6.874, 6.802, 20.390, 20.490, HST.506
Computational Systems Biology
Deep Learning in the Life Sciences

Lecture 09: Predicting gene expression and splicing

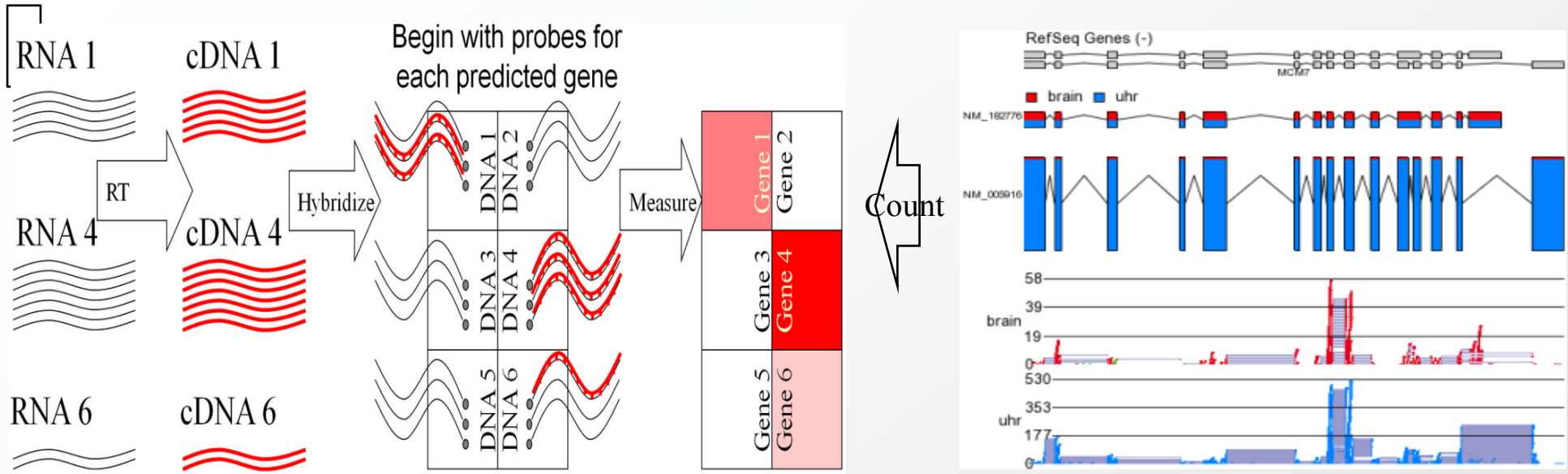
Prof. Manolis Kellis

Massachusetts Institute of Technology

http://mit6874.github.io

Slides credit: David Gifford, et al

RNA-Seq: De novo tx reconstruction / quantification



Microarray technology

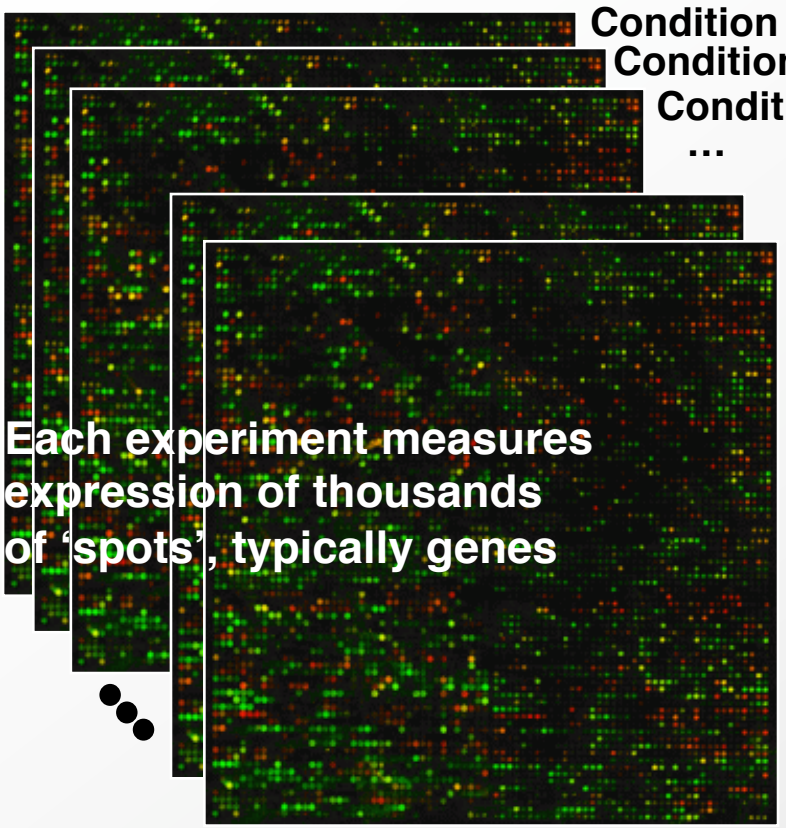
- Synthesize DNA probe array, complementary hybridization
- Variations:
  - One long probe per gene
  - Many short probes per gene
  - Tiled k-mers across genome
- Advantage:
  - Can focus on small regions, even if few molecules / cell

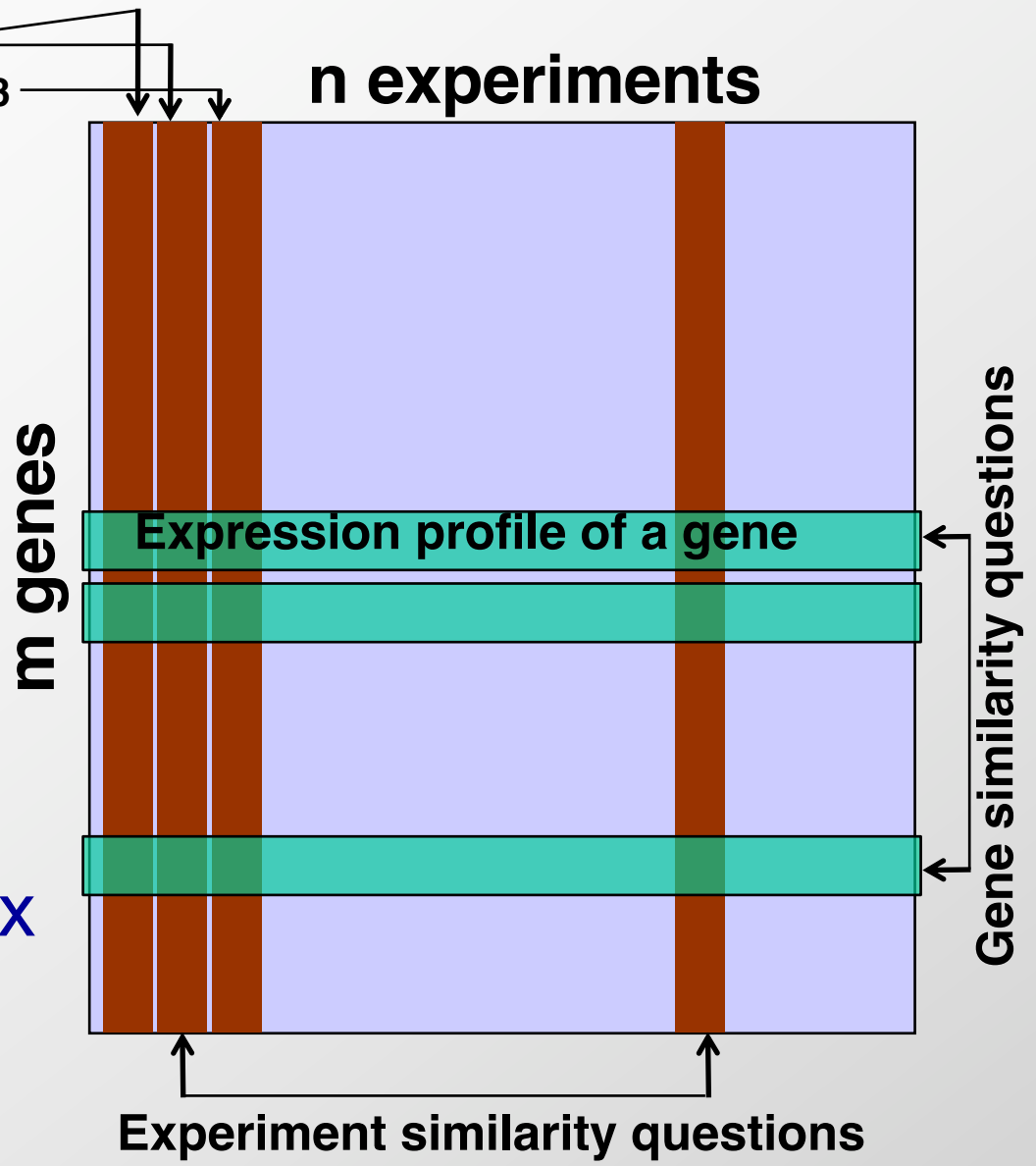
RNA-Seq technology:

- Sequence short reads from mRNA, map to genome
- Variations:
  - Count reads mapping to each known gene
  - Reconstruct transcriptome *de novo* in each experiment
- Advantage:
  - Digital measurements, de novo

Expression Analysis Data Matrix

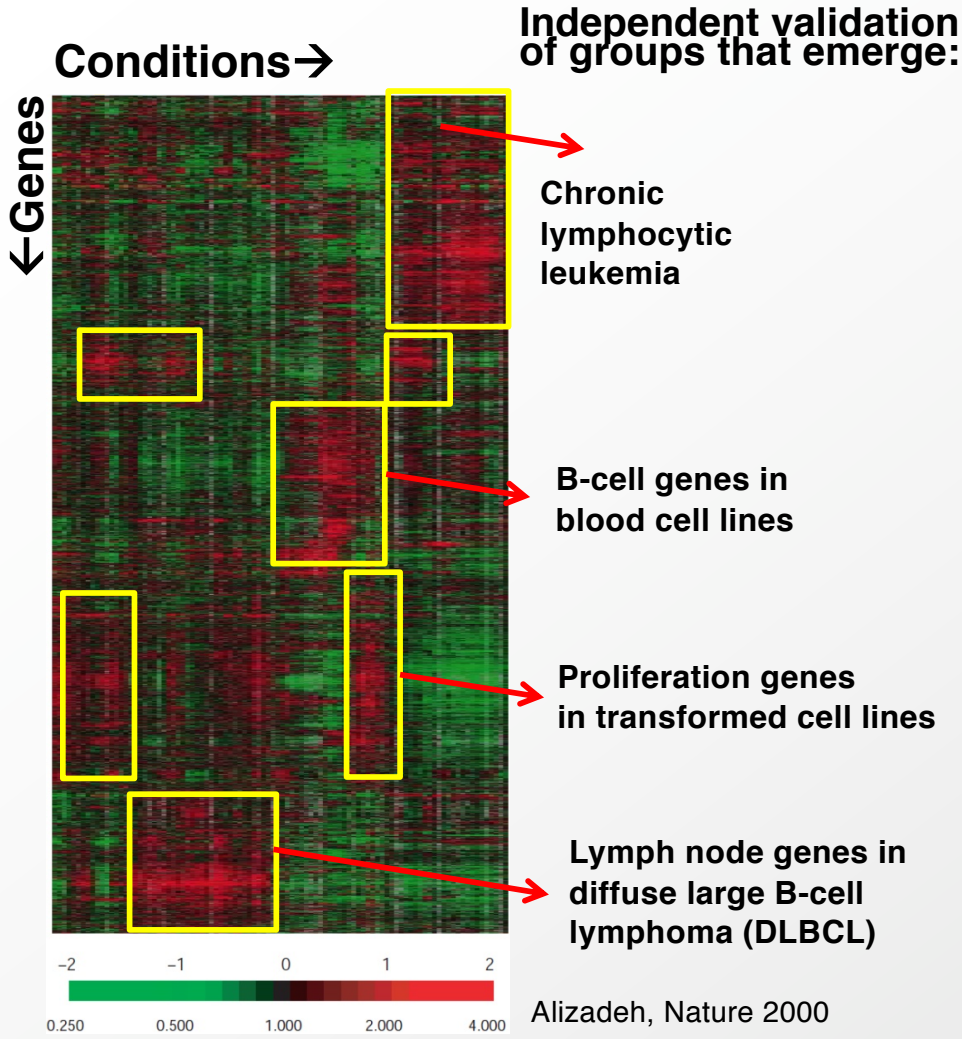
- Measure 20,000 genes in 100s of conditions

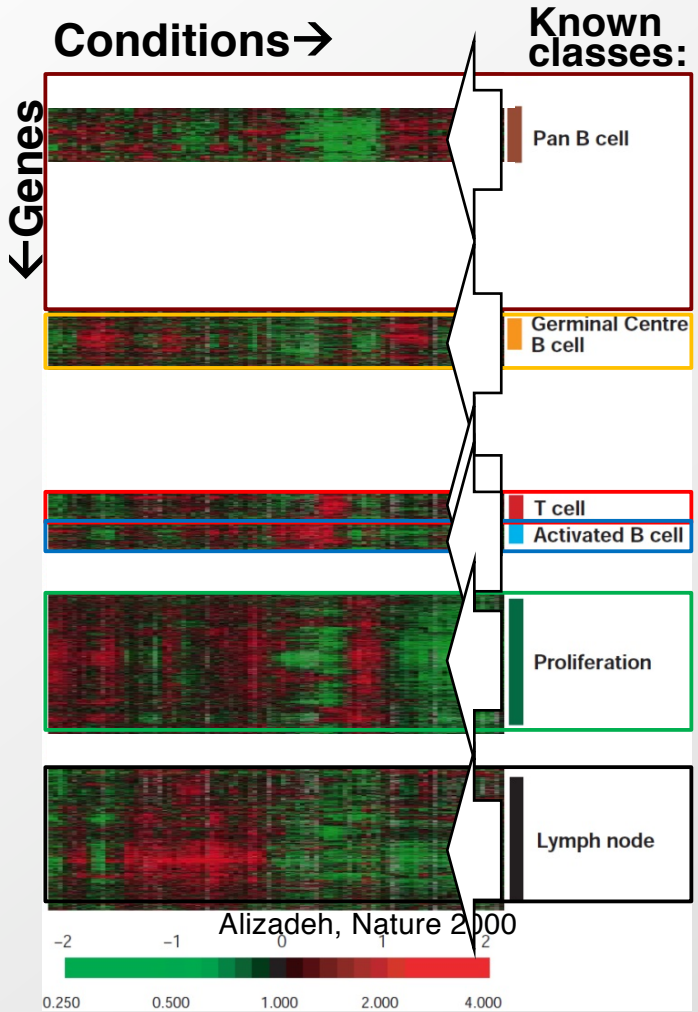




- Study resulting matrix

Clustering vs. Classification



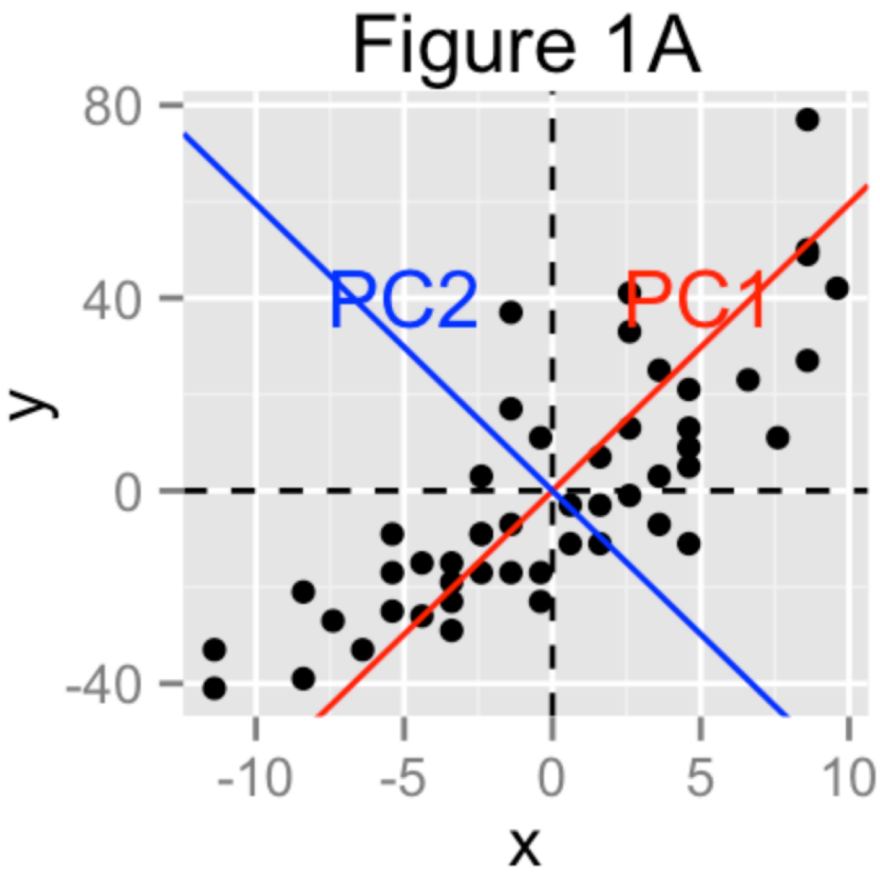


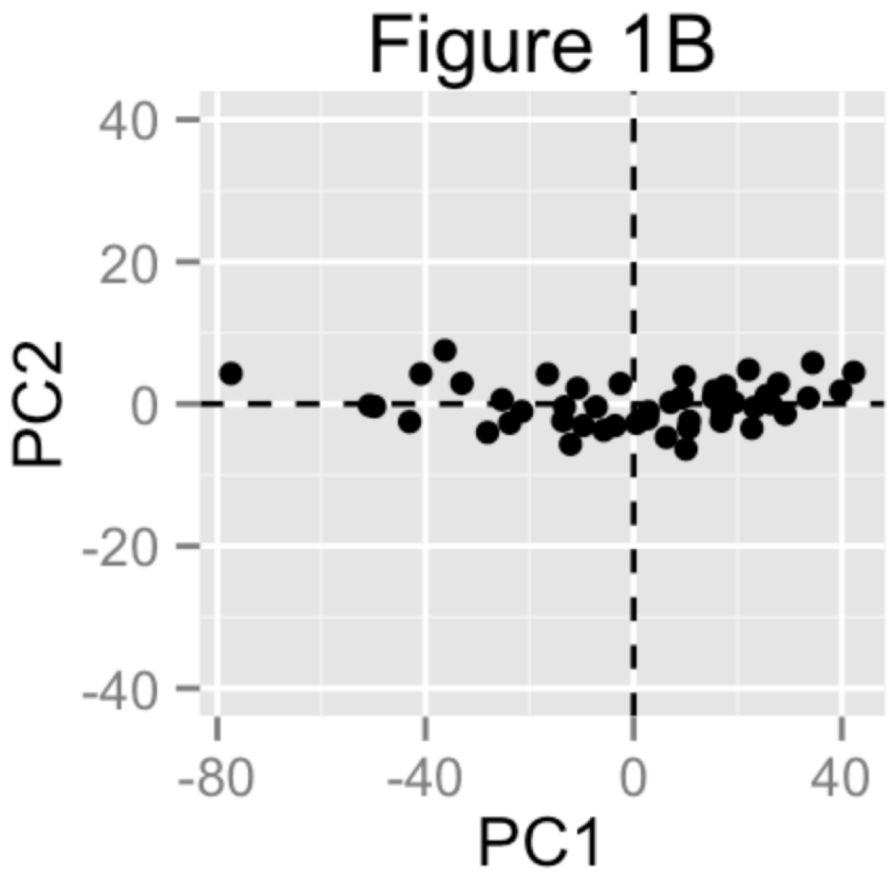
Goal of Clustering: **Group similar items** that likely come from the same category, and in doing so **reveal hidden structure**

Goal of Classification: Extract features from the data that best **assign new elements** to ≥1 of **well-defined classes**

