Lecture 14. Functional Genomics 3

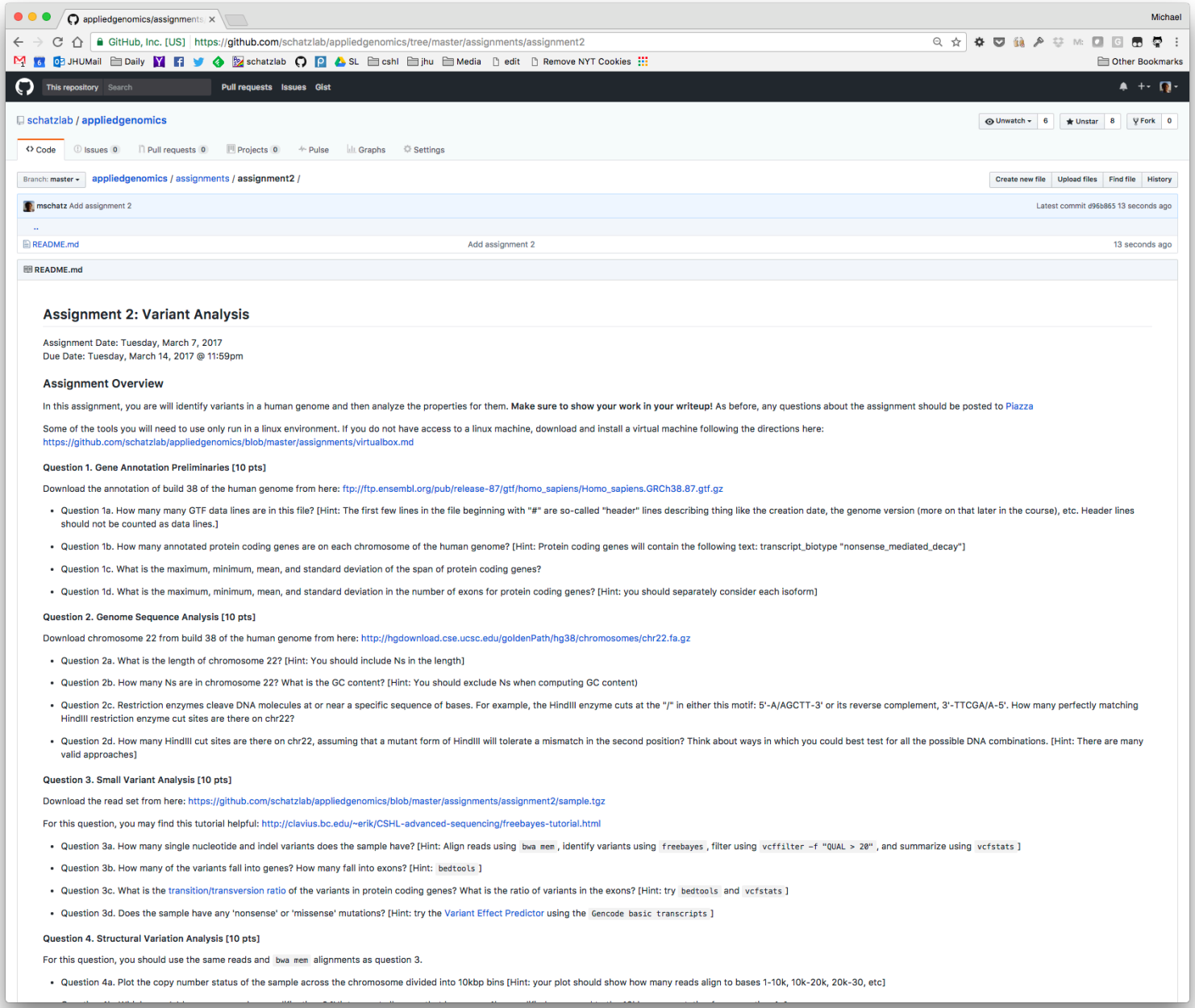
Michael Schatz

March 28, 2017

JHU 600.649: Applied Comparative Genomics

Assignment 2

Due: Thursday March 16 @ 11:59pm



appliedgenomics/assignments

GitHub, Inc. [US] https://github.com/schatzlab/appliedgenomics/tree/master/assignments/assignment2

schatzlab / appliedgenomics

Code | Issues 0 | Pull requests 0 | Projects 0 | Pulse | Graphs | Settings

Branch: master appliedgenomics / assignments / assignment2 /

mschatz Add assignment 2 — Latest commit d96b865 13 seconds ago

README.md — Add assignment 2 — 13 seconds ago

README.md

Assignment 2: Variant Analysis

Assignment Date: Tuesday, March 7, 2017
Due Date: Tuesday, March 14, 2017 @ 11:59pm

Assignment Overview

In this assignment, you are will identify variants in a human genome and then analyze the properties for them. **Make sure to show your work in your writeup!** As before, any questions about the assignment should be posted to Piazza

Some of the tools you will need to use only run in a linux environment. If you do not have access to a linux machine, download and install a virtual machine following the directions here:
https://github.com/schatzlab/appliedgenomics/blob/master/assignments/virtualbox.md

Question 1. Gene Annotation Preliminaries [10 pts]

Download the annotation of build 38 of the human genome from here: ftp://ftp.ensembl.org/pub/release-87/gtf/homo_sapiens/Homo_sapiens.GRCh38.87.gtf.gz

- Question 1a. How many many GTF data lines are in this file? [Hint: The first few lines in the file beginning with "#" are so-called "header" lines describing thing like the creation date, the genome version (more on that later in the course), etc. Header lines should not be counted as data lines.]
- Question 1b. How many annotated protein coding genes are on each chromosome of the human genome? [Hint: Protein coding genes will contain the following text: transcript_biotype "nonsense_mediated_decay"]
- Question 1c. What is the maximum, minimum, mean, and standard deviation of the span of protein coding genes?
- Question 1d. What is the maximum, minimum, mean, and standard deviation in the number of exons for protein coding genes? [Hint: you should separately consider each isoform]

Question 2. Genome Sequence Analysis [10 pts]

Download chromosome 22 from build 38 of the human genome from here: http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/chr22.fa.gz

- Question 2a. What is the length of chromosome 22? [Hint: You should include Ns in the length]
- Question 2b. How many Ns are in chromosome 22? What is the GC content? [Hint: You should exclude Ns when computing GC content)
- Question 2c. Restriction enzymes cleave DNA molecules at or near a specific sequence of bases. For example, the HindIII enzyme cuts at the "/" in either this motif: 5'-A/AGCTT-3' or its reverse complement, 3'-TTCGA/A-5'. How many perfectly matching HindIII restriction enzyme cut sites are there on chr22?
- Question 2d. How many HindIII cut sites are there on chr22, assuming that a mutant form of HindIII will tolerate a mismatch in the second position? Think about ways in which you could best test for all the possible DNA combinations. [Hint: There are many valid approaches]

Question 3. Small Variant Analysis [10 pts]

Download the read set from here: https://github.com/schatzlab/appliedgenomics/blob/master/assignments/assignment2/sample.tgz

For this question, you may find this tutorial helpful: http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html

- Question 3a. How many single nucleotide and indel variants does the sample have? [Hint: Align reads using `bwa mem`, identify variants using `freebayes`, filter using `vcffilter -f "QUAL > 20"`, and summarize using `vcfstats`]
- Question 3b. How many of the variants fall into genes? How many fall into exons? [Hint: `bedtools`]
- Question 3c. What is the transition/transversion ratio of the variants in protein coding genes? What is the ratio of variants in the exons? [Hint: try `bedtools` and `vcfstats`]
- Question 3d. Does the sample have any 'nonsense' or 'missense' mutations? [Hint: try the Variant Effect Predictor using the `Gencode basic transcripts`]

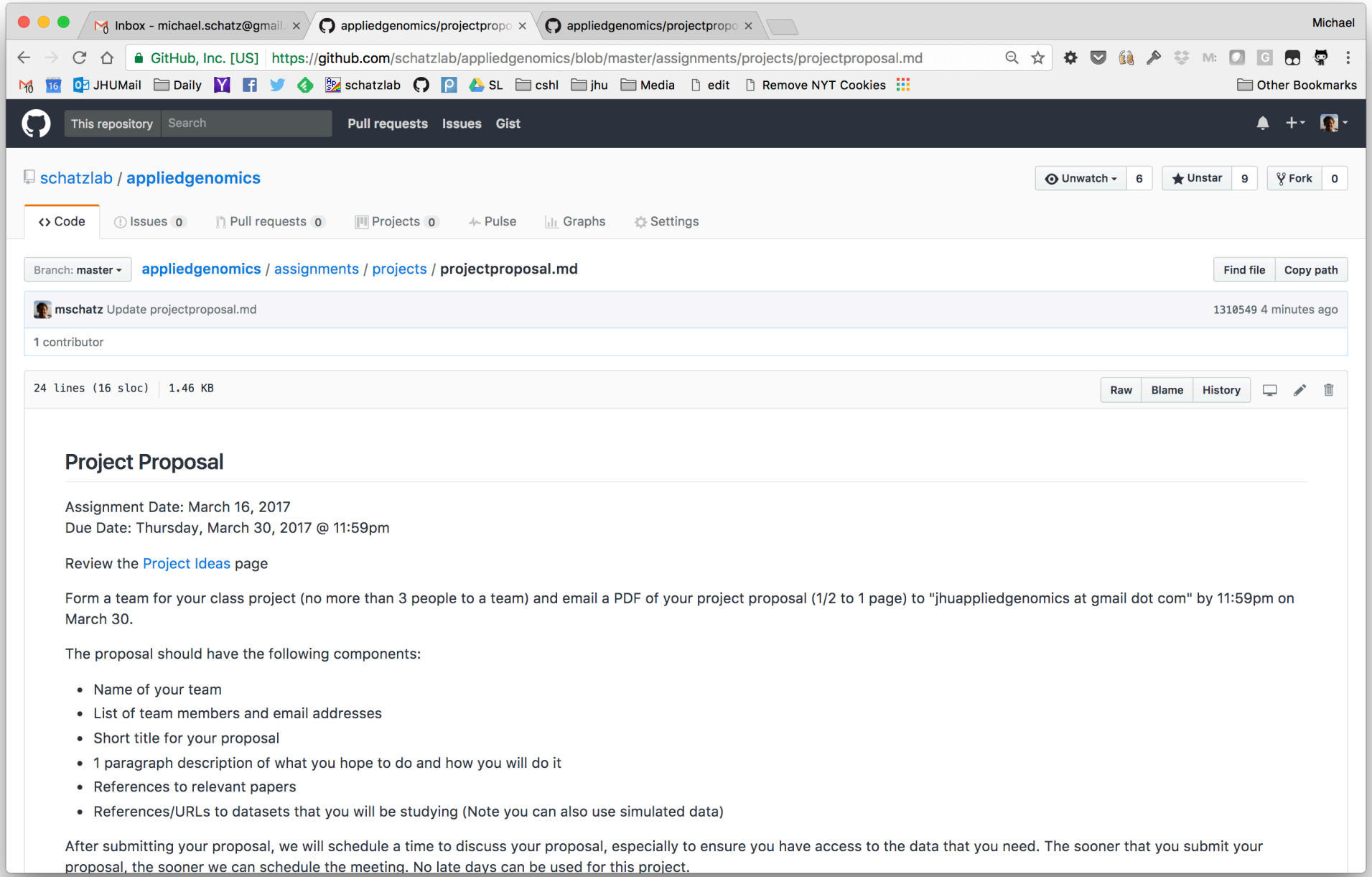
Question 4. Structural Variation Analysis [10 pts]

For this question, you should use the same reads and `bwa mem` alignments as question 3.

- Question 4a. Plot the copy number status of the sample across the chromosome divided into 10kbp bins [Hint: your plot should show how many reads align to bases 1-10k, 10k-20k, 20k-30, etc]

Project Proposal

Due: Thursday March 30 @ 11:59pm



Inbox - michael.schatz@gmail. × | appliedgenomics/projectpropo × | appliedgenomics/projectpropo × | Michael

GitHub, Inc. [US] https://github.com/schatzlab/appliedgenomics/blob/master/assignments/projects/projectproposal.md

JHUMail | Daily | schatzlab | SL | cshl | jhu | Media | edit | Remove NYT Cookies | Other Bookmarks

This repository | Search | Pull requests | Issues | Gist

schatzlab / appliedgenomics

Unwatch 6 | Unstar 9 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Pulse | Graphs | Settings

Branch: master | appliedgenomics / assignments / projects / projectproposal.md | Find file | Copy path

mschatz Update projectproposal.md — 1310549 4 minutes ago

1 contributor

24 lines (16 sloc) | 1.46 KB | Raw | Blame | History

Project Proposal

Assignment Date: March 16, 2017
Due Date: Thursday, March 30, 2017 @ 11:59pm

Review the Project Ideas page

Form a team for your class project (no more than 3 people to a team) and email a PDF of your project proposal (1/2 to 1 page) to "jhuappliedgenomics at gmail dot com" by 11:59pm on March 30.

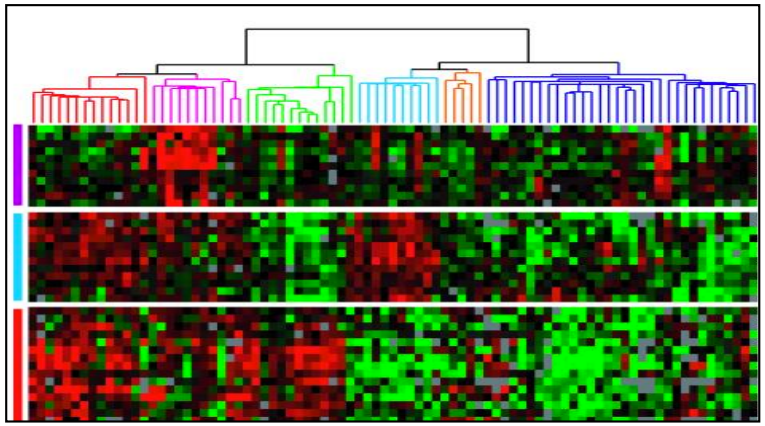
The proposal should have the following components:

- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)

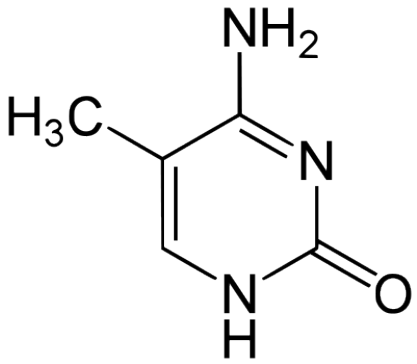
After submitting your proposal, we will schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we can schedule the meeting. No late days can be used for this project.

*-seq in 4 short vignettes

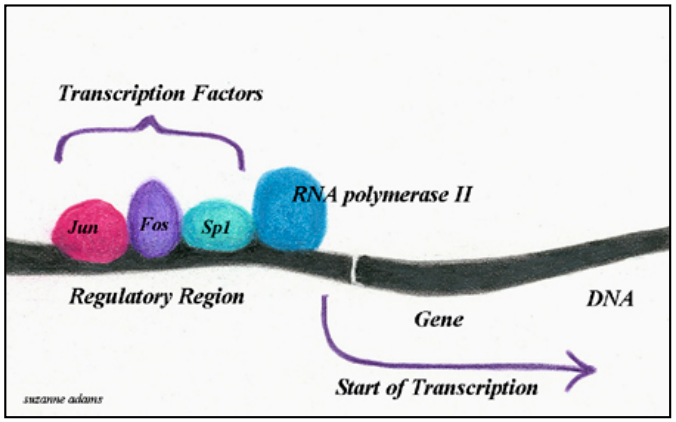
RNA-seq



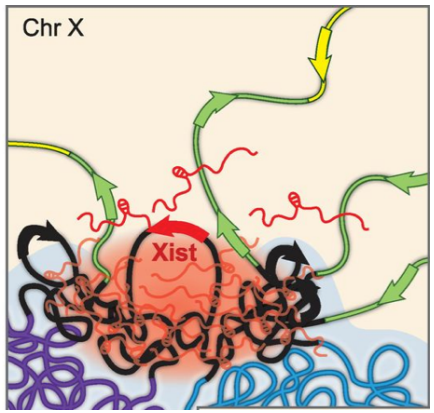
Methyl-seq



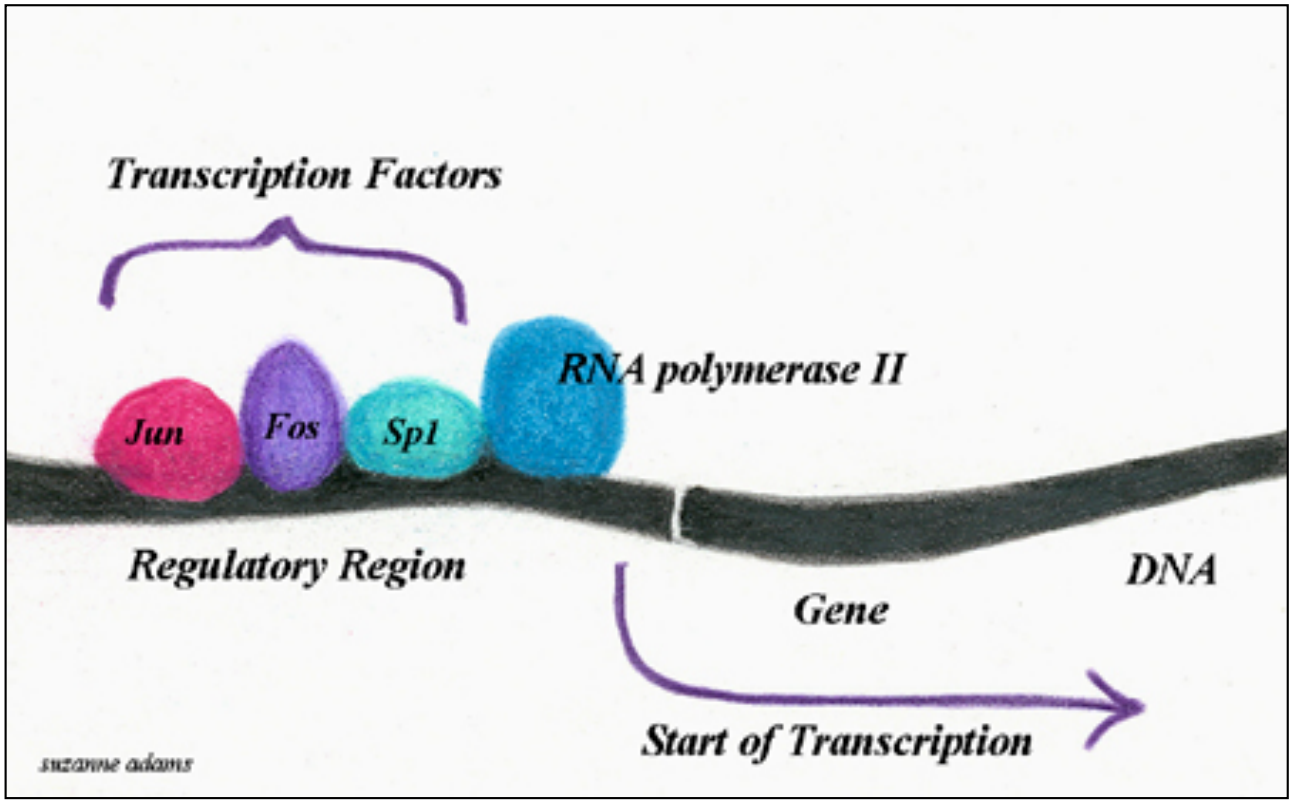
ChIP-seq



Hi-C

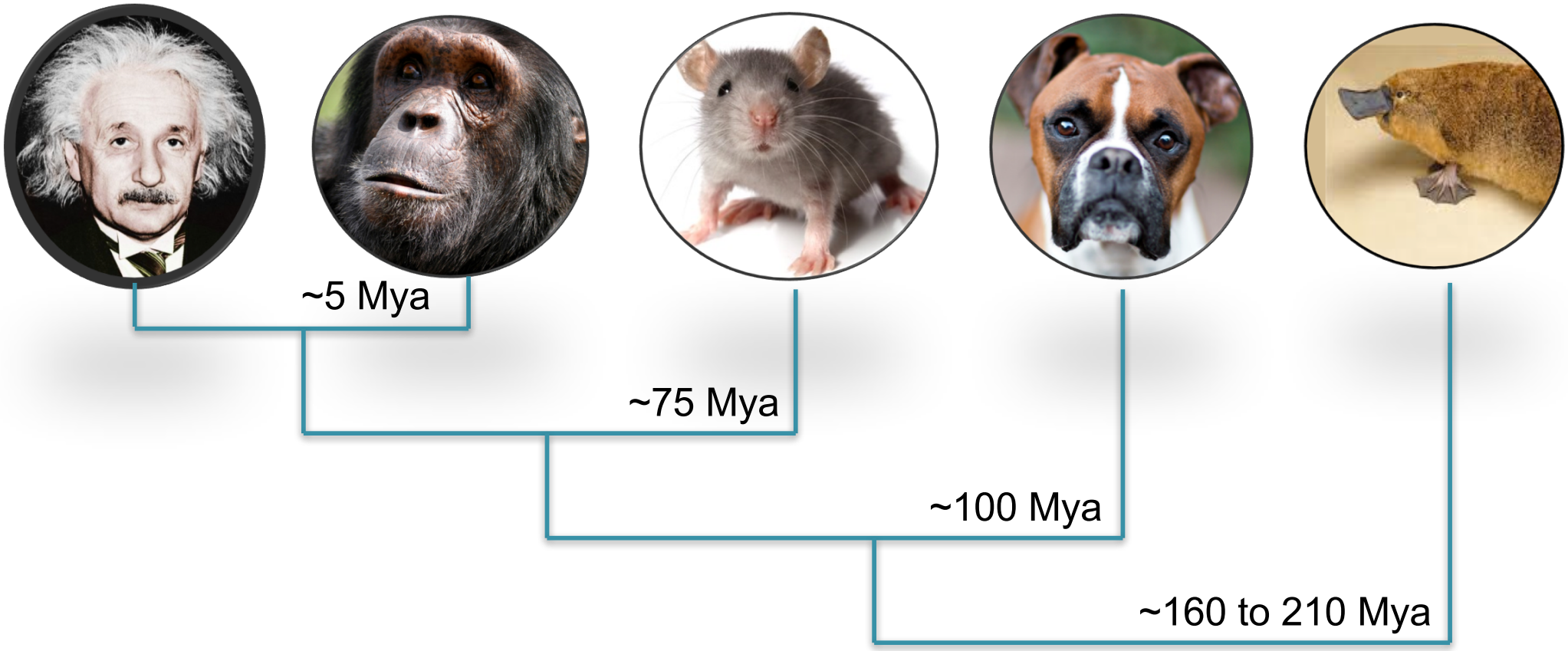


ChIP-seq



Genome-wide mapping of in vivo protein-DNA interactions.
Johnson *et al* (2007) *Science*. 316(5830):1497-502

