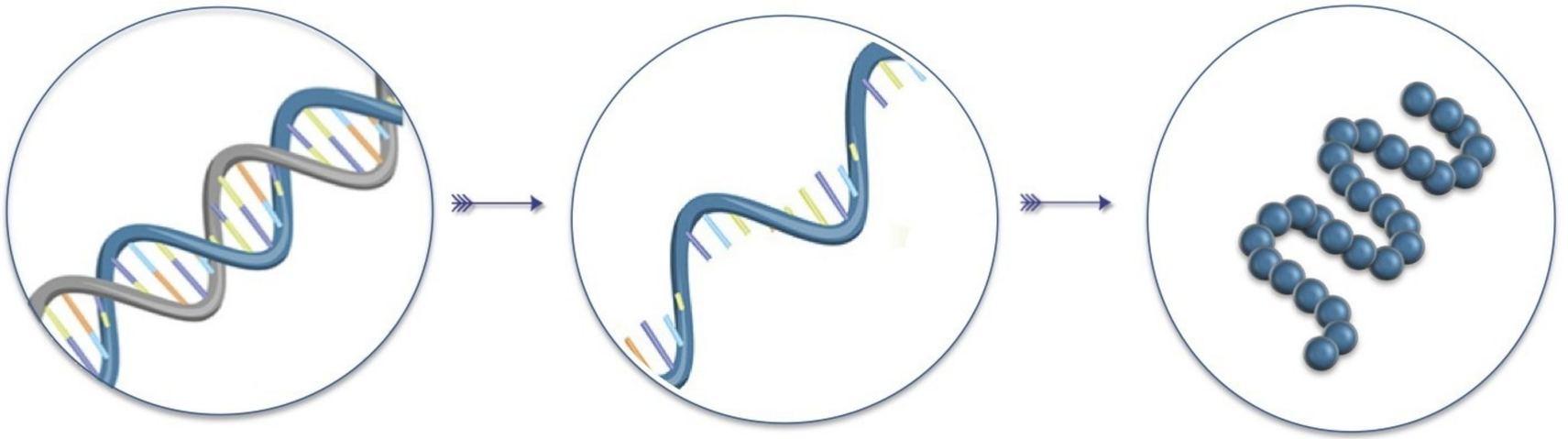
How does genotype determine phenotype?

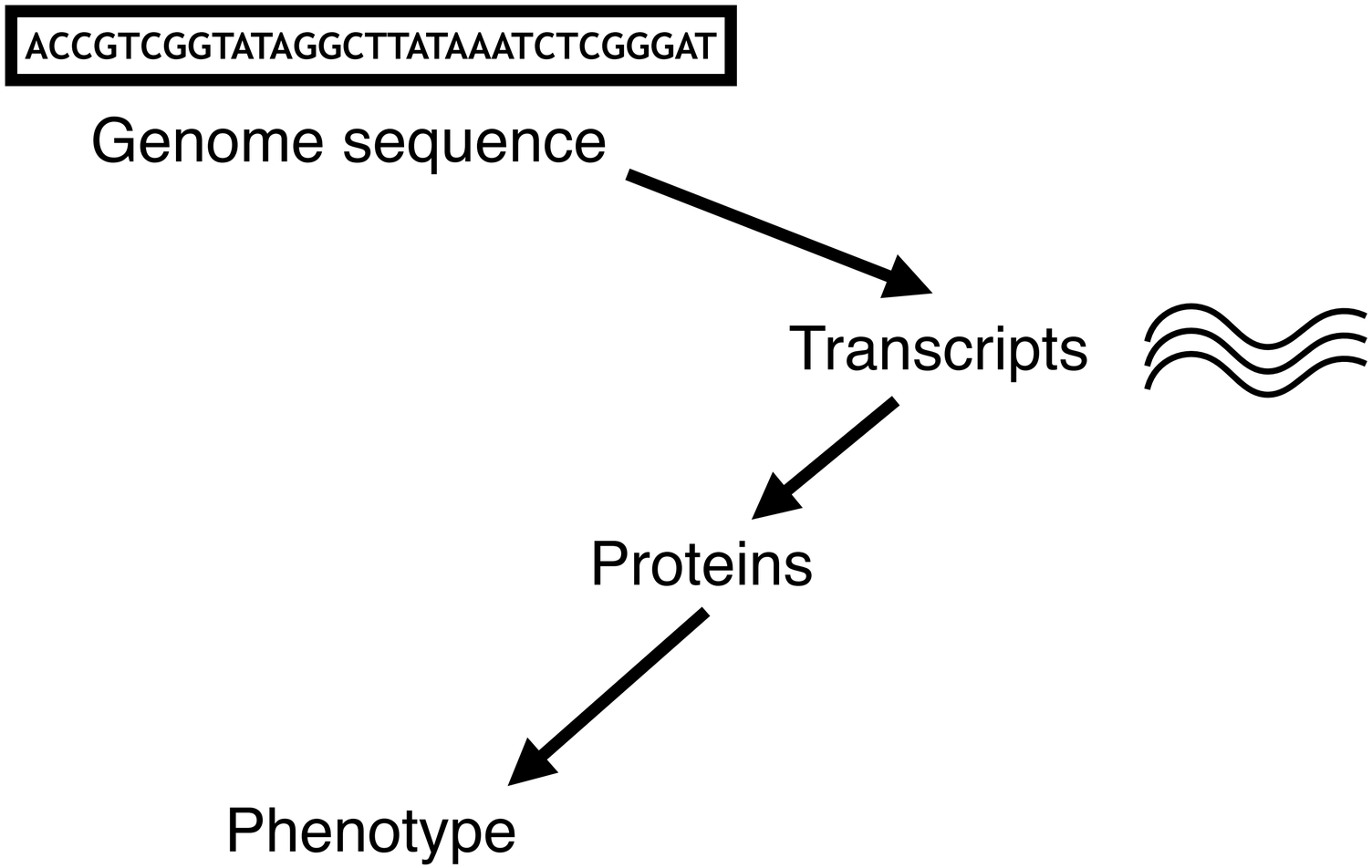
How does genotype affect phenotype?

ACCGTCGGTATAGGCTTATAAATC**A**TCGGGATCCTATTAATGAGGAAAA

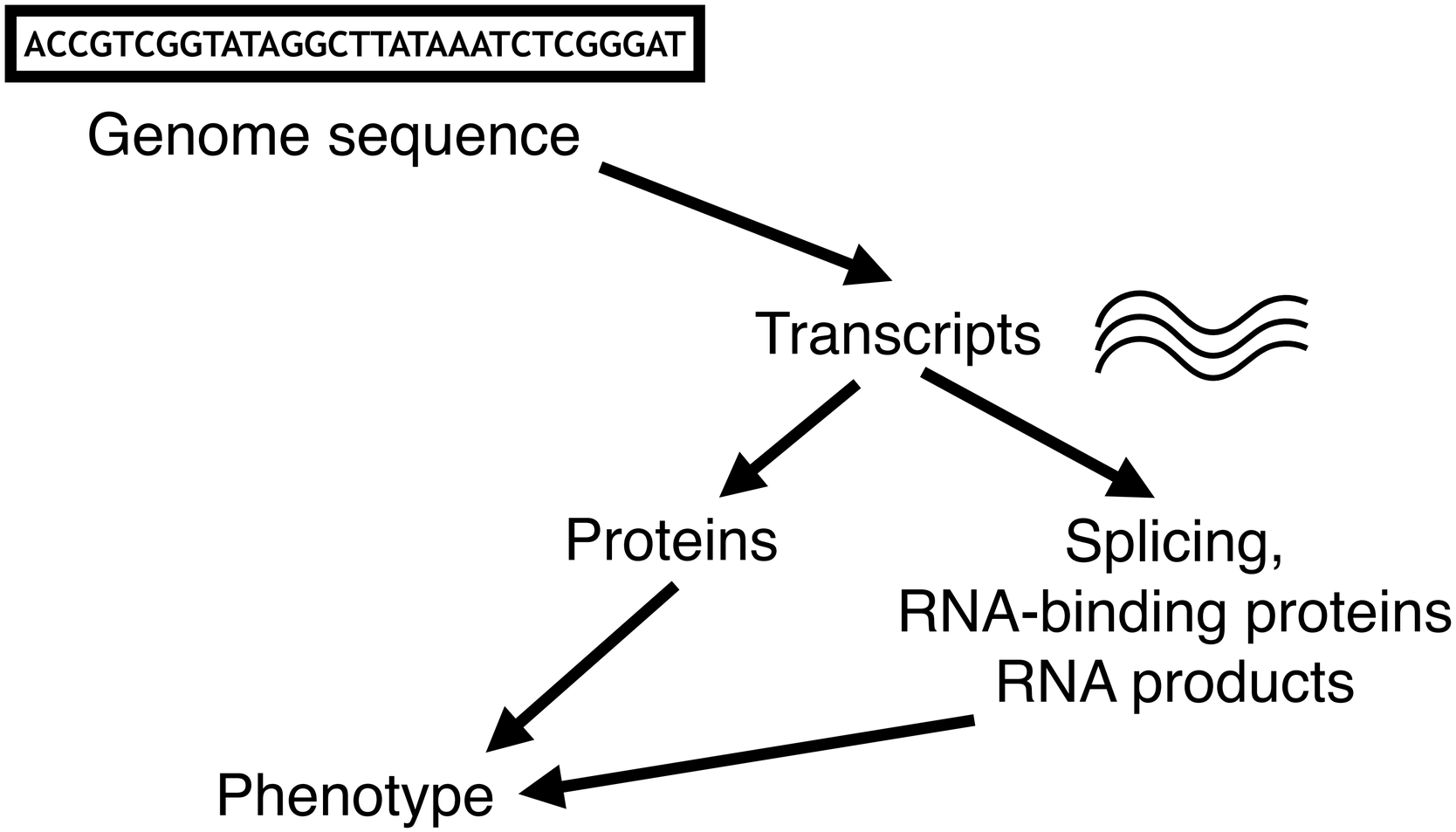
Genetic traits
Disease
Evolutionary fitness


DNA
RNA
Protein

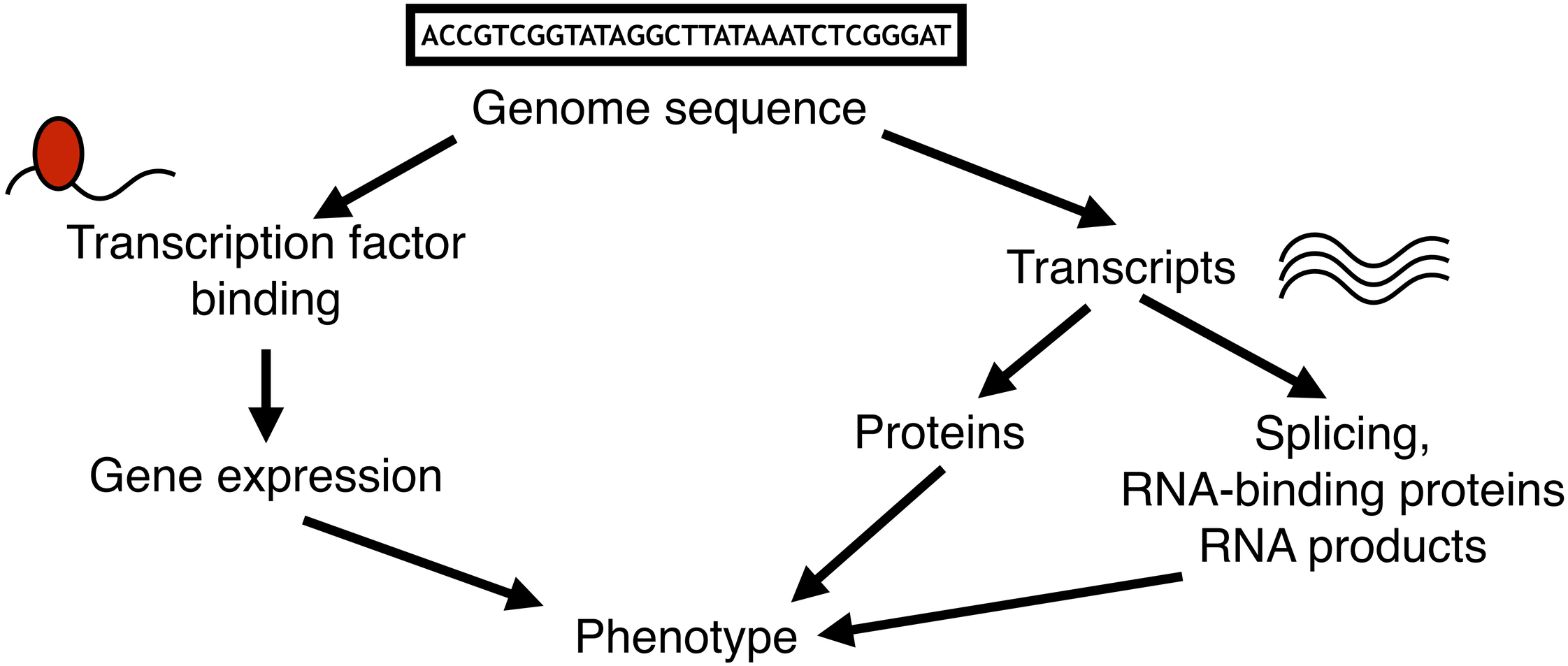
Genotype usually determines phenotype either through (1) protein-coding sequence or (2) gene regulation



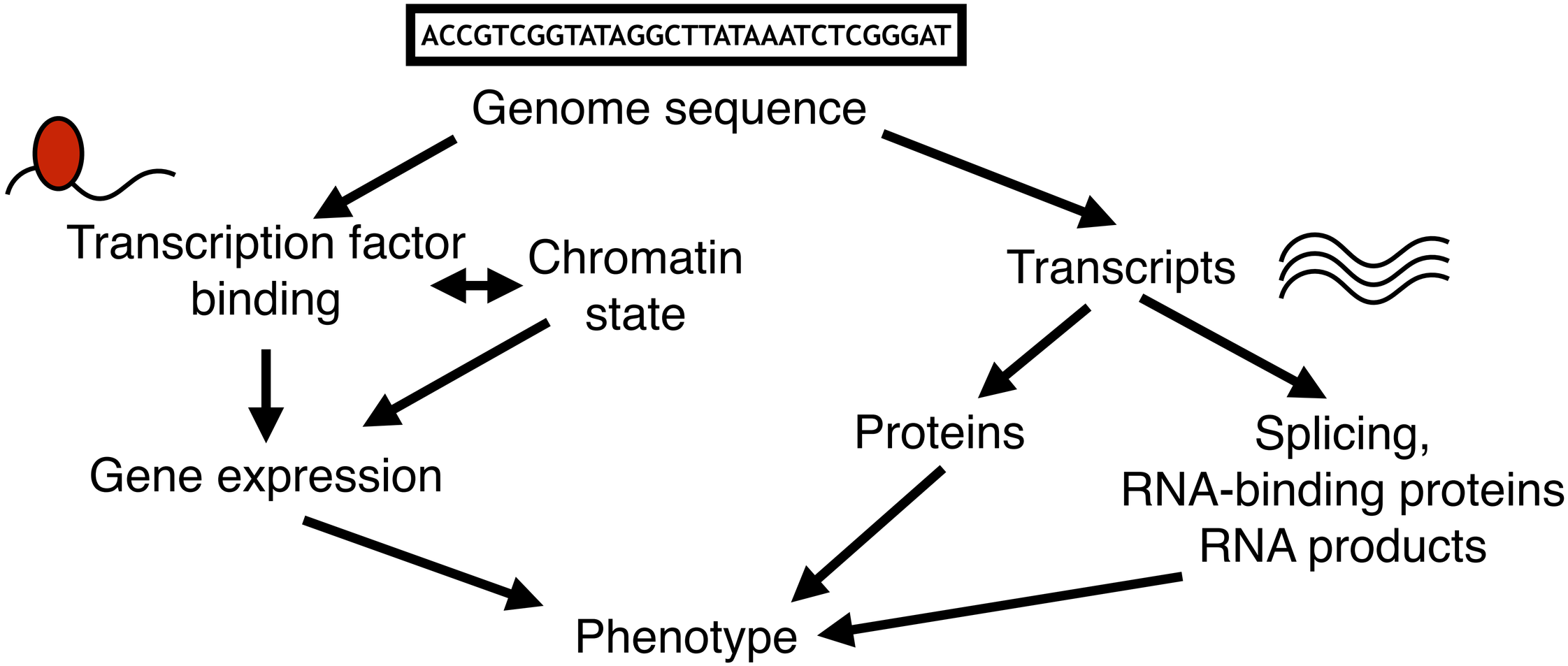
Genotype usually determines phenotype either through (1) protein-coding sequence or (2) gene regulation



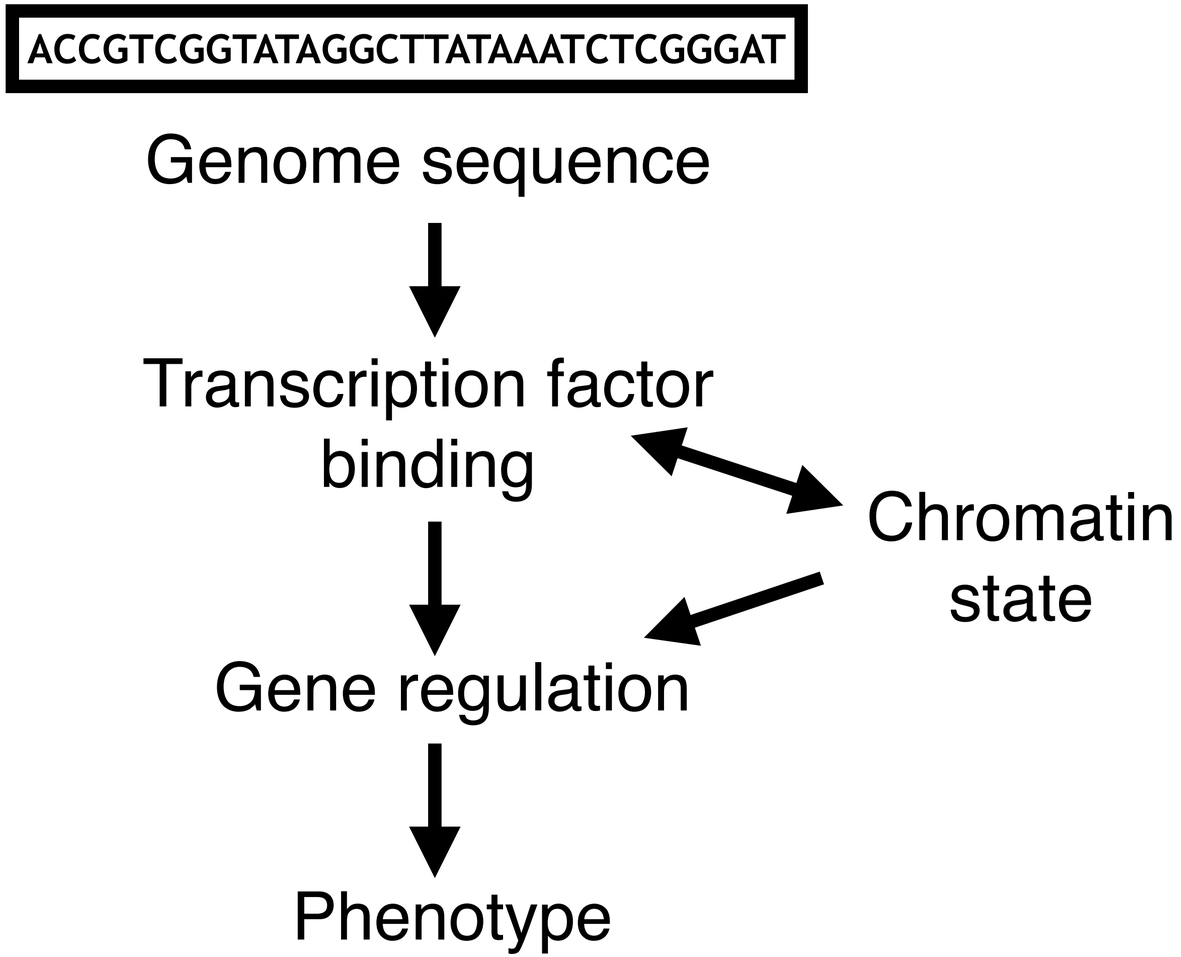
Genotype usually determines phenotype either through (1) protein-coding sequence or (2) gene regulation



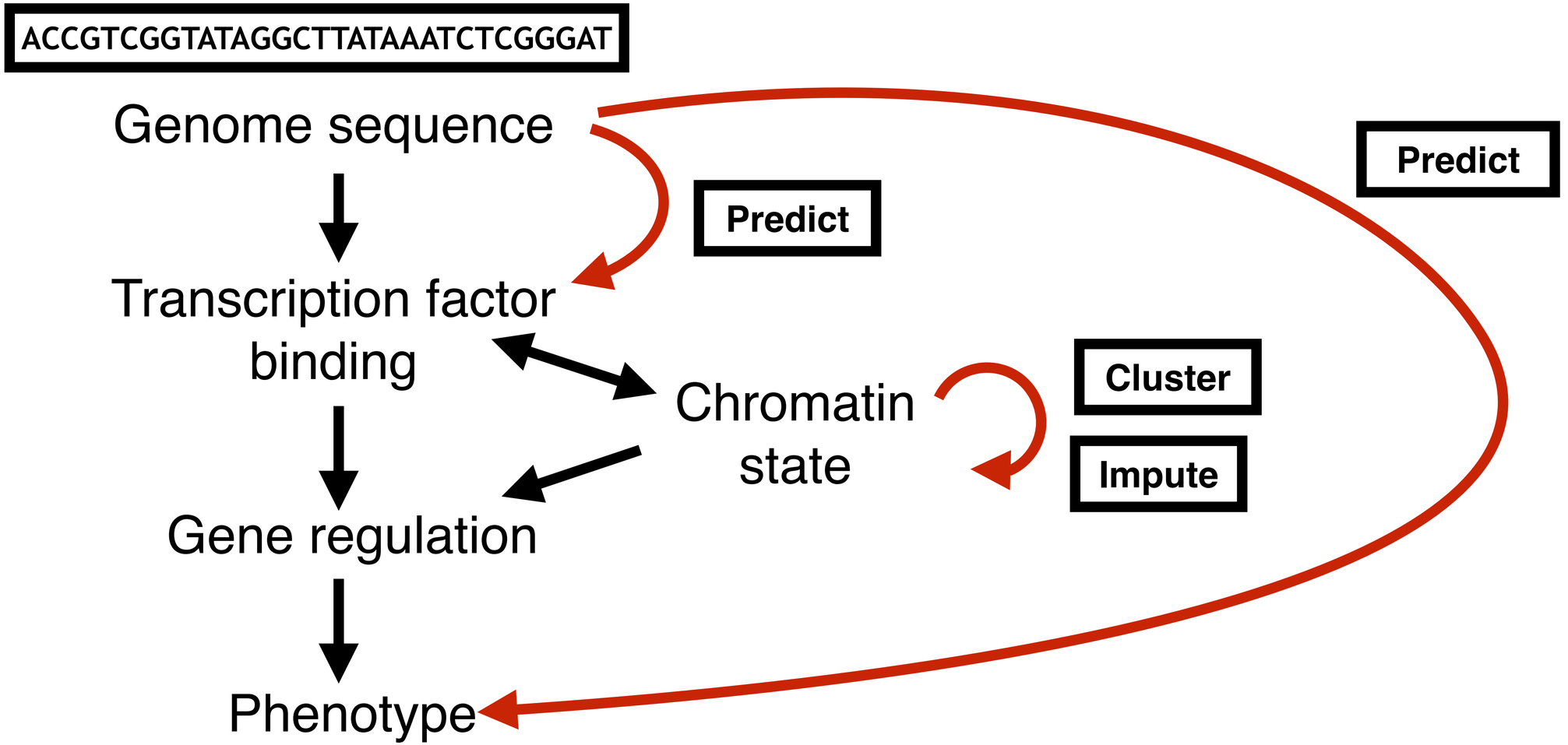
Genotype usually determines phenotype either through (1) protein-coding sequence or (2) gene regulation



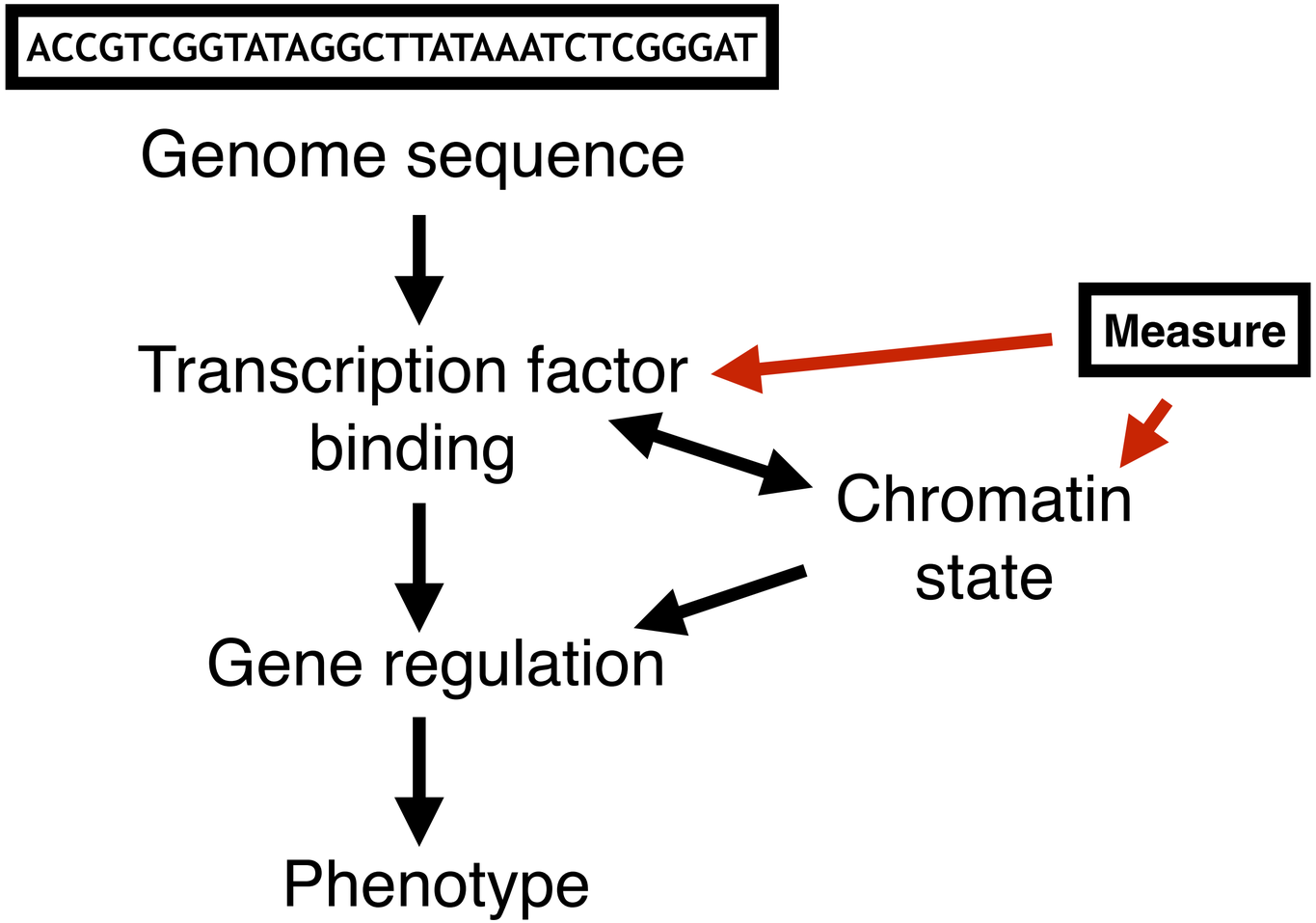
Machine learning methods for the genotype-phenotype relationship, gene regulation and epigenomics



Machine learning methods for the genotype-phenotype relationship, gene regulation and epigenomics

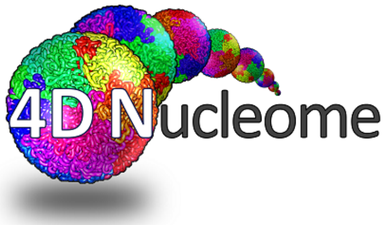


Machine learning methods for the genotype-phenotype relationship, gene regulation and epigenomics



What are the functional elements in the human genome?

