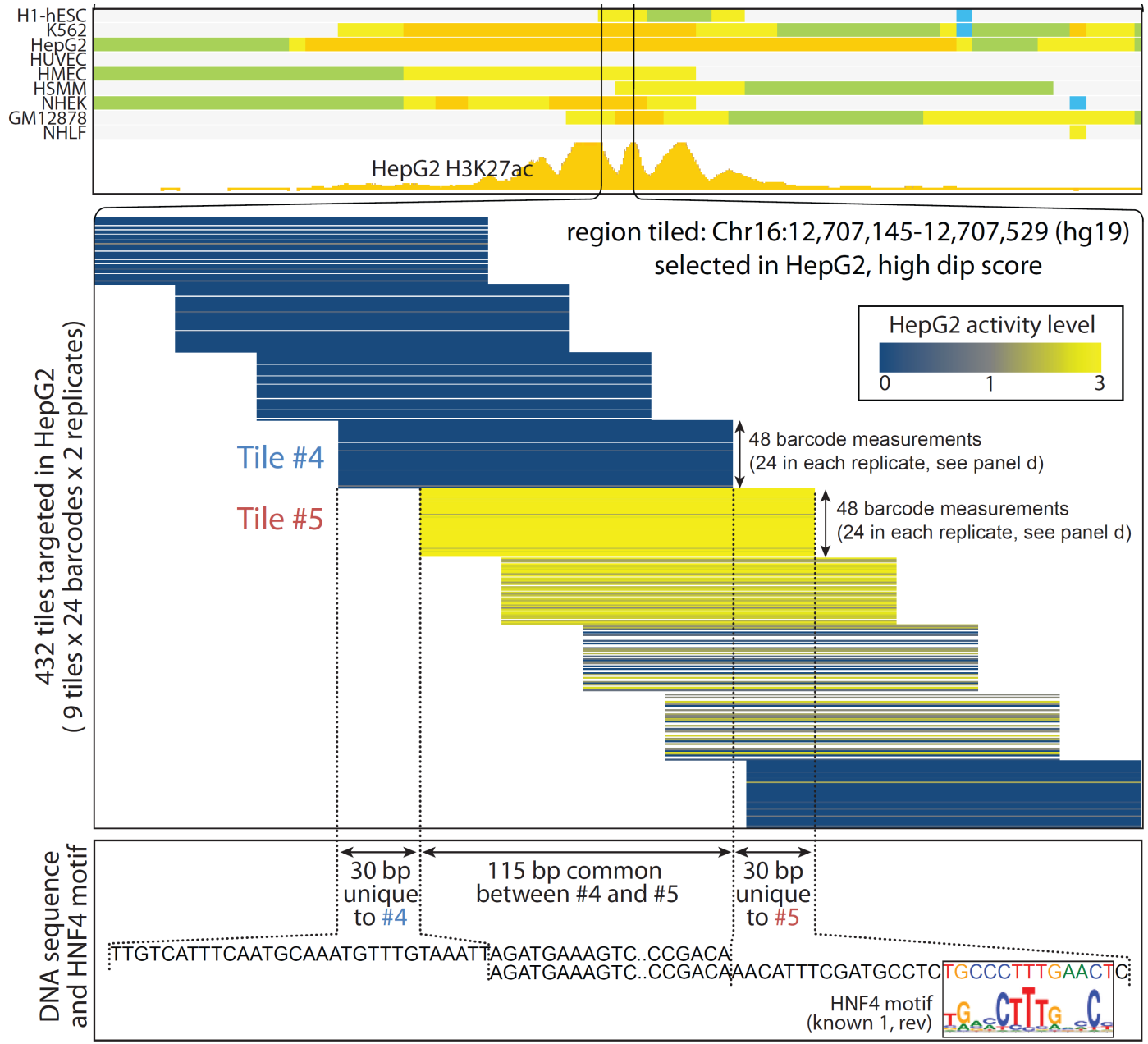



***Replicates of same tile are highly consistent***

***Consecutive tiles can differ greatly***

# Consecutive tile diffs due to motif inclusion/exclusion



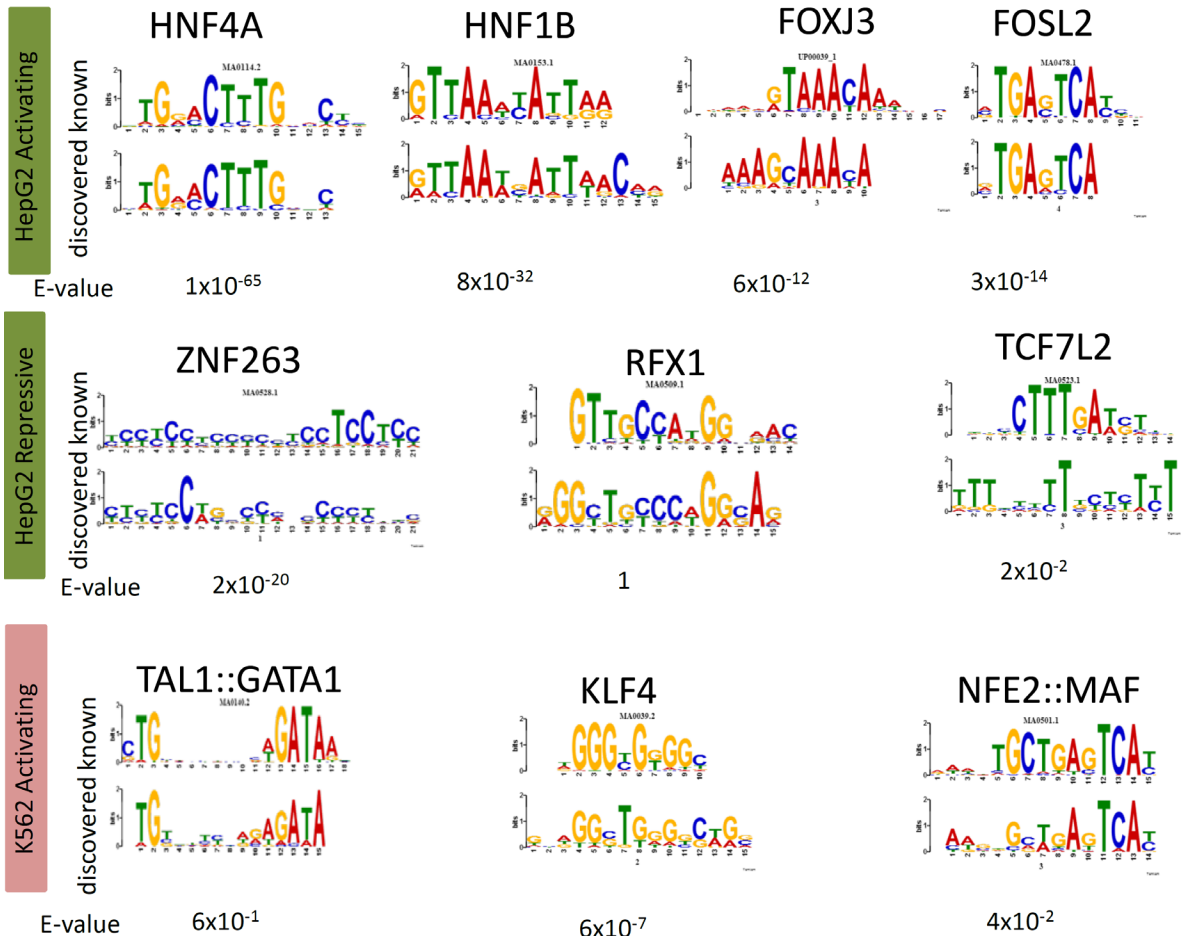

- **Inclusion/ exclusion of 30-bp intervals**
  - ➢ **Akin to systematic disruption**
  - ➢ **Increase resolution from tile (145bp) to offset (30bp)**

**Applications:**

- **Use to discover motifs?**
- **Further increase resolution?**

# Tile differences allow motif discovery

| | HepG2 Activating | HepG2 Repressing | K562 Activating | K562 Repressing |
|---|---|---|---|---|
| GATA_known14 | 1.0 | 1.0 | 6.4 | 1.0 |
| LMO2_2 | 1.0 | 1.0 | 4.4 | 1.0 |
| TAL1_disc1 | 1.0 | 1.0 | 3.0 | 1.0 |
| CPHX_1 | 1.0 | 1.0 | 2.7 | 1.0 |
| JDP2_2 | 1.0 | 1.0 | 2.4 | 1.0 |
| NFE2L2_3 | 1.0 | 1.0 | 2.0 | 1.0 |
| HNF4_known9 | 4.6 | 0.3 | 1.0 | 1.0 |
| NR2F6_2 | 3.7 | 1.0 | 1.0 | 1.0 |
| HNF1B_4 | 3.5 | 0.6 | 1.0 | 1.0 |
| RXRA_known10 | 3.4 | 0.9 | 1.0 | 1.0 |
| PPARA_4 | 3.1 | 1.0 | 1.0 | 1.0 |
| HNF1A_4 | 2.8 | 1.0 | 1.0 | 1.0 |
| HNF1_4 | 2.6 | 1.0 | 1.0 | 1.0 |
| TCF7L2_disc1 | 2.4 | 1.0 | 1.0 | 1.0 |
| AP1_known4 | 2.3 | 1.0 | 1.6 | 1.0 |
| SMARC_disc1 | 2.2 | 1.0 | 1.1 | 1.0 |
| CEBPB_known1 | 2.1 | 1.0 | 1.0 | 1.0 |
| TLX2_2 | 2.0 | 1.0 | 1.0 | 1.0 |
| FOXJ2_4 | 2.0 | 1.0 | 1.0 | 1.0 |
| EGR1_disc2 | 2.0 | 1.0 | 1.0 | 1.0 |
| RFX2_3 | 1.0 | 3.0 | 1.0 | 1.0 |
| RFX5_known9 | 1.0 | 2.9 | 1.0 | 1.0 |

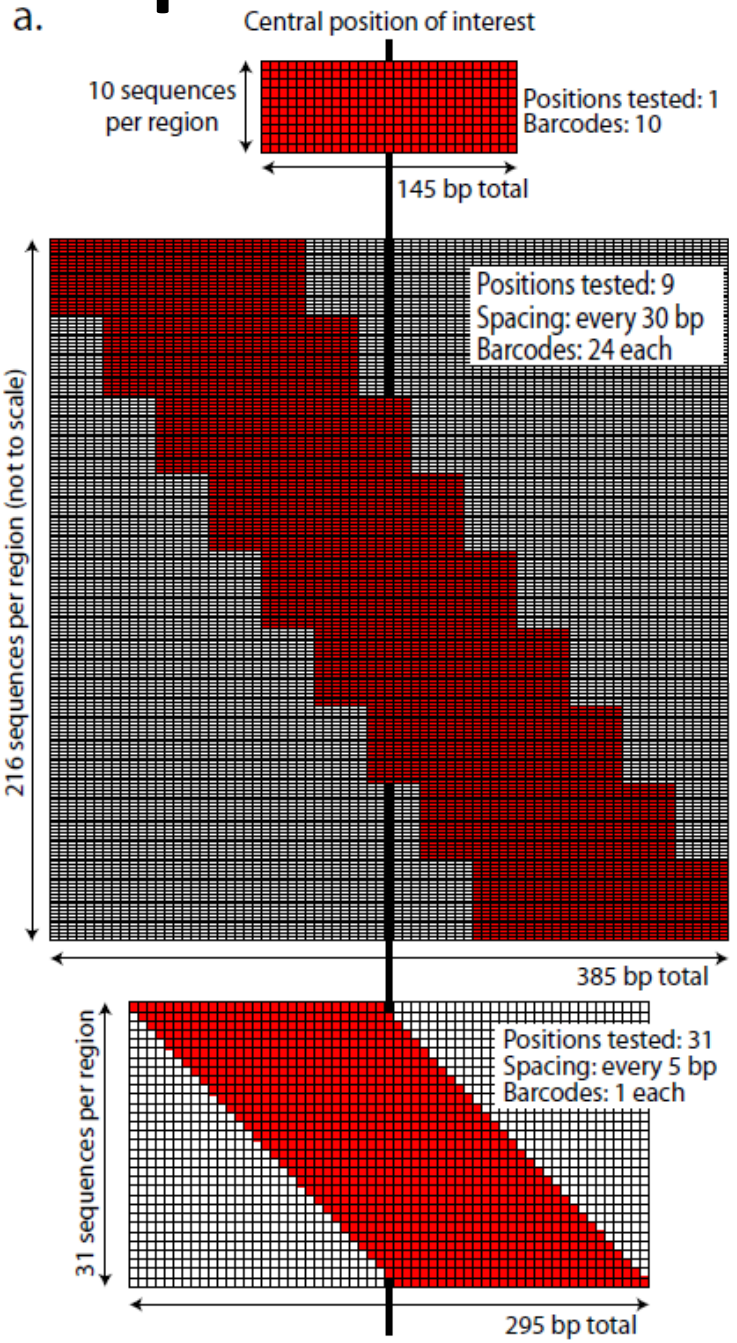


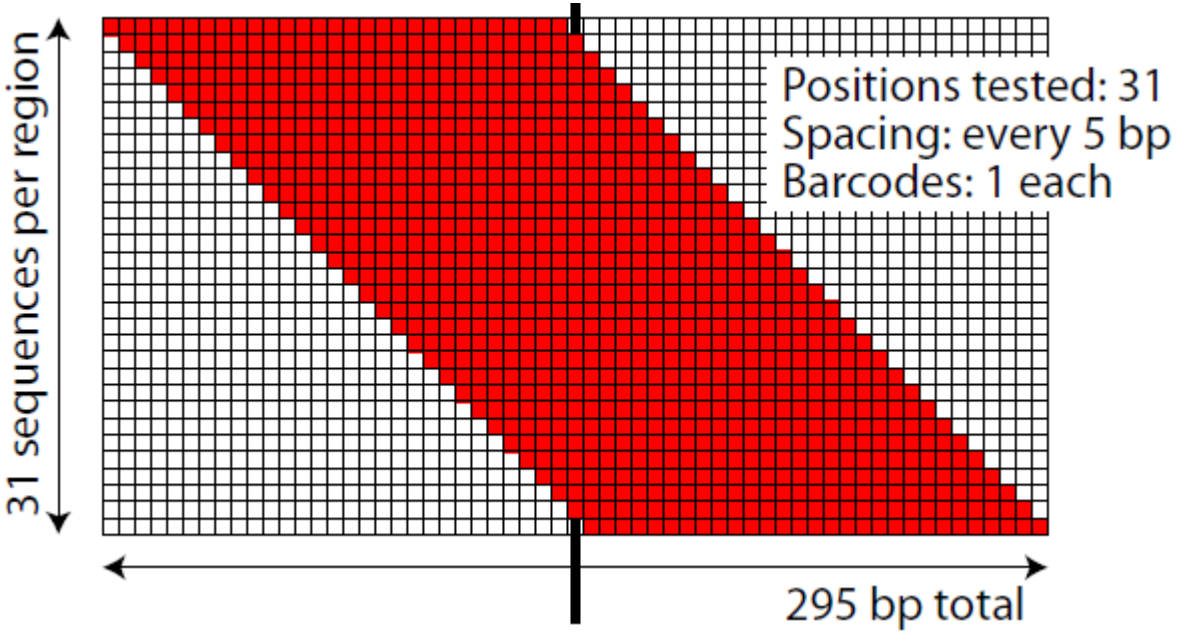

- Increased resolution allows testing of only 30-bp intervals
- *De novo* discovered motifs match known motifs
- Discovery distinguishes activating vs. repressive factors

# Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint (M,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

# Experimental design for high-resolution tiling

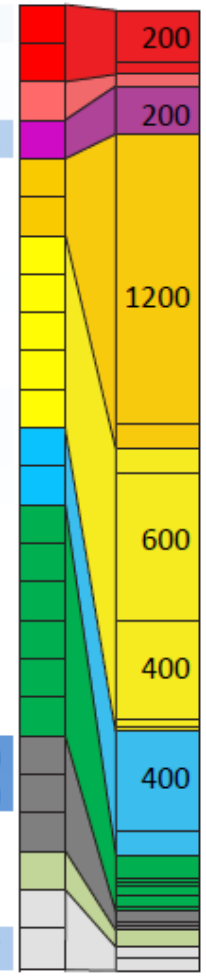



- ***Tile 295-bp regions (vs. 385)***
- ***Tile @ 5-bp offsets (vs @30)***
- ***Center on DNase peaks (vs dips)***
- ***Single barcode / offset (vs 24)***
- ***244K spot array (vs 54K)***
- ***25 chromatin states (vs. Enh)***
- ***Both minP and SV40 promoters***

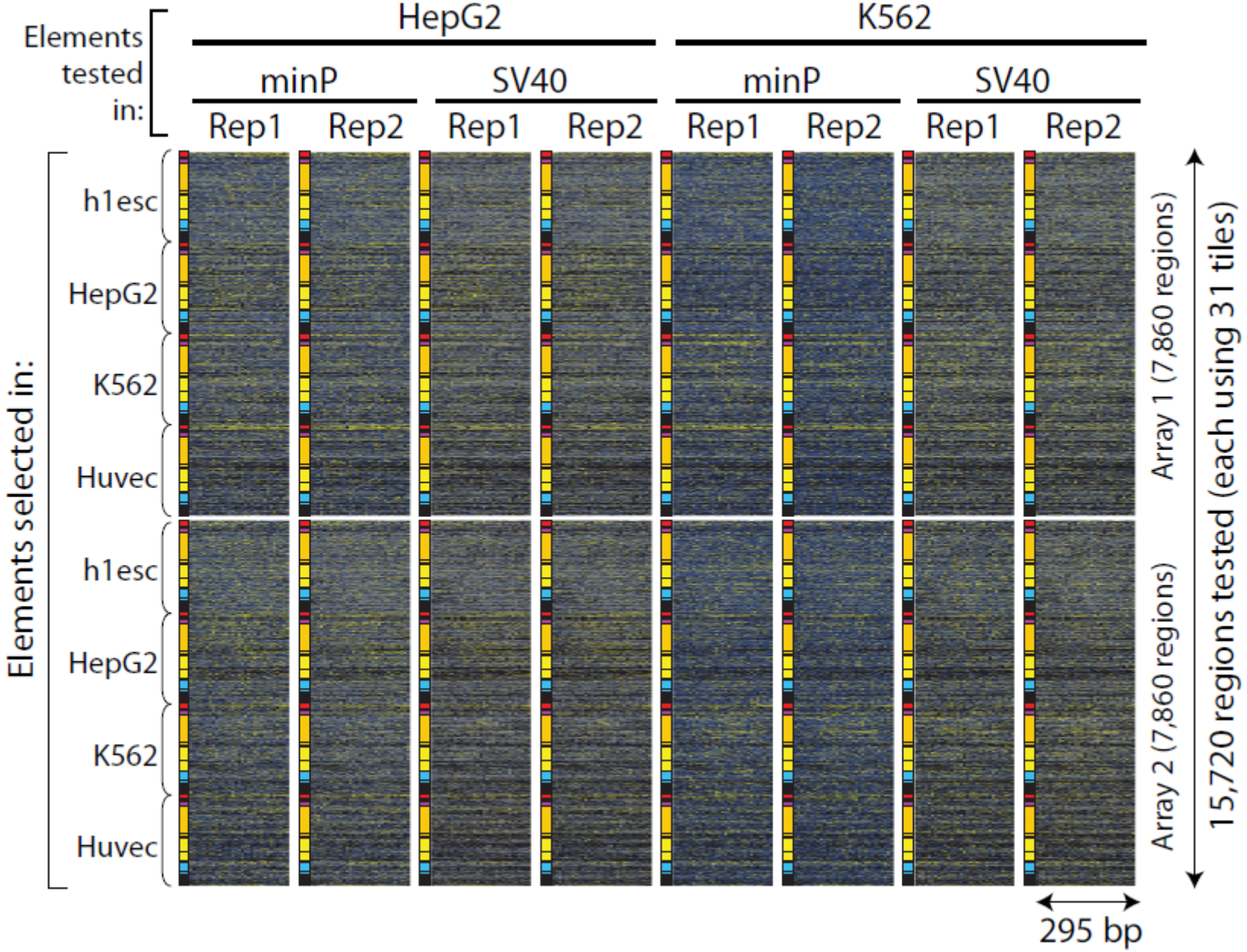
# Chromatin state vs. reporter activity of DNase elmts

| State | H3K36me3 | H4K20me1 | Control | FAIRE | POL2 | H3K27ac | H3K9ac | H3K4me3 | H3K4me2 | H3K4me1 | DNase UW | DNase DUKE | CTCF | H3K27me3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Tss | 2 | 6 | 3 | 33 | 98 | 96 | 99 | 100 | 96 | 43 | 86 | 92 | 36 | 5 |
| 2 TssF | 21 | 24 | .7 | 4 | 55 | 72 | 86 | 98 | 99 | 73 | 24 | 15 | 4 | 2 |
| 3 PromF | 3 | 6 | .2 | .6 | 7 | 2 | 8 | 32 | 89 | 44 | 5 | 4 | 2 | 2 |
| 4 PromP | .1 | 7 | .5 | 8 | 38 | 2 | 37 | 88 | 97 | 61 | 51 | 87 | 32 | 42 |
| 5 Enh | 11 | 7 | 4 | 49 | 69 | 98 | 73 | 34 | 97 | 98 | 89 | 67 | 17 | .1 |
| 6 EnhF | 5 | 2 | .6 | 4 | 11 | 94 | 29 | 1 | 54 | 93 | 20 | 8 | 2 | .1 |
| 7 EnhWF | .9 | 2 | .2 | .8 | 1 | .6 | 1 | 0 | 7 | 88 | 4 | 4 | 1 | .4 |
| 8 EnhW | .2 | 8 | .7 | 17 | 19 | 11 | 6 | .6 | 62 | 94 | 63 | 60 | 17 | .7 |
| 9 DNaseU | .4 | 1 | .1 | 3 | 3 | 2 | .5 | .1 | 3 | 9 | 95 | 14 | .3 | .3 |
| 10 DNaseD | 1 | 3 | .2 | .3 | 2 | .5 | .7 | .1 | .3 | 3 | 2 | 69 | .9 | .3 |
| 11 FaireW | 1 | 2 | 8 | 22 | .8 | 10 | 6 | 2 | .2 | .3 | .5 | 1 | .6 | 2 |
| 12 CtcfO | 6 | 9 | 1 | 35 | 19 | 2 | 1 | .3 | 10 | 17 | 71 | 87 | 100 | 2 |
| 13 Ctcf | .3 | 2 | .1 | .5 | .3 | .1 | .2 | 0 | .7 | 1 | 6 | 16 | 99 | 2 |
| 14 Gen5' | 70 | 36 | .7 | 3 | 52 | 26 | 9 | 1 | 23 | 84 | 9 | 12 | 4 | 0 |
| 15 Elon | 93 | 4 | .3 | .7 | 1 | .3 | .2 | 0 | .1 | .3 | .4 | 1 | 1 | 0 |
| 16 ElonW |  | .4 |  | .3 | .8 | .1 | 0 | 0 |  | .1 | .1 | .2 | 0 | 0 |
| 17 Gen3' | 93 | 18 | .4 | 1 | 83 | 1 | .6 | .1 | .1 | .7 | 2 | 7 | 3 | 0 |
| 18 Pol2 | .7 | 6 | .1 | .7 | 83 | .9 | .6 | 0 | .1 | 2 | 3 | 7 | 1 | .3 |
| 19 H4K20 | 3 | 86 | .3 | 1 | 2 | .2 | .3 | 0 | .2 | 2 | .4 | 1 | .5 | .6 |
| 20 ReprD | .5 | 15 | .5 | 1 | 2 | .5 | .6 | 1 | 11 | 21 | 4 | 48 | 12 | 90 |
| 21 Repr | 0 | 2 | .2 | .1 | .1 | .1 | 0 | 0 | 0 | 0 | 0 | 1 | .1 | 90 |
| 22 ReprW | 0 | .2 | 0 | .1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | .4 | 0 | 0 |
| 23 Low | .1 | 0 | 0 | .2 | .1 | 0 | 0 | 0 | 0 | 0 | 0 | .2 | 0 | 0 |
| 24 Quies | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 Art | 77 | 73 | 41 | 70 | 59 | 35 | 38 | 25 | 32 | 52 | 41 | 63 | 64 | 47 |



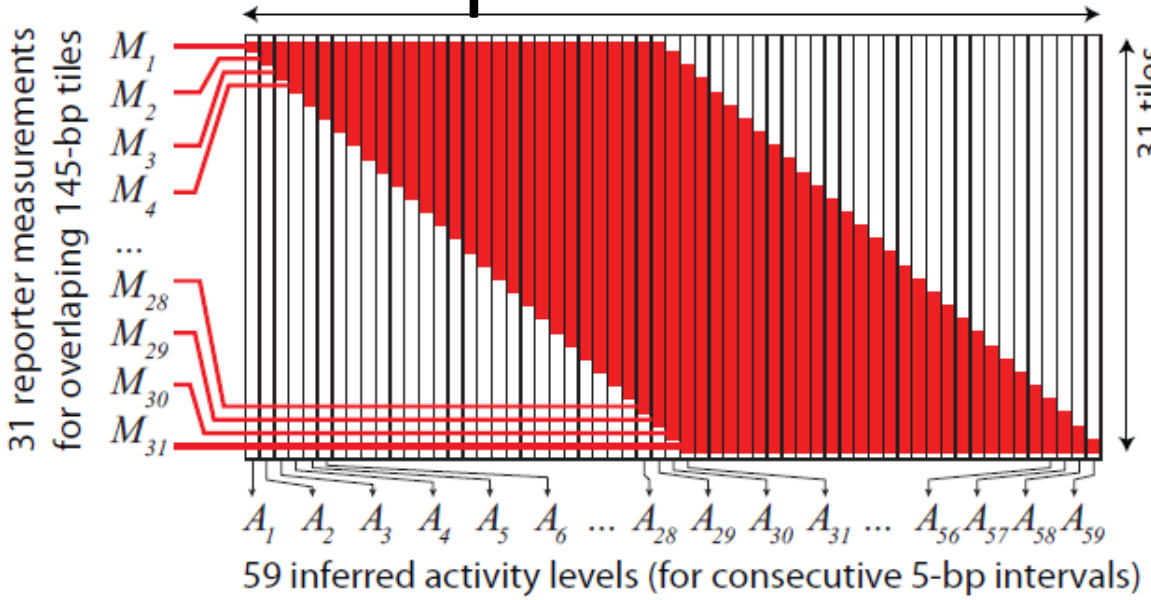

***Select 15,720 DNase elements across all 25 chromatin states***

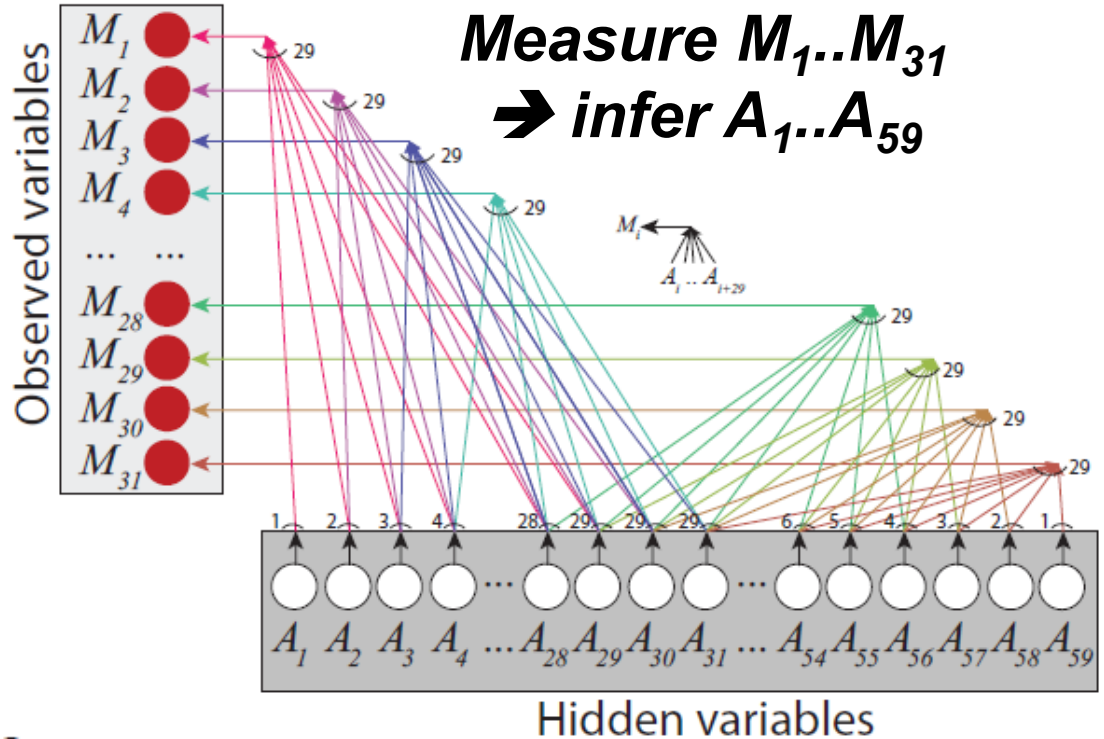
# Regions selected in 4 cell types, tiled in HepG2,K562




***15,720 regions x 31 offsets x 2 promoters x 2 reps x 2 cell lines***

# Computational inference model




***Measure $M_1..M_{31}$ ➔ infer $A_1..A_{59}$***

*Joint samples from multivariate normal*

$$\begin{bmatrix} A_{r,t} \\ M_{r,t} \end{bmatrix} \sim N(\mu_{x_{r,t}}, \Sigma_{x_{r,t}})$$

*True activity normally distributed*

$$A_{r,t,k} \sim N(\mu_{a_r}, \sigma_a^2) \quad k = 1, \ldots, K$$

*Observation is mean of true signals*

$$\mu_{a_r} = \frac{1}{|\widehat{M_r}|} \sum_{m \in \widehat{M_r}} m$$

*Measurements normally distributed*

$$M_{r,t,j} \sim N(\mu_{m_{r,t,j}}, \sigma_{m_r}^2)$$

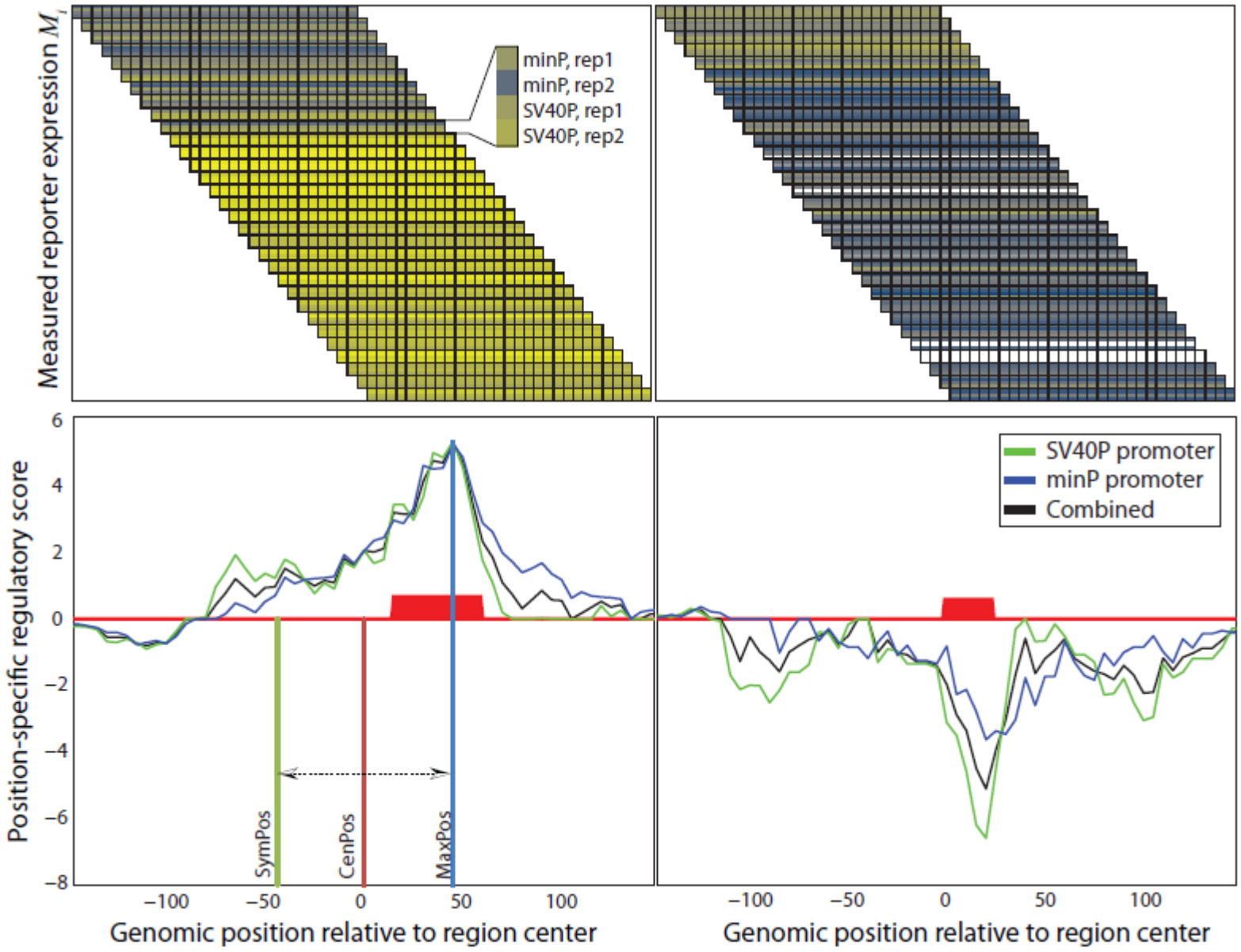
*Mean set to empirical mean*

$$\mu_{m_{r,t,j}} = \frac{1}{N} \sum_{i=0}^{N-1} A_{r,t,j+i}$$

*Variance set to empirical variance*

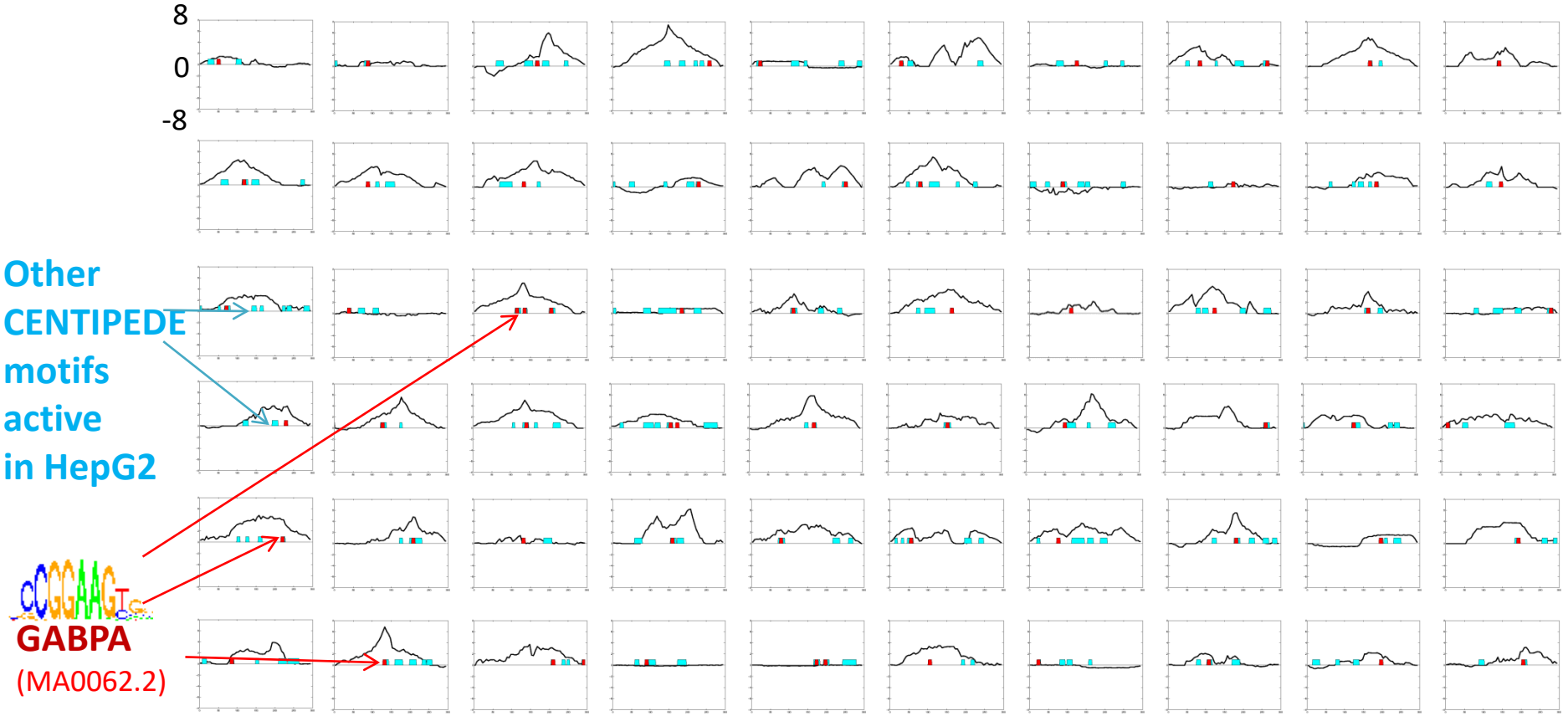
$$\sigma_{m_r}^2 = \frac{1}{|\widehat{M_r}|} \sum_{m \in \widehat{M_r}} (m - \frac{1}{|\widehat{M_r}|} \sum_{m \in \widehat{M_r}} m)^2$$