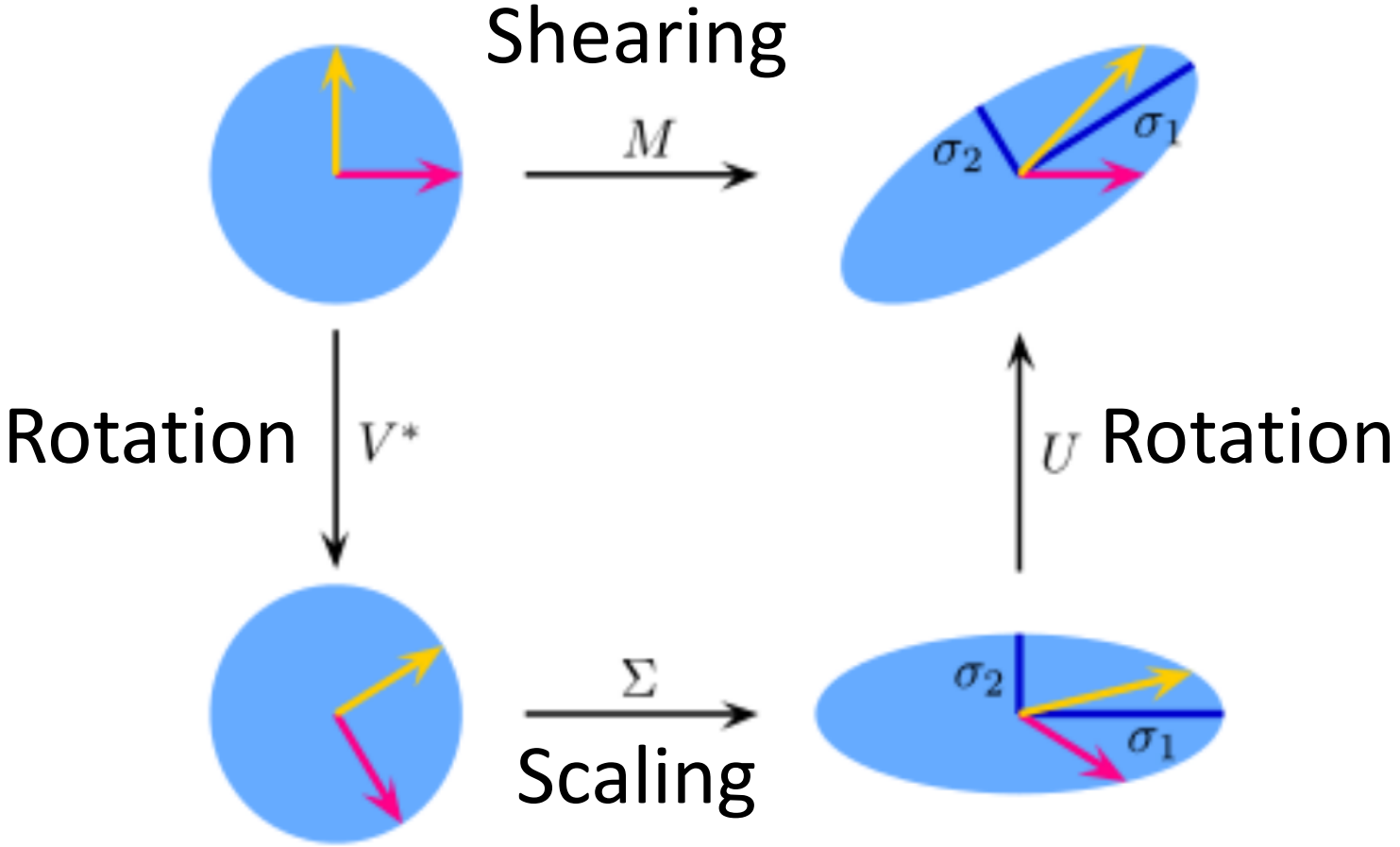
- **Unsupervised learning**
- **Supervised learning**

PCA, Dimensionality reduction





Geometric interpretation of SVD



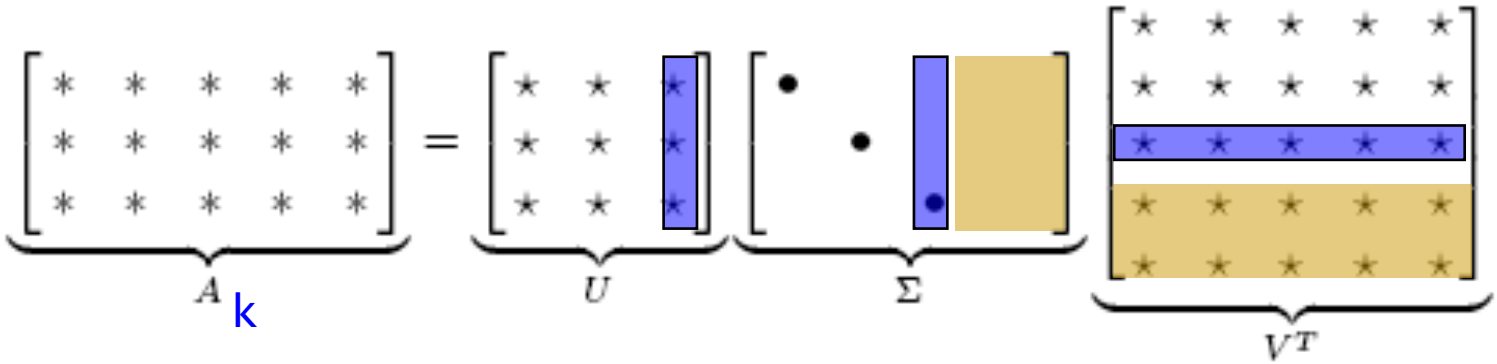
$$M = U \cdot \Sigma \cdot V^*$$

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

Low-rank Approximation

- Solution via SVD

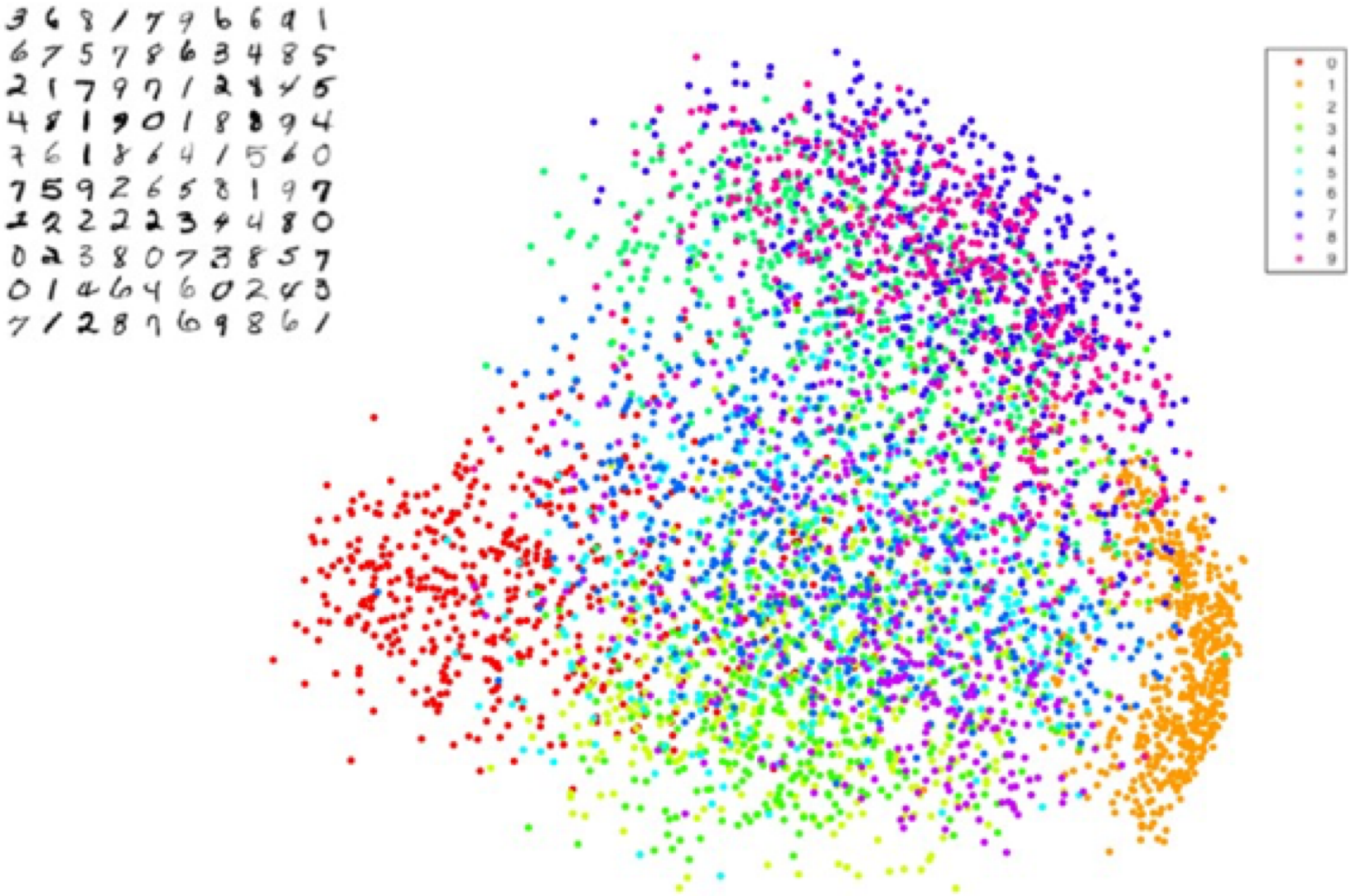
$$A_k = U\,\mathrm{diag}(\sigma_1,\ldots,\sigma_k,\underbrace{0,\ldots,0}_{\text{set smallest r-k singular values to zero}})V^T$$



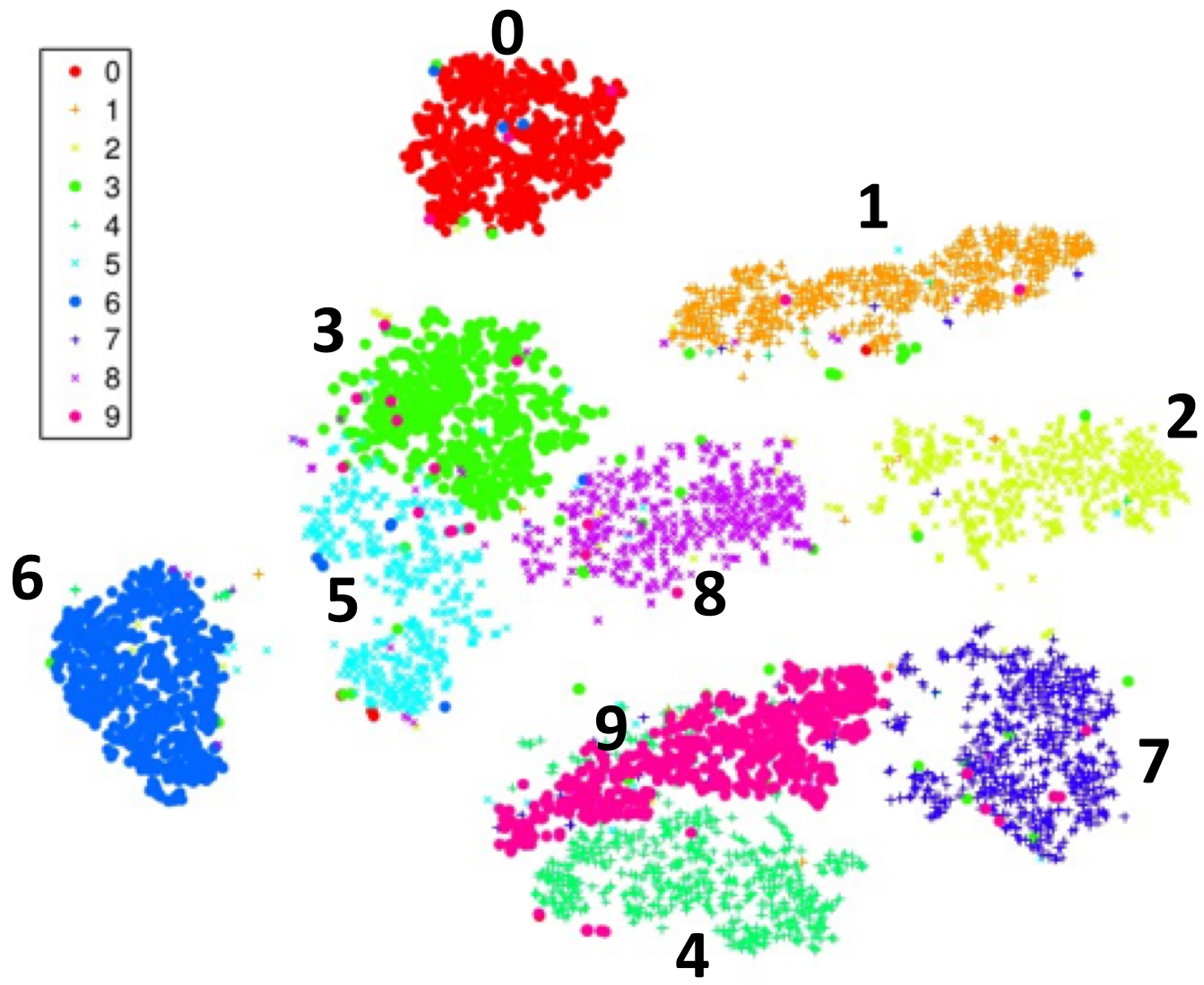
$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$ ← *column notation: sum of rank 1 matrices*

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

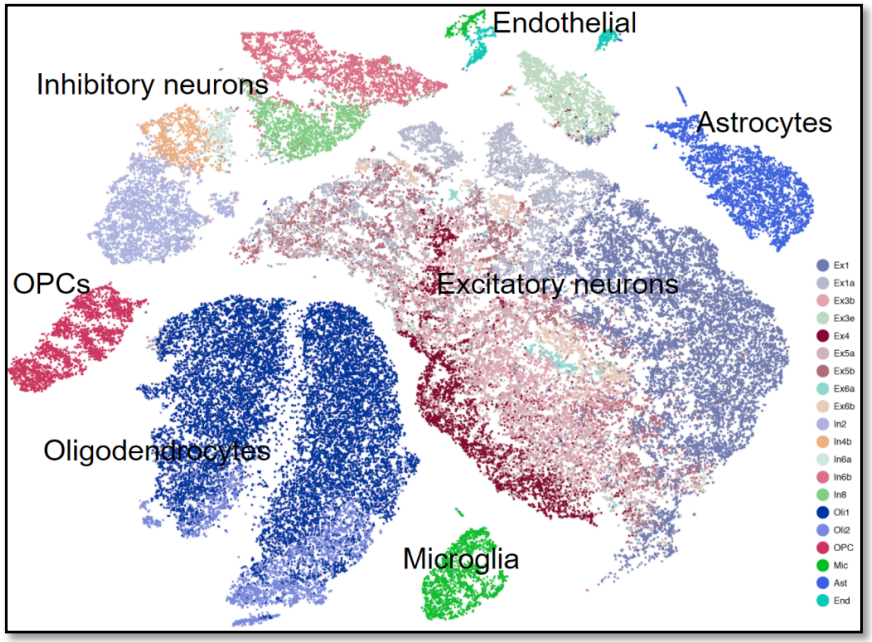
PCA of MNIST digits



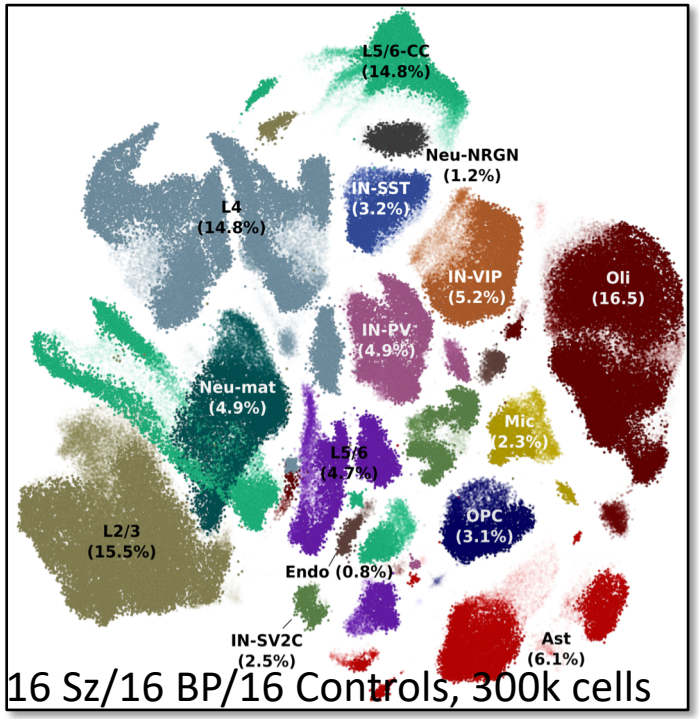
t-SNE of MNIST digits



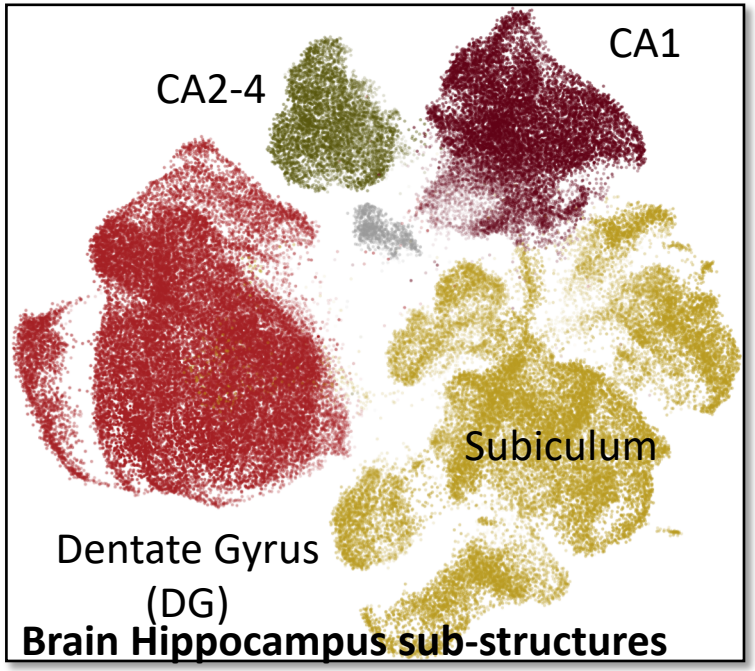
t-SNEs of single-cell Brain data



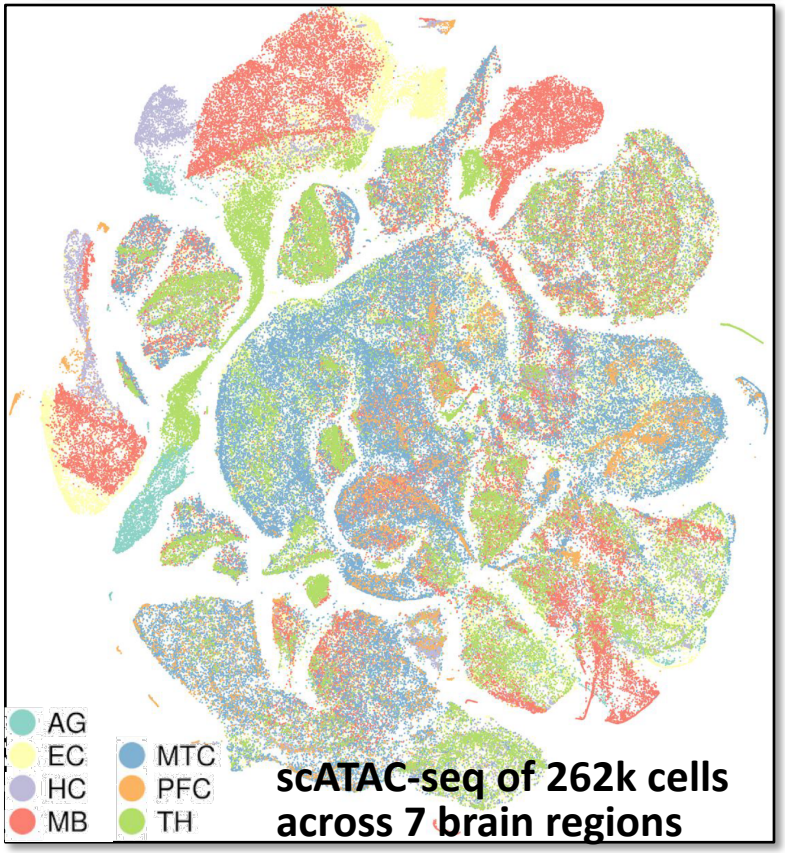
scRNA-seq in 48 individuals, 84k cells, Nature, 2019