CEEHRC NETWORK
Canadian Epigenetics, Environment and Health Research Consortium Network



Primary tool: performing *sequencing-based genomics assays*

ChIP-seq measures where a given protein binds along the genome

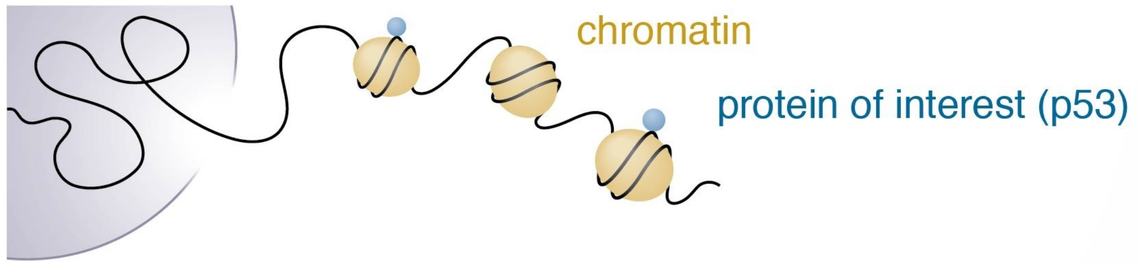


Figure derived from Brookhaven National Laboratory Newsroom 2011-11-30

ChIP-seq measures where a given protein binds along the genome

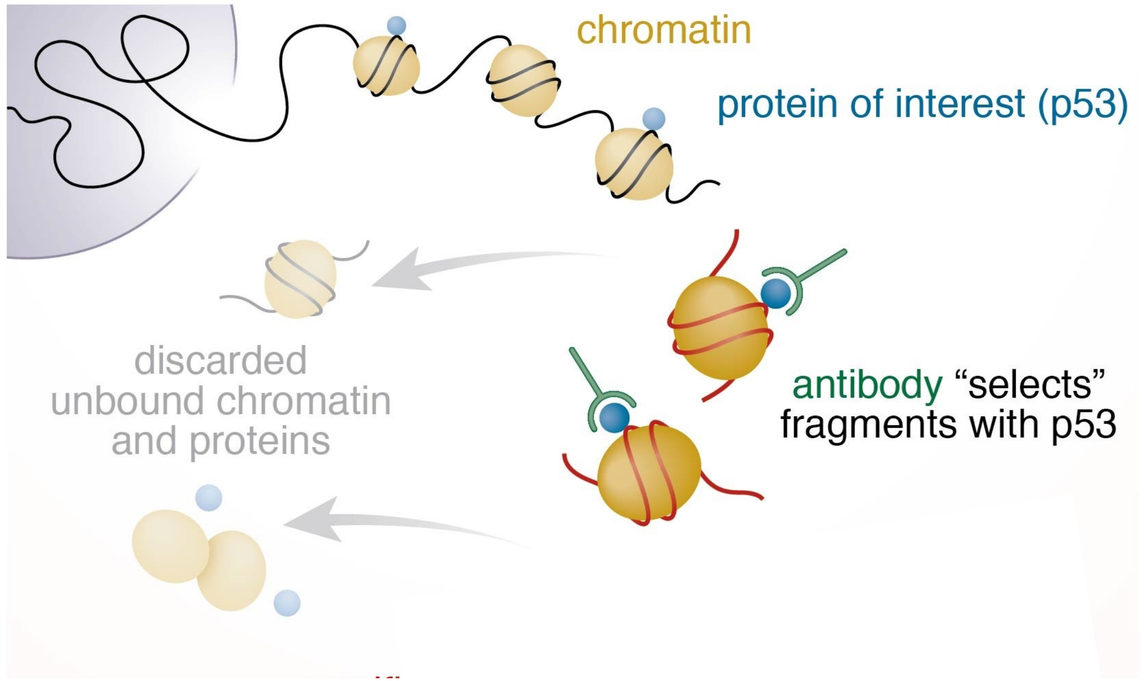


Figure derived from Brookhaven National Laboratory Newsroom 2011-11-30

ChIP-seq measures where a given protein binds along the genome

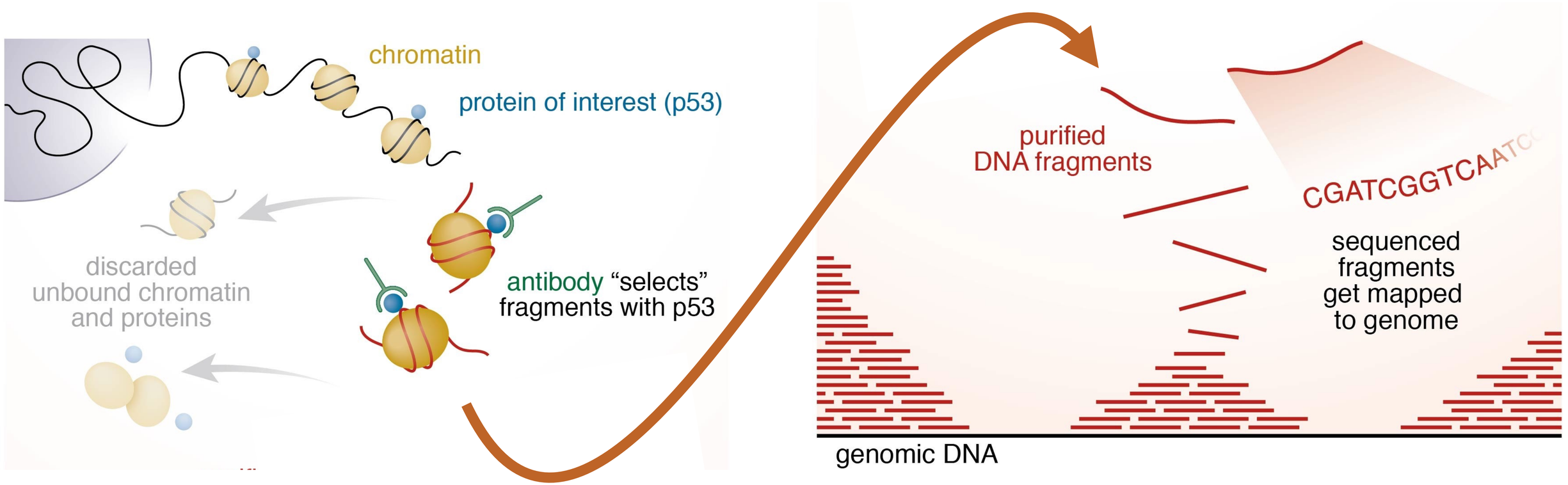


Figure derived from Brookhaven National Laboratory Newsroom 2011-11-30

ChIP-seq measures where a given protein binds along the genome

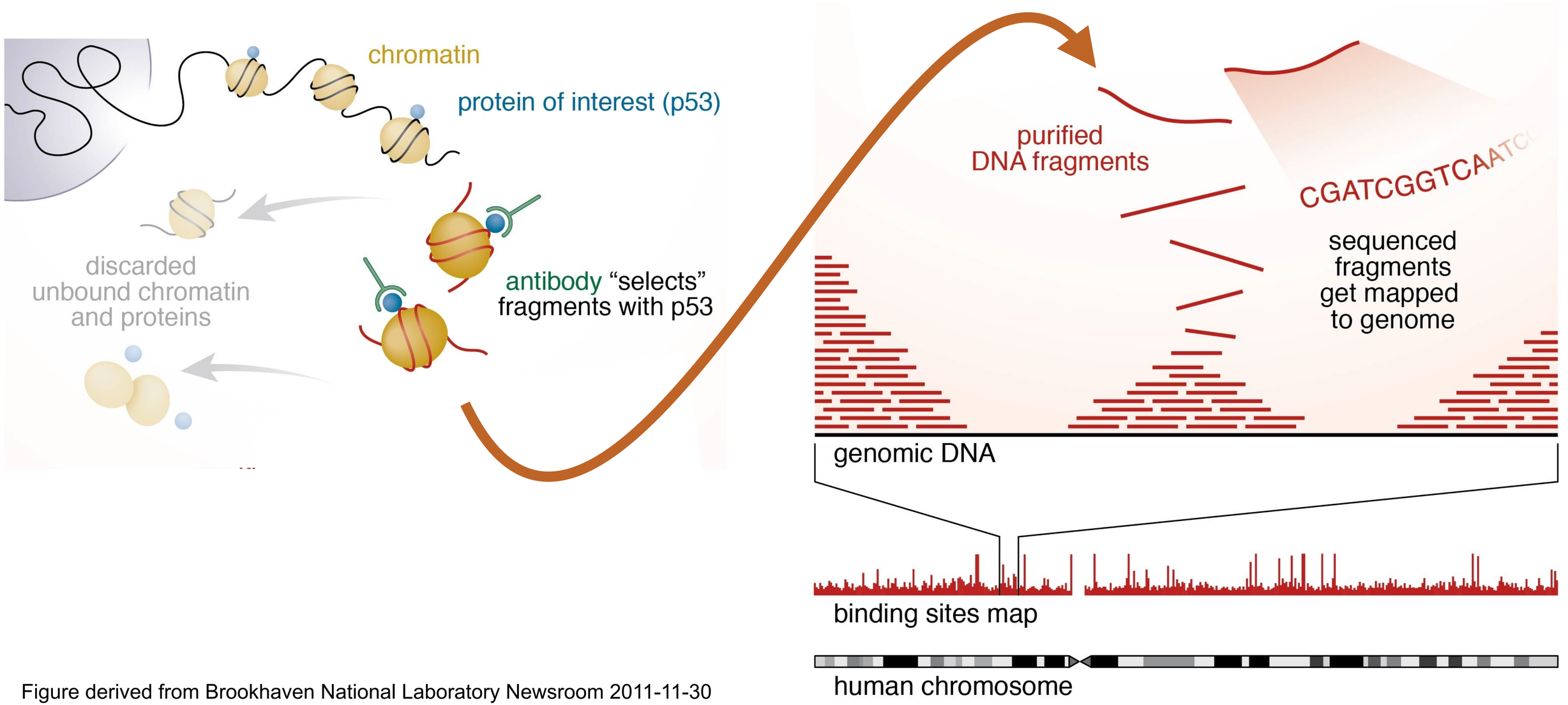
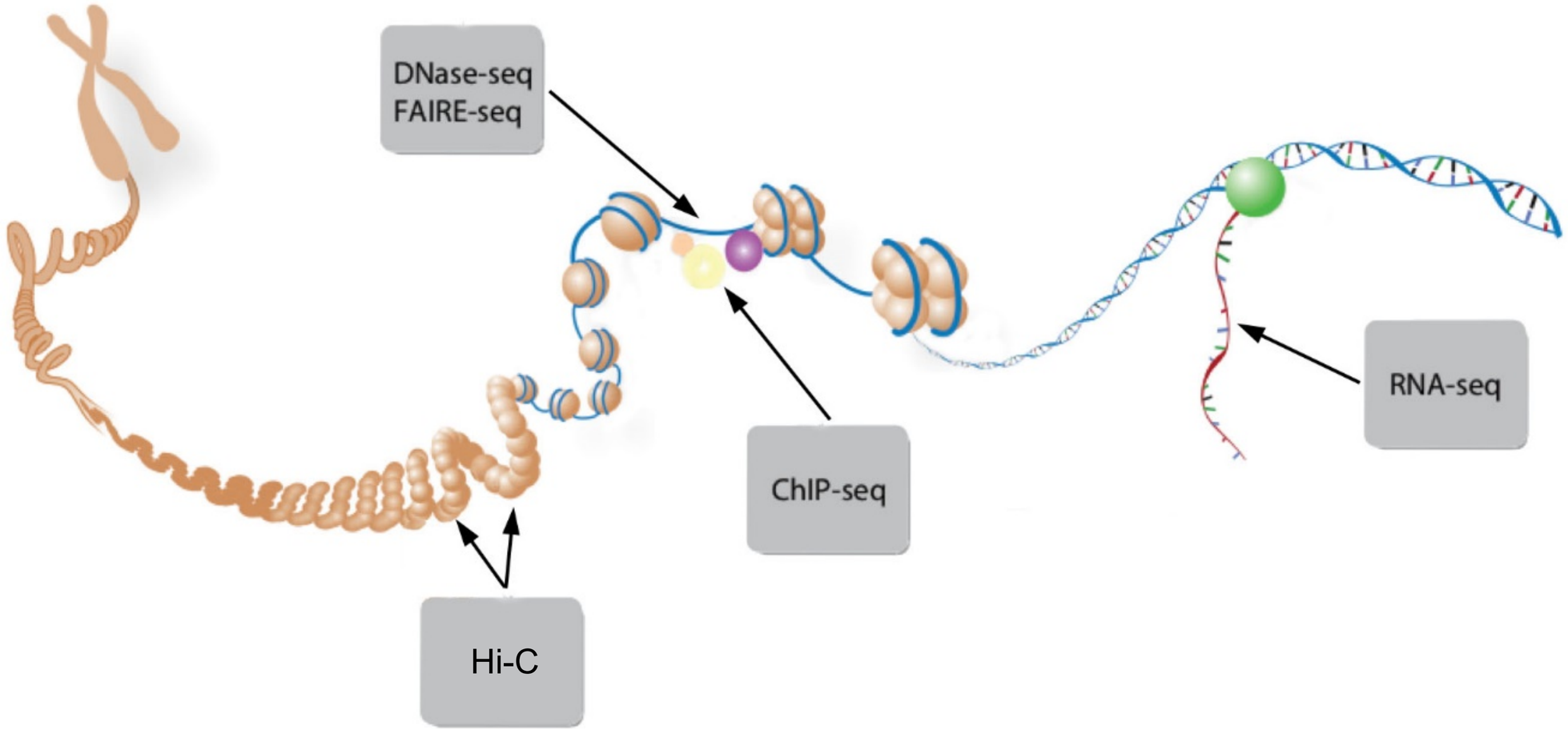


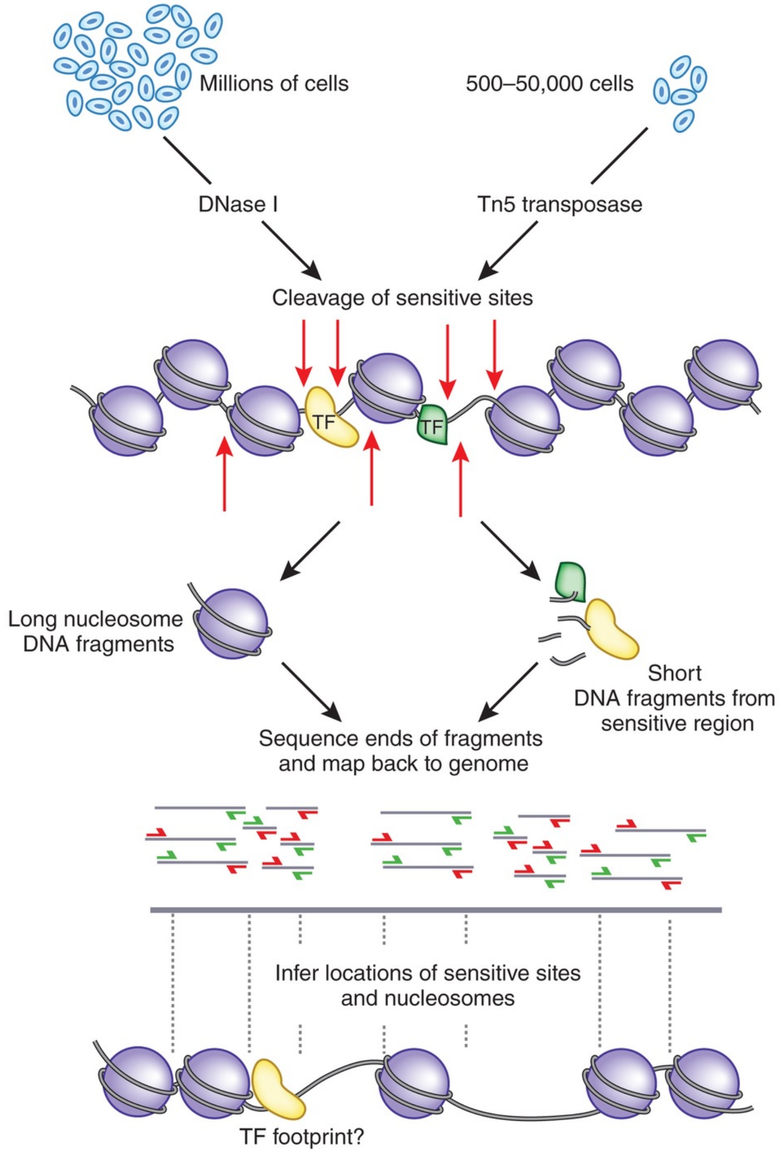
Figure derived from Brookhaven National Laboratory Newsroom 2011-11-30

Sequencing-based genomics assays measure many types of genomic activity

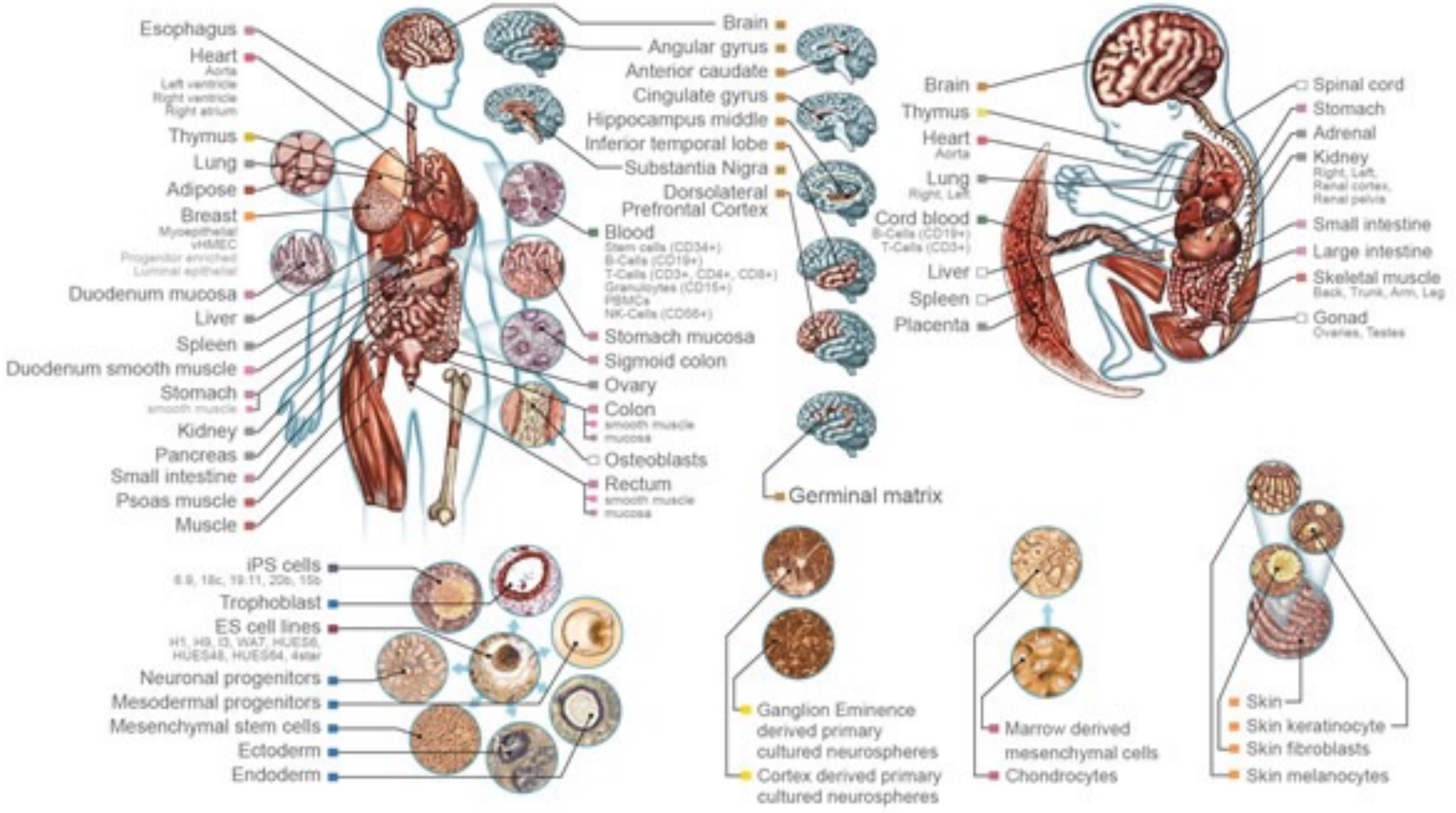


ENCODE Project Consortium 2011. *PLoS Biol* 9:e1001046

DNase-seq and ATAC-seq measure DNA accessibility

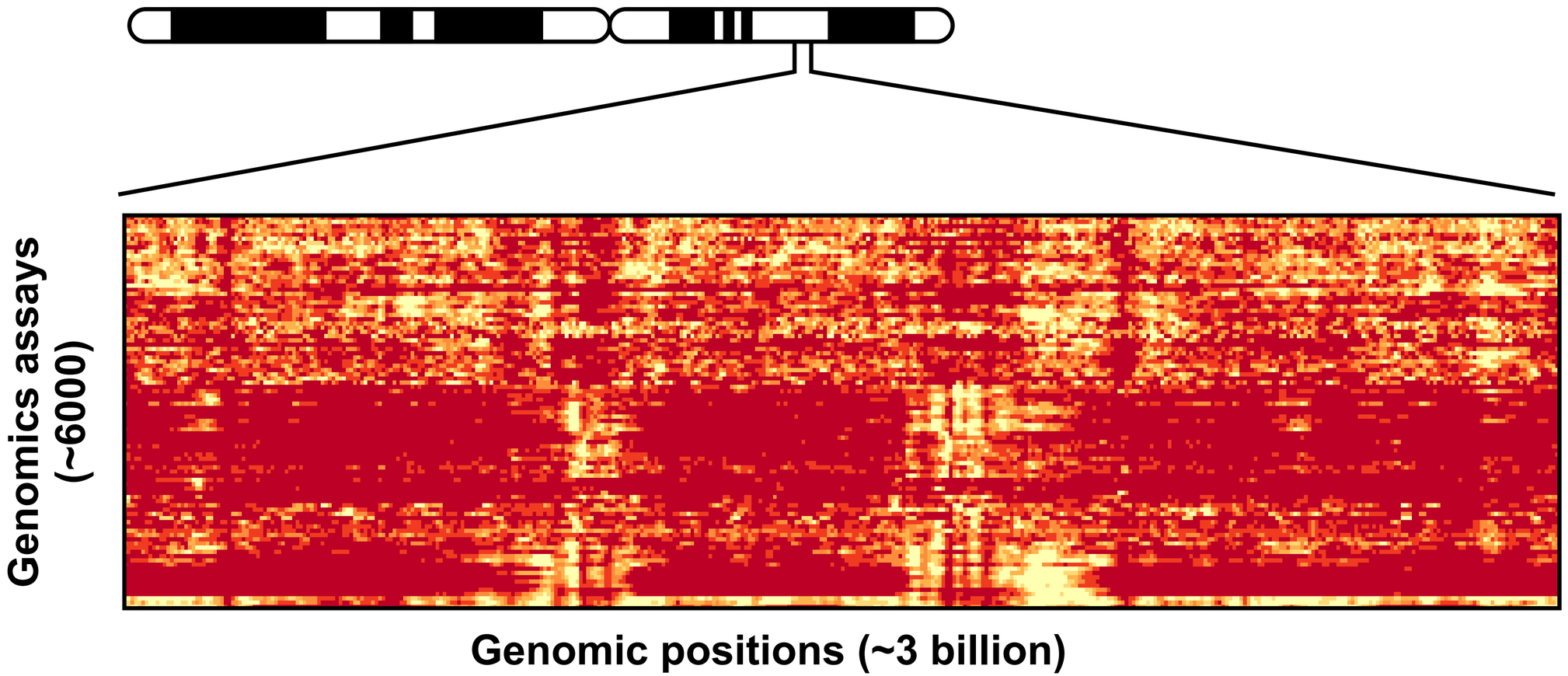


Hundreds of human tissues have been profiled with genomics assays

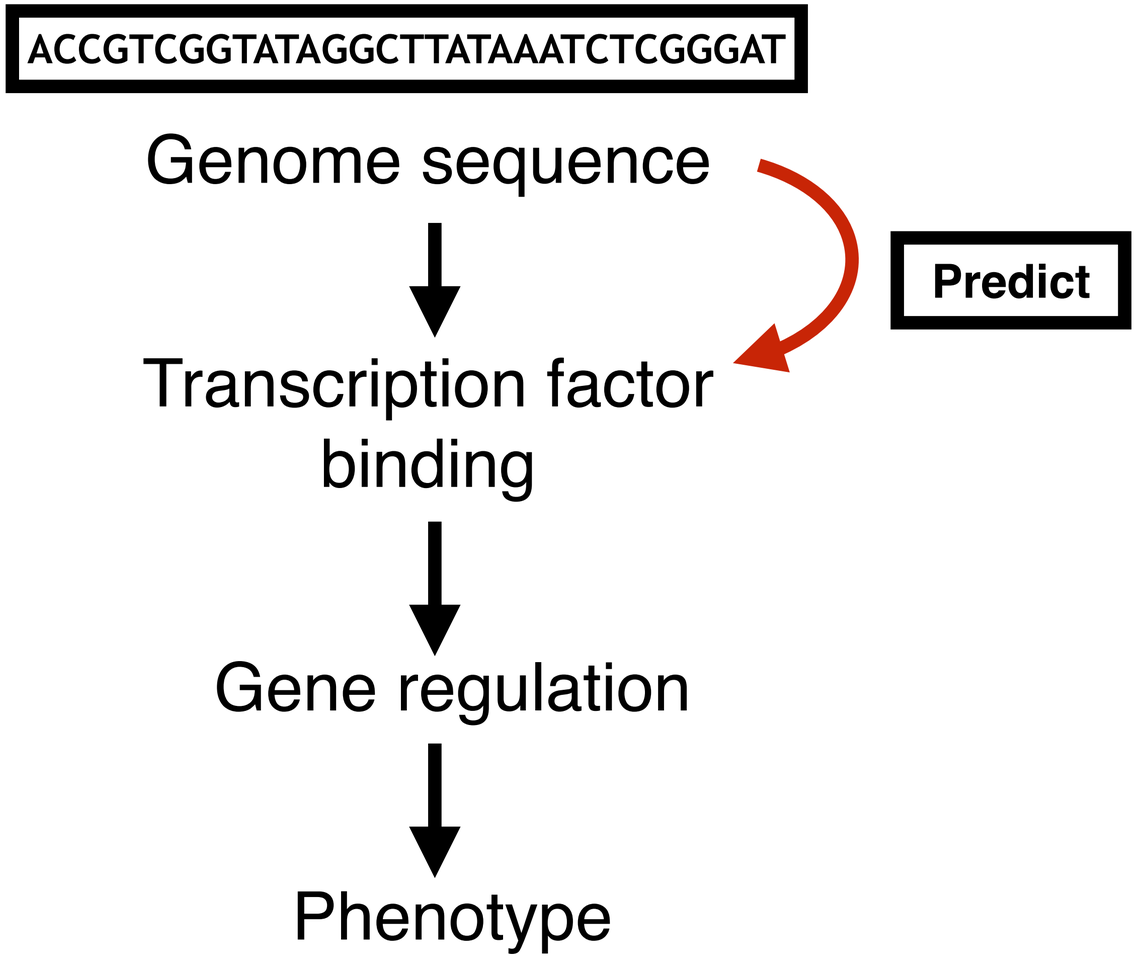


Roadmap Epigenomics Consortium, *Nature* 2015

Sequencing-based assays are a rich data set for understanding the genome



Machine learning methods for the genotype-phenotype relationship, gene regulation and epigenomics



Predicting effects of noncoding variants with deep learning-based sequence model

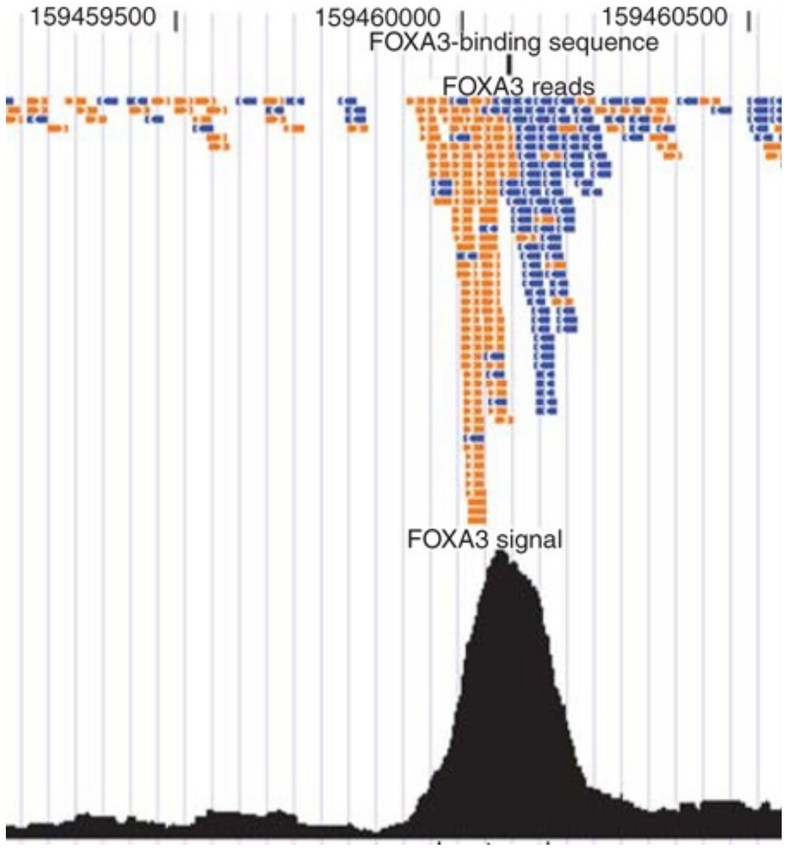
Jian Zhou & Olga G Troyanskaya

Nature Methods 2015

ChIP-seq peak calls indicate confident transcription factor binding sites

- Peak calling: Stack up the reads in the genome; choose the tall stacks.

- **Issues to consider**:
  - Sequencing fragment lengths
  - Sequencing read lengths
  - Experimental biases
  - Mappability
  - GC bias
  - How to pick a threshold and assign statistical confidence



Problem setup

Reference sequence

ACCGTCGGTATAGGCTATAAATCTCGGGAT

CTCF binding in liver cells (ChIP-seq)

Deep neural network classifier

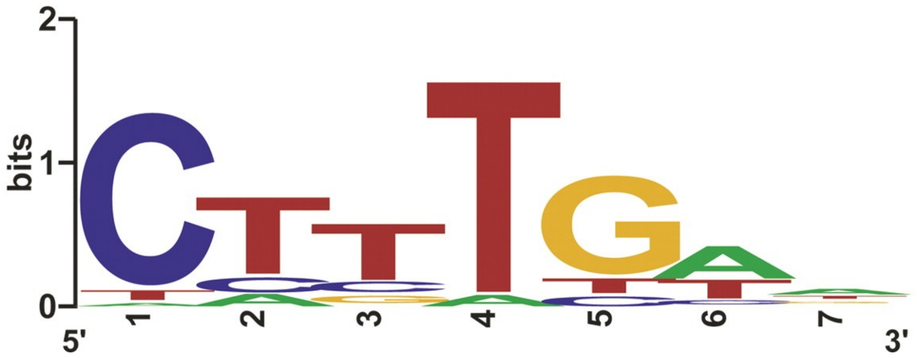
Observed variant

ACCGTCGGTATAGGC**A**TATAAATCTCGGGAT

Does variant affect CTCF binding in liver?