# Examples of tiling data deconvolution




***Detect activating/repressive elements at high resolution***

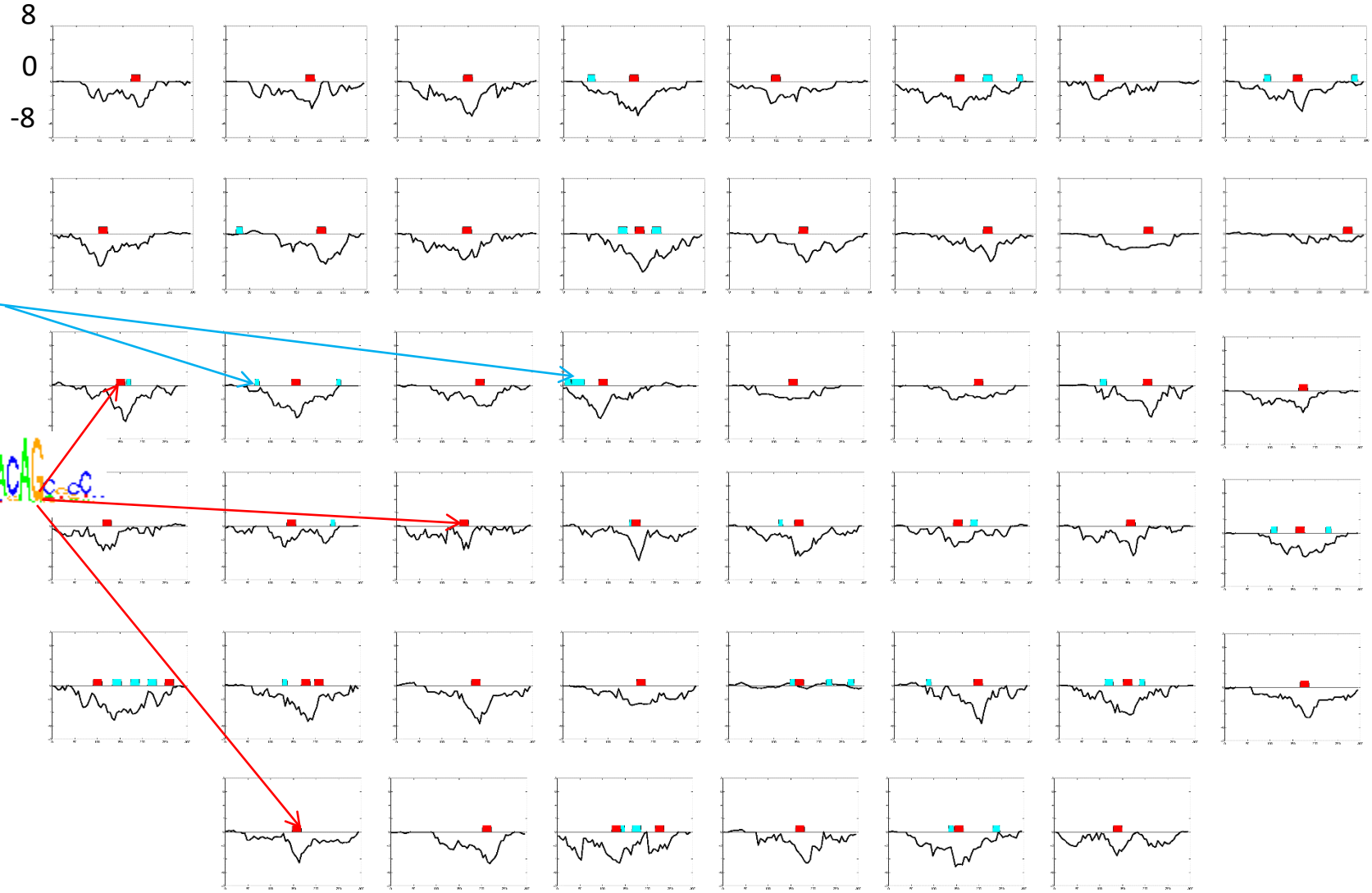
# Deconvolved regulatory signal vs. activator motif



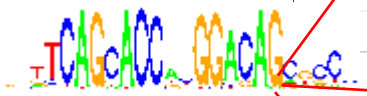

60 sites containing GABPA HepG2 motifs predicted by CENTIPEDE

CENTIPEDE predictions (Pique-Regi, et al, 2011)

# Deconvolved regulatory signal vs. repressor motif




Other CENTIPEDE motifs active in HepG2

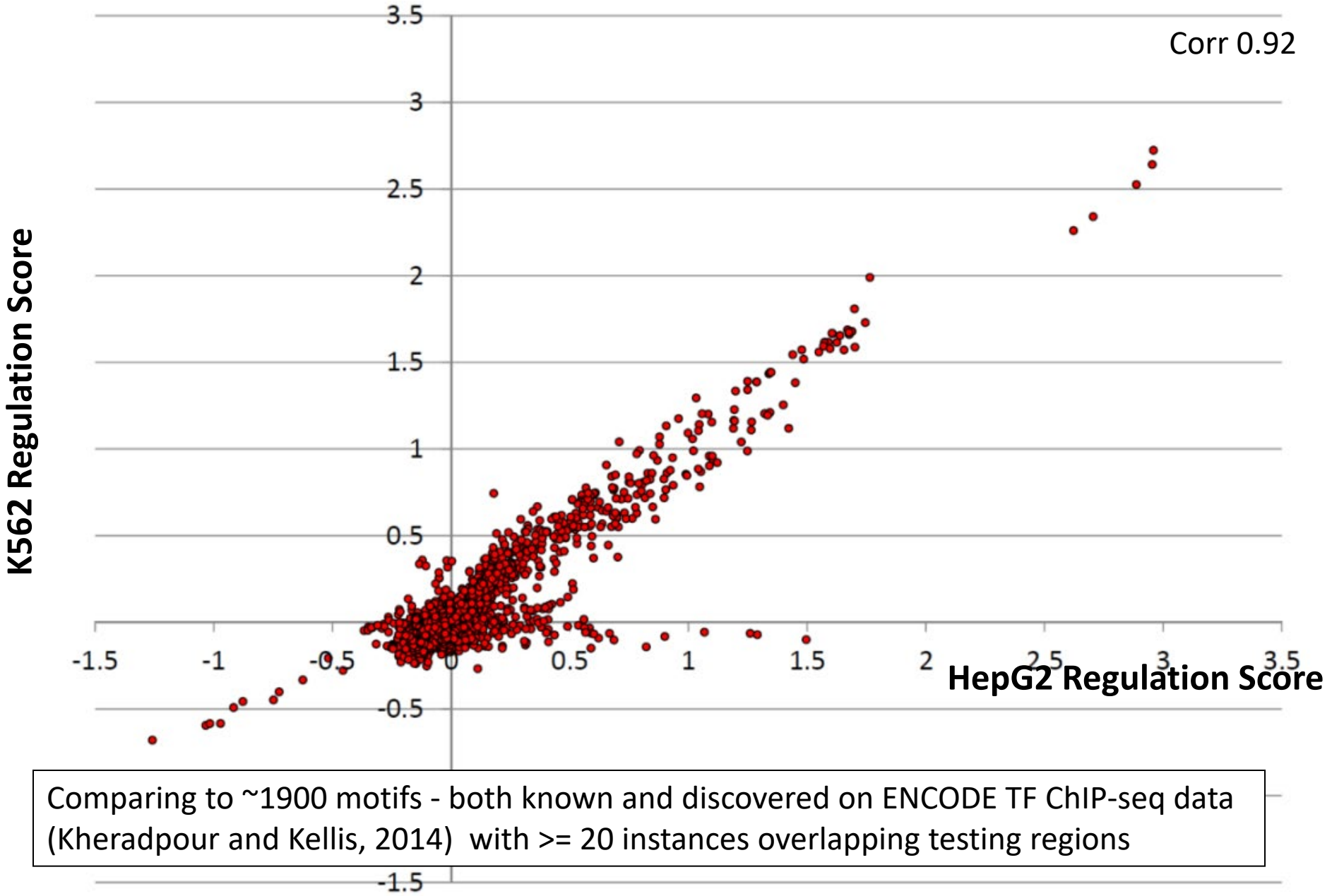


**NRSF**
(M00256)

46 sites containing NRSF HepG2 motifs predicted by CENTIPEDE

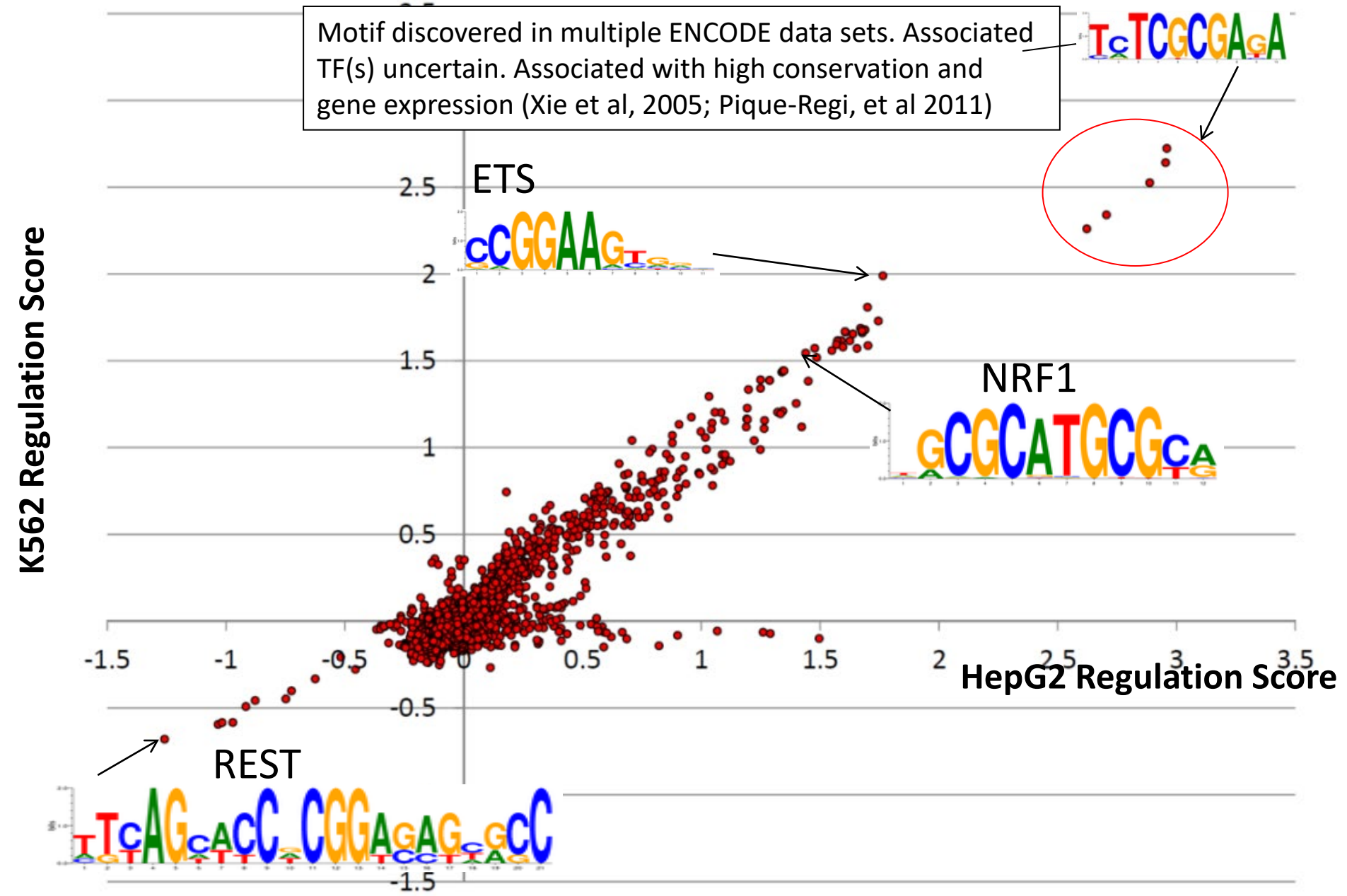
CENTIPEDE predictions (Pique-Regi, et al, 2011)

# Aggregate Motif Score Highly Correlated between K562 and HepG2

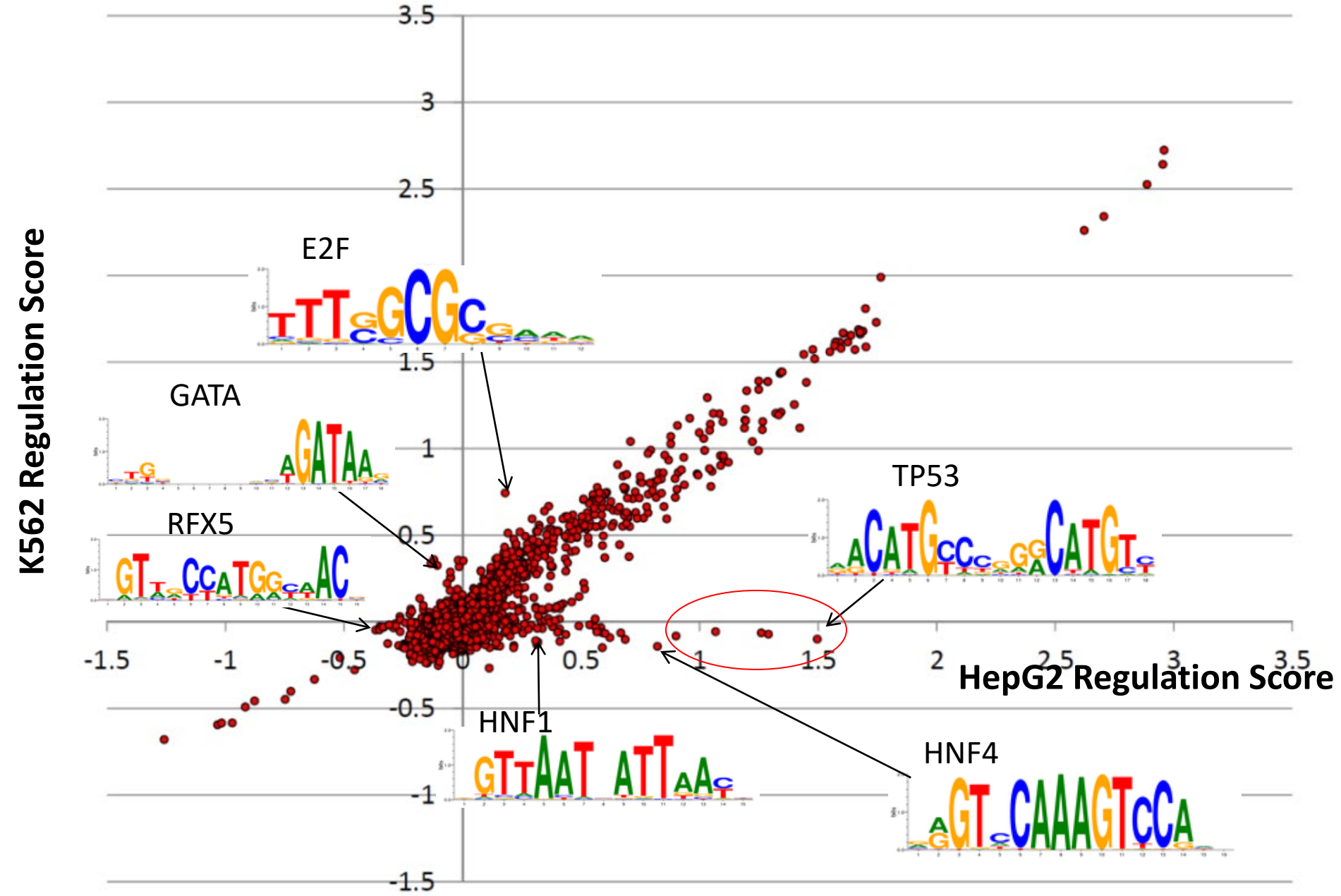



Comparing to ~1900 motifs - both known and discovered on ENCODE TF ChIP-seq data (Kheradpour and Kellis, 2014) with >= 20 instances overlapping testing regions

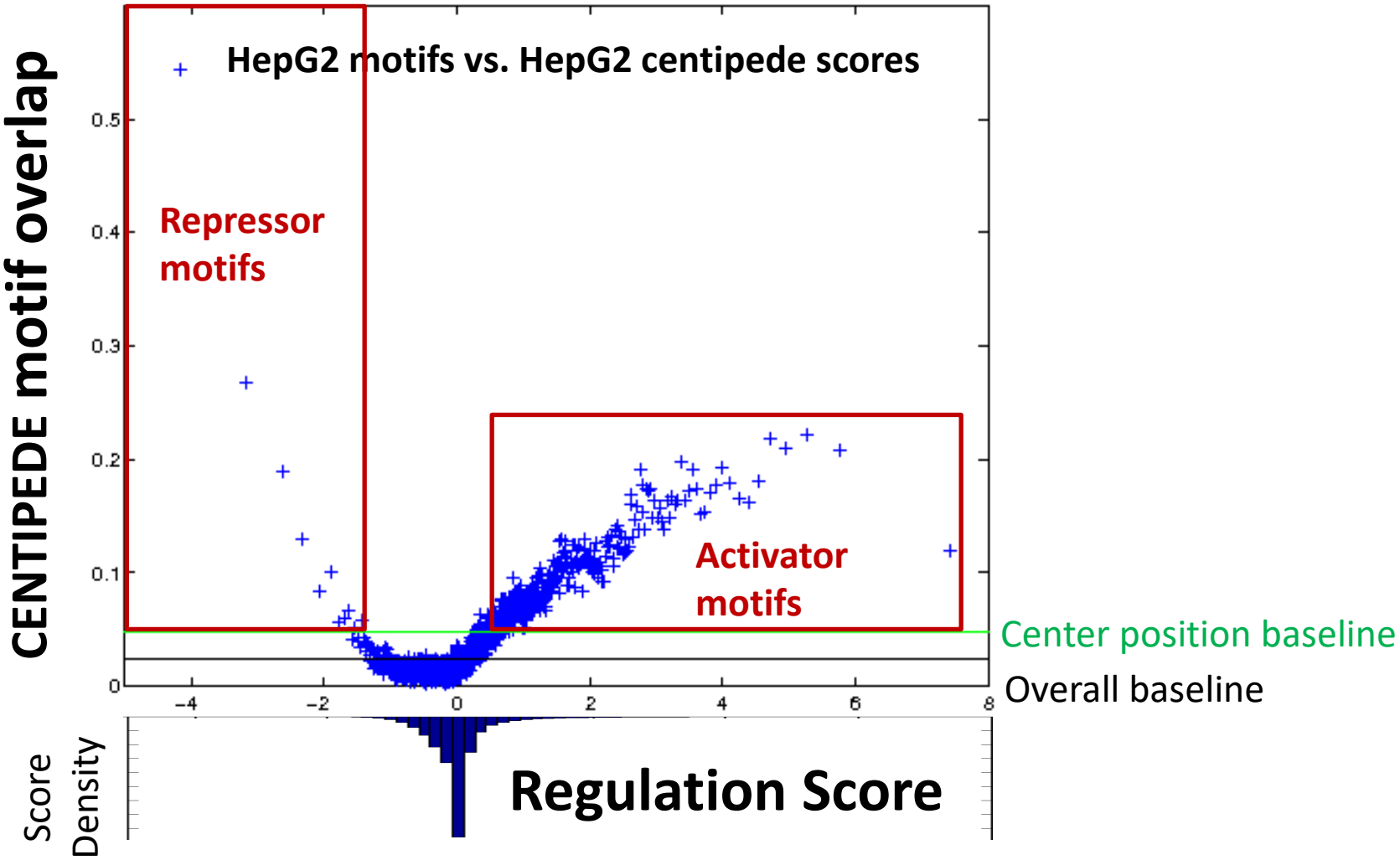
# Top Activating and Repressive Motifs Revealed



# Cell Type Specific Motifs Revealed



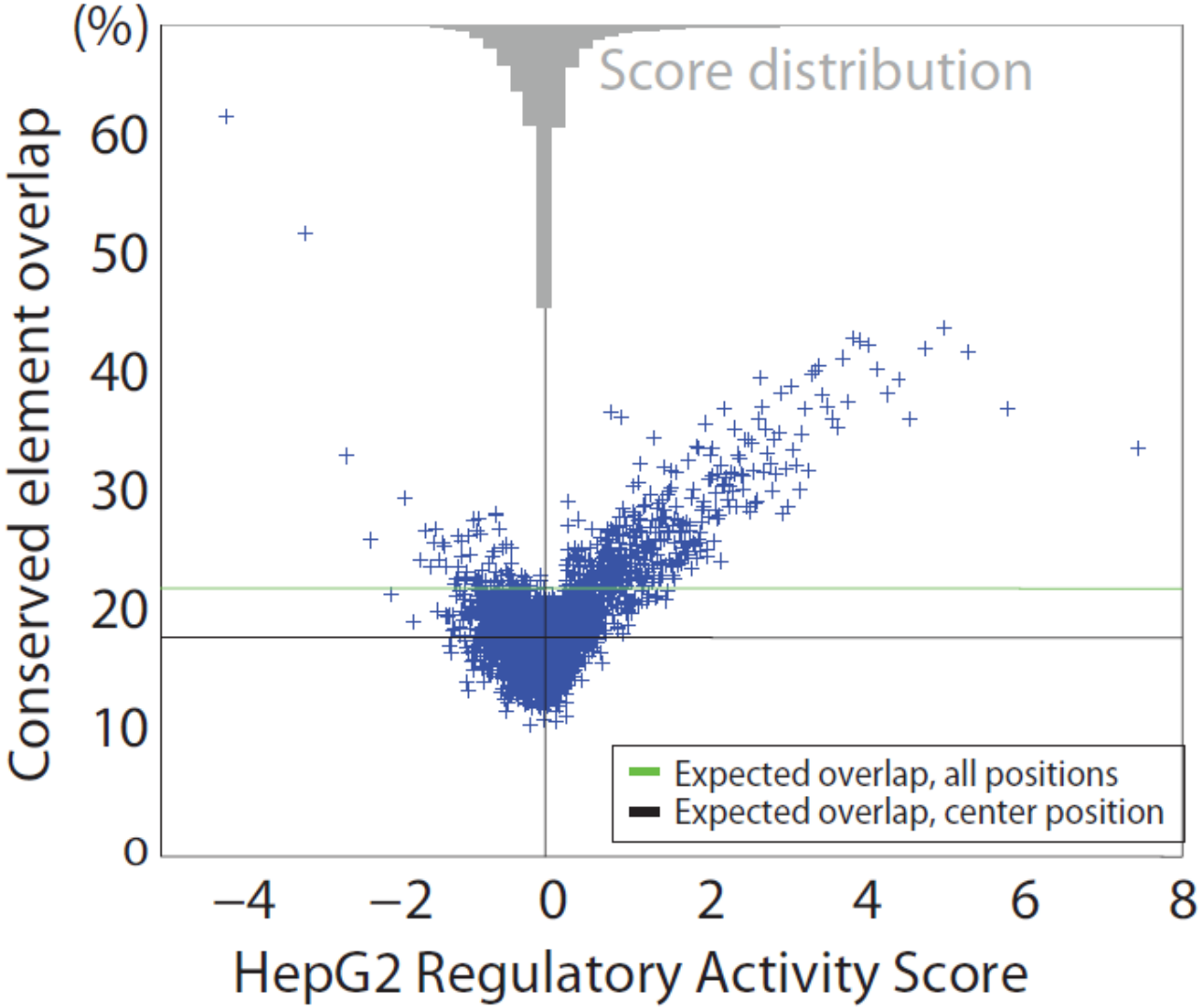
# Inferred positions match regulatory motifs






***Predicted Activation and Repressive Bases Strongly Enrich for Predicted Binding Sites in HepG2 + K562***

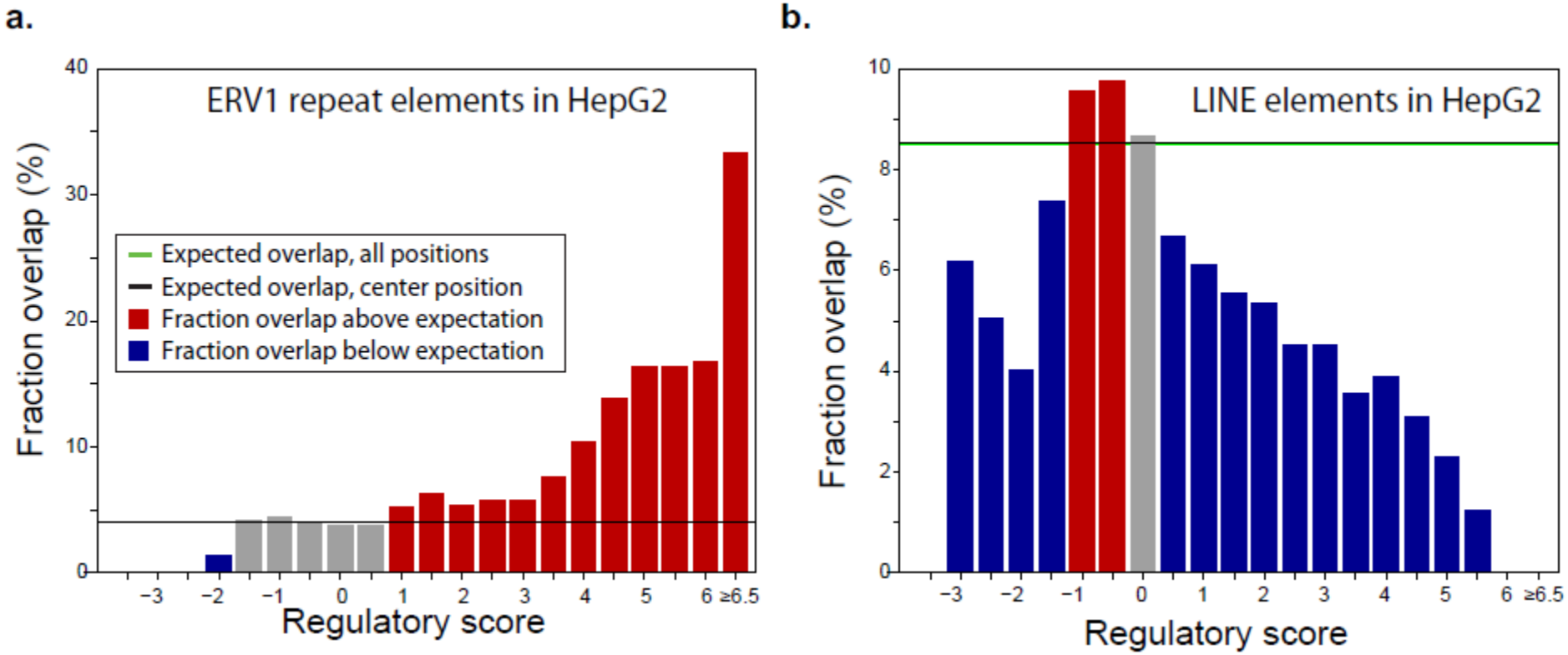
# Active/repressed positions are evol. conserved




***Strongest enrichment for repressive positions***

***Slight depletion at strongest activating positions***

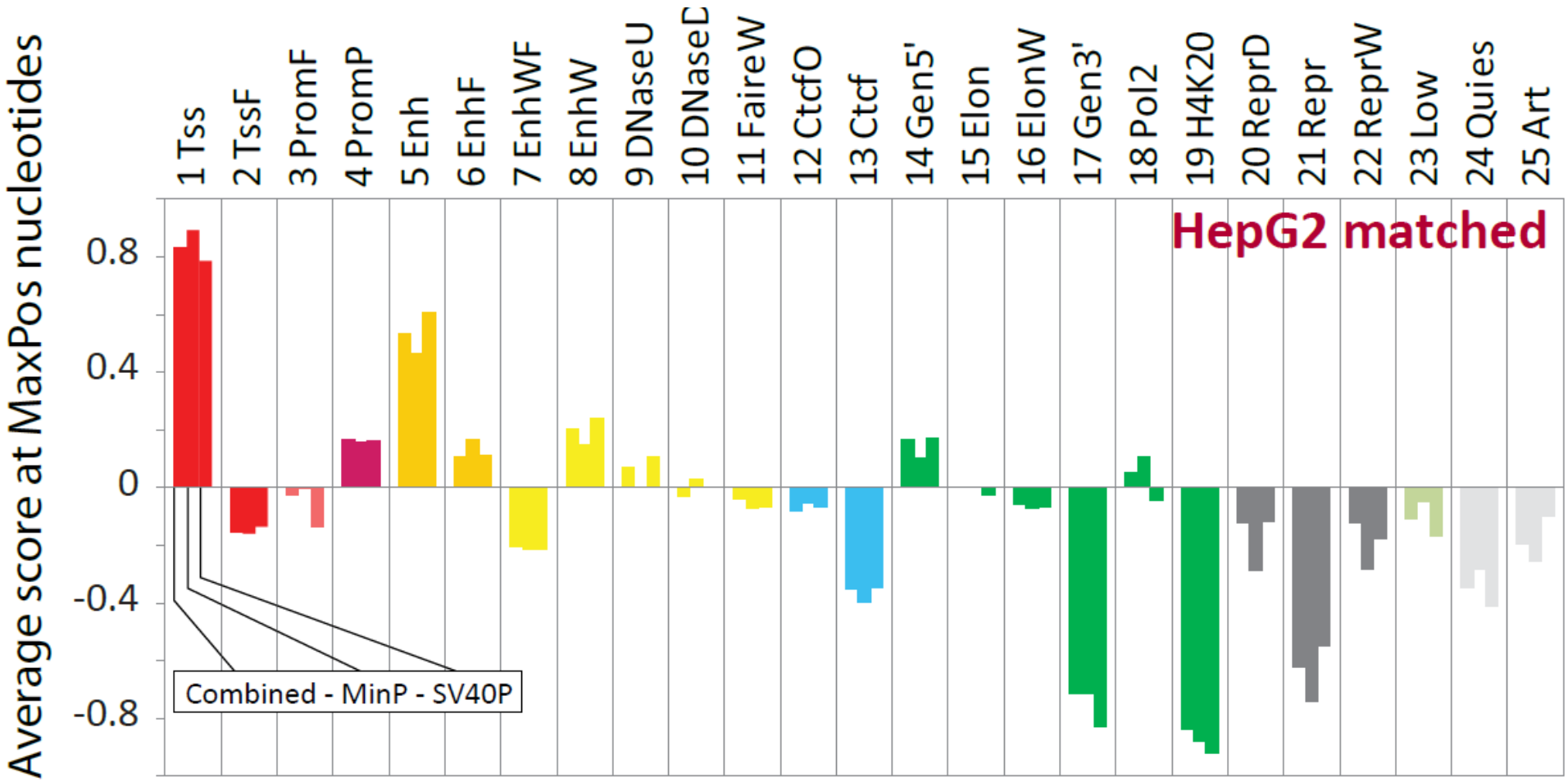
# ERV1 repeat elements can drive activity




***Strongest activating nucleotides match ERV1 repeats (by contrast, LINE elements strongly depleted)***

***Enable rapid evolution of gene-regulatory networks***

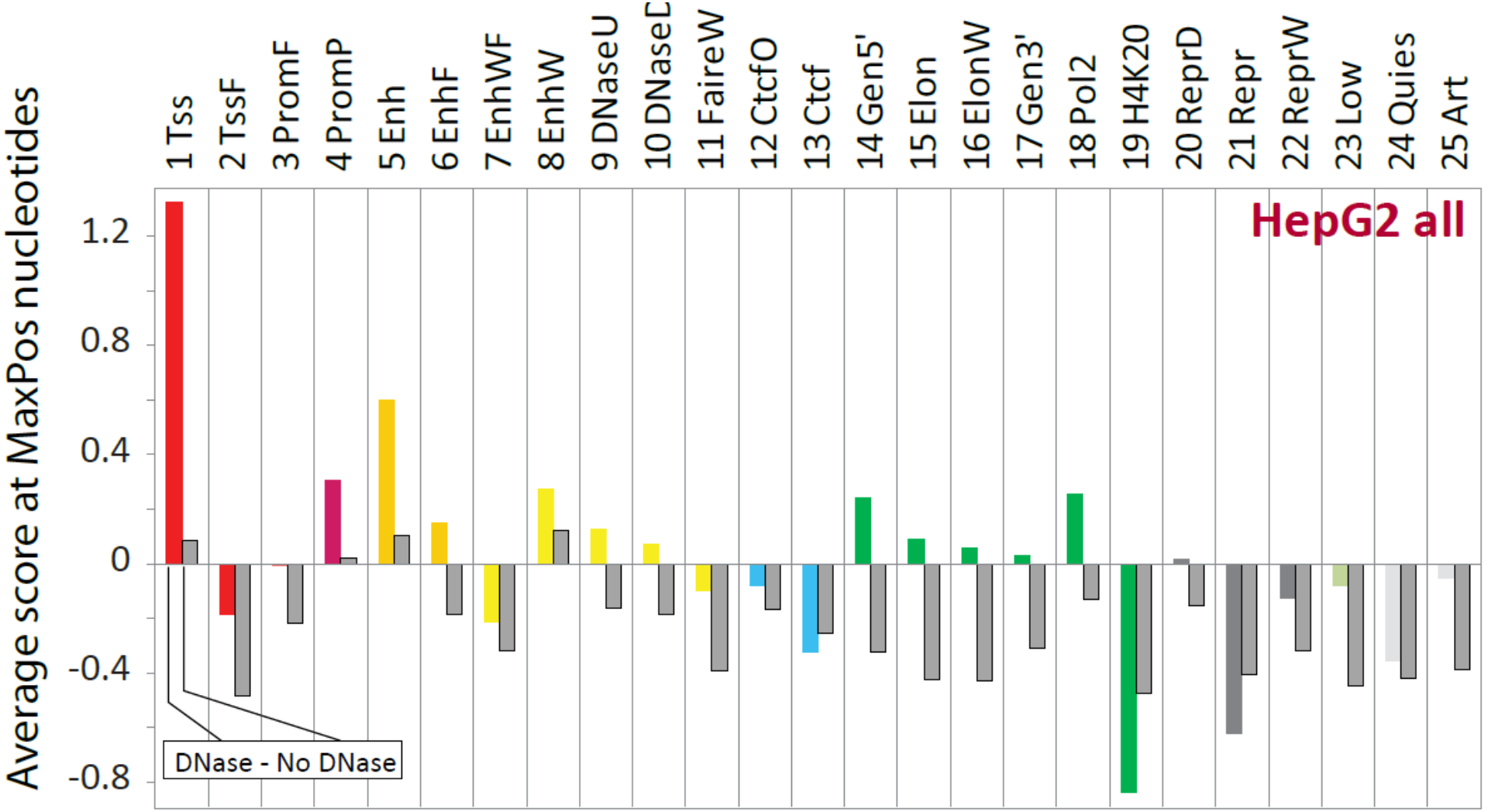
# DNase elements in different chromatin states differ in their activity levels




**Promoter, Enhancer regions highly activating.**

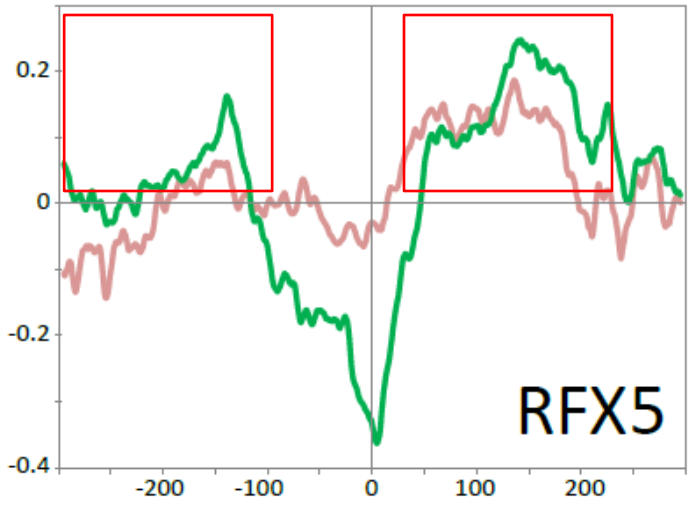
**ReprPC regions highly repressive**

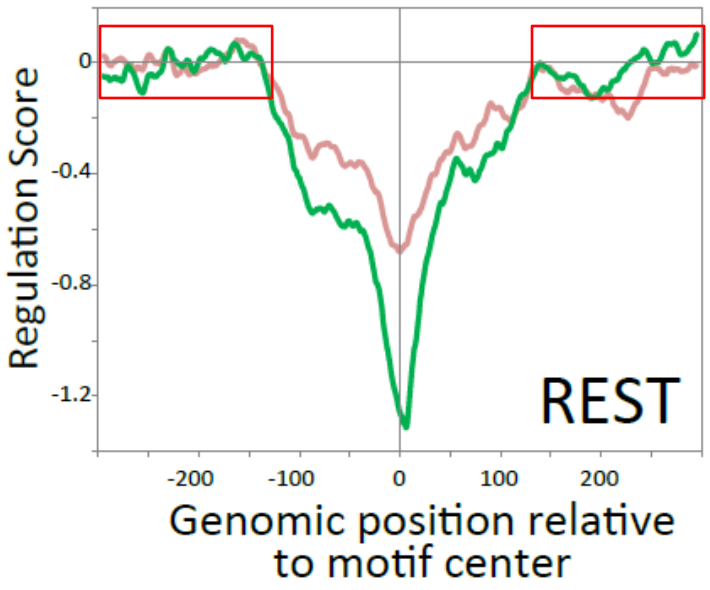
# Accessible regions drive stronger activity