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

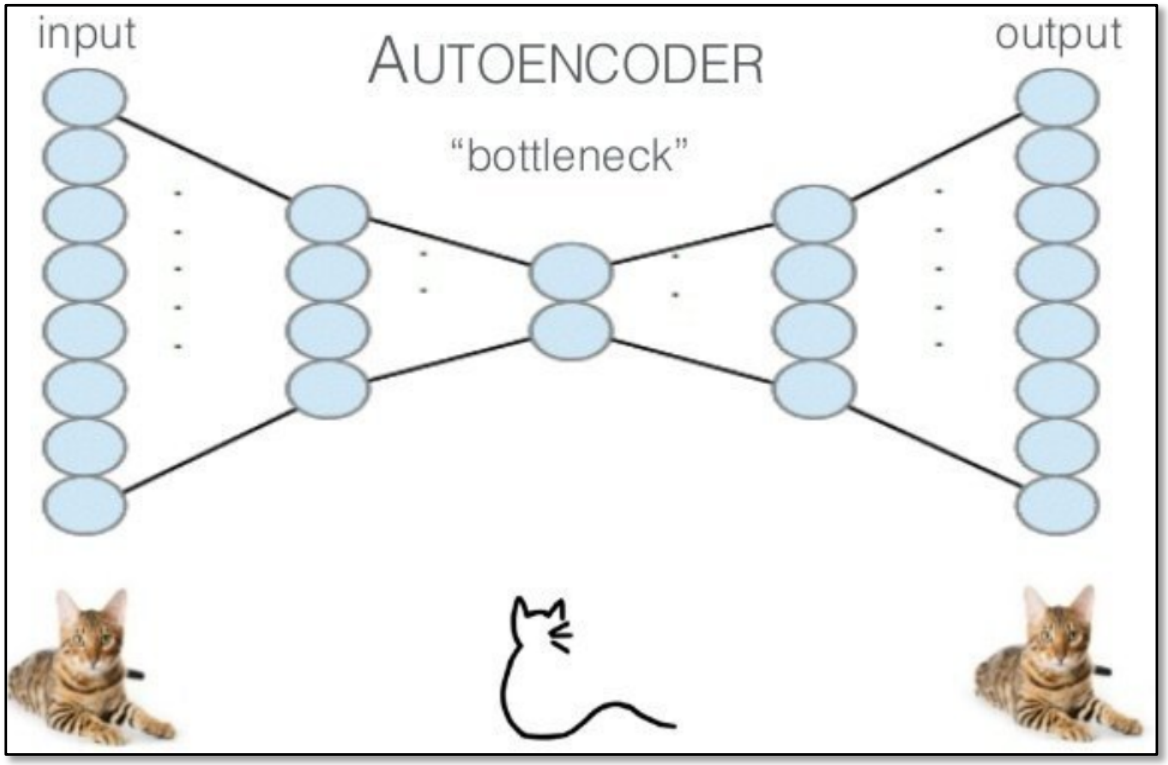


Brain Hippocampus sub-structures



scATAC-seq of 262k cells across 7 brain regions

Autoencoder: dimensionality reduction with neural net



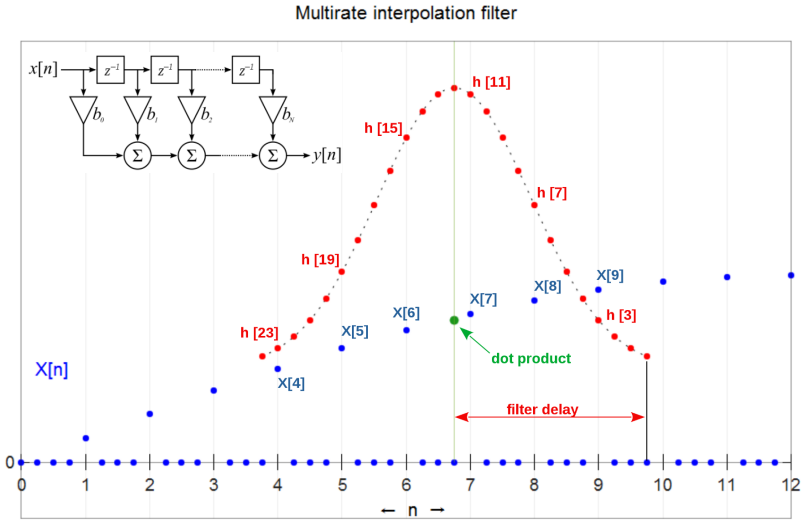
- Tricking a **supervised** learning algorithm to work in **unsupervised** fashion
- Feed input as output function to be learned. **But!** Constrain model complexity

- **Pretraining** with RBMs to learn representations for future supervised tasks. Use RBM output as "data" for training the next layer in stack
- After pretraining, "unroll" RBMs to create deep autoencoder
- Fine-tune using backpropagation

[Hinton *et al*, 2006]

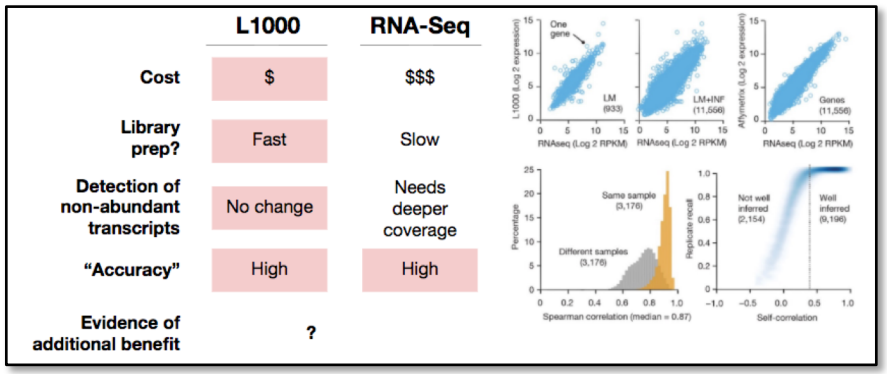
1. Up-sampling gene expression patterns

Challenge: Measure few values, infer many values

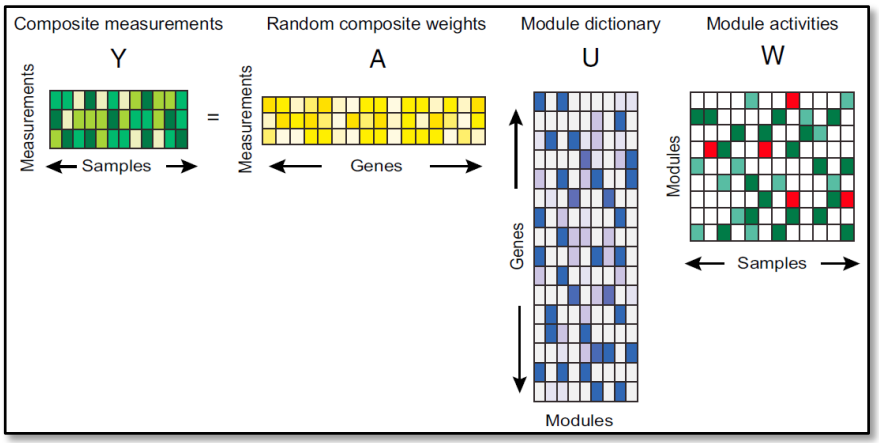


- Digital signal upscaling
  - Interpolating low-pass filter (e.g. FIR finite impulse response)
  - Low-dim. capture of higher-dim. signal
  - Nyquist rate (discrete) / freq. (contin.)

- Image up-scaling
  - Inverse of convolution (de-convolution)
  - Transfer learning from corpus of images
  - Low-dim. re-projection to high-dim. img

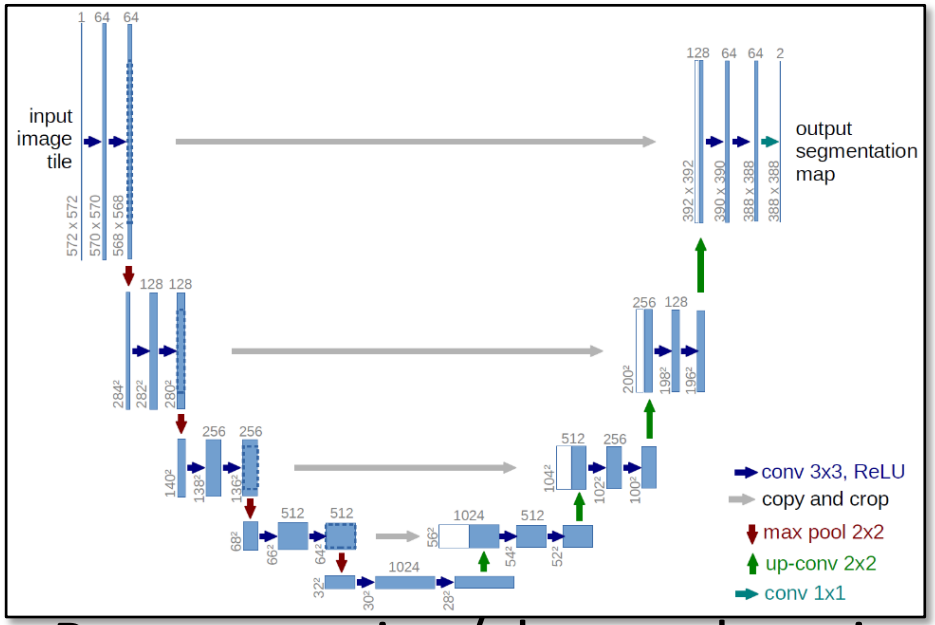


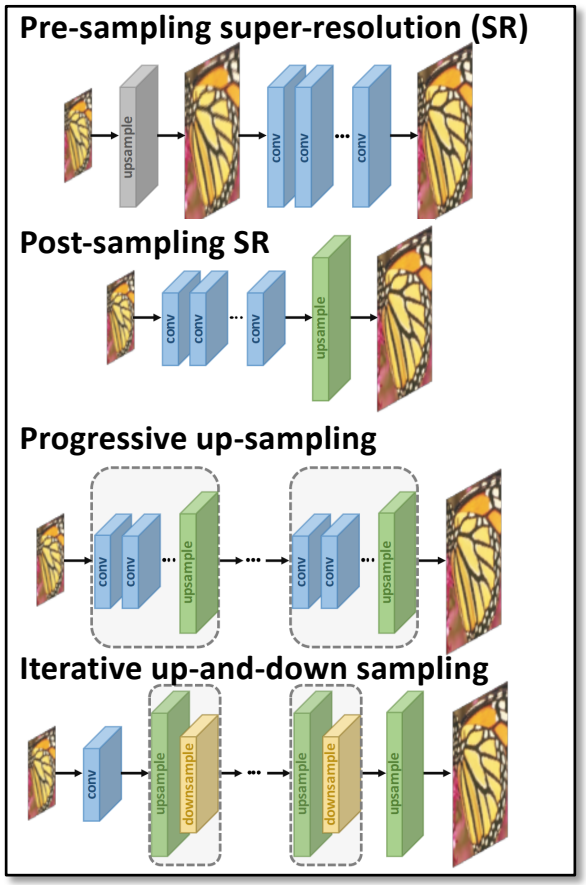
- Gene expression measurements
  - Measure 1000 genes, infer the rest
  - Rapid, cheap, reference assay
  - Apply to millions of conditions



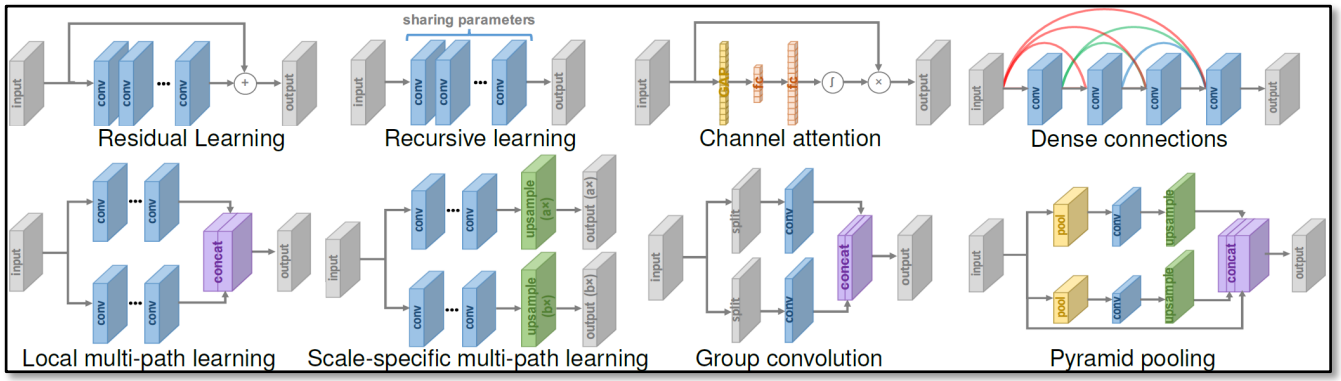
- ***Which*** 1000 genes? Compressed sensing
  - Measure few ***combinations*** of genes
  - Better capture high-dimensional vector

Deep Learning architectures for up-sampling images

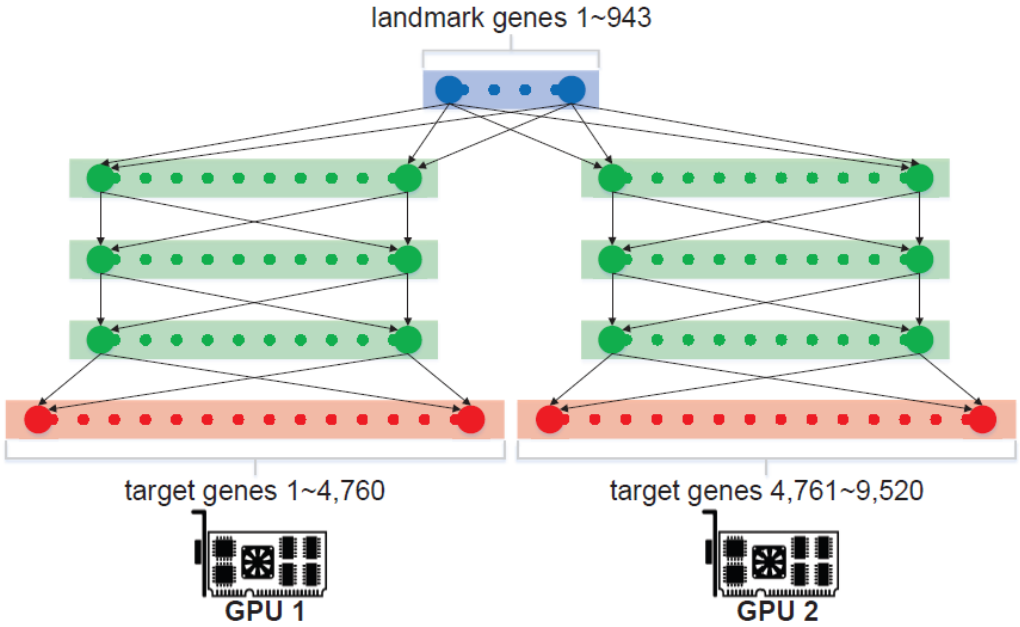




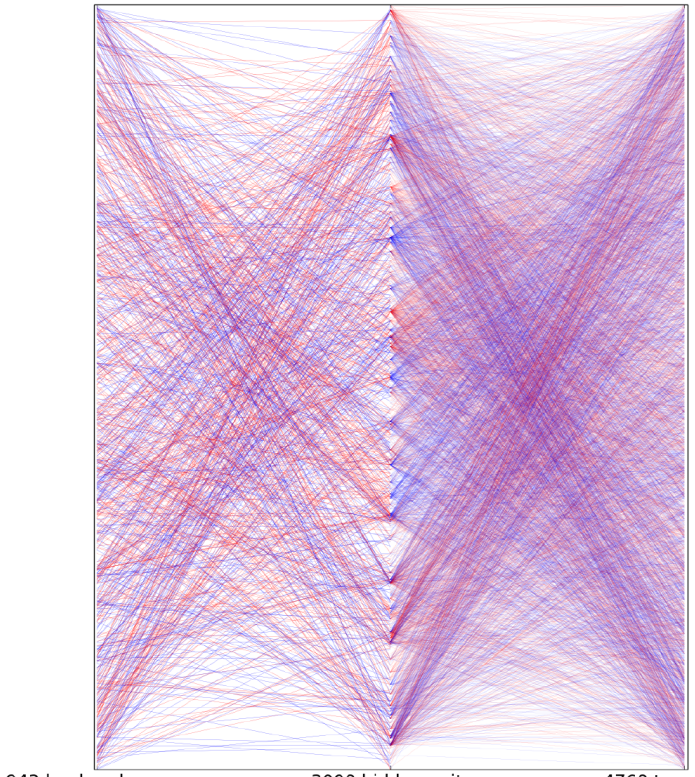
- Representation/abstract learning
  - Enables compression, re-upscaling, denoising
  - Example: autoencoder bottleneck. High-low-high
  - Modification: de-compression, up-scaling, low-high only



D-GEX - Deep Learning for up-scaling L1000 gene expression



Parameters	
# of hidden layers	[1, 2, 3]
# of hidden units in each hidden layer	[3000, 6000, 9000]
Dropout rate	[0%, 10%, 25%]
Momentum coefficient	0.5
Initial learning rate [a]	5e-4 or 3e-4
Minimum learning rate	1e-5
Learning rate decay factor	0.9
Learning scale [b]	3.0
Mini-batch size	200
Training epoch	200
Weights initial range [c]	$\left[-\frac{\sqrt{6}}{\sqrt{n_i+n_o}}, \frac{\sqrt{6}}{\sqrt{n_i+n_o}}\right]$



Gene expression inference with deep learning FREE

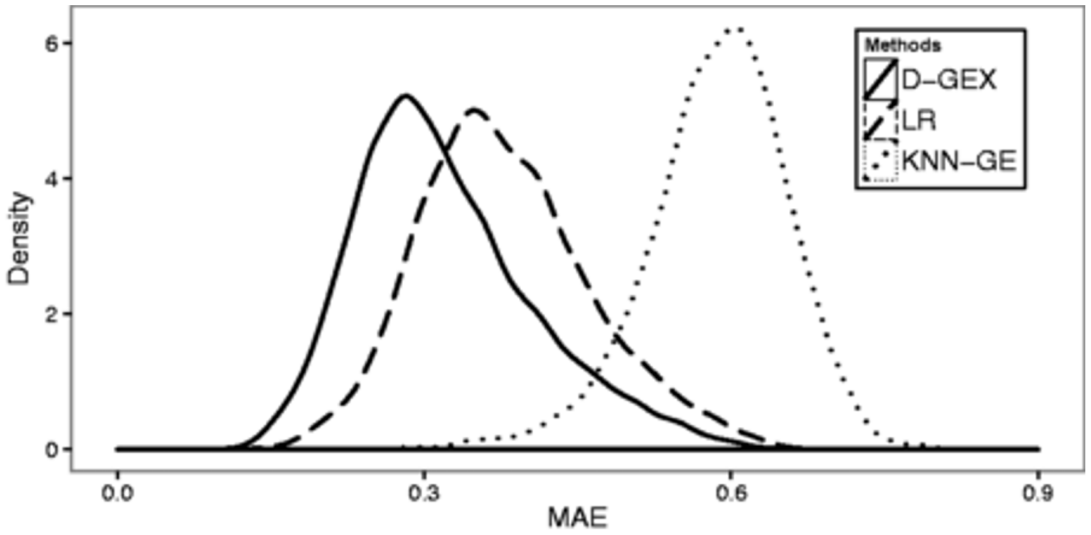
Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, Xiaohui Xie ✉ Author Notes