The traditional model for understanding transcription factor binding is the position-weight matrix (PWM)

	1	2	3	4	5	6	7
A	1	4	1	2	0	17	13
C	28	5	5	0	3	3	2
G	0	0	4	0	25	1	7
T	2	22	21	29	4	10	9

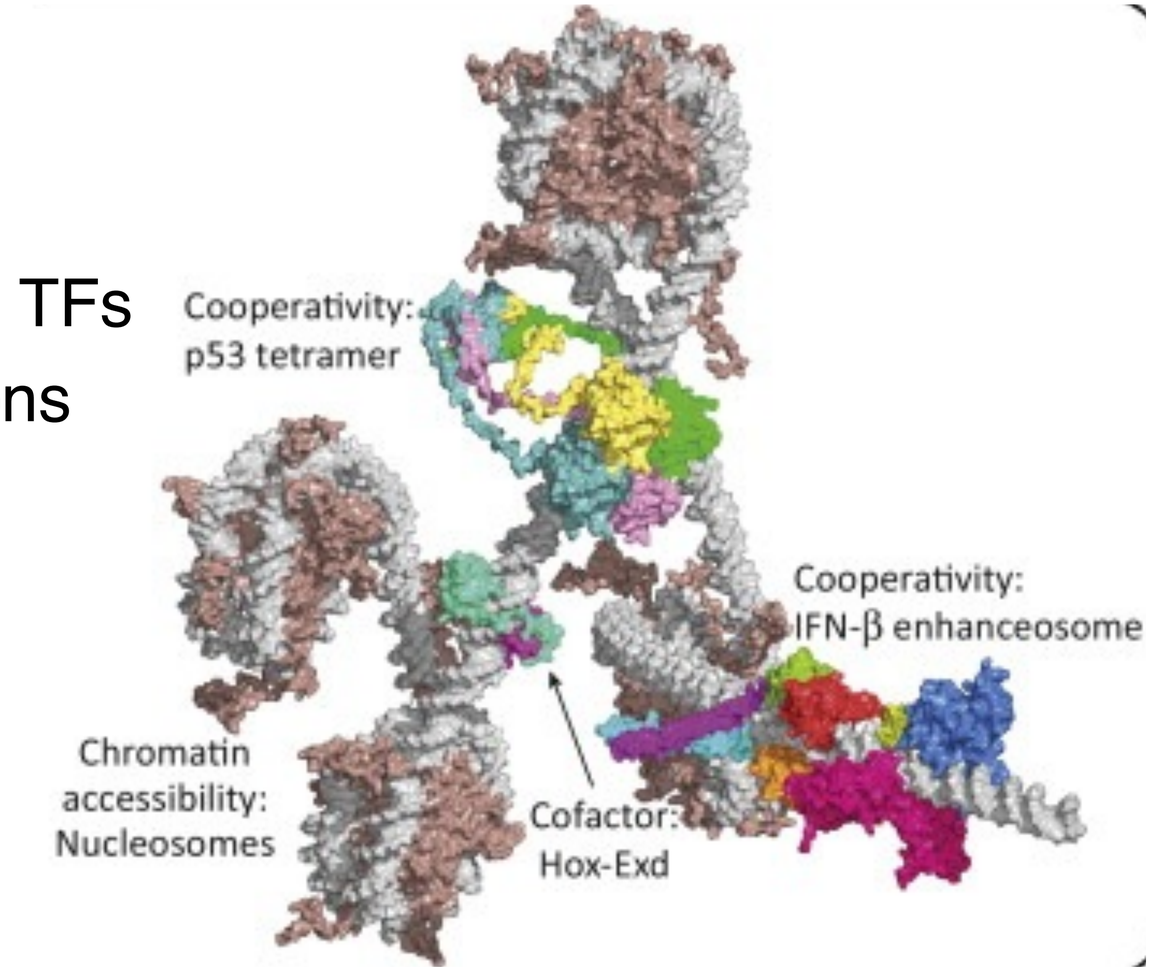


Binds?

ACCGTCGGTATAGGCTTATAAATCTCGGGAT

How can we get a better model than sequence motifs?

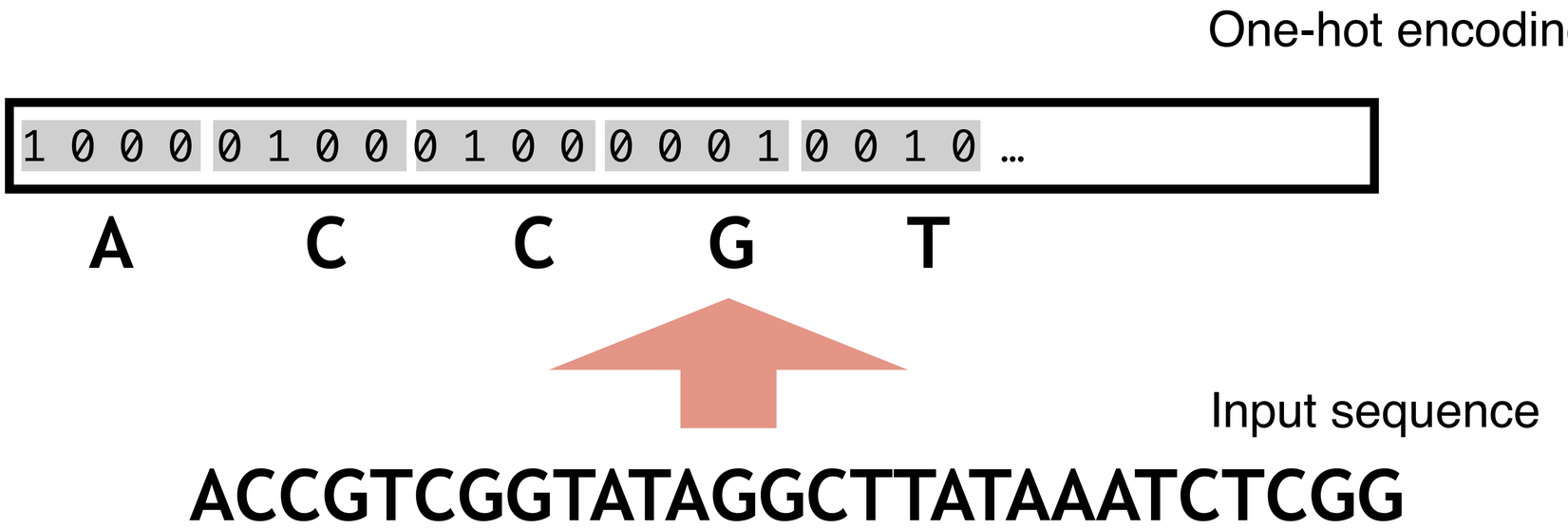
- DNA physical shape
- Variable gaps
- Cooperativity between TFs
- Nucleosome interactions



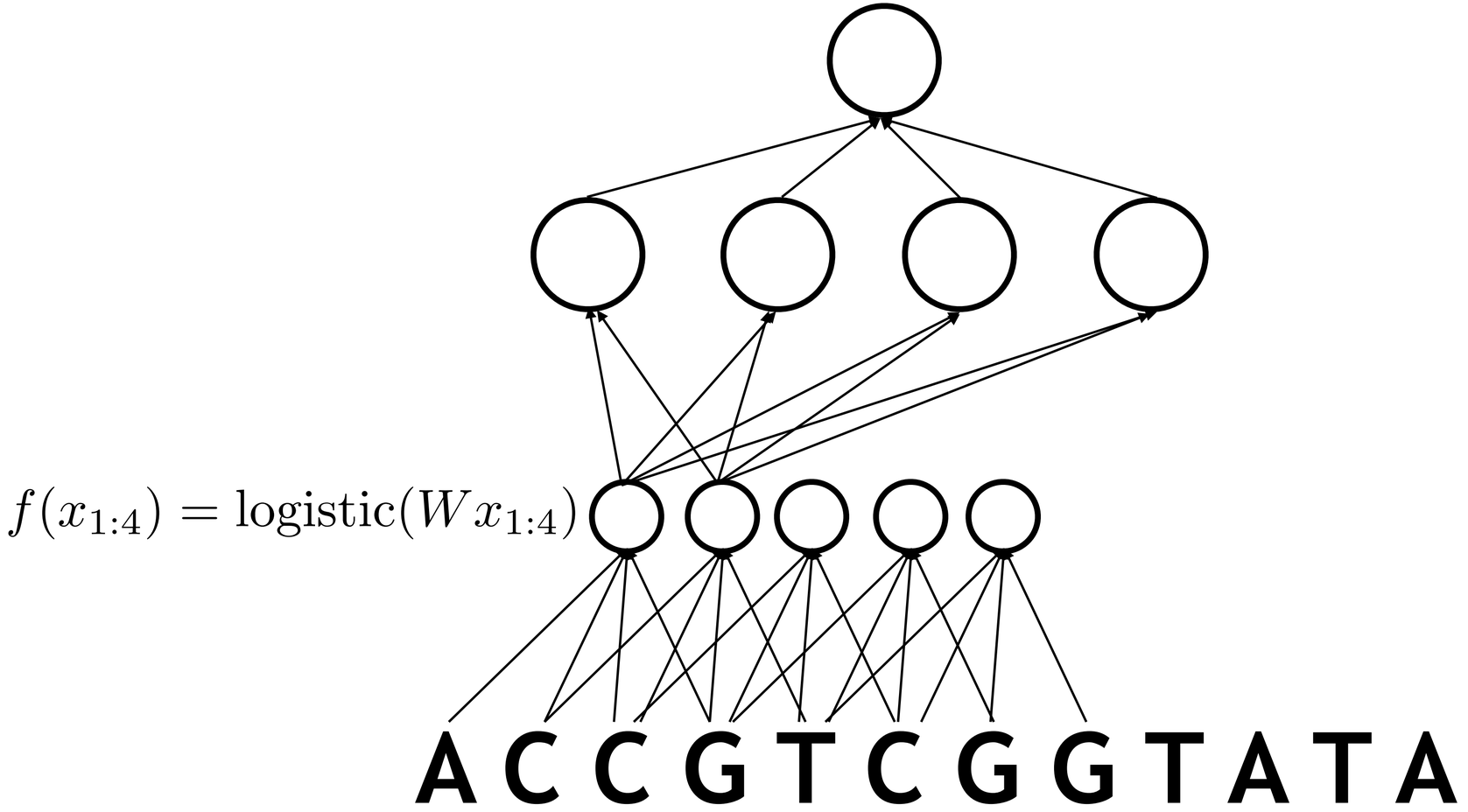
Why deep neural networks?

Deep learning is best when you have more data than sense.
— Jacob Schreiber

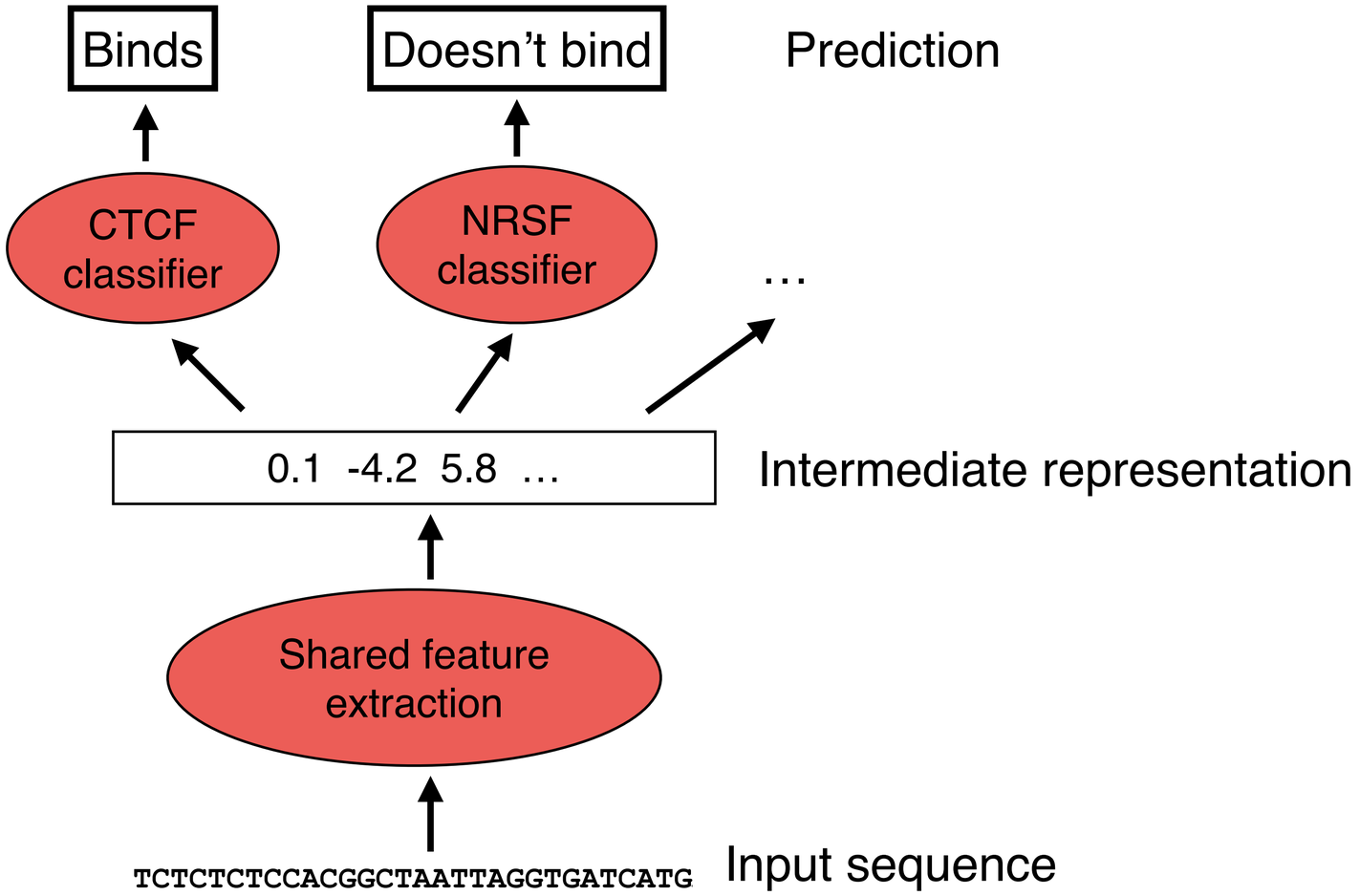
Sequence representation: one-hot encoding



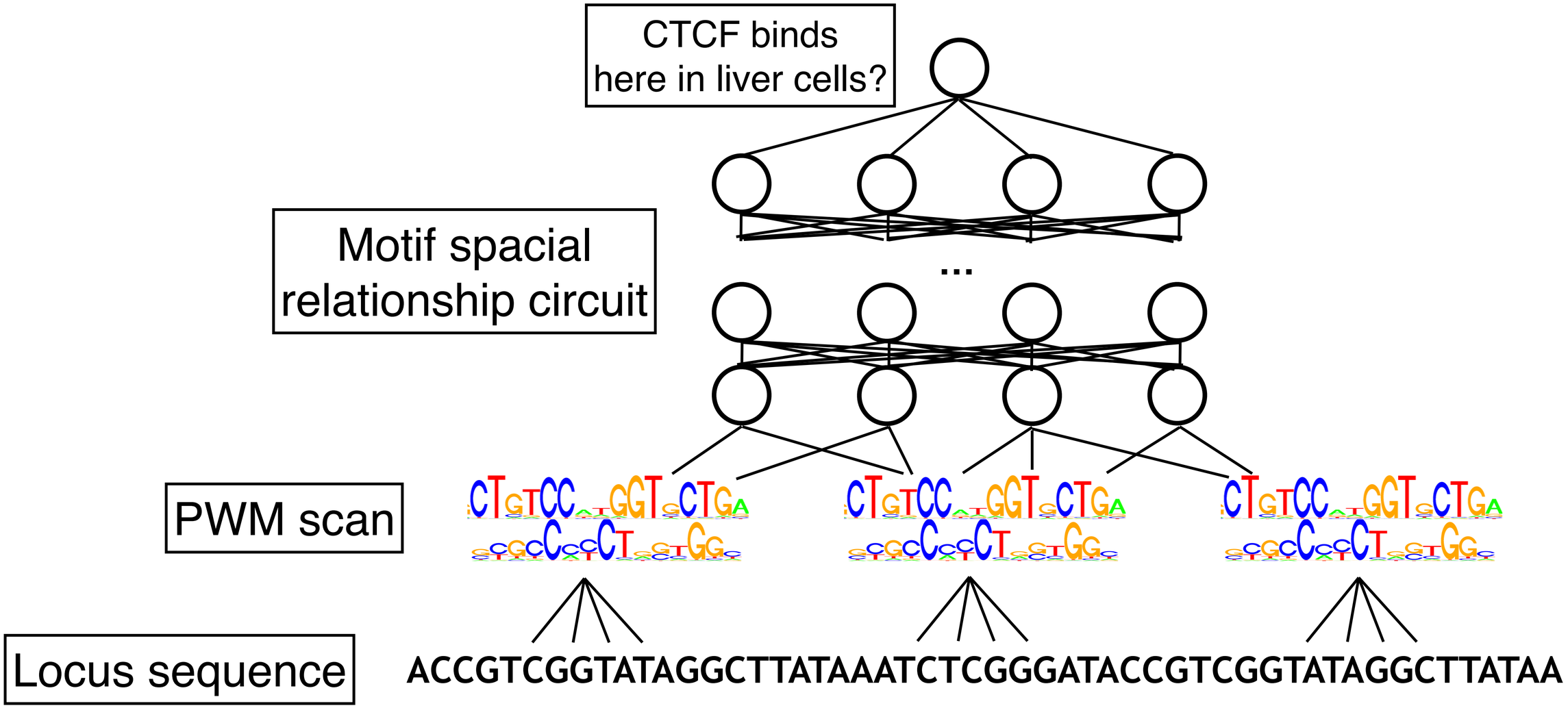
A convolutional network reduces parameters by applying the same function across each portion of the input



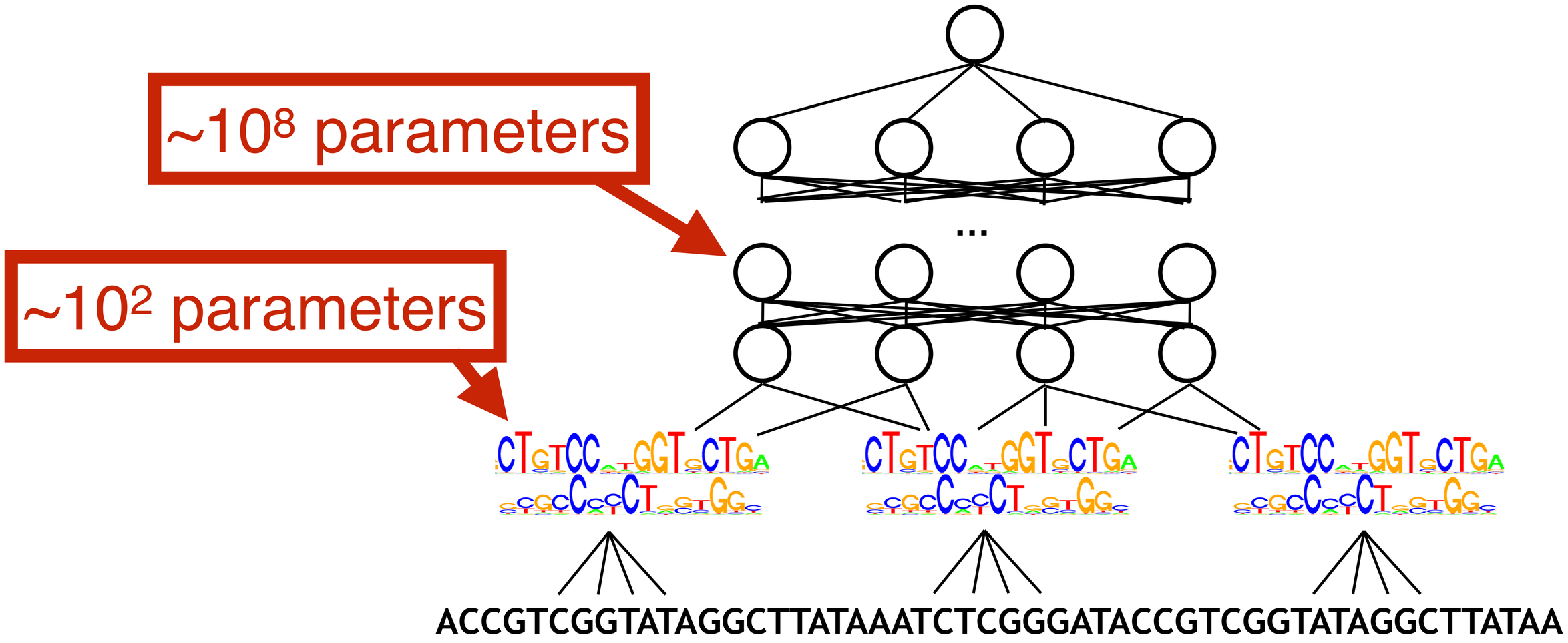
A multi-task approach shares representations between factors



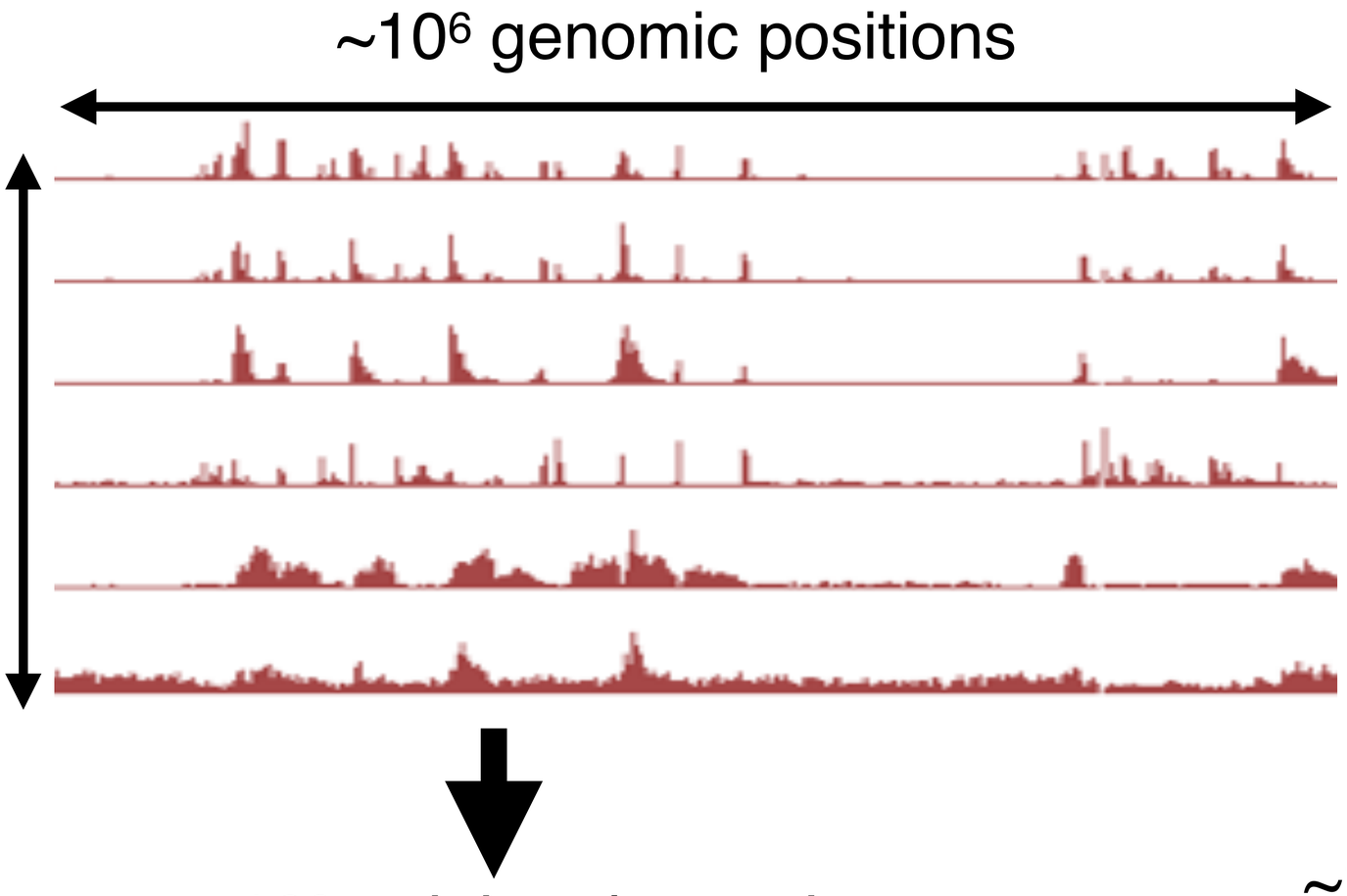
The deep neural network captures complex patterns of motif occurrence



Concern: Deep neural networks have a lot of parameters to train



We have plenty of data to train a deep model



~10^9 training data points > ~10^8 neural network parameters

Supplementary Note. DeepSEA model configuration

Model Architecture:

1. Convolution layer (320 kernels. Window size: 8. Step size: 1.)
2. Pooling layer (Window size: 4. Step size: 4.)
3. Convolution layer (480 kernels. Window size: 8. Step size: 1.)
4. Pooling layer (Window size: 4. Step size: 4.)
5. Convolution layer (960 kernels. Window size: 8. Step size: 1.)
6. Fully connected layer (925 neurons)
7. Sigmoid output layer

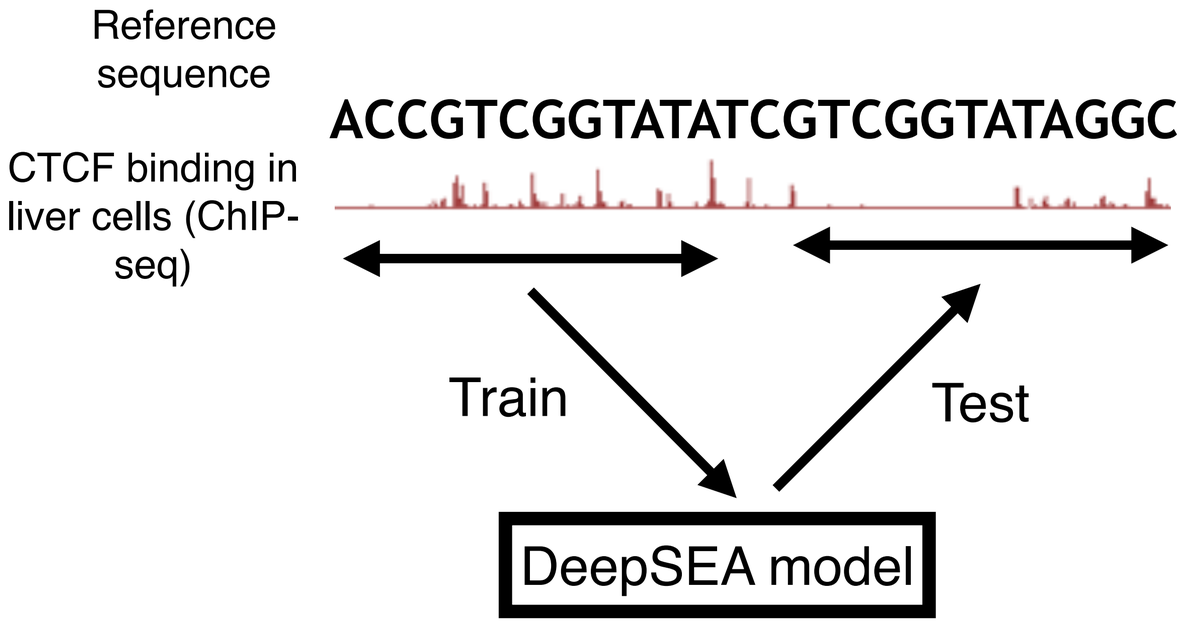
Regularization Parameters:

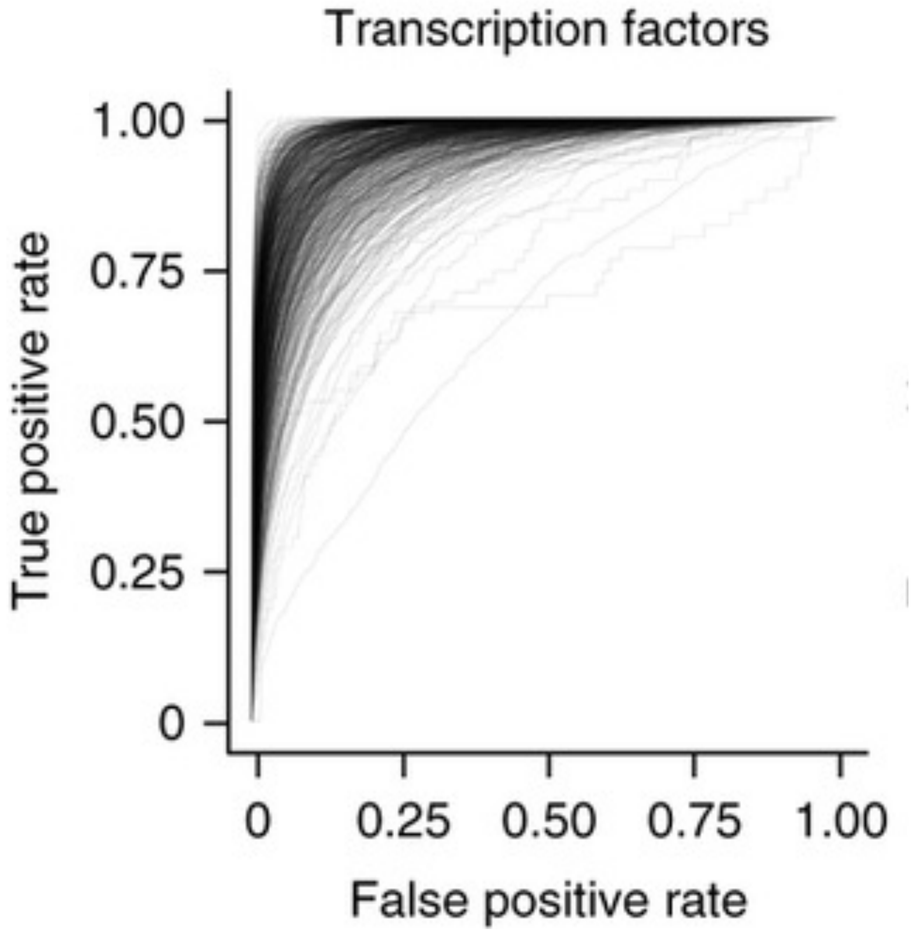
Dropout proportion (proportion of outputs randomly set to 0):