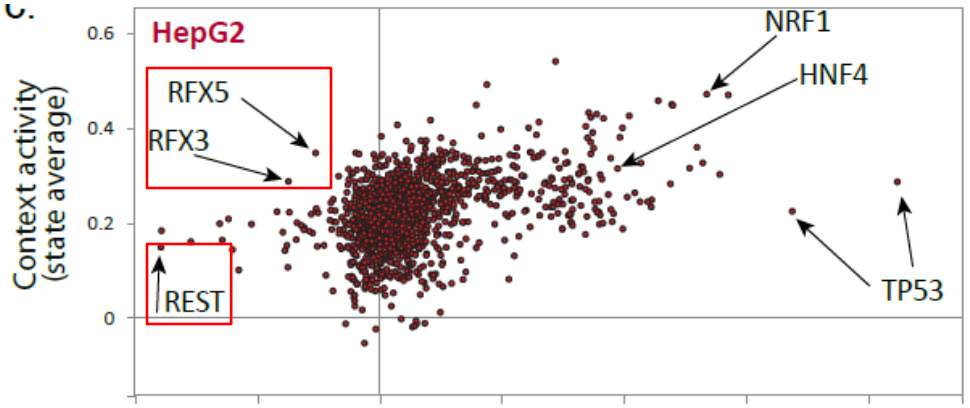


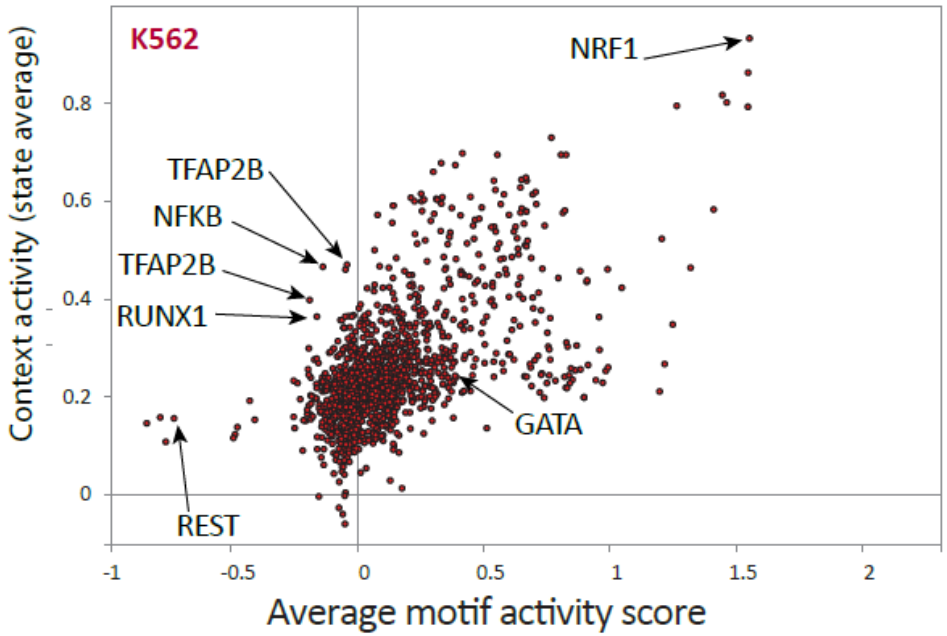

**For both activating and repressive positions**

# Discovery of repressors that act in active regions



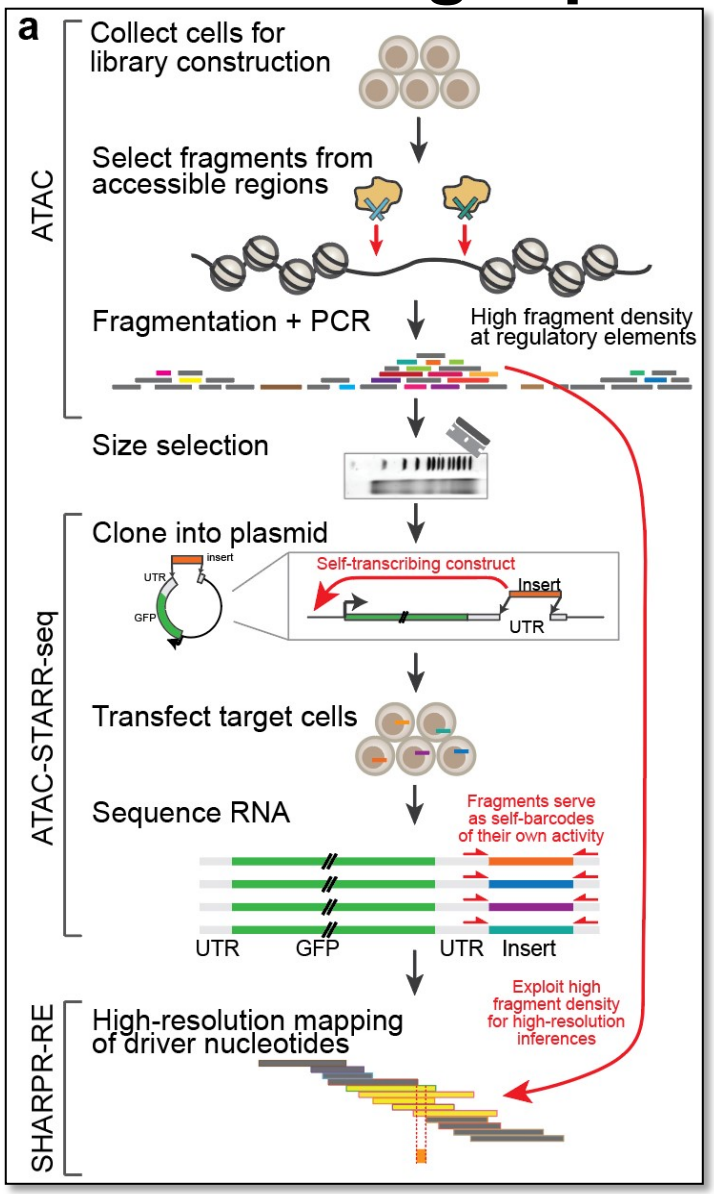



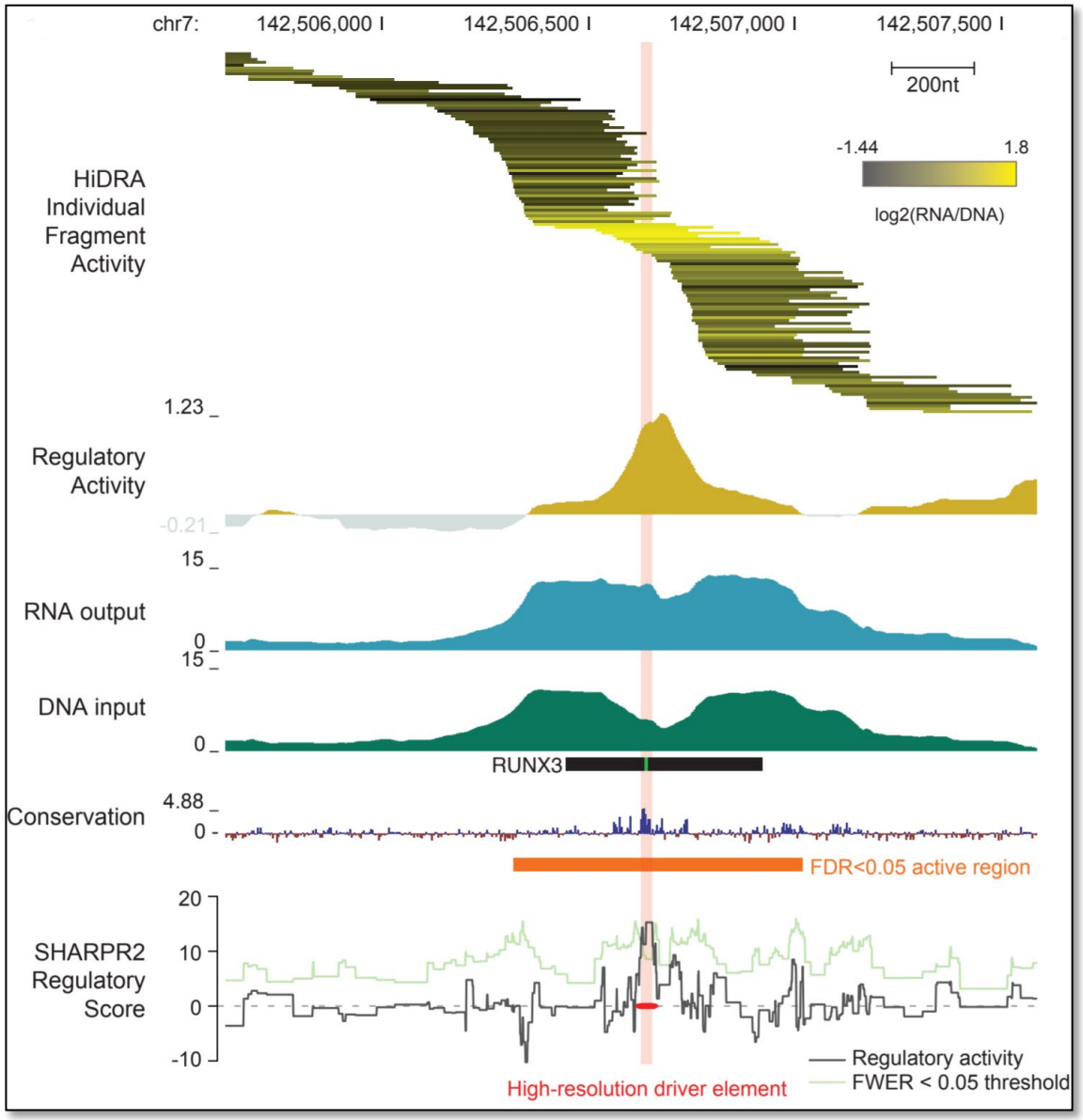

- ***REST acts as a repressor in repressive regions (as expected)***
- ***But RFX5 acts as a repressor only in active regions (modulator?)***

# Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint (M,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

# HiDRA: Longer probes + Hi-res dissection + 7M tests

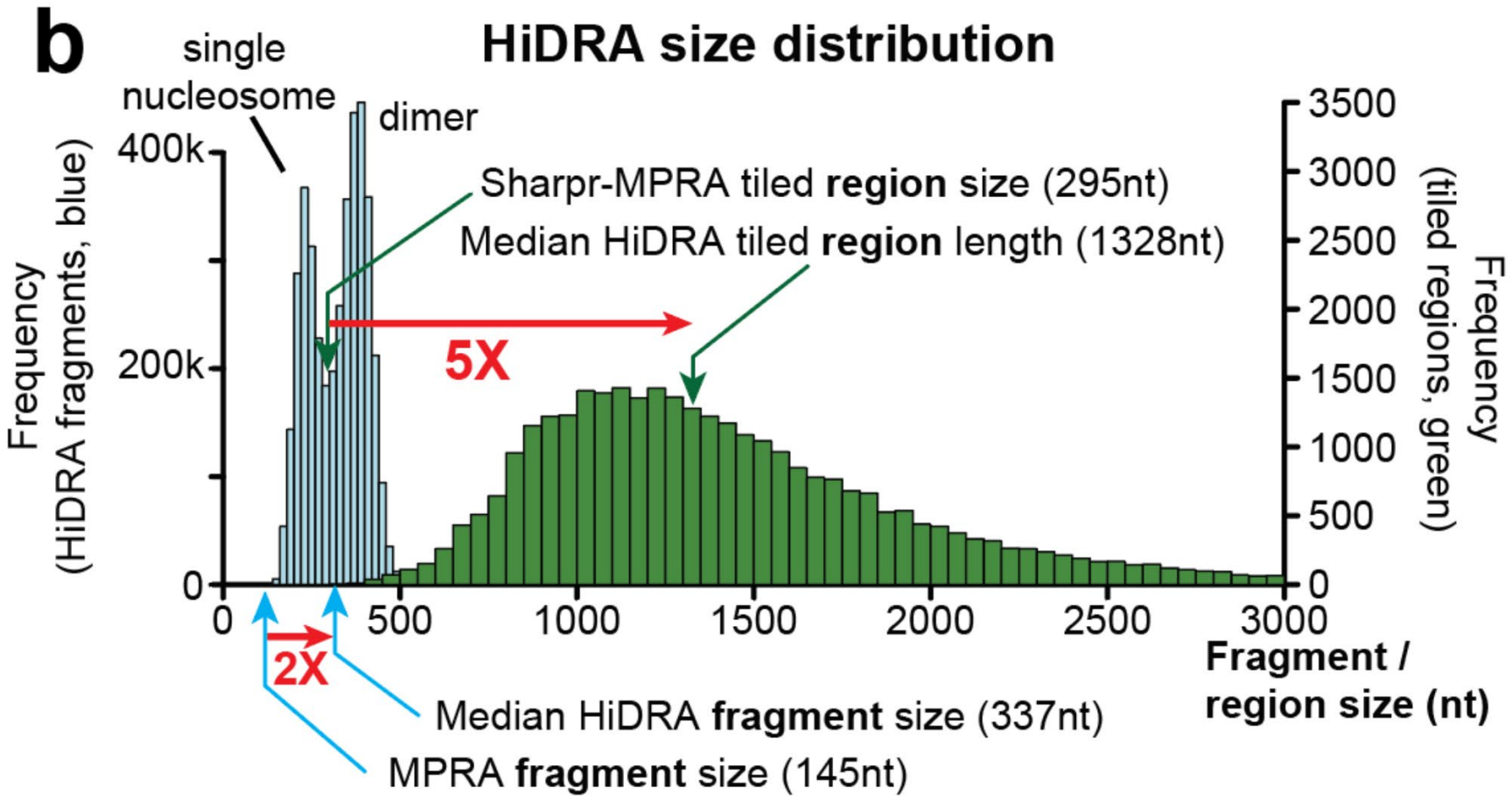


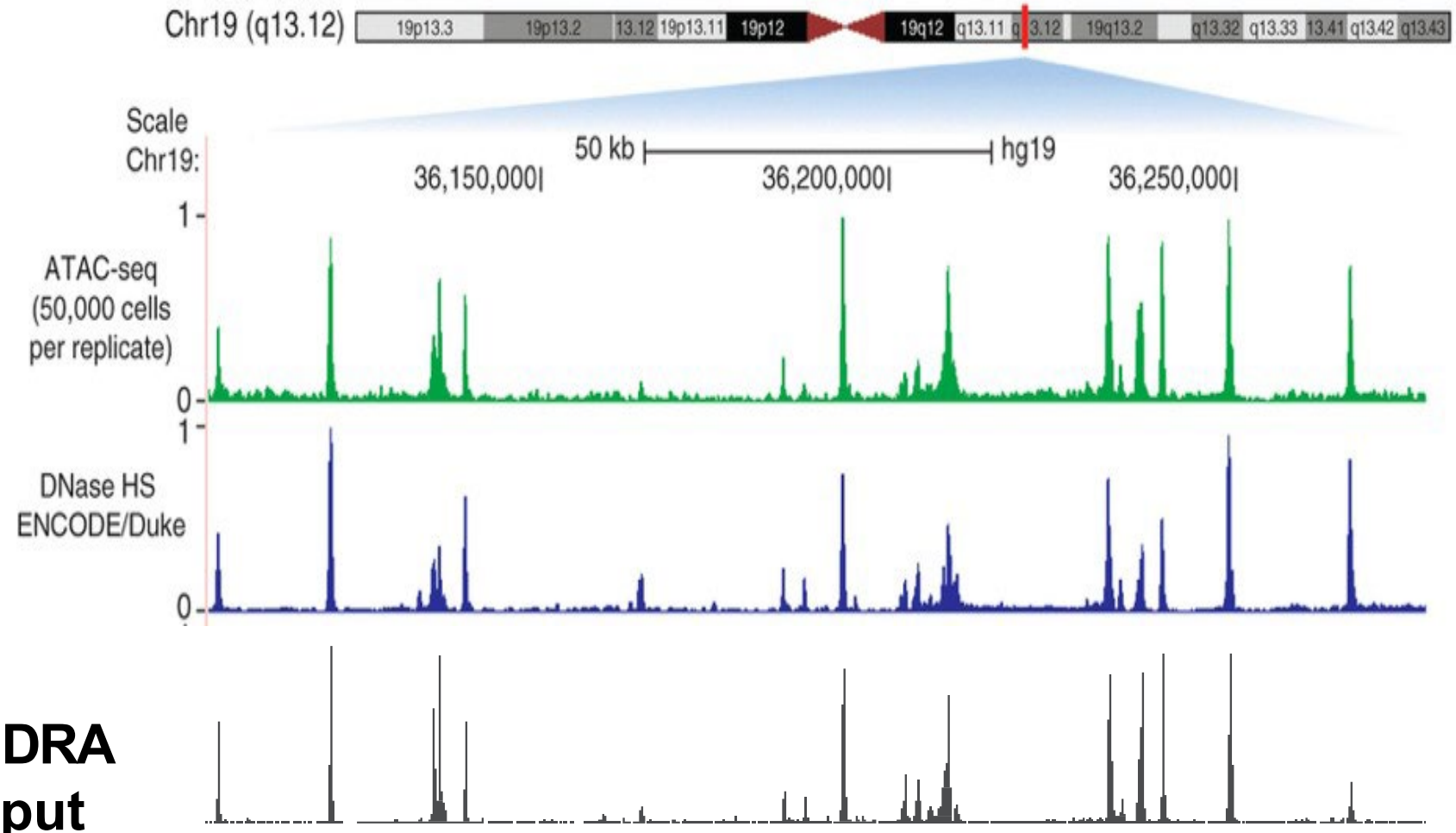

*ATAC selection* ➔ *No synthesis* ➔ ***7M tests***
*3'UTR incorp.* ➔ *Self-transcribe* ➔ ***No barcode***
*Dense, random start/end* ➔ ***Region tiling***

***High-resolution inference of driver nucleotides***
➔*Exploit differences between neighboring fragments*
➔*Driver nucleotides match motifs, evolut. conservation*

# HiDRA enables testing of larger fragments



# HiDRA input DNA library recapitulates DNase/ATAC-Seq

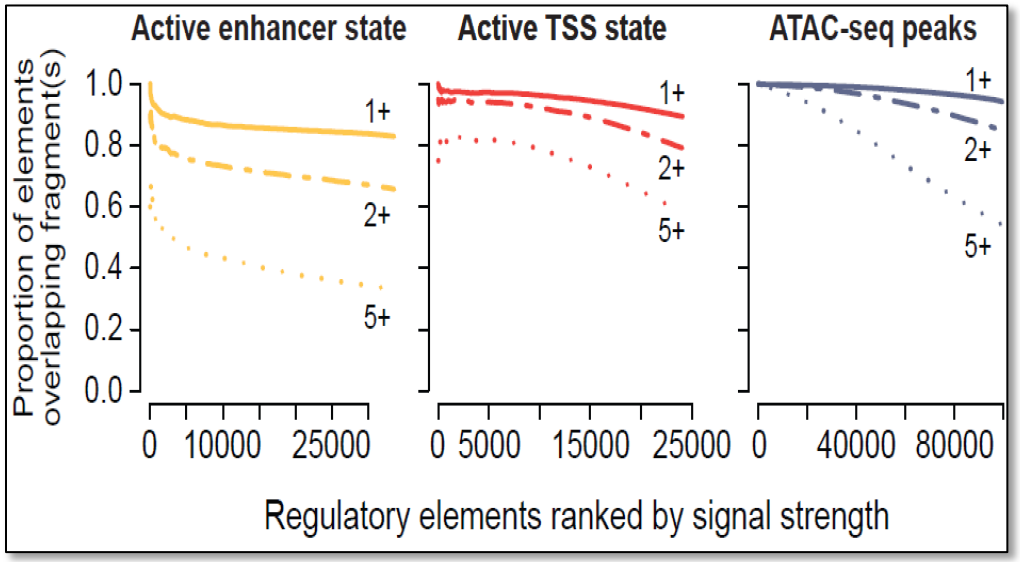



**HiDRA input DNA library**

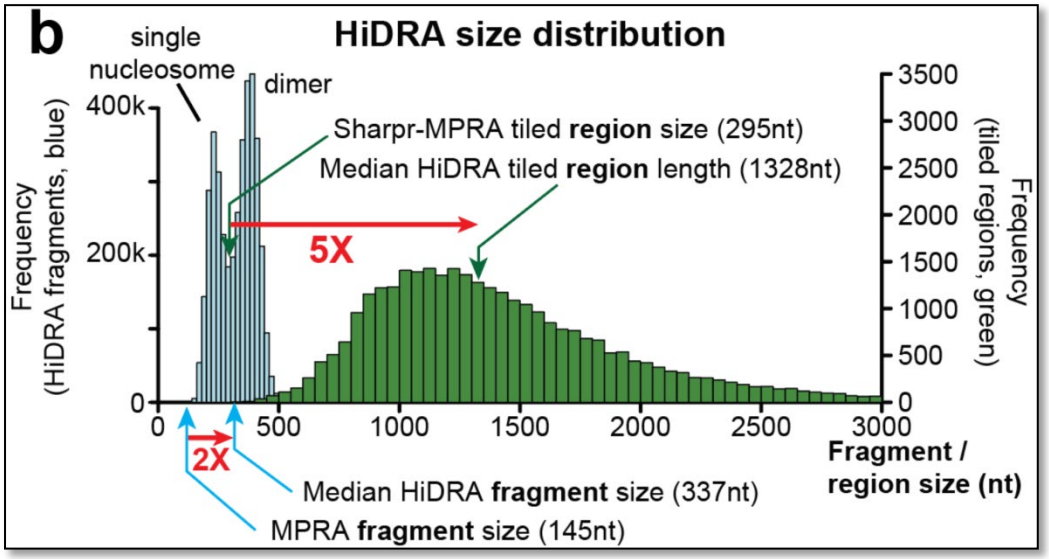
Fig 1c from Buenrostro *et al.* Nature Methods 2013

***Preferential selection of putative regulatory elements***

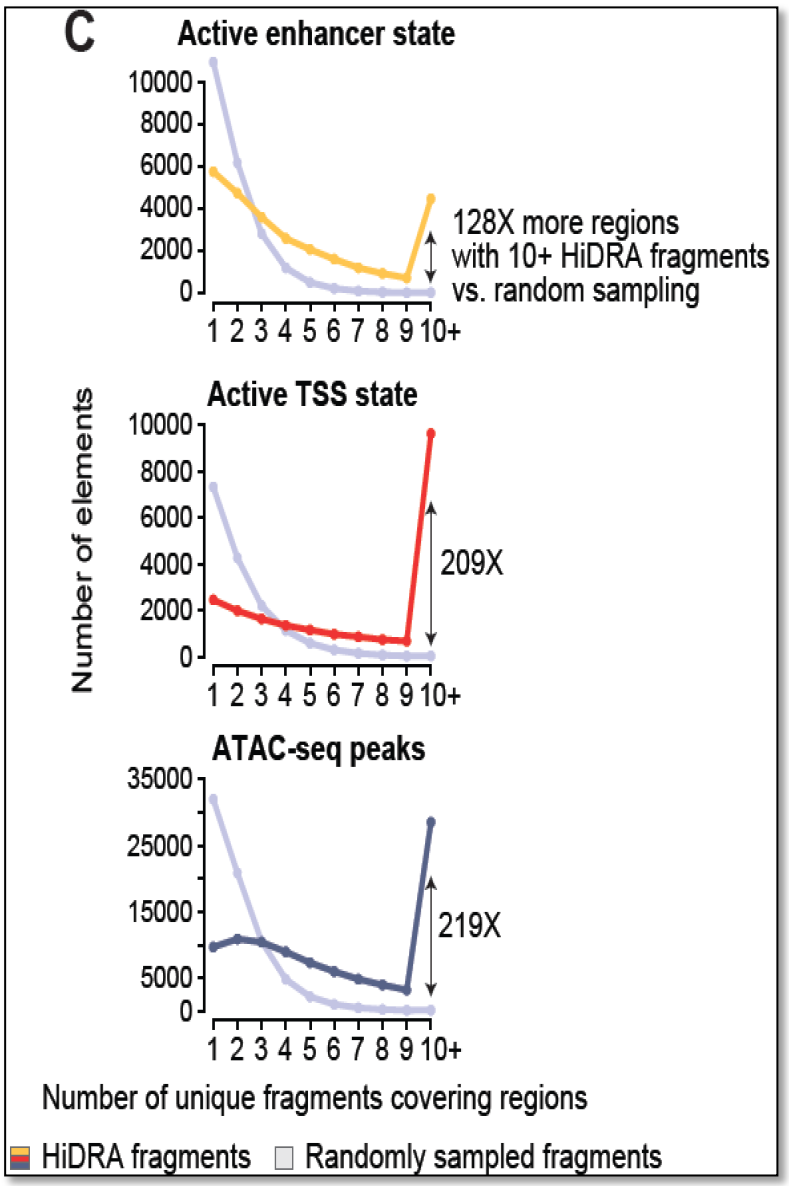
# HiDRA input DNA library: long, active, densely-covered regions




***HiDRA DNA library captures more active elements***

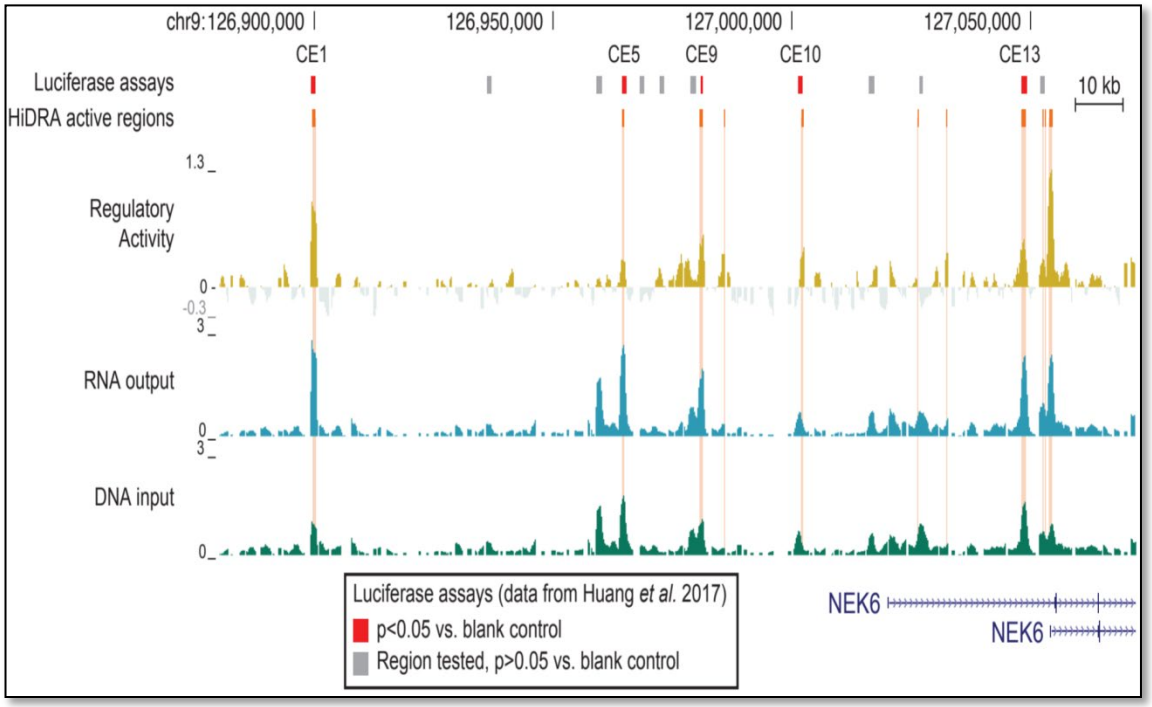


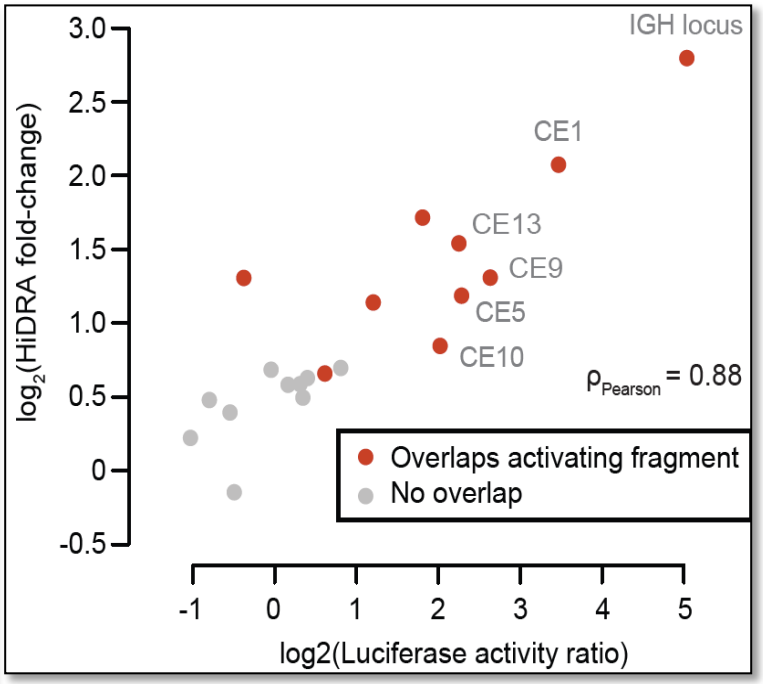

***Fragments: 99% are 169-477 nt (median: 337nt)***
***Regions: 99% are 513-4,036 nt (median: 1,328nt)***




***Up to 200-fold higher coverage for putative regulatory elements***

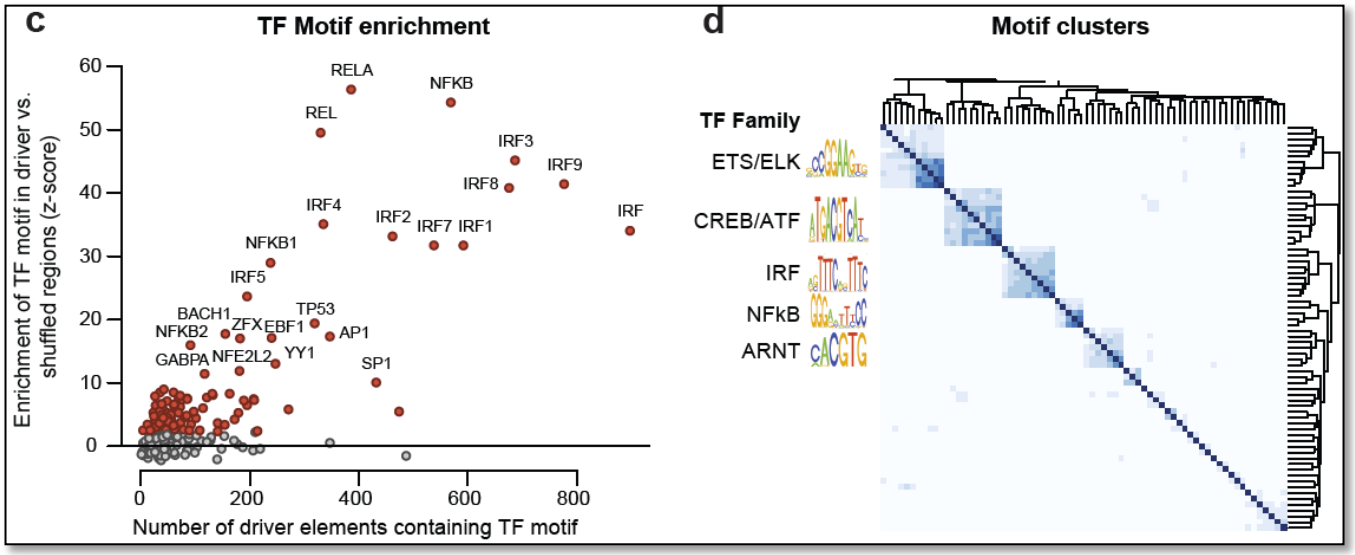
# HiDRA captures known enhancers, known motifs



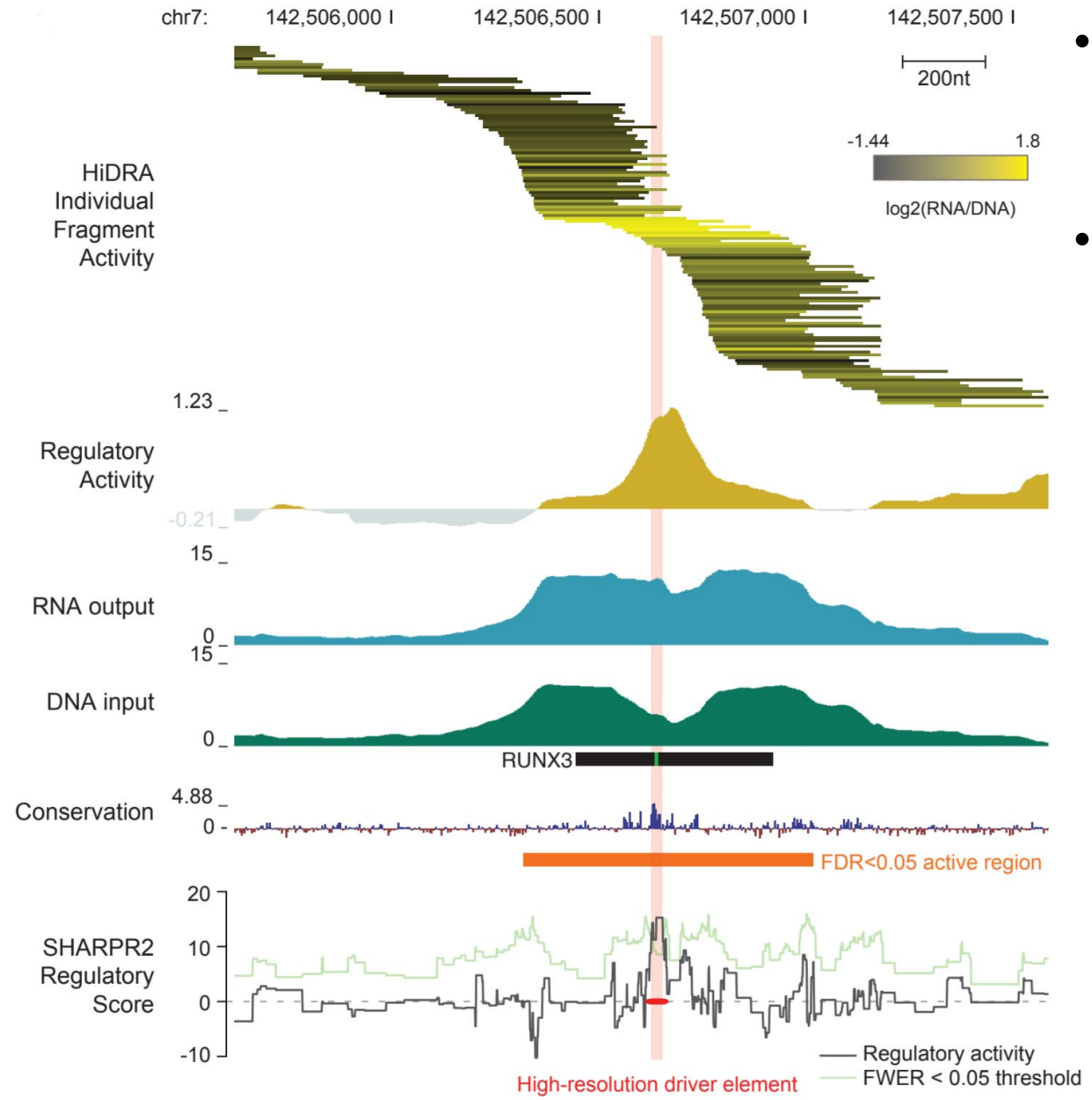

***High sensitivity / high specificity vs. Luciferase assays***

***Quantitative read-out***

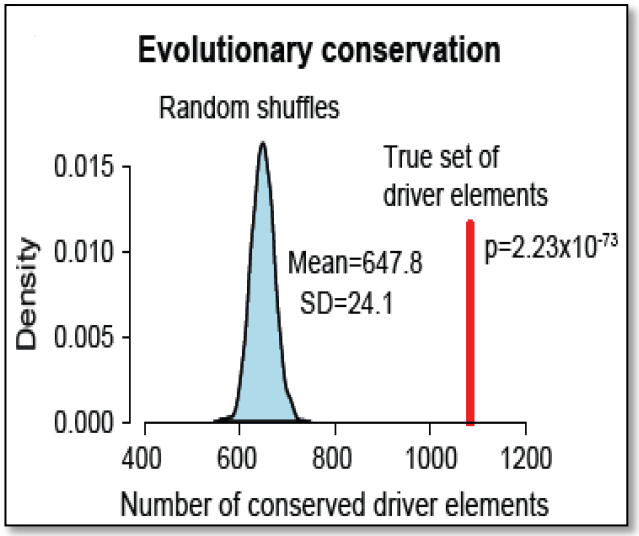



***Capture known motifs***

# Sharpr2 algorithm infers high-resolution driver nucleotides

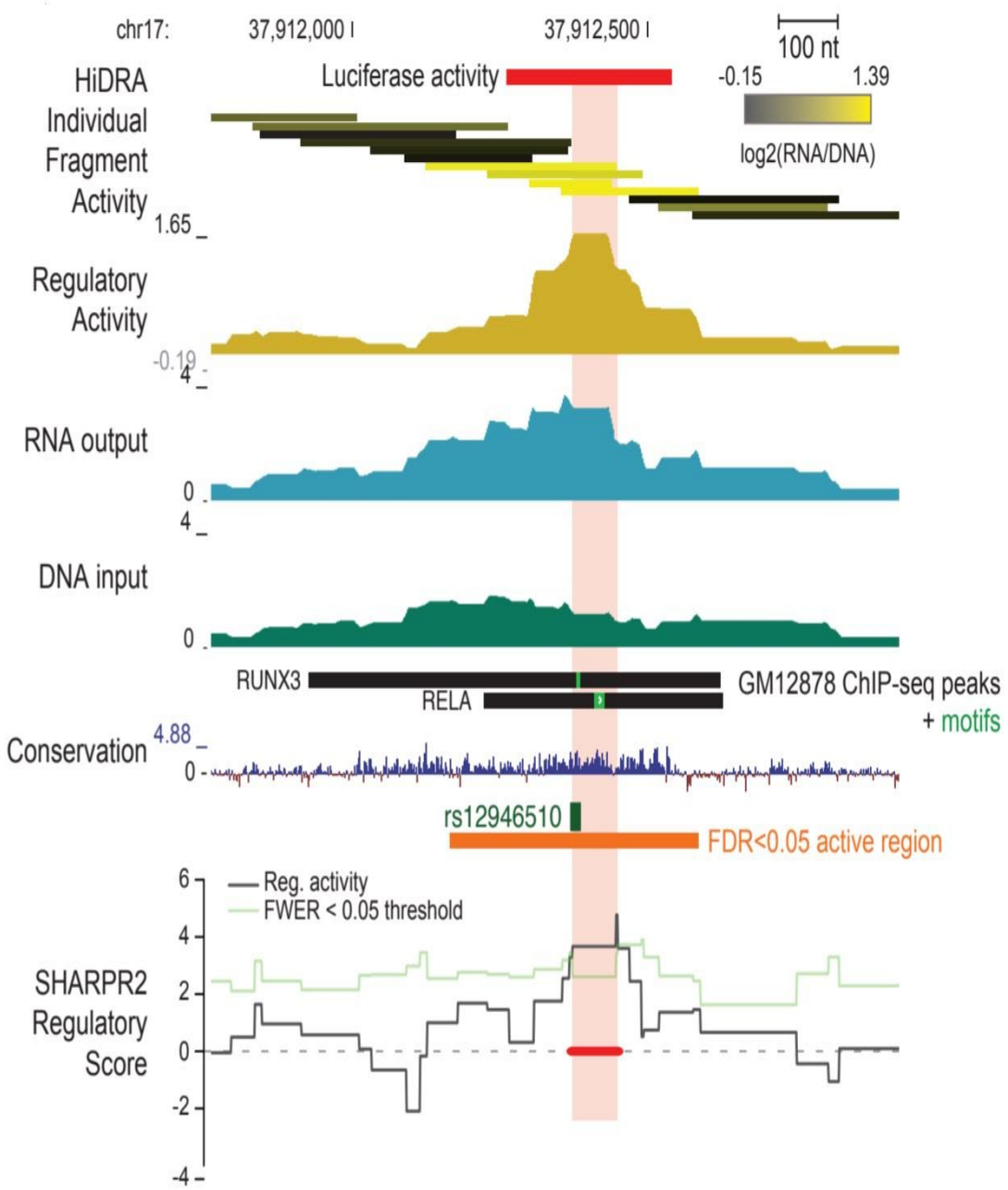



- **Exploit differences between neighboring fragments**
- **Driver nucleotides match motifs, evolutionary conservation**