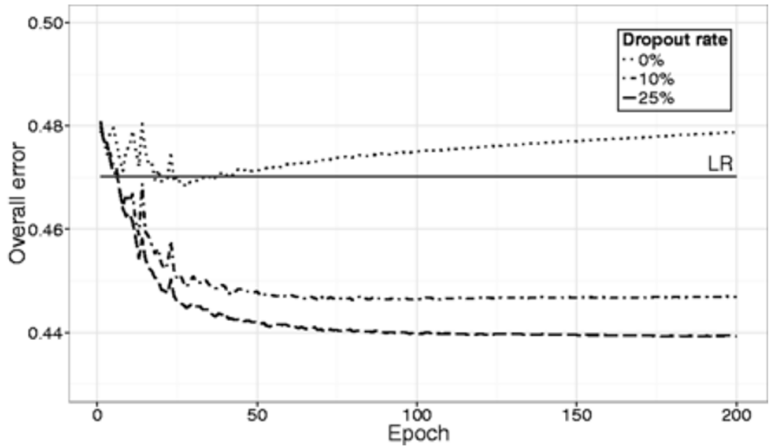
Bioinformatics, Volume 32, Issue 12, 15 June 2016, Pages 1832–1839,
https://doi.org/10.1093/bioinformatics/btw074

Published: 11 February 2016 **Article history** ▾

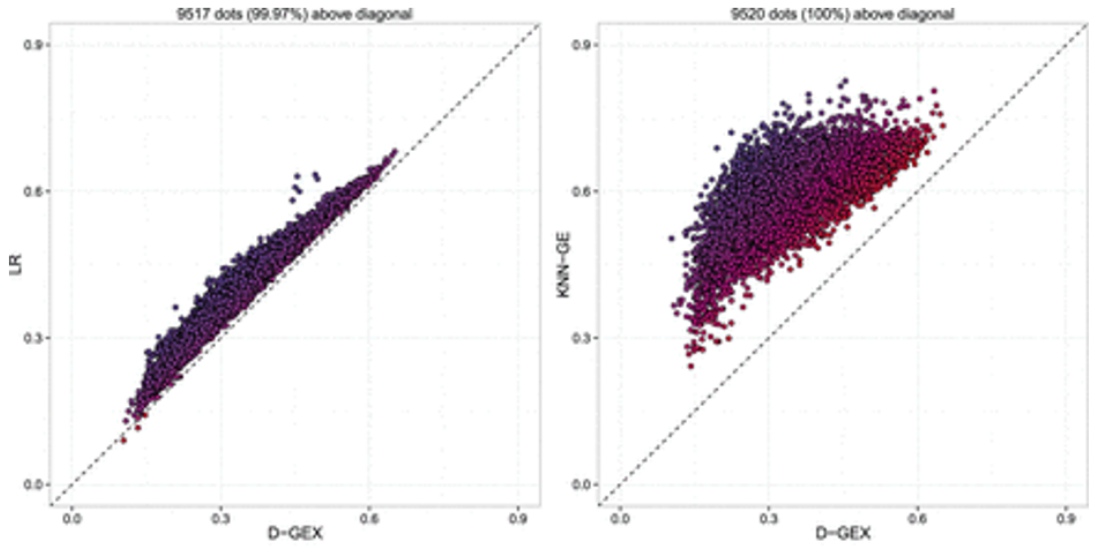
- Multi-task Multi-Layer Feed-Forward Neural Net
- Non-linear activation function (hyperbolic tangent)
- Input: 943 genes, Output: 9520 targets (partition to fit in memory)

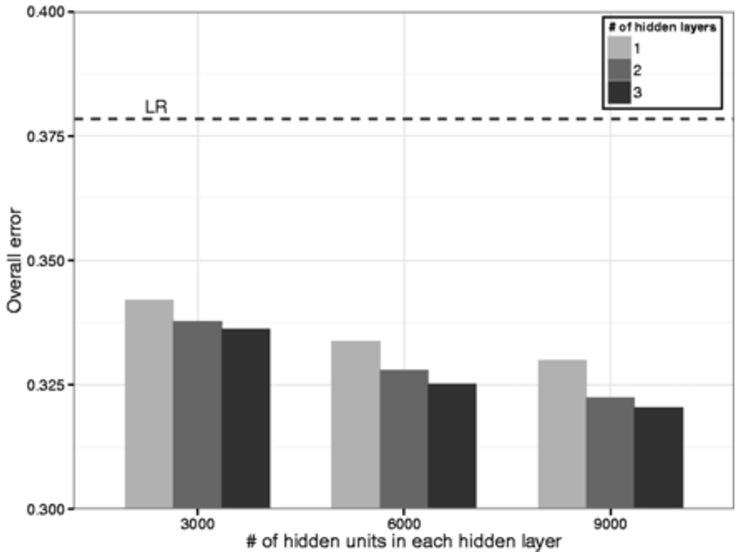
D-GEX outperforms Linear Regression or K-nearest-Neighbors





- Lower error than LR or KNN
- Training rapidly converges





- Strictly better for nearly all genes
- Deeper = better

However: performance still not great, computational limitations

2. Composite measurements for compressed sensing

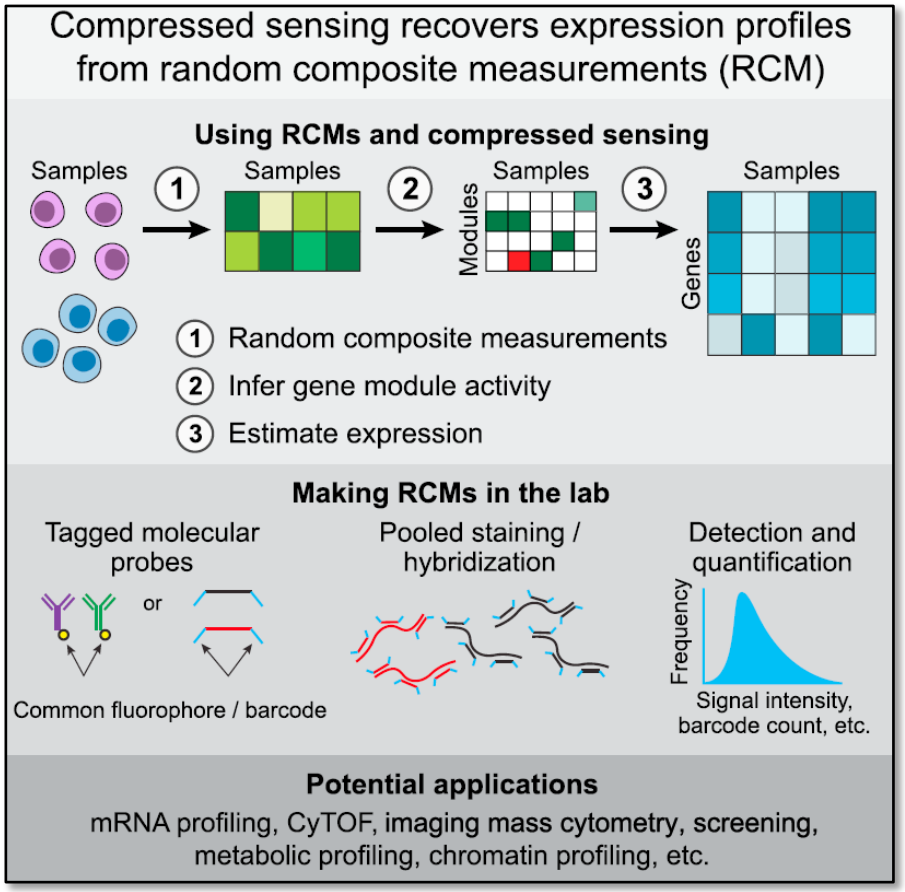
Key insight: Composite measurements better capture modules

Cell

Theory

Efficient Generation of Transcriptomic Profiles by Random Composite Measurements

Brian Cleary,[1,2] Le Cong,[1] Anthea Cheung,[1] Eric S. Lander,[1,3,4] and Aviv Regev[1,3,5,6,*]
[1]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[2]Computational and Systems Biology Program, MIT, Cambridge, MA, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA
[4]Department of Systems Biology, Harvard Medical School, Boston, MA, USA



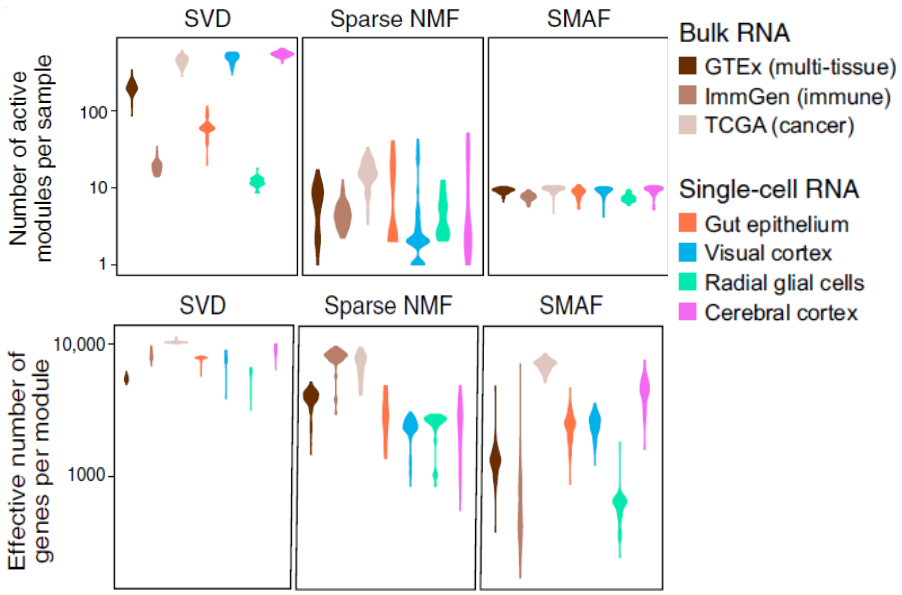
gene expression matrix $X \in \mathbb{R}^{gxn}$

module dictionary $U \in \mathbb{R}^{gxd}$

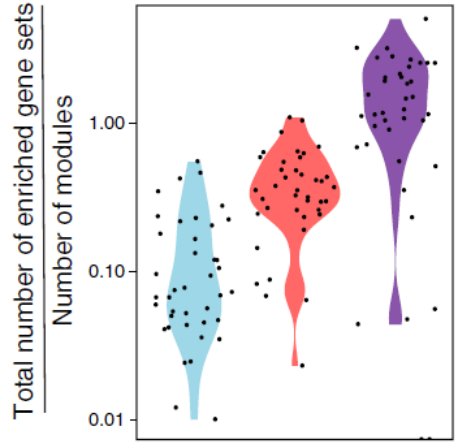
module activity matrix $W \in \mathbb{R}^{dxn}$

$$\min_{U,W} \|X - UW\|^2 + \lambda \|U\|_1$$

such that $u_{ij} \geq 0, \|u_{:j}\| = 1,$ and $\|w_i\|_0 \leq k \ \forall \ i \in \{1, \ldots, n\}$



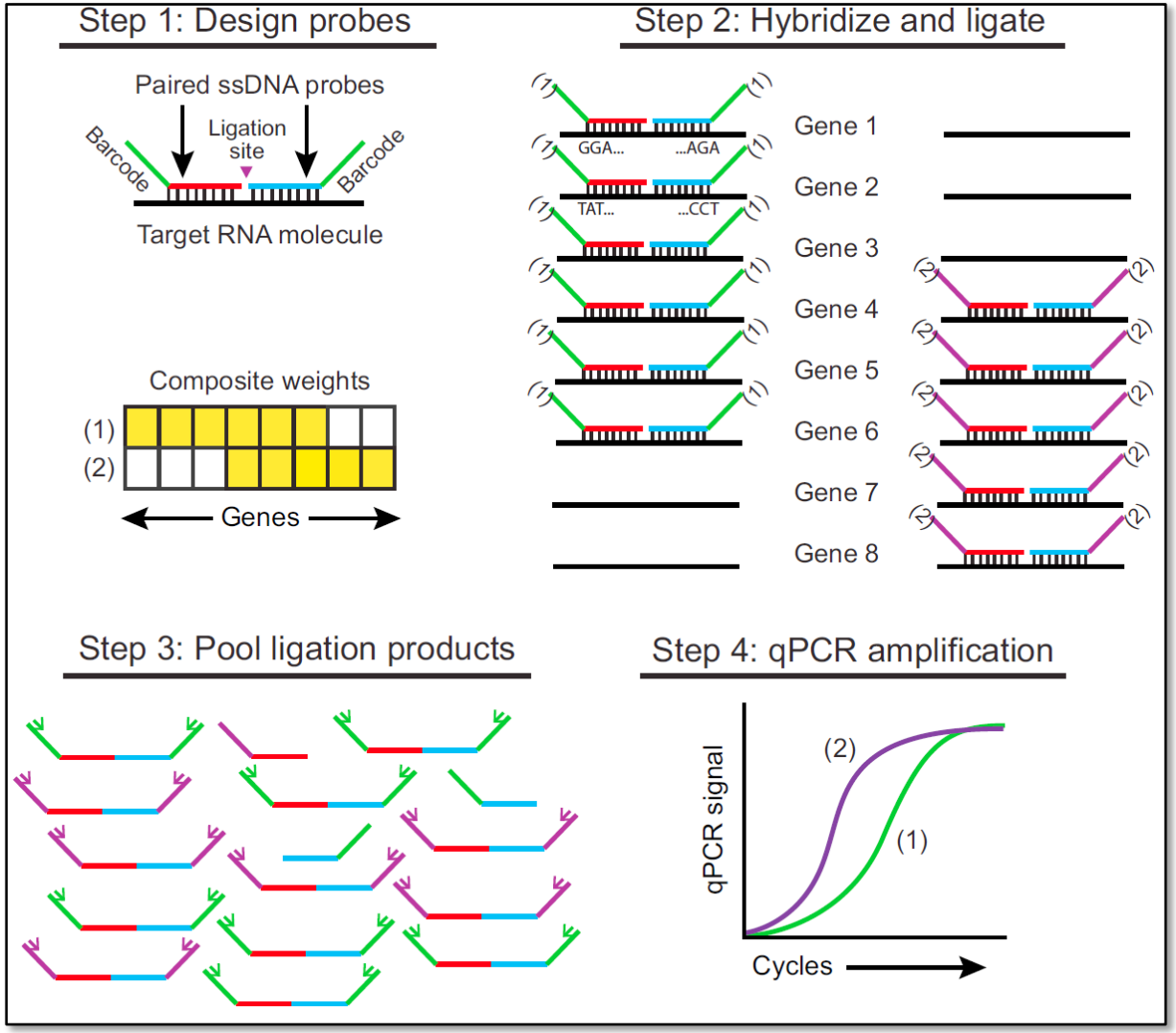
- Sparse Module Activity Factorization (SMAF)

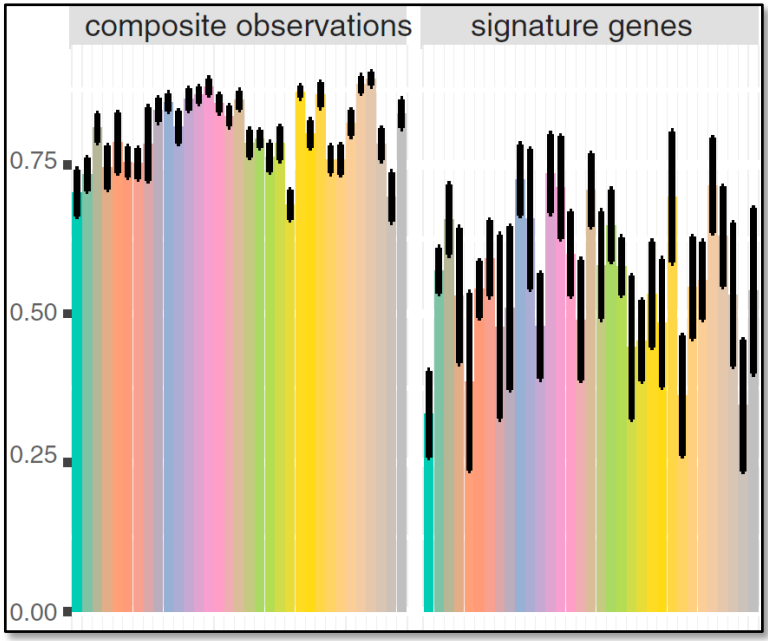


Algorithm: Sparse Module Activity Factorization

1. SMAF(X, d, λ, k)
2. Initialize $U \in \mathbb{R}^{g \ x \ d}$ and $W \in \mathbb{R}^{d \ x \ n}$ randomly.
3. For 10 iterations:
   a. Update the module dictionary as $U = LassoNonnegative(X, W, \lambda)$.
   b. Normalize each module so that $\|u_i\|_2 = 1$.
   c. Update the activity levels as $W = OMP(X, U, k)$.
4. Return U, W.