- Layer 2: 20%
- Layer 4: 20%
- Layer 5: 50%
- All other layers: 0%

L2 regularization (λ_1): 5e-07
L1 sparsity (λ_2): 1e-08
Max kernel norm (λ_3): 0.9

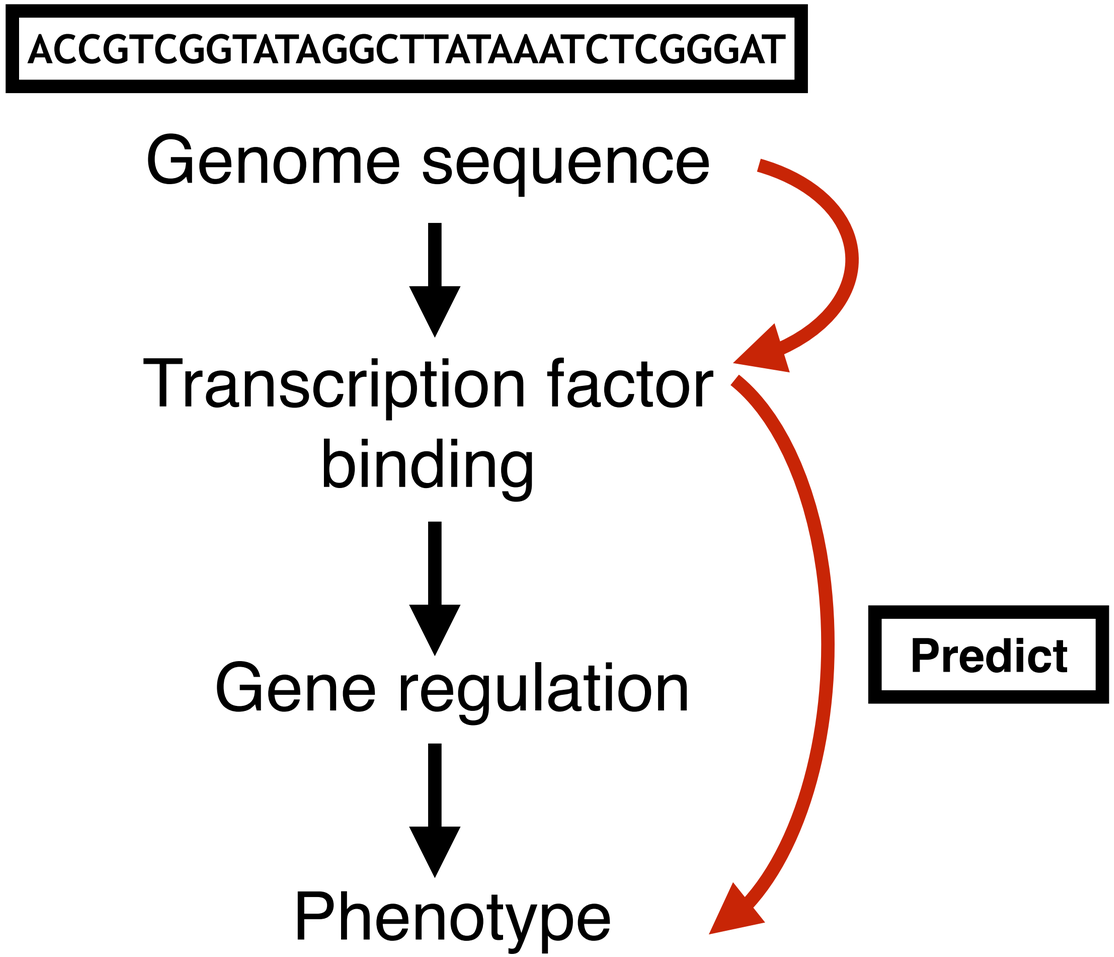
DeepSEA accurately predicts TF binding and DNase hypersensitivity



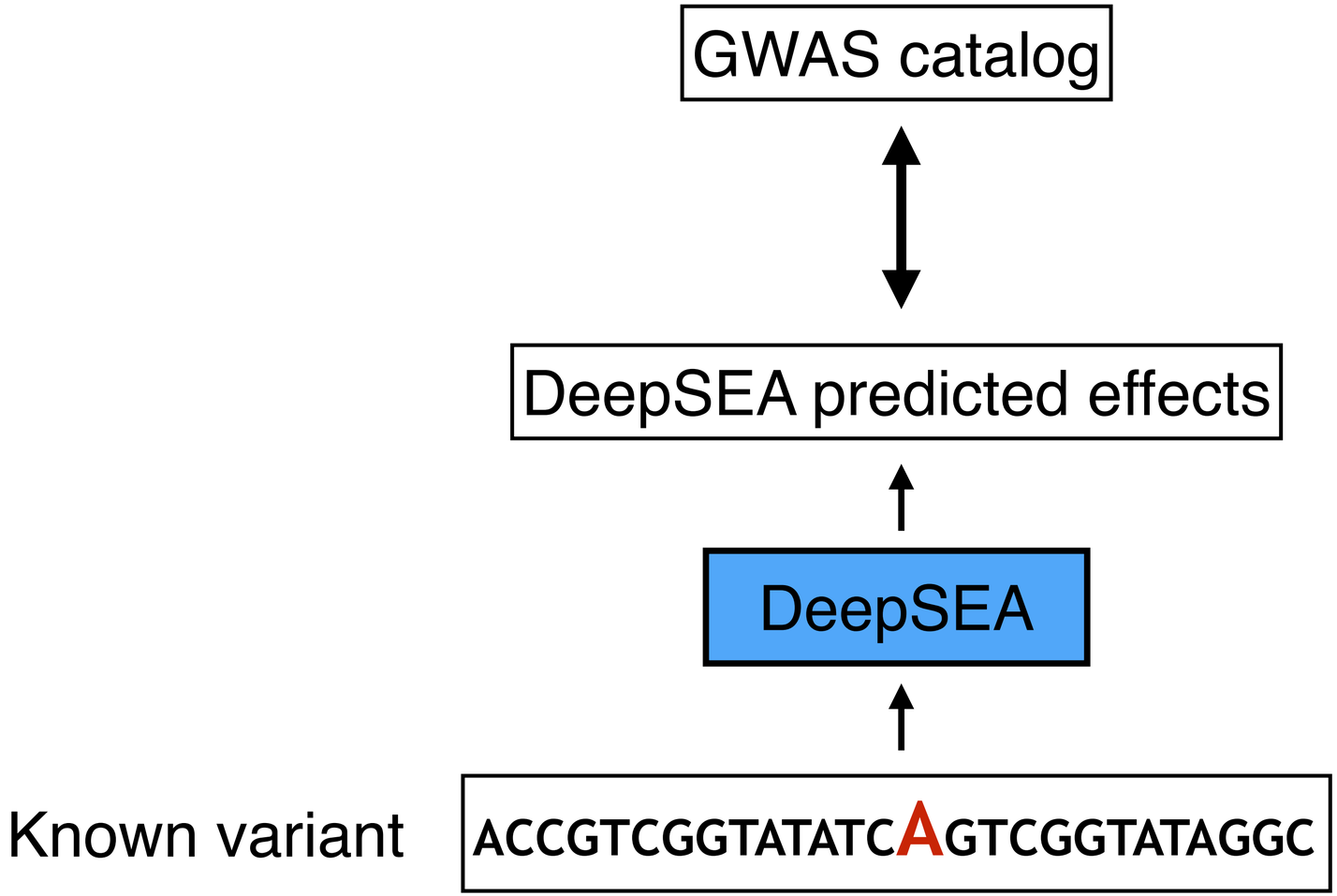


Mean AUC: 0.958

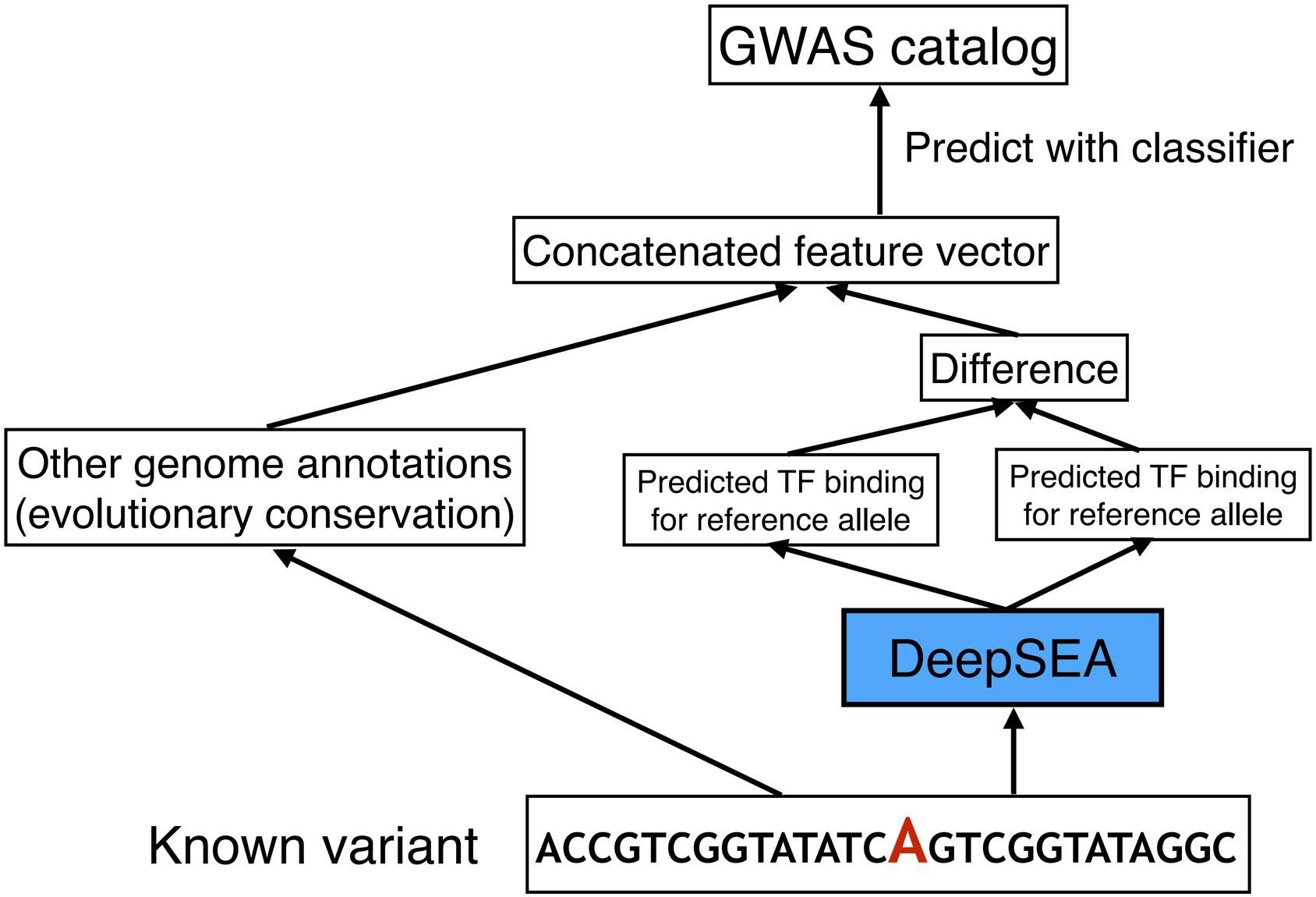
DeepSEA can perform in-silico mutagenesis



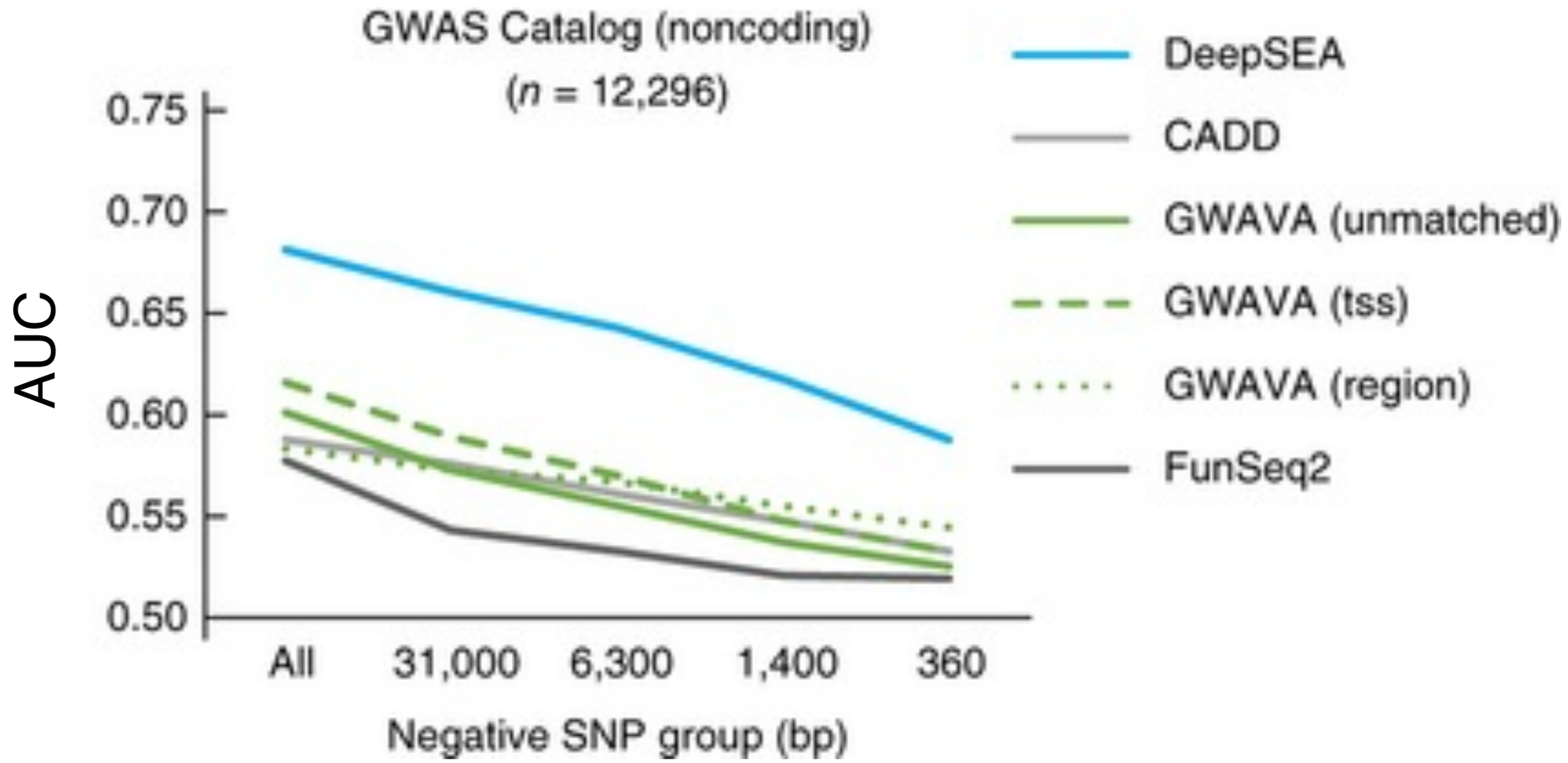
Can DeepSEA predict known regulatory variants?



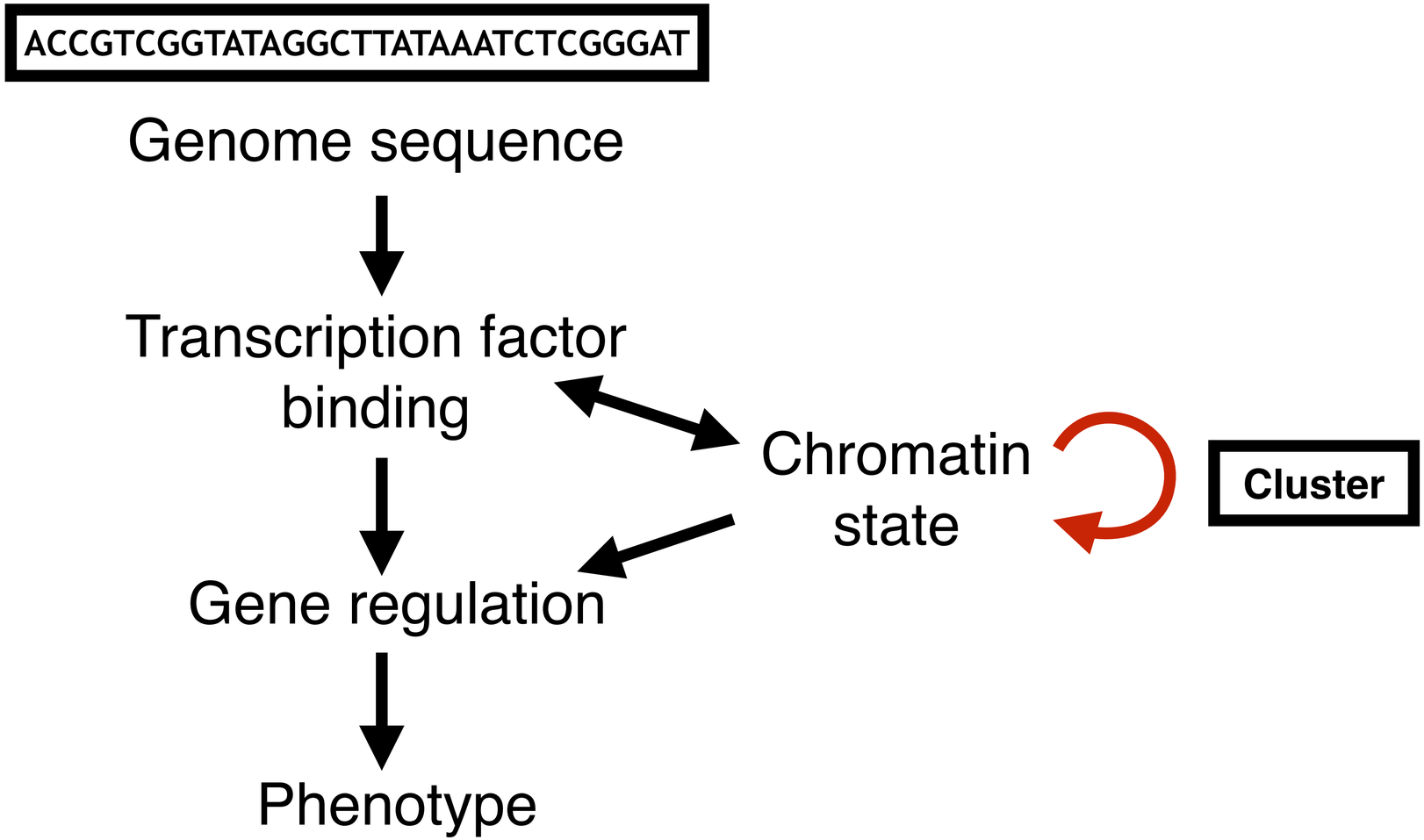
Can DeepSEA predict known regulatory variants?



DeepSEA accurately predicts known regulatory variants



Machine learning methods for the genotype-phenotype relationship, gene regulation and epigenomics

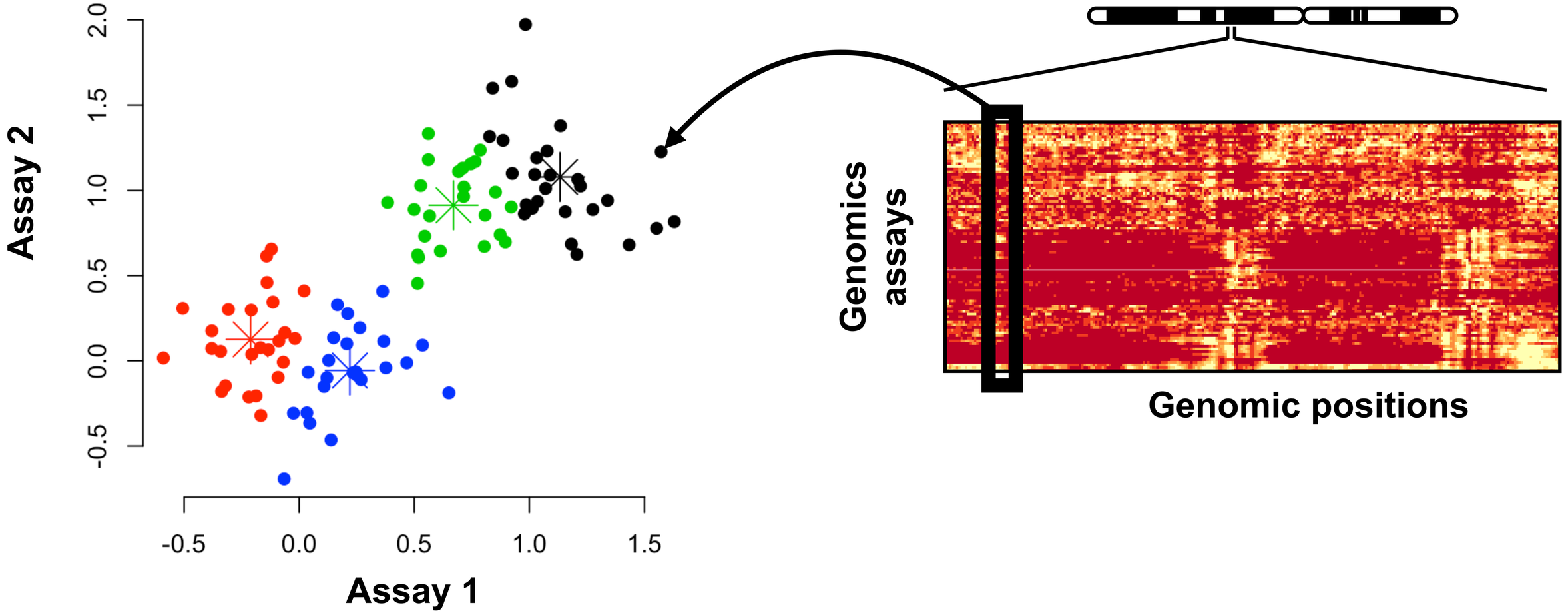


Segmentation and genome annotation

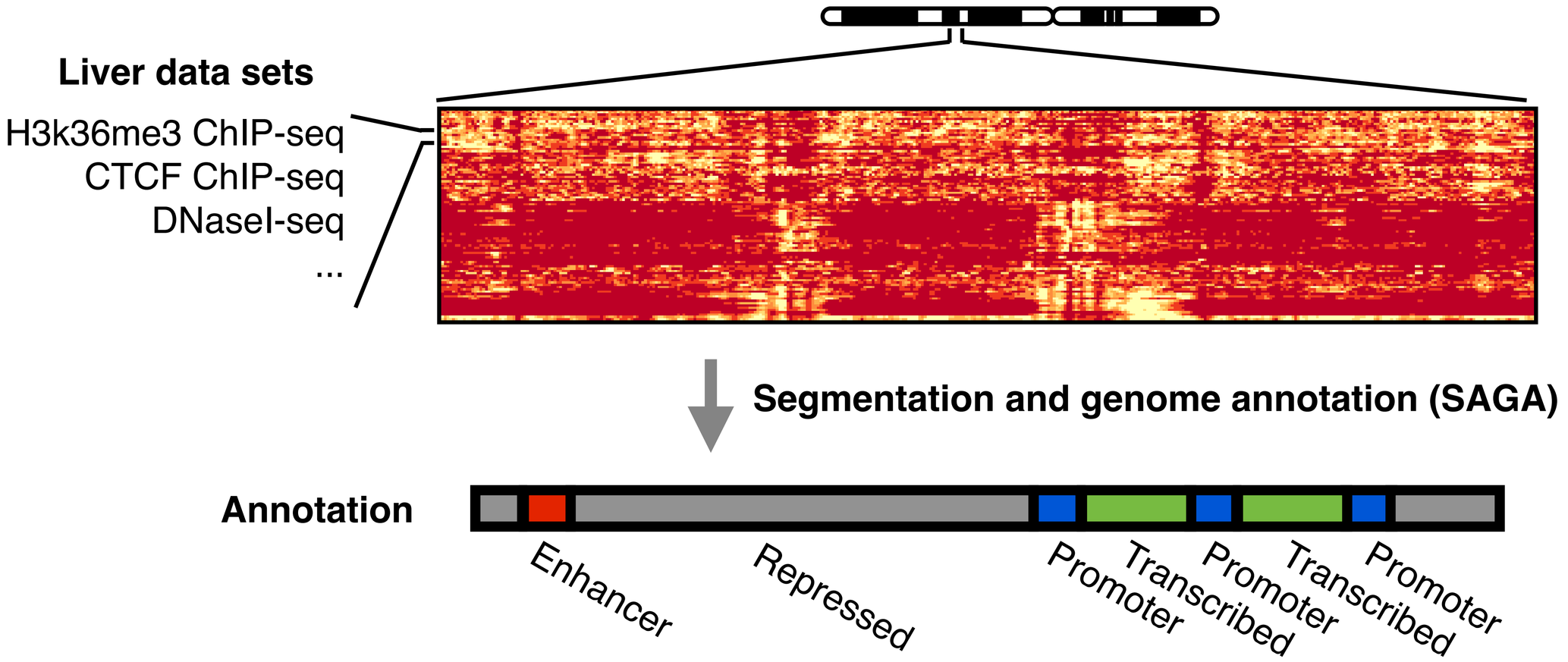
Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes J, Noble WS. Nature Methods 2012. *Unsupervised pattern discovery in human chromatin structure through genomic segmentation*

Maxwell W Libbrecht, Oscar L Rodriguez, Zhiping Weng, Jeffrey A Bilmes, Michael M Hoffman, William Stafford Noble. Genome Biology 2019. *A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types*

Unsupervised machine learning is a way to find patterns in a data set

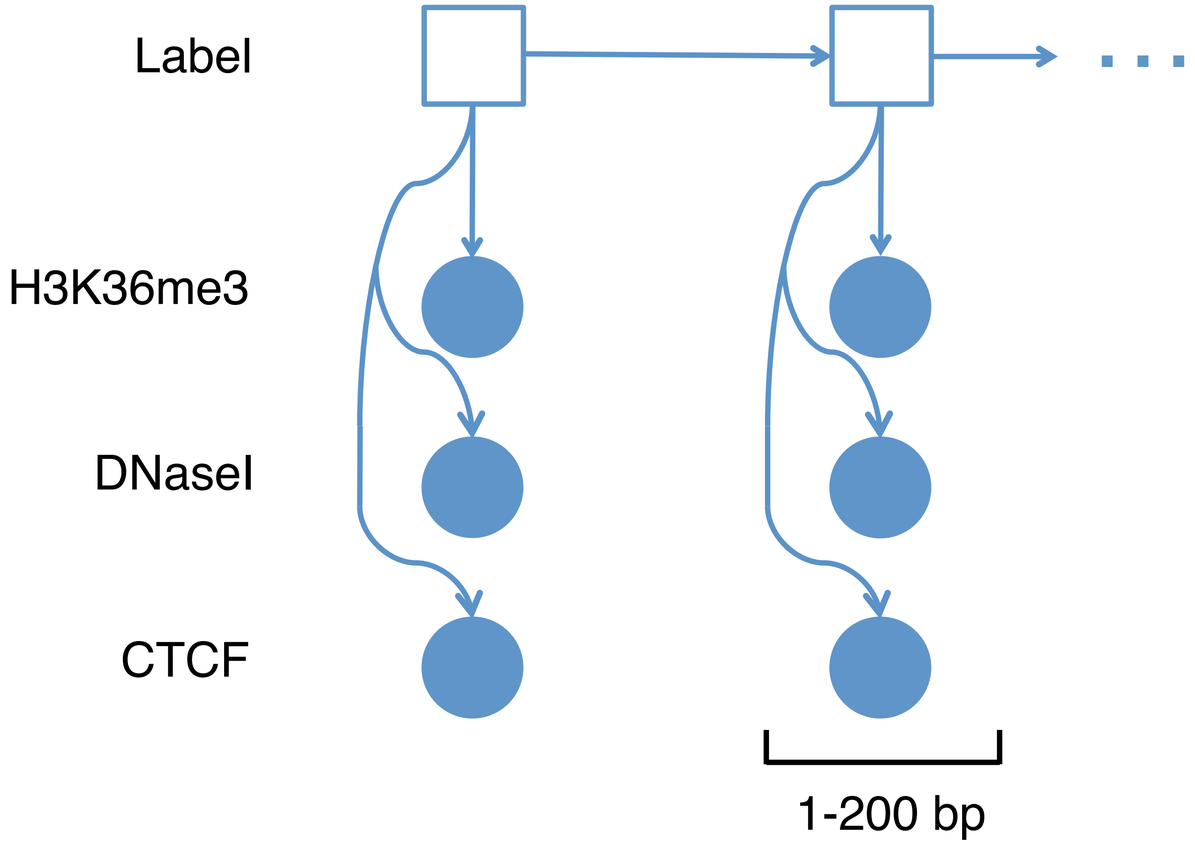


Segmentation and genome annotation (SAGA) algorithms partition and label the genome on the basis of genomics data sets