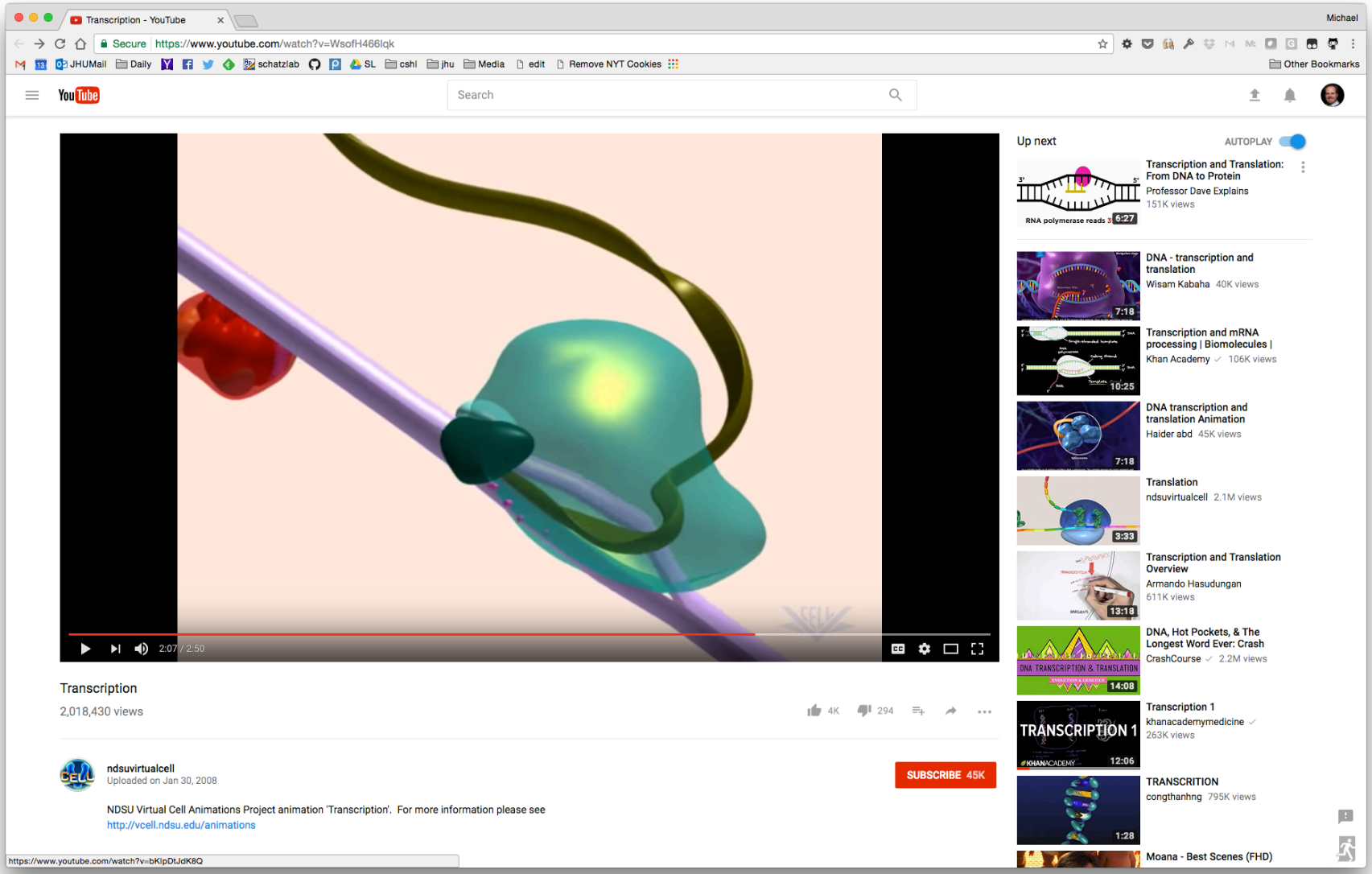
Human Evolution



As expected, the majority of platypus genes (82%; 15,312 out of 18,596) have orthologues in these five other amniotes (Supplementary Table 5). The remaining 'orphan' genes are expected to primarily reflect rapidly evolving genes, for which no other homologues are discernible, erroneous predictions, and true lineage-specific genes that have been lost in each of the other five species under consideration.

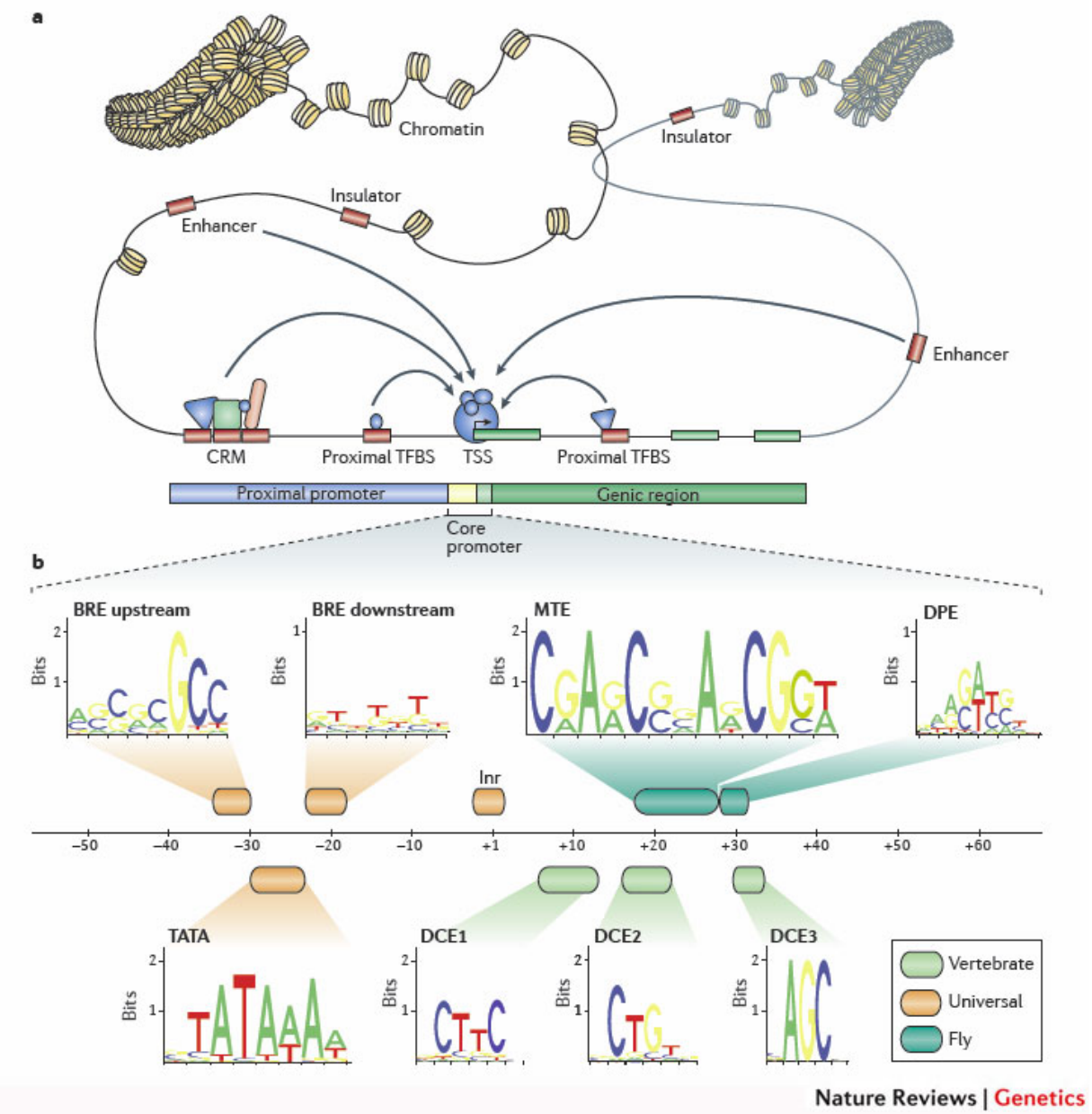
Genome analysis of the platypus reveals unique signatures of evolution
(2008) *Nature.* 453, 175-183 doi:10.1038/nature06936

Transcriptional Regulation



https://www.youtube.com/watch?v=WsofH466lqk

Promoters

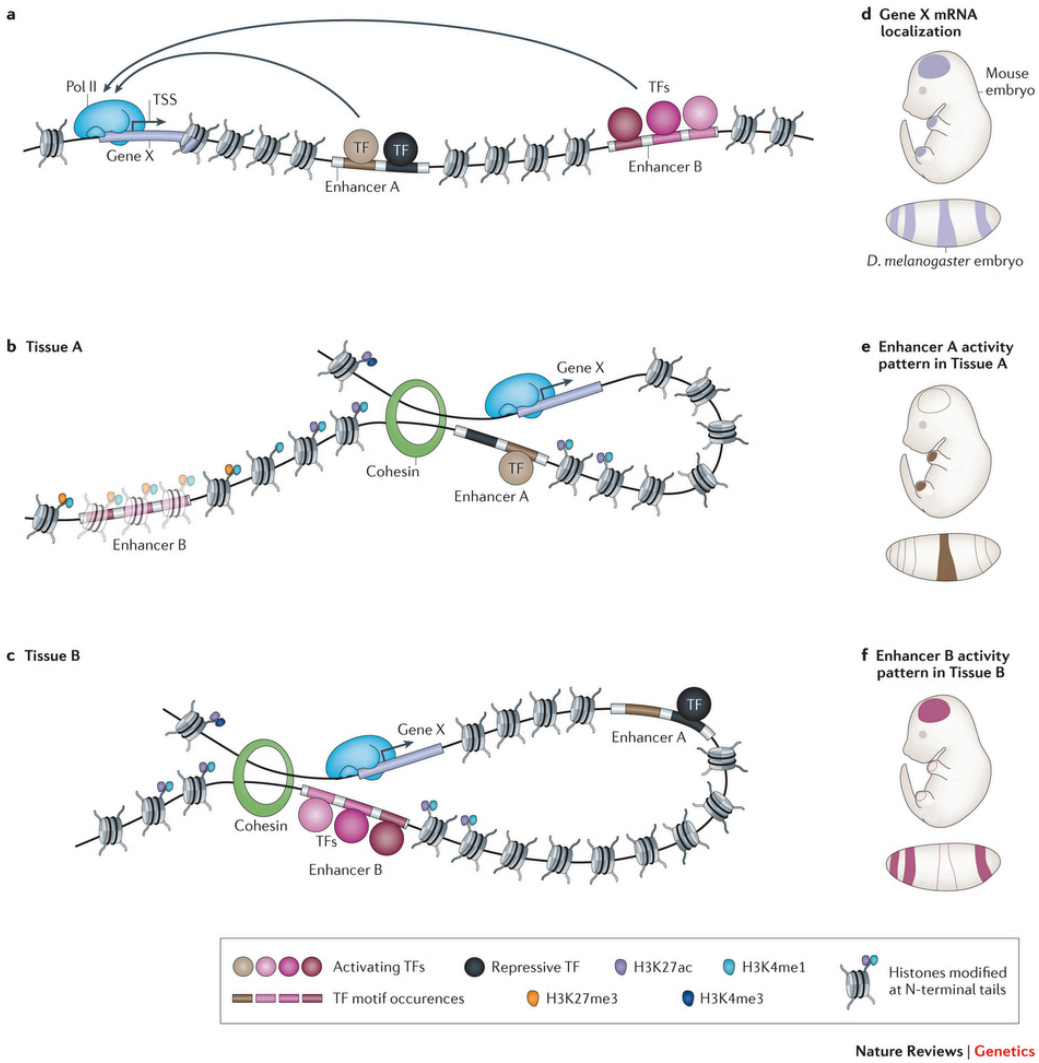


Metazoan promoters: emerging characteristics and insights into transcriptional regulation
Lenhard et al (2014) *Nature Reviews Genetics* 15, 272–286

Enhancers

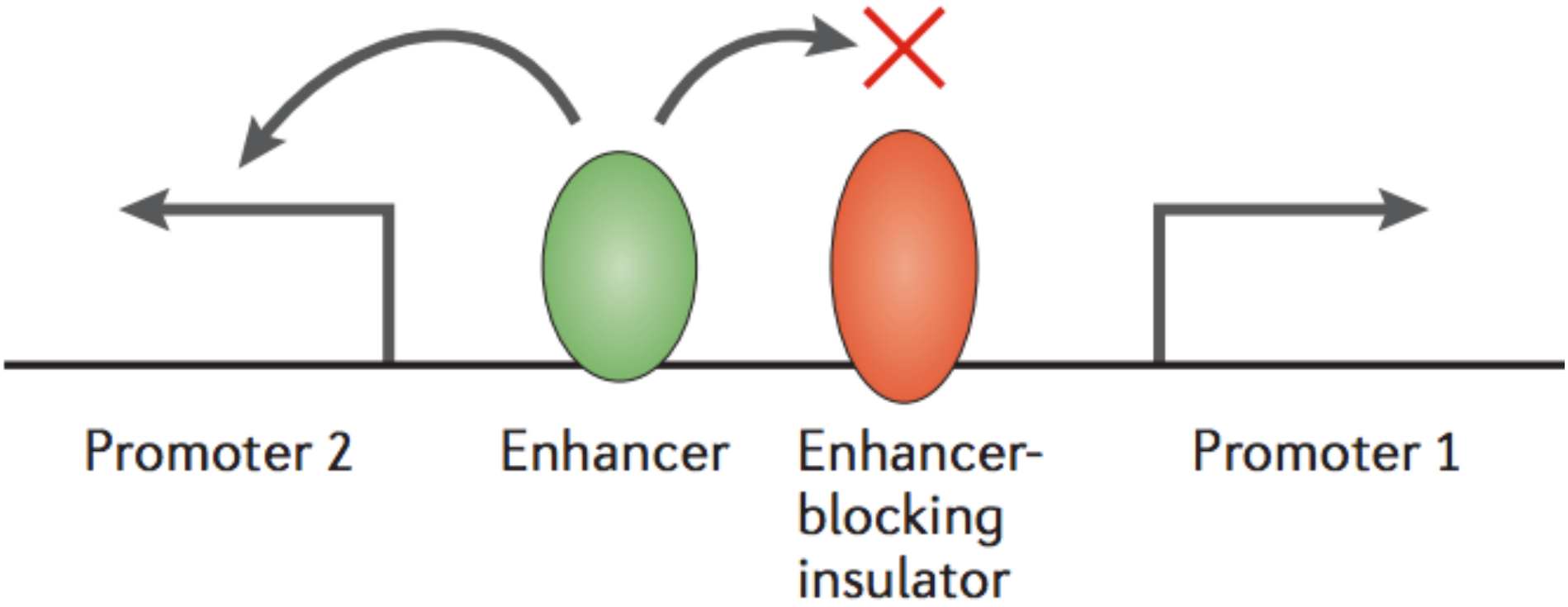
Enhancers are genomic regions that contain binding sites for transcription factors (TFs) and that can upregulate (enhance) the transcription of a target gene.

- Enhancers can be located at any distance from their target genes (up to ~1Mbp)
- In a given tissue, active enhancers (Enhancer A in part b or Enhancer B in part c) are bound by activating TFs and are brought into proximity of their respective target promoters by looping
- Active and inactive gene regulatory elements are marked by various biochemical features
- Complex patterns of gene expression result from the additive action of different enhancers with cell-type- or tissue-specific activities



Transcriptional enhancers: from properties to genome-wide predictions
Shlyueva et al (2014) *Nature Reviews Genetics* 15, 272–286

Insulators



Insulators are DNA sequence elements that prevent "inappropriate interactions" between adjacent chromatin domains.

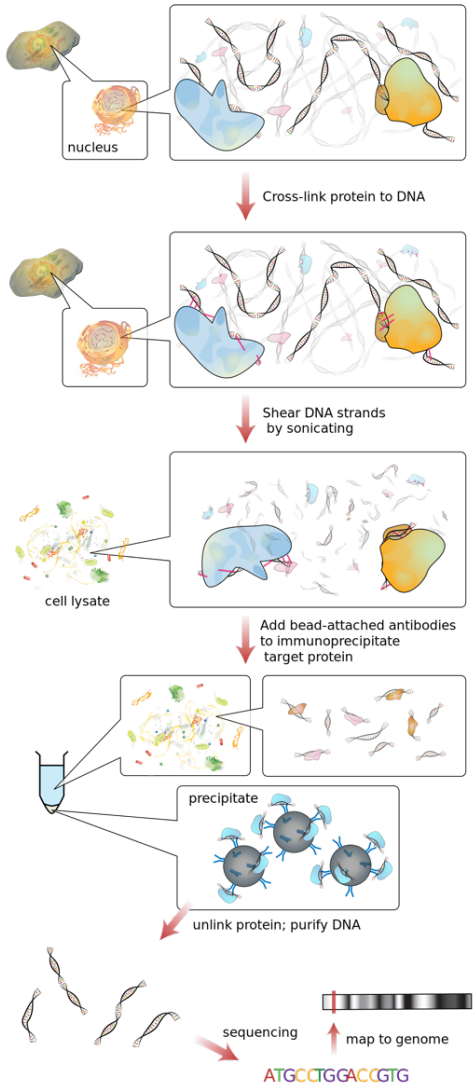
- One type of insulator establishes domains that separate enhancers and promoters to block their interaction,
- Second type creates a barrier against the spread of heterochromatin.

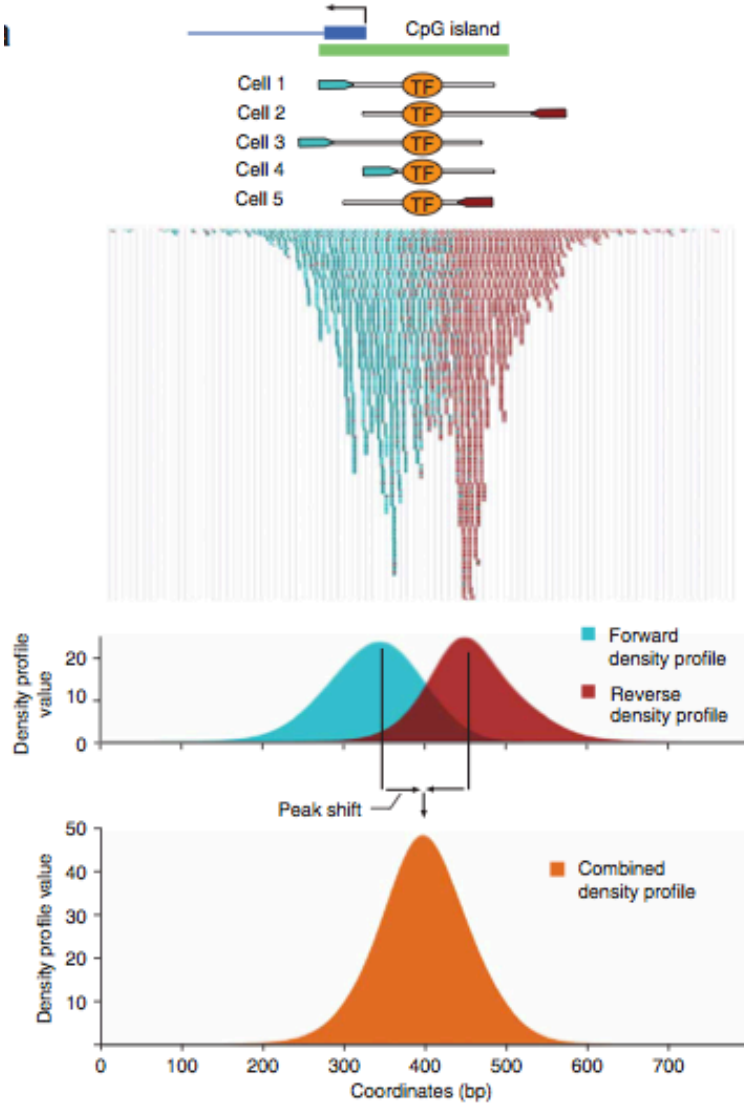
Insulators: exploiting transcriptional and epigenetic mechanisms
Gaszner & Felsenfeld (2006) *Nature Reviews Genetics* 7, 703-713. doi:10.1038/nrg1925

ChIP-seq: TF Binding

Goals:

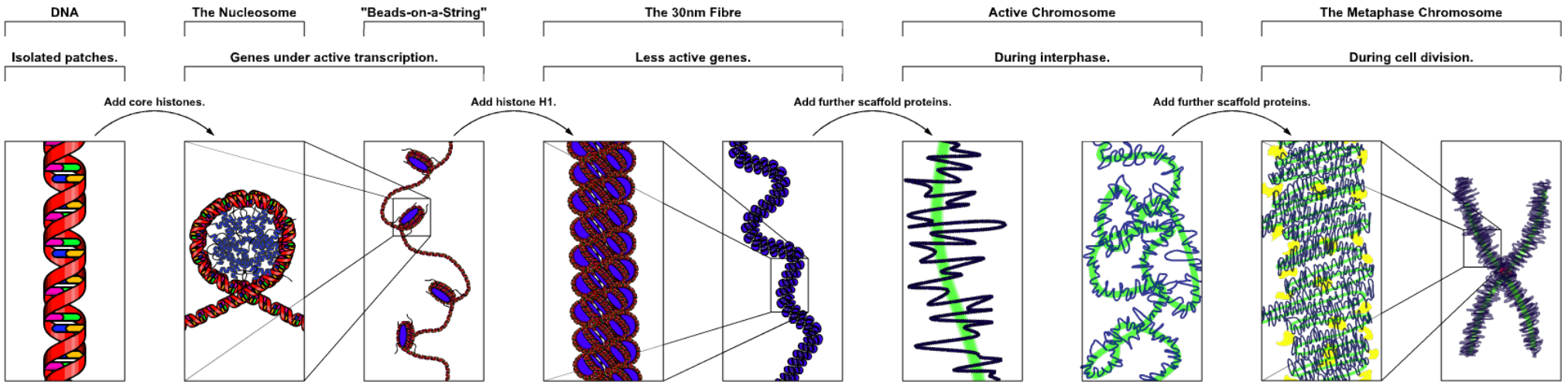
- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?





Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data
Valouev et al (2008) *Nature Methods*. 5, 829 - 834

Chromatin compaction model



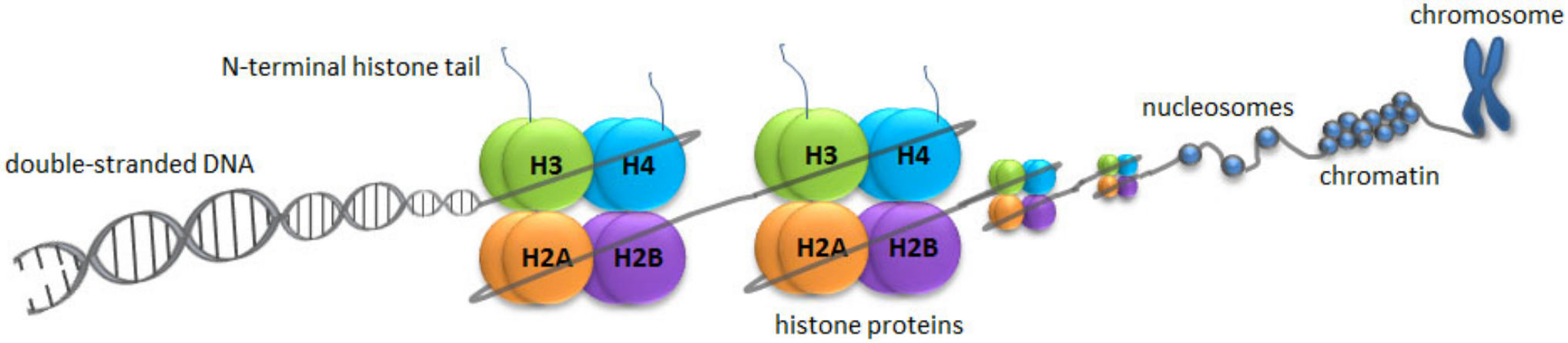
Nucleosome is a basic unit of DNA packaging in eukaryotes

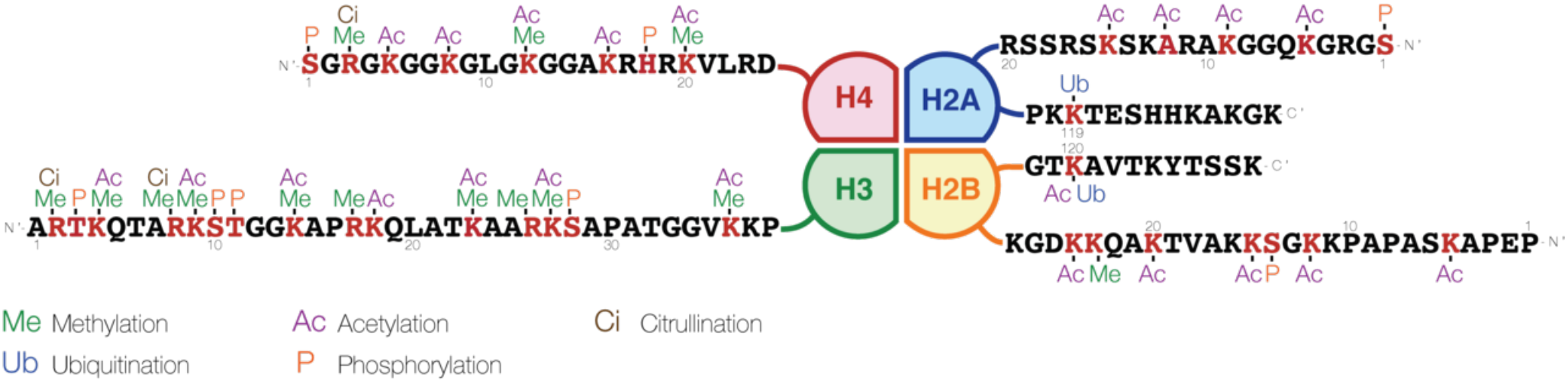
- Consists of a segment of 146bp DNA wound in sequence around eight histone protein cores (thread wrapped around a spool) followed by a ~38bp linker
- Under active transcription, nucleosomes appear as “beads-on-a-string”, but are more densely packed for less active genes

Nucleosomes form the fundamental repeating units of eukaryotic chromatin

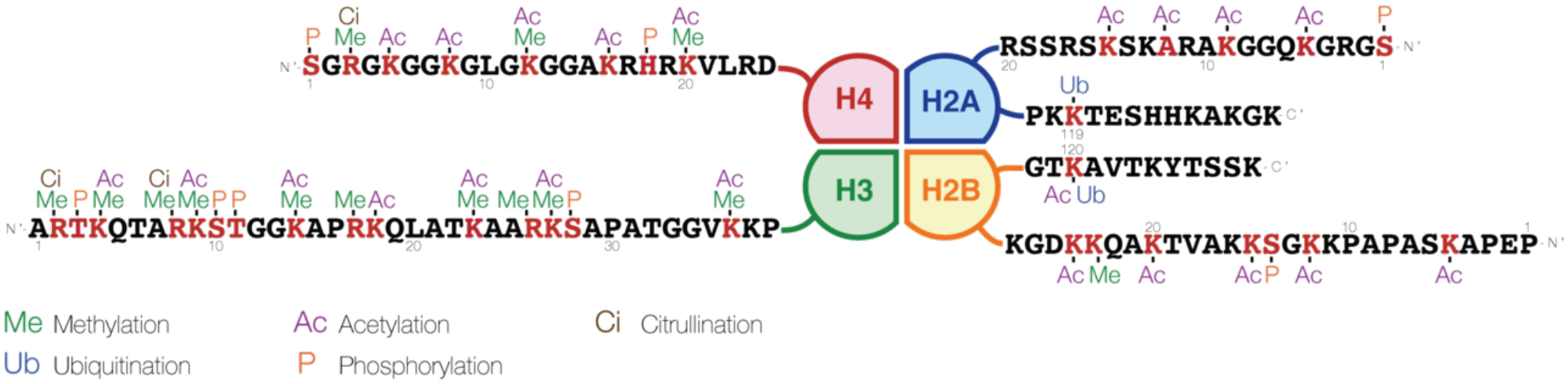
- Used to pack the large eukaryotic genomes into the nucleus while still ensuring appropriate access to it (in mammalian cells approximately 2 m of linear DNA have to be packed into a nucleus of roughly 10 μm diameter).

ChIP-seq: Histone Modifications





ChIP-seq: Histone Modifications

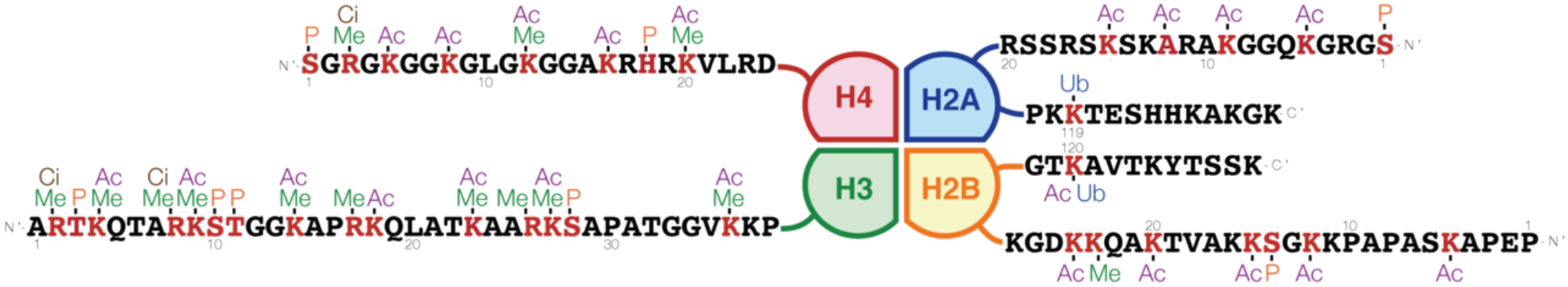


The common nomenclature of histone modifications is:

- The name of the histone (e.g., H3)
- The single-letter amino acid abbreviation (e.g., K for Lysine) and the amino acid position in the protein
- The type of modification (Me: methyl, P: phosphate, Ac: acetyl, Ub: ubiquitin)
- The number of modifications (only Me is known to occur in more than one copy per residue. 1, 2 or 3 is mono-, di- or tri-methylation)

So H3K4me1 denotes the monomethylation of the 4th residue (a lysine) from the start (i.e., the N-terminal) of the H3 protein.

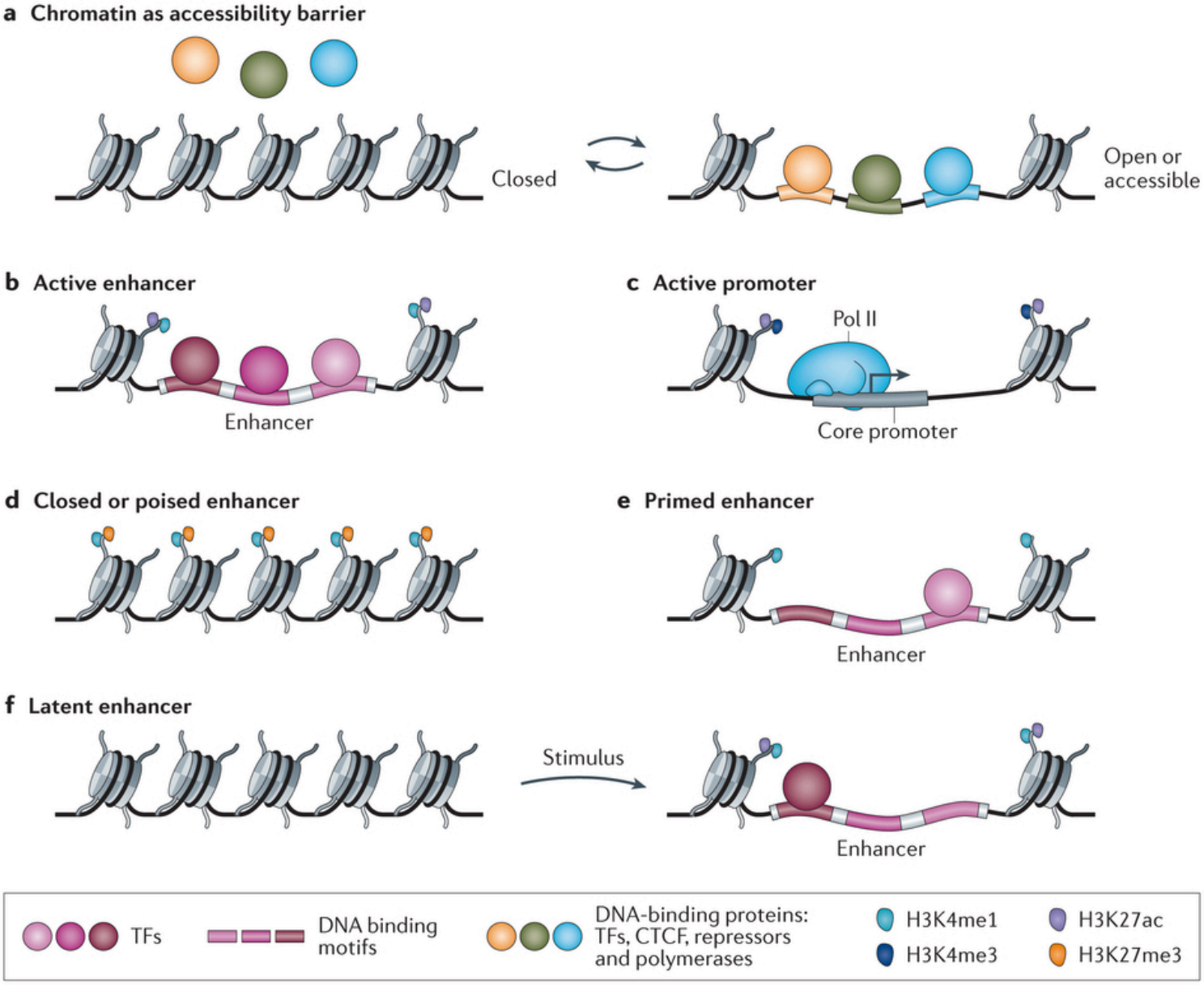
ChIP-seq: Histone Modifications



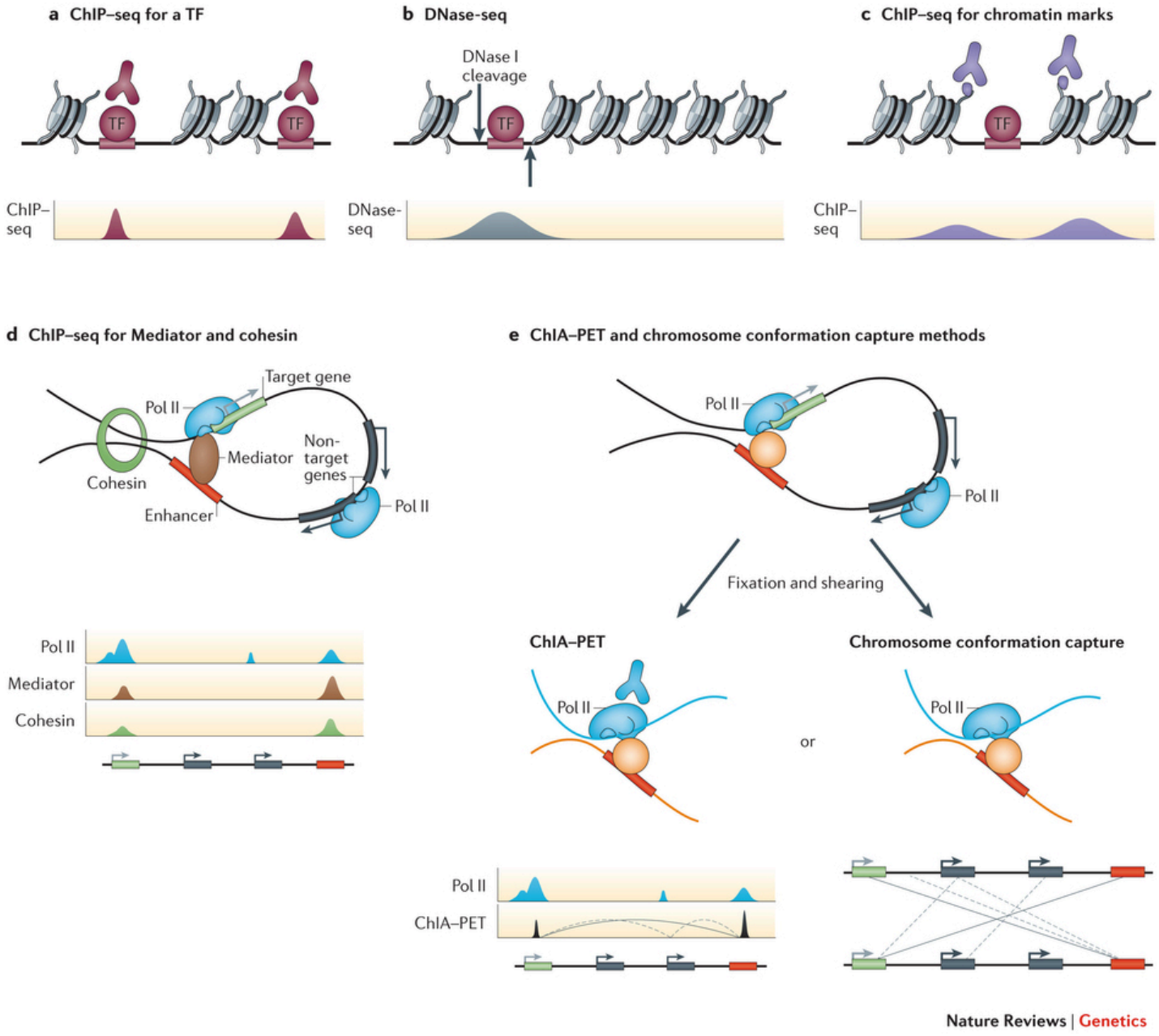
Type of modification	Histone							
	H3K4	H3K9	H3K14	H3K27	H3K79	H3K122	H4K20	H2BK5
mono-methylation	activation[6]	activation[7]		activation[7]	activation[7][8]		activation[7]	activation[7]
di-methylation	activation	repression[3]		repression[3]	activation[8]			
tri-methylation	activation[9]	repression[7]		repression[7]	activation,[8] repression[7]			repression[3]
acetylation		activation[9]	activation[9]	activation[10]		activation[11]		

- H3K4me3 is enriched in transcriptionally active promoters.[12]
- H3K9me3 is found in constitutively repressed genes.
- H3K27me is found in facultatively repressed genes.[7]
- H3K36me3 is found in actively transcribed gene bodies.
- H3K9ac is found in actively transcribed promoters.
- H3K14ac is found in actively transcribed promoters.
- H3K27ac distinguishes active enhancers from poised enhancers.
- H3K122ac is enriched in poised promoters and also found in a different type of putative enhancer that lacks H3K27ac.