- **Enrichment: $P<10^{-73}$**

# HiDRA high-resolution drivers help dissect GWAS loci

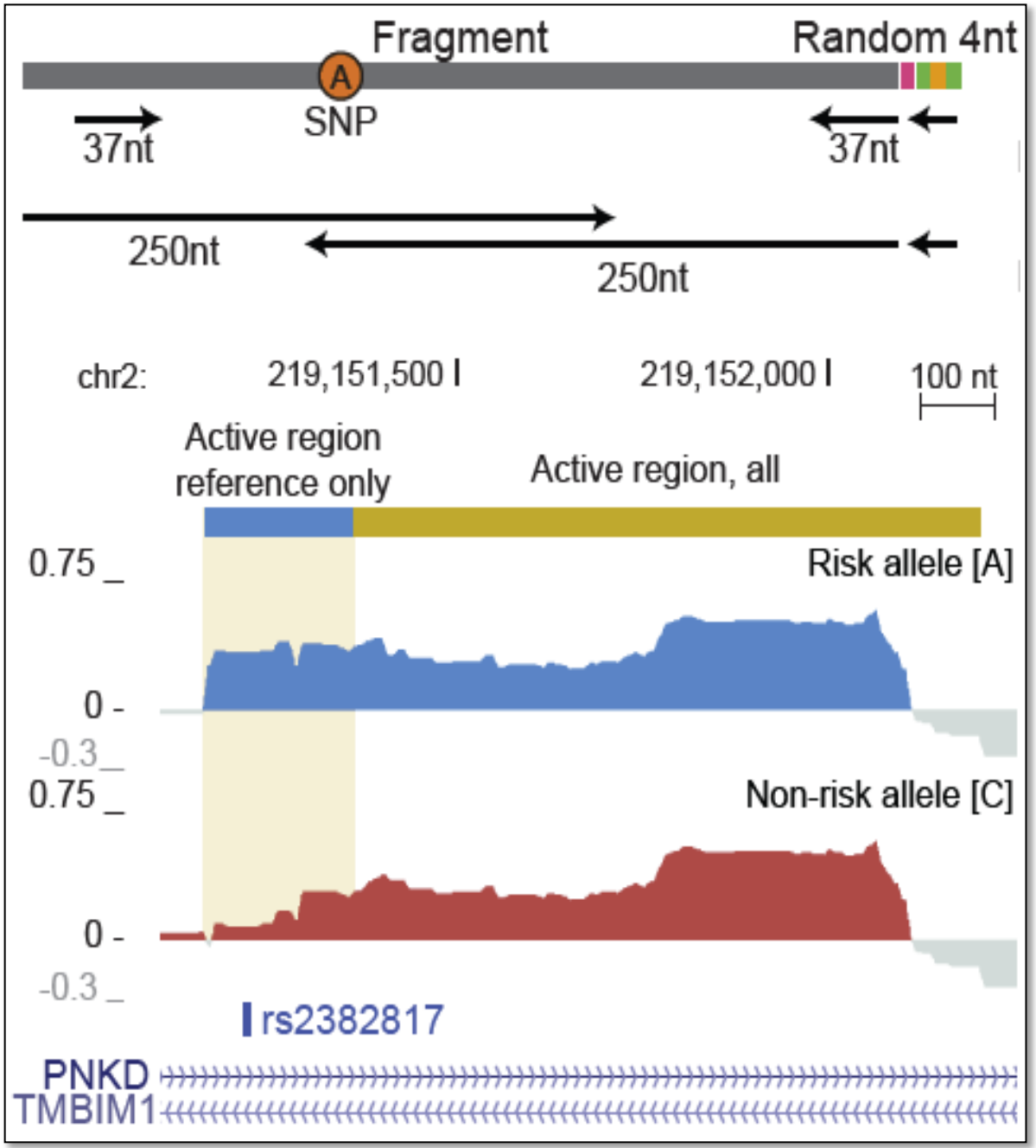


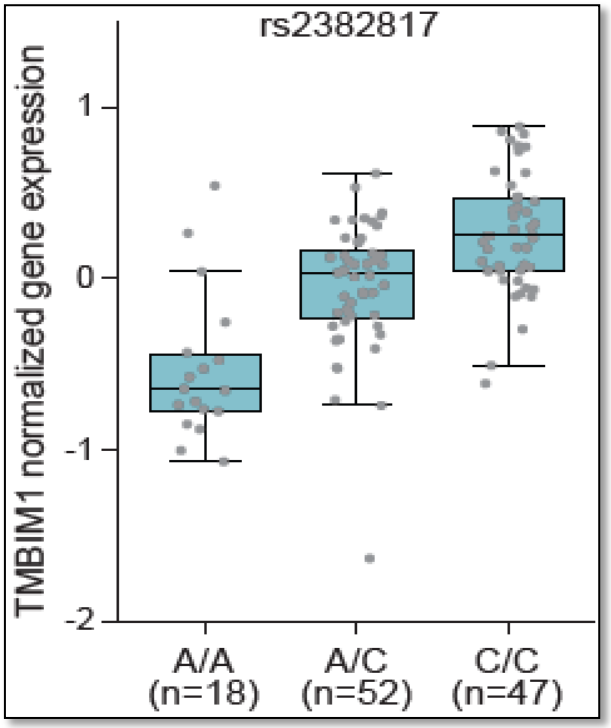

➔**General method to dissect non-coding variation**

➔**Applicable to millions of genomic regions simultaneously**

***Pinpoint causal GWAS variants***

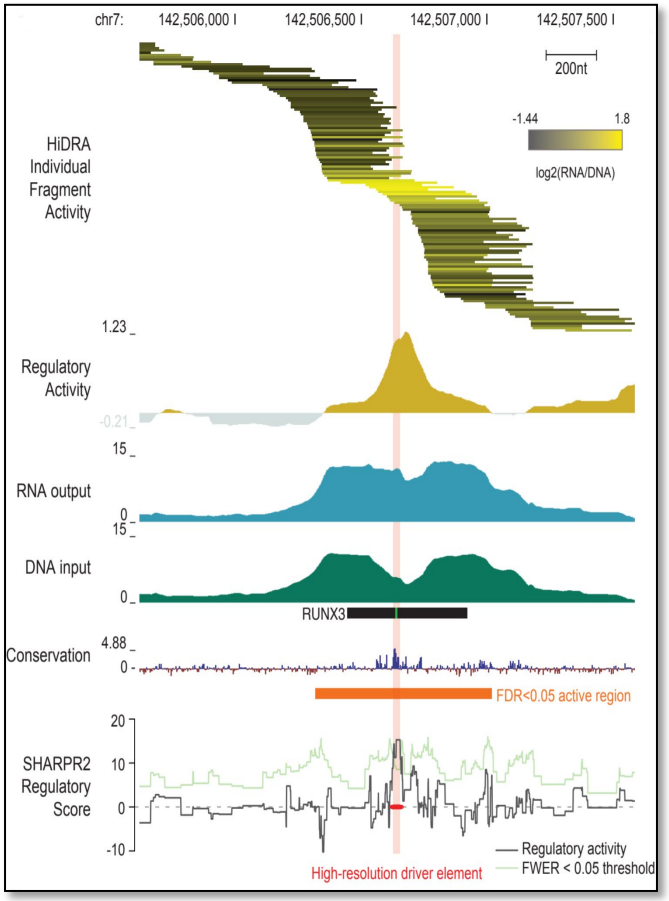
# HiDRA activity differences between risk and non-risk alleles




***Allele-specific activity for IBD-associated rs2382817***

# HiDRA summary




- 7M fragments tested in one experiment
- Longer fragments (~350nt on average)
- High reproducibility, 0.95 for higher-activity elmt
- Up to 200-fold enrichment for regulatory regions
- High-resolution dissection of driver nts
- Captures known motifs, conserved nucleotides
- Pinpoints driver SNPs in GWAS loci
- Reveals diffs between risk and non-risk alleles
- **General tool for testing regulatory regions**

bioRxiv beta

doi.org/10.1101/193136

**High-resolution genome-wide functional dissection of transcriptional regulatory regions in human**

Xinchen Wang[1,2,3,†], Liang He[2,3], Sarah M. Goggin[2], Alham Saadat[2], Li Wang[2], Melina Claussnitzer[2,4,*], Manolis Kellis[2,3,*]

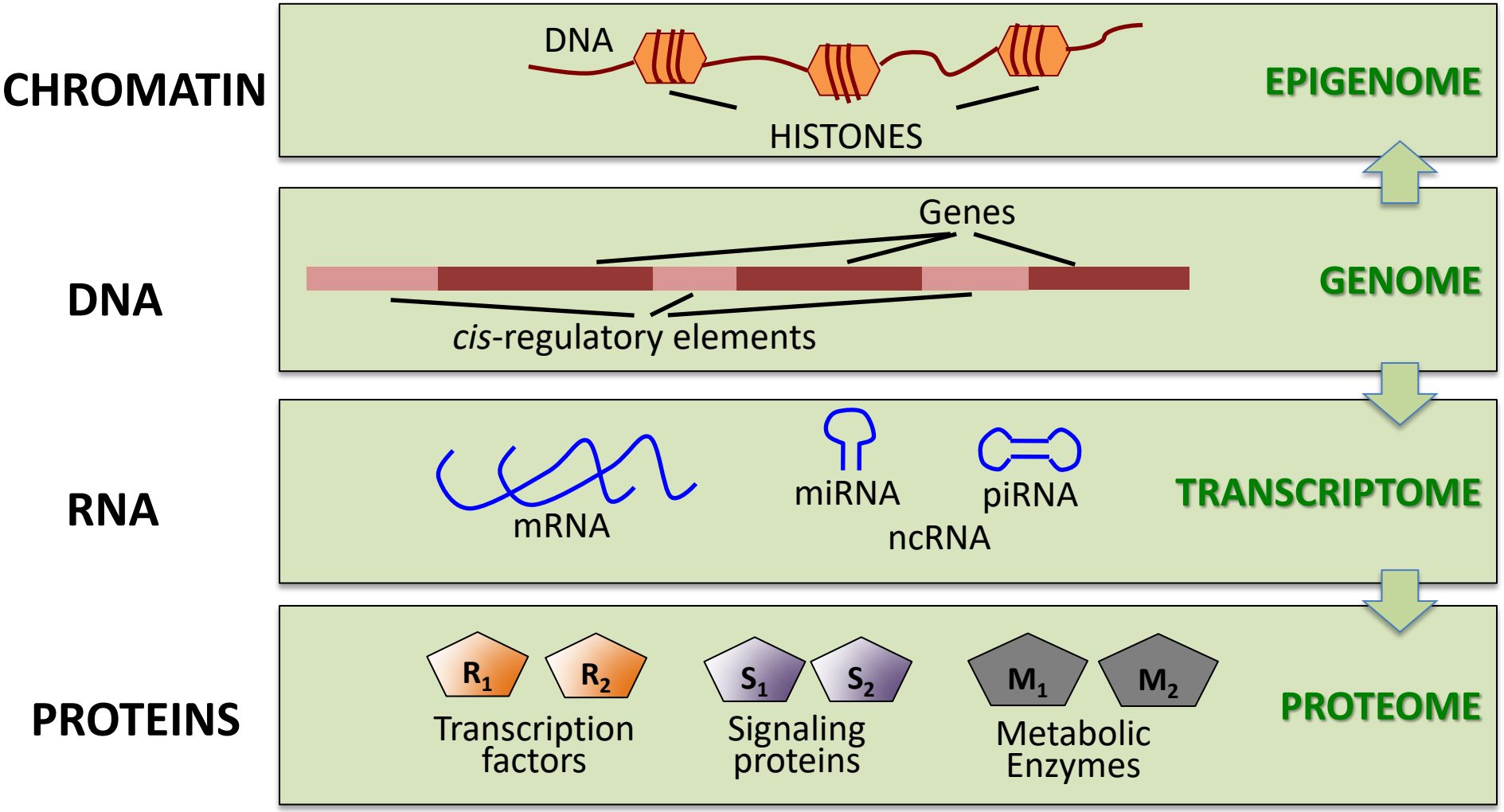
# Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint ($M$,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Positioning matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

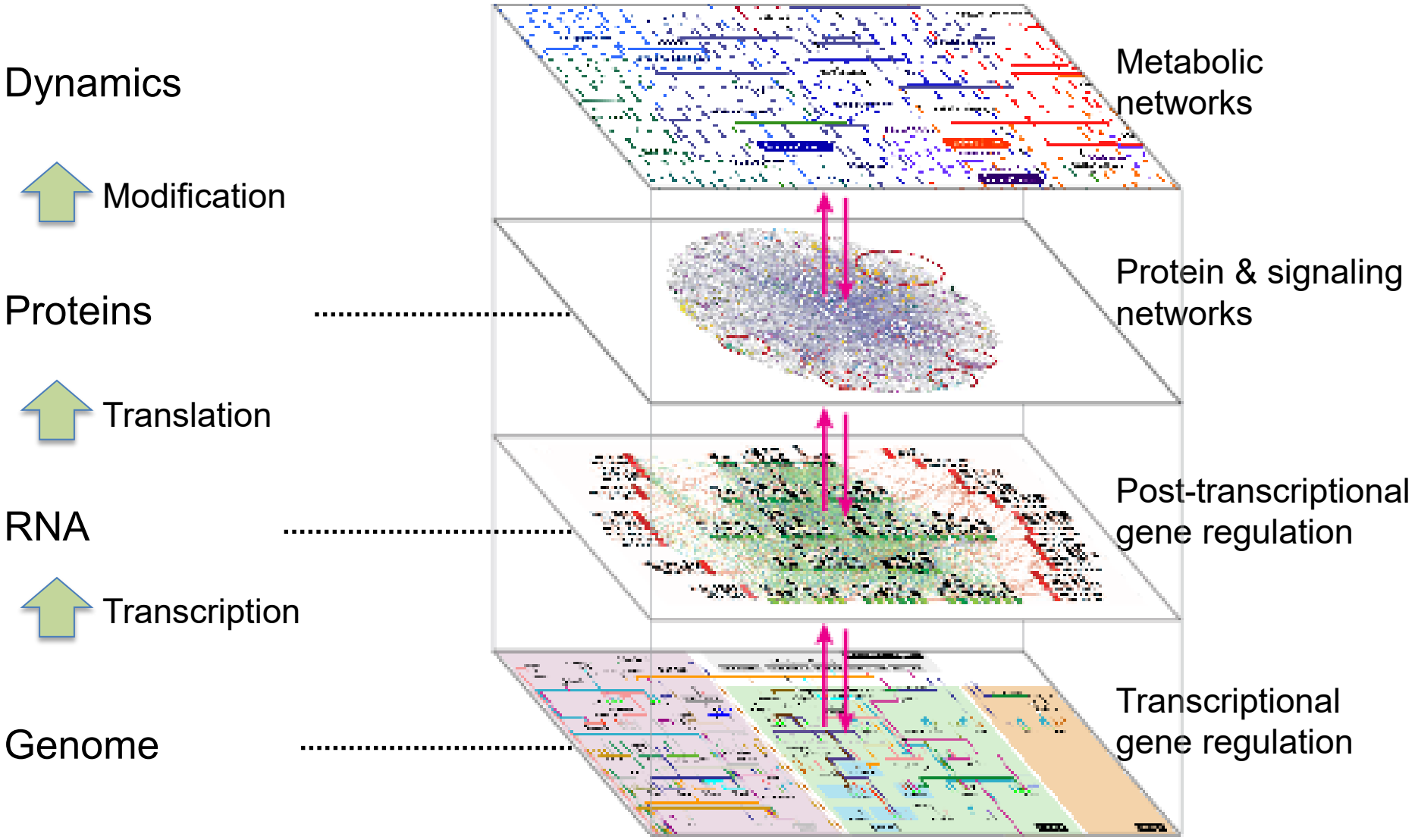
# Goals for today: Network analysis

1. **Introduction to networks**
   - Network types: regulatory, metab., signal., interact., func.
   - Bayesian (probabilistic) and Algebraic views
2. **Network Centrality Measures**
   - Local centrality metrics (degree, betweenness, closeness, etc)
   - Global centrality metrics (eigenvector centrality, page-rank)
3. **Linear Algebra Review: eigenvalues, SVD, low-rank approximations**
   - Eigenvector and singular vector decomposition
   - Low rank approximations, Wigner semicircle law
4. **Sparse Principal Component Analysis**
   - Lasso and Elastic lasso
   - PCA and Sparse PCA
5. **Machine Learning in Networks**
   - Guilt by association
   - Maximum cliques, density-based modules and spectral clustering
6. **Network Diffusion Kernels and Deconvolution**
   - Network diffusion kernels
   - Network deconvolution

# The multi-layered organization of information in living systems

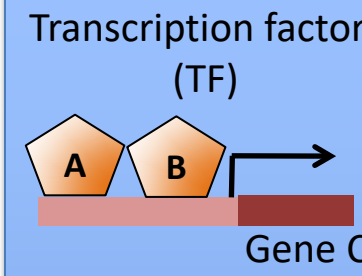


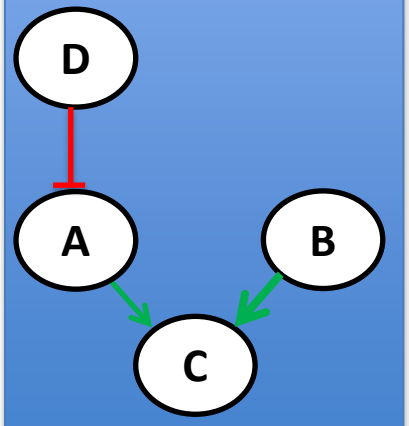
# Biological networks at all cellular levels



# Five major types of biological networks

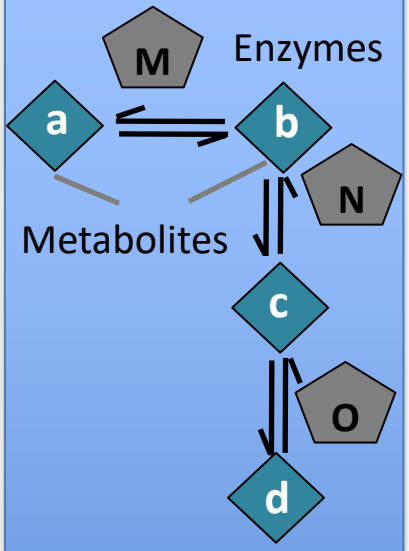
## Regulatory network

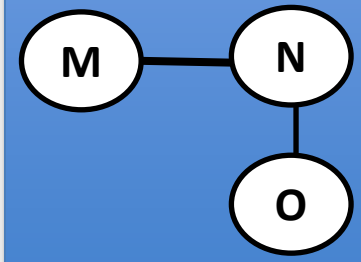



Directed, Signed, weighted

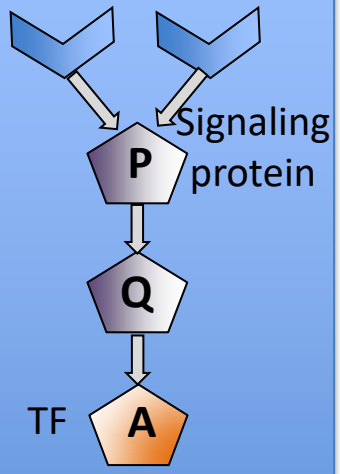
## Metabolic network




Undirected, weighted

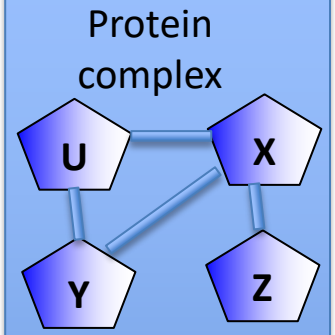
## Signaling network

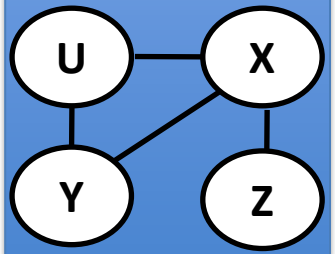




Directed, unweighted

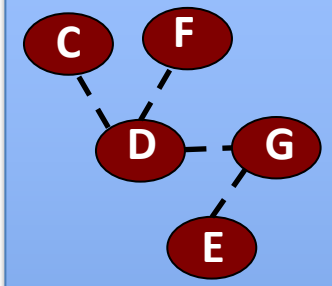
## PPI, Protein interaction network

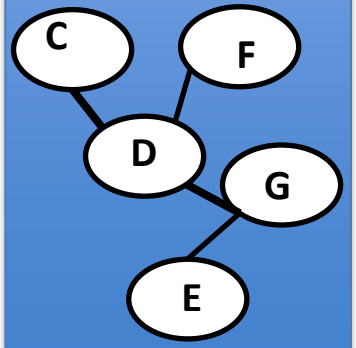



Undirected, unweighted

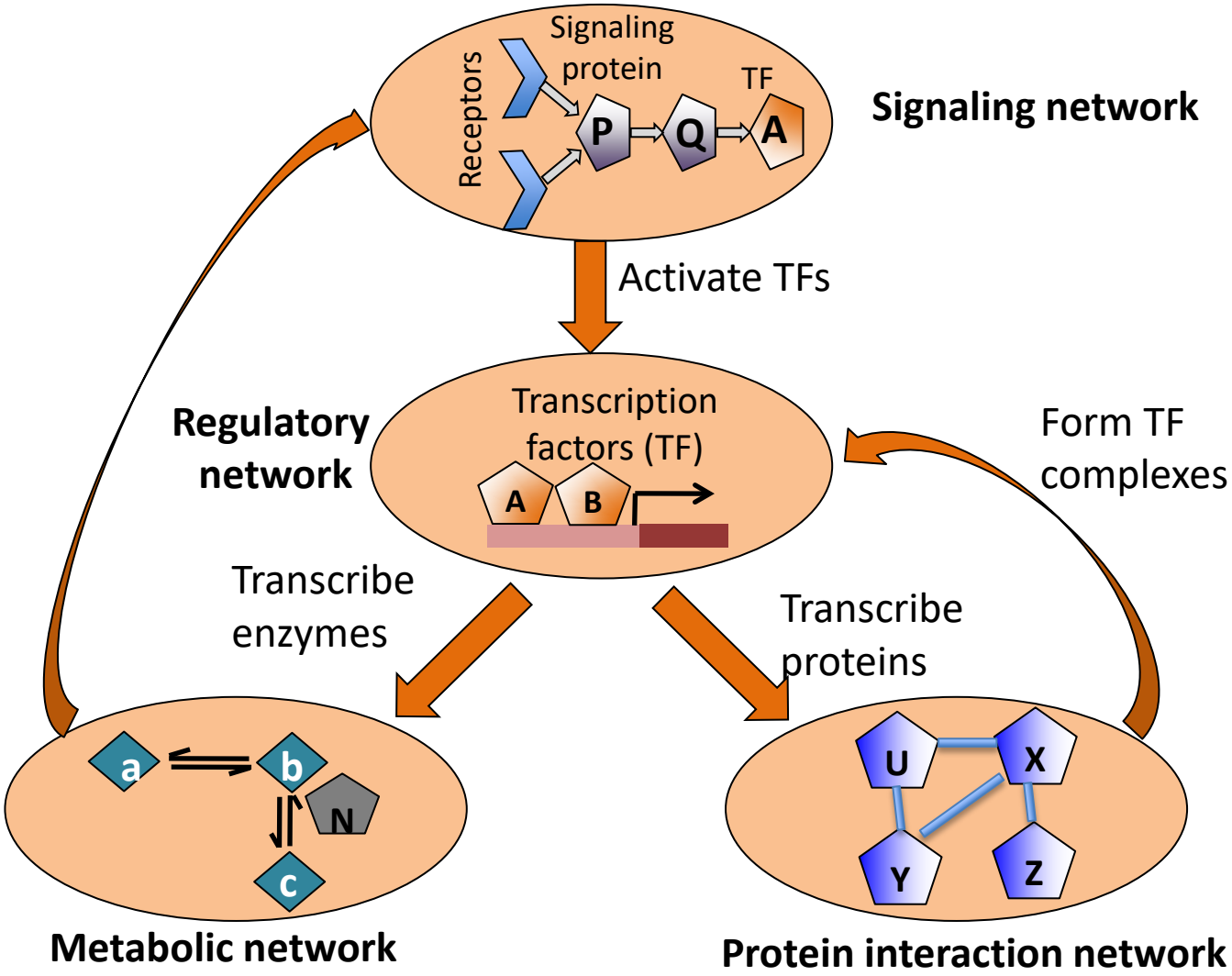
## Functional network (Co-expression)



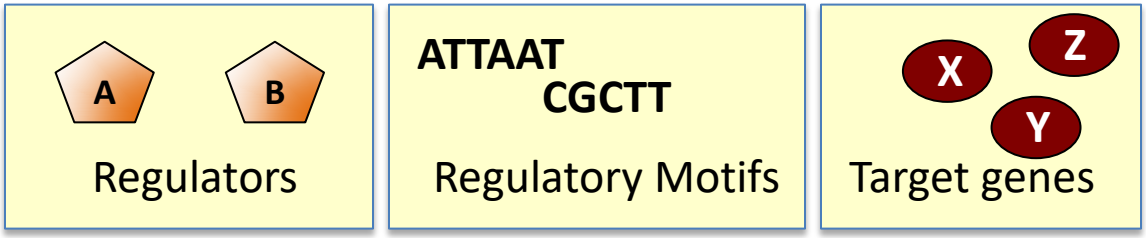

Undirected, weighted

# Information exchange across networks

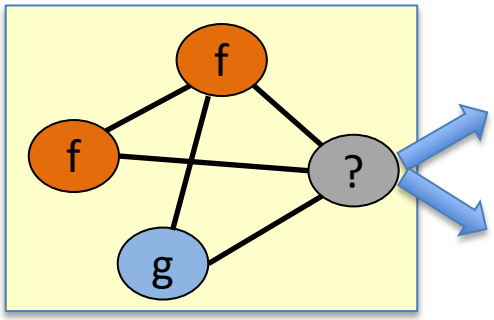


# Network applications and challenges

1. **Element Identification (motif finding lecture)**



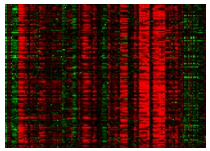

2. **Using networks to predict cellular activity**



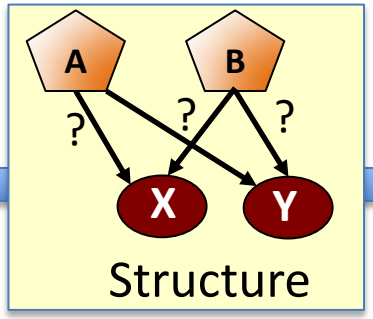

Predict expression levels

Predict gene ontology (GO) functional annotation terms

3. **Inferring networks from functional data**



Activity patterns

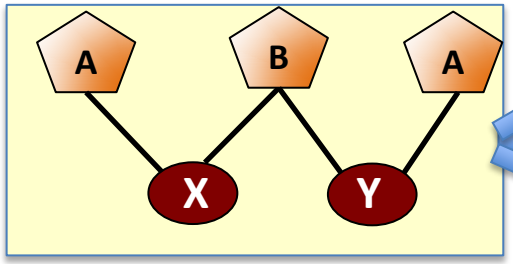



**X=f(A,B)**

**Y=g(B)**

Function

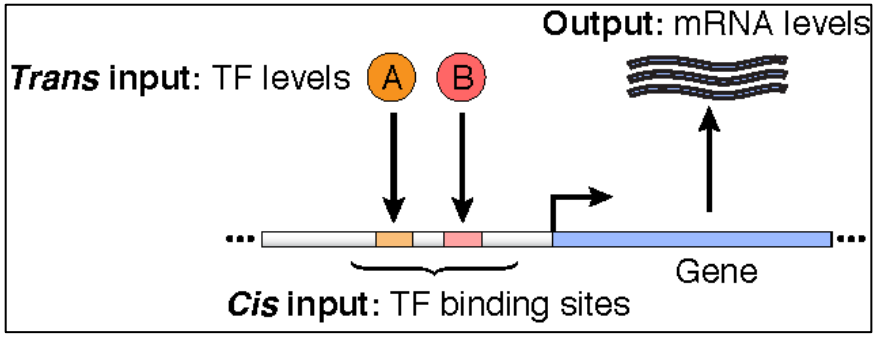
4. **Network Structure Analysis**



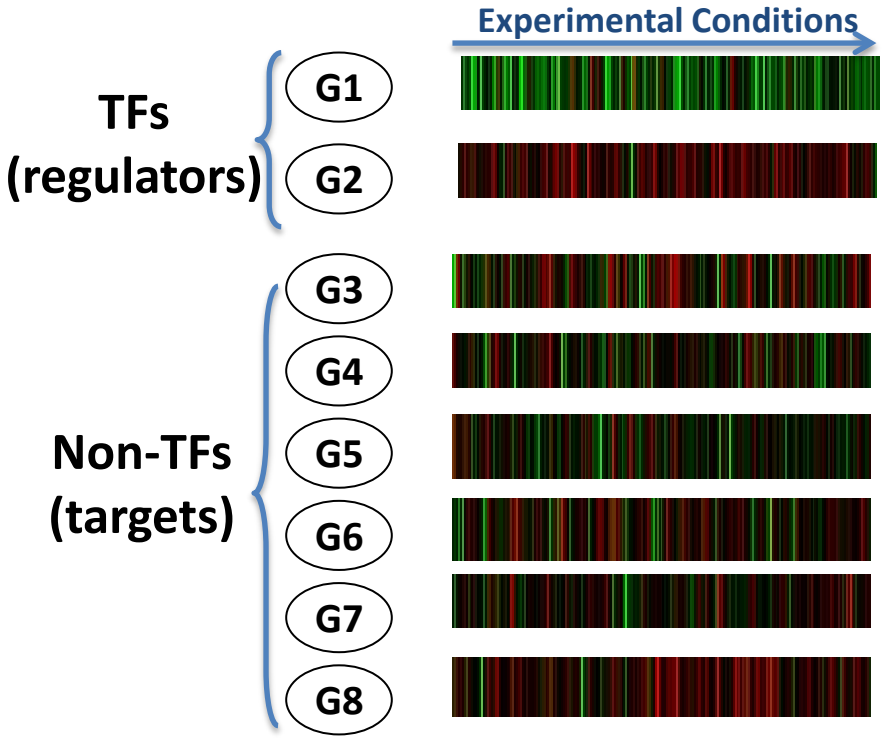

Hubs (degree-distribution)
Network motifs
Functional modules

# Physical and Relevance Networks

- **Physical Networks:**
  - edges represent "physical interaction" among nodes
  - Example: physical regulatory networks

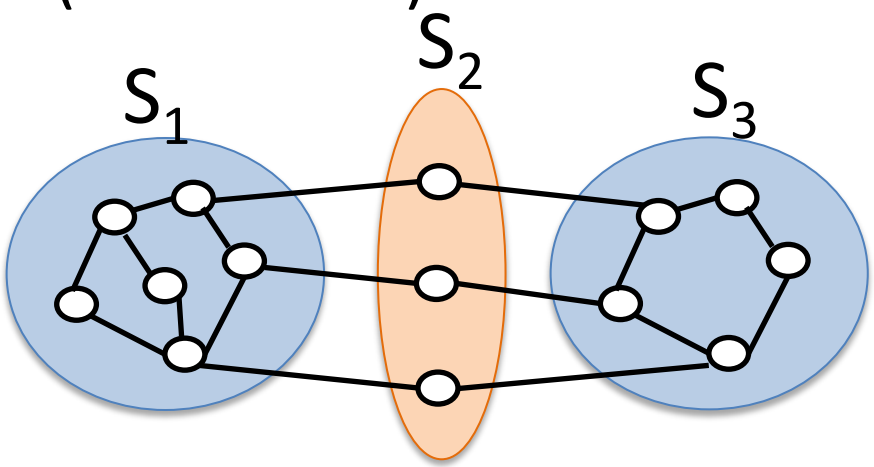



- **Relevance Networks:**
  - edge weights represent node similarities
  - Example: functional regulatory networks



# Probabilistic networks and graphical model

- There are several types of networks, with different meanings, and different applications
- Networks as graphical models:
  - modeling joint probability distribution of variables using graphs
  - Bayesian networks (directed), Markov Random Fields (undirected)



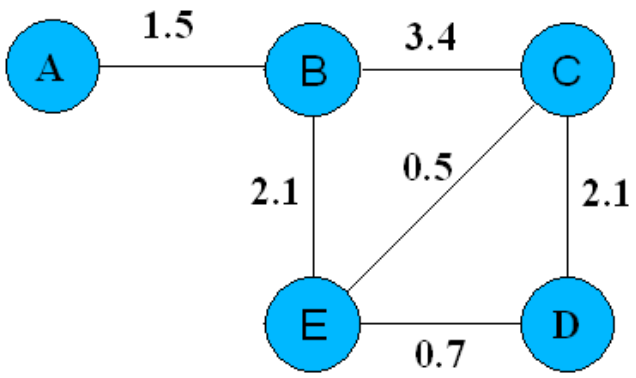

$$X_{S_1} \perp\!\!\!\perp X_{S_3} | X_{S_2}$$

# Goals for today: Network analysis

1. **Introduction to networks**
   - Network types: regulatory, metab., signal., interact., func.
   - Bayesian (probabilistic) and Algebraic views
2. **Network Centrality Measures**
   - Local centrality metrics (degree, betweenness, closeness, etc)
   - Global centrality metrics (eigenvector centrality, page-rank)
3. **Linear Algebra Review: eigenvalues, SVD, low-rank approximations**
   - Eigenvector and singular vector decomposition
   - Low rank approximations, Wigner semicircle law
4. **Sparse Principal Component Analysis**
   - Lasso and Elastic lasso
   - PCA and Sparse PCA
5. **Machine Learning in Networks**
   - Guilt by association
   - Maximum cliques, density-based modules and spectral clustering
6. **Network Diffusion Kernels and Deconvolution**
   - Network diffusion kernels
   - Network deconvolution

# Matrix representation of networks

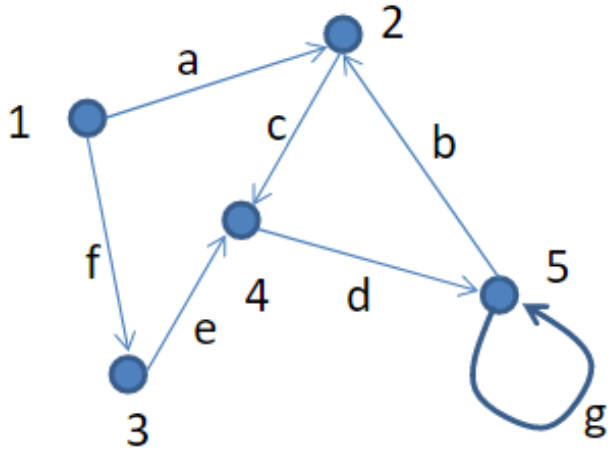
- A matrix representation of a network:
  - **Unweighted network:** binary adjacency matrix
  - **Weighted network:** real-valued matrix




| | A | B | C | D | E | ***Degree*** |
|---|---|---|---|---|---|---|
| A | 0 | 1.5 | 0 | 0 | 0 | *1.5* |
| B | 1.5 | 0 | 3.4 | 0 | 0 | *4.9* |
| C | 0 | 3.4 | 0 | 2.1 | 0.5 | *6* |
| D | 0 | 0 | 2.1 | 0 | 0.7 | *2.8* |
| E | 0 | 2.1 | 0.5 | 0.7 | 0 | *3.3* |

# Matrix interpretation of graphs

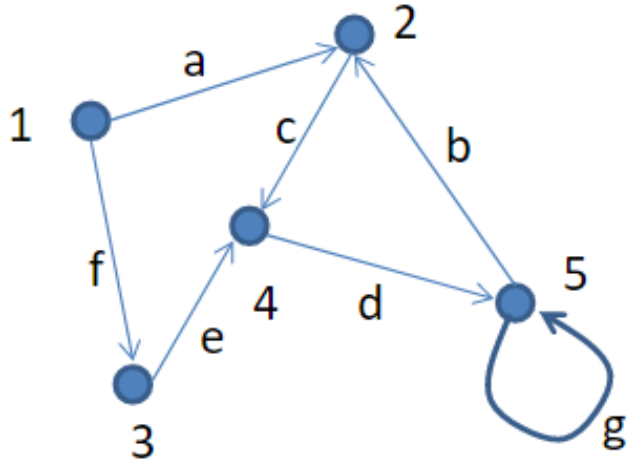
- Graph (V,E) as a matrix
  - Choose an ordering of vertices
  - Number them sequentially
  - Fill in |V|x|V| matrix
  - Called "incidence matrix" of graph
- Observations:
  - Diagonal entries: weights on self-loops
  - Symmetric matrix ←→ undirected graph
  - Lower triangular matrix ←→ no edges from lower numbered nodes to higher numbered nodes
  - Dense matrix ←→ clique (edge between every pair of nodes)




| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | a | f | 0 | 0 |
| 2 | 0 | 0 | 0 | c | 0 |
| 3 | 0 | 0 | 0 | e | 0 |
| 4 | 0 | 0 | 0 | 0 | d |
| 5 | 0 | b | 0 | 0 | g |

# Matrix operations on graphs

- Matrix computation: y = Ax
- Graph interpretation:
  - Each node i has two values (labels) x(i) and y(i)
  - Each node i updates its label y using the x value from each of its neighbors j, scaled by the label on edge (i,j)
- Observation:
  - Graph perspective shows dense MVM is just a special case of sparse MVM




| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | a | f | 0 | 0 |
| 2 | 0 | 0 | 0 | c | 0 |
| 3 | 0 | 0 | 0 | e | 0 |
| 4 | 0 | 0 | 0 | 0 | d |
| 5 | 0 | b | 0 | 0 | g |

A

# Linear Algebra: Eigenvectors/Eigenvalues

- **Eigenvectors** (for a square $m \times m$ matrix $\mathbf{S}$)

$$\mathbf{S}\mathbf{v} = \lambda\mathbf{v}$$

(right) eigenvector $\mathbf{v} \in \mathbb{R}^m \neq \mathbf{0}$

eigenvalue $\lambda \in \mathbb{R}$

*Example*

$$\begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

- How many eigenvalues are there at most?

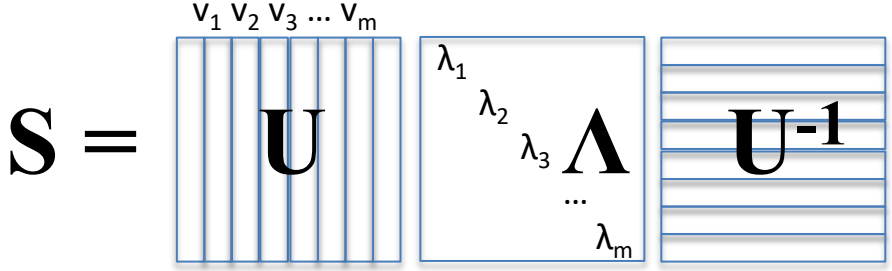
$$\mathbf{S}\mathbf{v} = \lambda\mathbf{v} \iff (\mathbf{S} - \lambda\mathbf{I})\,\mathbf{v} = \mathbf{0}$$

only has a non-zero solution if $|\mathbf{S} - \lambda\mathbf{I}| = 0$

this is a $m$-th order equation in $\lambda$ which can have **at most $m$ distinct solutions** (roots of the characteristic polynomial) – can be complex even though **S** is real.