ChromHMM: Ernst, J. and Kellis, M. *Nature Biotechnology*, 2010
Segway: Hoffman, M et al. *Nature Methods*, 2012

Method: unsupervised probabilistic graphical model

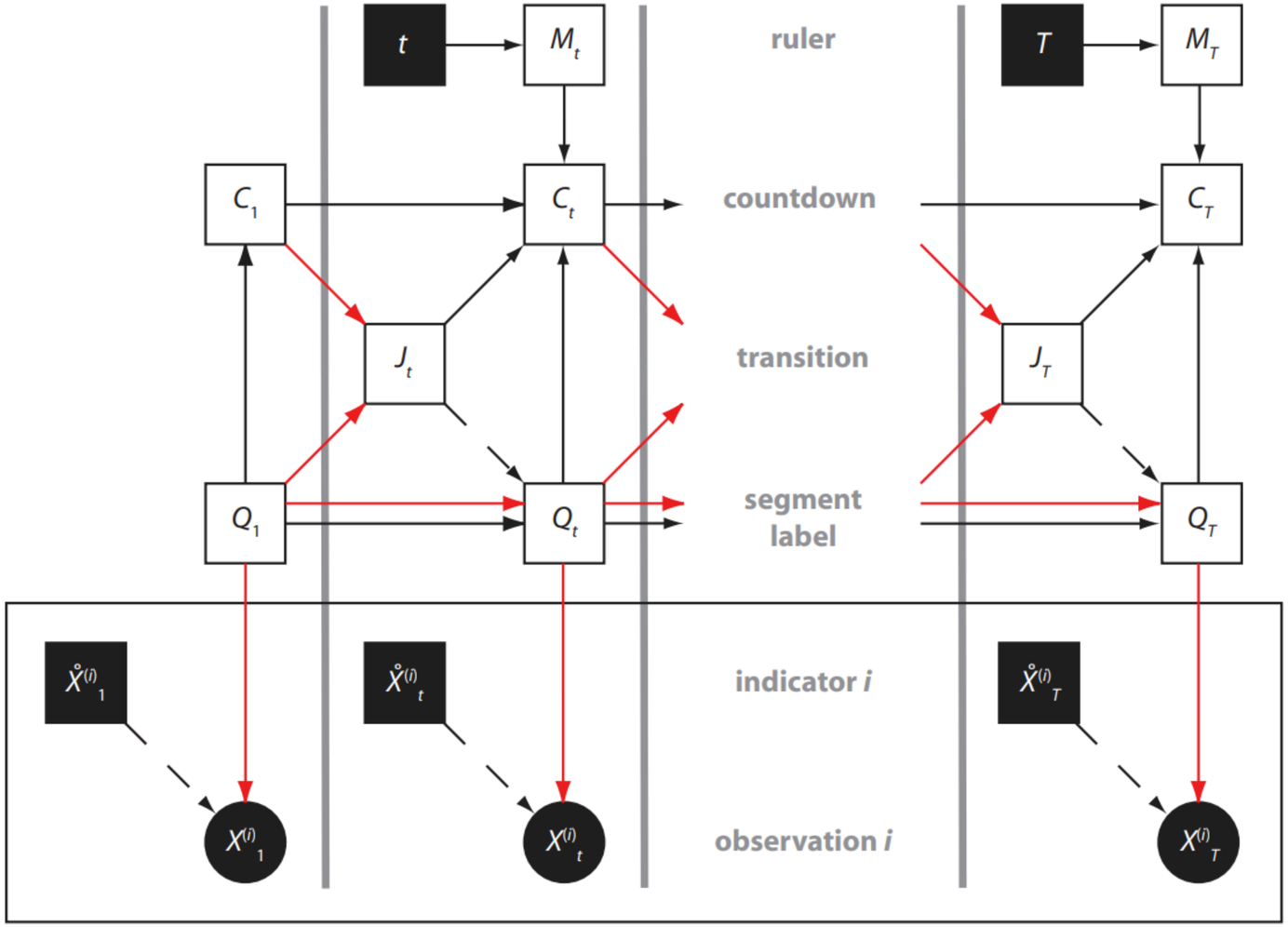


hidden random variable

observed random variable

Training: Expectation-Maximization (EM) algorithm

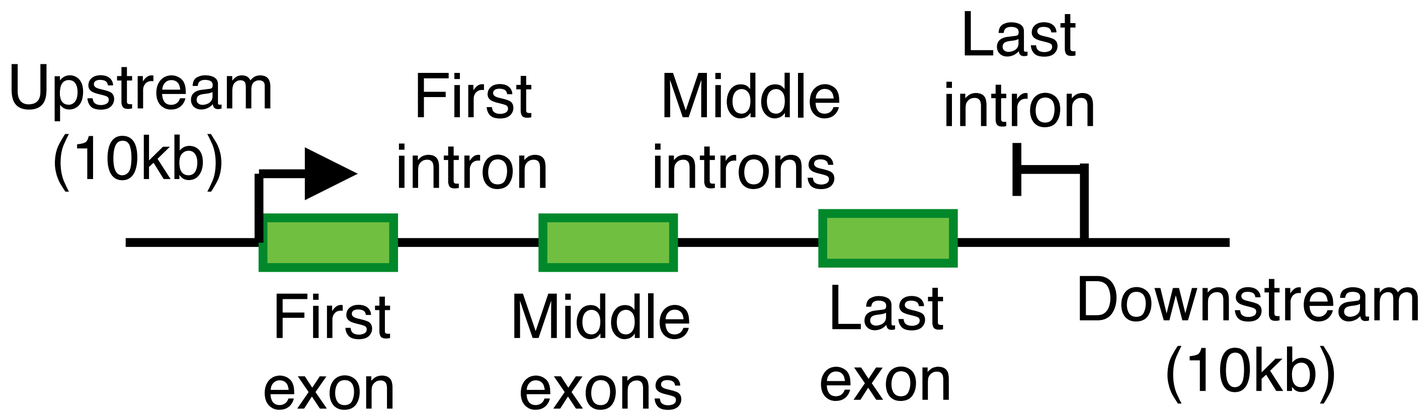
Full dynamic Bayesian network model



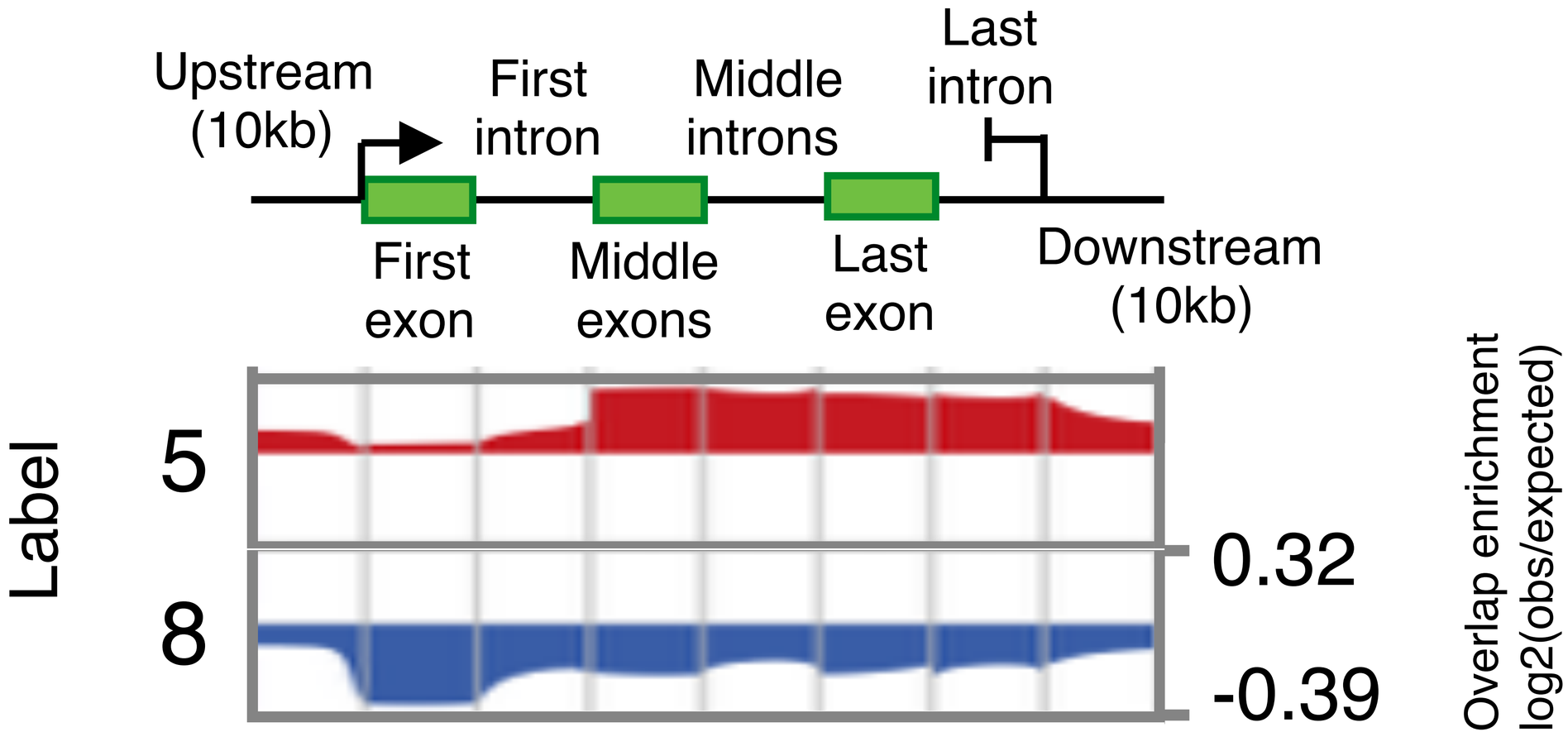
$i \in [1, n]$

What types of genomic elements did the algorithm find?

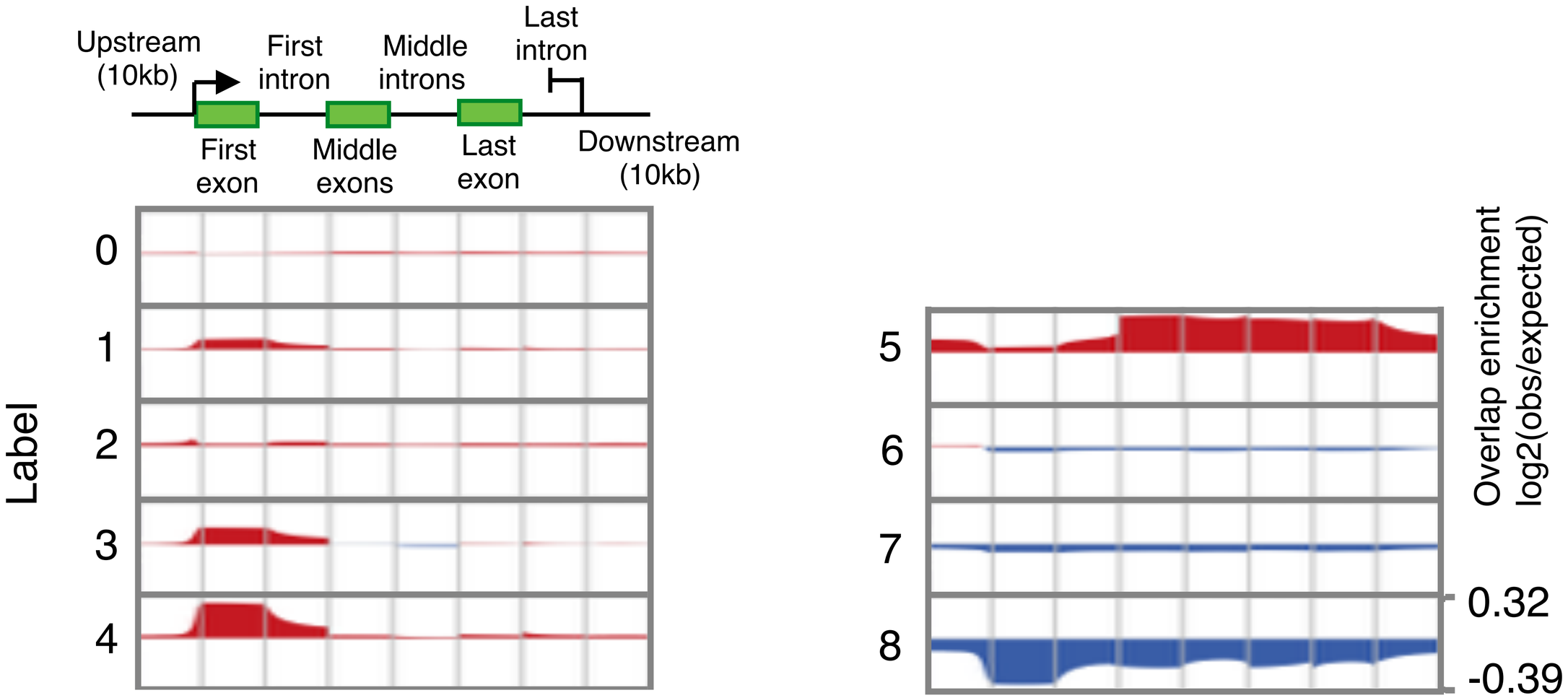
What types of genomic elements did the algorithm find?



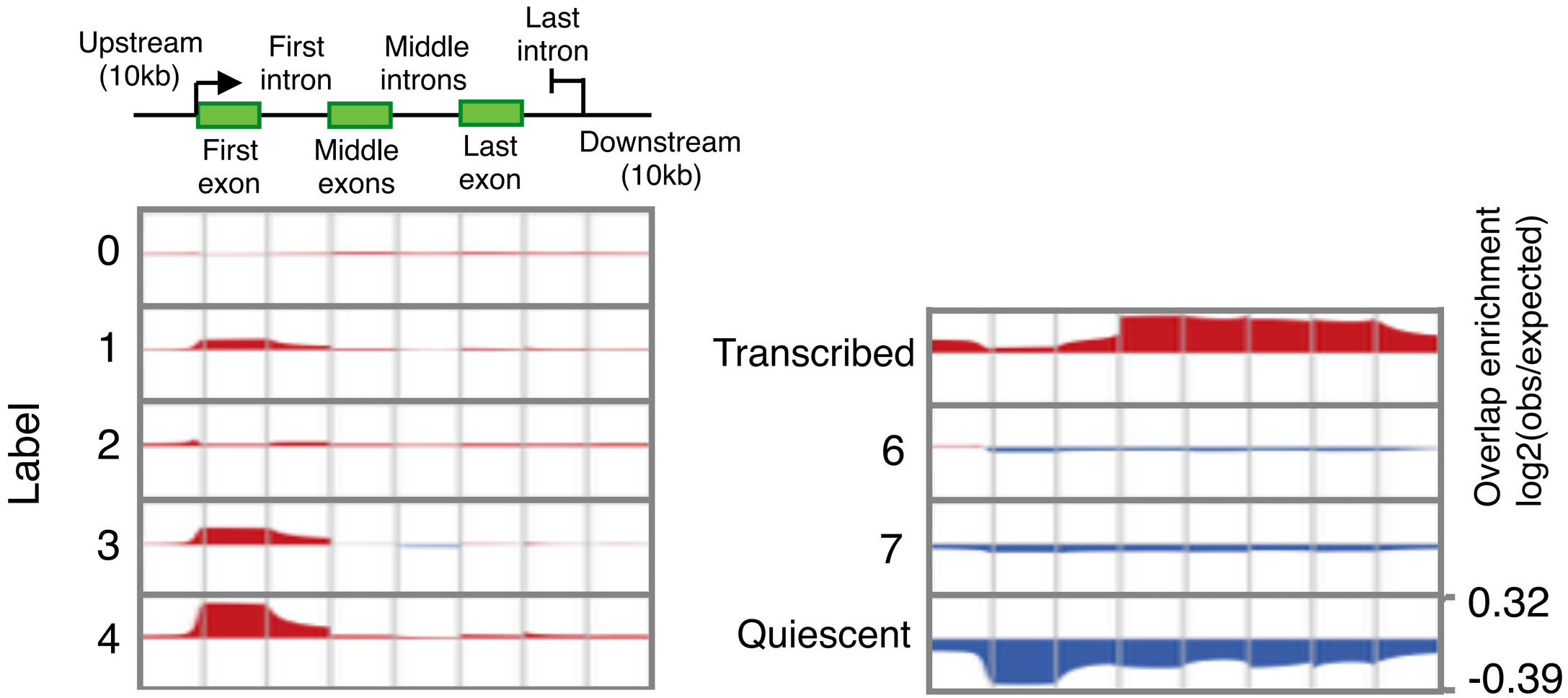
What types of genomic elements did the algorithm find?



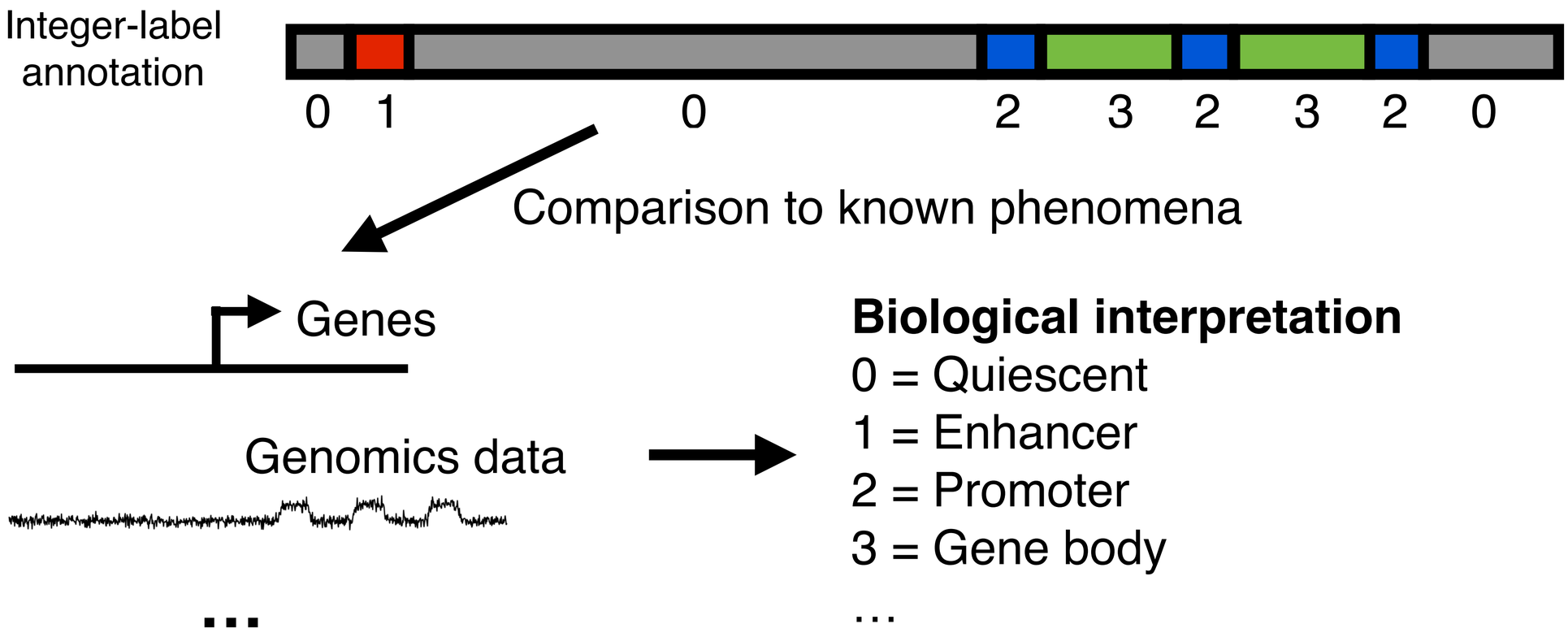
What types of genomic elements did the algorithm find?



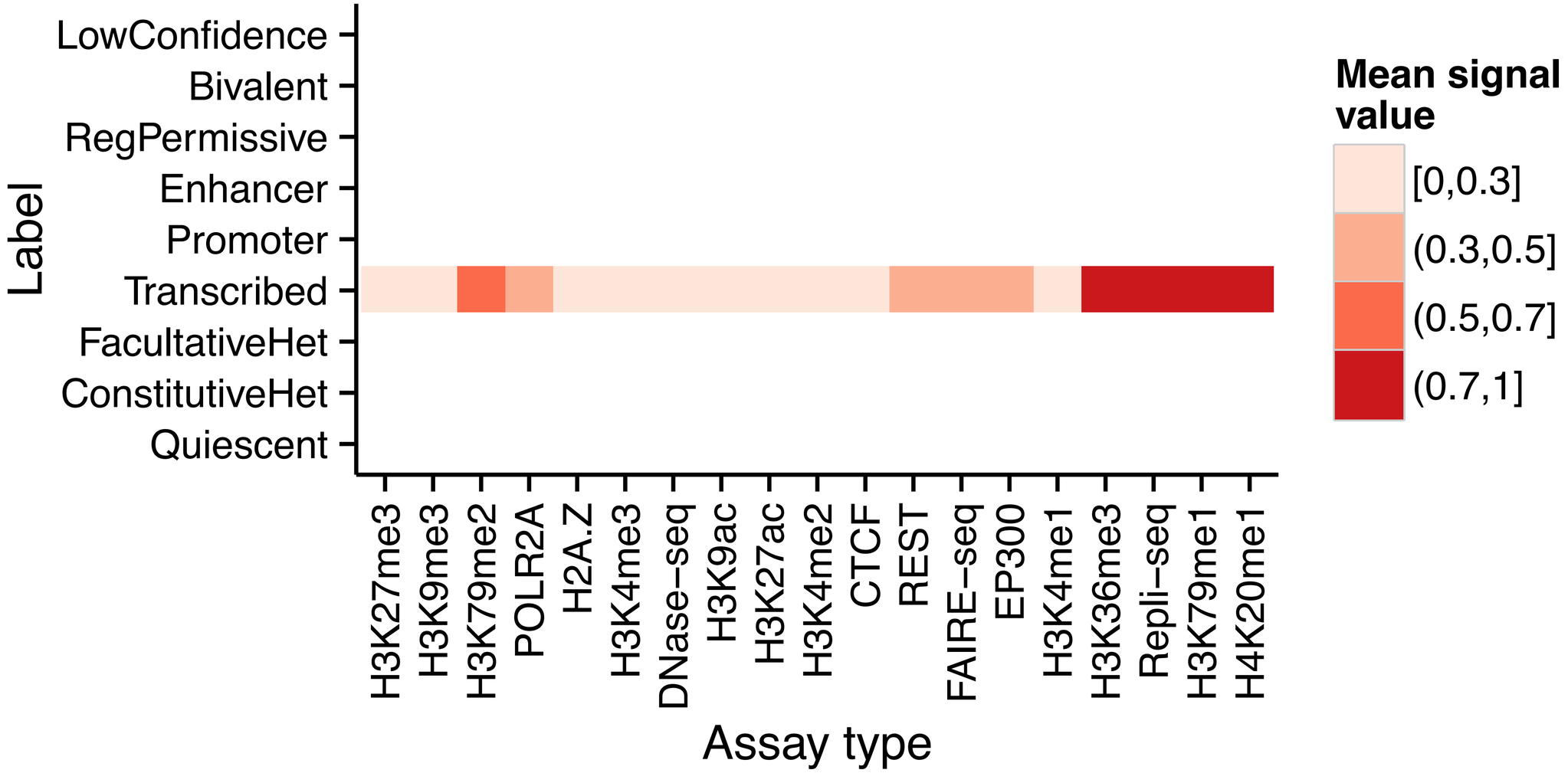
What types of genomic elements did the algorithm find?



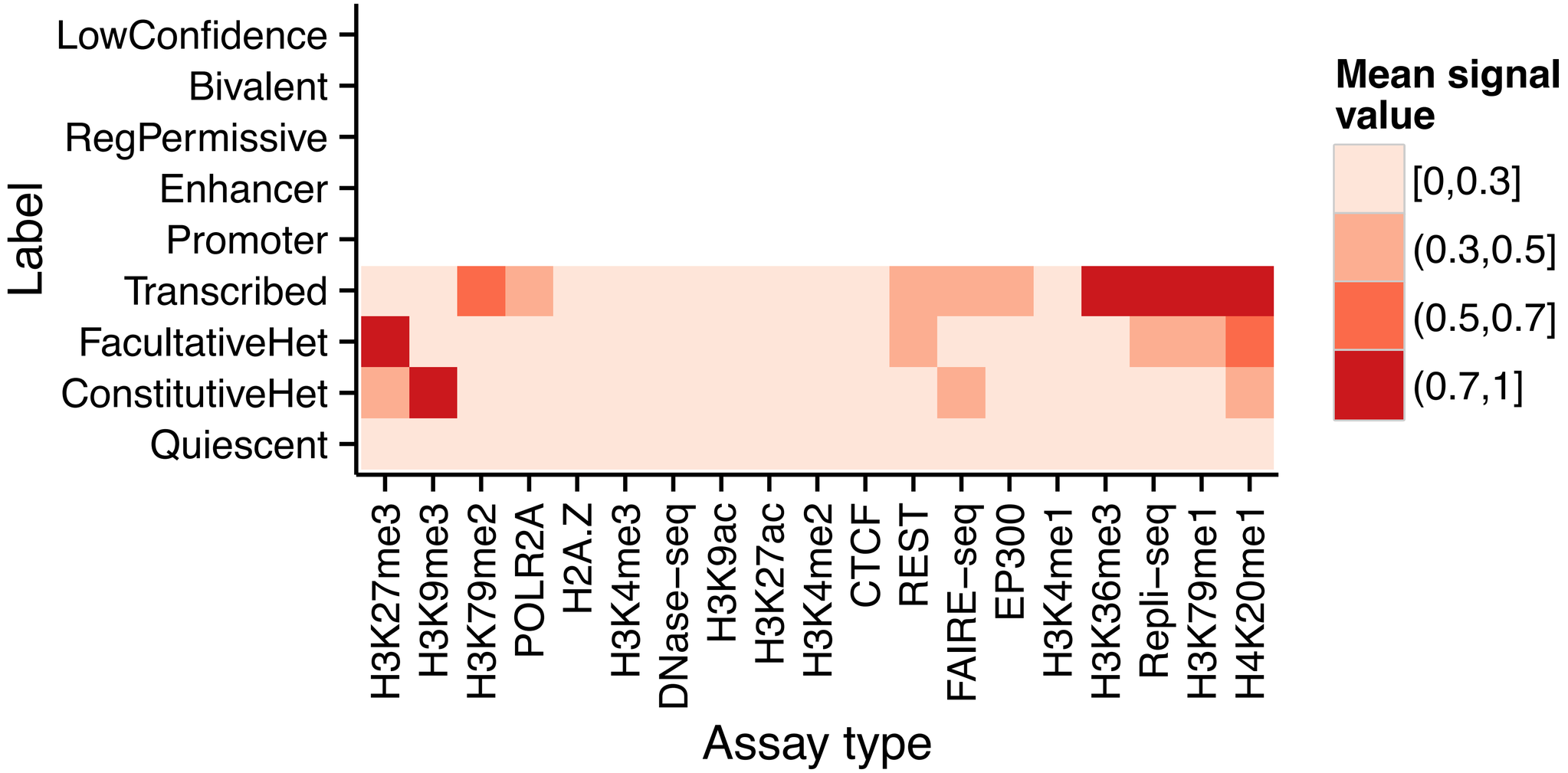
What biological phenomenon does each unsupervised label correspond to?