Enhancer States

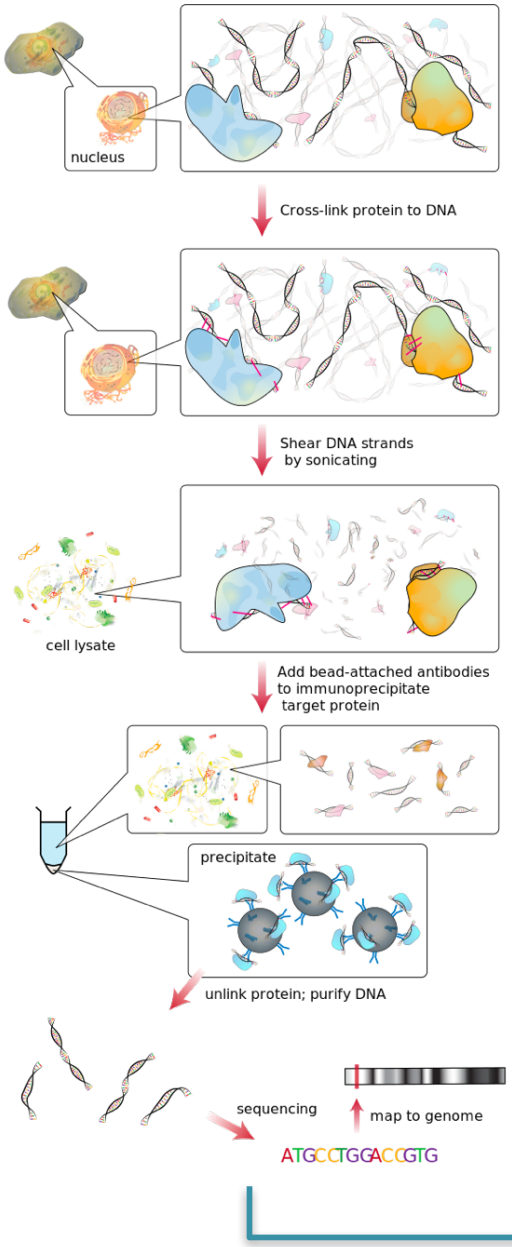


Related Assays



Nature Reviews | Genetics

General Flow of ChIP-seq Analysis



Peak Calling

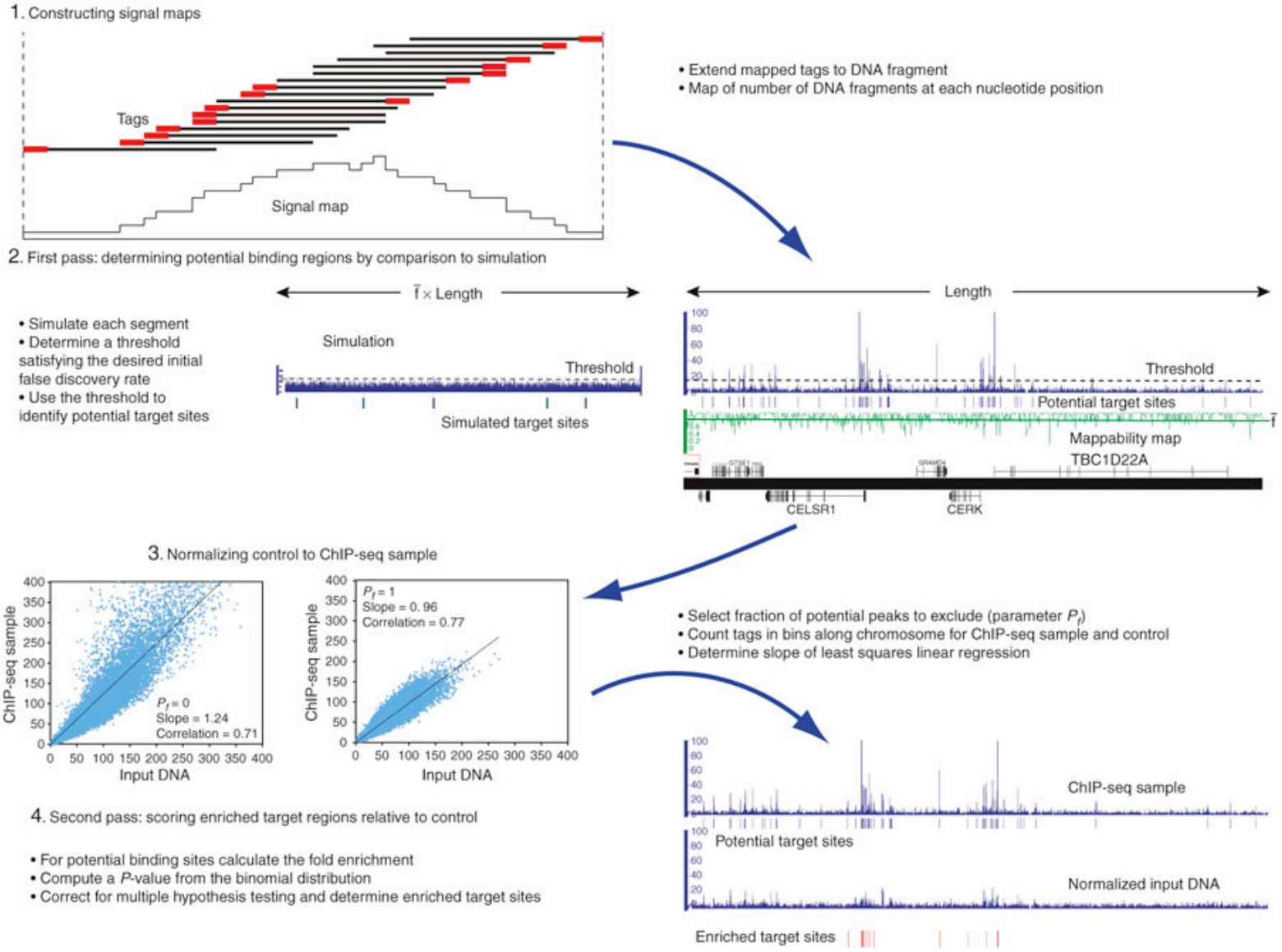
Analyze Enriched Regions

Visualization

Peak-peak comparisons

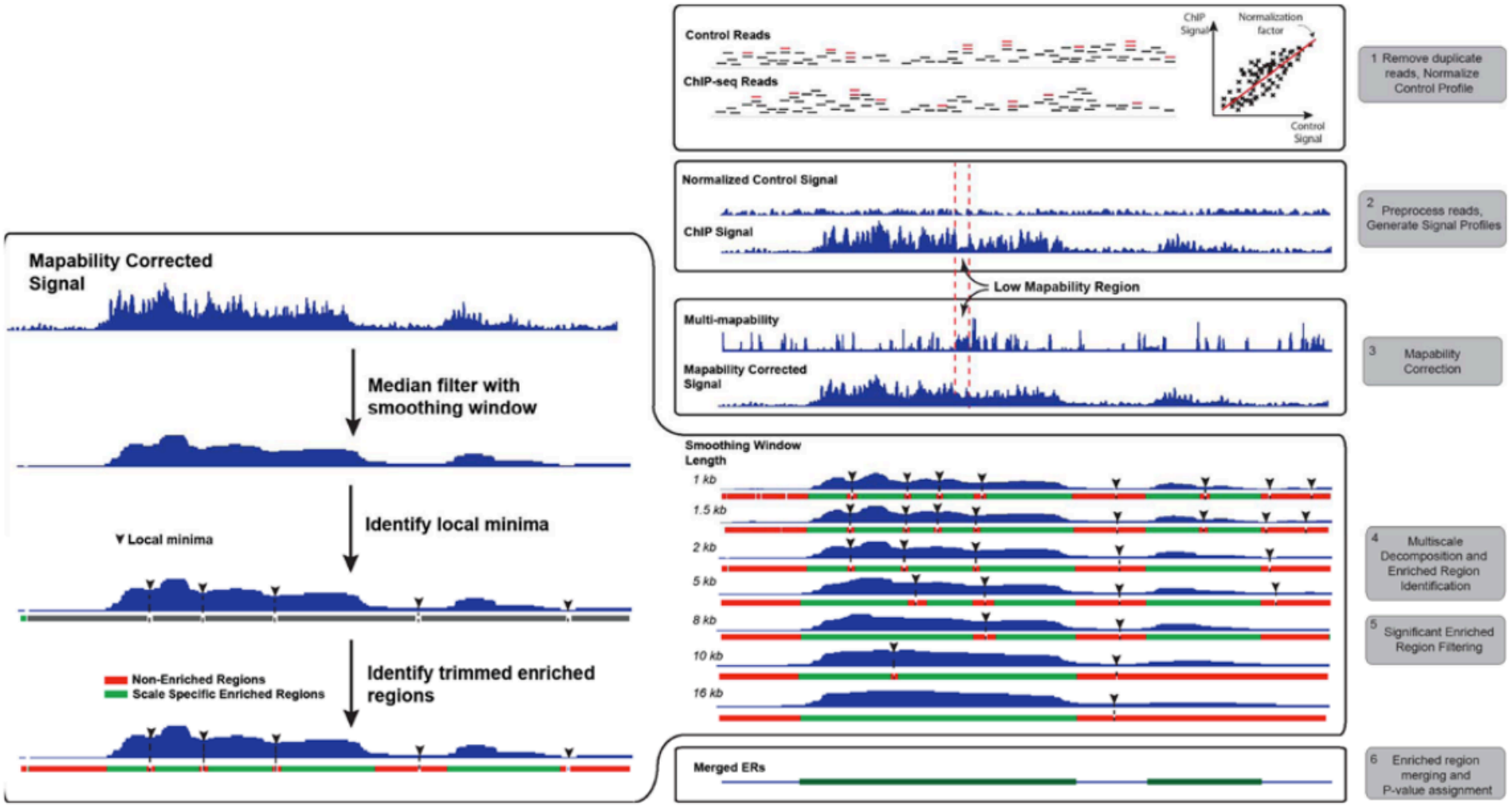
Annotation / Interpretation

PeakSeq



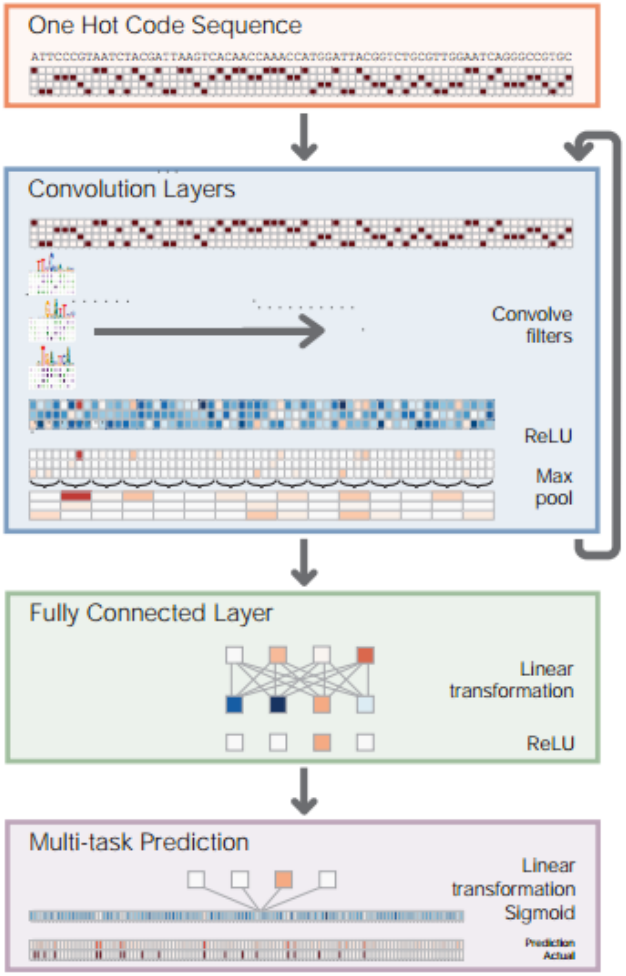
PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls
Rozowsky et al (2009) Nature Biotechnology 27, 66 - 75

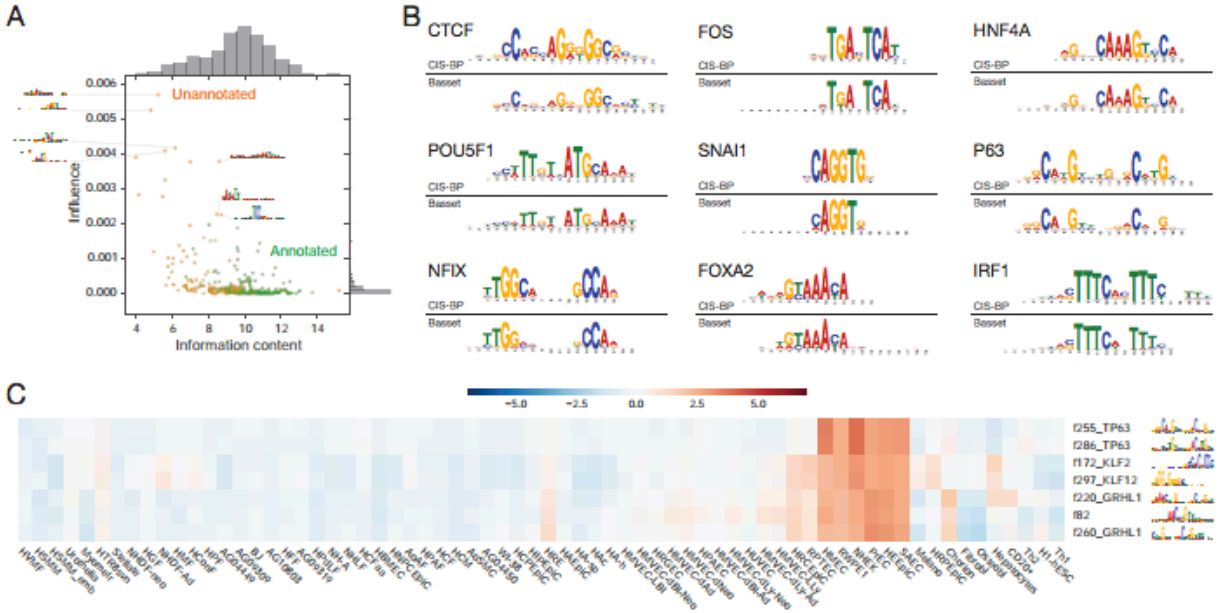
MUSIC



MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework
Harmanci et al. Genome Biology 2014, 15:474

Basset

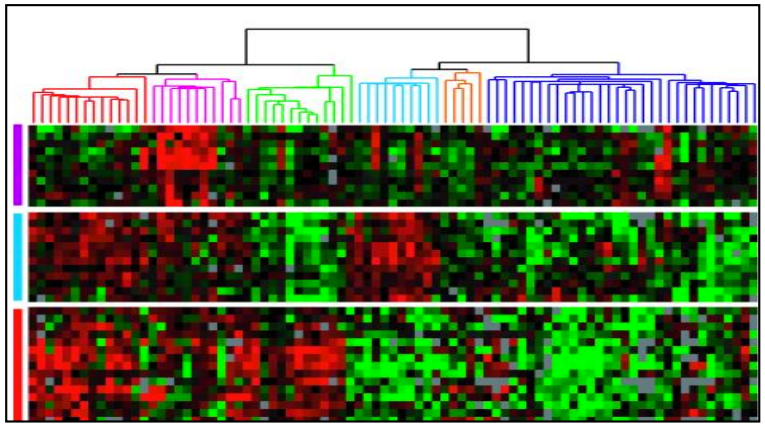




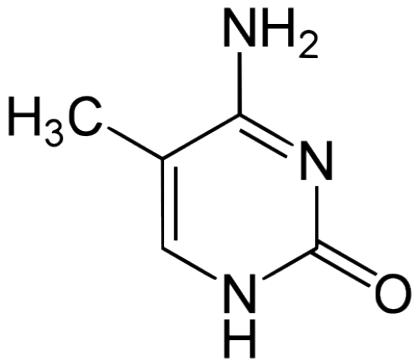
Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks
Kelley et al. (2016) Genome Research doi: 10.1101/gr.200535.115