# Eigen/diagonal Decomposition

- Let $\mathbf{S} \in \mathbb{R}^{m \times m}$ be a **square** matrix with ***m*** **linearly independent eigenvectors** (a "non-defective" matrix)




- **Theorem**: Exists an **eigen decomposition**

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$$

*diagonal*

Unique for distinct eigen-values

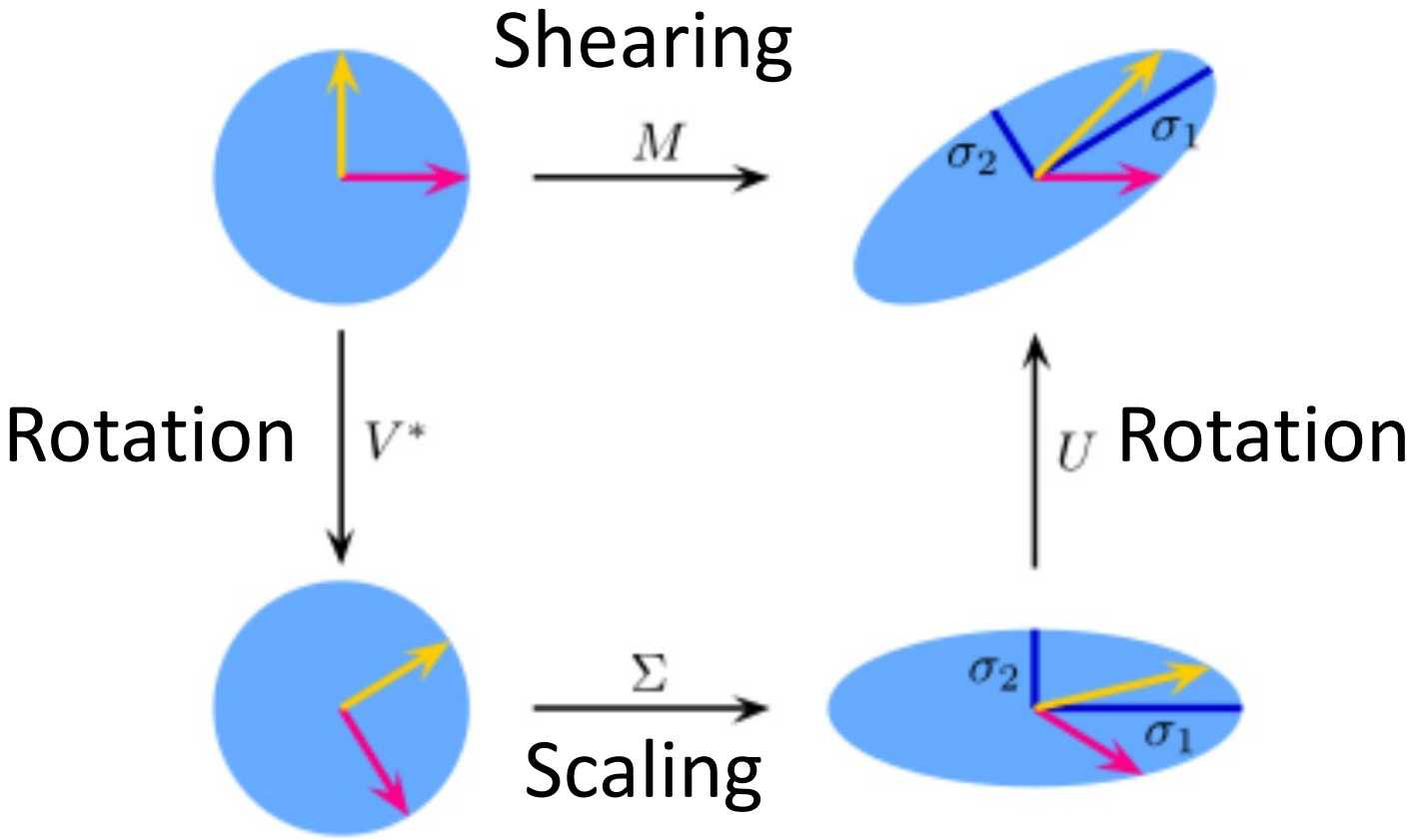
  - (cf. matrix diagonalization theorem)

- Columns of ***U*** are **eigenvectors** of ***S***

- Diagonal elements of $\mathbf{\Lambda}$ are **eigenvalues** of $\mathbf{S}$

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \quad \lambda_i \geq \lambda_{i+1}$$

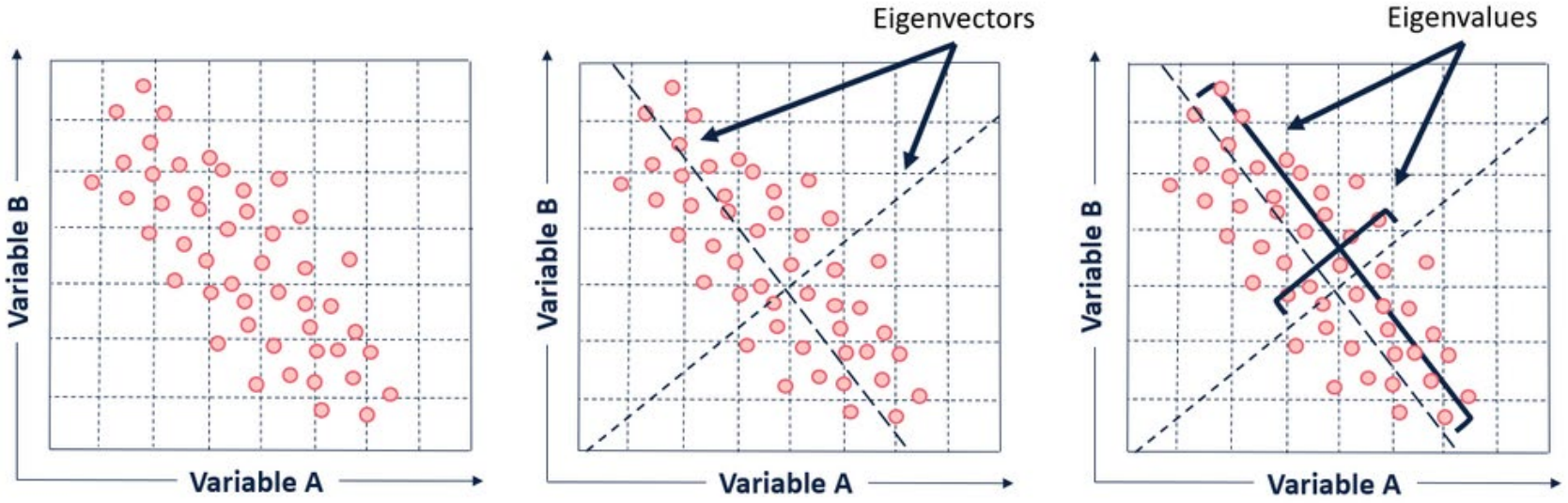
# Geometric interpretation of SVD




$$M = U \cdot \Sigma \cdot V^*$$

Mx = M(x) = U( S( V*(x) ) )

# PCA and Eigenvectors/Eigenvalues




**1. Covariance Matrix:**

First, compute covariance matrix of data: capture variance and relationships between variables in the data.

**2. Eigenvalues and Eigenvectors:**

Next, find eigenvalues and eigenvectors of covariance matrix.

•**Eigenvectors**: Directions of maximum variation ⇔ principal components

•**Eigenvalues**: Amount of variance in direction of corresponding eigenvectors: How "spread out" the data is in direction. Larger ⇔ more variance.

**3. Dimensionality Reduction:**

Reduce dimensionality of data while retaining as much variance as possible:

- Sort eigenvalues in descending order, pick top *k* eigenvalues (*k*: number of dimensions to keep).
- Use eigenvectors corresponding to these top *k* eigenvalues.
- Projecting data onto these eigenvectors.

**4. Transformation:**

- Project data onto new space defined by top *k* eigenvectors: dot product of data points with eigenvectors

# Singular Value Decomposition

For an $m \times n$ matrix $\mathbf{A}$ of rank $r$ there exists a factorization (Singular Value Decomposition = **SVD**) as follows:

$$A = U\Sigma V^T$$

$m \times m$ $m \times n$ $V$ is $n \times n$

The columns of $\boldsymbol{U}$ are orthogonal eigenvectors of $\boldsymbol{AA^T}$.

The columns of $\boldsymbol{V}$ are orthogonal eigenvectors of $\boldsymbol{A^TA}$.

Eigenvalues $\lambda_1 \ldots \lambda_r$ of $\boldsymbol{AA^T}$ are the eigenvalues of $\boldsymbol{A^TA}$.

$$\sigma_i = \sqrt{\lambda_i}$$

$$\Sigma = diag(\sigma_1 \ldots \sigma_r)$$ ← *Singular values.*

# Low-rank Approximation

- SVD can be used to compute optimal **low-rank approximations**.
- Approximation problem: Find $\boldsymbol{A_k}$ of rank $\boldsymbol{k}$ such that

$$A_k = \min_{X:rank(X)=k} \|A - X\|_F$$

*Frobenius norm (aka Euclidian norm)*

$$\|A\|_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}.$$

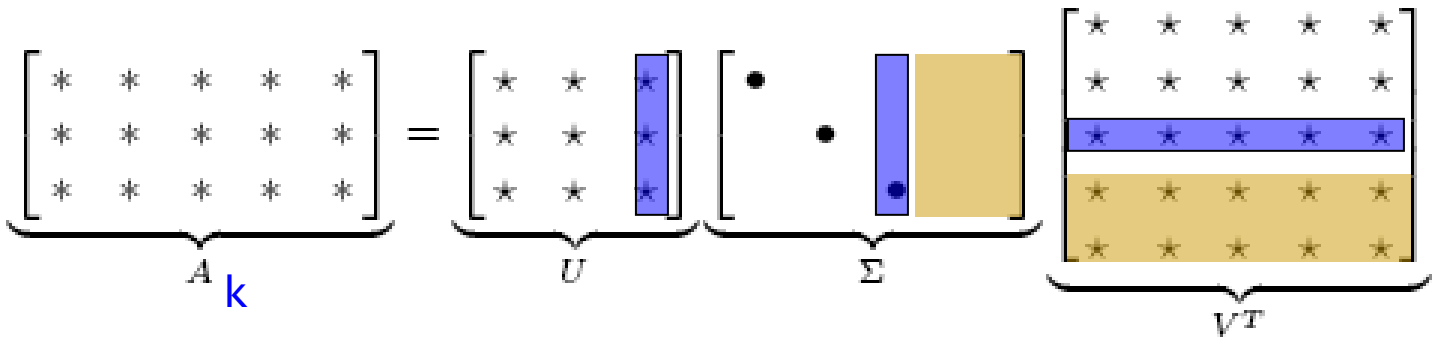
$A_k$ and $X$ are both $m \times n$ matrices.

Typically, want $k << r$.

# Low-rank Approximation

- Solution via SVD

$$A_k = U\ \mathrm{diag}(\sigma_1,\ldots,\sigma_k,\underbrace{0,\ldots,0}_{\text{set smallest r-k singular values to zero}})V^T$$




$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$ ← *column notation: sum of rank 1 matrices*

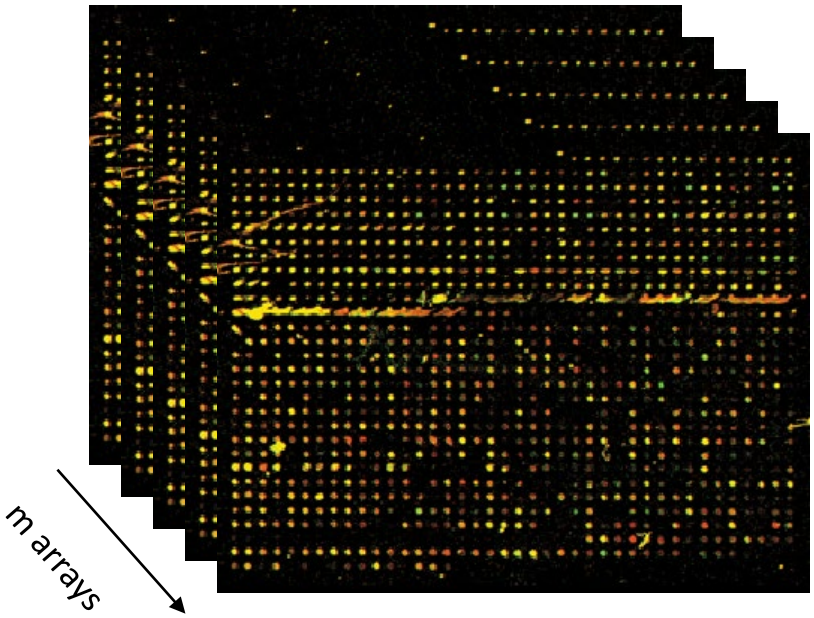
- Error: $\min_{X:rank(X)=k} \|A - X\|_F = \|A - A_k\|_F = \sigma_{k+1}$

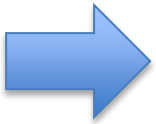
# Goals for today: Network analysis

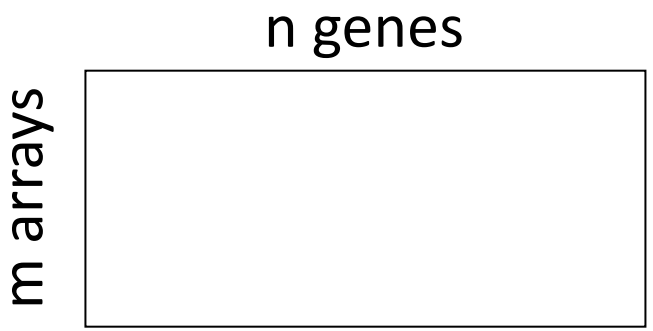
1. **Introduction to networks**
   - Network types: regulatory, metab., signal., interact., func.
   - Bayesian (probabilistic) and Algebraic views
2. **Network Centrality Measures**
   - Local centrality metrics (degree, betweenness, closeness, etc)
   - Global centrality metrics (eigenvector centrality, page-rank)
3. **Linear Algebra Review: eigenvalues, SVD, low-rank approximations**
   - Eigenvector and singular vector decomposition
   - Low rank approximations, Wigner semicircle law
4. **Sparse Principal Component Analysis**
   - Lasso and Elastic lasso
   - PCA and Sparse PCA
5. **Machine Learning in Networks**
   - Guilt by association
   - Maximum cliques, density-based modules and spectral clustering
6. **Network Diffusion Kernels and Deconvolution**
   - Network diffusion kernels
   - Network deconvolution

# Sparse Principal Component Analysis

Gene Expression Data
(RNA-Seq, Microarray, …)

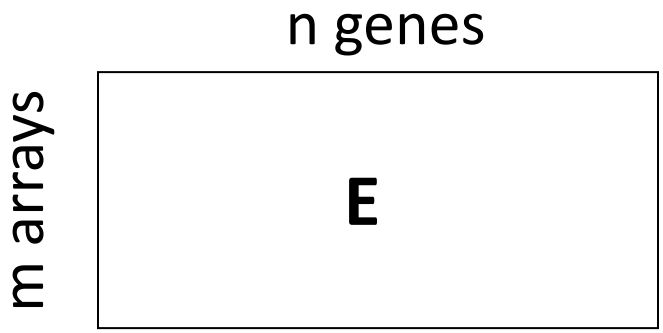


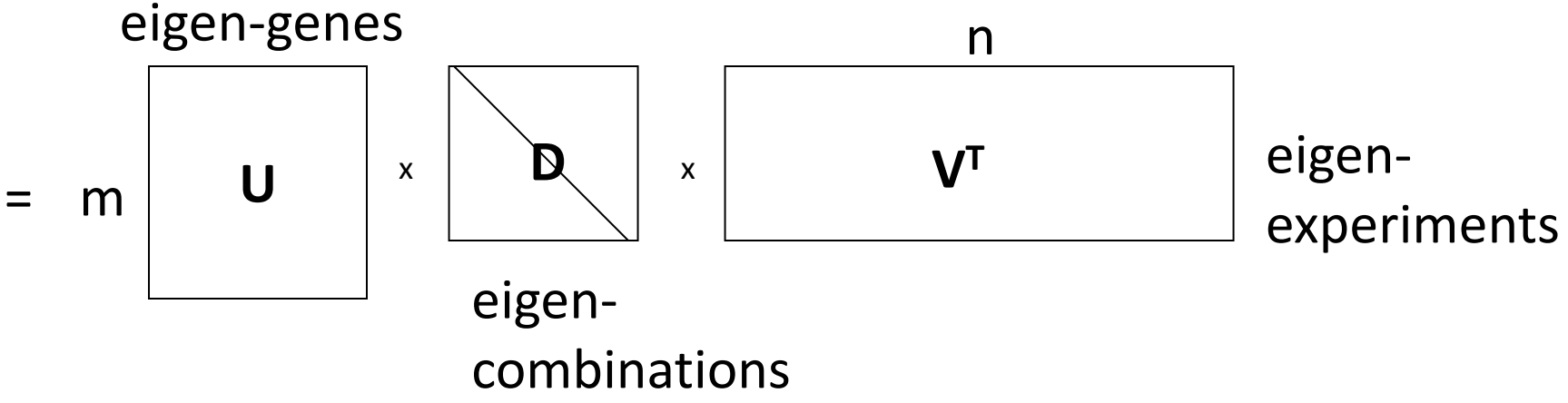



- n = 20k genes, m = 100 arrays
- n >> m

# Applications of eigenvector analysis: PCA on Expression Data

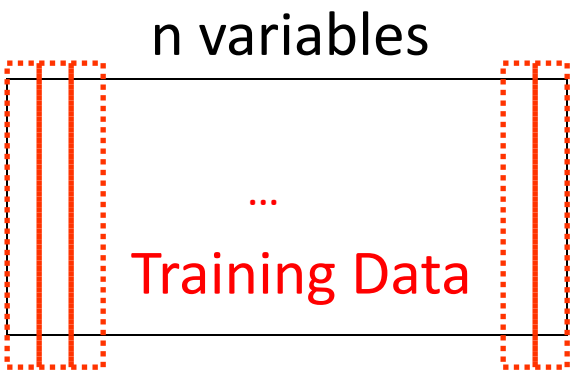



- Each eigen-gene is expressed only in the corresponding eigen-array with the corresponding eigen-expression level.

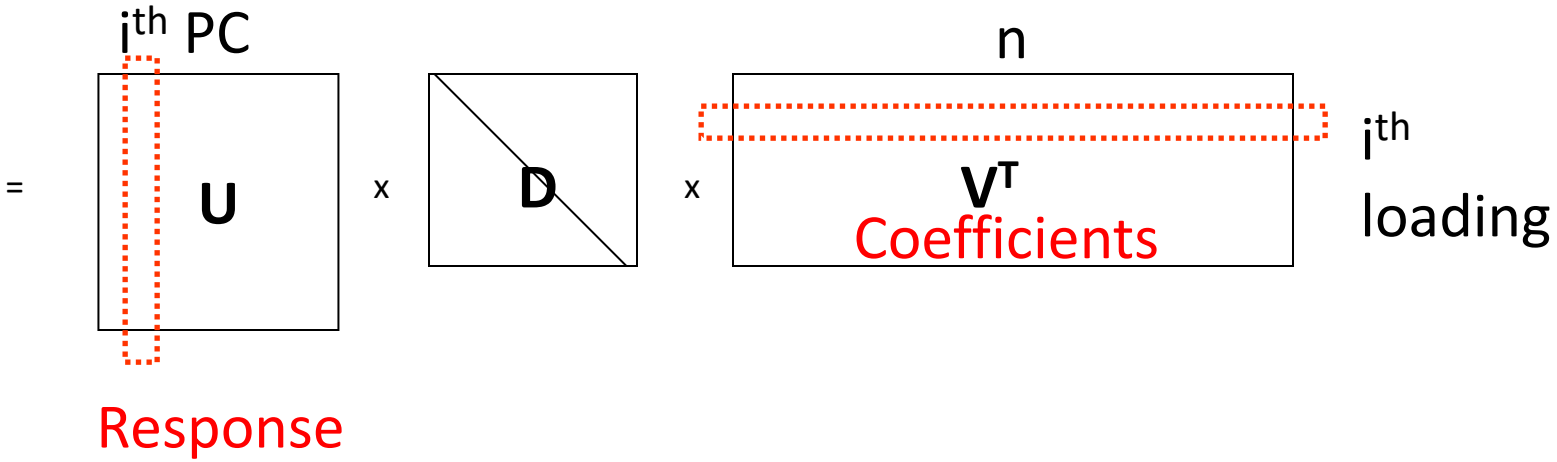



(Alter et al., 2000)

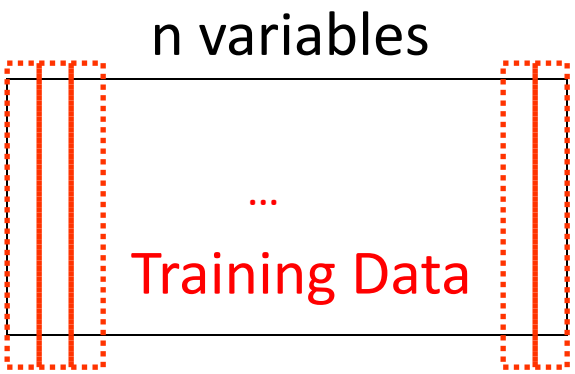
# Drawbacks of PCA



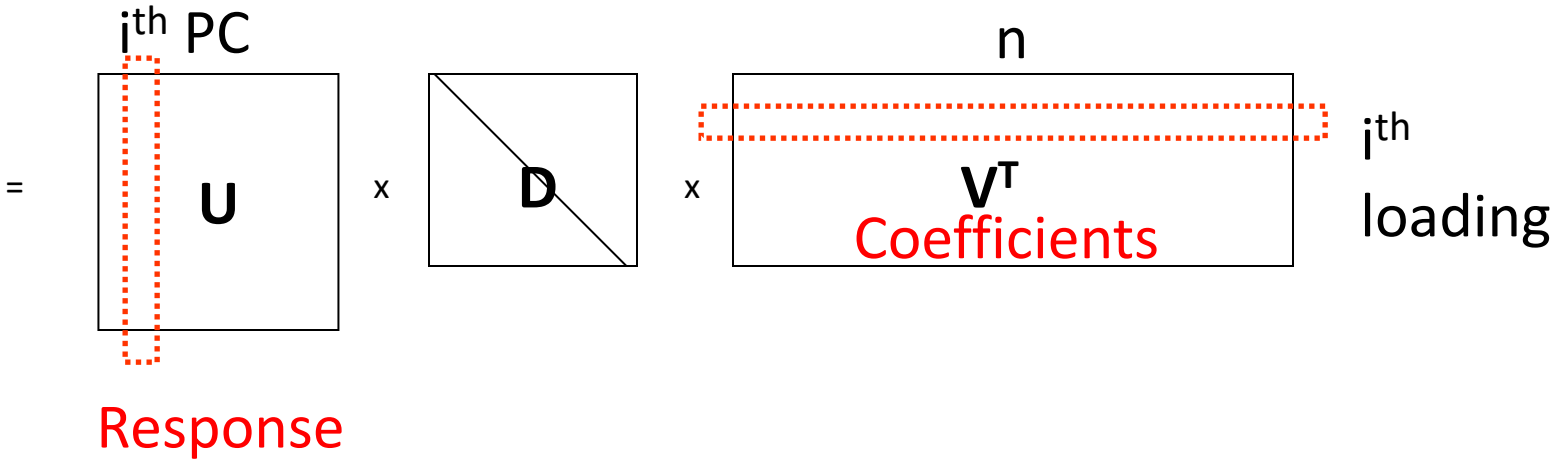

Each PC is a linear combination of the n variables → Difficult to interpret



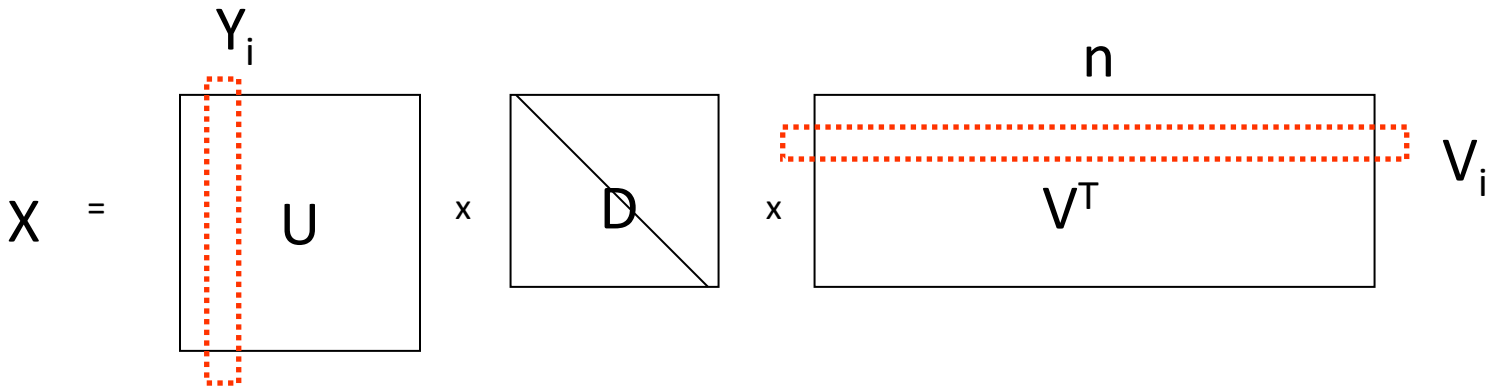
# Reconstruction of PCA in a Regression Framework




Each PC is a linear combination of the n variables. Its loadings can be recovered by regressing PC on the n variables.



# Theorem- solving PCA by regression




Let $X = UDV^T$.

$\forall i$, denote $Y_i = U_i D_{i,i}$. $Y_i$ is the i$^{th}$ principal component of X.

$\forall \lambda > 0$, suppose $\hat{\beta}_{ridge}$ is the ridge estimate given by

$$\hat{\beta}_{ridge} = \arg\min_{\beta} | Y_i - X\beta |^2 + \lambda | \beta |^2 .$$

Let $\hat{v} = \dfrac{\hat{\beta}_{ridge}}{| \hat{\beta}_{ridge} |}$, then $\hat{v} = V_{i.}$

# Reconstruction of PCA in a Regression Framework

- By theorem 1, we can reconstruct the loadings of PCs exactly by a linear regression problem.
  - not an alternative to PCA as it uses its results.

- Ridge penalty does not penalize the coefficients, but ensure the reconstruction of PCs.

- **Sparse PCA:** add lasso penalty to the problem to penalize for the absolute values of coefficients.

# Sparse PCA

- Idea: Reduce the number of explicitly used variables (genes).
- Approach: Modify PCA so that PCs have sparse loadings = sparse PCA (SPCA)
- Writes PCA as a regression-type optimization problem.
- Uses lasso (Least Absolute Shrinkage and Selection Operator)
  - Tibshirani 1996
  - Both variable selection and regularization
  - Produces sparse models
- Result: Modified PCs with sparse loadings.

# Construction of SPCA in a Regression Framework

Let $X = UDV^T$.

$\forall i$, denote $Y_i = U_i D_{i,i}$. $Y_i$ is the $\mathrm{i}^{\mathrm{th}}$ principal component of X.

Solve $\hat{\beta} = \arg\min_{\beta} | Y_i - X\beta |^2 + \lambda | \beta |^2 + \lambda_1 | \beta |$.

$$\hat{V}_i = \frac{\hat{\beta}}{| \hat{\beta} |} \approx V_i$$

$X\hat{V}_i \approx Y_i = i^{th}$ principal component

$\hat{V}_i =$ sparse loading, $X\hat{V}_i = \mathrm{i}^{\mathrm{th}}$ sparse principal component

# Simulation Example: PCA vs. SPCA

- Data points $X = (X_1, X_2, \ldots, X_{10})$
  - 10 variables

- Model to generate data:
  - 3 hidden factors: $V_1, V_2, V_3$
  - $V_1 \sim N(0, 290)$
  - $V_2 \sim N(0, 300)$
  - $V_3 = -0.3\, V_1 + 0.925\, V_2 + e,\ e \sim N(0, 1)$

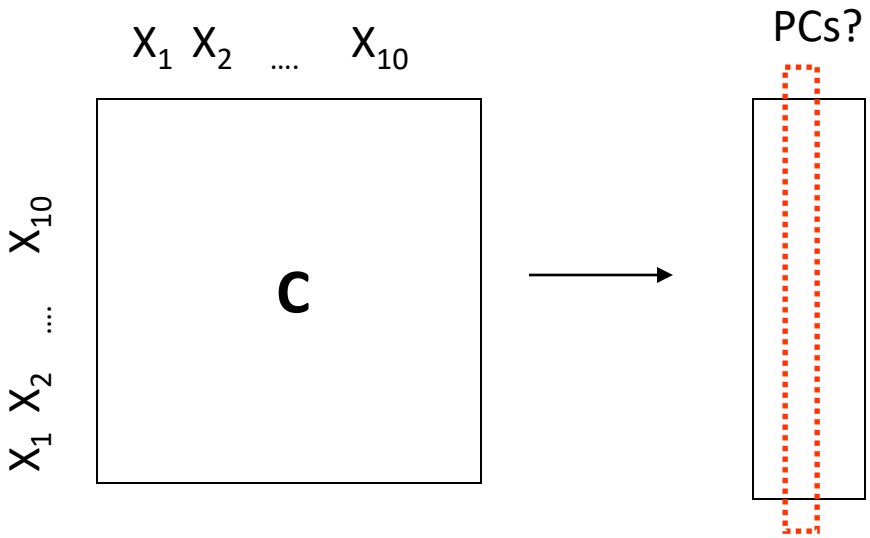
  } variances: 290,300,298

  - $X_i = V_1 + e_i^1,\ e_i^1 \sim N(0, 1),\ i=1,2,3,4$
  - $X_i = V_2 + e_i^2,\ e_i^2 \sim N(0, 1),\ i=5,6,7,8$
  - $X_i = V_3 + e_i^3,\ e_i^3 \sim N(0, 1),\ i=9,10$

  } 4 vars associated with $V_1$
  4 vars associated with $V_2$
  2 vars associated with $V_3$

# Simulation Example: PCA vs. SPCA

- How many observations?
  - PCA and SPCA performed on exact covariance matrix. => infinitely many data points.




  - We expect to derive 2 PCs with right sparse loadings:
    - One from ($X_5$, $X_6$, $X_7$, $X_8$) recovering $V_2$
    - One from ($X_1$, $X_2$, $X_3$, $X_4$) recovering $V_1$

# Table of Loadings

| | PCA | | | SPCA ($\lambda = 0$) | |
|---|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC1 | PC2 |
| $X_1$ | 0.116 | -0.478 | -0.087 | 0.0 | 0.5 |
| $X_2$ | 0.116 | -0.478 | -0.087 | 0.0 | 0.5 |
| $X_3$ | 0.116 | -0.478 | -0.087 | 0.0 | 0.5 |
| $X_4$ | 0.116 | -0.478 | -0.087 | 0.0 | 0.5 |
| $X_5$ | -0.395 | -0.145 | 0.270 | 0.5 | 0.0 |
| $X_6$ | -0.395 | -0.145 | 0.270 | 0.5 | 0.0 |
| $X_7$ | -0.395 | -0.145 | 0.270 | 0.5 | 0.0 |
| $X_8$ | -0.395 | -0.145 | 0.270 | 0.5 | 0.0 |
| $X_9$ | -0.401 | 0.010 | -0.582 | 0.0 | 0.0 |
| $X_{10}$ | -0.401 | 0.010 | -0.582 | 0.0 | 0.0 |
| Adjusted Variance (%) | 60.0 | 39.6 | 0.08 | 40.9 | 39.5 |

$V_1$

$V_2$

# Linear Regression Problem

- Input variables $x = (1, x_1, \ldots, x_p)$
- Response variable $y \equiv f(x) + \varepsilon$
- Regression coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$

- Multivariate linear model

$$f(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = x\beta$$

# Linear Regression Problem

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

**Y** **X** **β** **ε**

Error

Training Data Coefficients

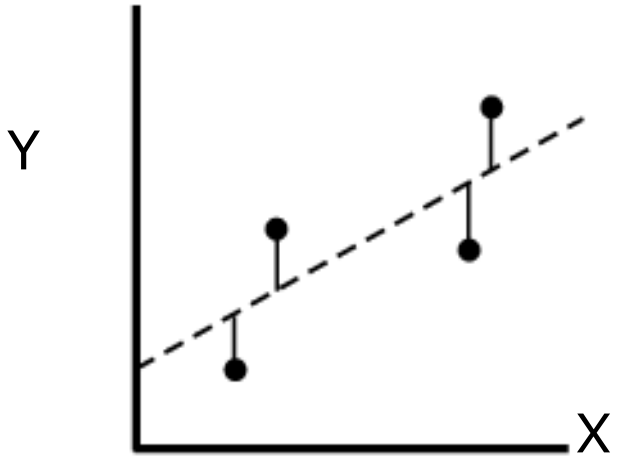
- N observations, p predictors.
- Goal: Estimate the coefficients, β.

# Least Squares Solution

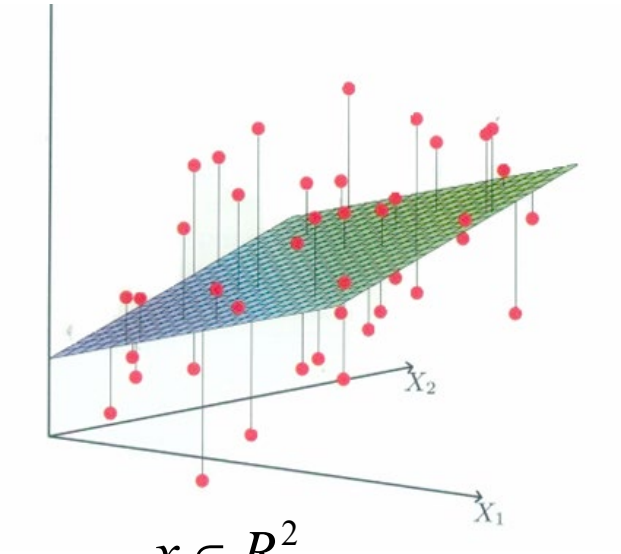
Training Set $D = \{(X_i, y_i)\}_{i=1}^{n}$

Residual Sum of Squares $RSS(\beta) \equiv \sum_{i=1}^{n} (y_i - \beta X_i)^2$

Objective: Find $\hat{\beta} = \arg\min_{\beta} \{RSS(\beta) \mid D\}$




$x \in R$




$x \in R^2$

# Sparse PCA solution with Lasso penalty

Training Set $D = \{(X_i, y_i)\}_{i=1}^{n}$

Residual Sum of Squares $RSS(\beta) \equiv \sum_{i=1}^{n} (y_i - \beta X_i)^2$

Lasso penalty $L_1(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|, \quad \lambda \geq 0$

Objective: Find $\hat{\beta}_{lasso} = \arg\min_{\beta} \{RSS(\beta) + L_1(\beta) \mid D\}$

- Pros:
  - Lasso continuously shrinks coefficients toward zero.
  - Produces a sparse model.
  - Variable selection method.
- Cons:
  - #selected variables limited by n, number of observations.

    e.g. microarray expression data n(arrays)<<p(genes)
  - Selects only one of the highly correlated variables, does not care which one is in the final model.

# L1 (Lasso) vs. L2 (Ridge) regularization

- L1 regularization (lasso)

Linear

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left( t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2 + \lambda \sum_{i=1}^{k} |w_i|$$

- L2 regularization (ridge)

Quadratic

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_j \left( t(\mathbf{x}_j) - \sum_i w_i h_i(\mathbf{x}_j) \right)^2 + \lambda \sum_{i=1}^{k} w_i^2$$

- L1 pros:
  - More robust (outliers don't matter more than small diffs)
- L1 cons:
  - Less stable gradient ascent
  - Multiple solutions

- L2 pros:
  - Always one solution
  - More stable gradient ascent
- L2 cons:
  - Less robust (square diffs, outliers have strong effect)

# Elastic Net Solution

$$\text{Lasso penalty } L_1(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j|, \quad \lambda_1 \geq 0$$

$$\text{Ridge penalty } L_2(\beta) = \lambda_2 \sum_{j=1}^{p} |\beta_j|^2, \quad \lambda_2 \geq 0$$

$$\text{Objective: Find } \hat{\beta}_{ridge} = \arg\min_{\beta} \{RSS(\beta) + L_2(\beta) + L_1(\beta) \mid D\}$$

- Pros:
  - Combines L1 (lasso) and L2 (ridge) regression
  - Limitation of lasso removed by ridge constraint. All variables are included in the model.
  - Grouping effect:
    - Selects a group of highly correlated variables once one variable among them is selected.

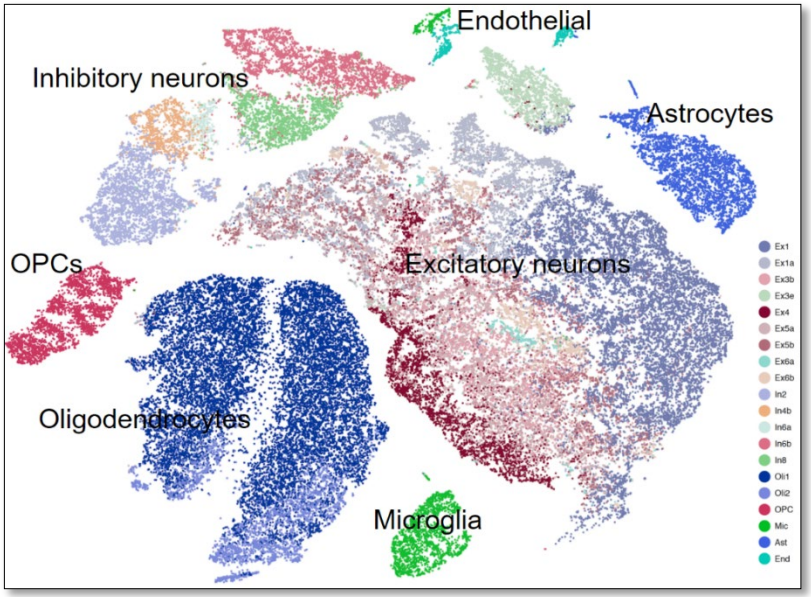
      (lasso selects only one of them, does not care which.)