Making composite measurements in practice

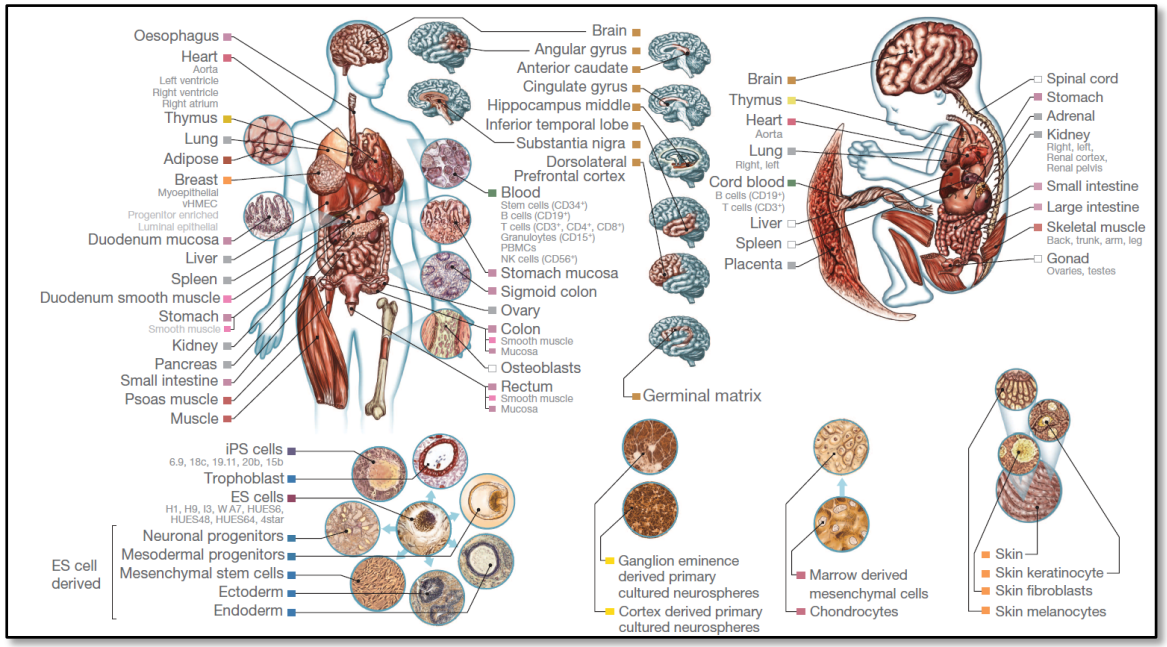


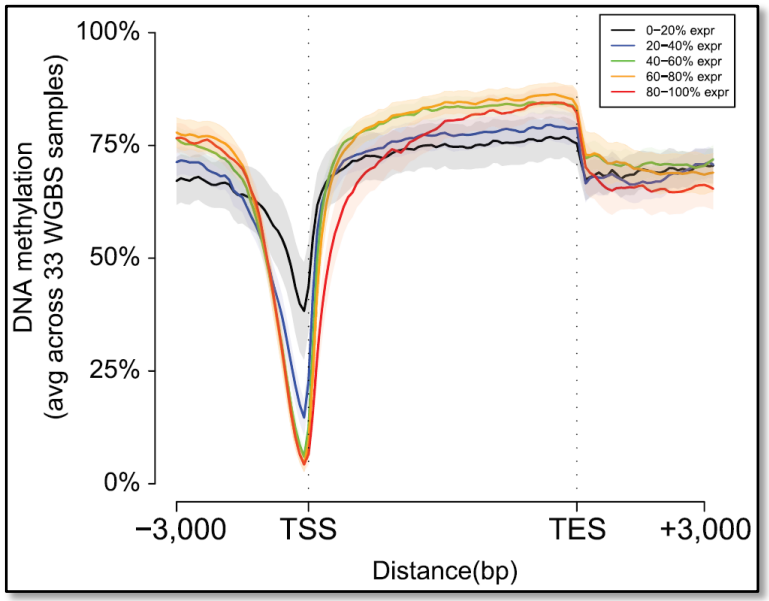


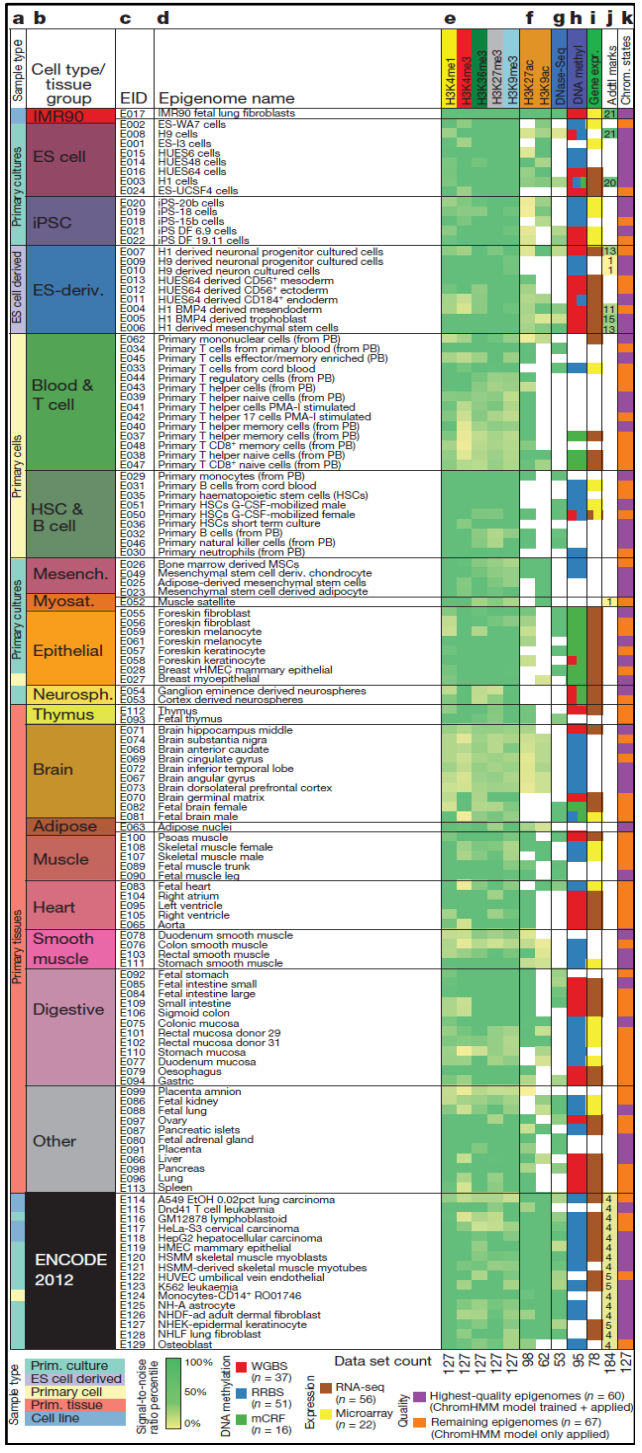
- Combinations of probes + barcodes for measurement
- More consistent signal-to-noise ratios

3. Predicting Expression from Chromatin

Can we predict gene expression from chromatin information?

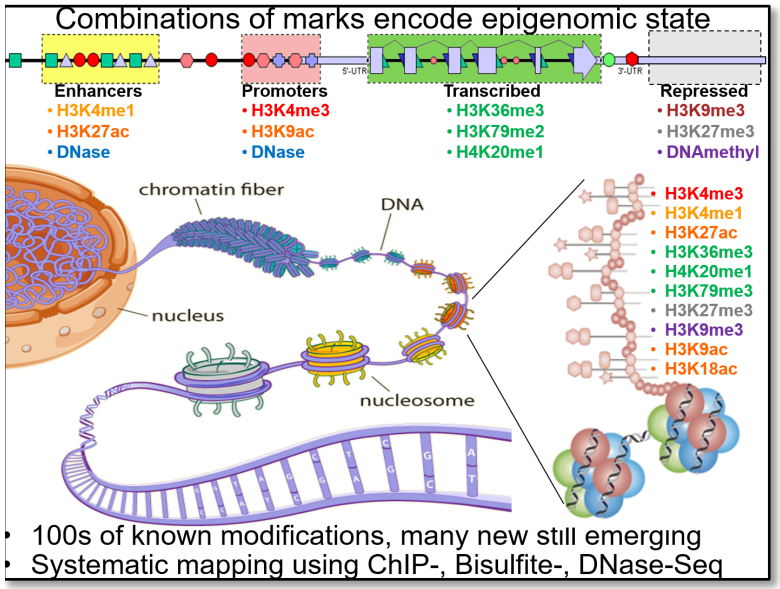


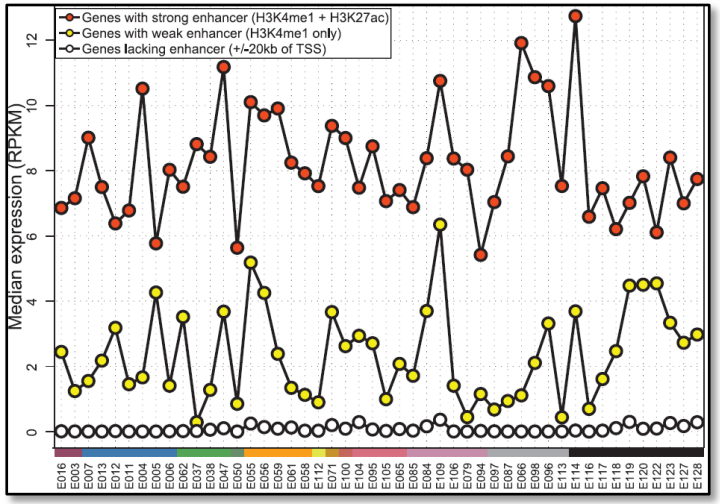


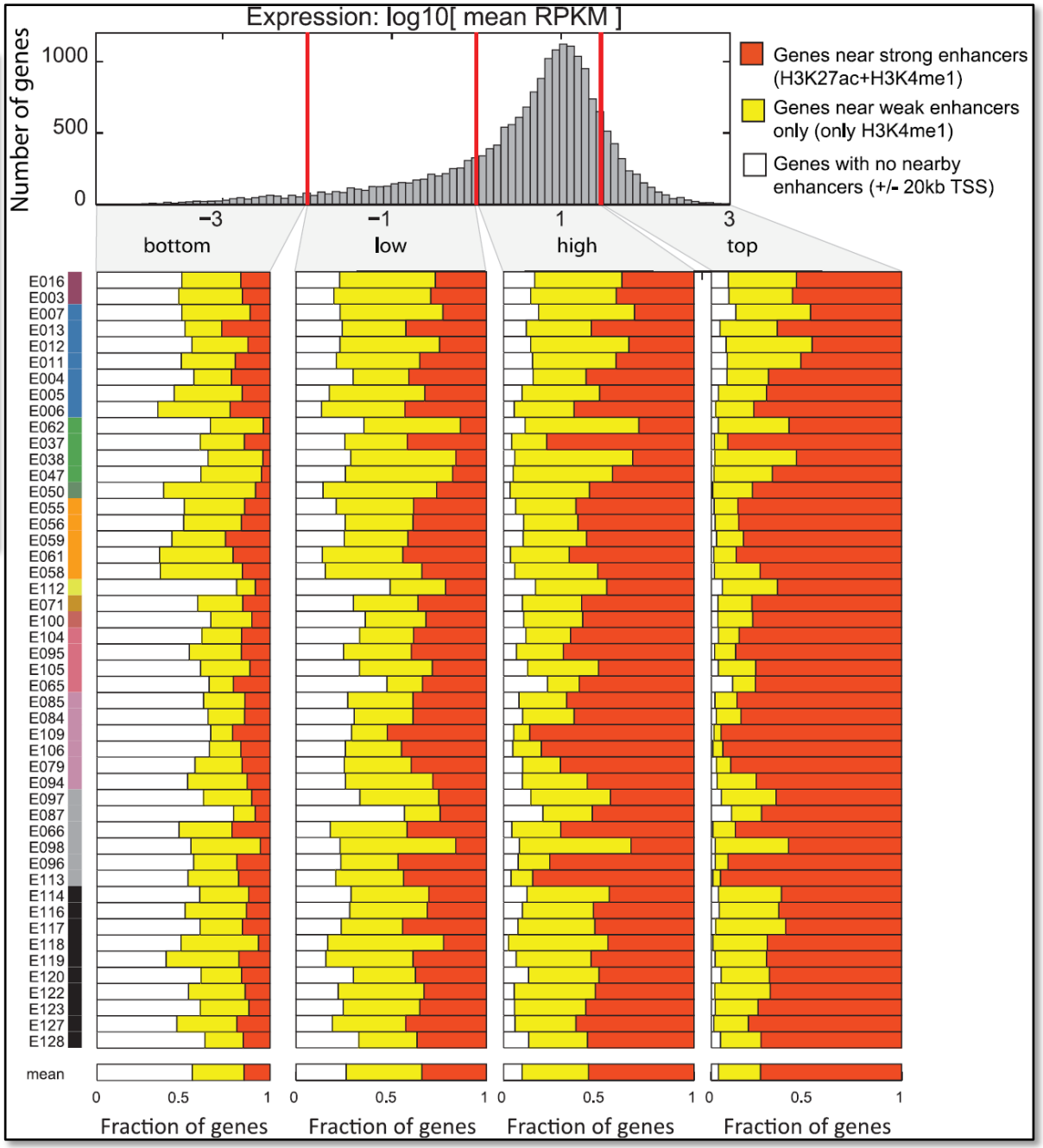


- DNA methylation vs. gene expression
- Promoters: high. Gene body: low

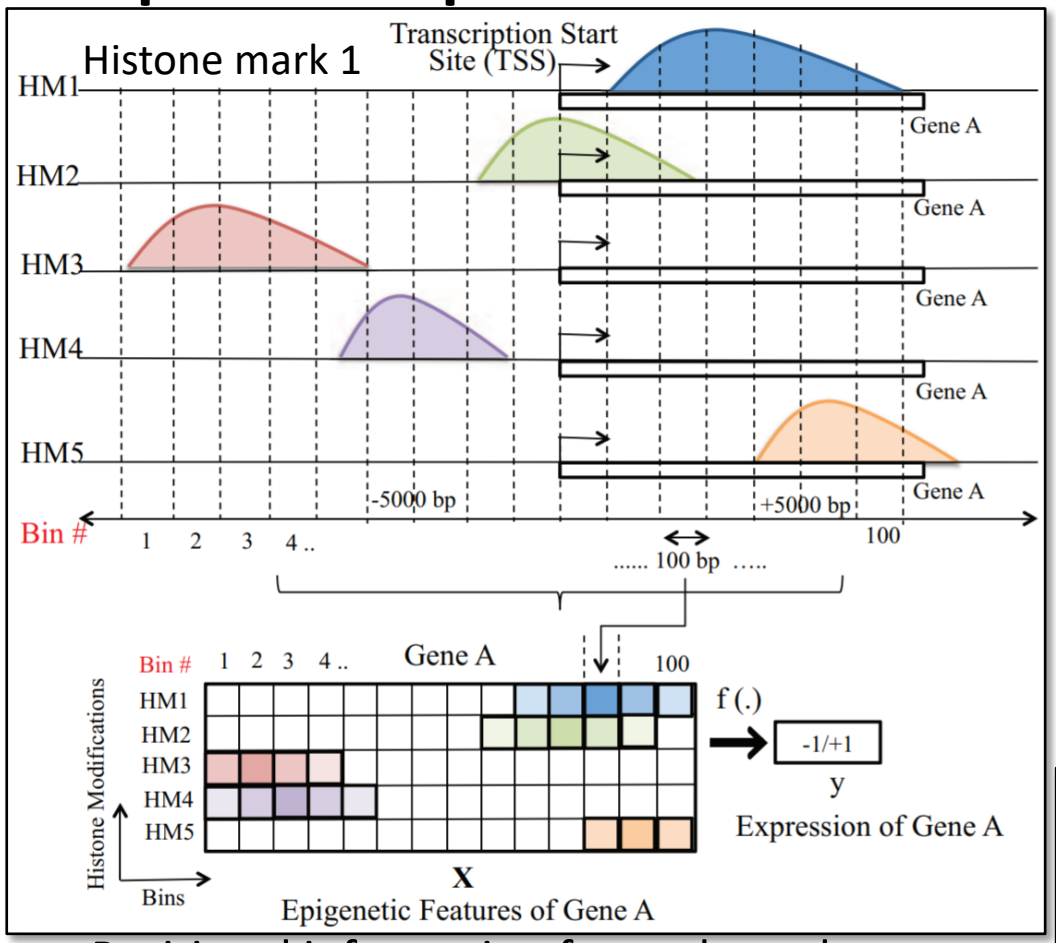
Strong enhancers (+H3K27ac) vs. weak enhancers (H3K4me1 only)



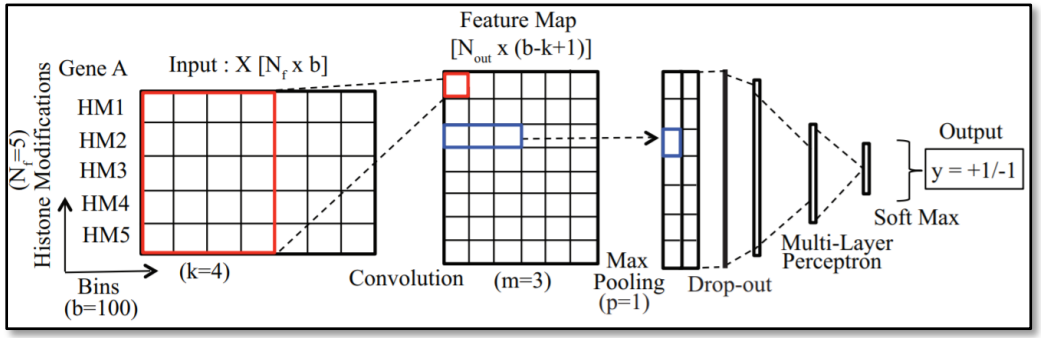




DeepChrome: positional histone features predictive of expression



- Positional information for each mark



- Convolution, pooling, drop-out, Multi-Layer-Perceptron (MLP) alternating lin/non-linear

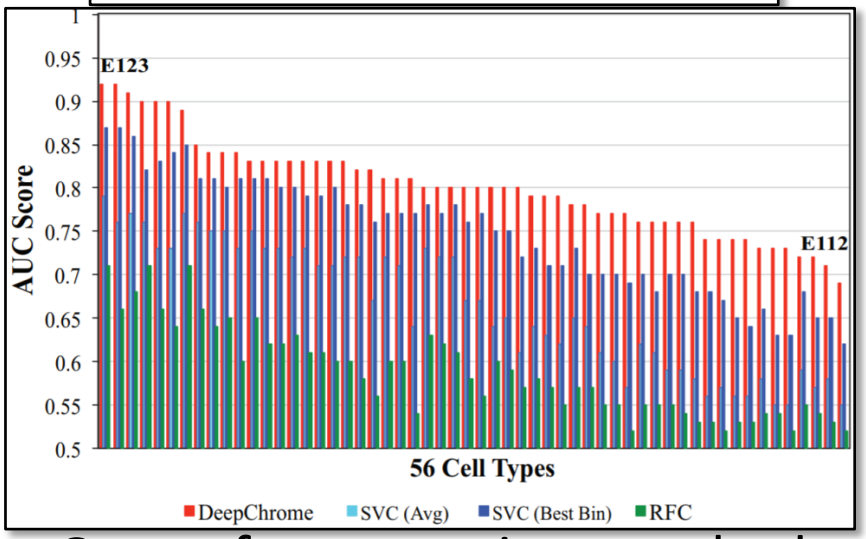
DeepChrome: Deep-learning for predicting gene expression from histone modifications.

Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi *

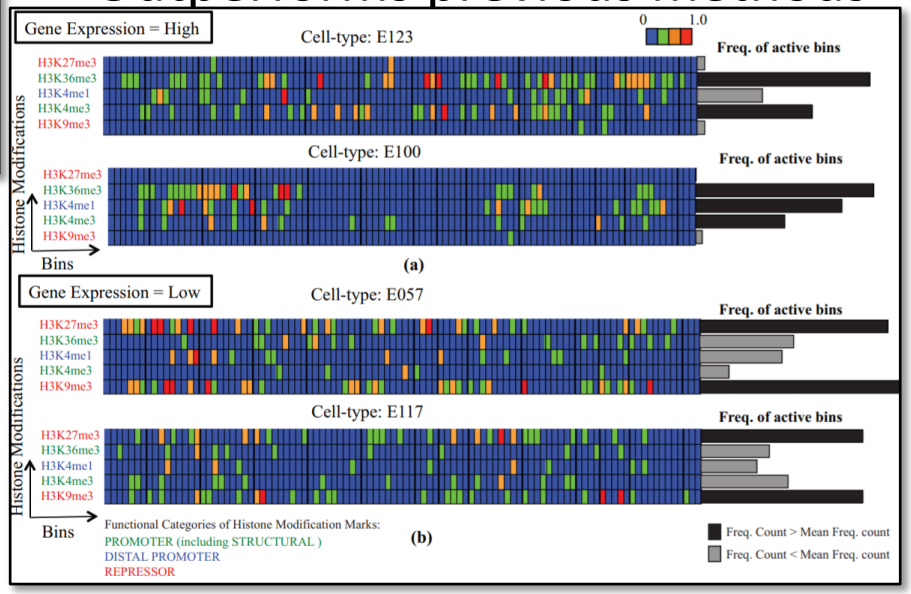
Department of Computer Science, University of Virginia, Charlottesville, VA, U.S.A

* To whom correspondence should be addressed.