*-seq in 4 short vignettes

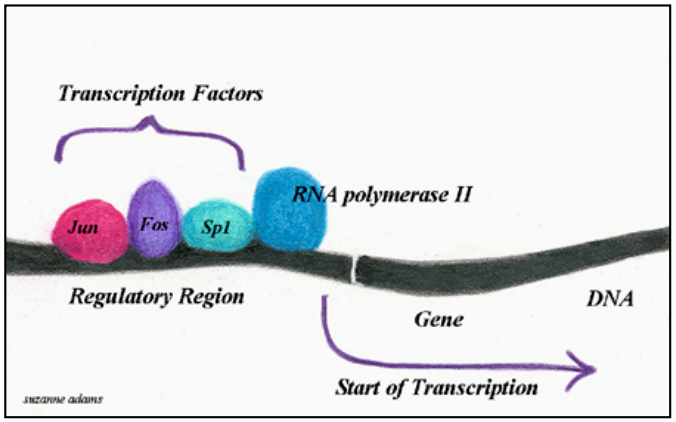
RNA-seq



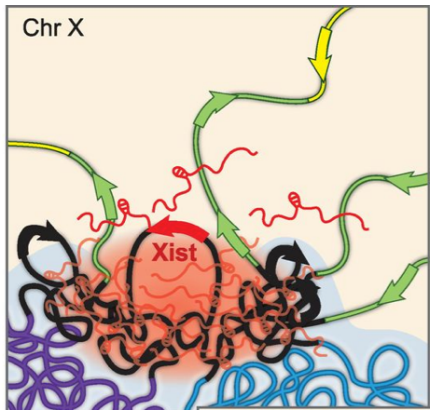
Methyl-seq



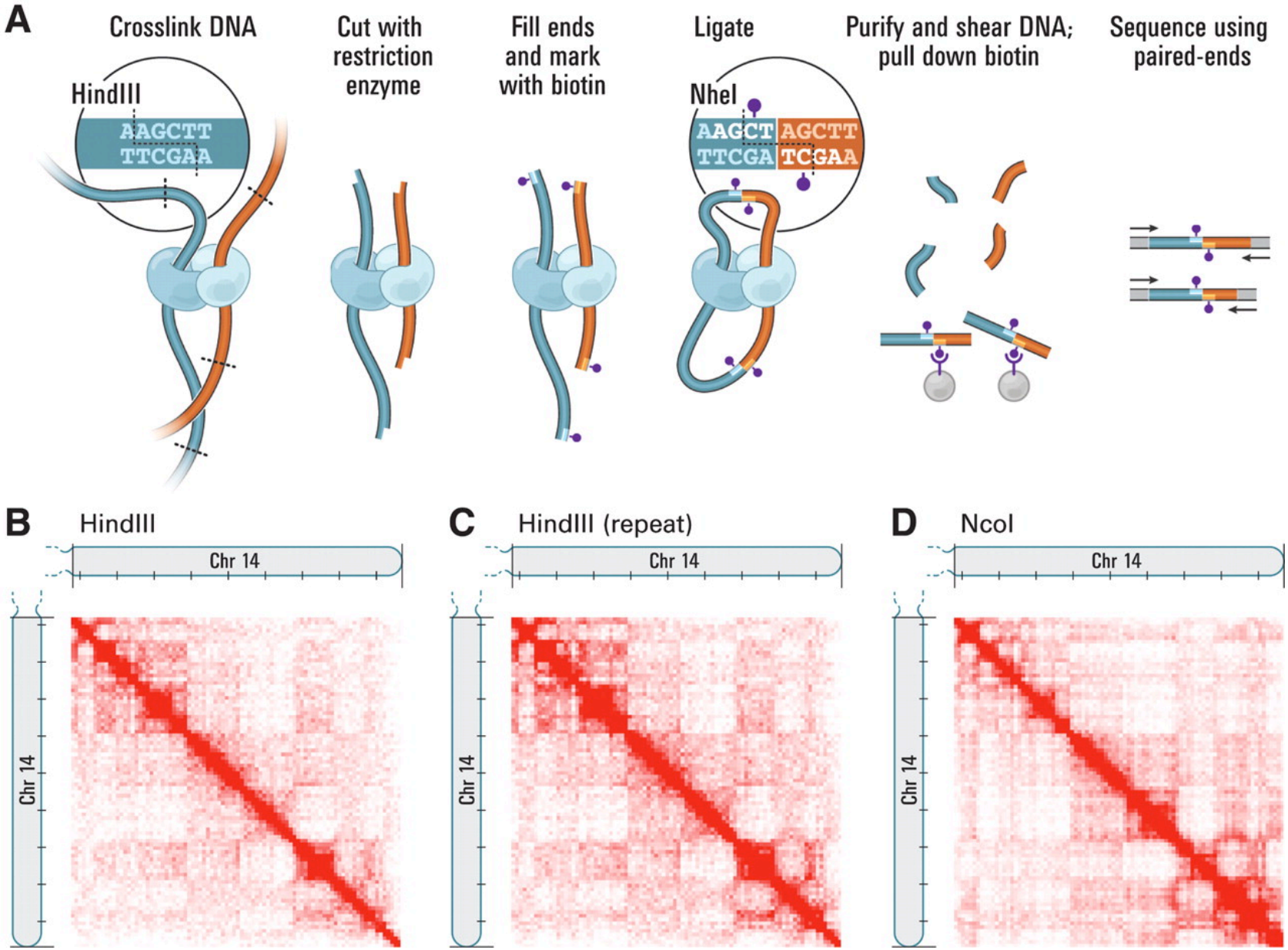
ChIP-seq



Hi-C

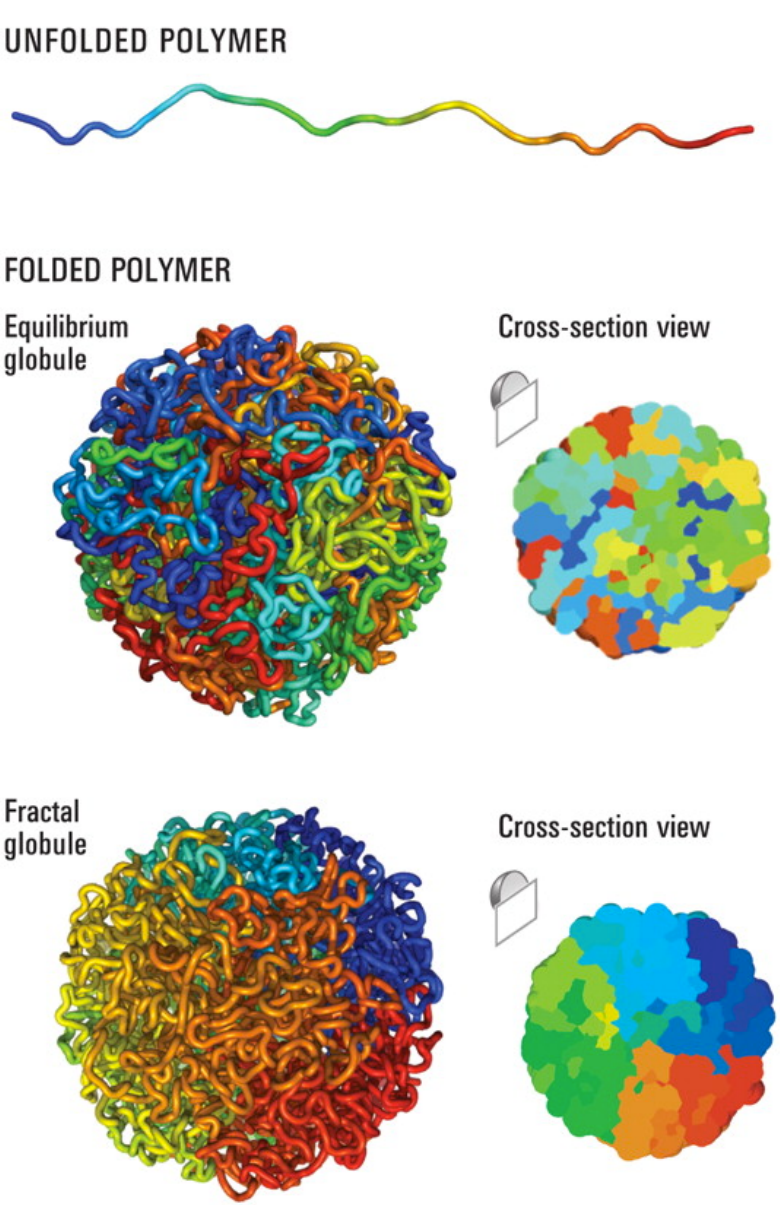


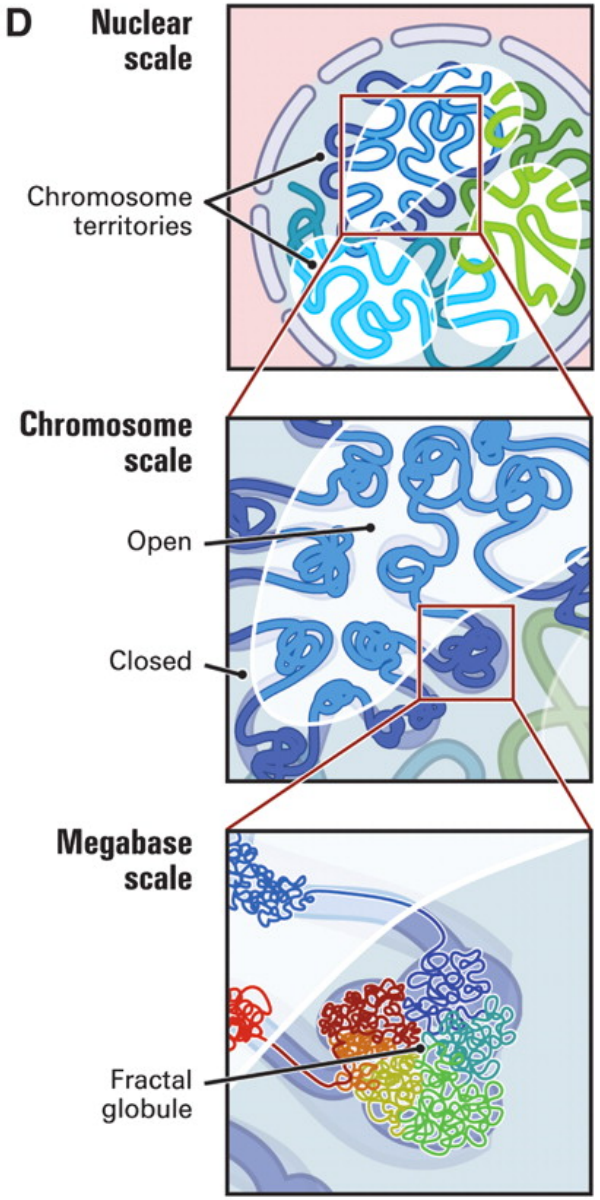
HI-C: Mapping the folding of DNA



Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome
Liberman-Aiden et al. (2009) *Science.* 326 (5950): 289-293

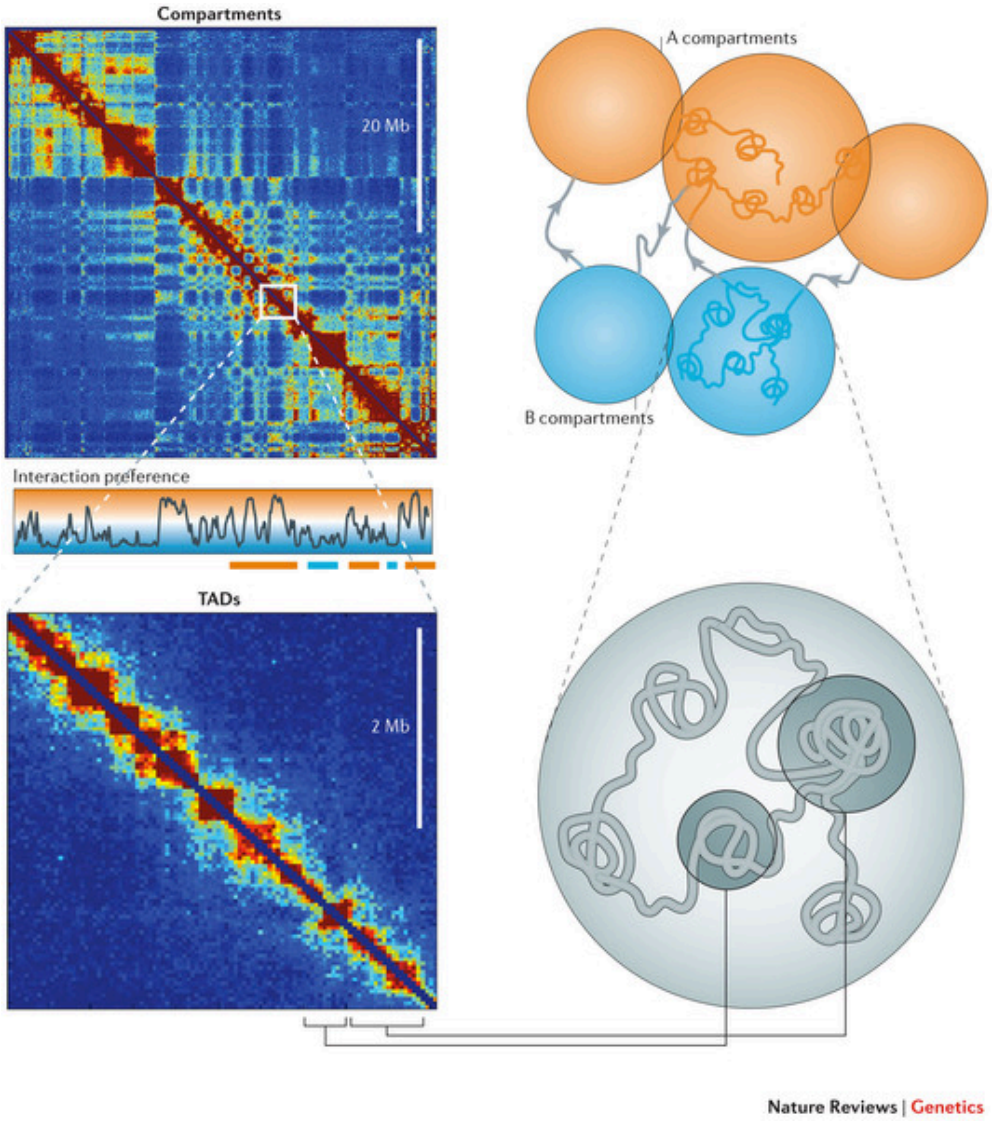
HI-C: Mapping the folding of DNA





Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome
Liberman-Aiden et al. (2009) *Science.* 326 (5950): 289-293

Genome compartments & TADs



Mammalian genomes have a pattern of interactions that can be approximated by two compartments called "A" and "B"

- alternate along chromosomes and have a characteristic size of ~5 Mb each.
- " A" compartments (orange) preferentially interact with other A compartments; "B" compartments (blue) associate with other "B" compartments.
- "A" compartments are largely euchromatic, transcriptionally active regions.

Topologically associating domains (TADs)

- TADs are smaller (~400–500 kb)
- Can be active or inactive, and adjacent TADs are not necessarily of opposite chromatin status.
- TADs are hard-wired features of chromosomes, and groups of adjacent TADs can organize in A and B compartments

Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data
Dekker et al. (2013) *Nature Reviews Genetics 14, 390–403*

Gene Regulation in 3-dimensions

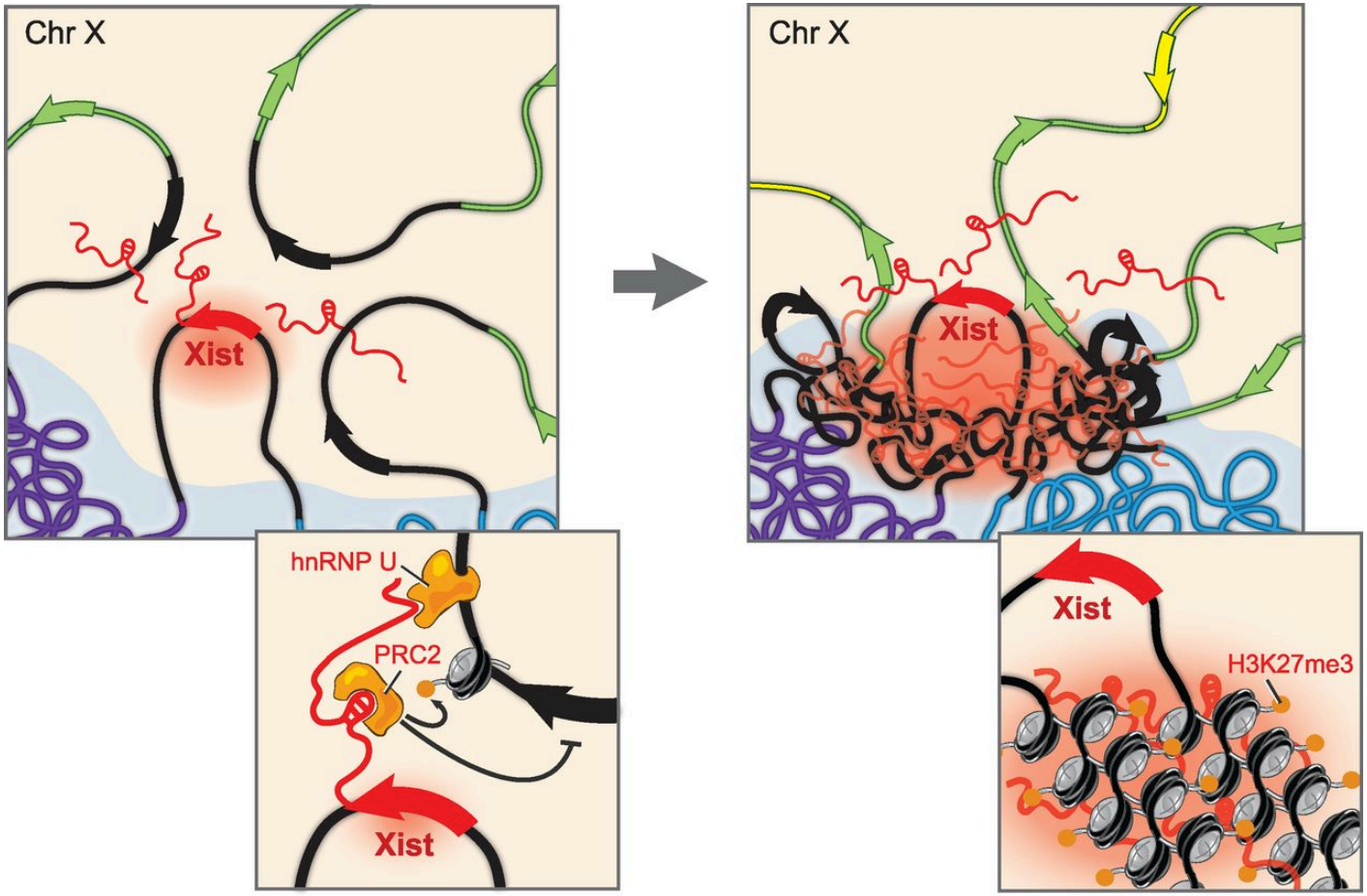
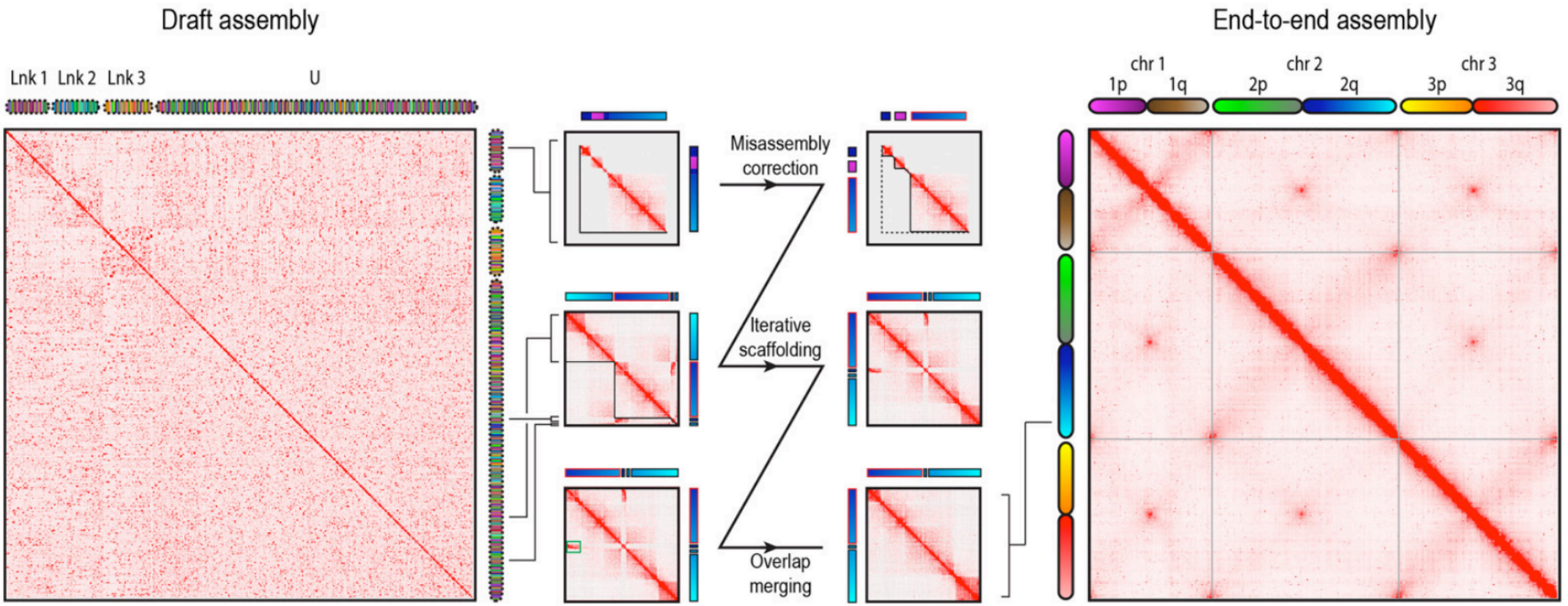


Fig 6. A model for how Xist exploits and alters three-dimensional genome architecture to spread across the X chromosome.

The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome
Engreitz et al. (2013) *Science*. 341 (6147)

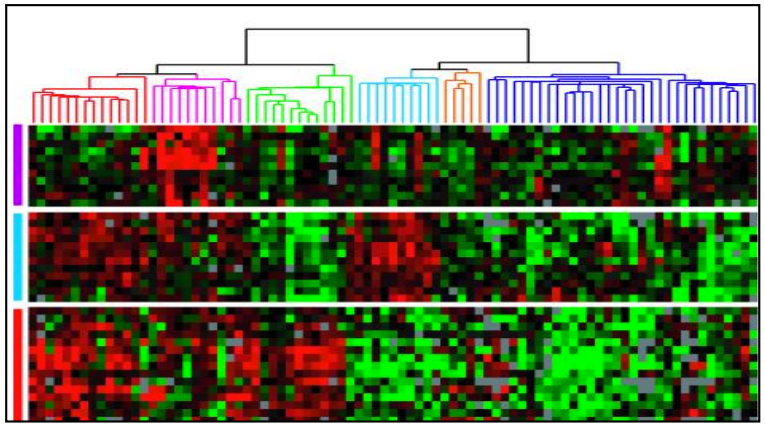
Scaffolding with Hi-C



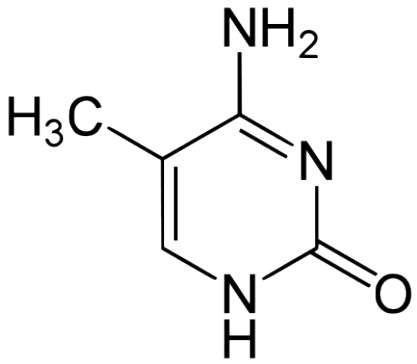
De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds
O. Dudchenko et al. (2017) Science 10.1126/science.aal3327

Putting it all together!

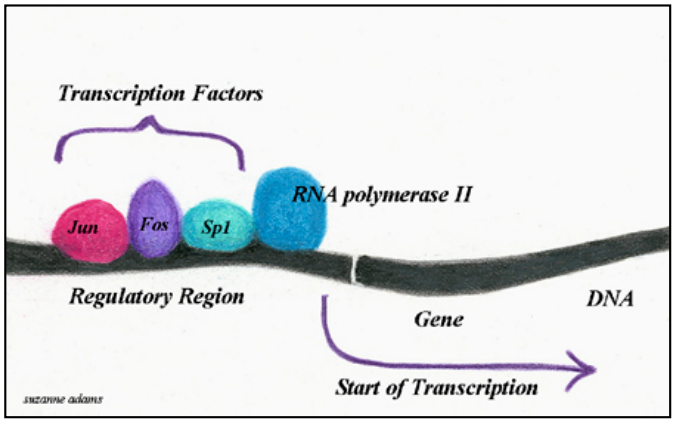
RNA-seq



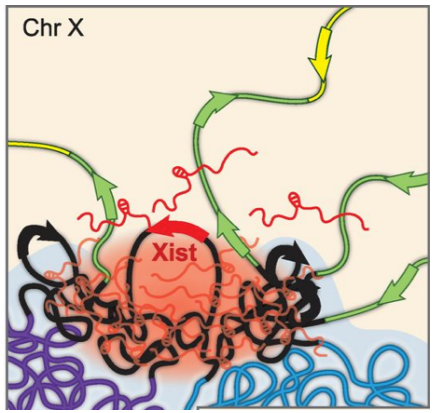
Methyl-seq



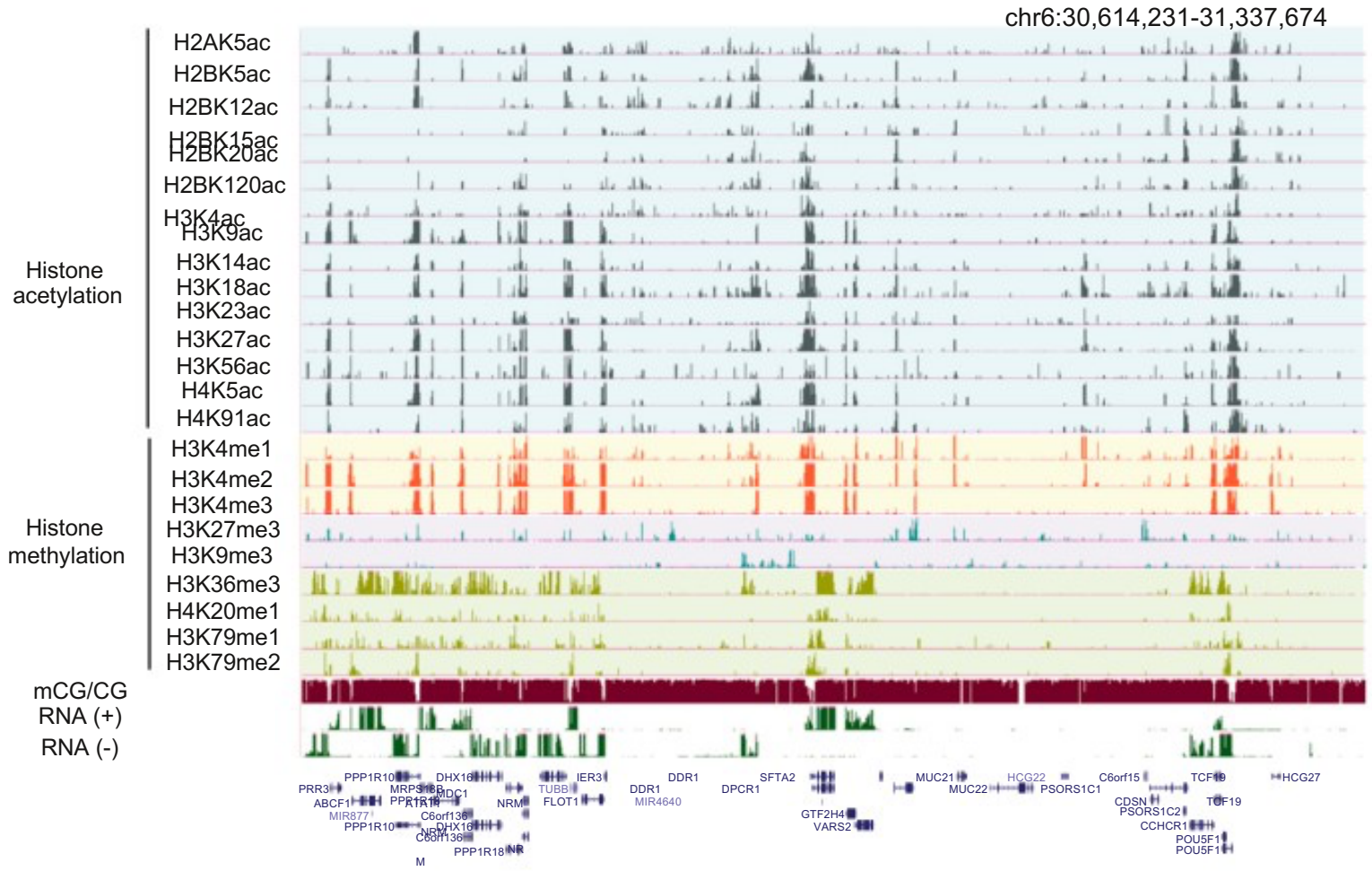
ChIP-seq



Hi-C

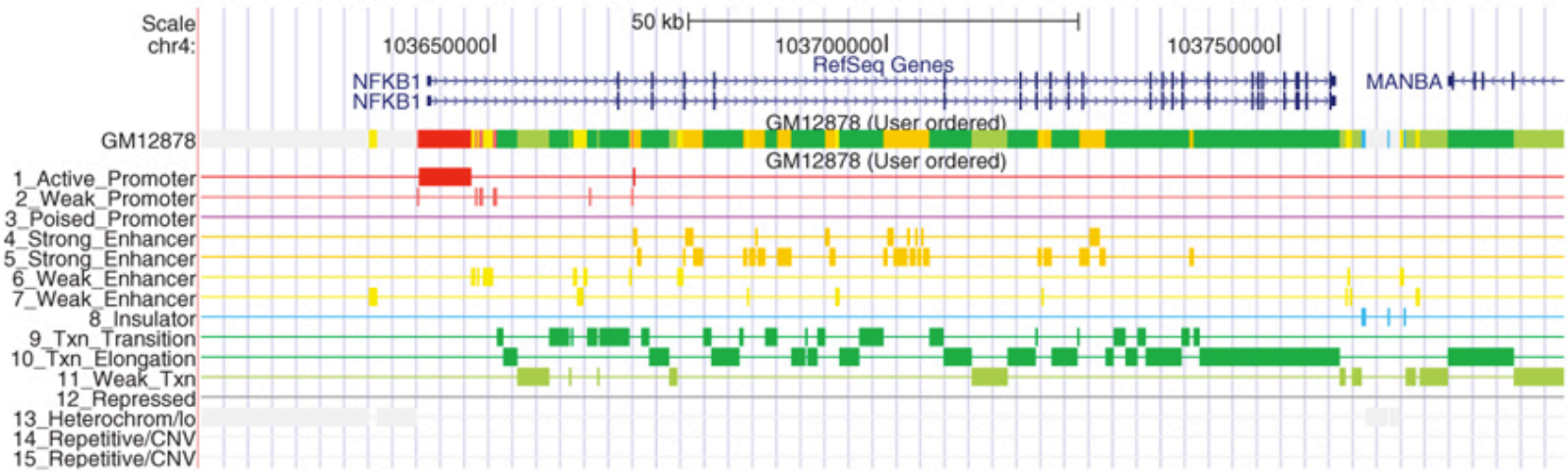


We can call peaks, but…



We need a way to summarize the combinatorial patterns of multiple histone marks into meaningful biological units