# SPCA

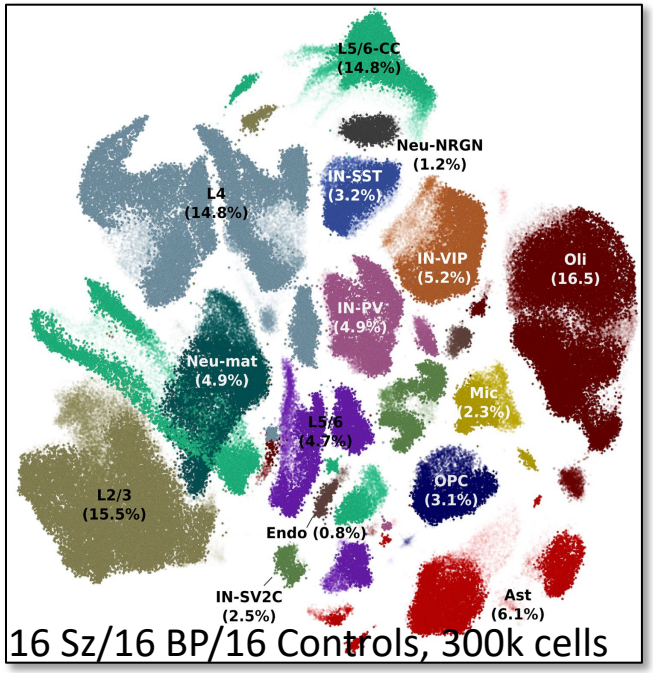
- Goal:
  - Construct a regression framework in which PCA can be reconstructed exactly.
  - Use lasso/ridge/elastic-net to construct modified PCs with sparse loadings.

# Non-linear embeddings

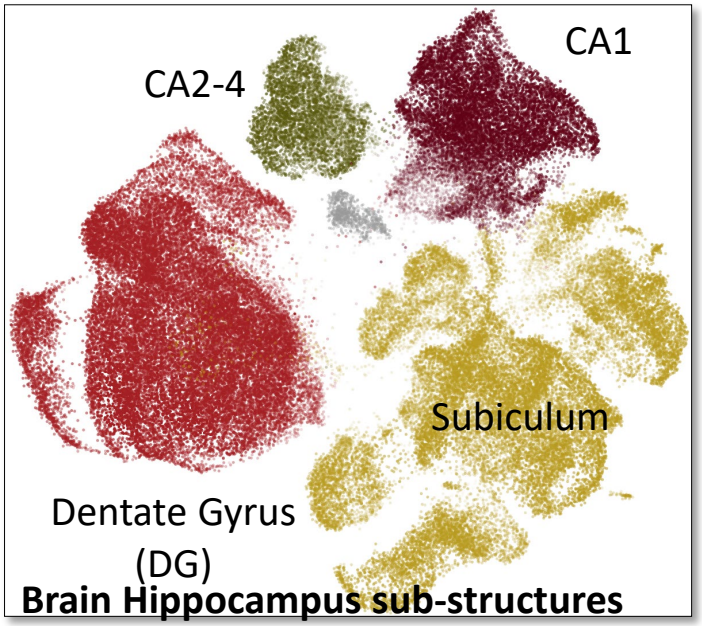
# t-SNEs of single-cell Brain data



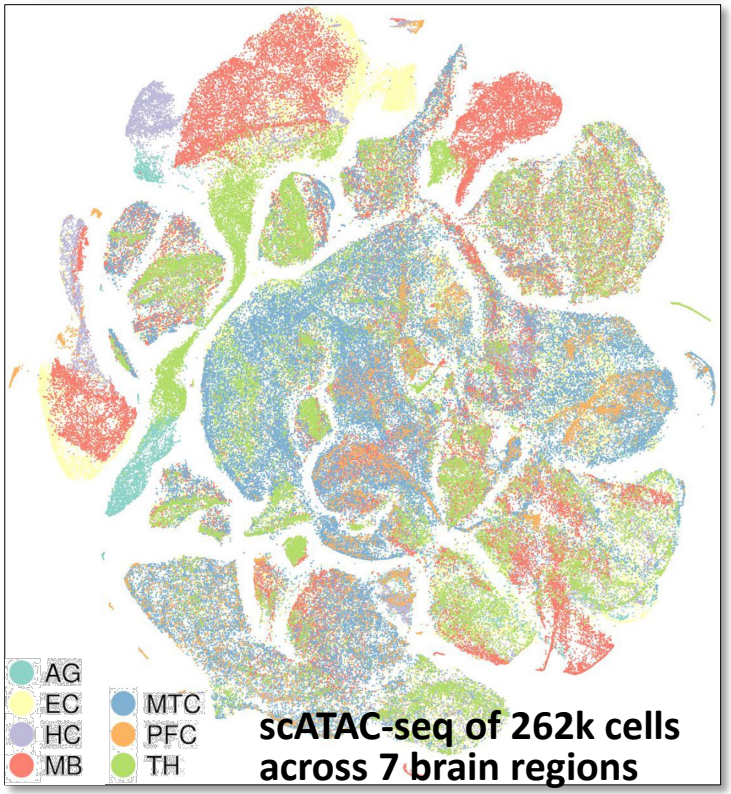

**scRNA-seq in 48 individuals, 84k cells, Nature, 2019**




16 Sz/16 BP/16 Controls, 300k cells

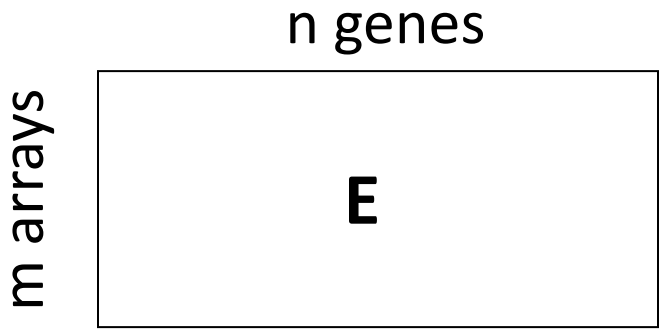



**Brain Hippocampus sub-structures**

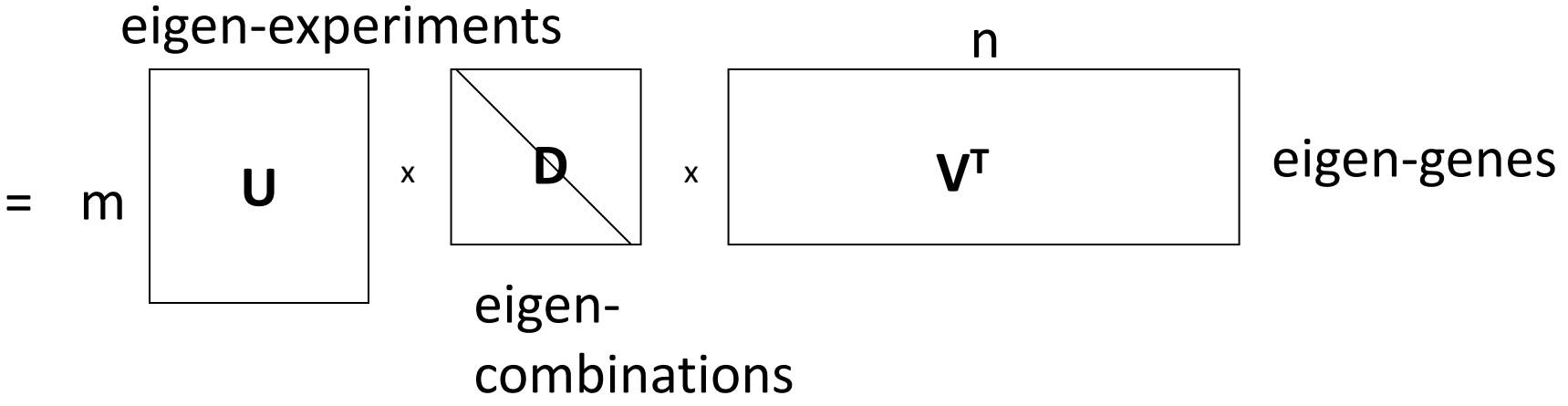



**scATAC-seq of 262k cells across 7 brain regions**

# PCA on Expression Data

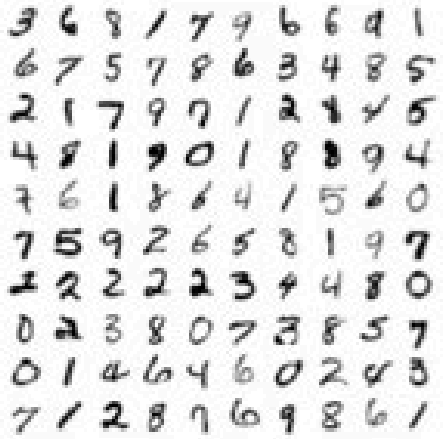


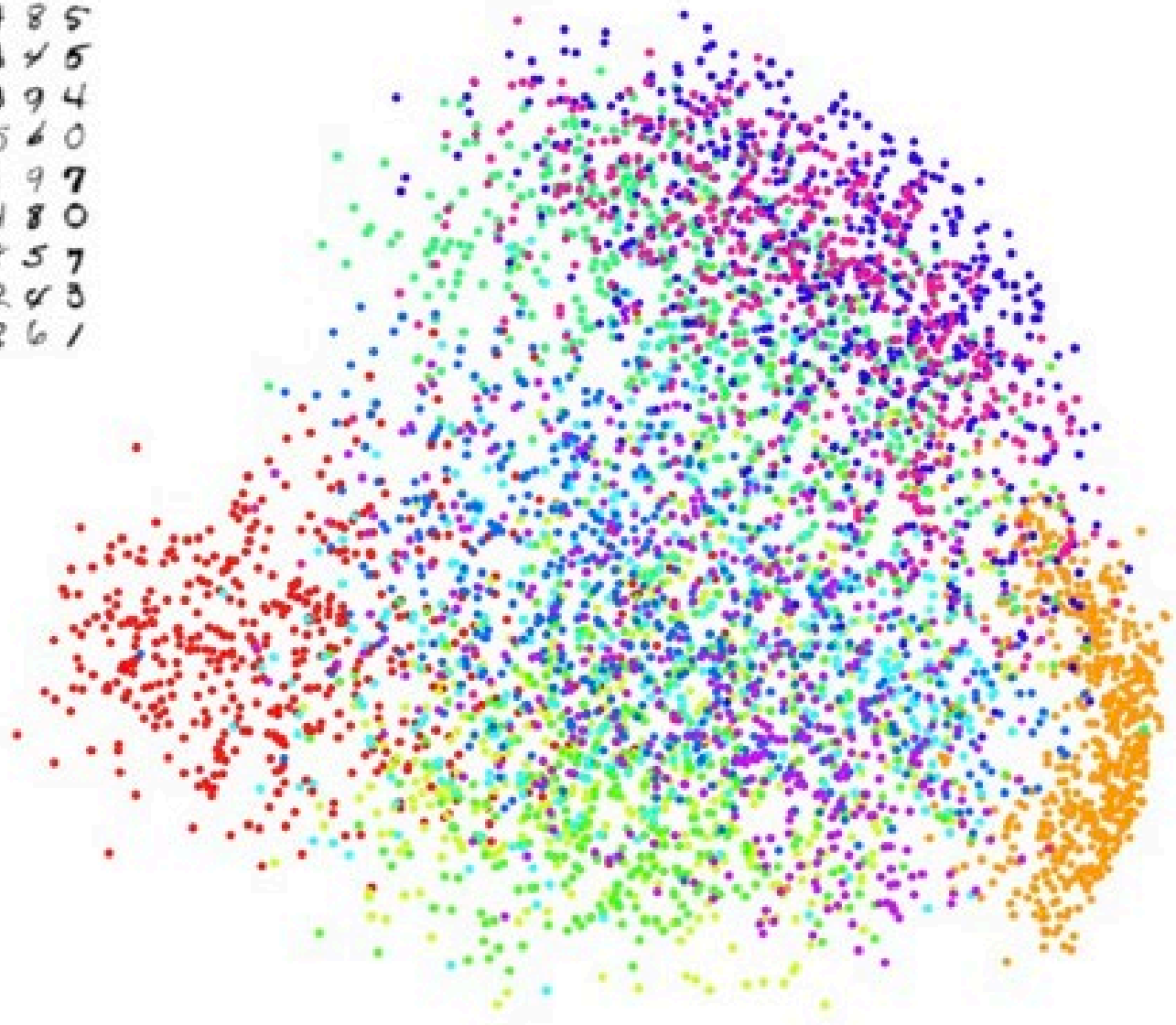

- Each eigen-gene is expressed only in the corresponding eigen-array with the corresponding eigen-expression level.



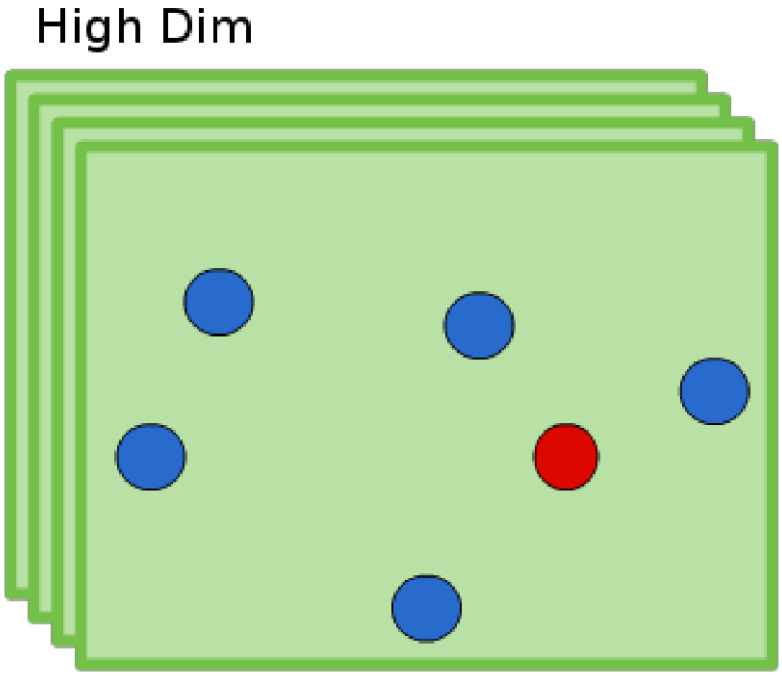

(Alter et al., 2000)

# PCA of MNIST handwritten-digits data

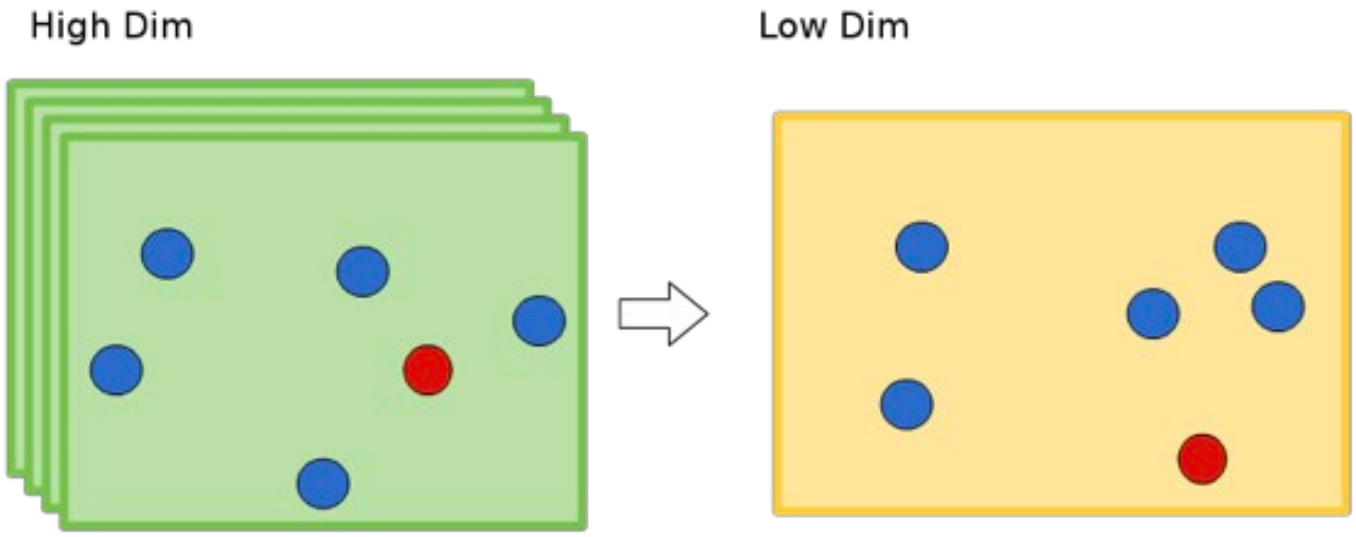




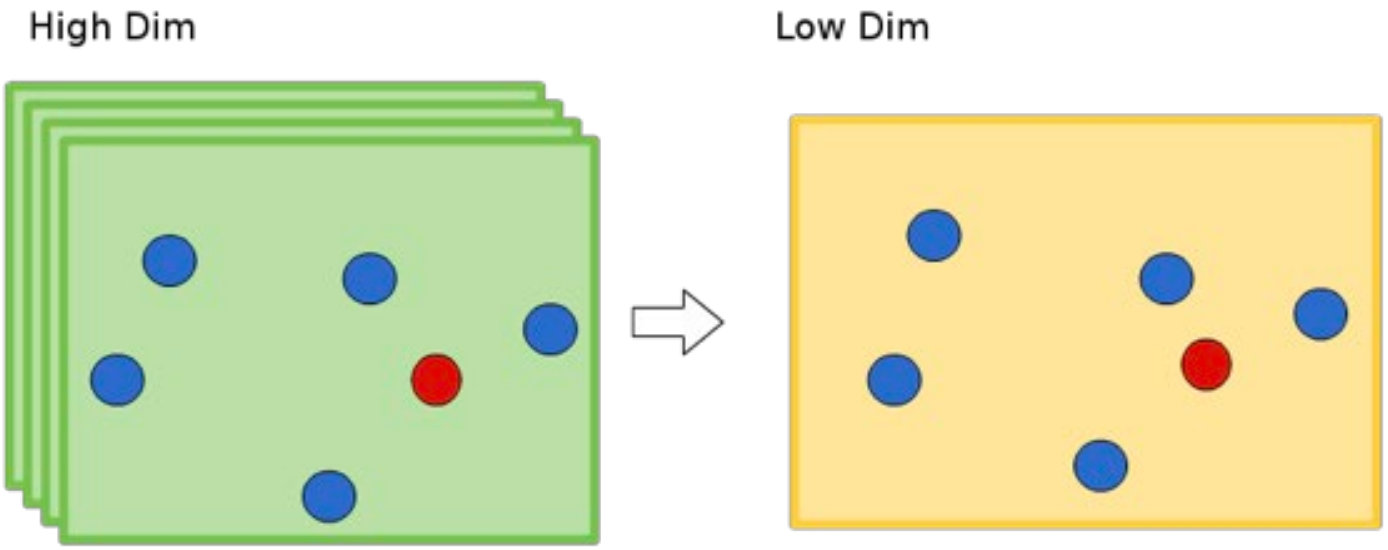
Distance Preservation Neighbor Preservation



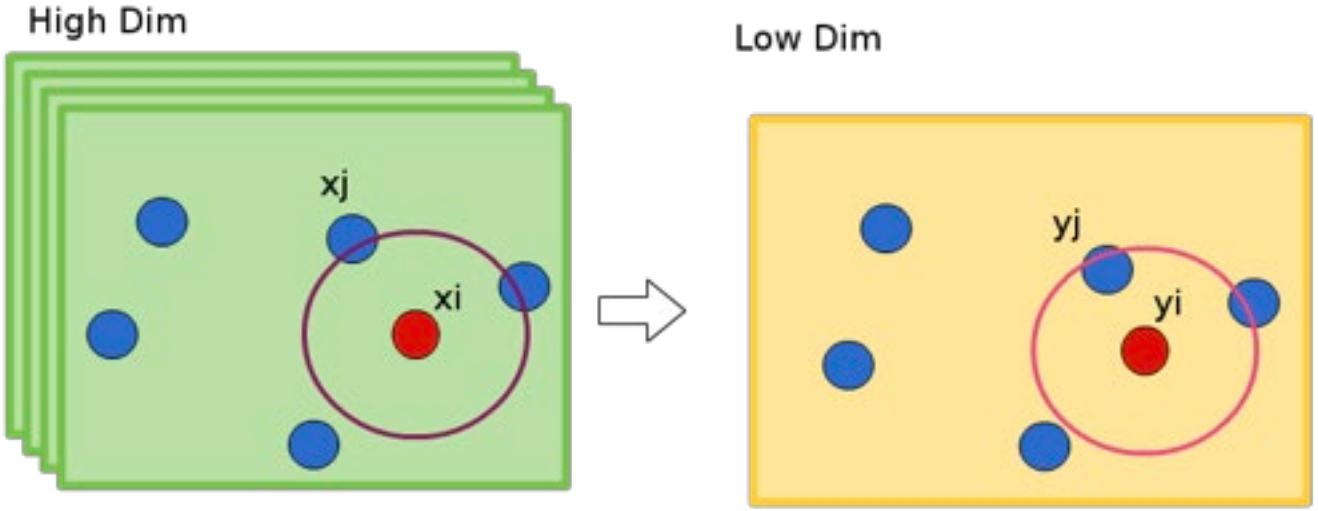
# Neighborhood not preserved



# Neighborhood preserved



# Measure pairwise distances in high dimensional space




$$p_{j|i} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

Set the bandwidth $\sigma_i$ such that the conditional has a fixed perplexity (effective number of neighbors) $Perp(P_i) = 2^{H(P_i)}$, typical value is about 5 to 50

Shannon entropy of $P_i$

# We want to choose an embedding that minimizes divergence between low and high dimension similarities

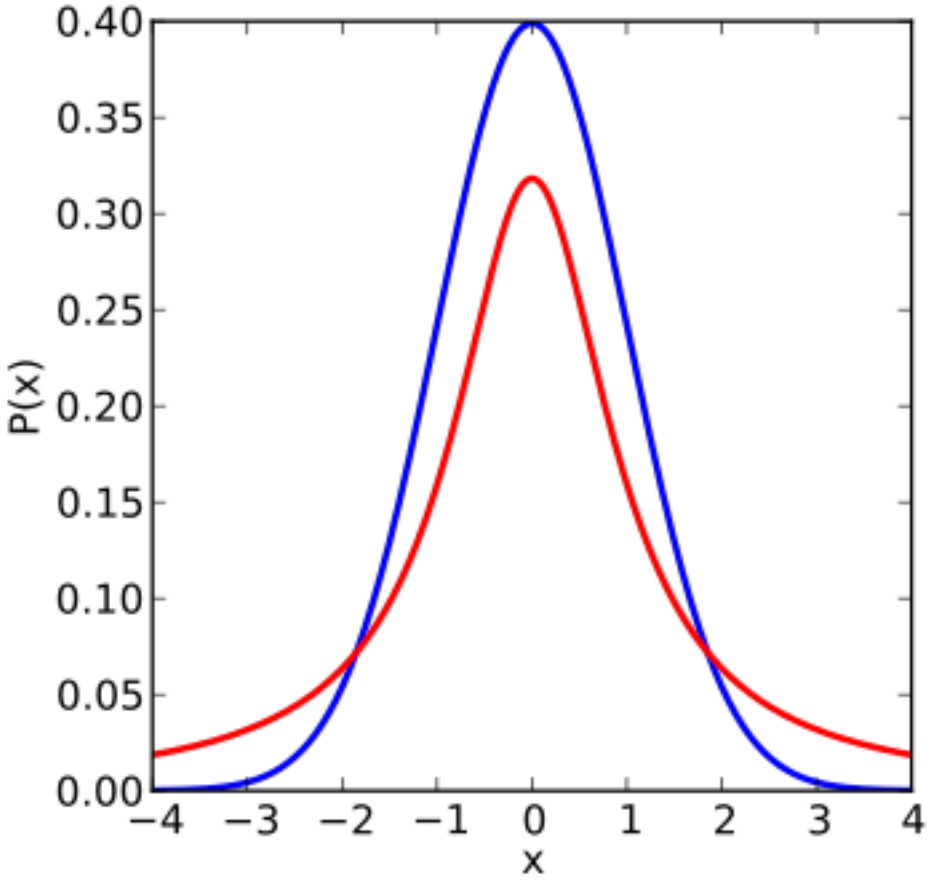
- Similarity of datapoints in High Dimension

$$p_{ij} = \frac{exp(-||x_i - x_j||^2/2\sigma^2)}{\sum_{k \neq l} exp(-||x_l - x_k||^2/2\sigma^2)}$$

- Similarity of datapoints in Low Dimension

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}$$

# Low dimensional embedding using a Student t-distribution to avoid overcrowding




Red – Student t-distribution (1 degree of freedom)

Blue - Gaussian

# We can use gradient methods to find an embedding

- Cost function

$p_{ij}$ = New (low) dimension distance

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

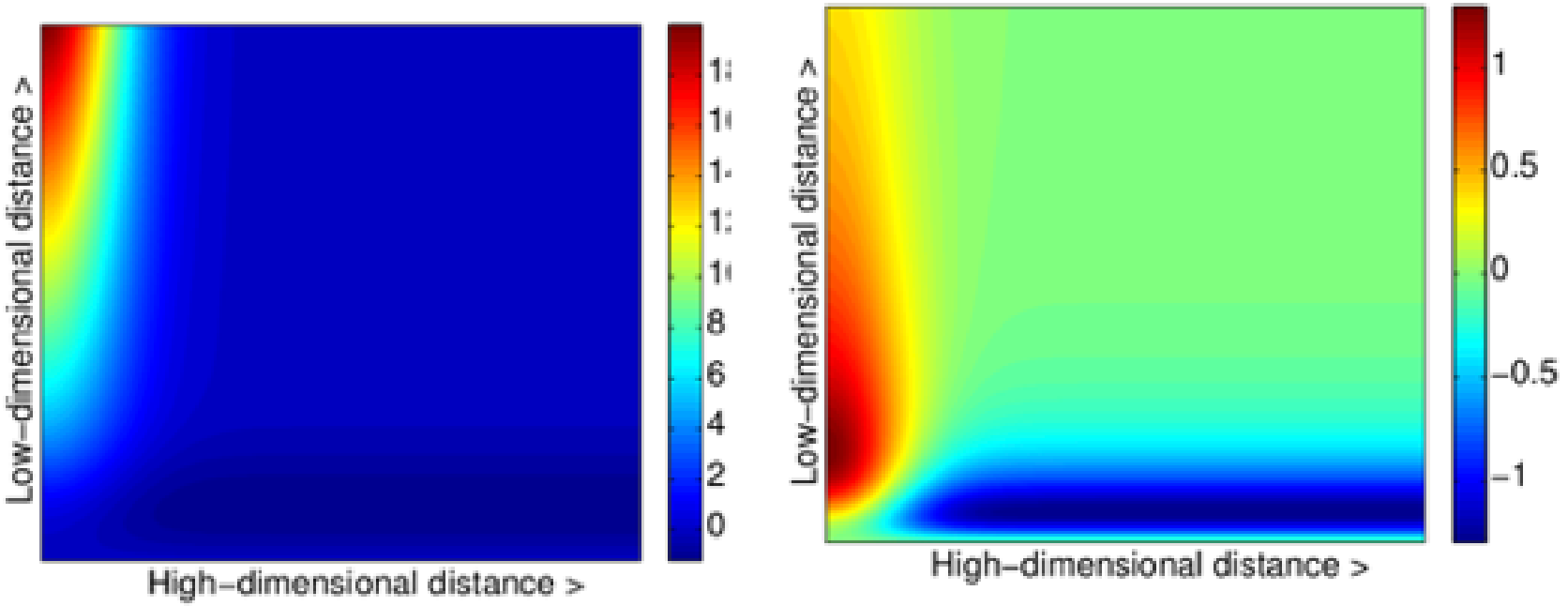
$q_{ij}$ = Original (high) dimension D

  - Large $p_{ij}$ modeled by small $q_{ij}$: Large penalty (not okay to bring distant points closer)
  - Small $p_{ij}$ modeled by large $q_{ij}$: Small penalty (okay to separate nearby points)
  - t-SNE mainly preserves local similarity structure of the data

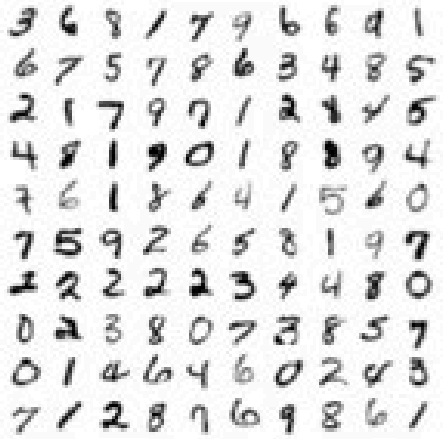
- Gradient

$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + ||y_i - y_j||^2)^{-1}(y_i - y_j)$$

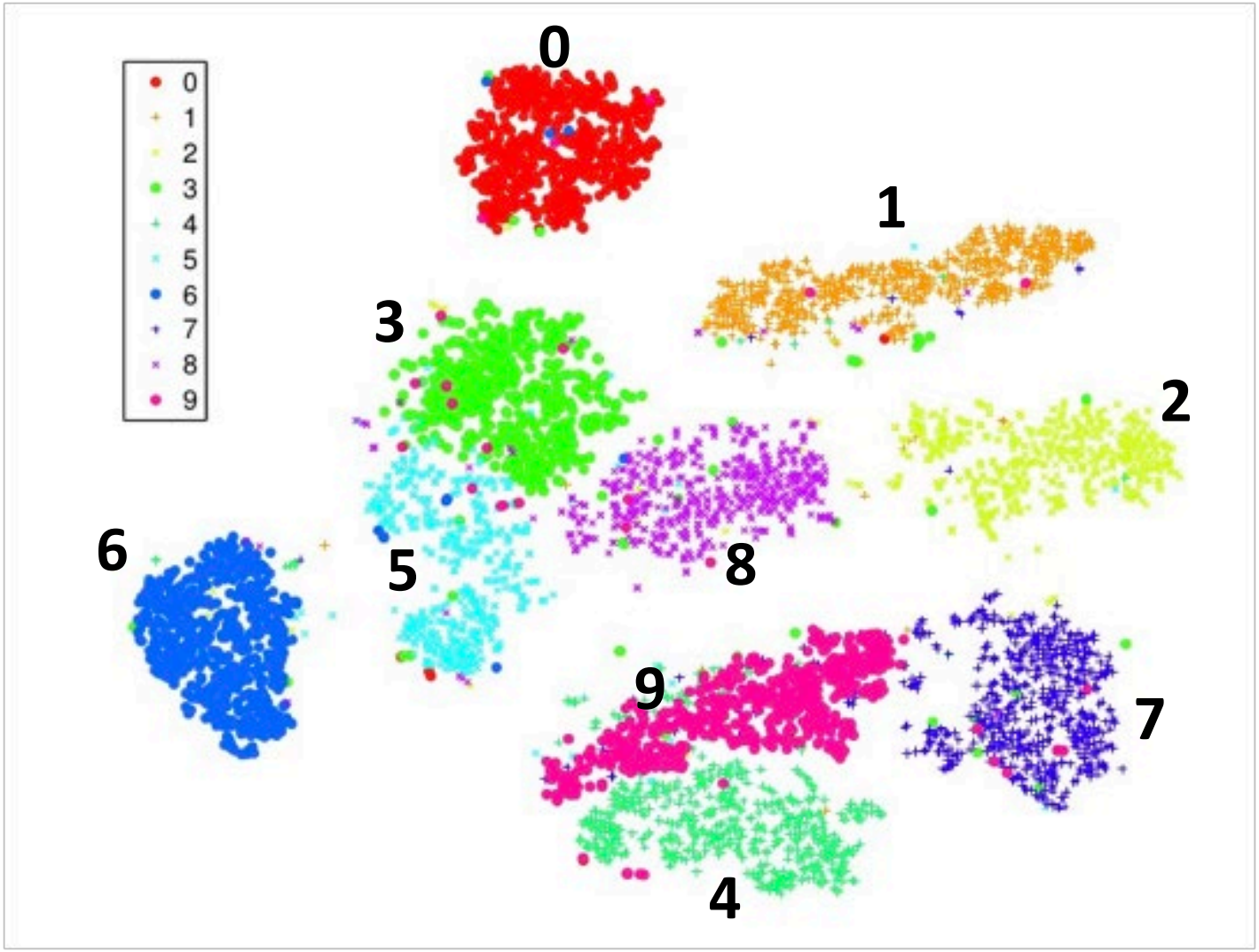
# Interpretation of SNE (left) and t-SNE (right) gradients



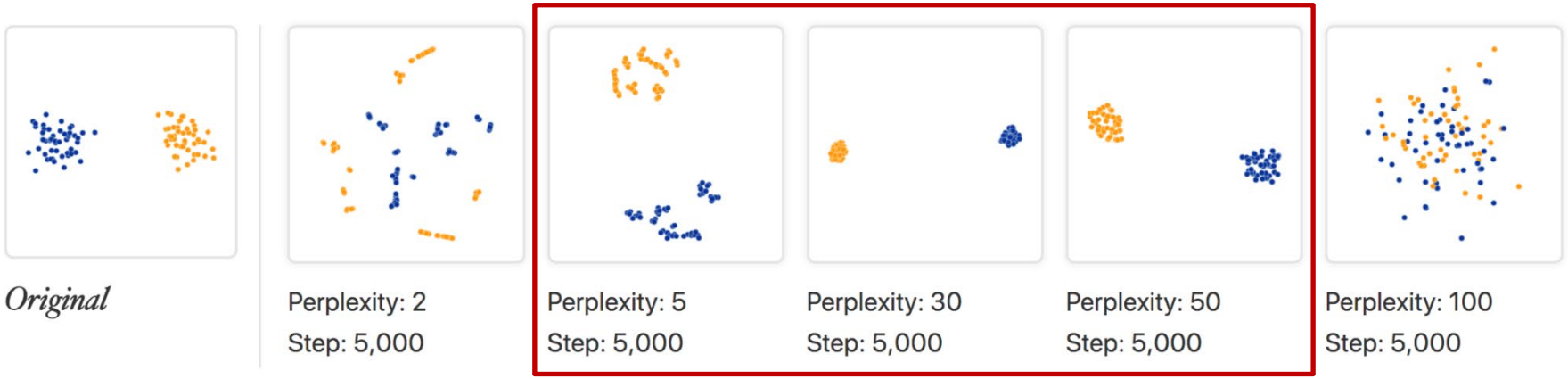
# Remember: PCA of MNIST handwritten-digits data



# By contrast: t-SNE of MNIST digits



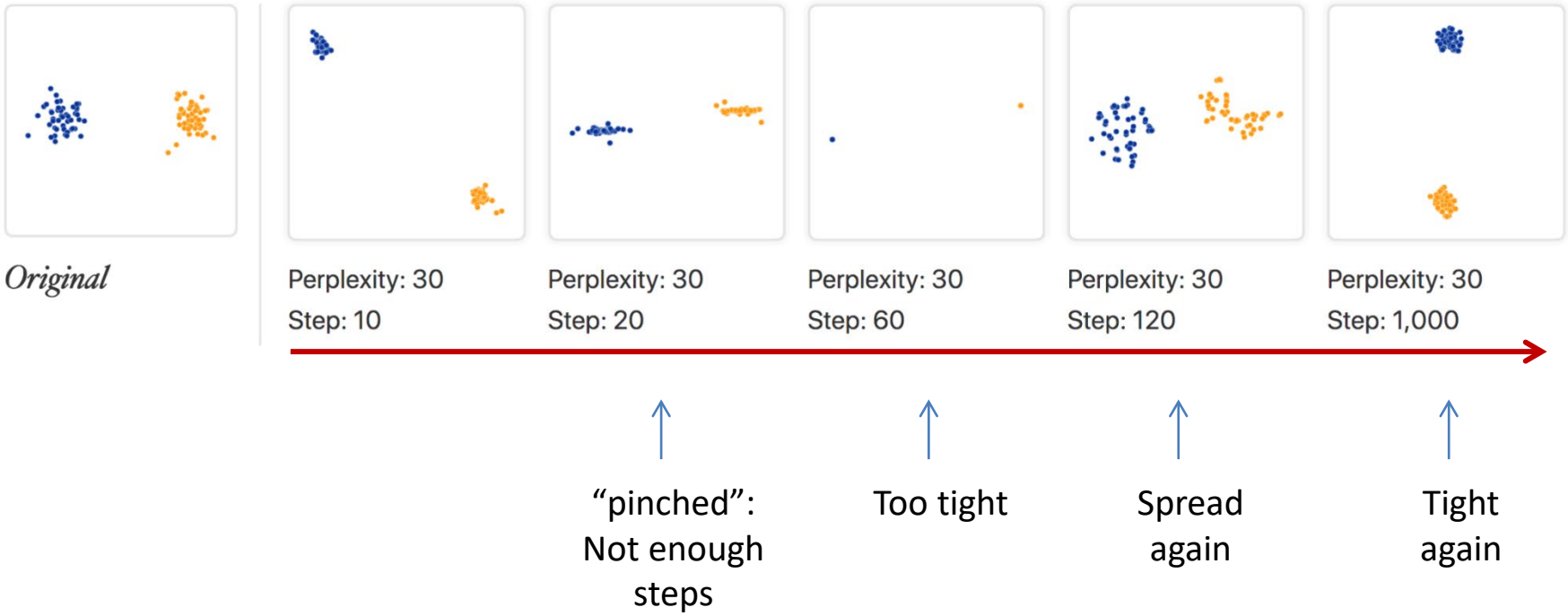
# Perplexity matters




Recommended range by Van Der Maaten and Hinton

https://distill.pub/2016/misread-tsne/

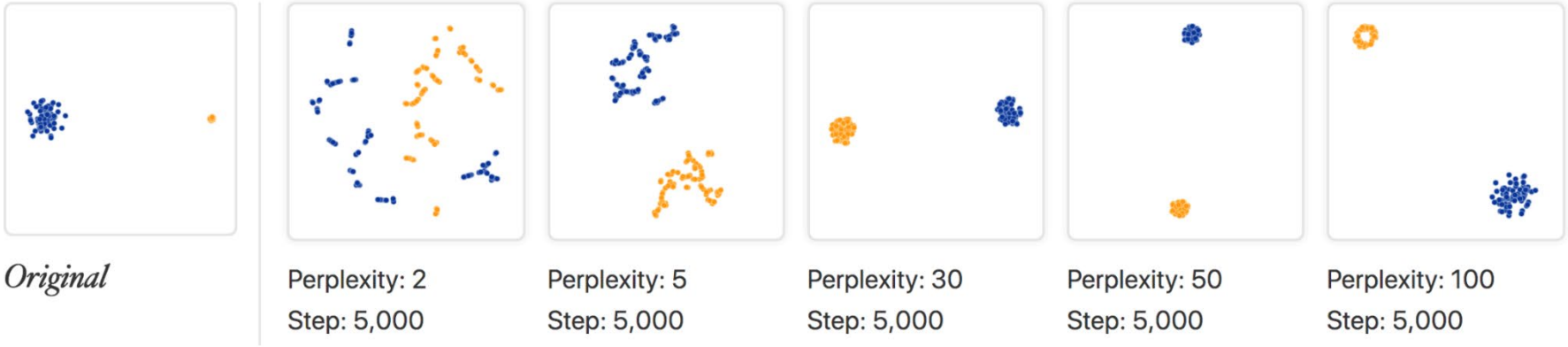
# Number of steps matter




https://distill.pub/2016/misread-tsne/

# Cluster sizes are not meaningful

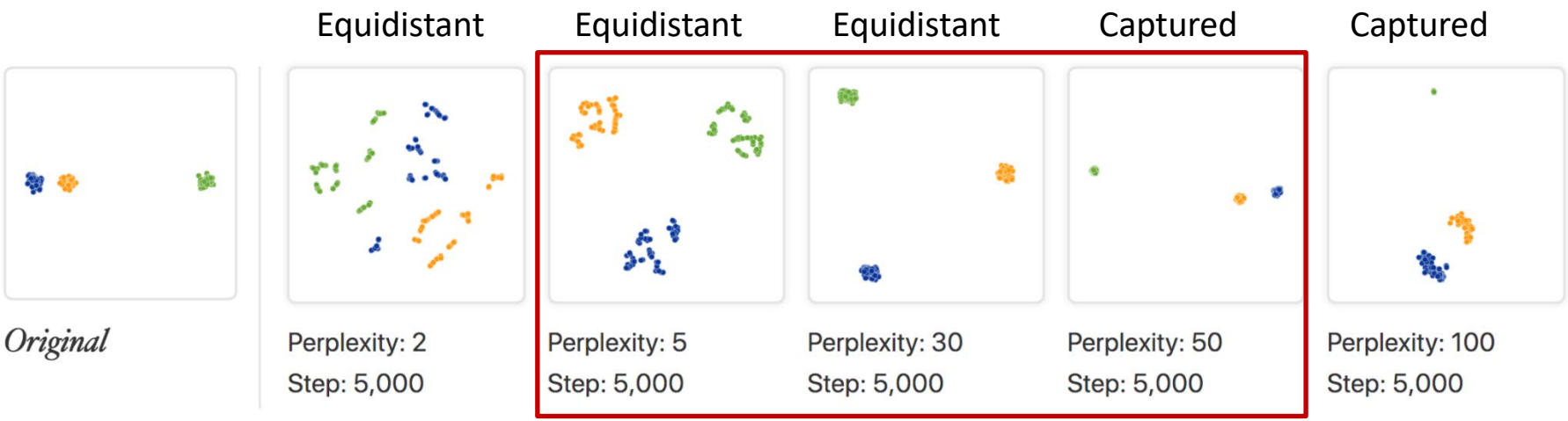
Original data: 2 Gaussians
Widely different (10-fold) dispersion




t-SNE loses that notion of distance.
By design, it adapts to regional variations in distance.