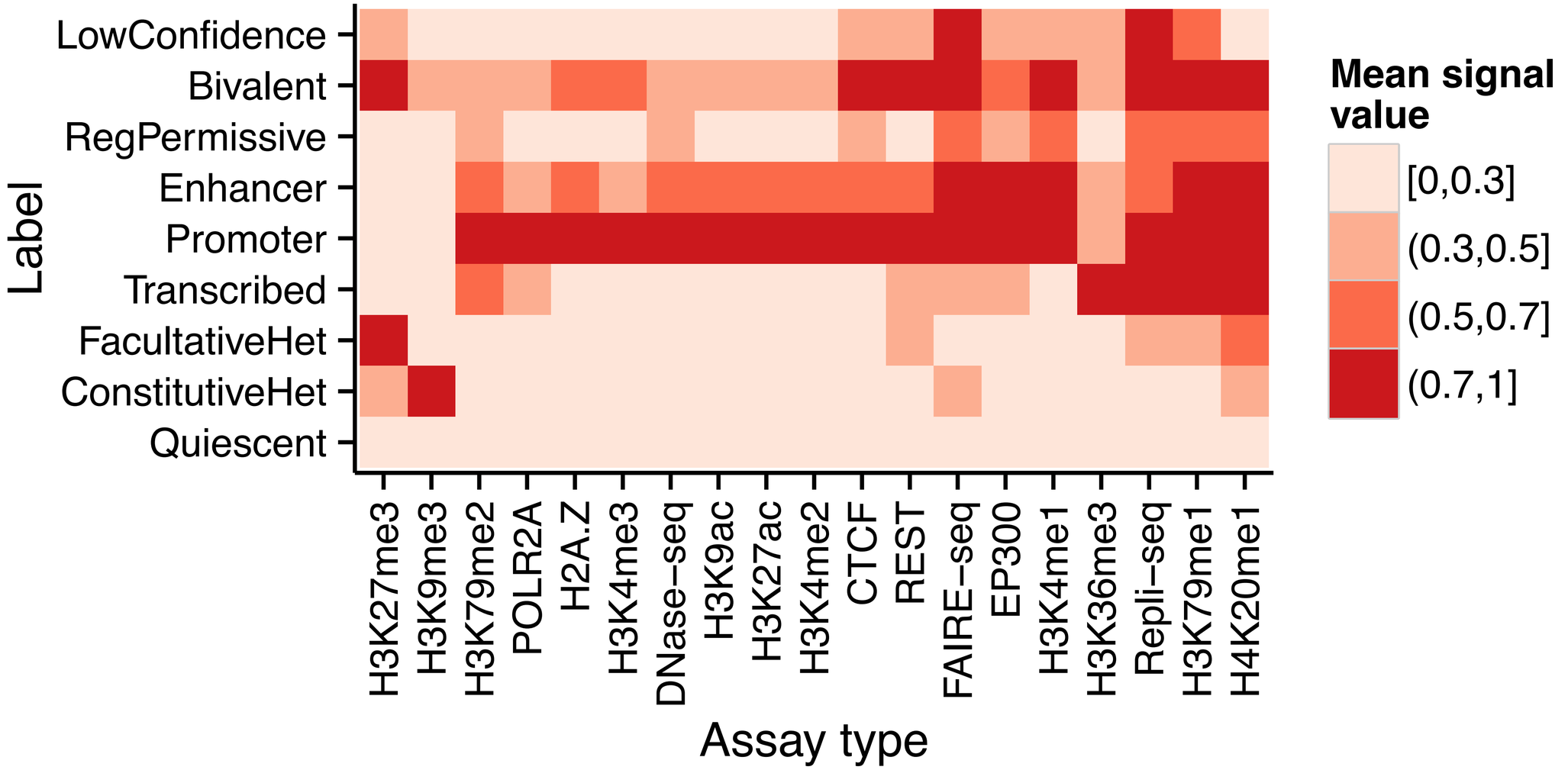
Unsupervised annotation discovers several types of regulatory elements



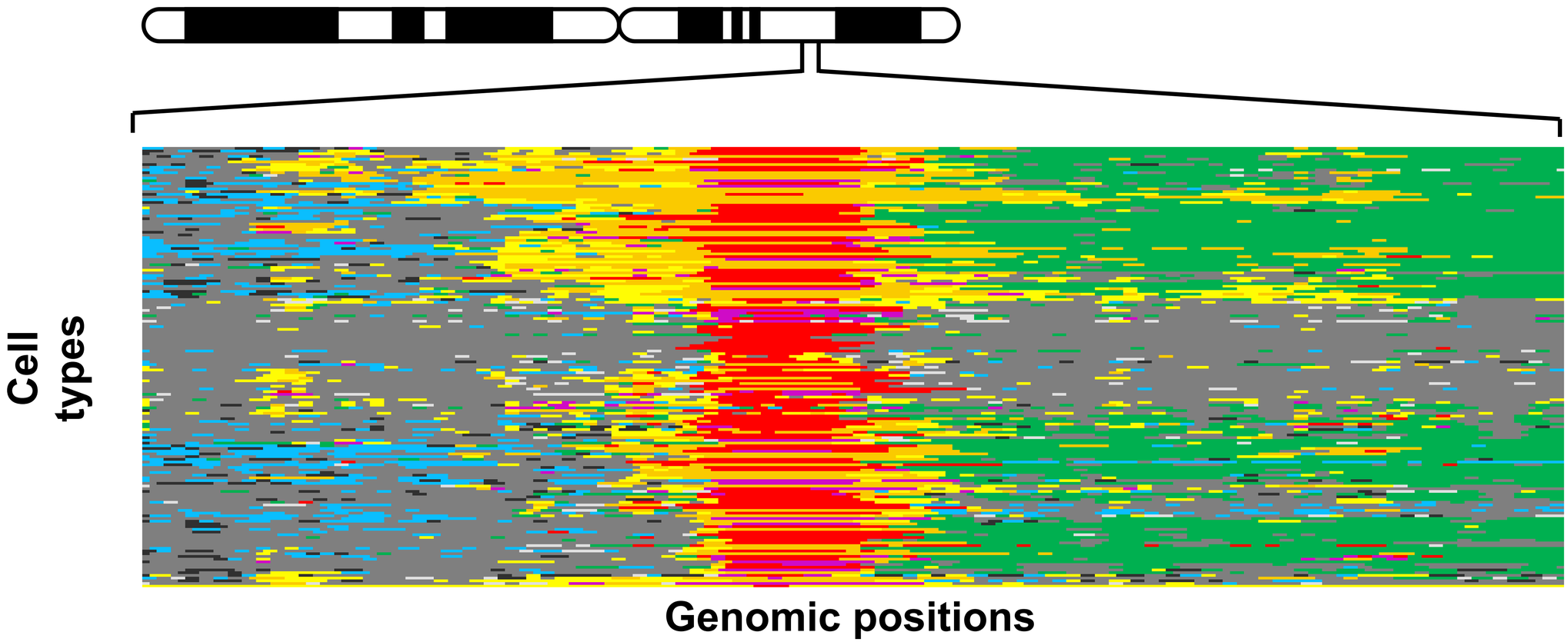
Unsupervised annotation discovers several types of regulatory elements



Unsupervised annotation discovers several types of regulatory elements



Annotations of hundreds of human cell types



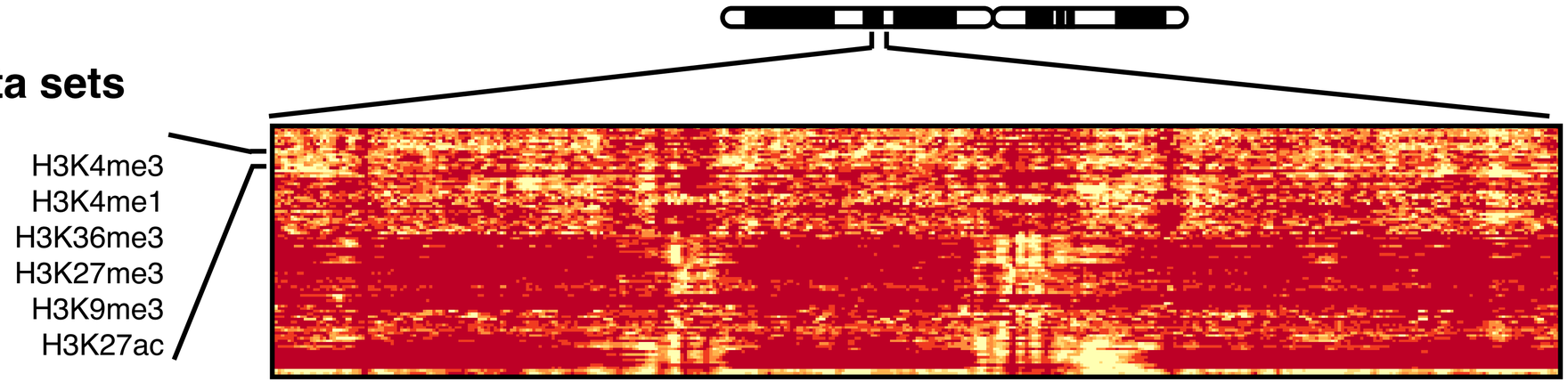
- 164 human cell types
- 1,615 genome-wide assays

Label type

- LowConfidence
- Quiescent
- ConstitutiveHet
- FacultativeHet
- Transcribed
- Promoter
- Enhancer
- RegPermissive
- Bivalent

Project option #1: Epigenome clustering

GM12878 data sets



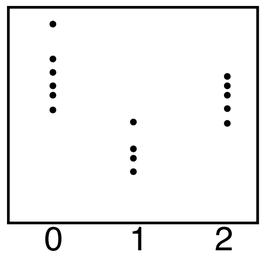
Selected 1% of the genome, binned to 100bp resolution

Your algorithm

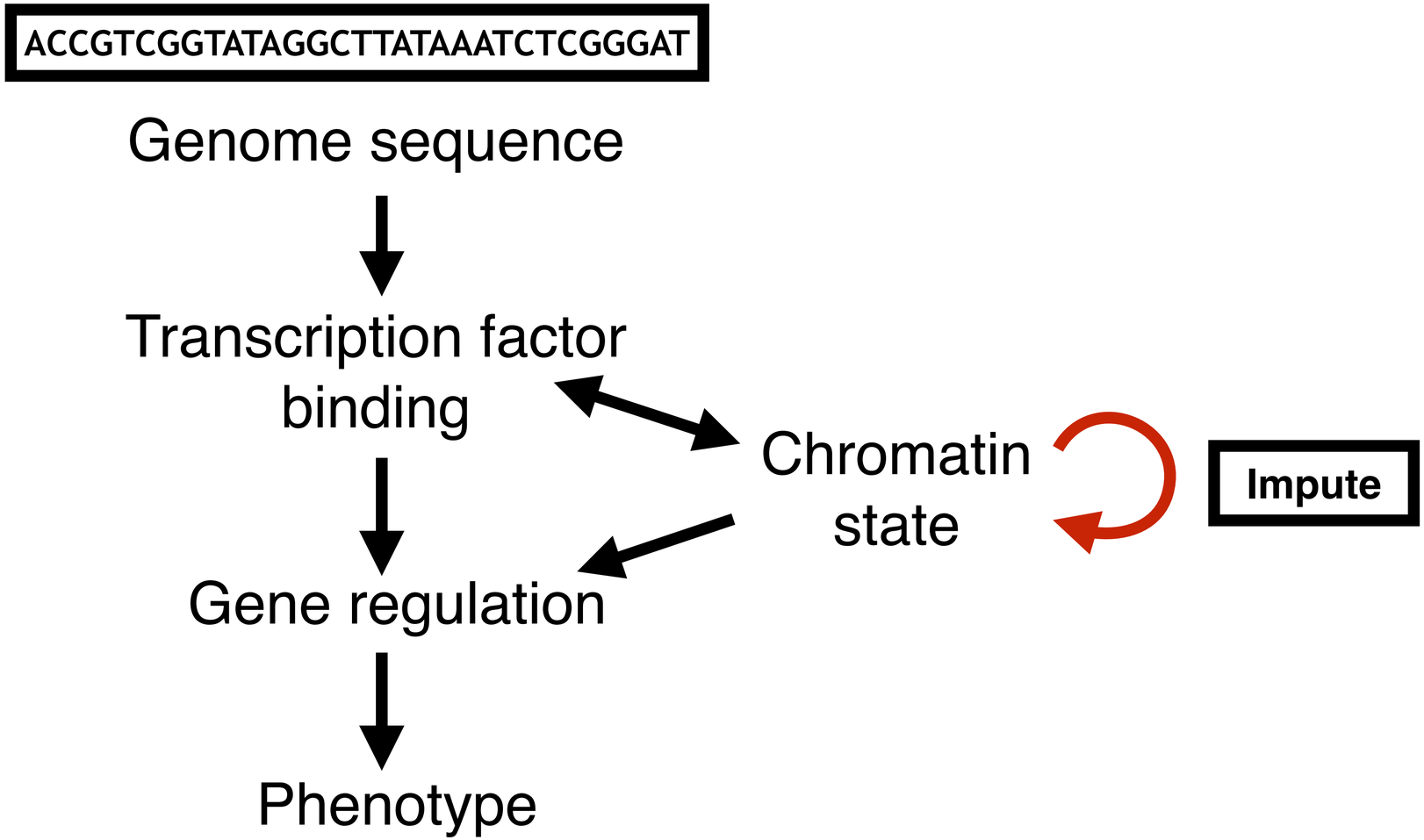
Annotation

Evaluation: Predict gene expression

RNA-seq gene expression



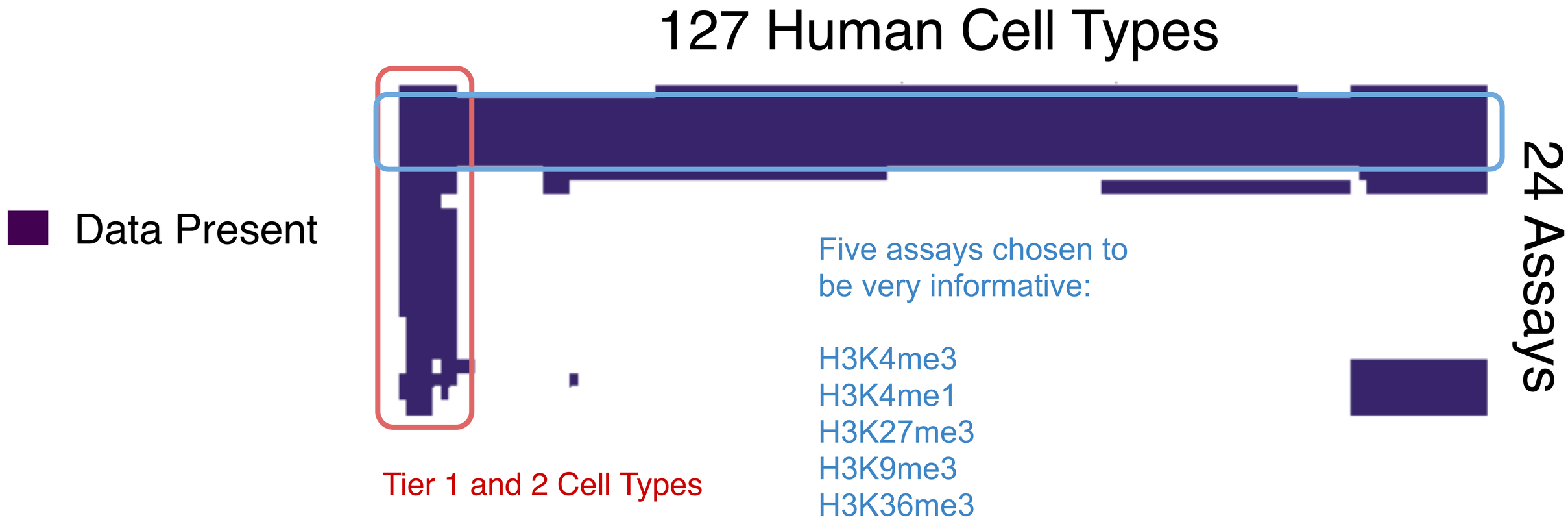
Machine learning methods for the genotype-phenotype relationship, gene regulation and epigenomics



Ernst J, Kellis M. Nature Biotechnology 2015. *Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues*.

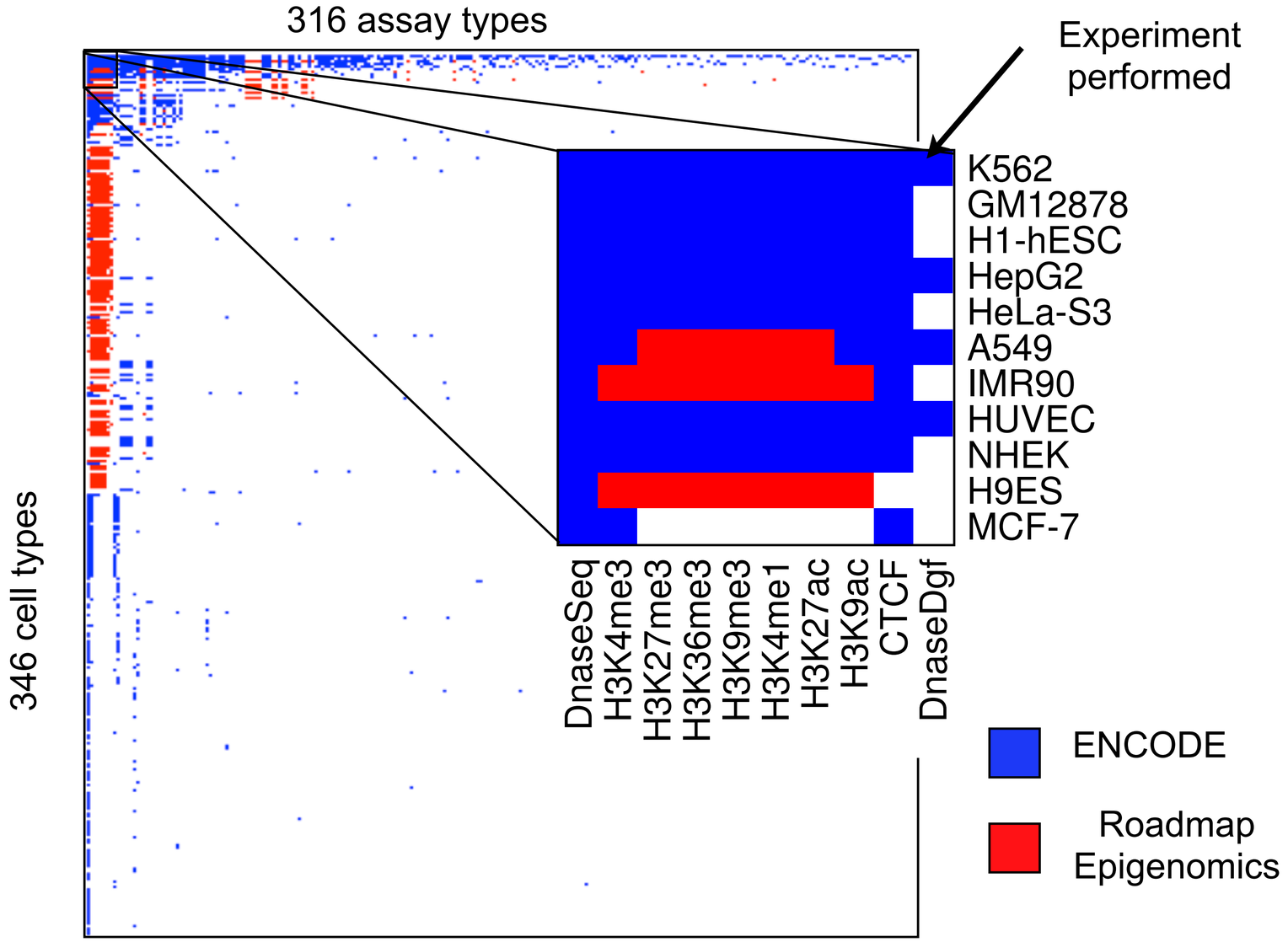
Jacob Schreiber, Timothy Durham, Jeffrey Bilmes & William Stafford Noble. Genome Biology 2020. *Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome*

Many experiments have been performed, but still only a fraction of possible experiments

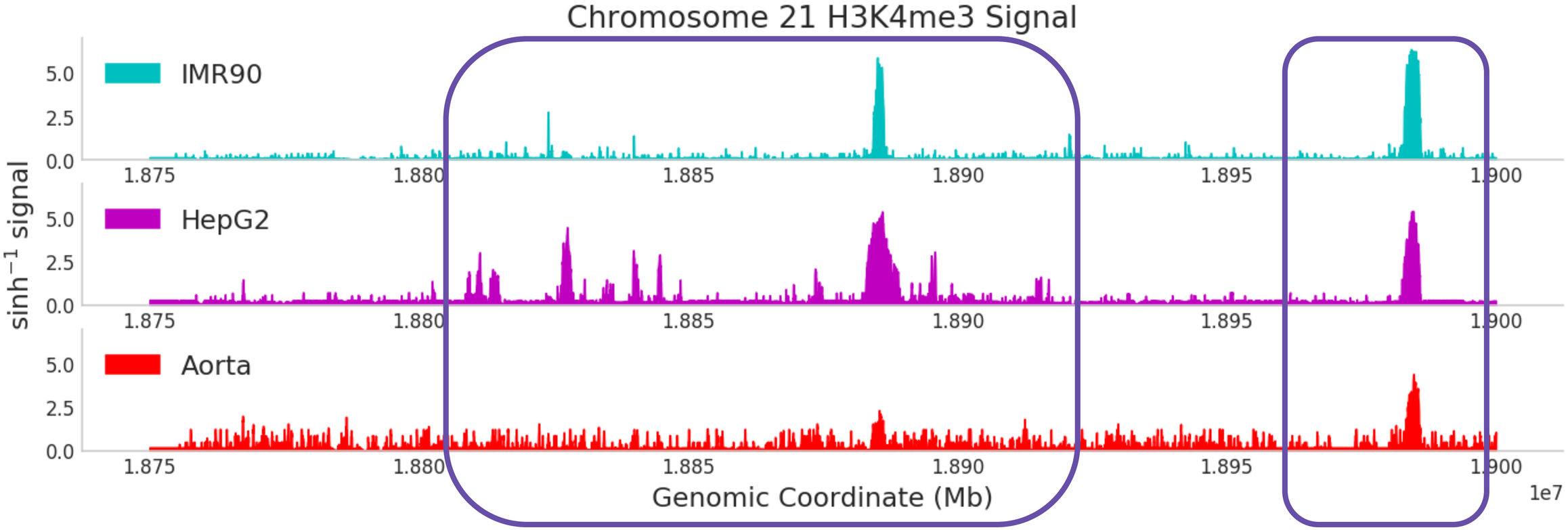


1,014 experiments performed out of a possible 3,048

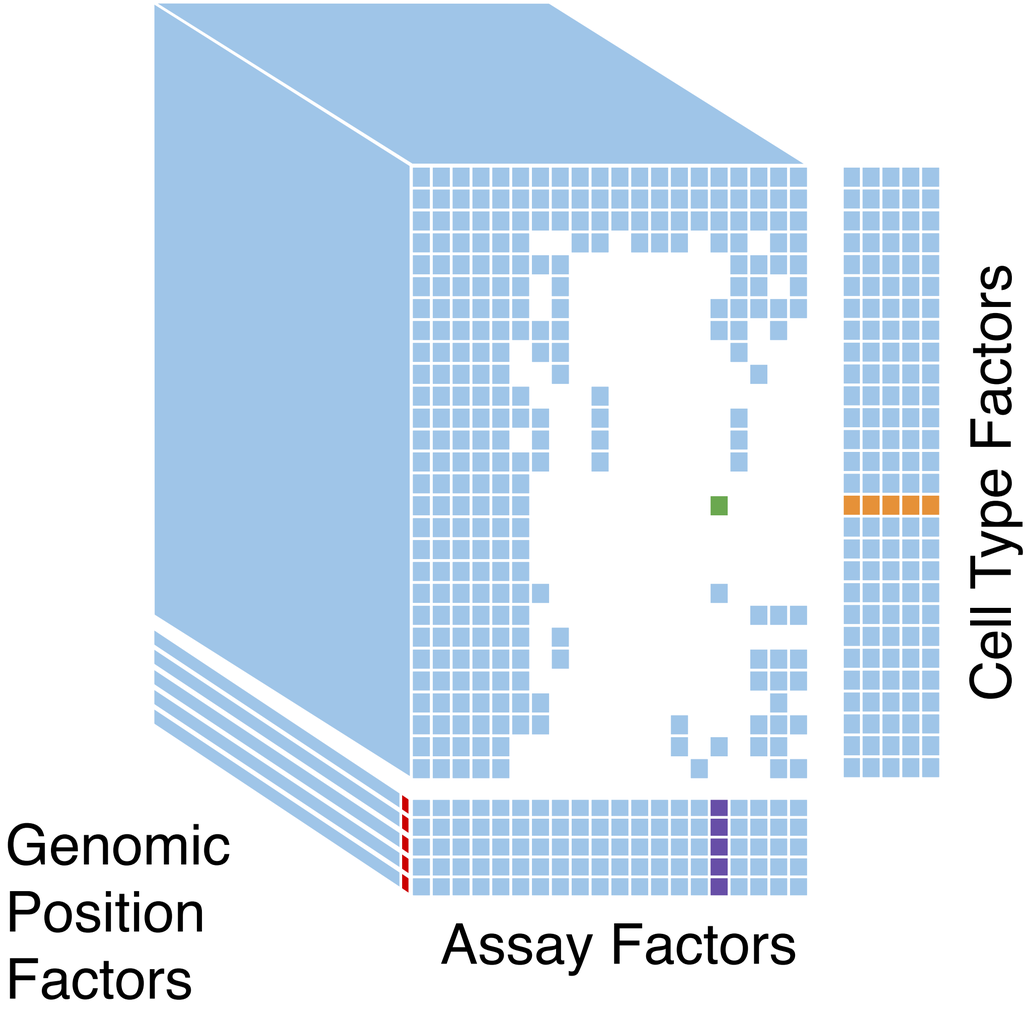
Problem: Can we impute the output of missing experiments?