ChromHMM



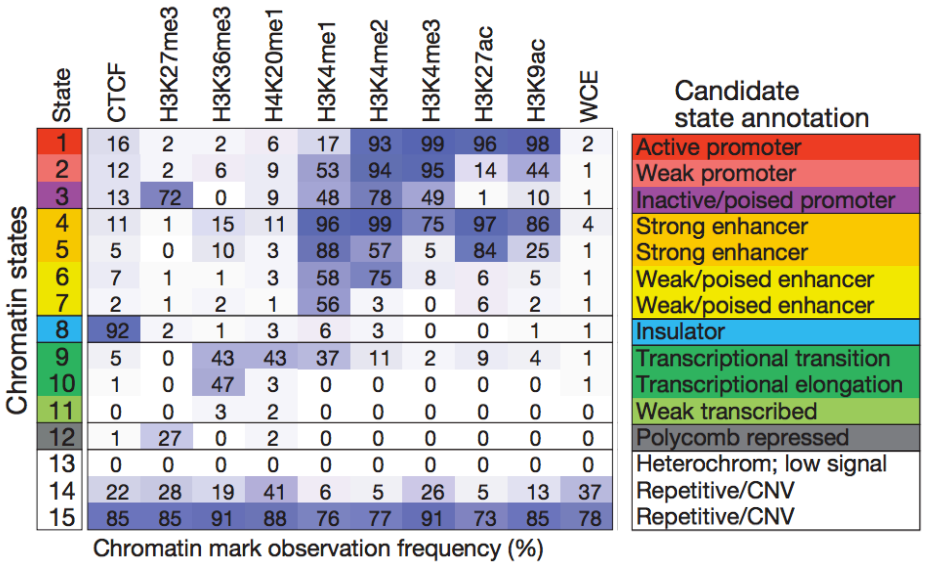
ChromHMM is software for learning and characterizing chromatin states.

- ChromHMM can integrate multiple chromatin datasets such as ChIP-seq data of various histone modifications to discover de novo the major re-occuring combinatorial and spatial patterns of marks.
- ChromHMM is based on a multivariate Hidden Markov Model that explicitly models the presence or absence of each chromatin mark.
- The resulting model can then be used to systematically annotate a genome in one or more cell types.

ChromHMM: automating chromatin-state discovery and characterization
Ernst & Kellis (2012) Nature Methods 9, 215–216. doi:10.1038/nmeth.1906

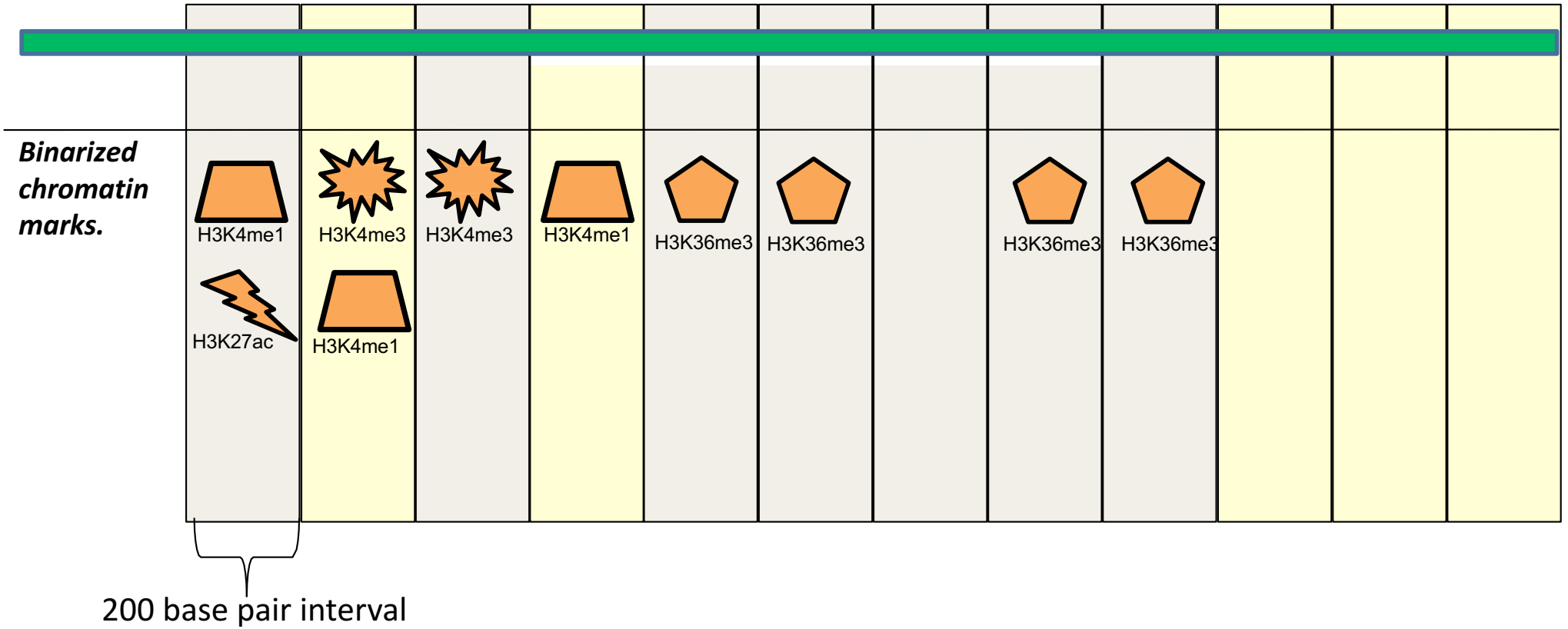
ChromHMM and Chromatin States

Chromatin states are defined based on different combinations of histone modifications and correspond to different functional regions



The goal is to segment every base of the the genome into biologically meaningful units: reveal & annotate *functional* elements

ChromHMM : Multivariate Hidden Markov Model

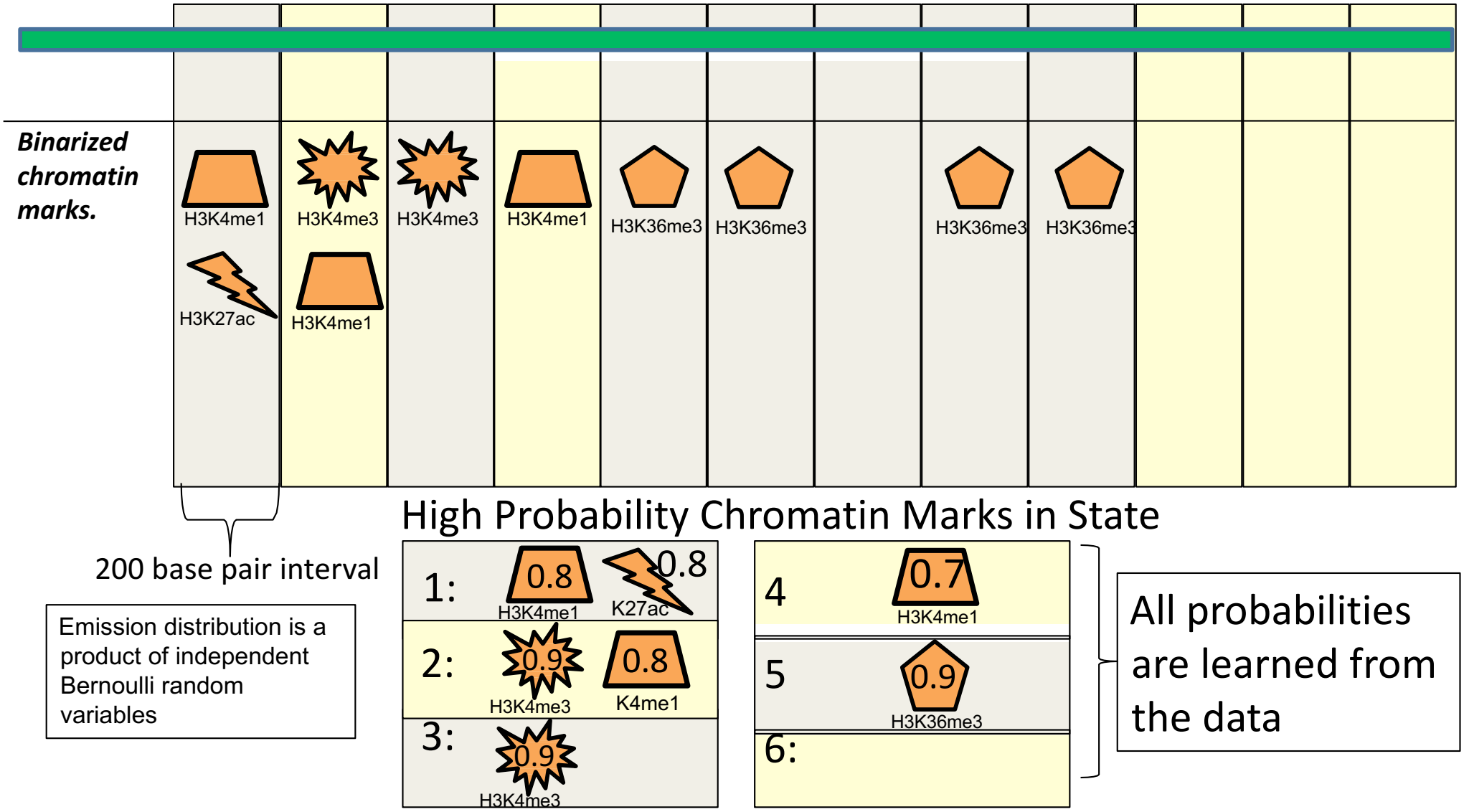


Emission distribution is a product of independent Bernoulli random variables

Binarization leads to explicit modeling of mark combinations and interpretable parameters

Ernst and Kellis, Nat Biotech 2010 ; Ernst and Kellis, Nature Methods 2012

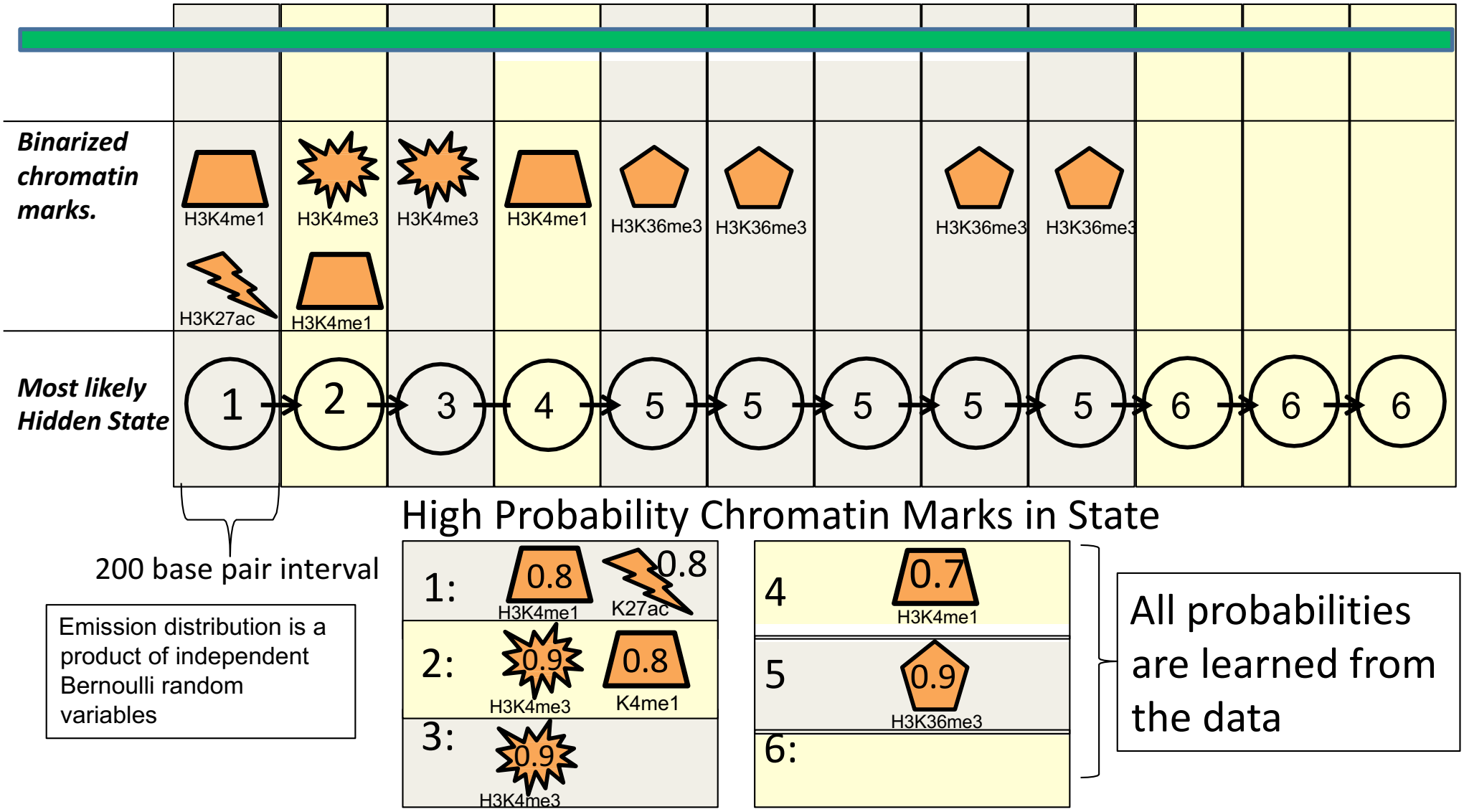
ChromHMM : Multivariate Hidden Markov Model



Binarization leads to explicit modeling of mark combinations and interpretable parameters

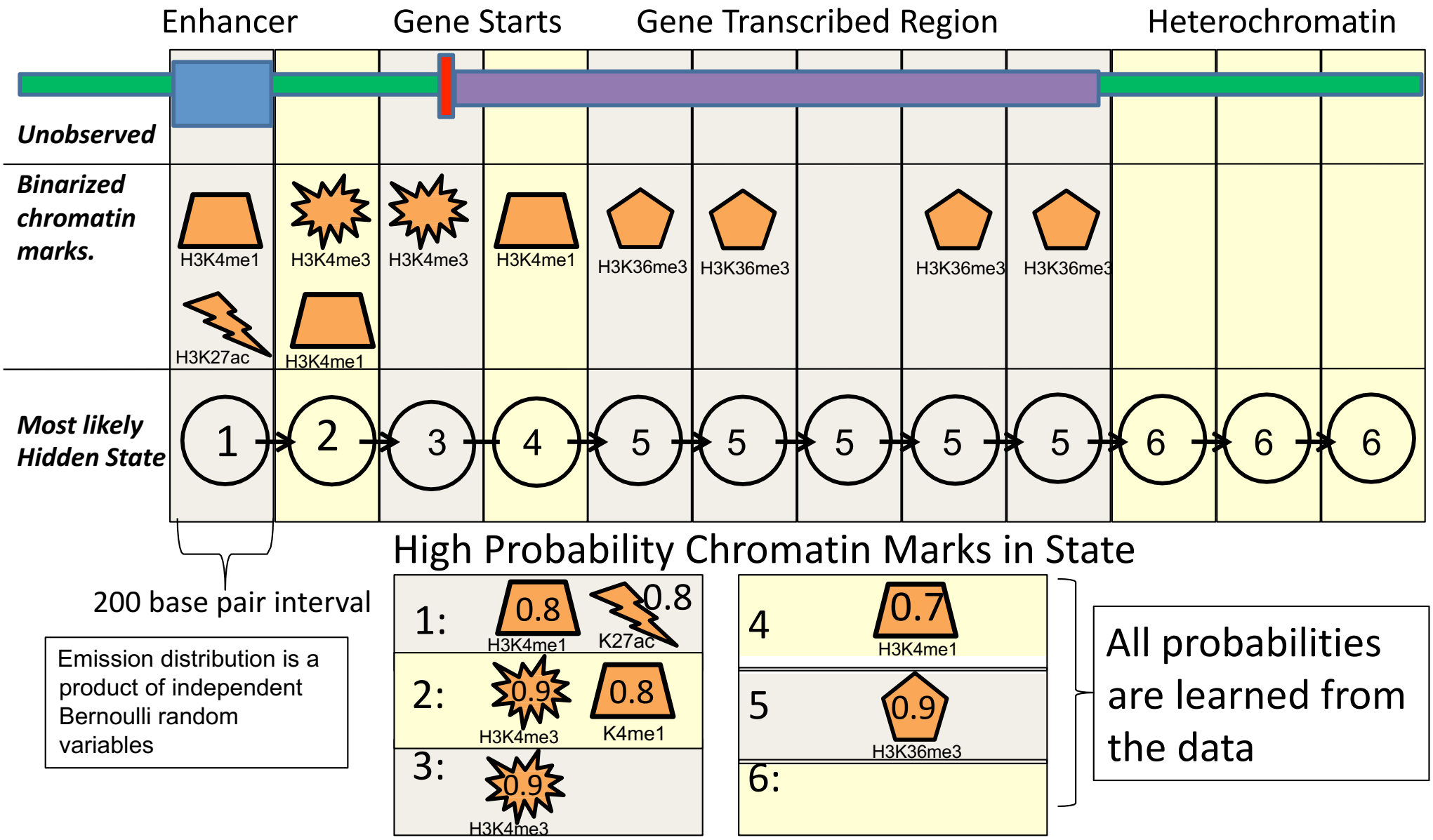
Ernst and Kellis, Nat Biotech 2010 ; Ernst and Kellis, Nature Methods 2012

ChromHMM : Multivariate Hidden Markov Model



Ernst and Kellis, Nat Biotech 2010 ; Ernst and Kellis, Nature Methods 2012

ChromHMM : Multivariate Hidden Markov Model



Binarization leads to explicit modeling of mark combinations and interpretable parameters

Ernst and Kellis, Nat Biotech 2010 ; Ernst and Kellis, Nature Methods 2012

The Workflow

1. Get ChIP-seq raw reads for different histone modifications
2. Align the reads to a reference genome
3. Convert aligned reads in bed format
4. Create Binned and Binarized Tracks
5. Train the model
6. Infer the states
7. Interpretation

Thanks Sri!

Align the ChIP-seq reads

- Starting from a file containing raw reads (usually a fastq file) you need to align them to a reference genome to get a .bam file (binary aligned file).

  You can use Bowtie or BWA.