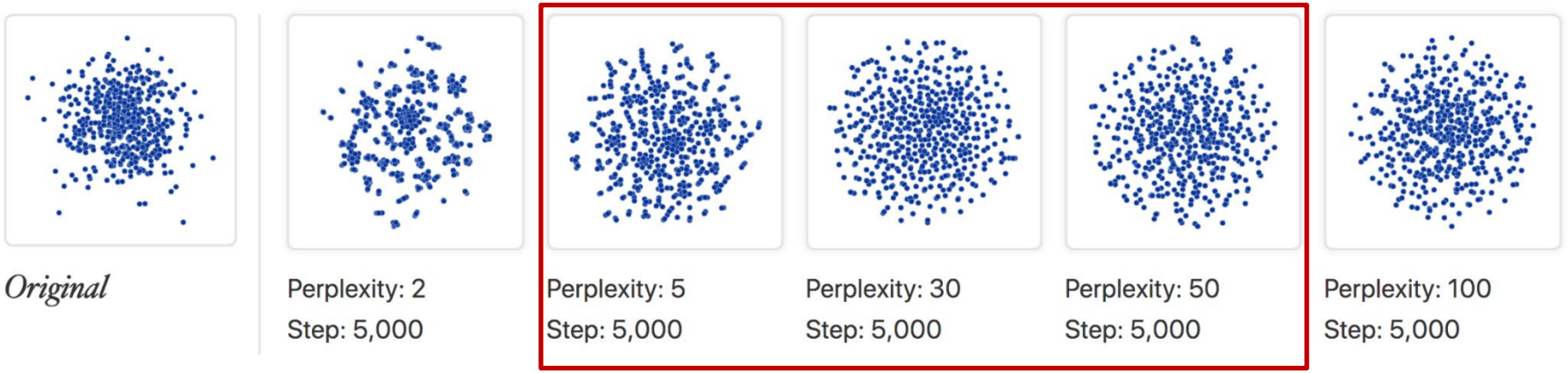
https://distill.pub/2016/misread-tsne/

# Between-cluster distance is not always preserved



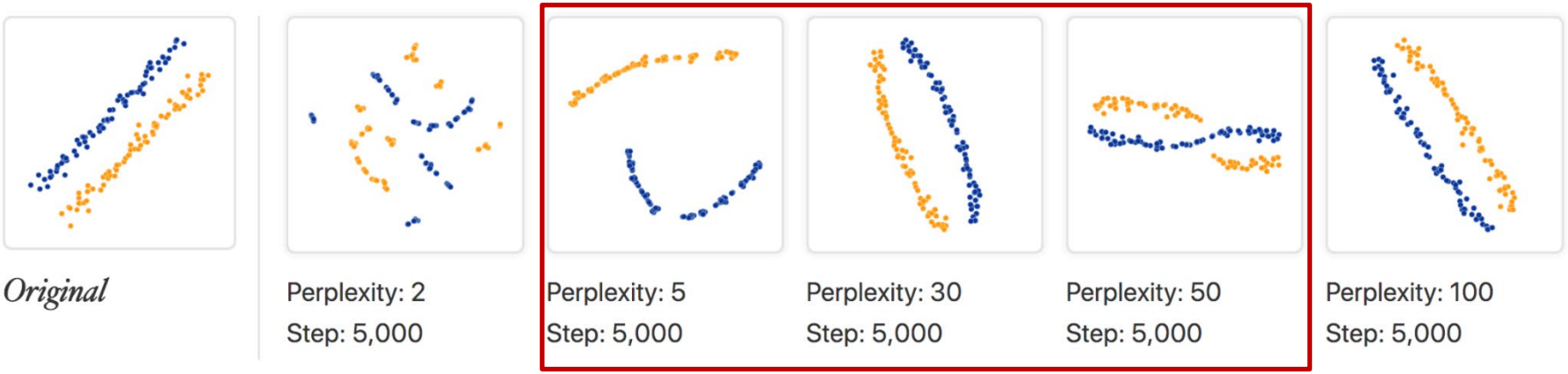

https://distill.pub/2016/misread-tsne/

# False clusters may appear



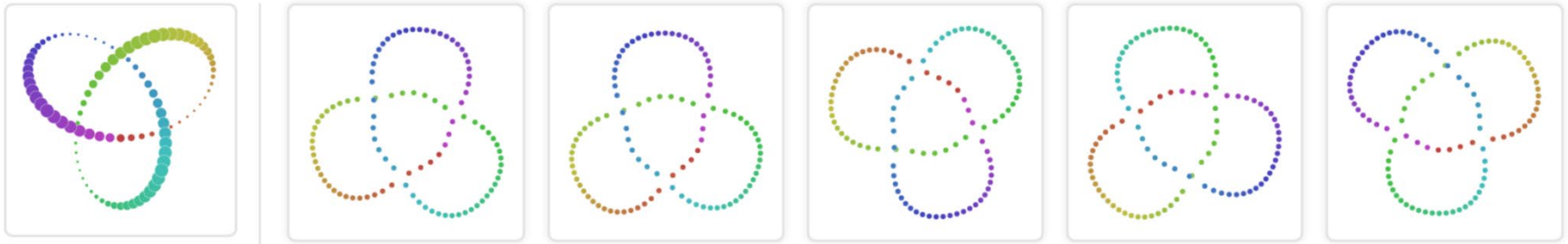

https://distill.pub/2016/misread-tsne/

# Relationships are not always preserved

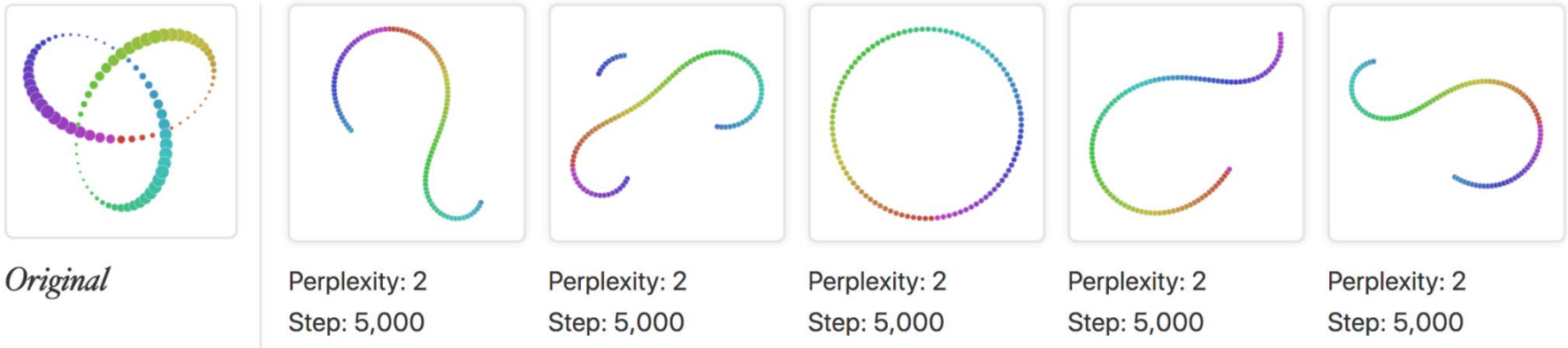



https://distill.pub/2016/misread-tsne/

# Different runs produce surprisingly similar results…

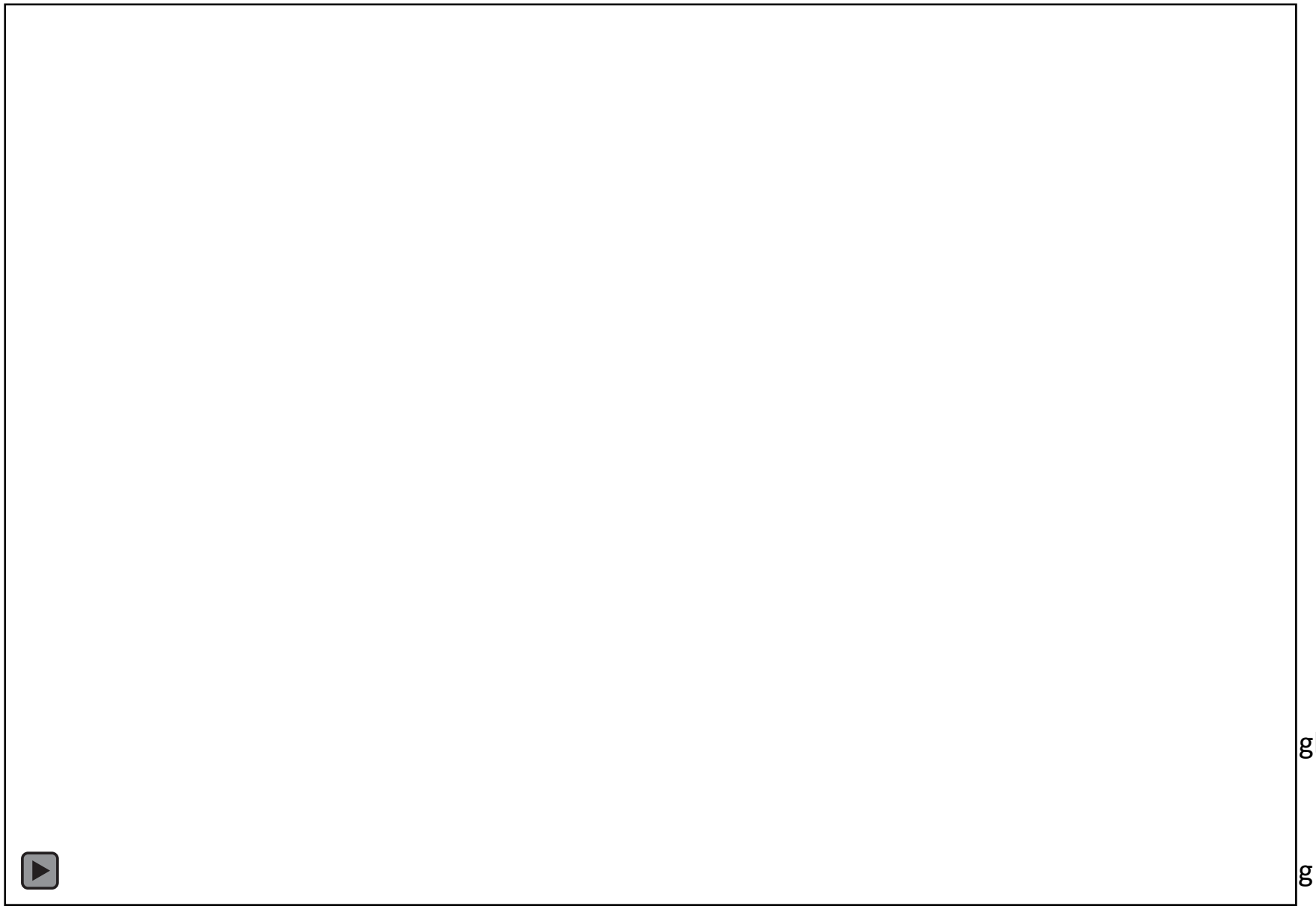



# (… but not at very low perplexity)




https://distill.pub/2016/misread-tsne/

# Demo of t-SNE in action

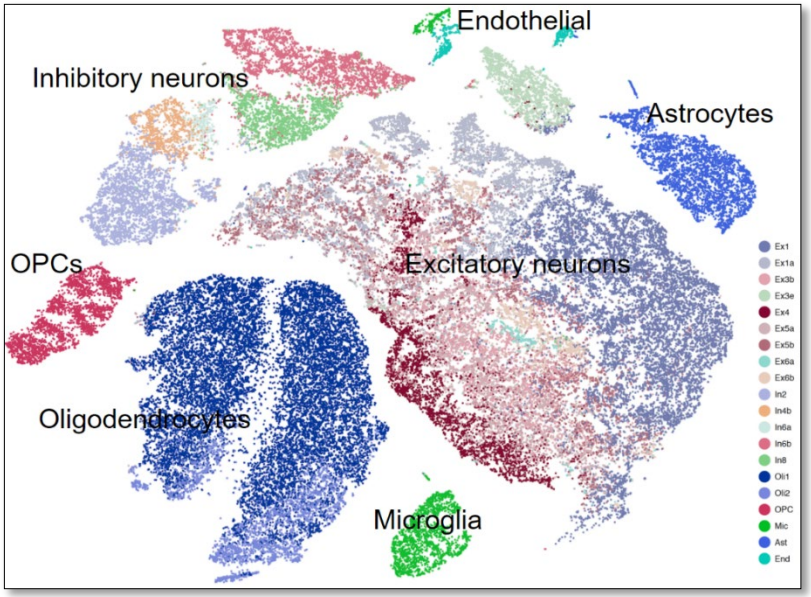


ghbors)

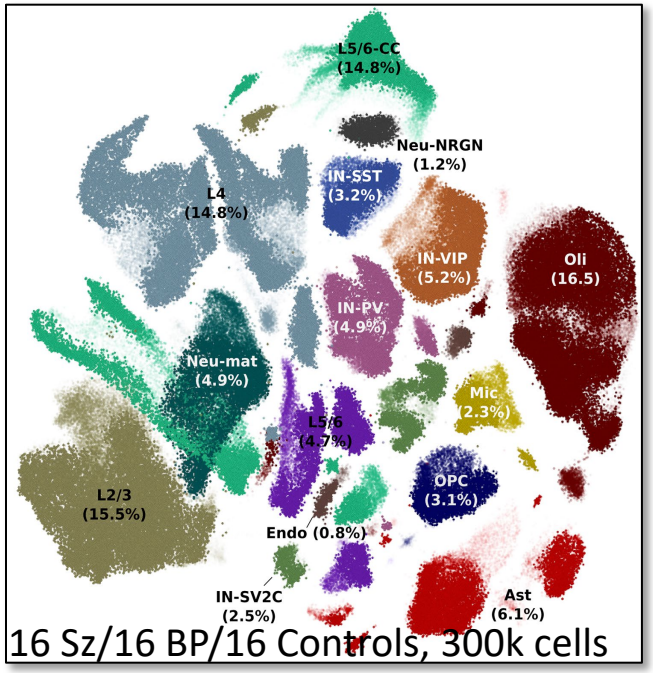
g rate)

https://distill.pub/2016/misread-tsne/

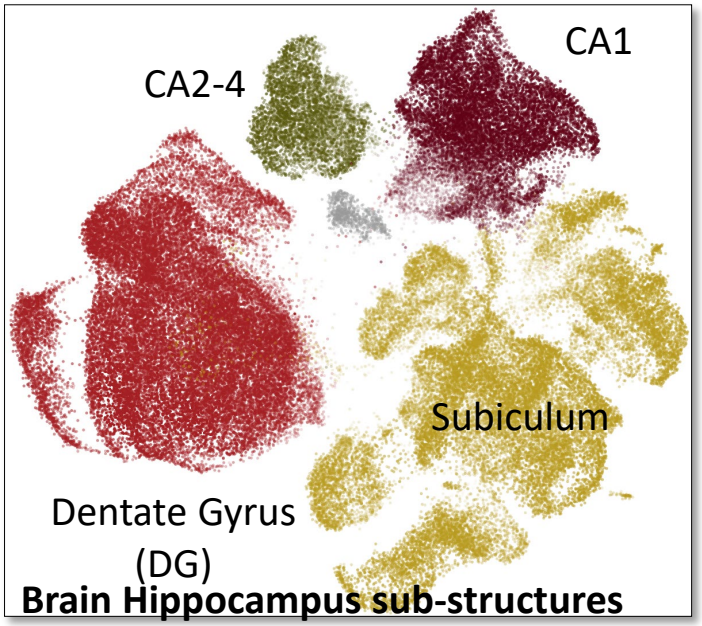
# t-SNEs of single-cell Brain data



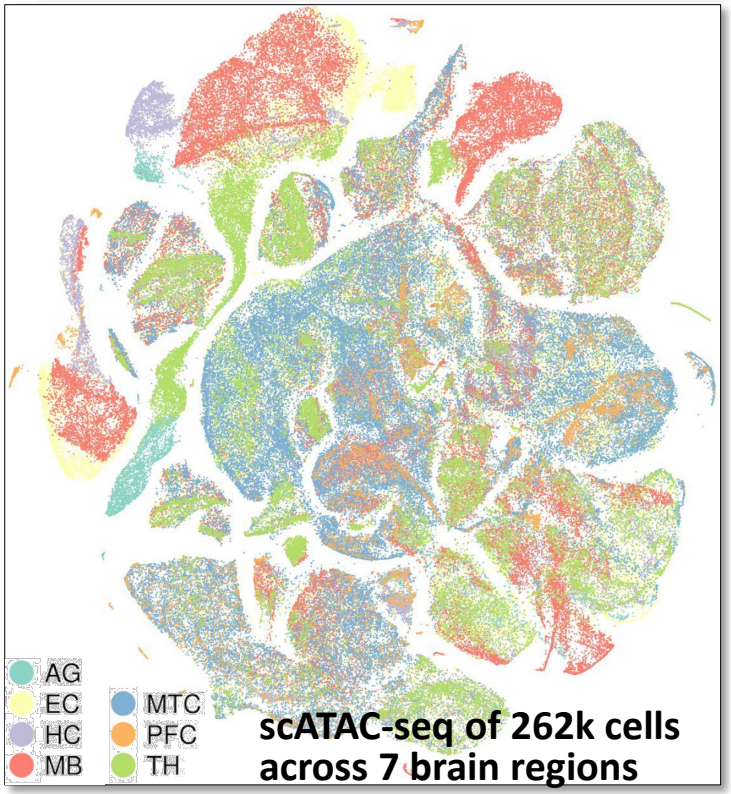

**scRNA-seq in 48 individuals, 84k cells, Nature, 2019**




16 Sz/16 BP/16 Controls, 300k cells




**Brain Hippocampus sub-structures**

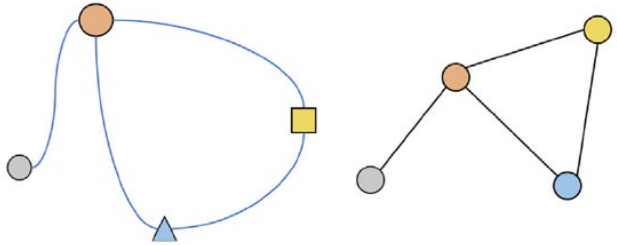



**scATAC-seq of 262k cells across 7 brain regions**

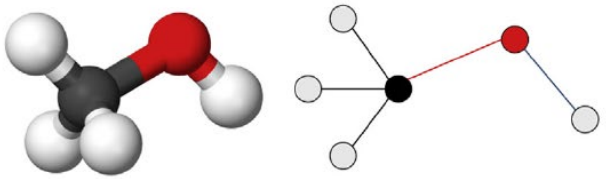
# Goals for today: Network analysis

1. **Introduction to networks**
   - Network types: regulatory, metab., signal., interact., func.
   - Bayesian (probabilistic) and Algebraic views
2. **Network Centrality Measures**
   - Local centrality metrics (degree, betweenness, closeness, etc)
   - Global centrality metrics (eigenvector centrality, page-rank)
3. **Linear Algebra Review: eigenvalues, SVD, low-rank approximations**
   - Eigenvector and singular vector decomposition
   - Low rank approximations, Wigner semicircle law
4. **Sparse Principal Component Analysis**
   - Lasso and Elastic lasso
   - PCA and Sparse PCA
5. **Machine Learning on Networks**
   - Guilt by association
   - Maximum cliques, density-based modules and spectral clustering
6. **Network Diffusion Kernels and Deconvolution**
   - Network diffusion kernels
   - Network deconvolution

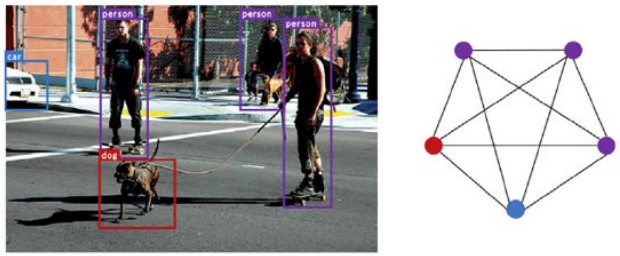
# Learning with Graph Neural Networks (GNN)



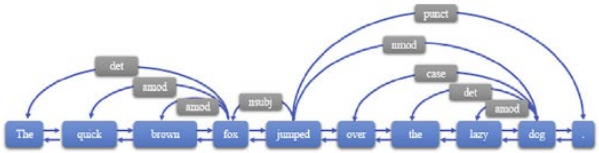
(a) Physics



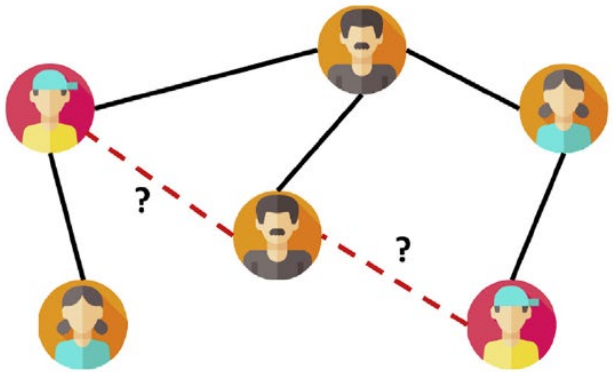
(b) Molecule



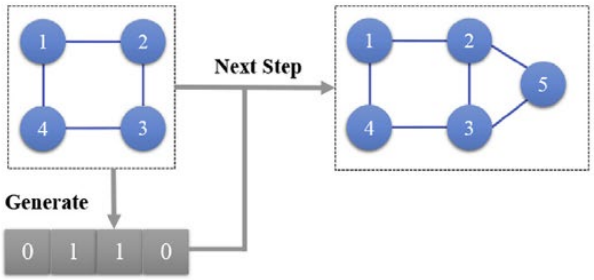

(c) Image



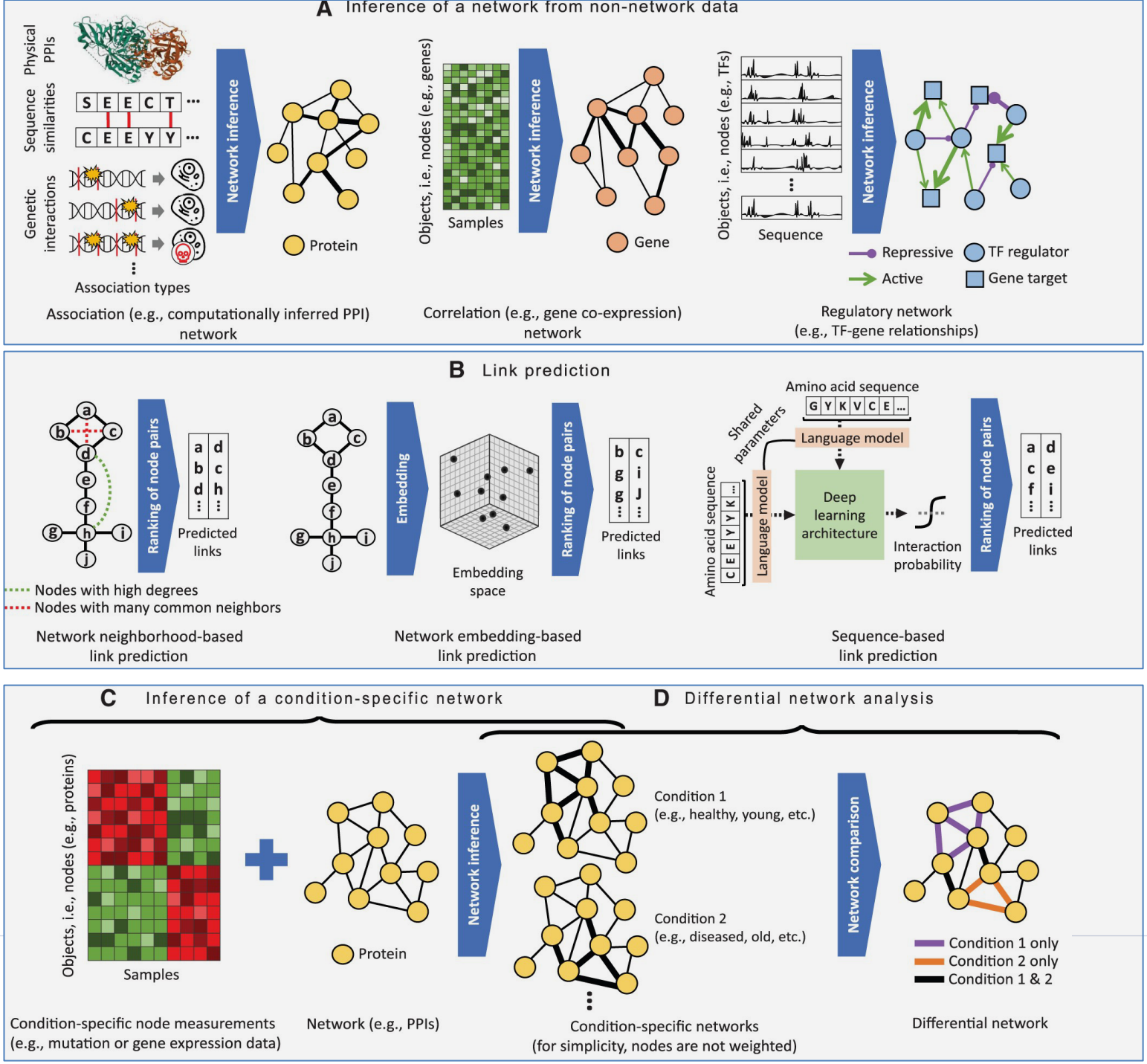

(d) Text




(e) Social Network




(f) Generation


A Inference of a network from non-network data
Physical PPIs
Sequence similarities
Genetic interactions
Association types
Network inference
Protein
Association (e.g., computationally inferred PPI) network
Objects, i.e., nodes (e.g., genes)
Samples
Gene
Correlation (e.g., gene co-expression) network
Objects, i.e., nodes (e.g., TFs)
Sequence
Repressive
Active
TF regulator
Gene target
Regulatory network (e.g., TF-gene relationships)
B Link prediction
Ranking of node pairs
Predicted links
Nodes with high degrees
Nodes with many common neighbors
Network neighborhood-based link prediction
Embedding
Embedding space
Network embedding-based link prediction
Amino acid sequence
Shared parameters
Language model
Deep learning architecture
Interaction probability
Sequence-based link prediction
C Inference of a condition-specific network
D Differential network analysis
Objects, i.e., nodes (e.g., proteins)
Samples
Condition-specific node measurements (e.g., mutation or gene expression data)
Network (e.g., PPIs)
Condition 1 (e.g., healthy, young, etc.)
Condition 2 (e.g., diseased, old, etc.)
Condition-specific networks (for simplicity, nodes are not weighted)
Network comparison
Condition 1 only
Condition 2 only
Condition 1 & 2
Differential network

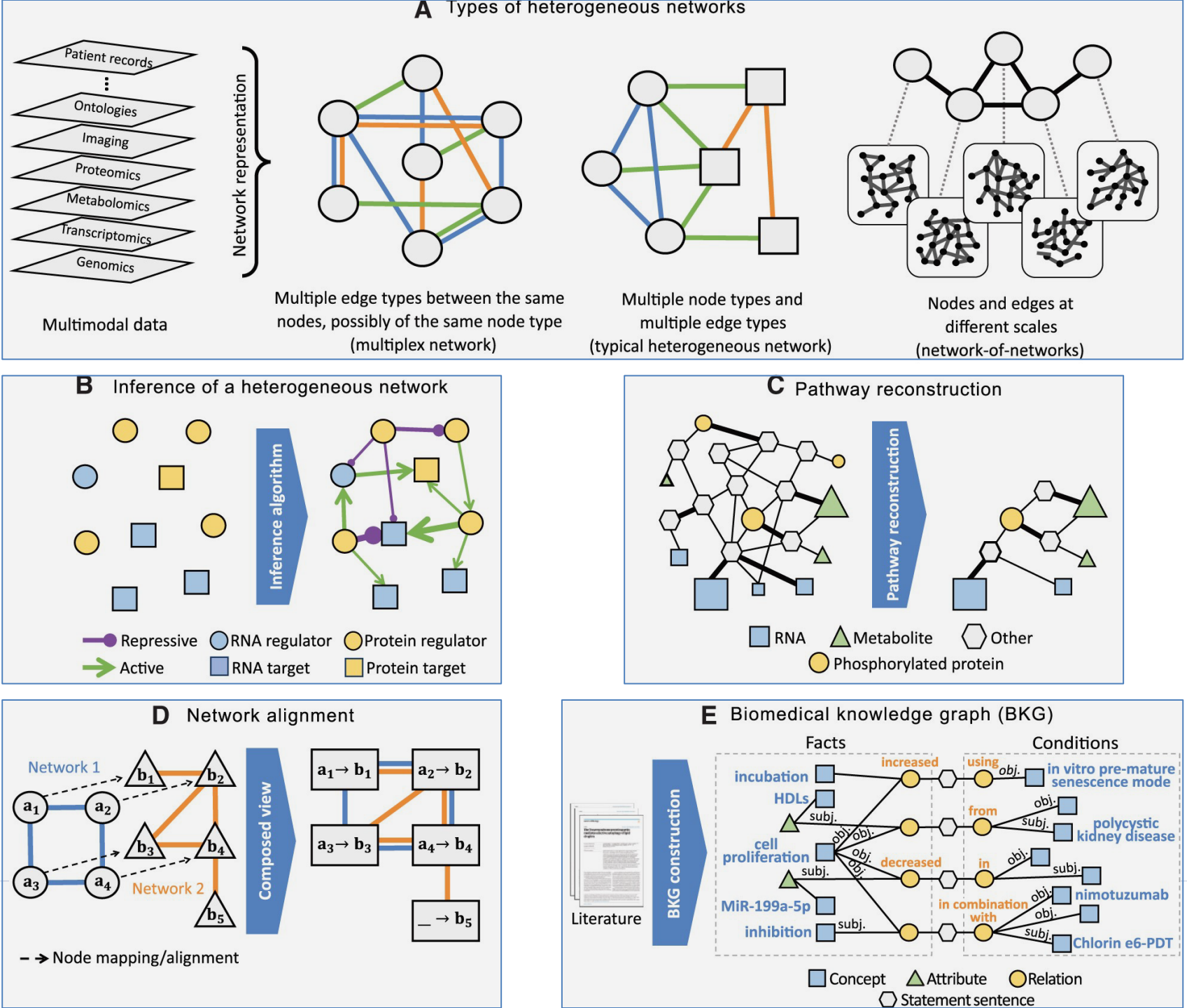



Figure 3. Prominent topics related to multimodal data integration and heterogeneous networks.

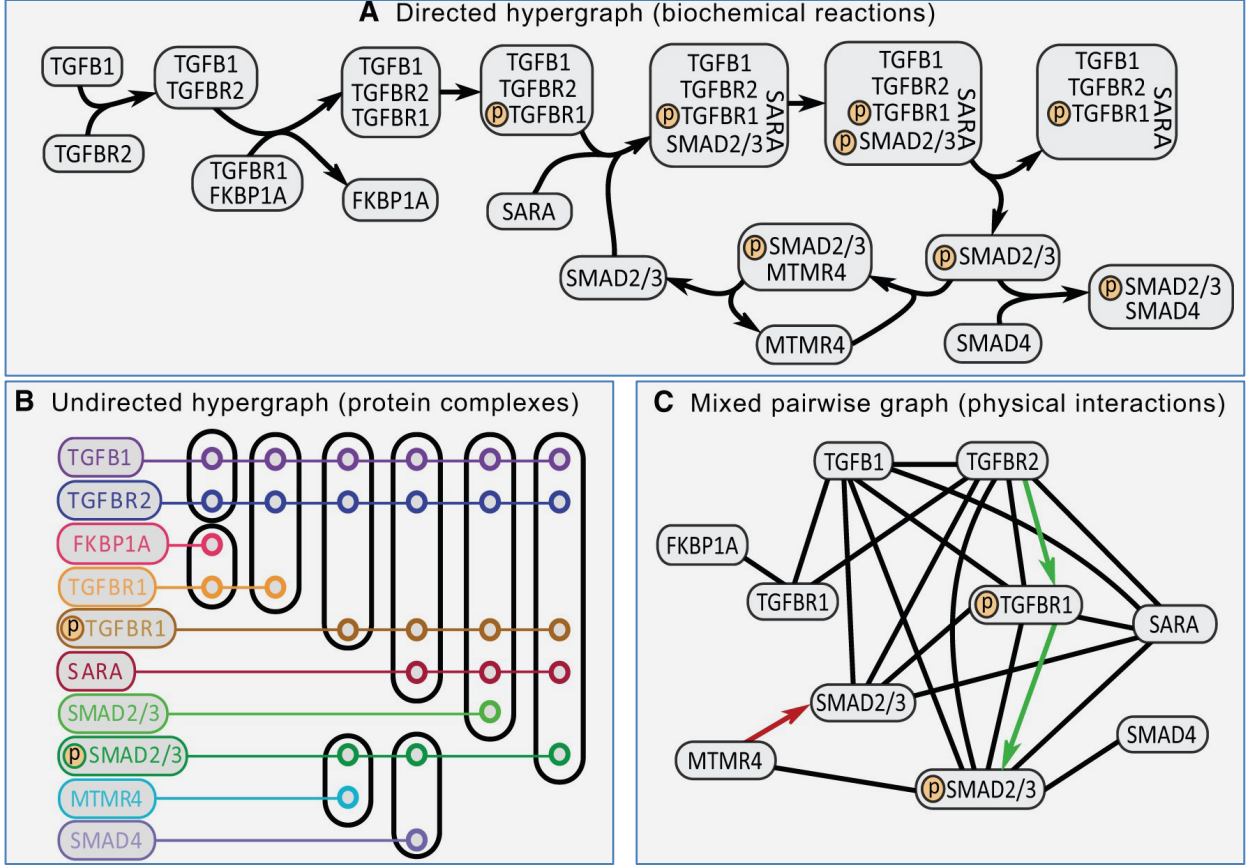
**(A) Heterogeneous networks** can naturally represent multimodal data. A heterogeneous network can have only a single node type, with different data modalities representing multiple edge types. Or, there can exist both multiple node and edge types. Different node types can exist at different biological scales; e.g. in a network-of-networks, nodes at a given scale are networks at the lower scale. (B–E) Prominent topics related to heterogeneous networks.

**(B) Inference** of a heterogeneous network aims to learn the graph topology from multimodal—to date, typically multi-omic—measurements.

**(C) Pathway reconstruction** for interpretation of multi-omic data: the input is multi-omic data and a background molecular network, and the output is a sparse subnetwork. Typically input biomolecules with higher scores (indicated by node sizes) and higher-quality connections (indicated by edge thickness) are prioritized in the output.

**(D) Network alignment**: input can be individual homogeneous networks (left) or heterogeneous networks. Even alignment of homogeneous networks leads to a heterogeneous network (right) whose "supernodes" contain mapped nodes and whose edge types indicate which edges of the original networks are conserved (e.g., between supernodes "a1→b1" and "a2→b2" where the edge exists in both network 1 and network 2) versus nonconserved (e.g. between supernodes "a1→b1" and "a3→b3" where the edge exists in network 1 but not in network 2) under the given node mapping.

**(E) Inference of and reasoning on BKGs**. Shown is a condition-aware BKG. The middle nodes (hexagons) are statement sentences. The layers on their left represent fact tuples and those on their right represent the conditions associated with the facts. The tuples have relation nodes (circles), concept nodes (squares), and optional attribute nodes (triangles).




## Graph representations

e.g. nine reactions from Reactome's TGFβ signaling pathway.