This work will be published originally in Bioinformatics Journal at
http://bioinformatics.oxfordjournals.org

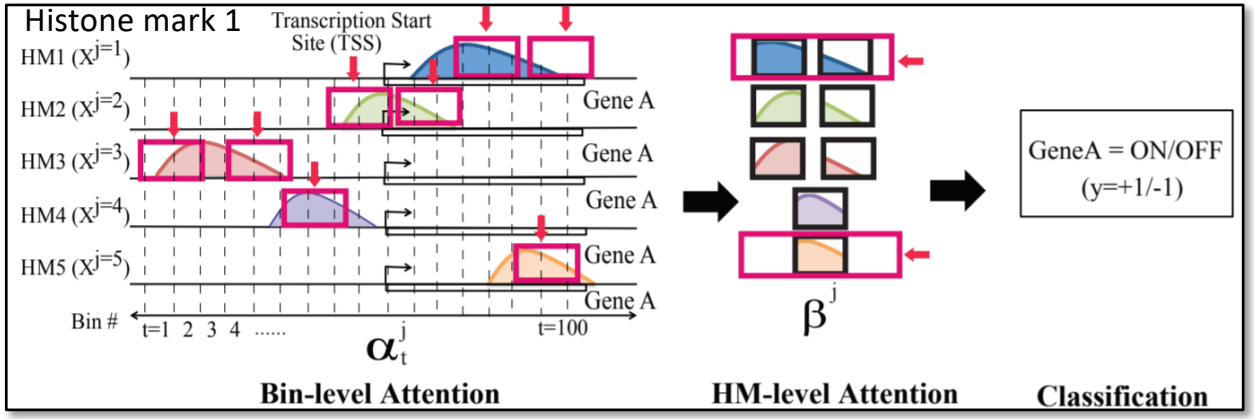


- Outperforms previous methods



- Meaningful features selected

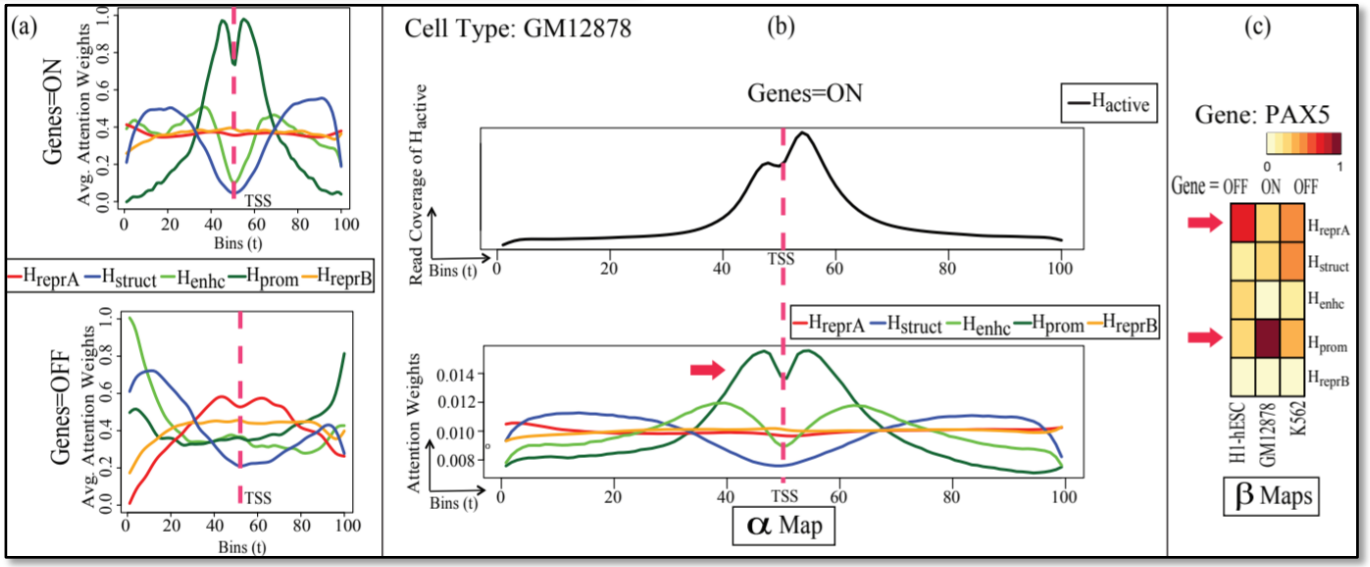
AttentiveChrome: Selectively attend to specific marks/positions



Attend and Predict: Understanding Gene Regulation by Selective Attention on Chromatin

Ritambhara Singh, Jack Lanchantin, Arshdeep Sekhon, Yanjun Qi
Department of Computer Science
University of Virginia
yanjun@virginia.edu

- Attention: LSTM: Long short-term memory module
- Hierarchical LSTM modules: interactions across marks



- Attention focuses on specific positions for specific marks

	Baselines		AttentiveChrome Variations				
Model	DeepChrome (CNN) [29]	LSTM	CNN-Attn	CNN-α,β	LSTM-Attn	LSTM-α	LSTM-α,β
Mean	0.8008	0.8052	0.7622	0.7936	0.8100	**0.8133**	0.8115
Median	0.8009	0.8036	0.7617	0.7914	0.8118	**0.8143**	0.8123
Max	**0.9225**	0.9185	0.8707	0.9059	0.9155	0.9218	0.9177
Min	0.6854	0.7073	0.6469	0.7001	**0.7237**	0.7250	0.7215
Improvement over DeepChrome [29] (out of 56 cell types)		36	0	16	49	**50**	49

- Consistent improvement over DeepChrome

Guest lecture: Xiaohui Xie

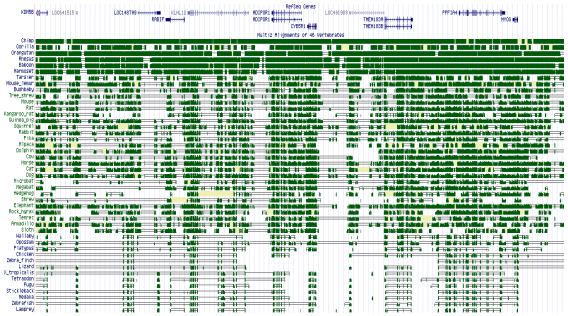
Deep Learning for Expression/Chromatin Prediction

Xiaohui Xie
Professor, UC Irvine

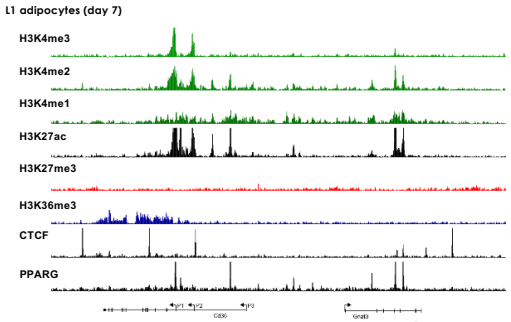
4. Predicting Reporter Expression from Chromatin Features

We can find regulatory elements … but we don't know how to read them

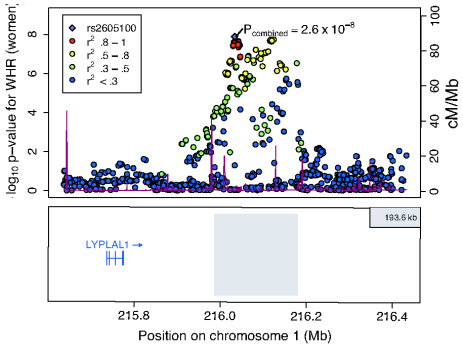
Comparative sequence analysis



Genome-wide chromatin/TF mapping



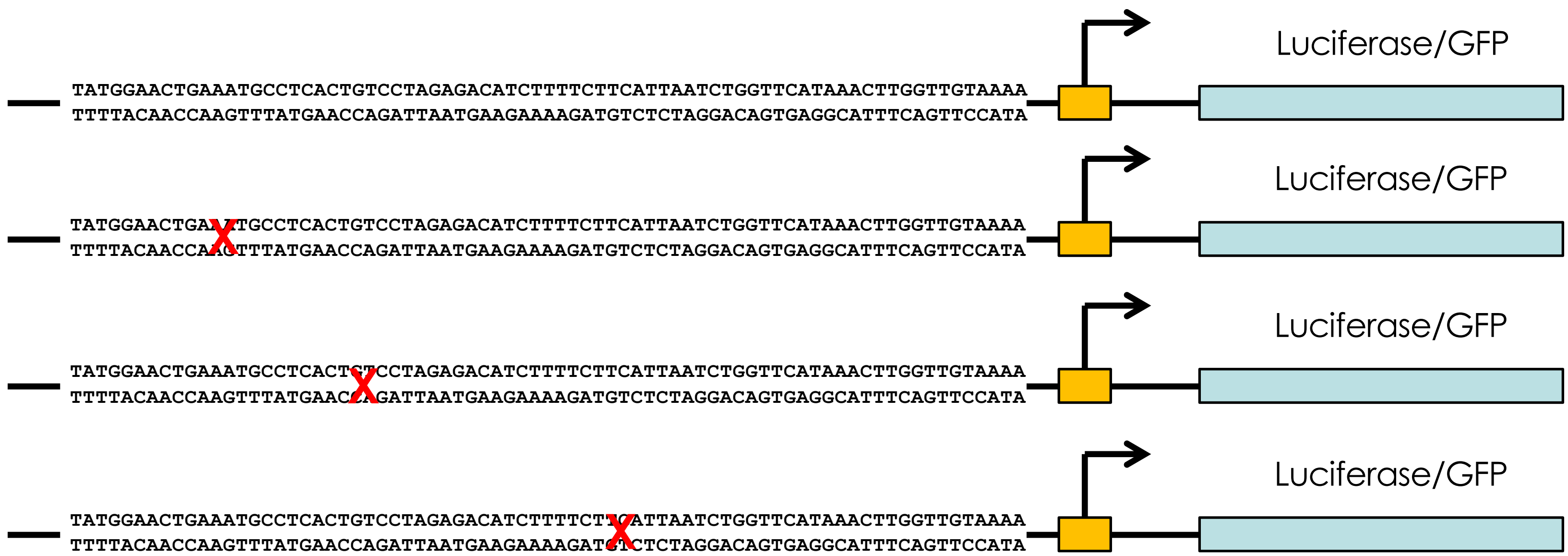
Genetics



100s of megabases of likely cis-regulatory elements

TATGGAACTGAAATGCCTCACTGTCCTAGAGACATCTTTTCTTCATTAATCTGGTTCATAAACTTGGTTGTAAAAAGCAAATTC
AAAGAATTCATCTAATAATGACAGAAAAAGAAACATTTCTGAATGAATTTGTGGAAGTGTACAATTTAATTTTTCATTTAATTT
TTATTTACAATTTTCATTTAATTTTCATATTAATTTTTCCTCTTTAATCCCATGCAAGGACACAGCAGTTAATAAATACTTAAG
AAATACTTCTCAATTGCATACCATTCTTA??????????GGATTTTATAAATCCAGTTATTAGTATCT
CAAACTTTAAATAATCAAAAGTCAATGGCAGC??????????AGGTATCAACTCACCCACCCACTCACCCC
AAAGAGACTTGTCGGTACGCCCTCTGGTGAG??????????AAATTTGAGAATGTGTAGCCCTCCAGTTCC
CAACCTTCATTCCATCTTTTTTTTTTTTTTGA??????????ACAGGTATATTAGAAAATACCAAATTCTCT
TAGAGAAAAAAAAAATTGCACAACAATCTCAA??????????AAGCAAGGTACTCCTAGGTTAGTAGTTT
TTCAAACCTATAGATGGCATTTGGAAAGAAGTACACGCATAGGCTTCAAAAAACCCTGAACACCAGAAACCTAATTAGCACTTT
CATTTTAGGGGCACGGGCTTTCCCTTATCCTTTTCTACAGCCAACCTTAATGGGGTCCAATTAAGGCAACAATGAAAGGTTACA
AAAGTCTACAAAAAGGTCCCATTACAGCATTTAAACAGGGTTCATTTTTTTTAATCAAATAGTTTACACTGTTTTTTTGTTGTT

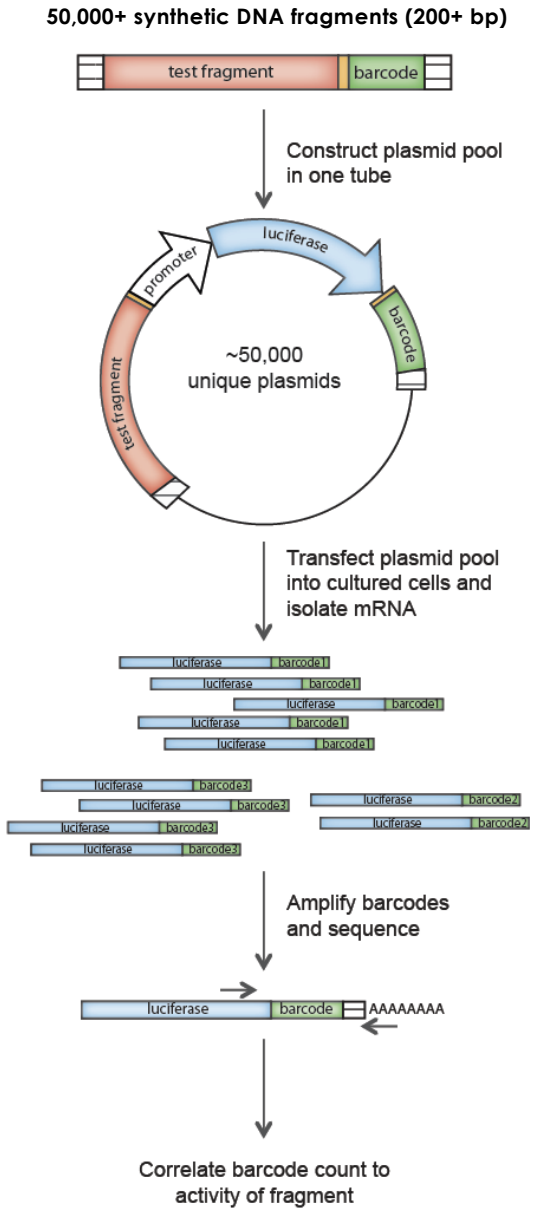
Traditional regulatory element “bashing”



Bottlenecks:

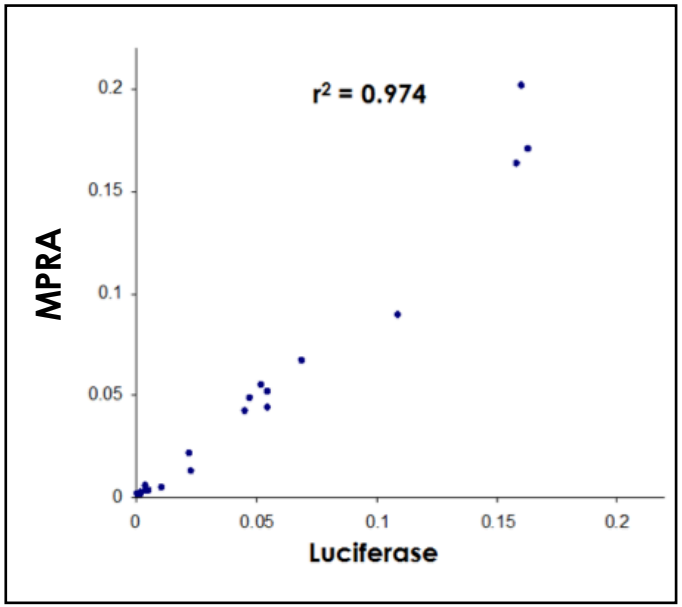
1. **Generating/cloning individual variants is tedious**
2. **Enzymatic/fluorescent reporters limit multiplexing**

Massively Parallel Reporter Assays (MPRA)



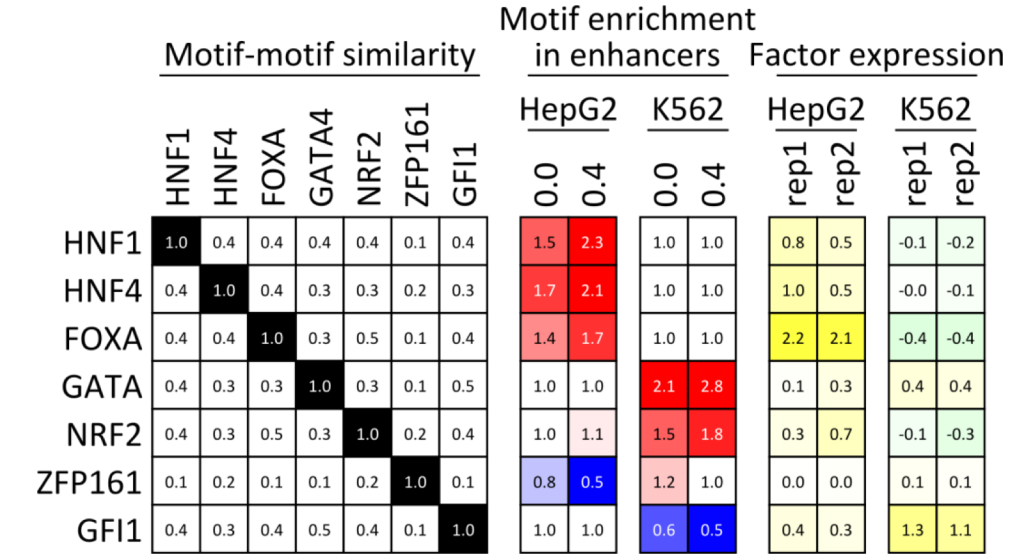
- Flexible assay format:
  Promoters, enhancers, silencers, Insulators, RNA stability elements, ++

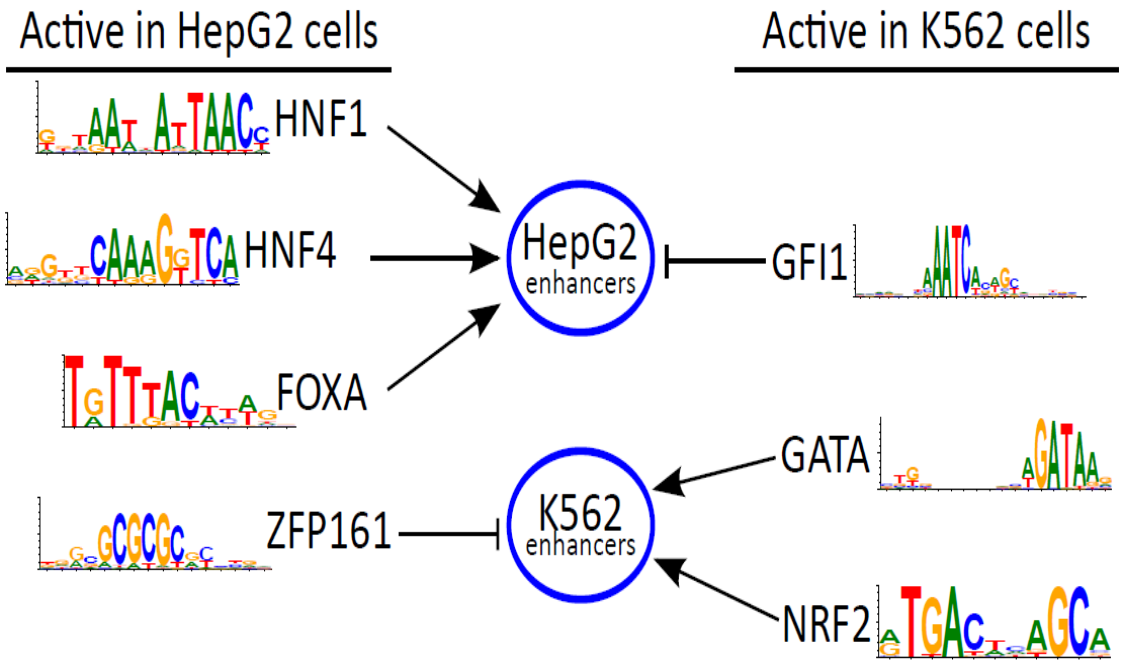
- Data is directly comparable to traditional reporter assays:

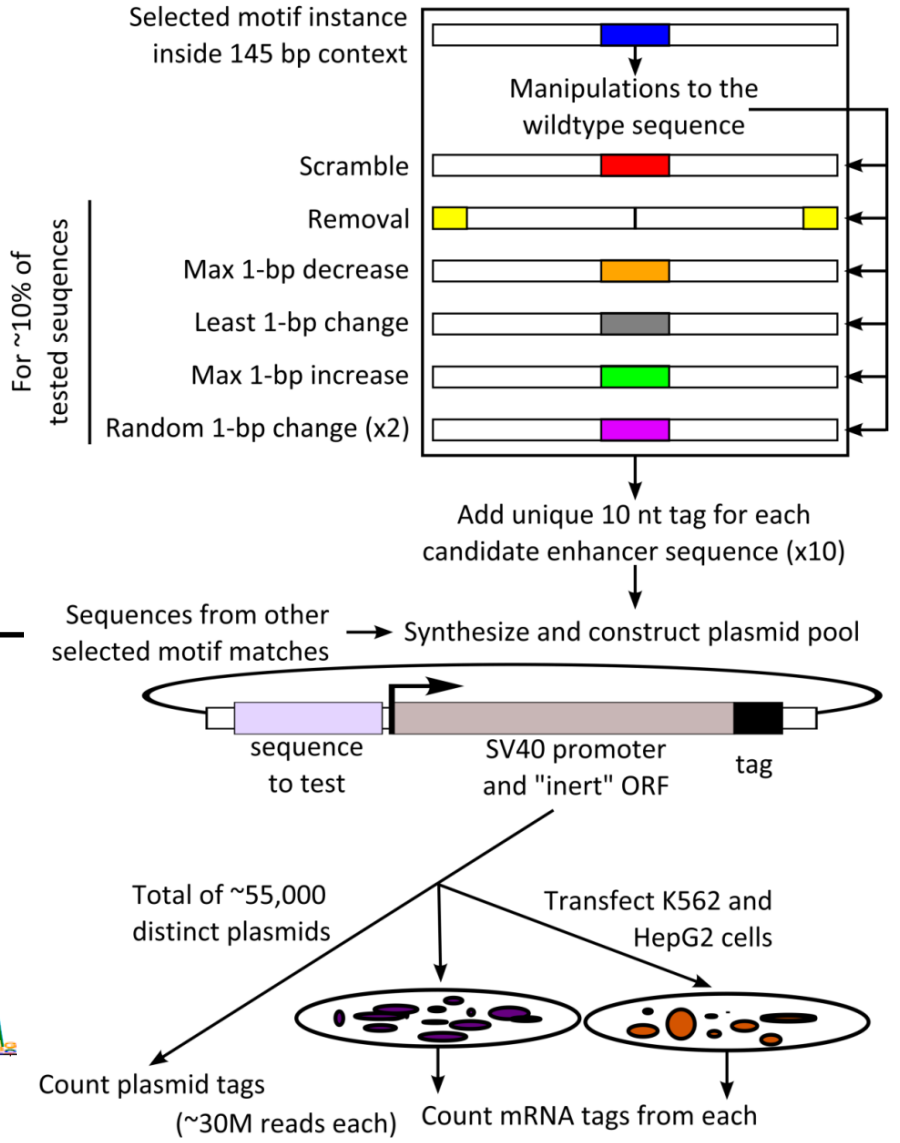


- Throughput increased by 3 orders of magnitude

Systematic motif disruption for 5 activators and 2 repressors in 2 human cell lines