Optionally one can also download the alignments from encode website: https://encode.project.org

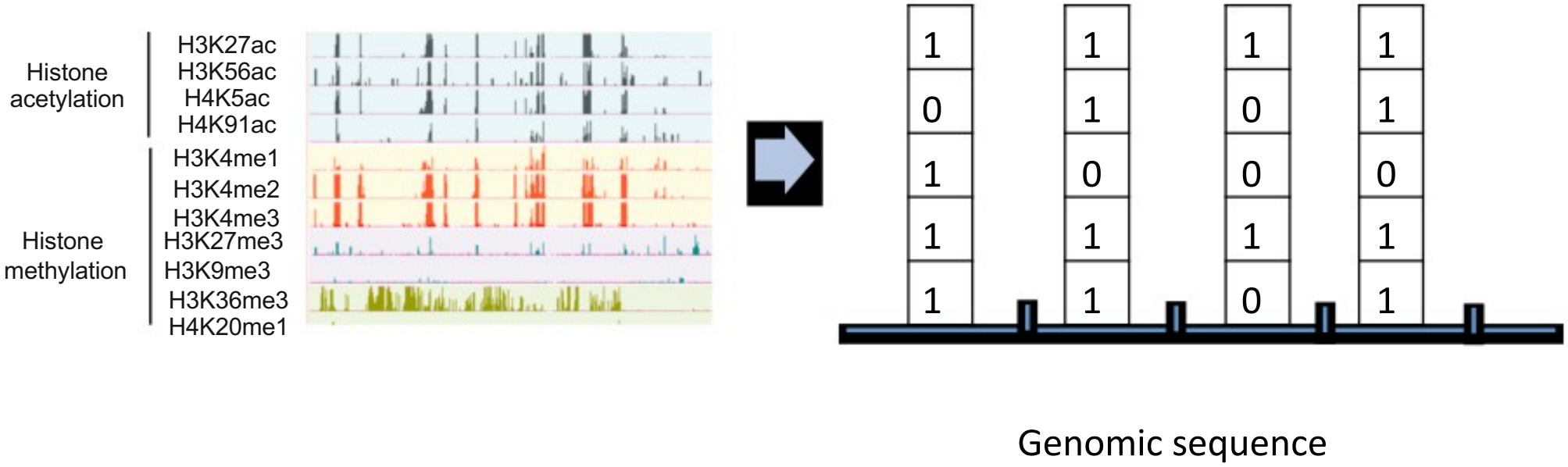
Convert aligned reads to bed format

- ChromHMM needs the aligned reads in .bed


```
bedtools bamtobed -i cell1_mark1.bam > ~/
data/cell1_mark1.bed
```

Create Binned and Binarized Tracks

- ChromHMM quantify the presence or absence of each mark in bins of fixed size



Genomic sequence

Create Binned and Binarized Tracks

- ```
  java  -mx4000M -jar ChromHMM.jar \
        BinarizeBed -b 200  CHROMSIZES/hg18 \
        ~/data/ cellmarkfiletable.txt \
        SAMPLEDATA_HG18
  ```

- Inside the cellmarkfiletable.txt:

```
cell1 mark1 cell1_mark1.bed cell1_control.bed
cell1 mark2 cell1_mark2.bed cell1_control.bed
cell2 mark1 cell2_mark1.bed cell2_control.bed
cell2 mark2 cell2_mark2.bed cell2_control.bed
```
```

Train the model and segment the genome

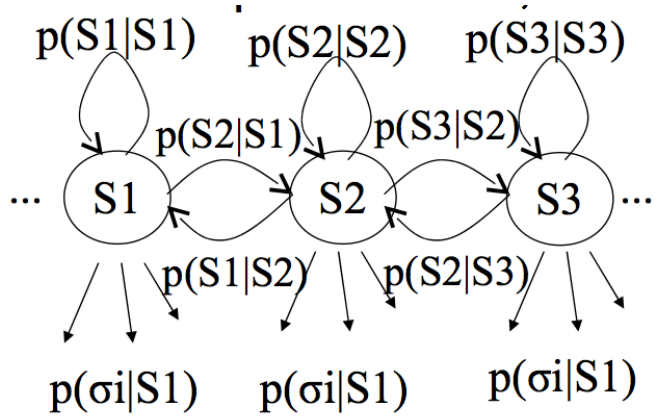