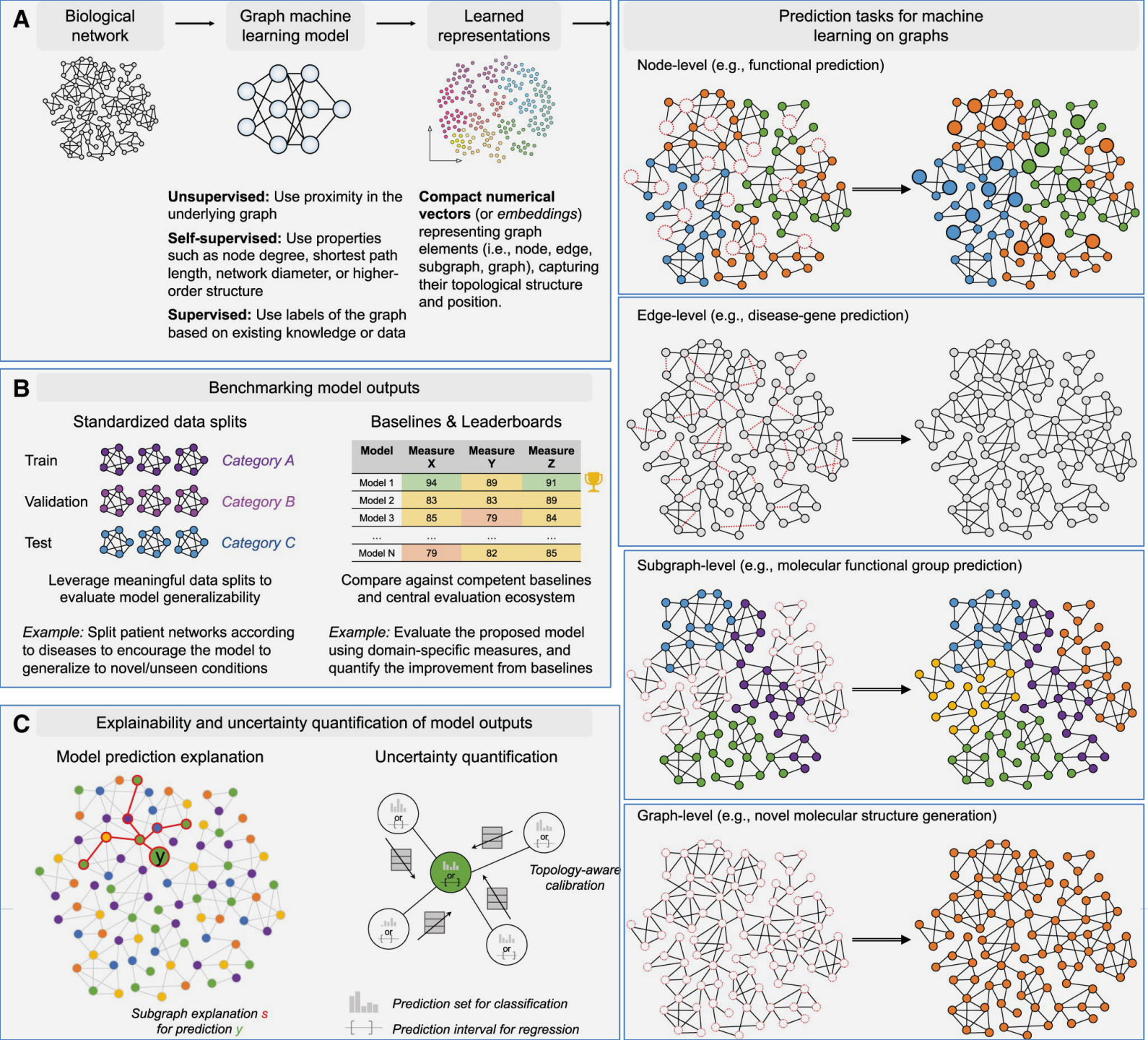
(A) In a directed hypergraph, each hyperedge captures a reaction ("p" denotes phosphorylation).

(B) In an undirected hypergraph, each hyperedge captures a protein complex.

(C) In a (mixed) pairwise graph, each edge captures a pairwise interaction. "Mixed" refers to having both directed and undirected edges in the graph. Undirected edges denote physical interactions; directed edges denote either phosphorylation (the two right-most directed edges) or dephosphorylation (the left-most directed edge).

(D) A node in a pairwise graph can be represented as a **vector of graphlet counts**. The number of 2-, 3-, and 4-node graphlet instances that include TGFB1 in the graph on the left are shown. (E) A node in an undirected hypergraph can be represented as a vector of hypergraphlet counts. The number of 2- and 3-node hypergraphlet instances that include TGFB1 in the hypergraph on the left are shown.

In panels (D and E), only the (hyper)graphlet-level counts are shown for simplicity, i.e. (hyper)graphlet orbits are not shown nor considered when doing the counting. However, in practice, the more detailed orbit-level counts are computed rather than the (hyper)graphlet-level counts.



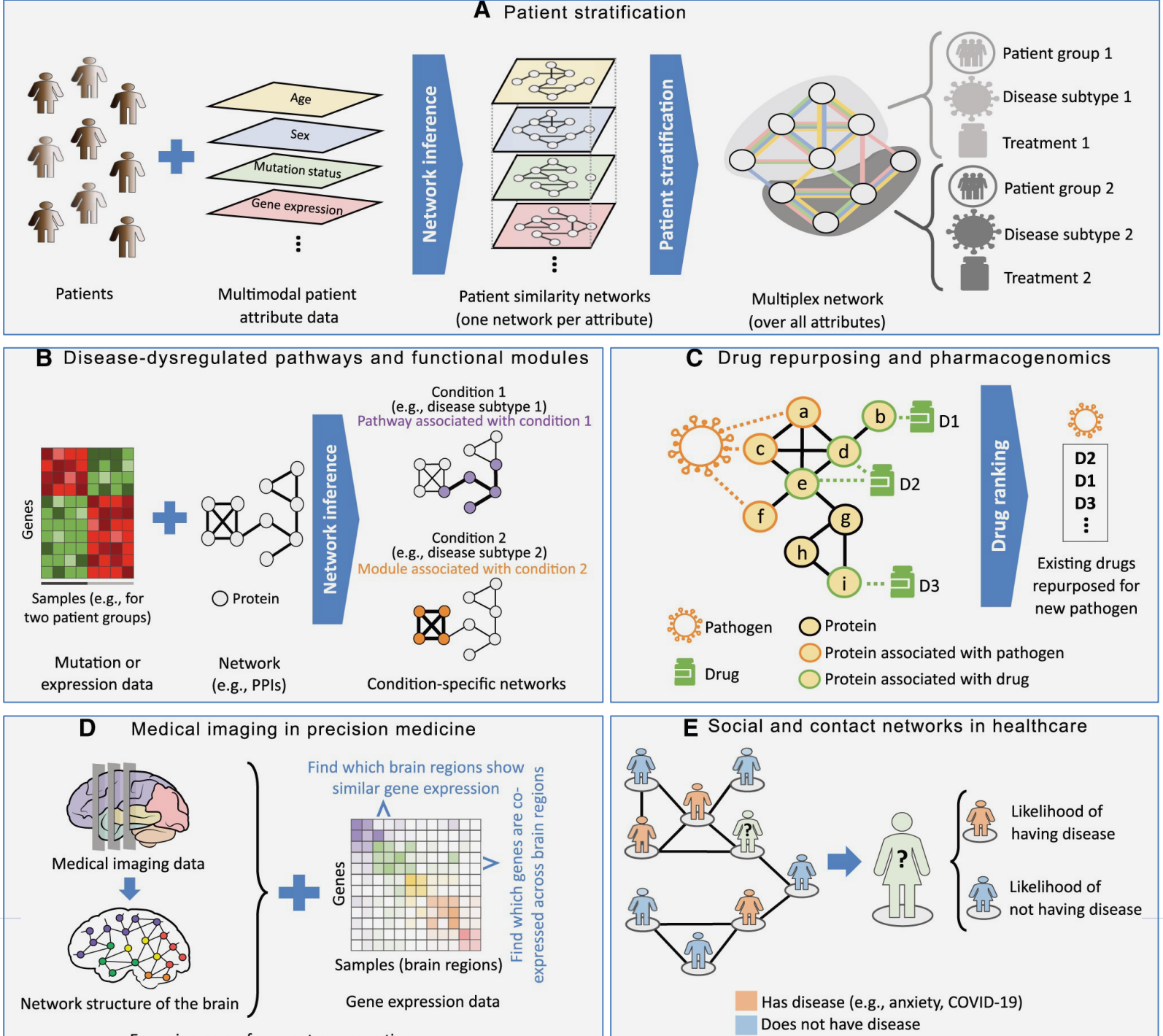

# Machine learning on networks.

**(A) The core of this approach is a machine learning model**, typically a neural network, that takes one or more biological networks as input and learns representations (i.e. embeddings) of various graph elements in an unsupervised, self-supervised, or supervised manner.

**Four types of prediction tasks** (denoted by the red dashed lines): node-, edge-, subgraph-, and graph-level predictions.
Colors of nodes for the node-, subgraph-, and graph-level tasks signify the label; white nodes indicate missing labels to be predicted by the model.

(B) Critical to continued development, wide adoption, and practical utility of network-based machine learning is a parallel improvement in frameworks for **rigorous benchmarking** via established data splits and baselines

(C) **Explainability** of model predictions (e.g. identifying a subgraph s, denoted by red lines, that best explains the prediction y for the query node, denoted in green) and uncertainty quantification (e.g. using the prediction set for a classification task or prediction interval for a regression task; Huang et al. 2023b).



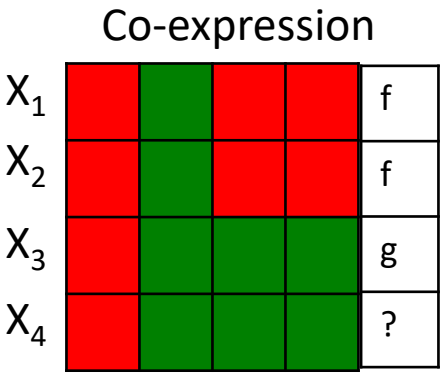

**Network-based precision medicine.** (A) Groups of patients that correspond to their communities (clusters) in a patient similarity network may shed light on distinct disease subtypes and thus lead to tailored, group-specific therapeutic strategies. (B) Identification of pathways (sparse, tree-like subnetworks) or functional modules (dense, clique-like subnetworks) associated with disease (subtypes) is related to inference of a condition-specific network (Section 2) and pathway reconstruction (Section 3). (C) Drug repurposing evaluates the fit of existing drugs to new diseases based on network "relatedness" between protein targets of the existing drugs and proteins associated with the new diseases, e.g. existing drug D2 may be a good treatment for the new pathogen because D2 targets two proteins (d and e), both of which directly interact with two of the proteins associated with the pathogen (a and c); the four proteins (a, c, d, e) form a clique, which further adds to their "relatedness." (D) An important application of medical imaging lies in brain disorders. In connectome genetics, network structure of the brain meets -omics data. (E) An individual's position in their social/contact network, along with demographic, personality, physical/mental health, etc. information about the other individuals, can give insights into the given individual's health.

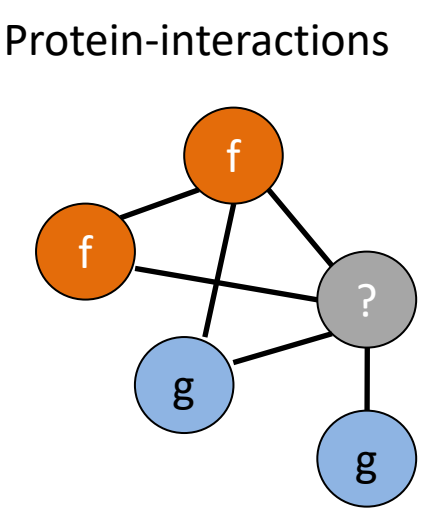
# Goals for today: Network analysis

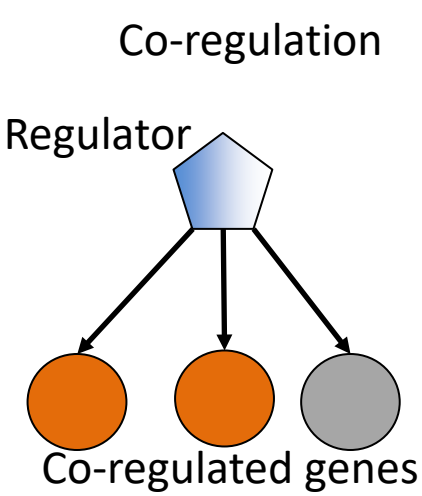
1. **Introduction to networks**
   - Network types: regulatory, metab., signal., interact., func.
   - Bayesian (probabilistic) and Algebraic views
2. **Network Centrality Measures**
   - Local centrality metrics (degree, betweenness, closeness, etc)
   - Global centrality metrics (eigenvector centrality, page-rank)
3. **Linear Algebra Review: eigenvalues, SVD, low-rank approximations**
   - Eigenvector and singular vector decomposition
   - Low rank approximations, Wigner semicircle law
4. **Sparse Principal Component Analysis**
   - Lasso and Elastic lasso
   - PCA and Sparse PCA
5. **Machine Learning in Networks**
   - Guilt by association
   - Maximum cliques, density-based modules and spectral clustering
6. **Network Diffusion Kernels and Deconvolution**
   - Network diffusion kernels
   - Network deconvolution

# Predicting functions of un-annotated genes

- Goal: Predict function of unannotated genes based on “guilt by association”
- Different types of “association”



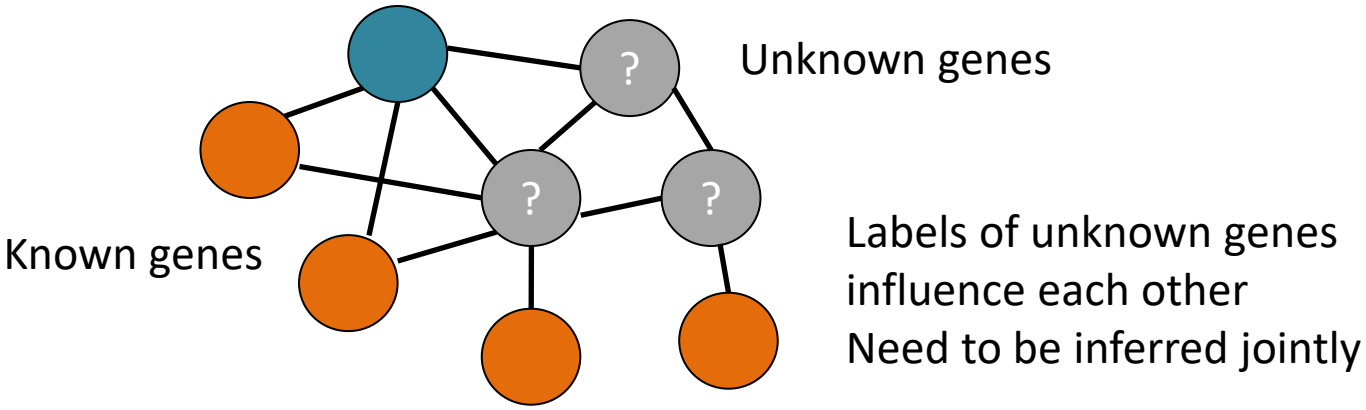




**However most approaches work with “functional networks”**

Deng et al 03, Sharan et al 07

# Iterative classification algorithm

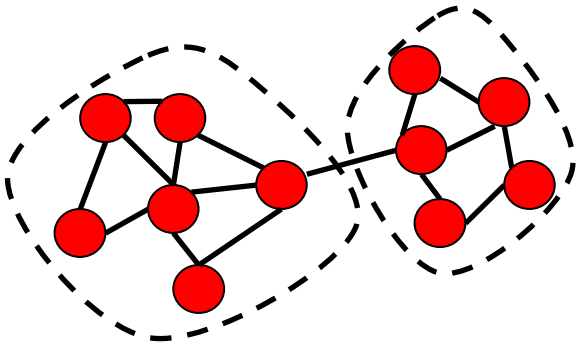



- Start with an initial assignment of labels
- Repeat iteratively
  - Update relational attributes
  - Re-infer the labels

Neville 03, Getoor 05

# Modularity of regulatory networks

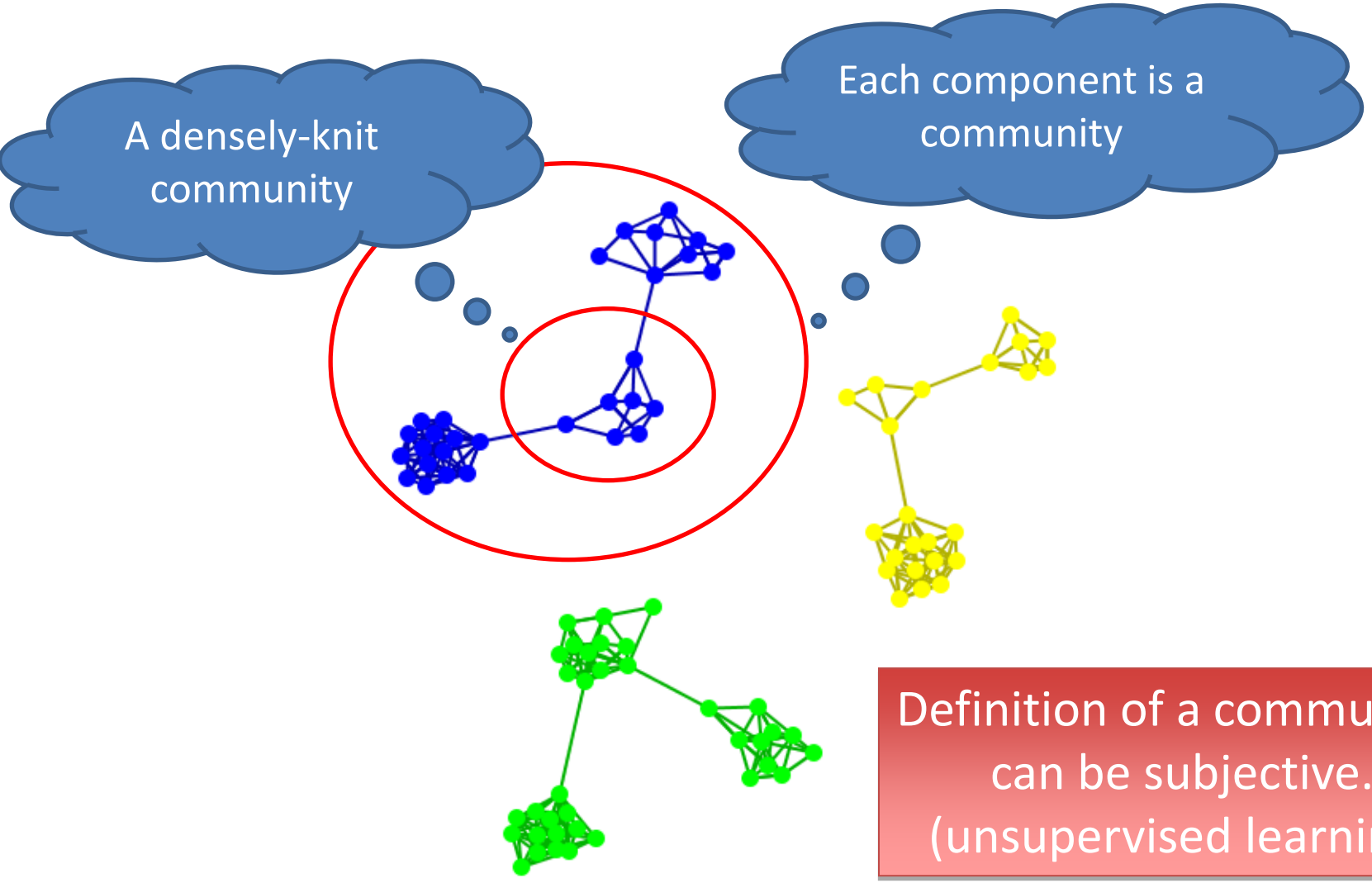
- Modular: Graph with densely connected subgraphs



- Genes in modules involved in similar functions and co-regulated
- Modules can be identified using graph partitioning algorithms
  - Markov Clustering Algorithm (random walks on graph)
  - Girvan-Newman Algorithm (hierarchical communities)
  - Spectral partitioning (eigenvalue of Laplacian matrix)

Newman PNAS 2007

# Subjectivity of Community Definition




Definition of a community can be subjective. (unsupervised learning)

# Taxonomy of Community Criteria

- Criteria vary depending on the tasks
- Roughly, community detection methods can be divided into 4 categories (not exclusive):
- Node-Centric Community
  - Each node in a group satisfies certain properties
- Group-Centric Community
  - Consider the connections within a group as a whole. The group has to satisfy certain properties without zooming into node-level
- **Network-Centric Community**
  - Partition the whole network into several disjoint sets
- Hierarchy-Centric Community
  - Construct a hierarchical structure of communities

# Group-Centric Community Detection: Density-Based Groups

- The group-centric criterion requires the whole group to satisfy a certain condition
  - E.g., the group density >= a given threshold
- A subgraph $G_s(V_s, E_s)$ is a $\gamma - dense$ quasi-clique if

$$\frac{2|E_s|}{|V_s|(|V_s|-1)} \geq \gamma$$

where the denominator is the maximum number of degrees.

- A similar strategy to that of cliques can be used
  - Sample a subgraph, and find a maximal $\gamma - dense$ quasi-clique (say, of size $|V_s|$)
  - Remove nodes with degree less than the average degree

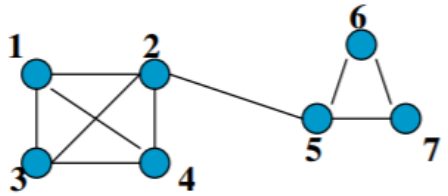
$$< \quad |V_s|\gamma \leq \frac{2|E_s|}{|V_s|-1}$$

# Network-Centric Community Detection

- Network-centric criterion needs to consider the connections within a network globally
- Goal: partition nodes of a network into disjoint sets
- Approaches:
  - (1) Clustering based on vertex similarity
  - (2) Latent space models (multi-dimensional scaling )
  - (3) Block model approximation
  - (4) Markov Clustering (MCL)
  - (5) Spectral clustering
  - (6) Modularity maximization

# Markov Clustering (MCL): Diffusion

- Example:




$$\begin{pmatrix} 0 & .25 & .33 & .33 & 0 & 0 & 0 \\ .33 & 0 & .33 & .33 & .33 & 0 & 0 \\ .33 & .25 & 0 & .33 & 0 & 0 & 0 \\ .33 & .25 & .33 & 0 & 0 & 0 & 0 \\ 0 & .25 & 0 & 0 & 0 & .5 & .5 \\ 0 & 0 & 0 & 0 & .33 & 0 & .5 \\ 0 & 0 & 0 & 0 & .33 & .5 & 0 \end{pmatrix} \xrightarrow{\text{eventually}} \begin{pmatrix} .15 & .15 & .15 & .15 & .15 & .15 & .15 \\ .2 & .2 & .2 & .2 & .2 & .2 & .2 \\ .15 & .15 & .15 & .15 & .15 & .15 & .15 \\ .15 & .15 & .15 & .15 & .15 & .15 & .15 \\ .15 & .15 & .15 & .15 & .15 & .15 & .15 \\ .1 & .1 & .1 & .1 & .1 & .1 & .1 \\ .1 & .1 & .1 & .1 & .1 & .1 & .1 \end{pmatrix}$$

- Avoid complete diffusion:

  ➔ increase strong links, reduce weak links

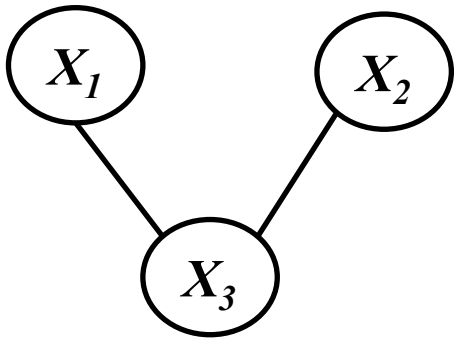
- Cycle between inflation and expansion
  - Inflation: raise a single column to a non-negative power (both strengthens strong, **and** weakens weak)
  - Expansion: take the power of the resulting matrix

# Laplacian matrix of a graph G

- **L=D-A (Degree matrix minus Adjacency matrix)**

- **Adjacency Matrix**

$$A=\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

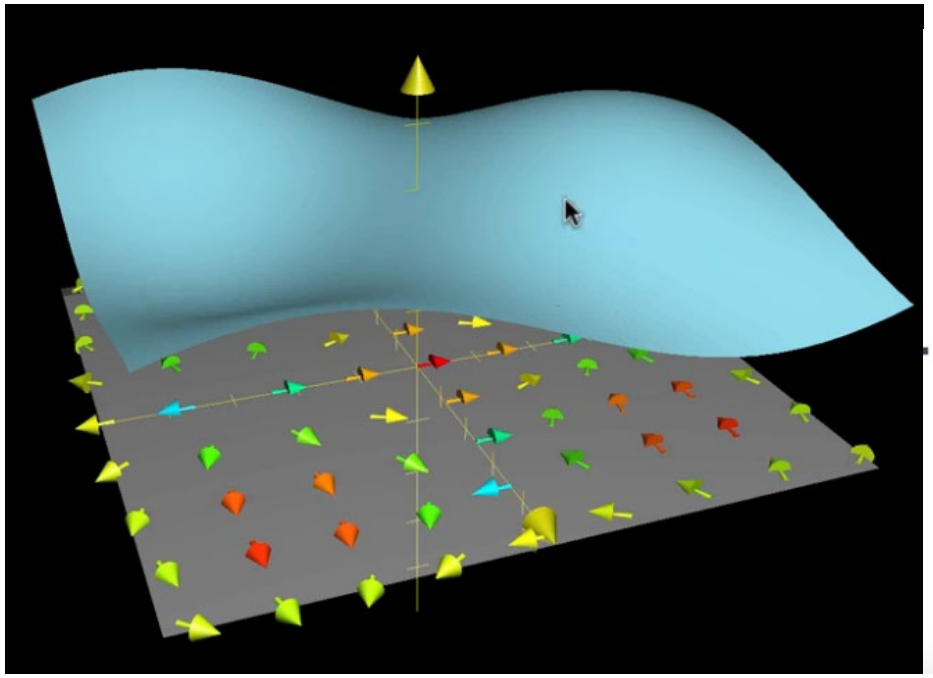


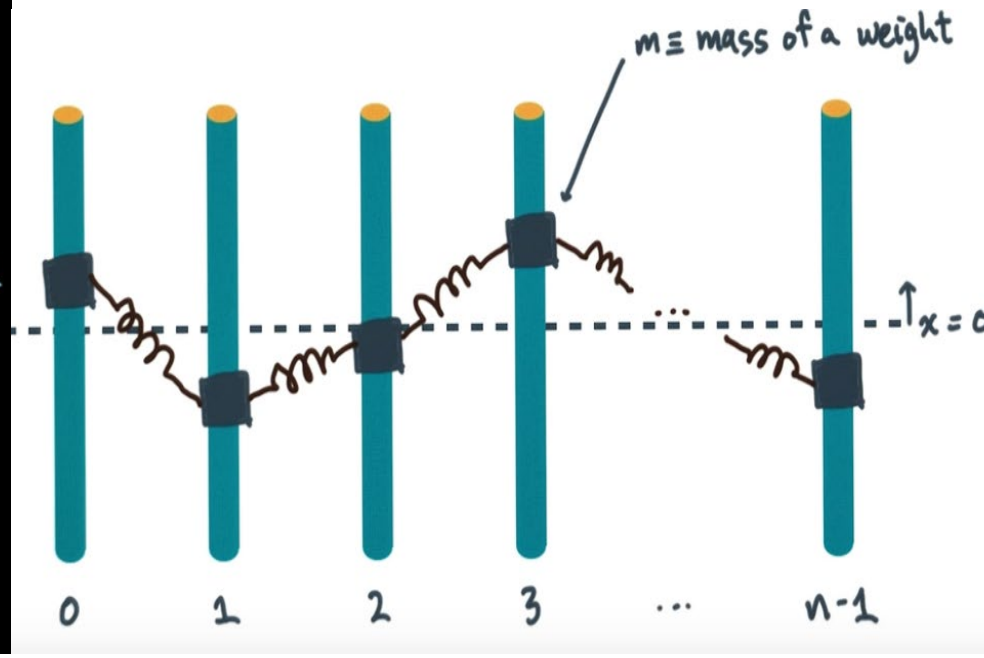

- **Laplacian Matrix**

$$L=\begin{bmatrix} K_1 & & -A_{ij} \\ & K_2 & \\ -A_{ij} & & K_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{bmatrix}$$

adjacency degree

# Laplacian intuition

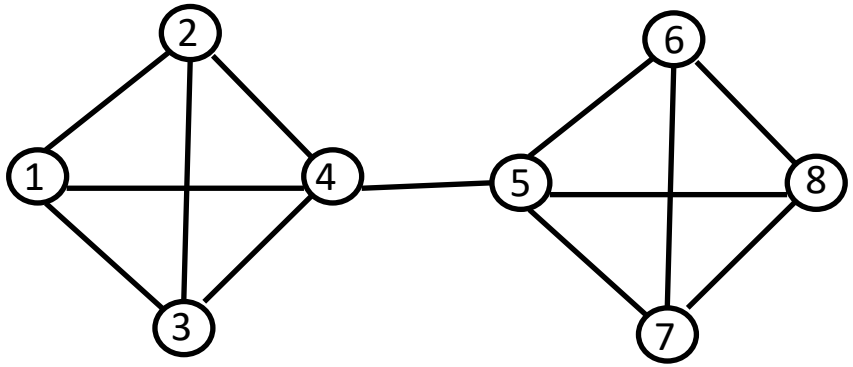





**Gradient concentrations**

**Spring forces**

# Network modularization-example




**A =**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |

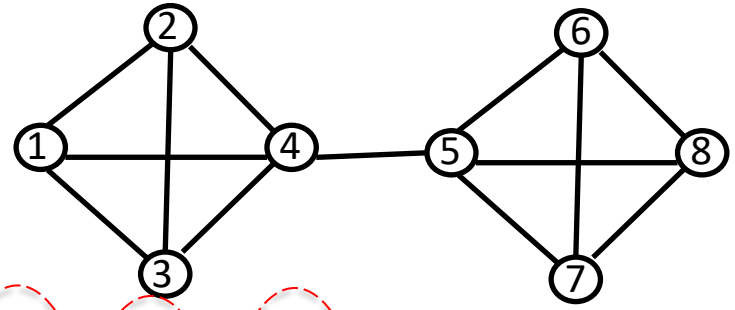
Adjacency Matrix

**L =**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | -1 | -1 | -1 | 0 | 0 | 0 | 0 |
| -1 | 3 | -1 | -1 | 0 | 0 | 0 | 0 |
| -1 | -1 | 3 | - 1 | 0 | 0 | 0 | 0 |
| -1 | -1 | -1 | 4 | - 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | -1 | 4 | -1 | - 1 | -1 |
| 0 | 0 | 0 | 0 | -1 | 3 | -1 | -1 |
| 0 | 0 | 0 | 0 | -1 | -1 | 3 | -1 |
| 0 | 0 | 0 | 0 | -1 | -1 | -1 | 3 |

Laplacian Matrix

# Eigen decomposition-example




$$L = U\Sigma U^{-1}$$

U=

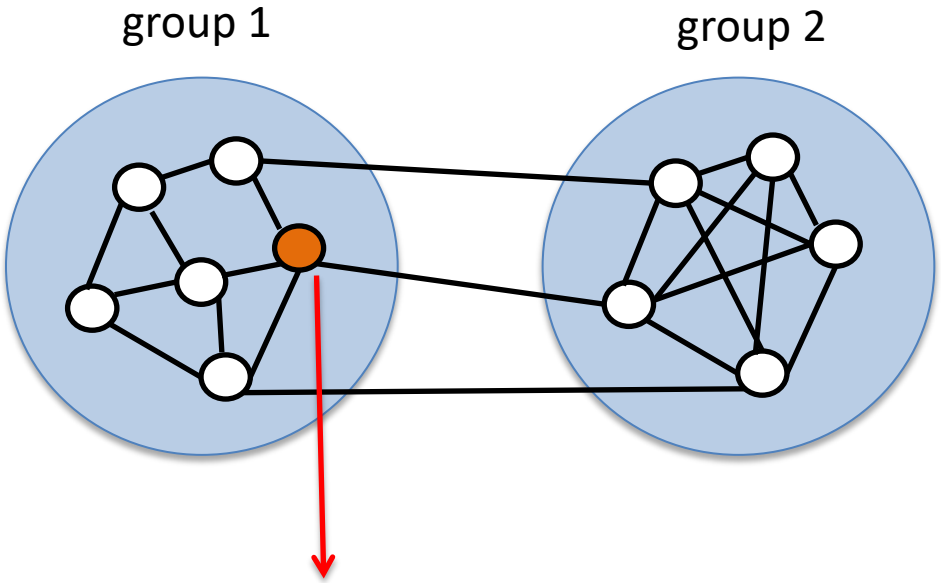
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.3536 | -0.3825 | 0.2714 | -0.1628 | -0.7783 | 0.0495 | -0.0064 | -0.1426 |
| 0.3536 | -0.3825 | 0.5580 | -0.1628 | 0.6066 | 0.0495 | -0.0064 | -0.1426 |
| 0.3536 | -0.3825 | -0.4495 | 0.6251 | 0.0930 | 0.0495 | -0.3231 | -0.1426 |
| 0.3536 | -0.2470 | -0.3799 | -0.2995 | 0.0786 | -0.1485 | 0.3358 | 0.6626 |
| 0.3536 | 0.2470 | -0.3799 | -0.2995 | 0.0786 | -0.1485 | 0.3358 | -0.6626 |
| 0.3536 | 0.3825 | 0.3514 | 0.5572 | -0.0727 | -0.3466 | 0.3860 | 0.1426 |
| 0.3536 | 0.3825 | 0.0284 | -0.2577 | -0.0059 | -0.3466 | -0.7218 | 0.1426 |
| 0.3536 | 0.3825 | 0.0000 | 0.0000 | 0.0000 | 0.8416 | -0.0000 | 0.1426 |

$\Sigma =$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0.3542 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 4.0000 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 4.0000 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 4.0000 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 4.0000 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 4.0000 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5.6458 |

- First smallest eigenvalue of Laplacian matrix is always zero.
- Second smallest eigenvector of Laplacian matrix characterizes a network partition.

# Laplacian matrix characterizes # of edges between groups




node *i* in group 1 => $s_i$=1
node *i* in group 2 => $s_i$=-1

*i*-th component of *L S*=

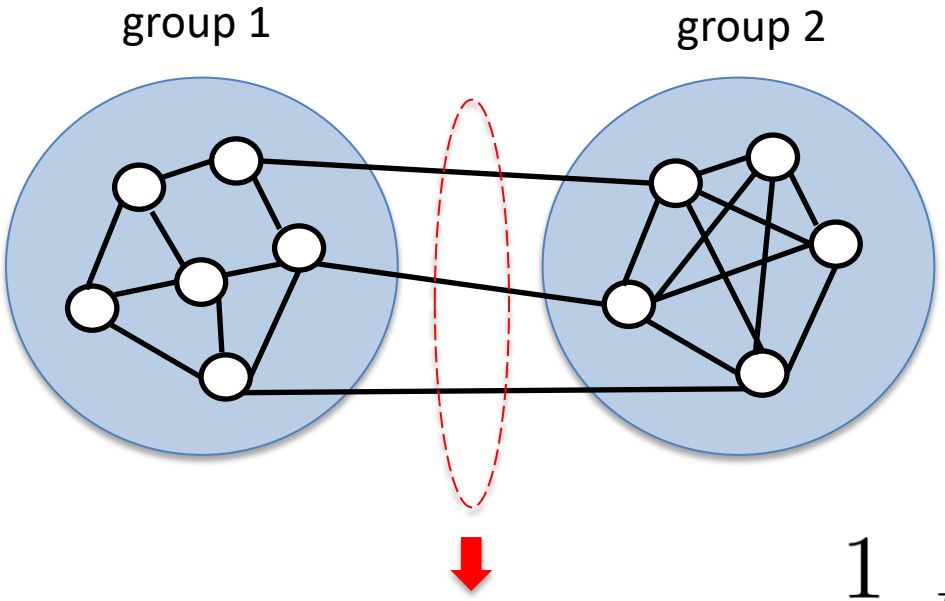
*= degree (i)- (# edges to group 1) + (# edges to group 2)*

**Diagonal term of Laplacian**

**Off-diagonal terms of Laplacian**

*= 4- 3+1=2=2*(# edges going out of group 1)*

# Laplacian matrix characterizes # of edges between groups




node *i* in group 1 => $s_i$=1
node *i* in group 2 => $s_i$=-1

**# of edges between groups =** $\frac{1}{4}\mathbf{s}^t L\mathbf{s}$

- Choose vector ***s*** to minimize the error term
- A trivial solution: if s=(1,1,…,1), error is zero.
- ***A non-trivial solution: s parallel to the second eigenvector of L (why?)***

# Network modularization by using decomposition of Laplacian matrix

$$\min_{\mathbf{s}} \mathbf{s}^t L \mathbf{s}$$

- Use eigen decomposition principles:

$$L \rightarrow (\mathbf{v}_i, \lambda_i) \qquad L = \sum_i \lambda_i \mathbf{v}_i^t \mathbf{v}_i$$

- Project *s* over eigenvectors of *L*: $\mathbf{s} = \sum_i a_i \mathbf{v}_i$

$$\mathbf{s}^t L \mathbf{s} = \sum_i a_i^2 \lambda_i$$

- Challenges in finding optimal $a_i$'s:
  - Without other conditions, a trivial solution exists
  - Second eigenvector characterizes partitioning
  - Vector s should be integer-valued => projection

# Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions $Z_{ij}$ from motif matrix
   - M step: Find max-likelihood motif from all positions $Z_{ij}$
3. Gibbs Sampling: Sample from joint ($M$,$Z_{ij}$) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Positioning matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

# Goals for today: Network analysis

1. **Introduction to networks**
   - Network types: regulatory, metab., signal., interact., func.
   - Bayesian (probabilistic) and Algebraic views
2. **Network Centrality Measures**
   - Local centrality metrics (degree, betweenness, closeness, etc)
   - Global centrality metrics (eigenvector centrality, page-rank)
3. **Linear Algebra Review: eigenvalues, SVD, low-rank approximations**
   - Eigenvector and singular vector decomposition
   - Low rank approximations, Wigner semicircle law
4. **Sparse Principal Component Analysis**
   - Lasso and Elastic lasso
   - PCA and Sparse PCA
5. **Machine Learning in Networks**
   - Guilt by association
   - Maximum cliques, density-based modules and spectral clustering
6. **Network Diffusion Kernels and Deconvolution**
   - Network diffusion kernels
   - Network deconvolution