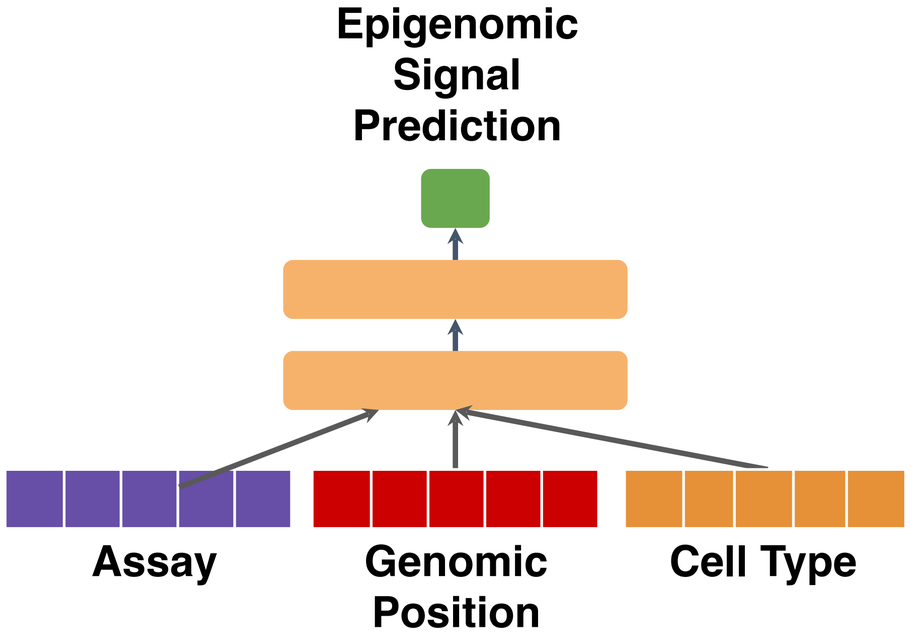


The signal of epigenomic assays vary across cell types

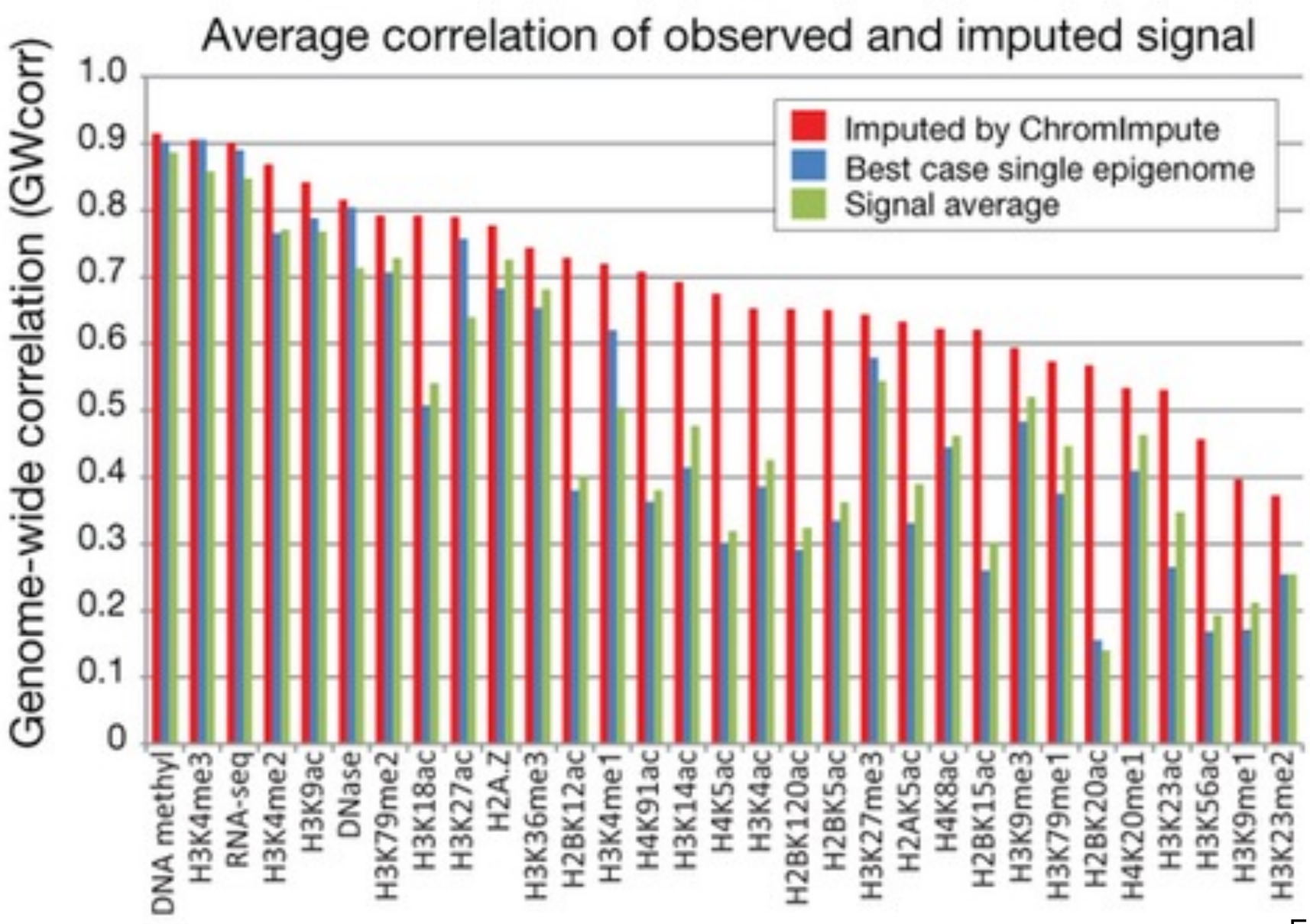


Avocado is a deep tensor factorization approach



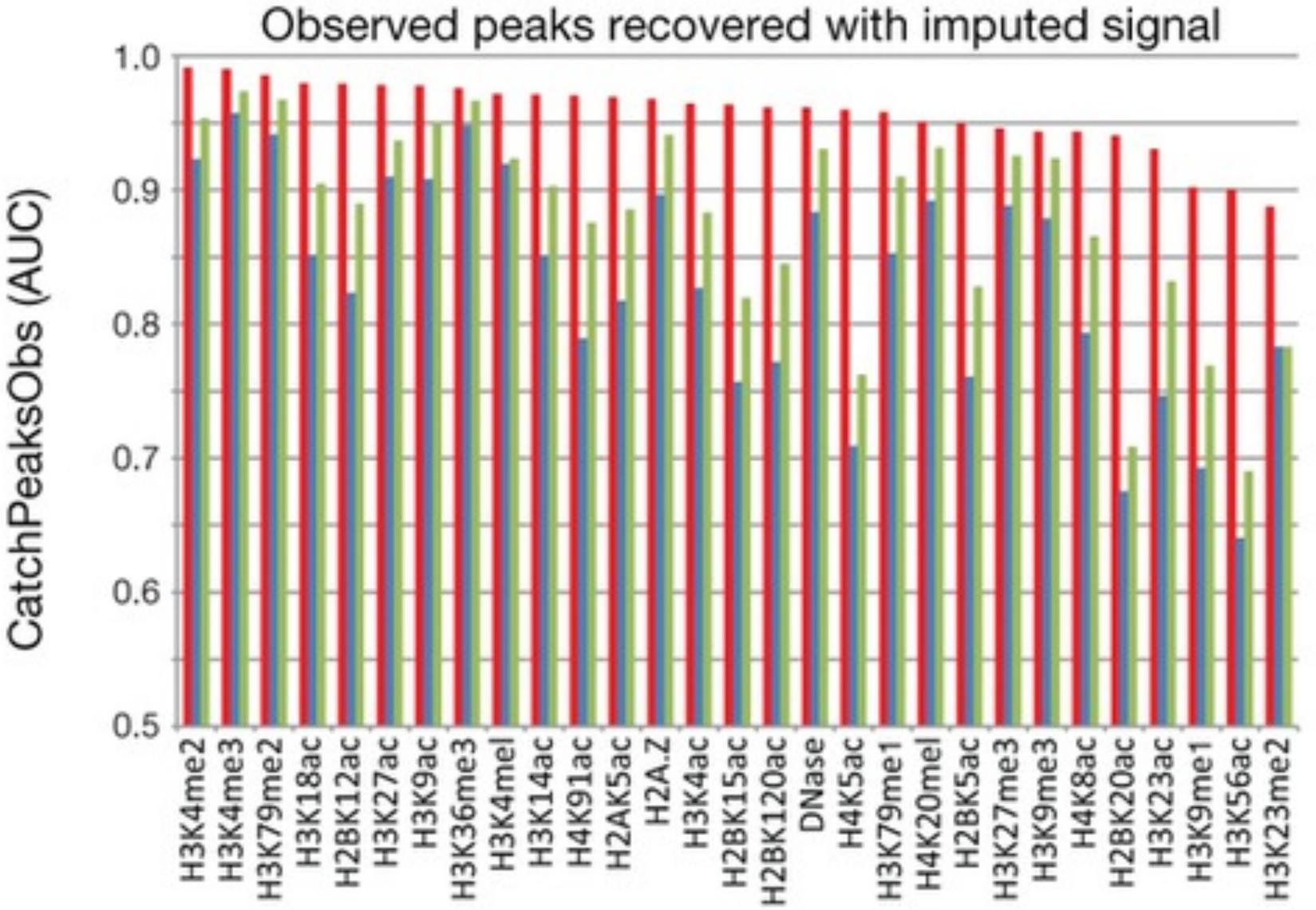


Imputed data has high correlation with observed data



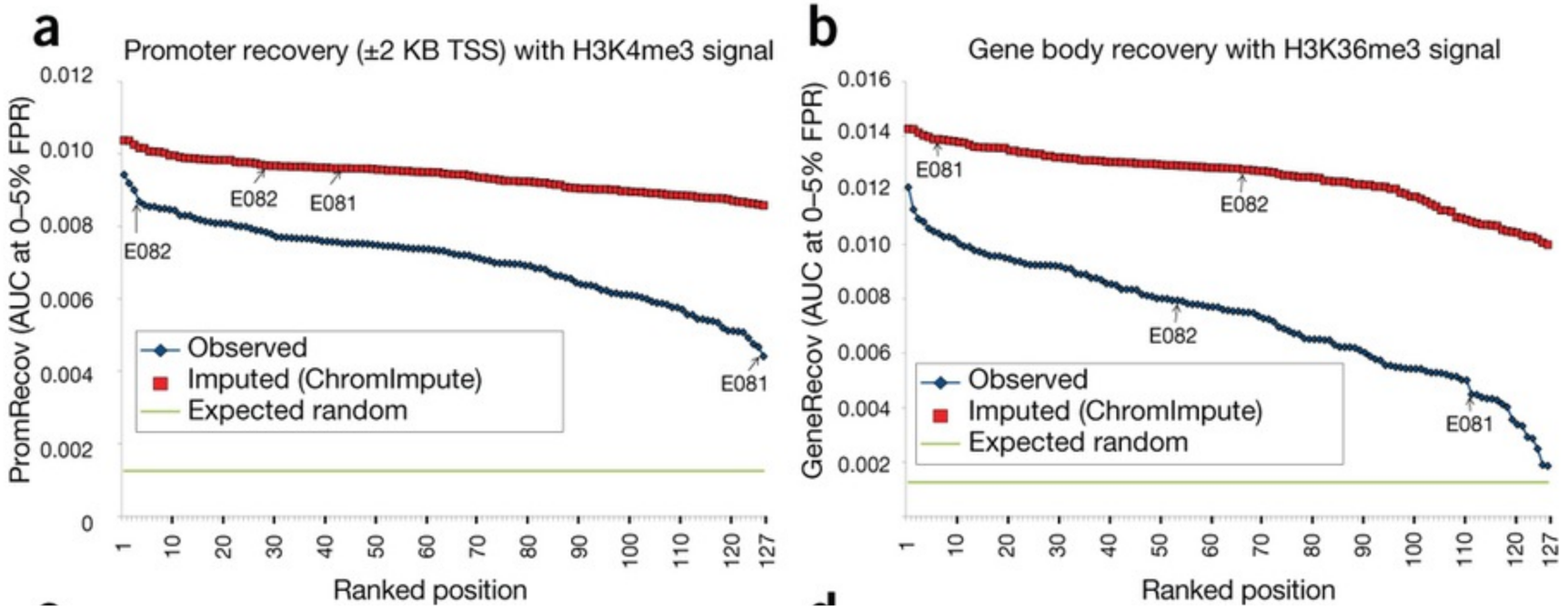
Ernst and Kellis. *Nature* 2015.

Imputed data has high correlation with observed data



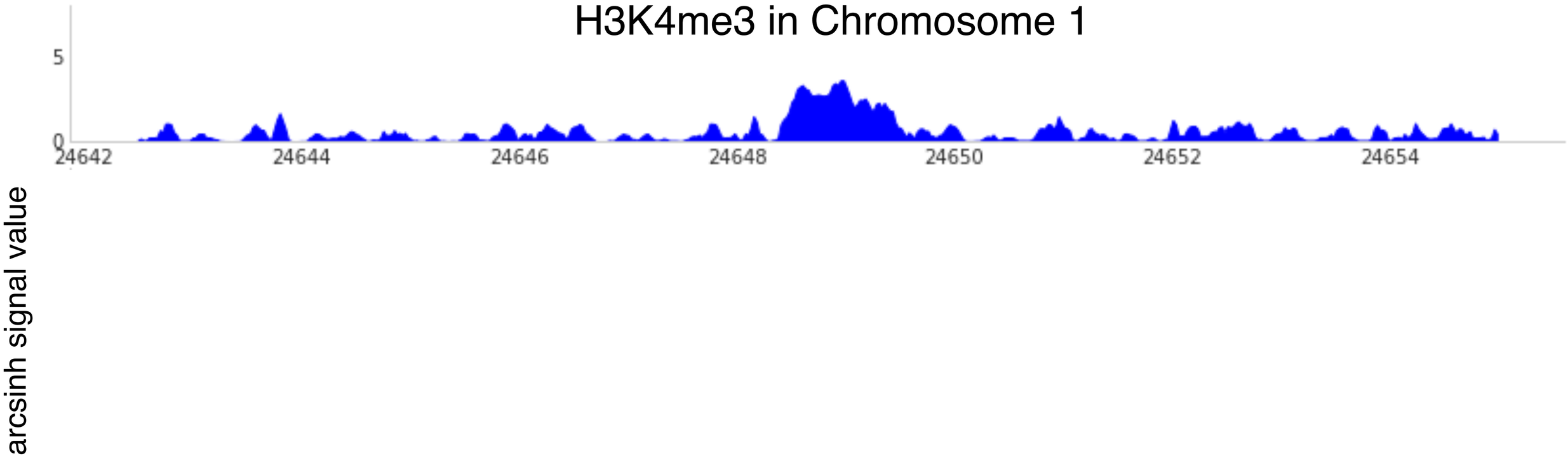
Ernst and Kellis. *Nature* 2015.

Imputed data recovers promoters and TSSs better than observed data



Ernst and Kellis. *Nature* 2015.

Initial inspection of the imputations suggest that Avocado performs well

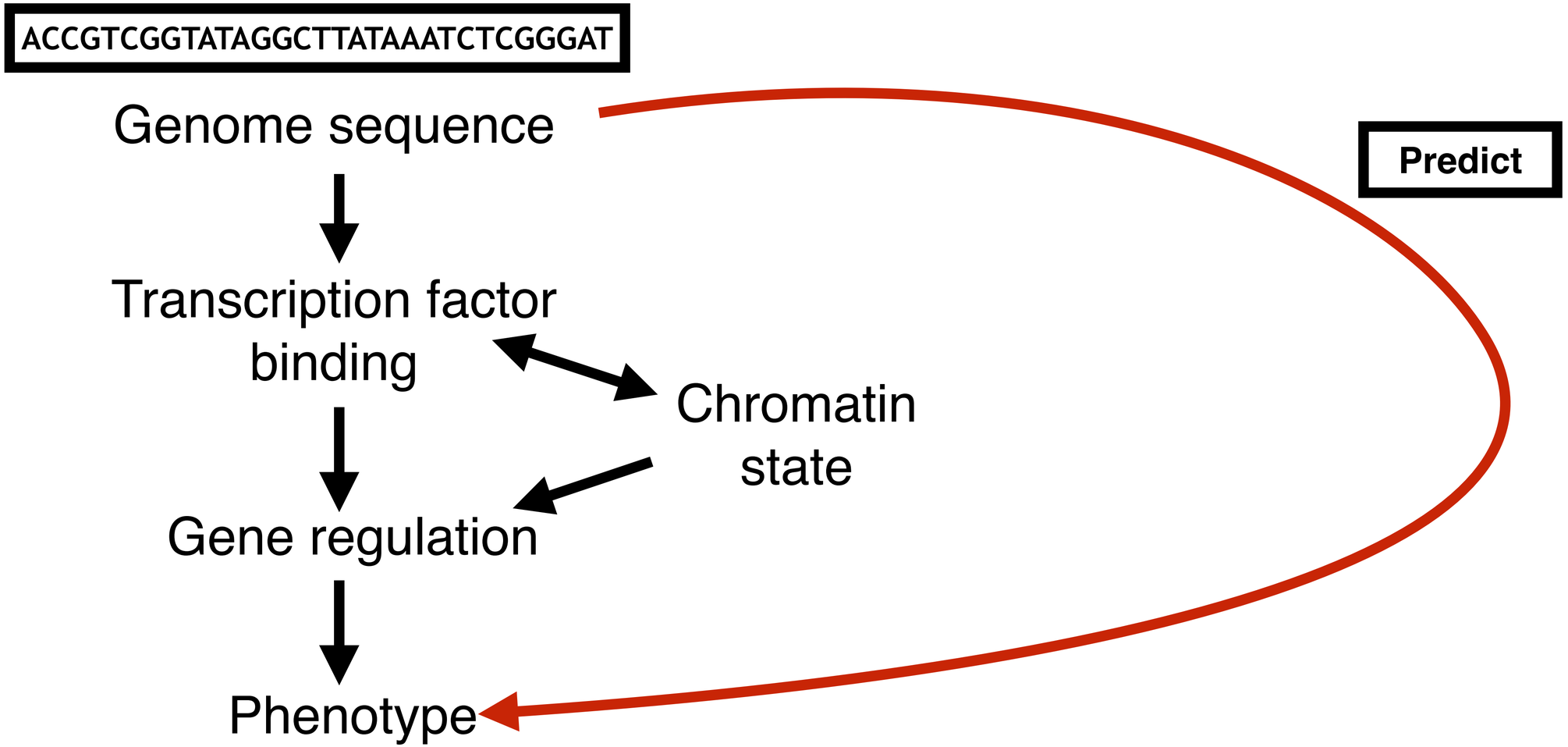


Avocado performs well well genome-wide

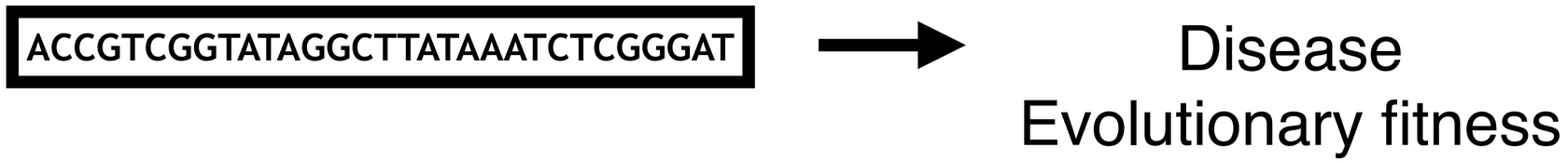
MSE-	global	1obs	1imp	Prom	Gene	Enh
ChromImpute	0.113	**0.941**	1.09	0.3246	0.1494	0.3164
PREDICTD	**0.1**	1.76	0.897	0.2576	**0.1295**	0.267
Avocado	**0.1**	1.66	**0.845**	**0.249**	**0.1295**	**0.26**

MSE-global: Mean squared error (MSE) across the full length of the genome
MSE-1obs: MSE at the top 1% of genomic positions ranked by experimental signal
MSE-1imp: MSE at the top 1% of genomic positions ranked by imputed signal
MSE-Prom: MSE at promoter regions defined by GENCODE
MSE-Gene: MSE at gene bodies defined by GENCODE
MSE-Enh: MSE at enhancer regions defined by FANTOM5

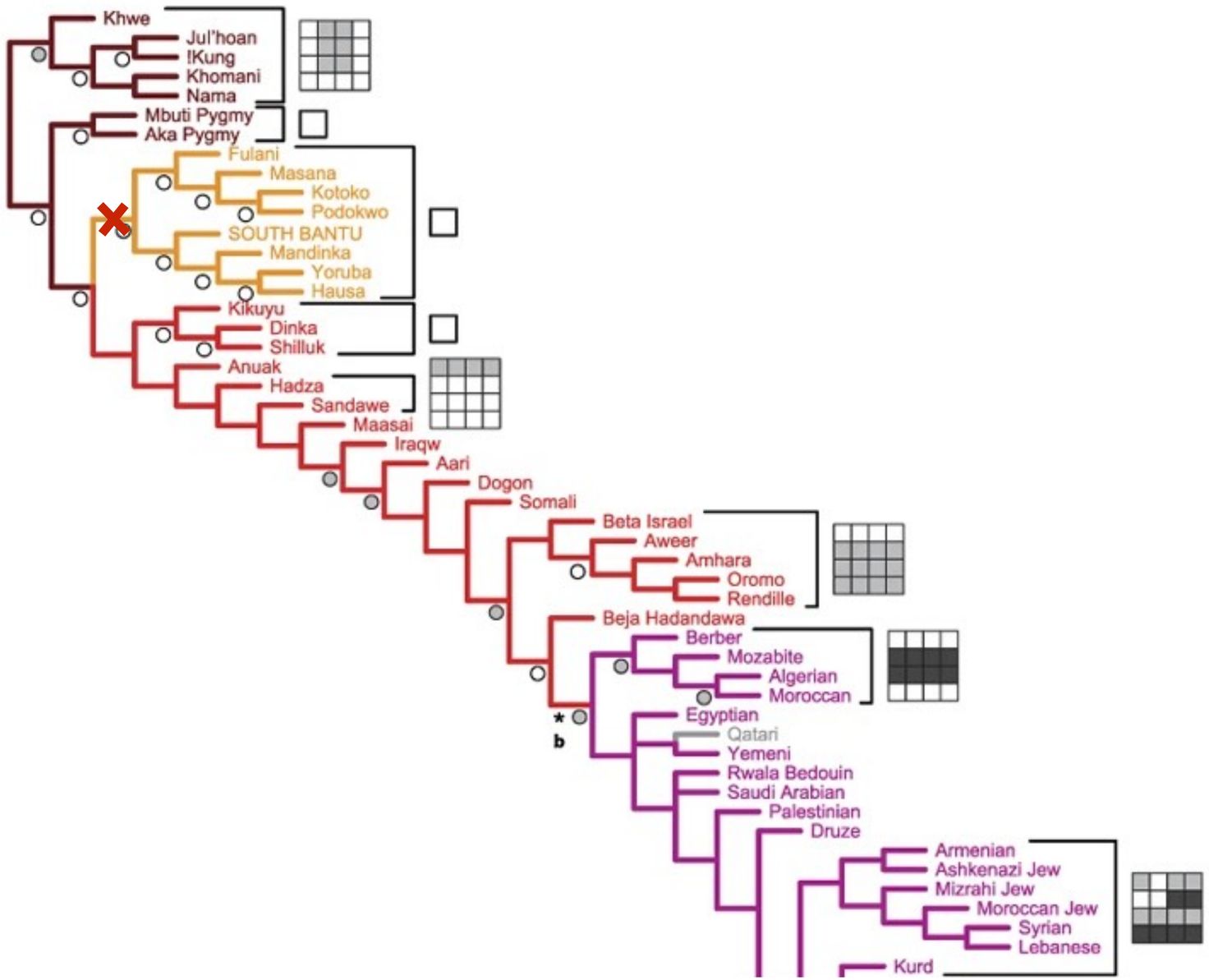
Machine learning methods for the genotype-phenotype relationship, gene regulation and epigenomics



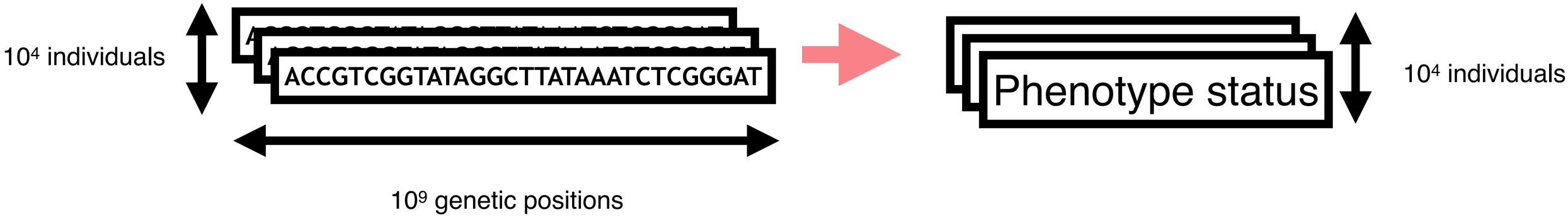
Predicting phenotype from genotype



Genetic variation is driven by phylogeny



There are far more features than labels for predicting phenotype



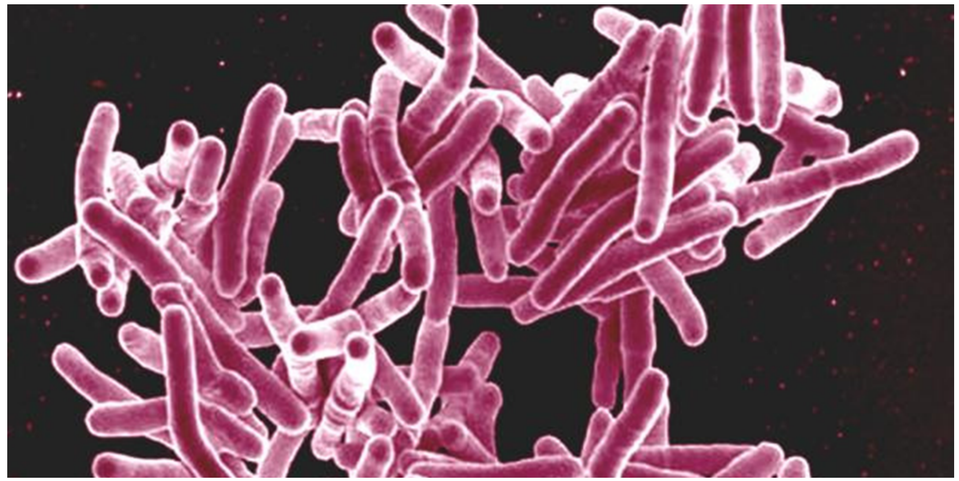
AH Safari, N Sedaghat, H Zabeti, A Forna, L Chindelevitch, M Libbrecht. *Predicting drug resistance in M. tuberculosis using a long-term recurrent convolutional network architecture*. Proceedings of ACM-BCB 2021

Drug resistant tuberculosis is a global health problem

10 million People got infected

1.5 million People died from TB

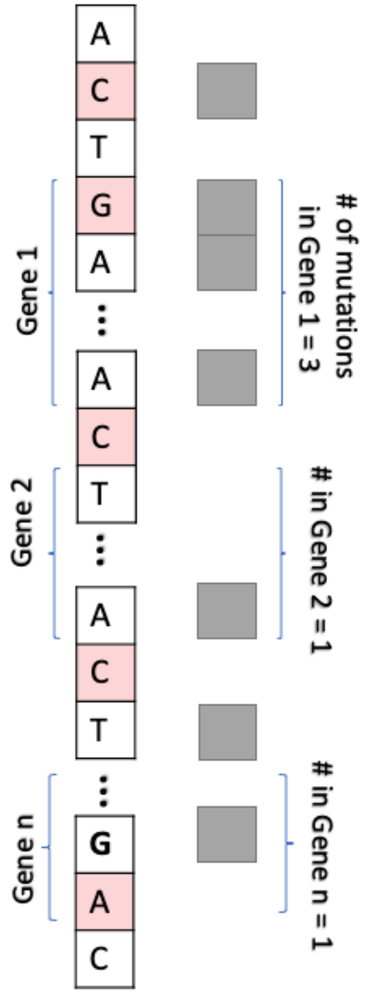
0.5 million New resistance cases



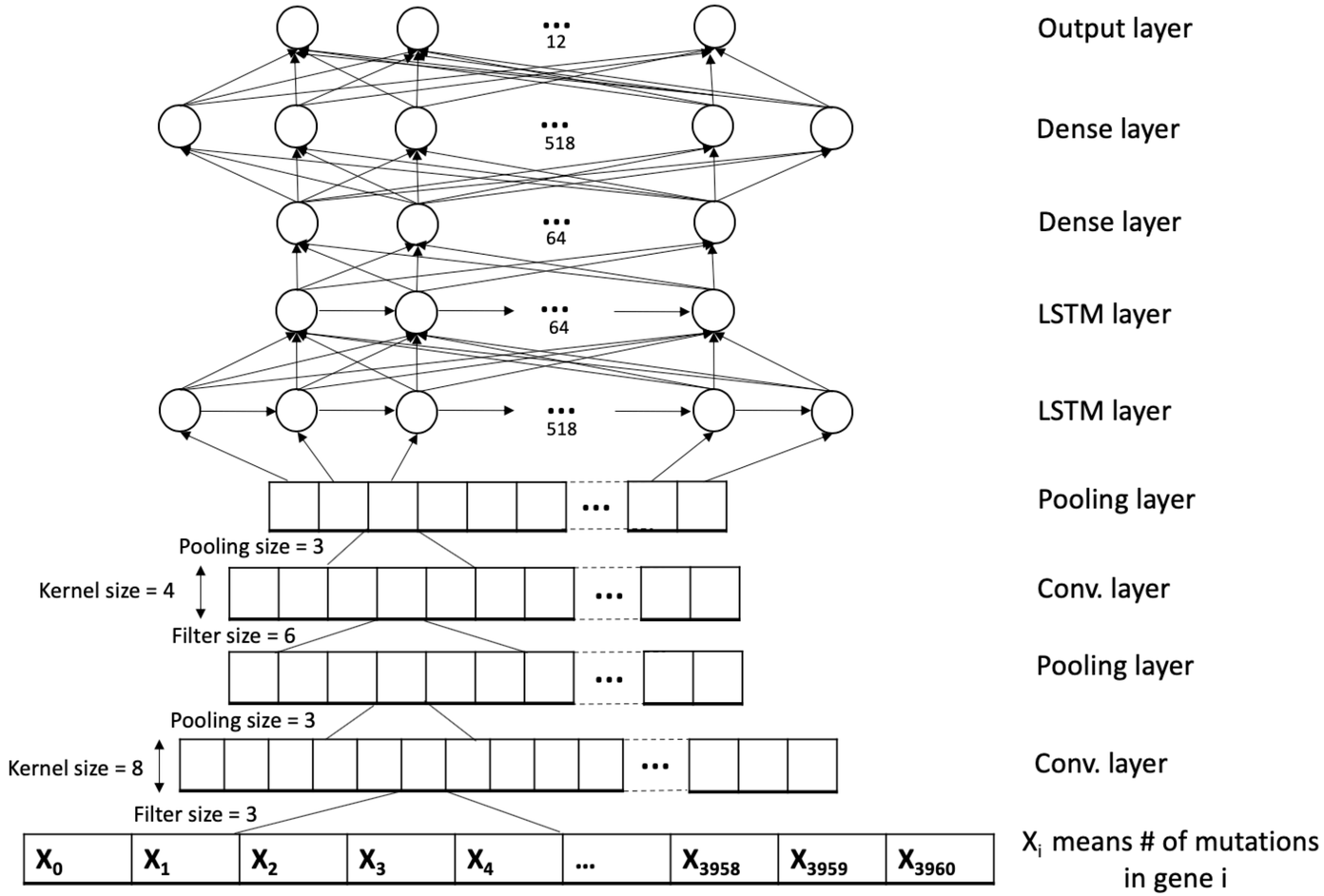
A gene burden-based method for predicting drug resistance in TB.