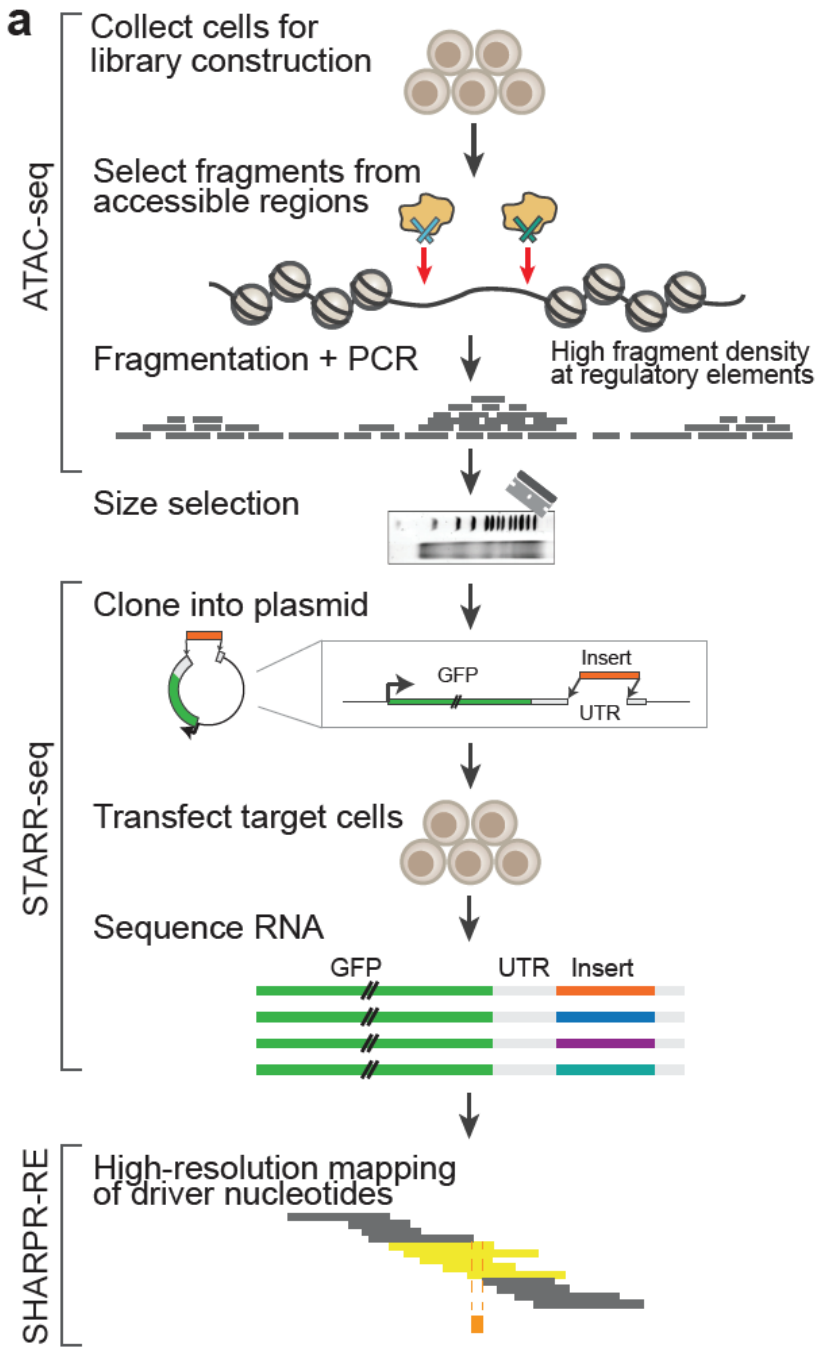






54000+ measurements (x2 cells, 2x repl)

HiDRA: High-Definition Reporter Assay



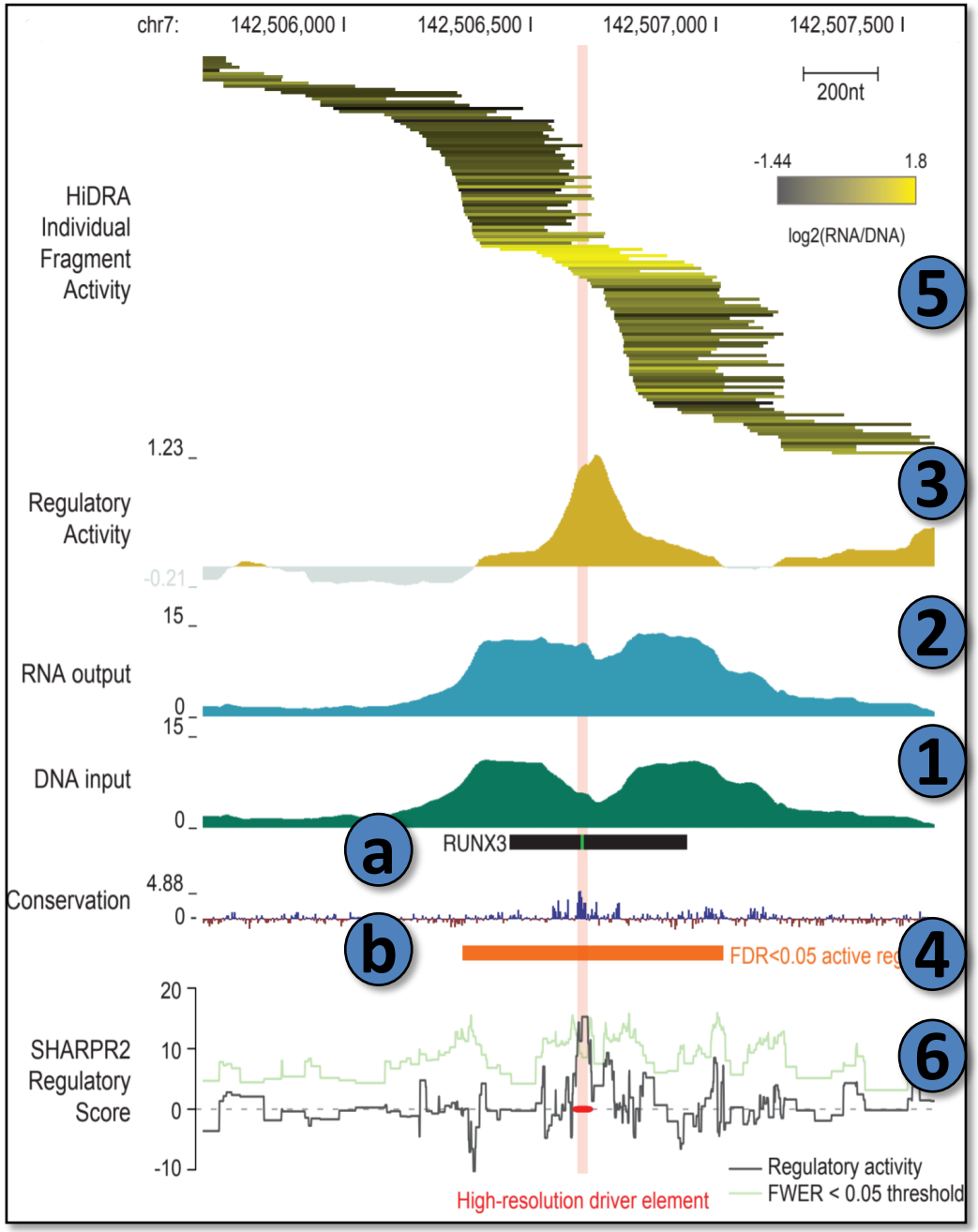
Key features:

- No synthesis → 7M fragments tested in 1 expt
- No synthesis, size-selection → Test long fragments
- Select accessible DNA regions → High sensitivity
- 3'UTR integration → self-transcribing → No barcode
- Densely-overlapping fragments → Region tiling
- Unbiased, random starts/ends → Sharpr dissection

Putting it all together:

- Testing 7M fragments in 1 experiment
- High sensitivity, high specificity, quantitative assay
- High-res inference pinpoints driver nucleotides

HiDRA data overview: DNA, RNA, Regulatory Activity, Sharpr2



1. **Sequence DNA library**
   - Effectively a DNase/ATAC-Seq expt
2. **Sequence RNA output**
   - How much expression does this drive
3. **Take RNA/DNA ratio**
   - Measures regulatory activity
4. **Pinpoint boundaries of active region**
   - FDR<0.05
5. **Study activity of individual fragments**
   - Random start/end cuts (Transposase)
6. **Infer high-resolution driver nucleotides**
   - Sharpr2 deconvolution algorithm
   - Exploit diffs btw overlapping fragments

a. Compare with evolutionary conservation

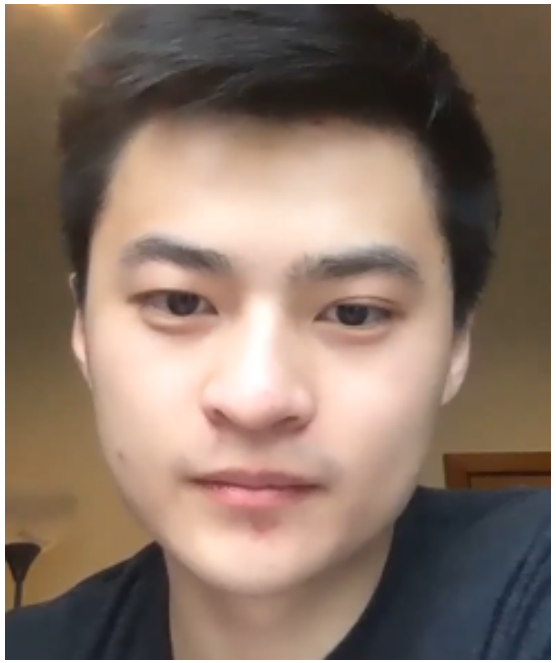
- Capture evolutionarily-conserved nts

b. Compare with bound regulatory motifs

- Driver nucleotides are highly accurate

Guest lecture: Flynn Chen, Mark Gerstein Lab

Deep Learning for Reporter Expression Prediction



Flynn Chen
Yale University
Statistics and Data Science

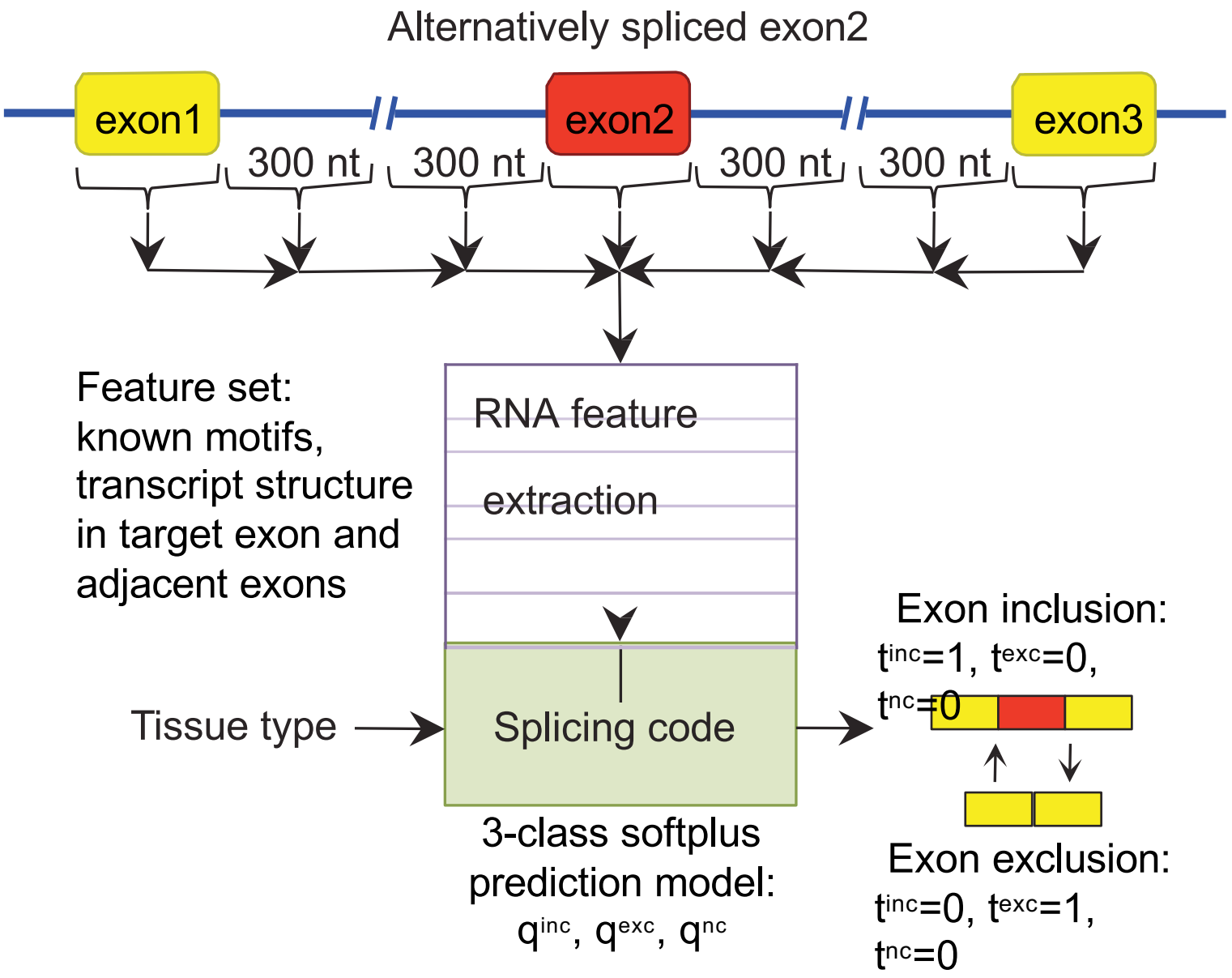
Prof. Mark Gerstein
Yale University
ENCODE, Data Science

Today: Predicting gene expression and splicing

0. Review: Expression, unsupervised learning, clustering
1. Up-sampling: predict 20,000 genes from 1000 genes
2. Compressive sensing: Composite measurements
3. DeepChrome+LSTMs: predict expression from chromatin
4. Predicting splicing from sequence: 1000s of features
5. Unsupervised deep learning: Restricted Boltzmann mach.
6. Multi-modal learning: Expr+DNA+miRNA RBMs in Cancer

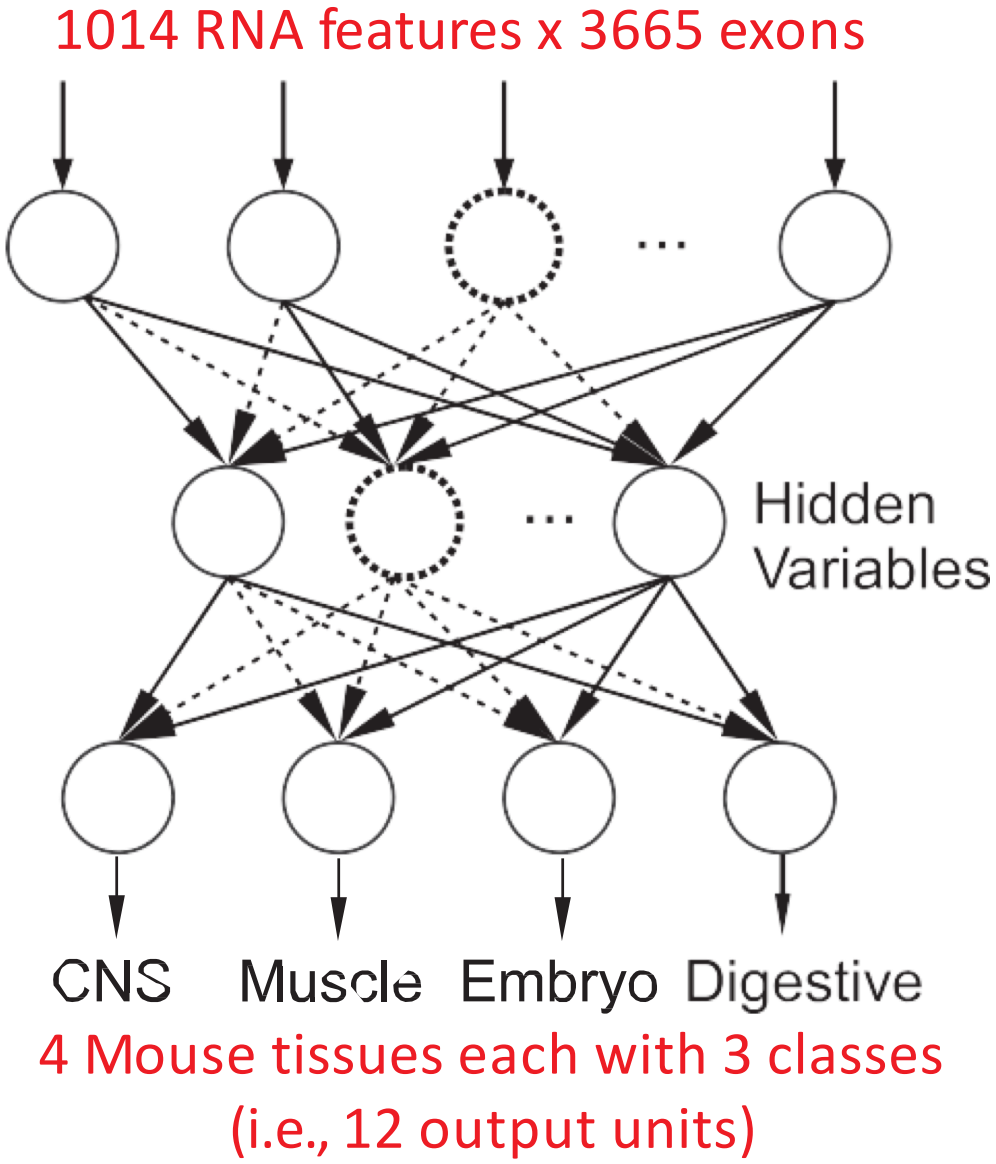
4. Predicting splicing from sequence

Deciphering tissue-specific splicing code



[Barash et al., 2010]