01. Gene burden-based features

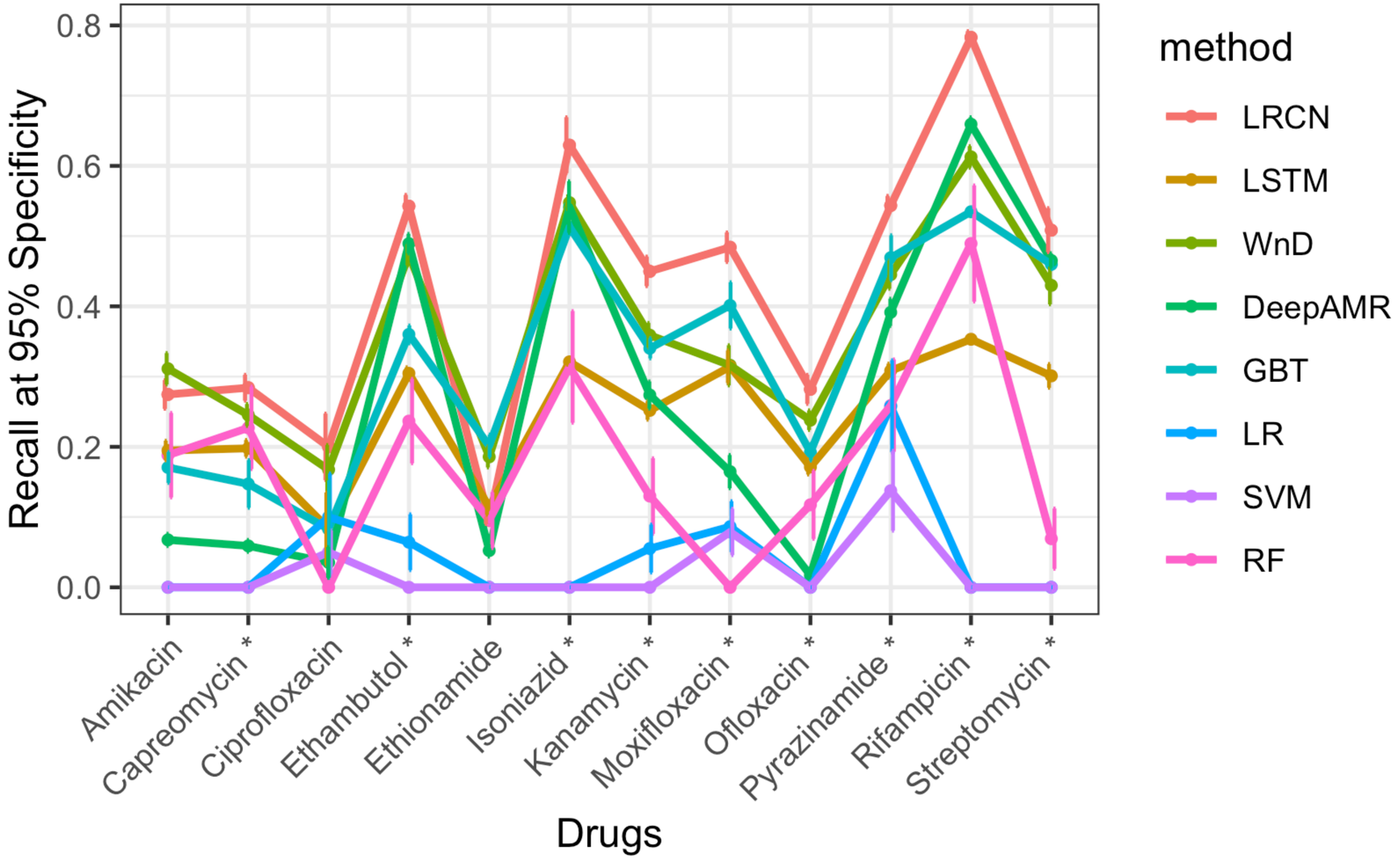
02. Long-term Recurrent Convolutional Network (LRCN)



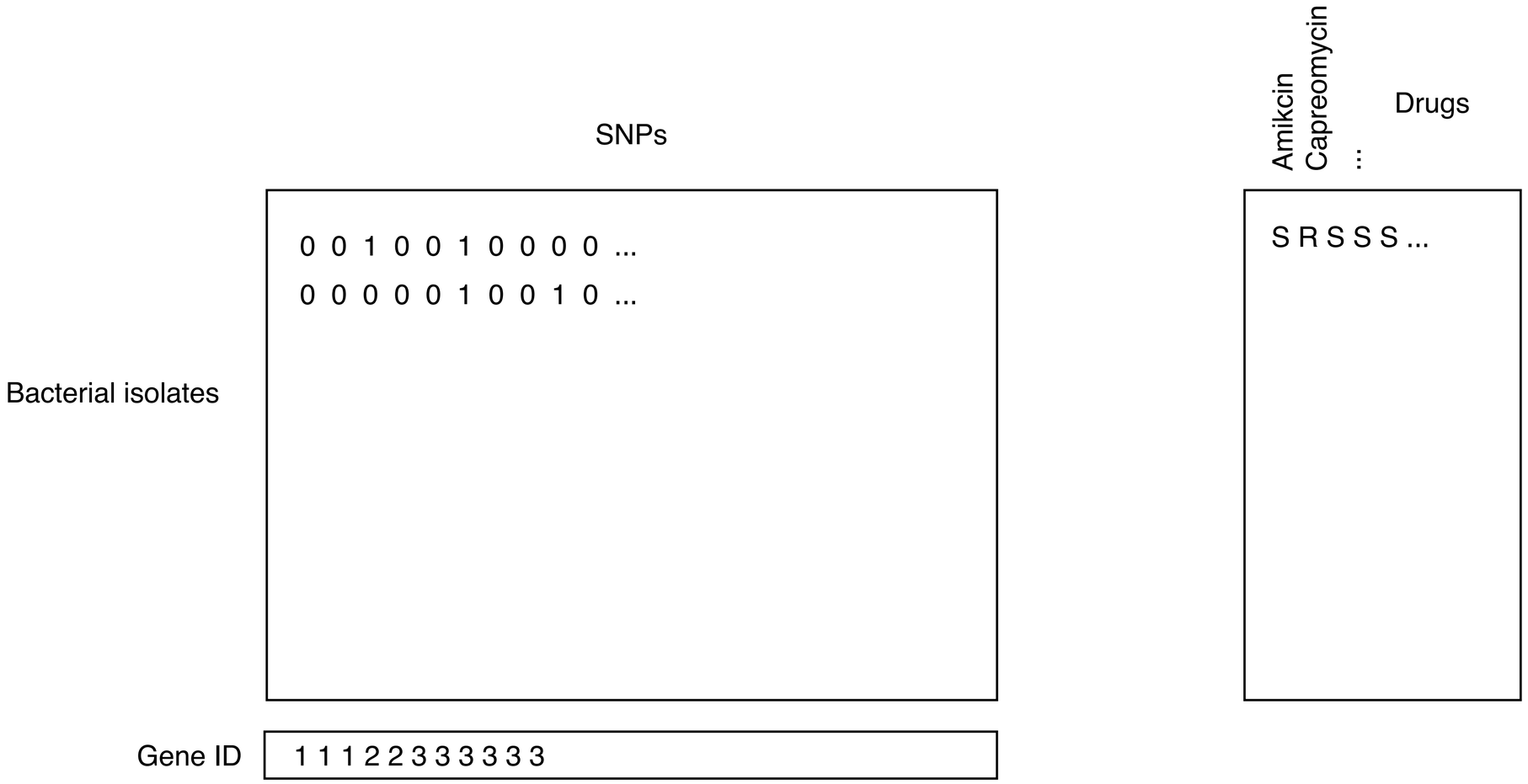
Combing a CNN and LSTM enables the model to take into account local arrangement of genes



An LRCN performs better than alternatives at a clinically-relevant false positive rate



Project option #2: Predict drug resistance in TB



Speaking

- Content
- Visual aids
  - Slides
  - Schematics
  - Data figures
- Delivery

Presentations should have a main thesis

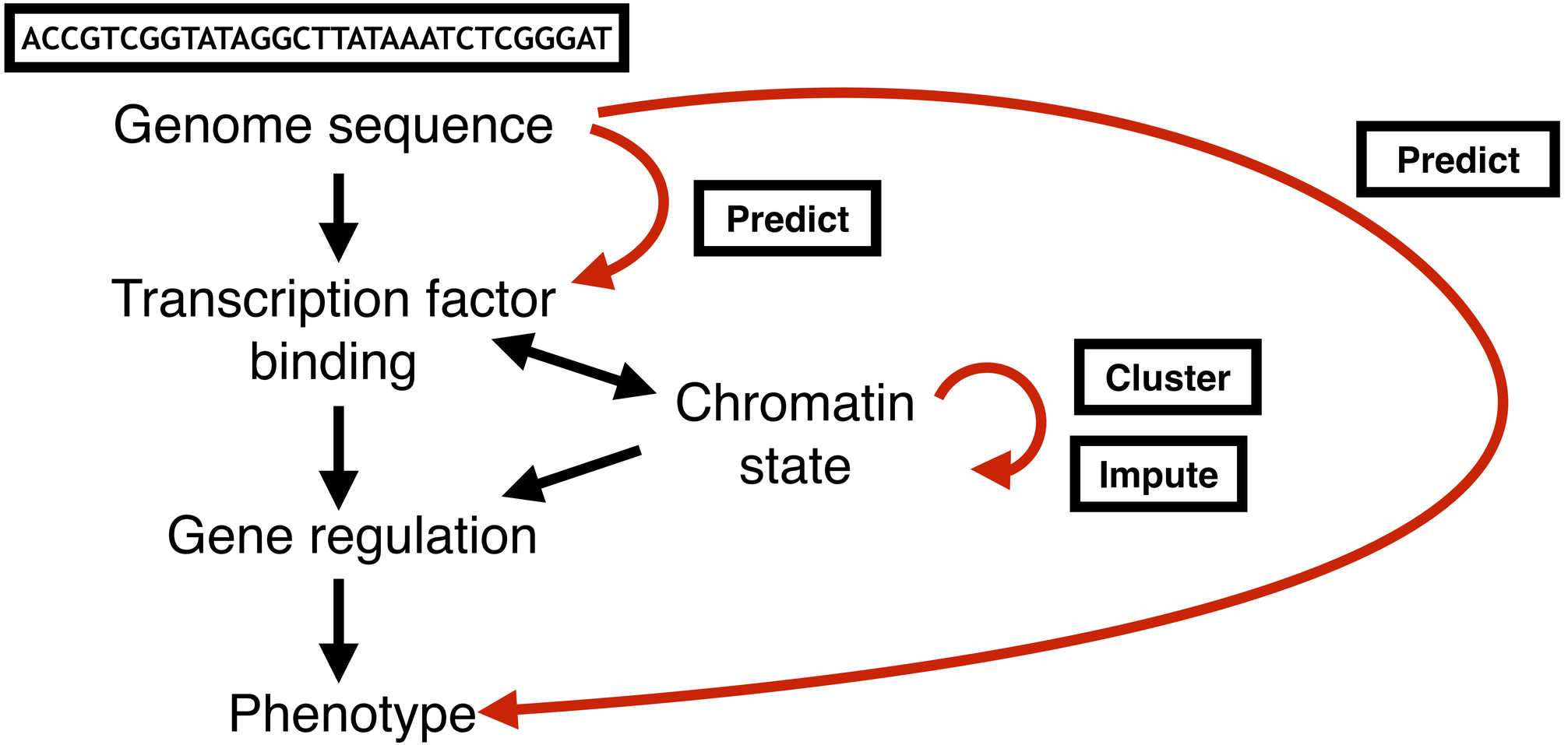
Err on the side of too much background

Re-engage the audience using a home slide

Outline

- Introduction
- Methods
- Results
- Conclusion

Machine learning methods for the genotype-phenotype relationship, gene regulation and epigenomics



Text should be in full sentences

Text should be in full sentences

"Method A crash rate too high"

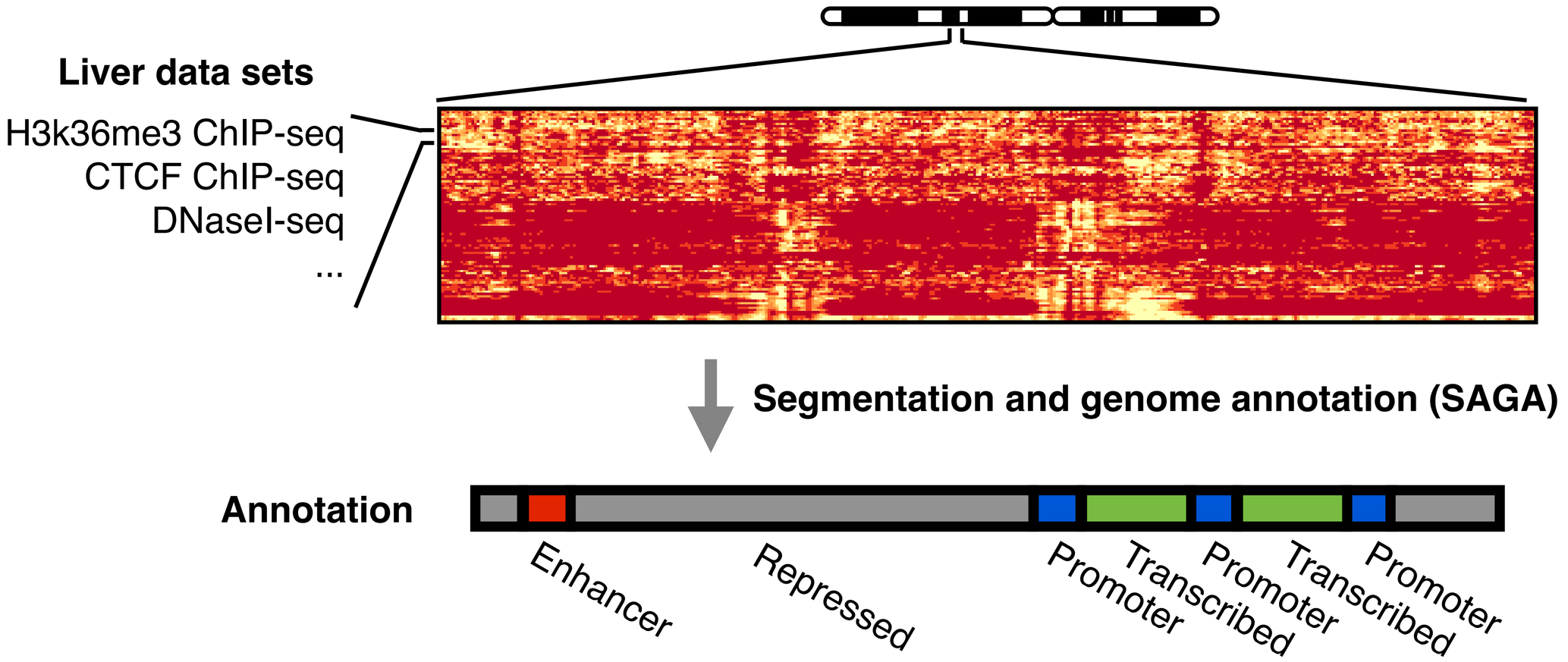
- Slide titles: Use a full sentence explaining the main point of the slide.
- Notation, acronyms: Re-introduce every time you use it.
- Animate in each slide element as you present it.
- Make every slide element legible (font size 18+)

Reduce text by translating it to schematics

Reduce text by translating it to schematics

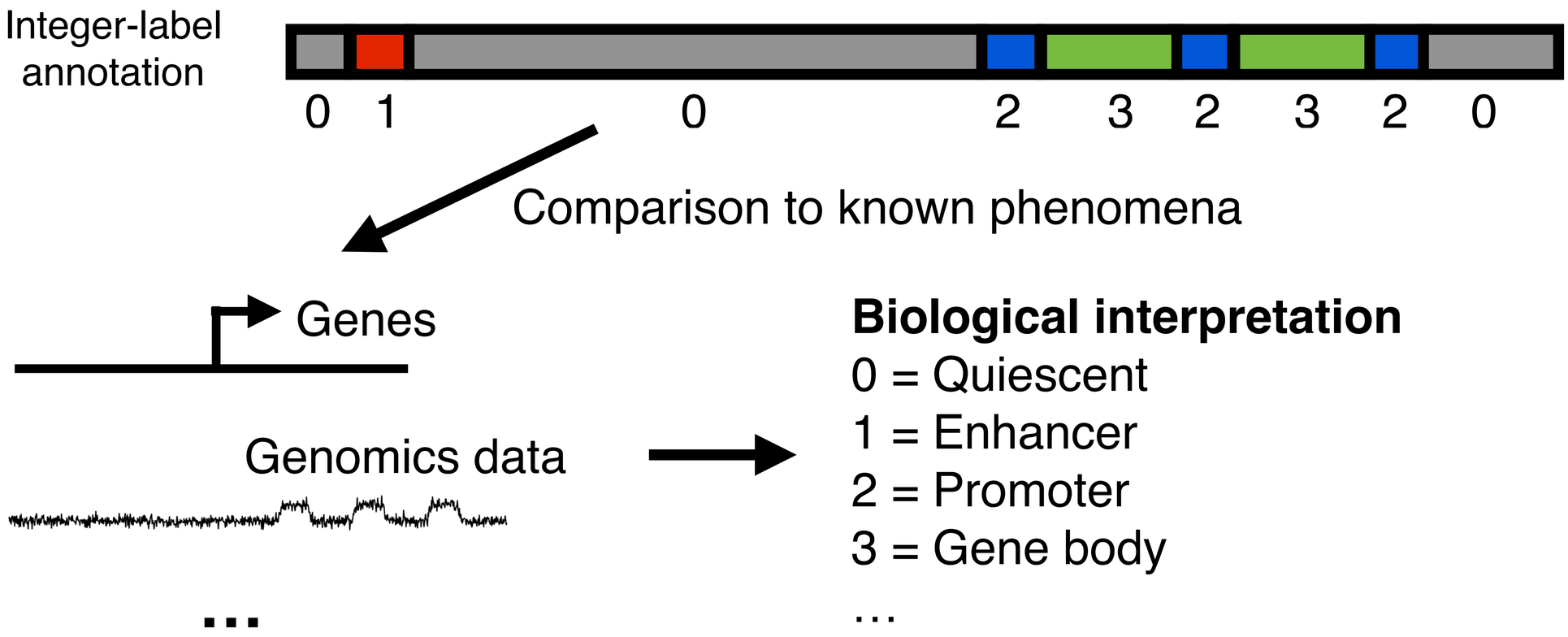
A class of methods known as *semi-automated genome annotation* (SAGA) algorithms are widely used to perform such integrative modeling of diverse genomics data sets. These algorithms take as input a collection of genomics data sets from a particular cell type. They output (1) a set of integer *state labels*, such that each state label putatively corresponds to a type of genomic activity (such as active promoter, active transcription or repressed region), and (2) a partition of the genome and annotation of each genomic segment with one state label. These methods are “semi-automated” because a human performs a functional interpretation of the state labels after the annotation process. In this interpretation step, the human assigns an *interpretation term* to each state label, such as “Promoter” or “Repressed”, indicating its putative function.

Segmentation and genome annotation (SAGA) algorithms partition and label the genome on the basis of genomics data sets



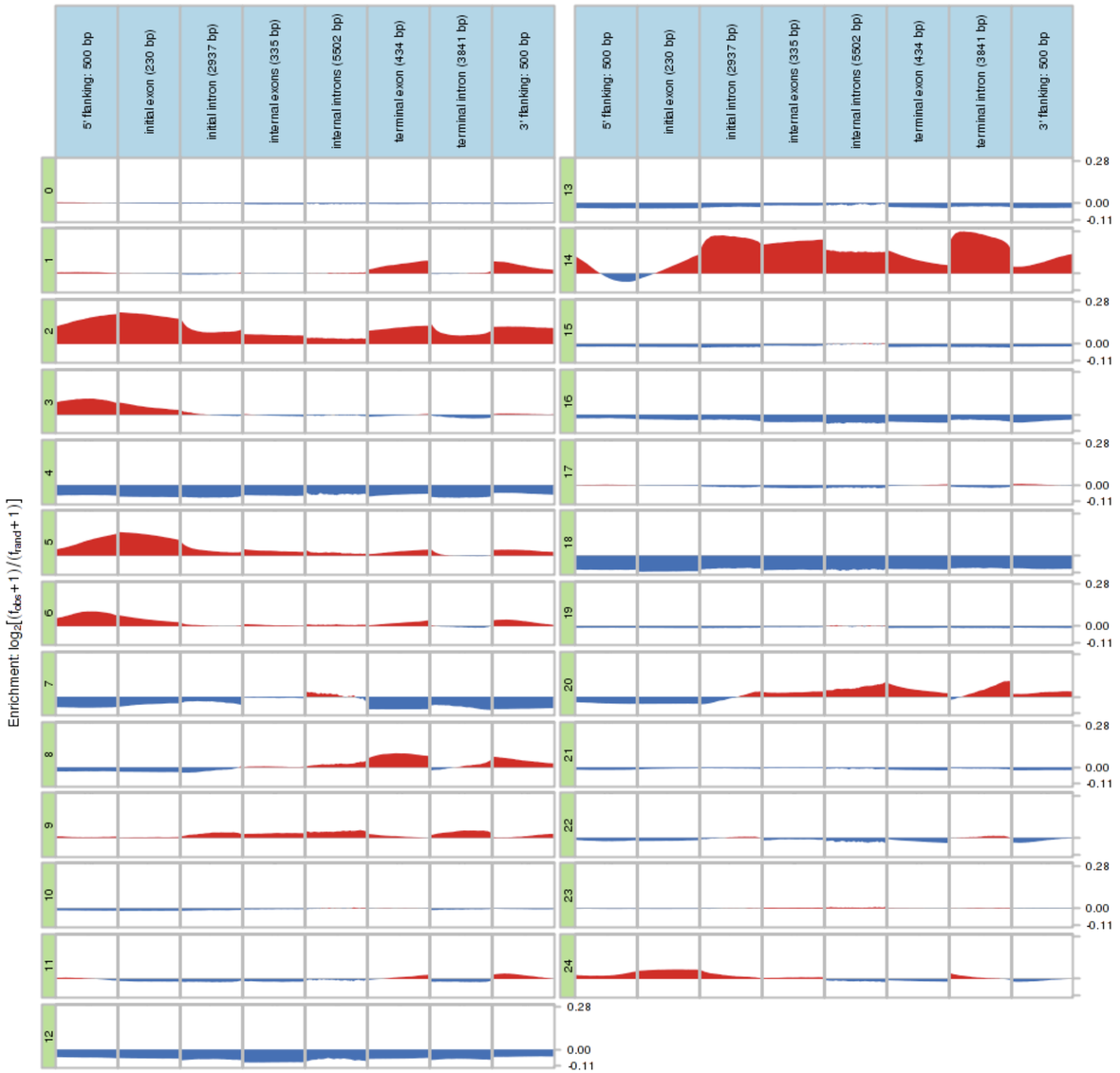
ChromHMM: Ernst, J. and Kellis, M. *Nature Biotechnology*, 2010
Segway: Hoffman, M et al. *Nature Methods*, 2012

What biological phenomenon does each unsupervised label correspond to?

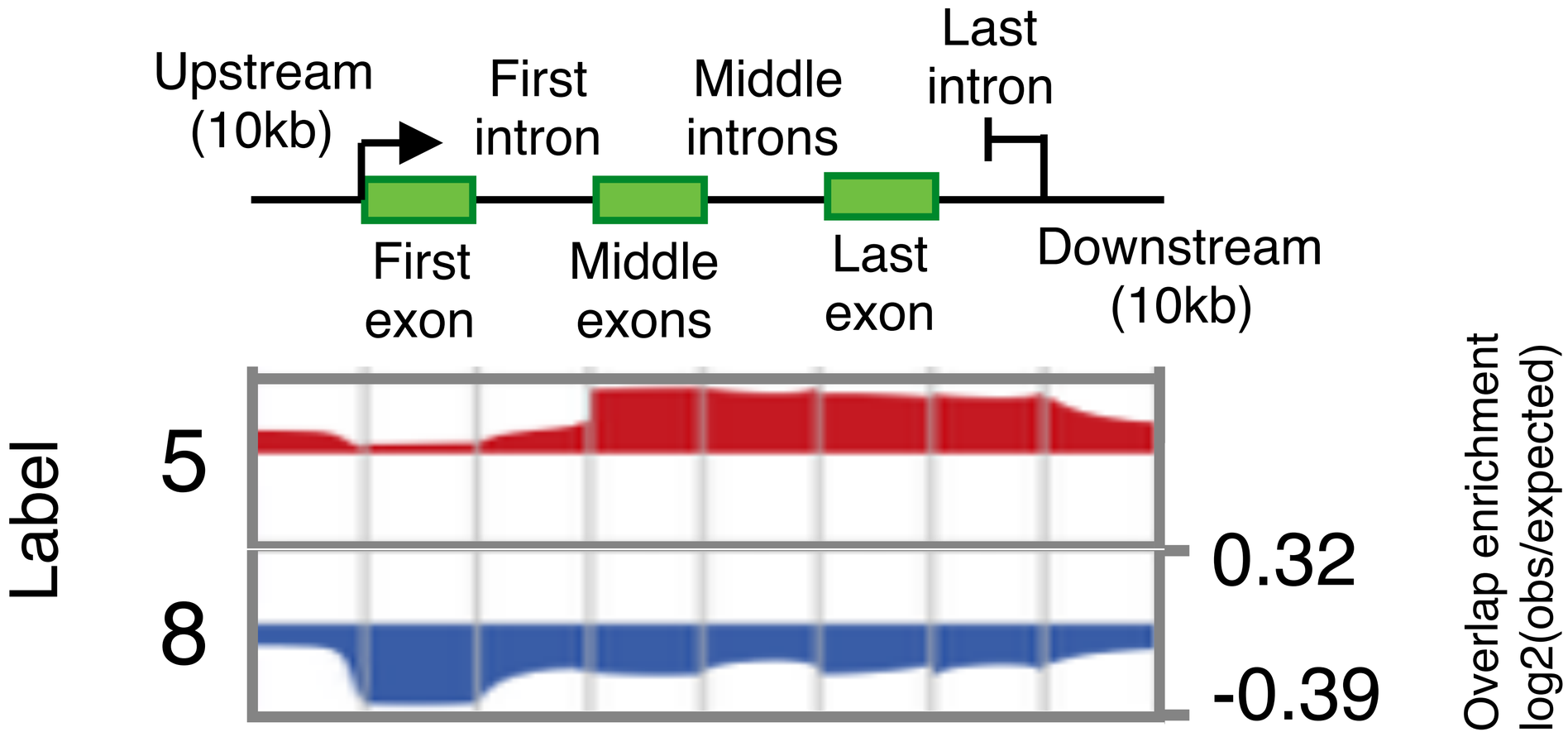


Figures should make a point

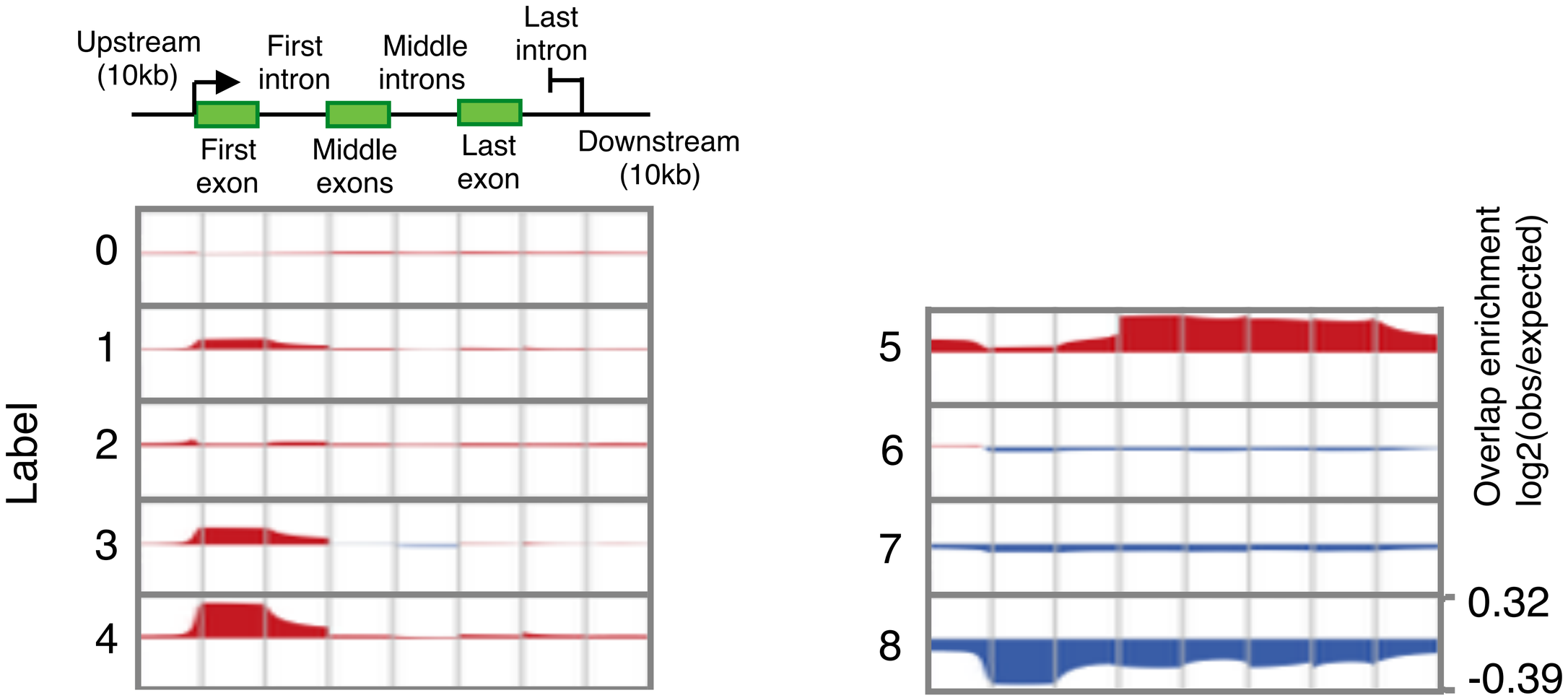
Walk the audience through each figure



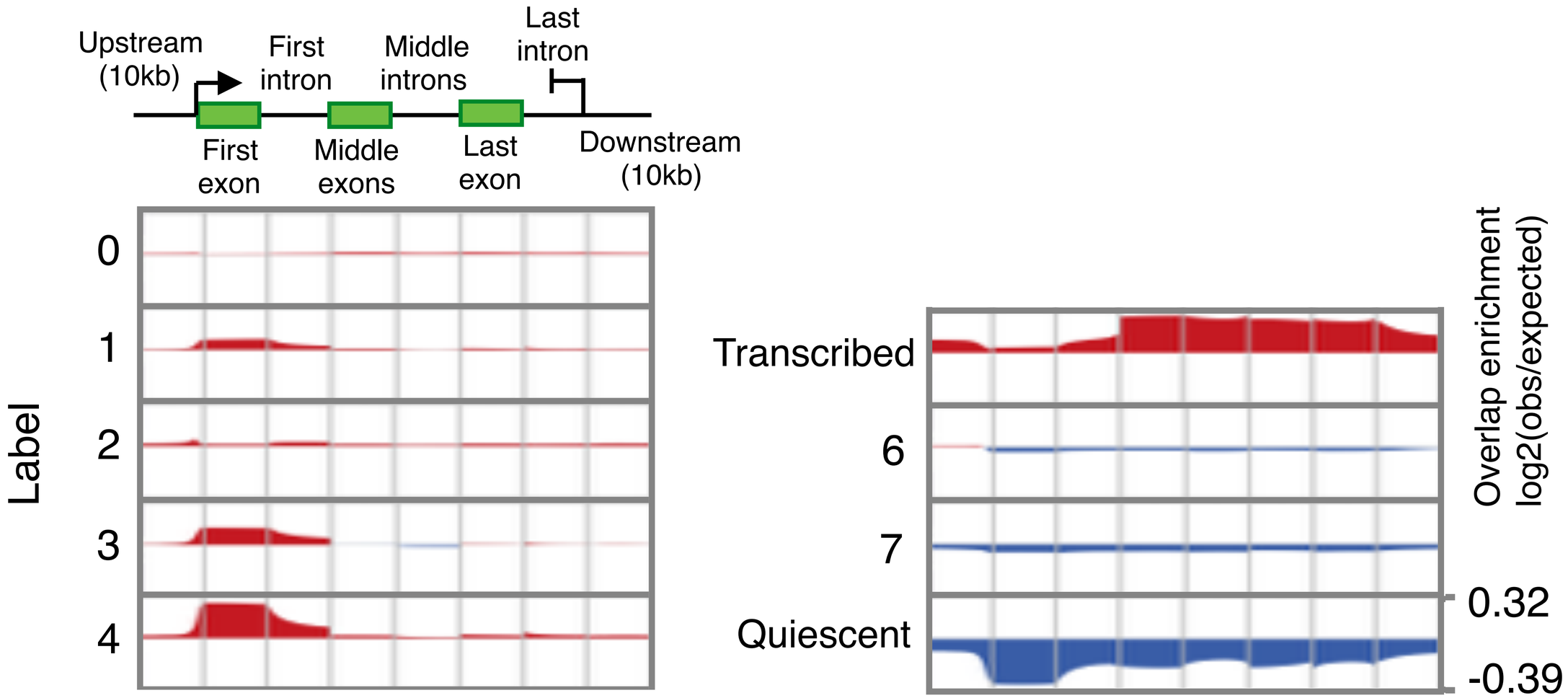
What types of genomic elements did the algorithm find?



What types of genomic elements did the algorithm find?



What types of genomic elements did the algorithm find?



Practice your delivery