Bayesian neural network splicing code

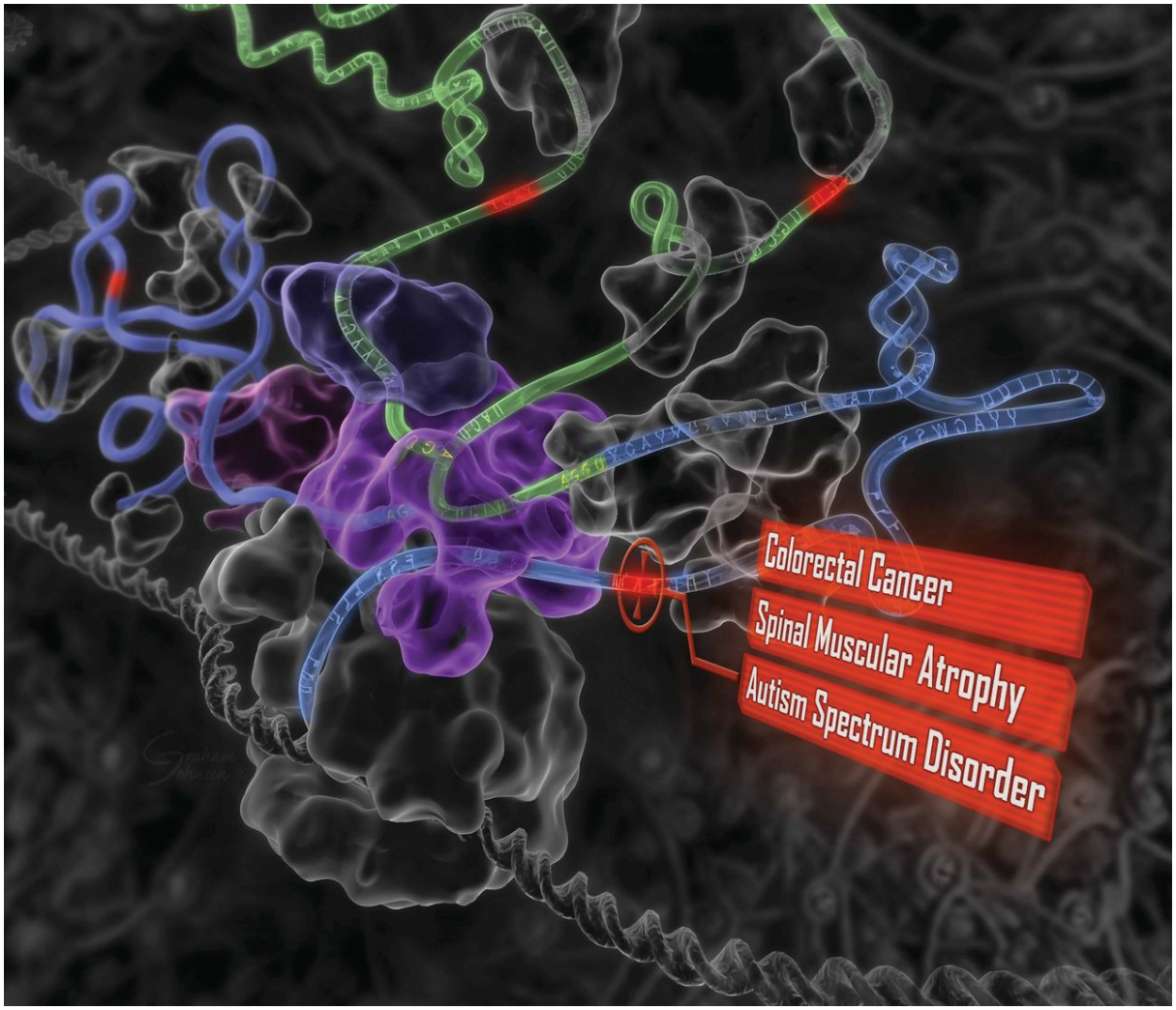


Bayesian neural network:

- # hidden units follows Poisson(λ)
- Network weights follows spike-and-slab prior Bern($1 - a$)
- Likelihood is cross-entropy
- Network weights are sampled from the posterior

[Xiong et al., 2011]

Predicts diseasing causing mutations from splicing code

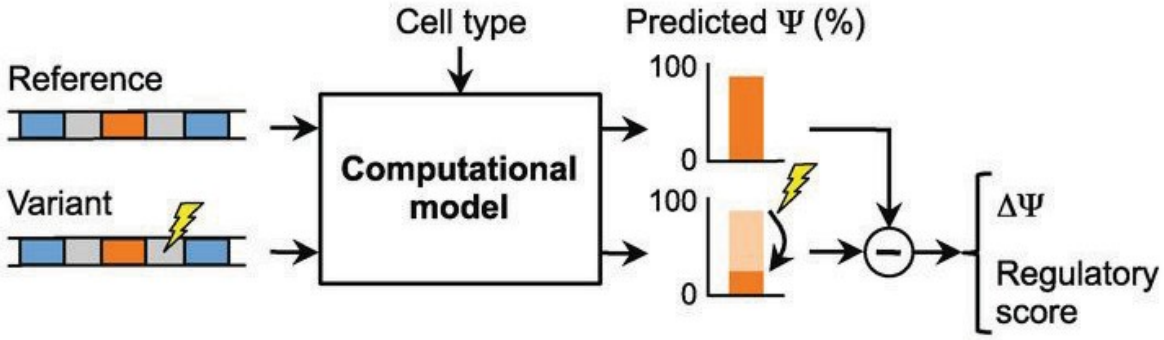


[Xiong et al., 2011]

Predicts diseasing causing mutations from splicing code

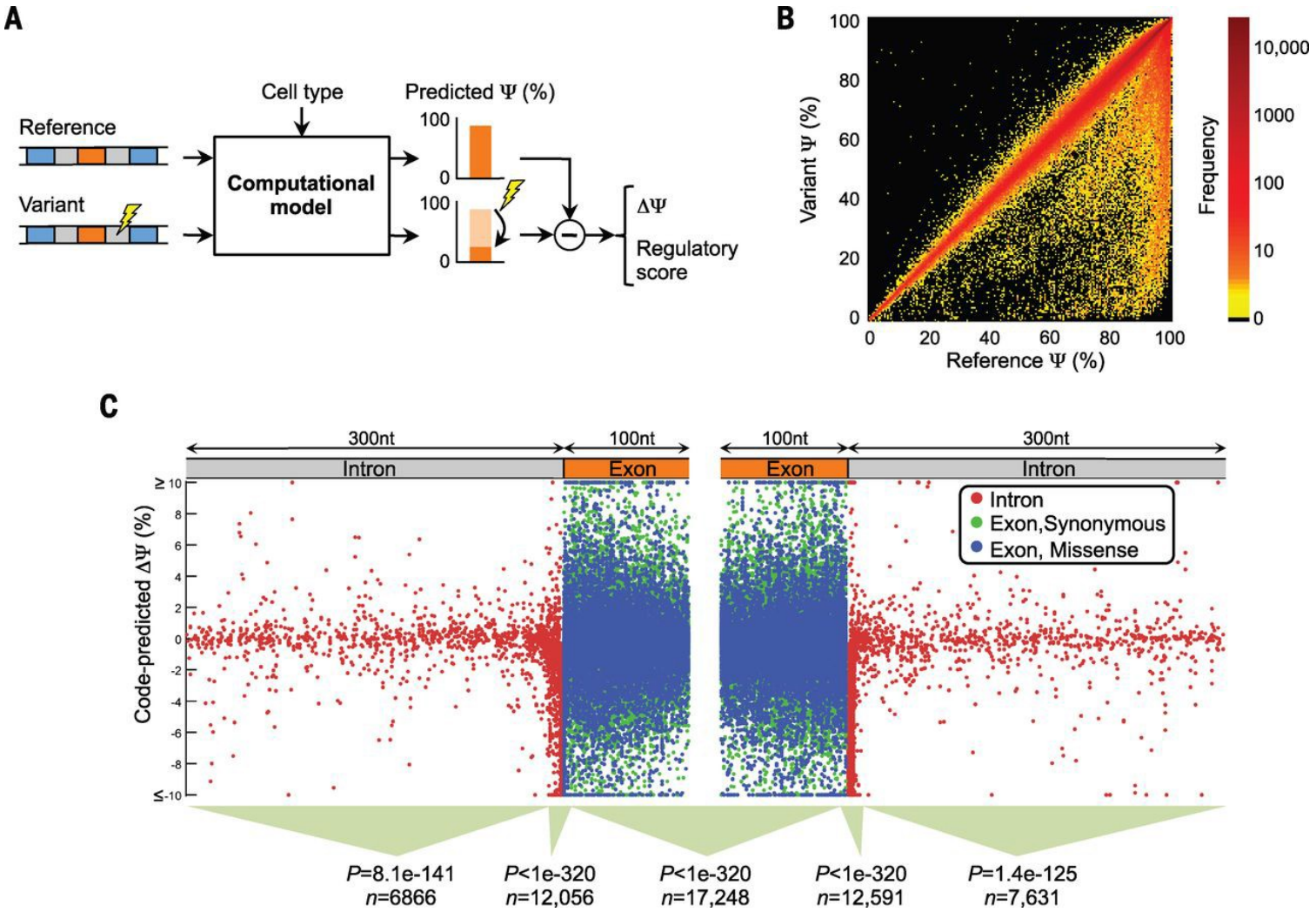
Scoring splicing changes due to SNP $\Delta\psi$:

- Train splice code model on 10,689 exons to predict the 3 splicing classes over 16 human tissues using 1393 sequence features (motifs & RNA structures)
- Score both the reference ψ_{ref} and alternative ψ_{alt} sequences harboring one of the 658,420 common variants
- Calculate $\Delta\psi t = \psi^t_{ref} - \psi^r_{alt}$ over each tissue t
- Obtain largest absolute or aggregate $\Delta\psi_t$ to score effects of SNPs

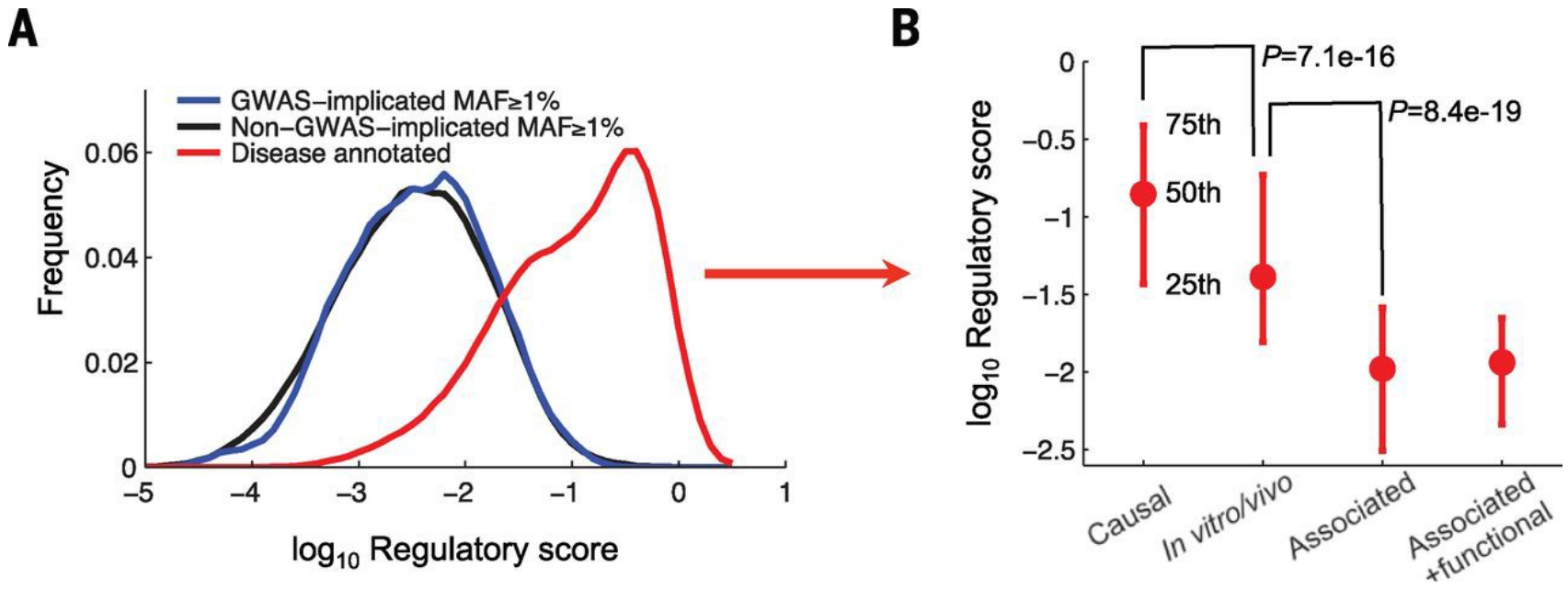


[Xiong et al., 2011]

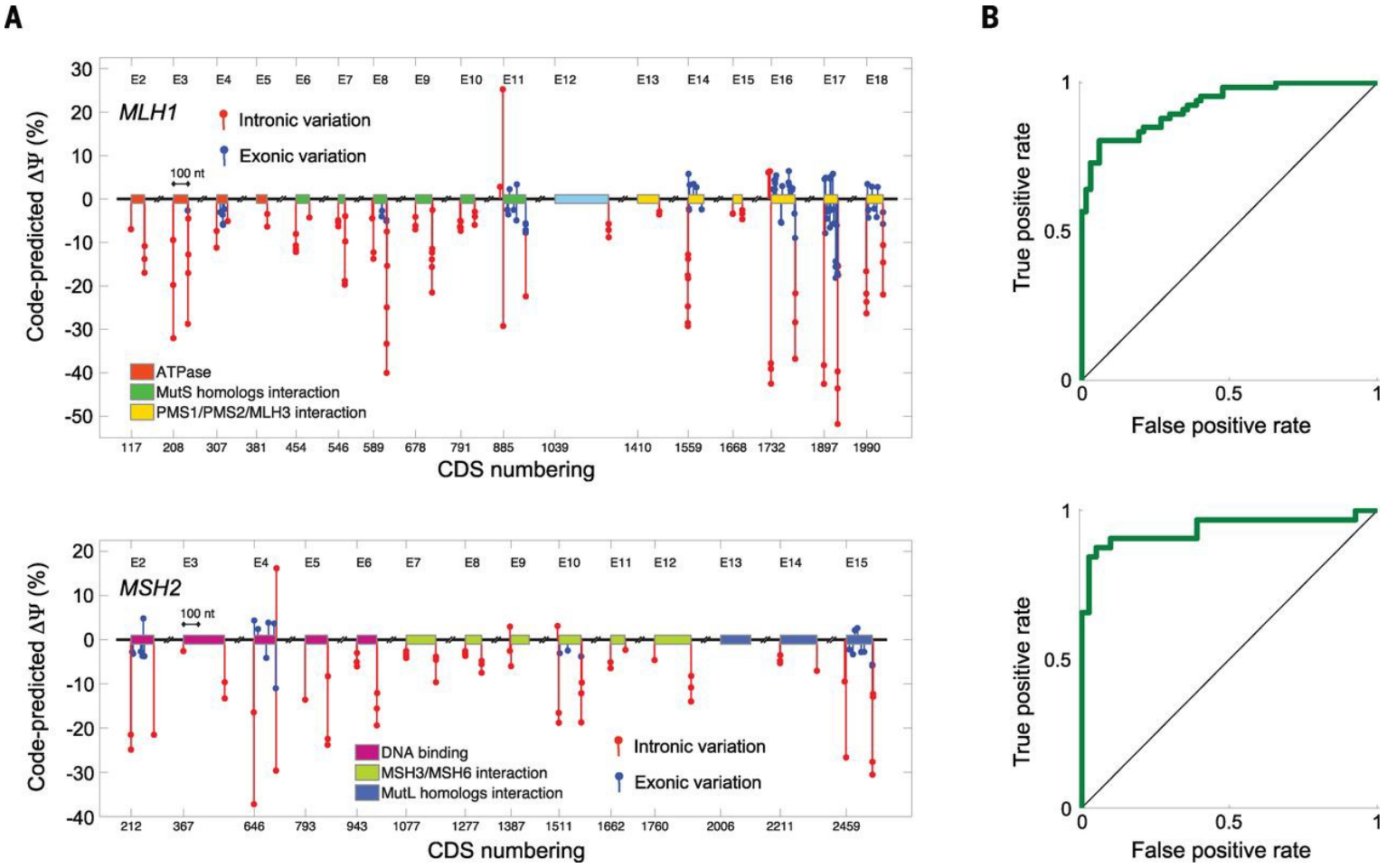
Predicted scores are indicative of disease causing mutations



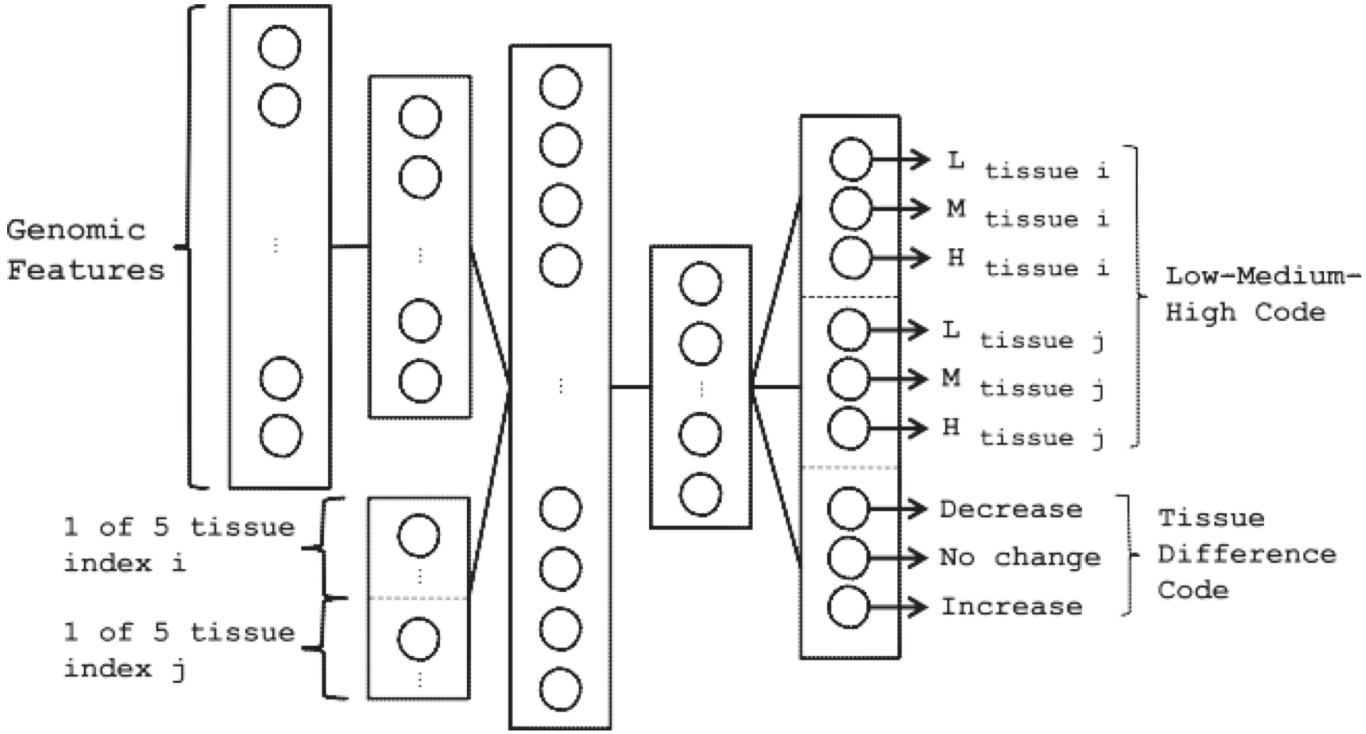
Predicted scores are indicative of disease causing mutations



Predicted mutations in MLH1,2 in nonpolyposis colorectal cancer patients are validated via RT-PCR



Splice code goes deep



Architecture of the new network to predict alternative splicing between two tissues. It contains three hidden layers, with hidden variables that jointly represent genomic features and tissue types.

[Leung et al., 2014]

Limitations of the splice code model

- Require threshold to define discrete splicing targets
- Not taking into account exon expression level in specific tissue types
- Fully connected neural network potentially impose a large number of parameters: (1393 inputs + 13 outputs) $\times$ 10 hidden units = 13000 parameters
- Although authors showed that neural network performs the best a softplus/Dirichlet multivariate linear regression may achieve similar performance
- The features are pre-defined and thus may be completely reflect the underlying splicing mechanism
- Interpretation of the importance of features is not trivial

Guest lecture: Kyle Farh, Illumina
Deep Learning for Splicing Prediction

Dr. Kyle Kai-How Farh
Illumina
Harvard/MIT/Broad alum

Today: Predicting gene expression and splicing

0. Review: Expression, unsupervised learning, clustering
1. Up-sampling: predict 20,000 genes from 1000 genes
2. Compressive sensing: Composite measurements
3. DeepChrome+LSTMs: predict expression from chromatin
4. Predicting splicing from sequence: 1000s of features
5. Unsupervised deep learning: Restricted Boltzmann mach.
6. Multi-modal learning: Expr+DNA+miRNA RBMs in Cancer

4. Predicting splicing from sequence

Today: Predicting gene expression and splicing

0. Review: Expression, unsupervised learning, clustering
1. Up-sampling: predict 20,000 genes from 1000 genes
2. Compressive sensing: Composite measurements
3. DeepChrome+LSTMs: predict expression from chromatin
4. Predicting splicing from sequence: 1000s of features
5. Unsupervised deep learning: Restricted Boltzmann mach.
6. Multi-modal learning: Expr+DNA+miRNA RBMs in Cancer