```
java -mx1600M -jar ChromHMM.jar LearnModel
SAMPLEDATA_HG18 OUTPUTSAMPLE 10 hg18
```

Hidden Markov Models (HMMs)

Steven Salzberg
JHU

Whats Hidden?

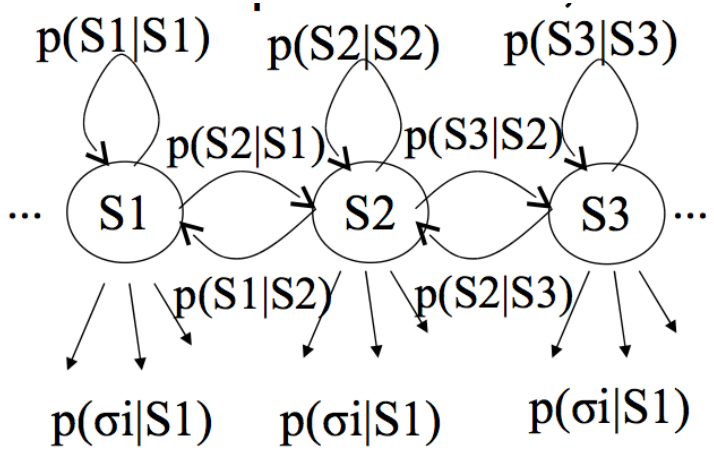
- Originally developed for time series data, were we can observe events (dice values, stock prices, weather data, neuronal spike network traffic, etc) occurring over time.
- Which "events" we see depends on which "state" the system is currently in
- We can only see the emitted symbols of an HMM (i.e., nucleotides) but not the states – they are hidden from us!
  - But we can ***infer*** the most likely hidden states of an HMM based on the given sequence of emitted symbols.



AAAGCATGCATTTAACGTGAGCACAATAGATTACA

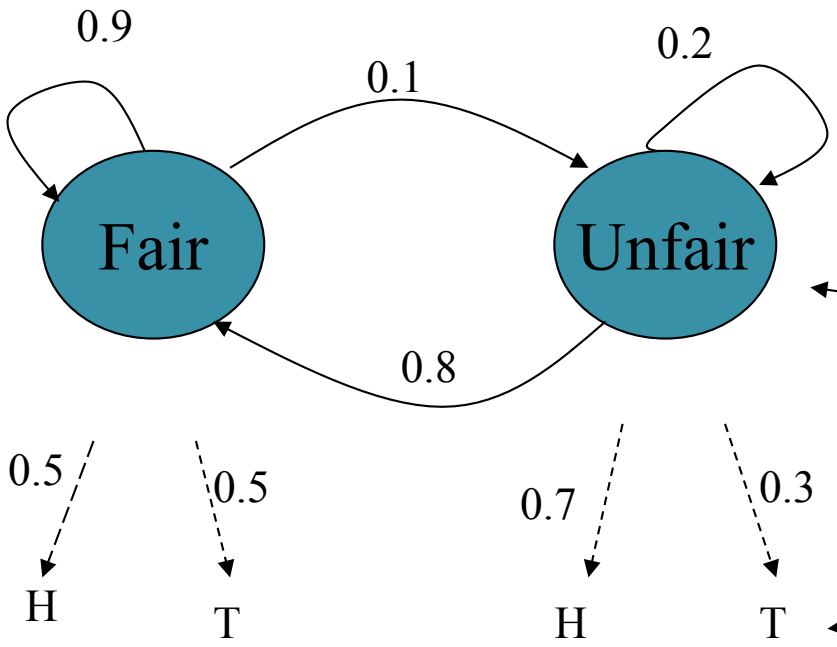
What is an HMM?

- Dynamic Bayesian Network
  - A set of states
    - {Fair, Biased} for coin tossing
    - {Gene, Not Gene} for Bacterial Gene
    - {Intergenic, Exon, Intron} for Eukaryotic Gene
    - {Promoter, Enhancer, Genetic, etc} for Regulation



  - A set of emission characters
    - E={H,T} for coin tossing
    - E={1,2,3,4,5,6} for dice tossing
    - E={A,C,G,T} for DNA
    - E={ChipSeq1, ChipSeq2, ChipSeq3, etc} for Regulation
  - State-specific emission probabilities
    - P(H | Fair) = .5, P(T | Fair) = .5, P(H | Biased) = .9, P(T | Biased) = .1
    - P(A | Gene) = .9, P(A | Not Gene) = .1 …
    - P(ChipSeq1 | Promoter) = .8, P(ChipSeq2 | Enhancer) = .5, …
  - A probability of taking a transition
    - $P(s_i=\text{Fair}|s_{i-1}=\text{Fair}) = .9$, $P(s_i=\text{Bias}|s_{i-1} = \text{Fair}) .1$
    - $P(s_i=\text{Exon} | s_{i-1}=\text{Intergenic})$, …

HMM Example - Casino Coin



State transition probs.

States

Symbol emission probs.

Observation Symbols

Observation Sequence

State Sequence

HTHHTTHHHTHTHTHHTHHHHHHTHTHH

FFFFFFUUUFFFFFFUUUUUUUFFFFFF

Motivation: Given a sequence of H & Ts, can you tell at what times the casino cheated?

Slide credit: Fatih Gelgi, Arizona State U.

Three classic HMM problems

1. **Evaluation**: given a model and an output sequence, what is the probability that the model generated that output?
2. **Decoding**: given a model and an output sequence, what is the most likely state sequence through the model that generated the output?
3. **Learning**: given a model and a set of observed sequences, how do we set the model's parameters so that it has a high probability of generating those sequences?

Three classic HMM problems

2. **Decoding**: given a model and an output sequence, what is the most likely state sequence through the model that generated the output?

- A solution to this problem gives us a way to match up an observed sequence and the states in the model.

AAAGCATGCATTTAACGAGAGCACAAGGGCTCTAATGCCG

The sequence of states is an annotation of the generated string – each nucleotide is generated in intergenic, start/stop, coding state

Three classic HMM problems

2. **Decoding**: given a model and an output sequence, what is the most likely state sequence through the model that generated the output?

- A solution to this problem gives us a way to match up an observed sequence and the states in the model.

AAAGC ATG CAT TTA ACG AGA GCA CAA GGG CTC TAA TGCCG

The sequence of states is an annotation of the generated string – each nucleotide is generated in intergenic, start/stop, coding state

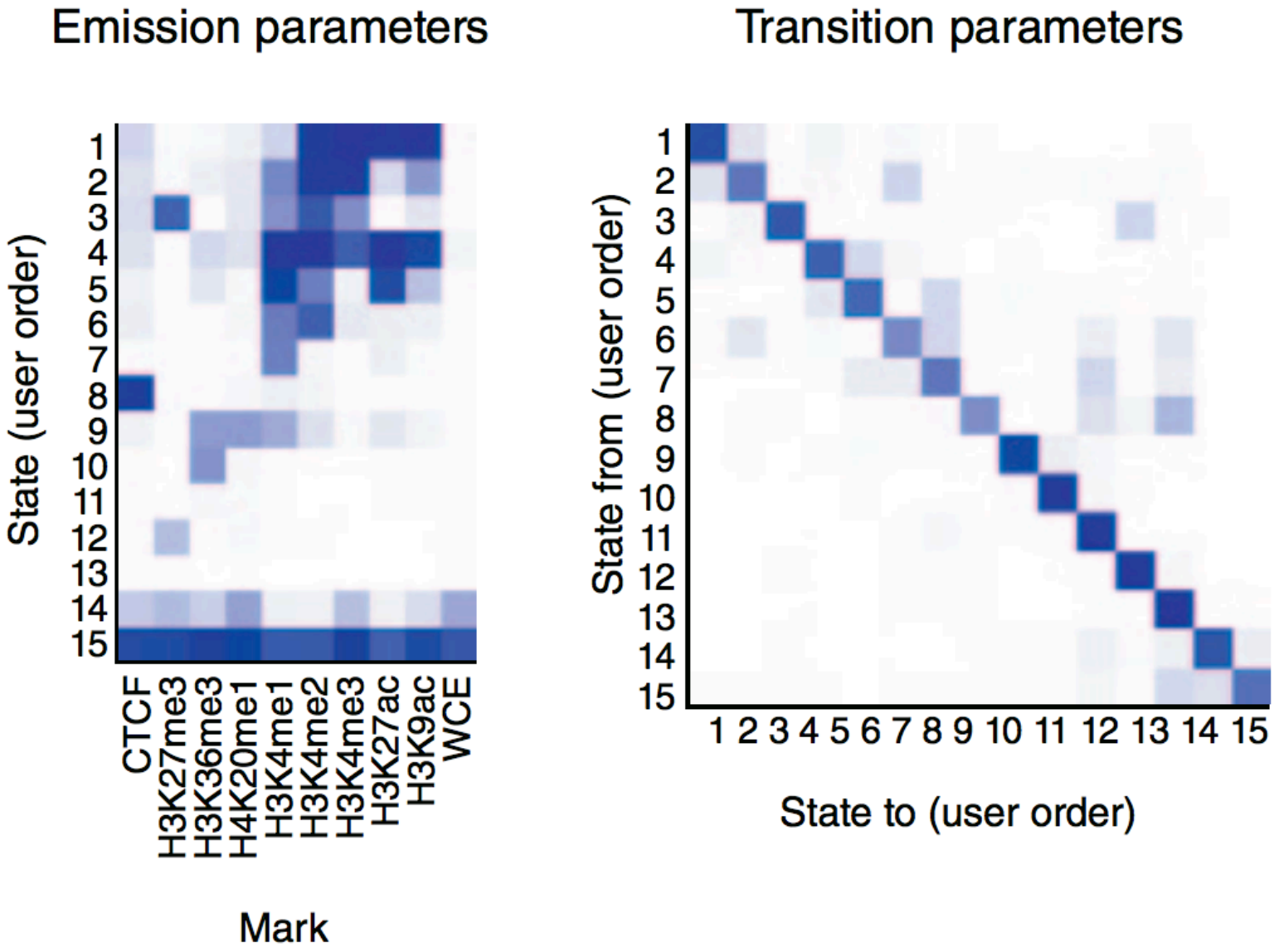
See bonus slides for more information!

Output of ChromHMM

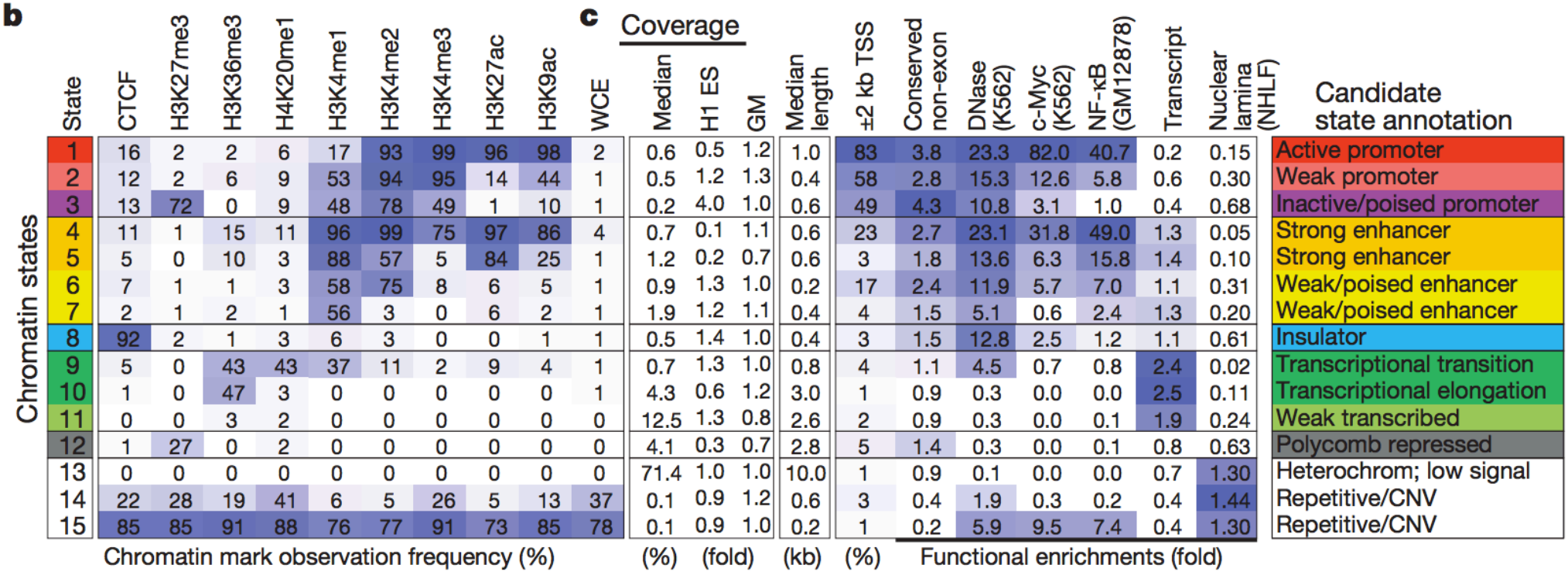
- ChromHMM generates an HTML report called **webpage_N.html** (N is the number of states used) with many useful information :

1. Model learned: transition and emission parameters
2. Enriched functional categories
3. Bed files to visualize the segmentation

Transition and emission Parameters



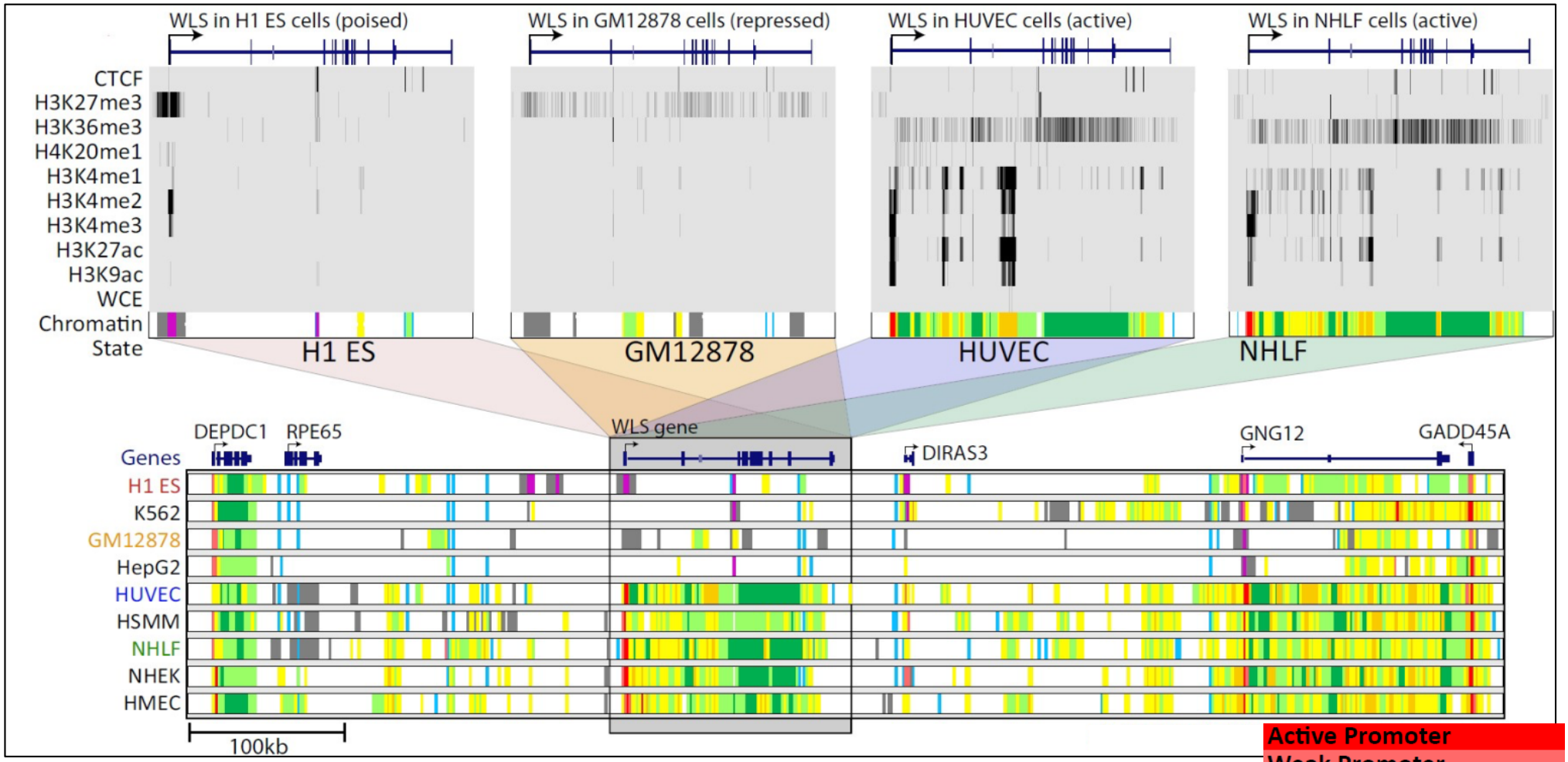
Enriched functional category



The states predicted by the HMM are ***statistical*** entities (#1 – #15)
The states we want are ***biological*** entities (Active/Weak/Poised promoter)

Investigate the properties of the statistical entities to label them with biological functions
=> Supervised learning problem ☺

Chromatin states dynamics across nine ENCODE cell types



- **Single annotation track for each cell type**
- **Summarize cell-type activity at a glance**
- **Can study 9-cell activity pattern across ↓**

Ernst et al, Nature 2011

Alternatives

Other methods are available to segment the genome in chromatin states

1. Segway: https://pmgenomics.ca/hoffmanlab/proj/segway/
   Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes J, Noble WS. 2012

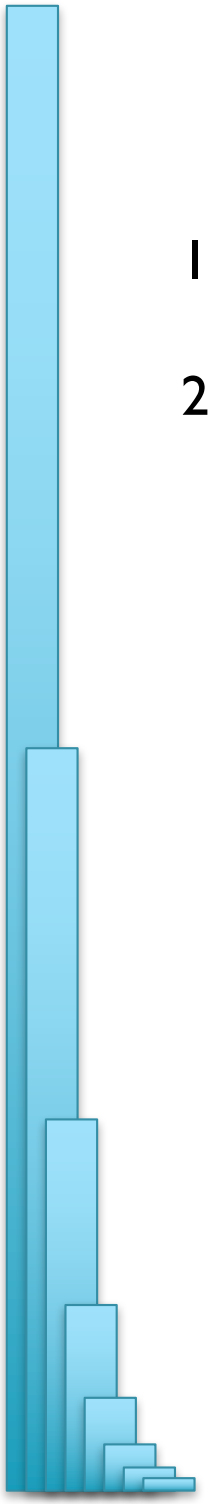
2. Spectacle: https://github.com/jiminsong/Spectacle
   Jimin Song and Kevin C Chen

3. DI-HMM: https://github.com/gcyuan/diHMM
   GC Yuan , M Kellis

Next Steps

1. Questions on project?

2. Check out the course webpage




Welcome to Applied Comparative Genomics
https://github.com/schatzlab/appliedgenomics

Questions?

Three classic HMM problems

1. **Evaluation**: given a model and an output sequence, what is the probability that the model generated that output?

- To answer this, we consider all possible paths through the model

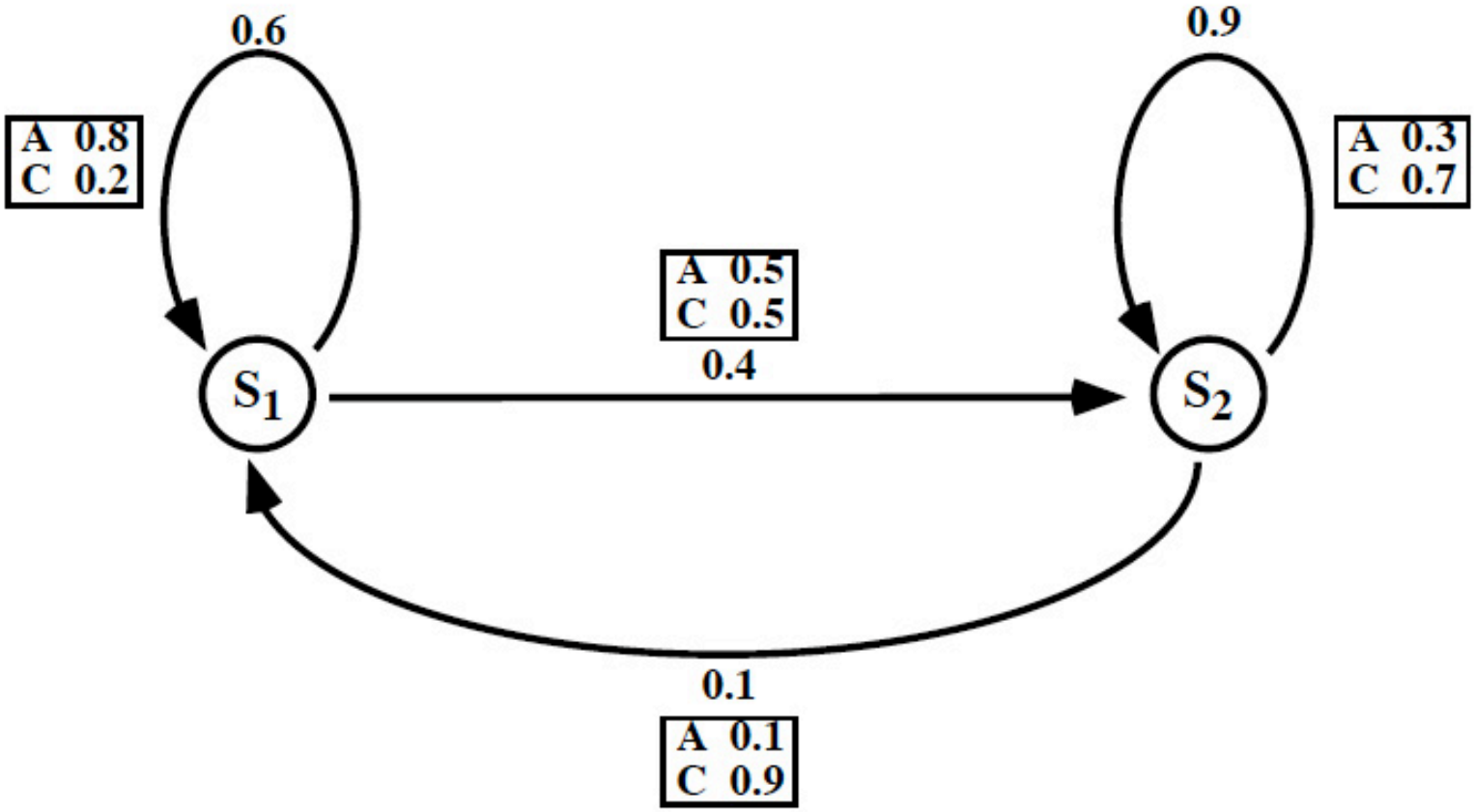
- Example: we might have a set of HMMs representing protein families -> pick the model with the best score

Solving the Evaluation problem: The Forward algorithm

- To solve the Evaluation problem (probability that the model generated the sequence), we use the HMM and the data to build a *trellis*

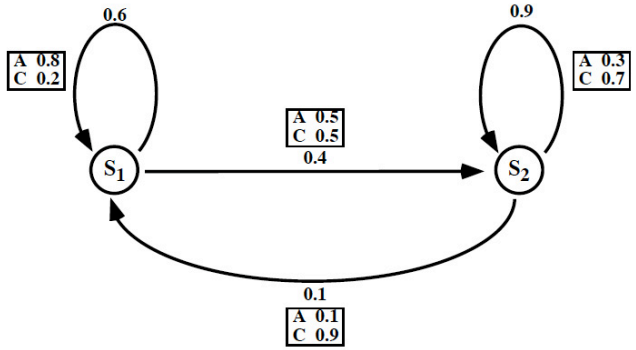
- Filling in the trellis will give tell us the probability that the HMM generated the data by finding all possible paths that could do it
  - Especially useful to evaluate from which models, a given sequence is most likely to have originated

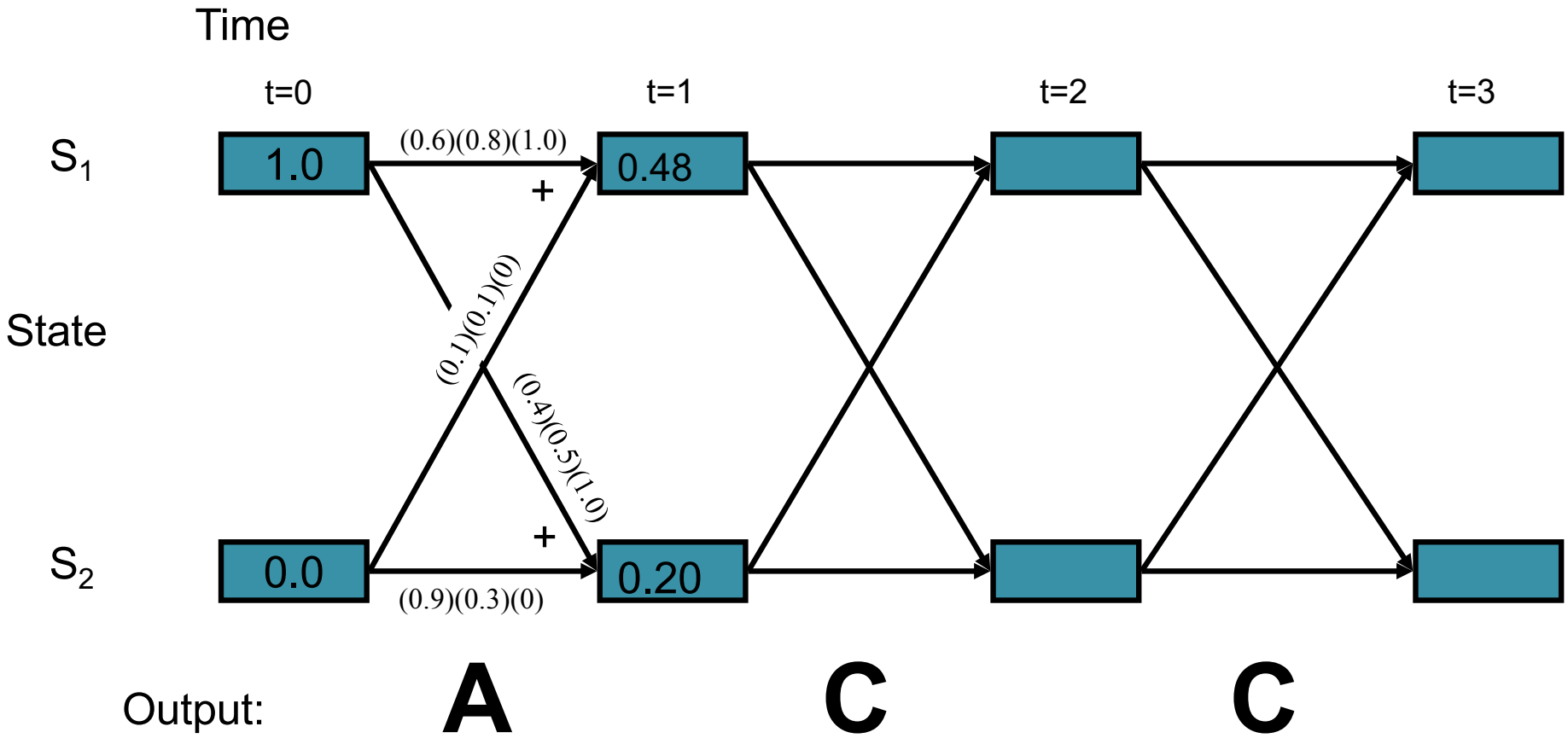
Our sample HMM



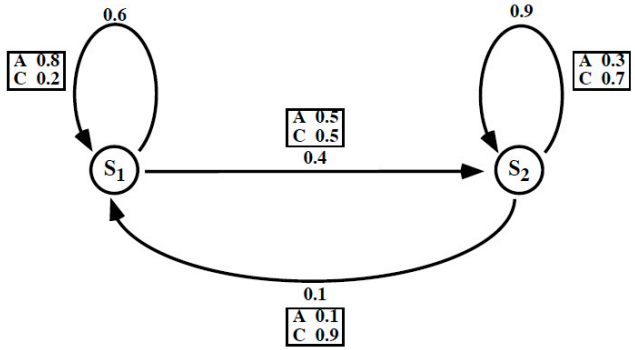
Let S_1 be initial state, S_2 be final state

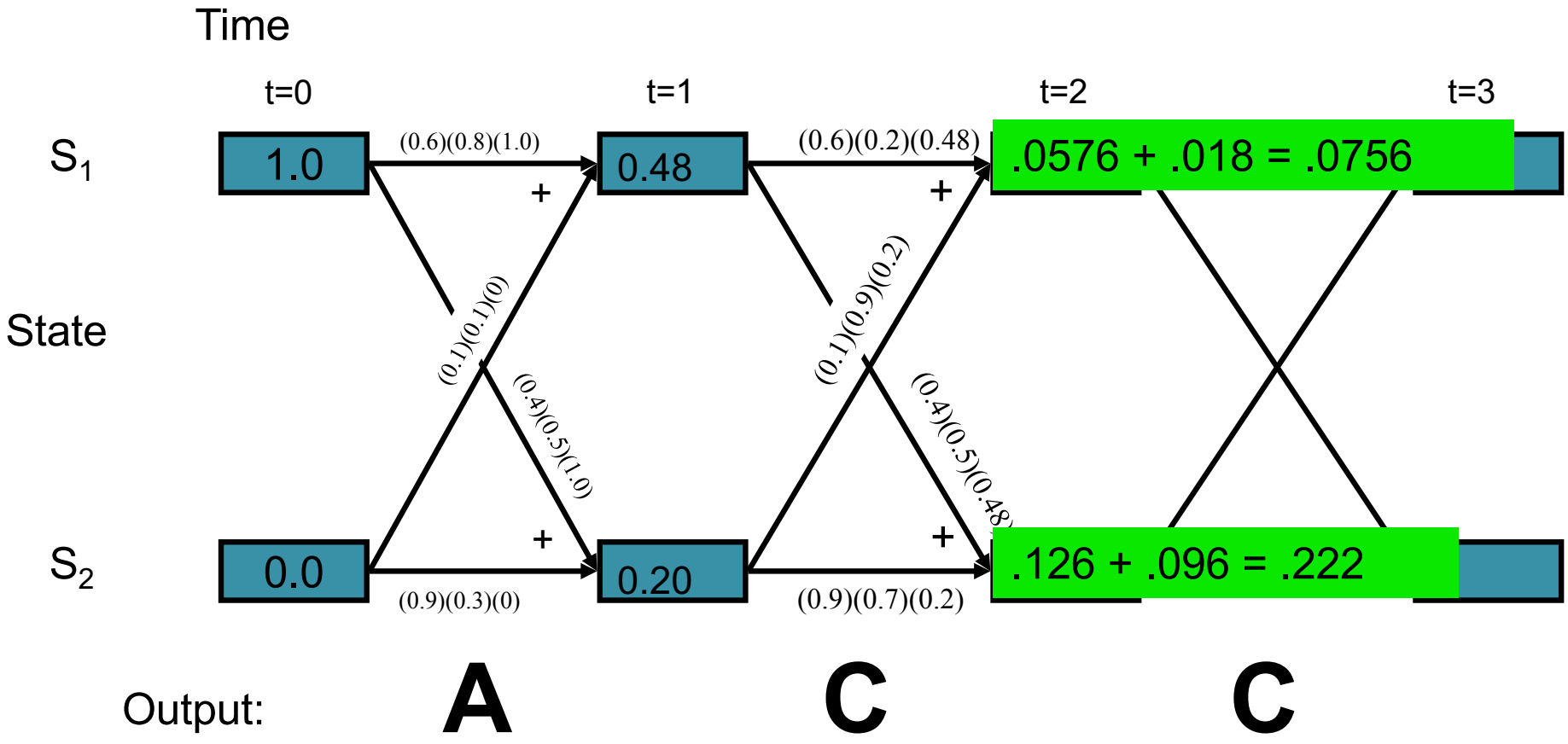
A trellis for the Forward Algorithm



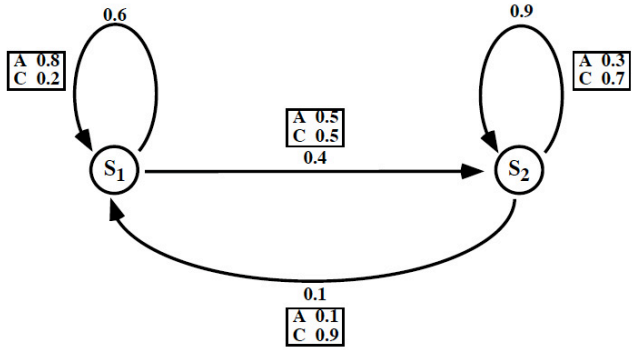


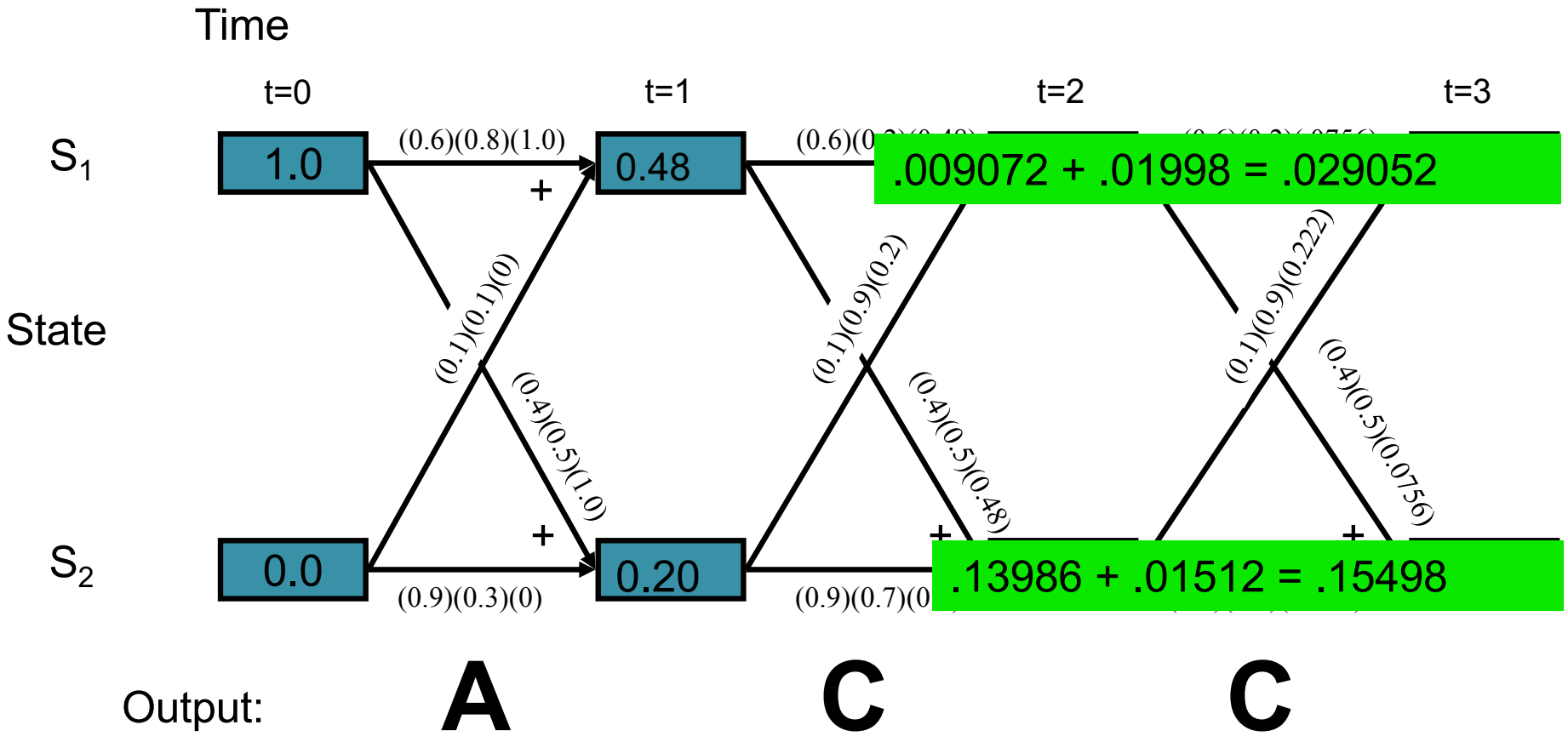
A trellis for the Forward Algorithm





A trellis for the Forward Algorithm





Probability of the model

- The Forward algorithm computes *P(y|M)*

- If we are comparing two or more models, we want the likelihood that each model generated the data: *P(M|y)*

  – Use Bayes' law: $$P(M \mid y) = \frac{P(y \mid M)P(M)}{P(y)}$$

  – Since P(y) is constant for a given input, we just need to maximize *P(y|M)P(M)*

Three classic HMM problems

2. **Decoding**: given a model and an output sequence, what is the most likely state sequence through the model that generated the output?

- A solution to this problem gives us a way to match up an observed sequence and the states in the model.

AAAGCATGCATTTAACGAGAGCACAAGGGCTCTAATGCCG

The sequence of states is an annotation of the generated string – each nucleotide is generated in intergenic, start/stop, coding state

Three classic HMM problems

2. **Decoding**: given a model and an output sequence, what is the most likely state sequence through the model that generated the output?

- A solution to this problem gives us a way to match up an observed sequence and the states in the model.

AAAGC ATG CAT TTA ACG AGA GCA CAA GGG CTC TAA TGCCG

The sequence of states is an annotation of the generated string – each nucleotide is generated in intergenic, start/stop, coding state

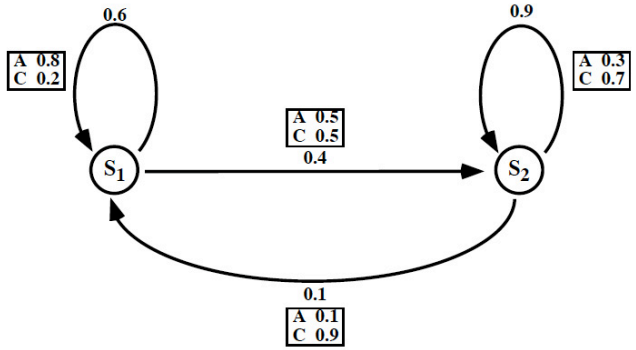
Solving the Decoding Problem: The Viterbi algorithm

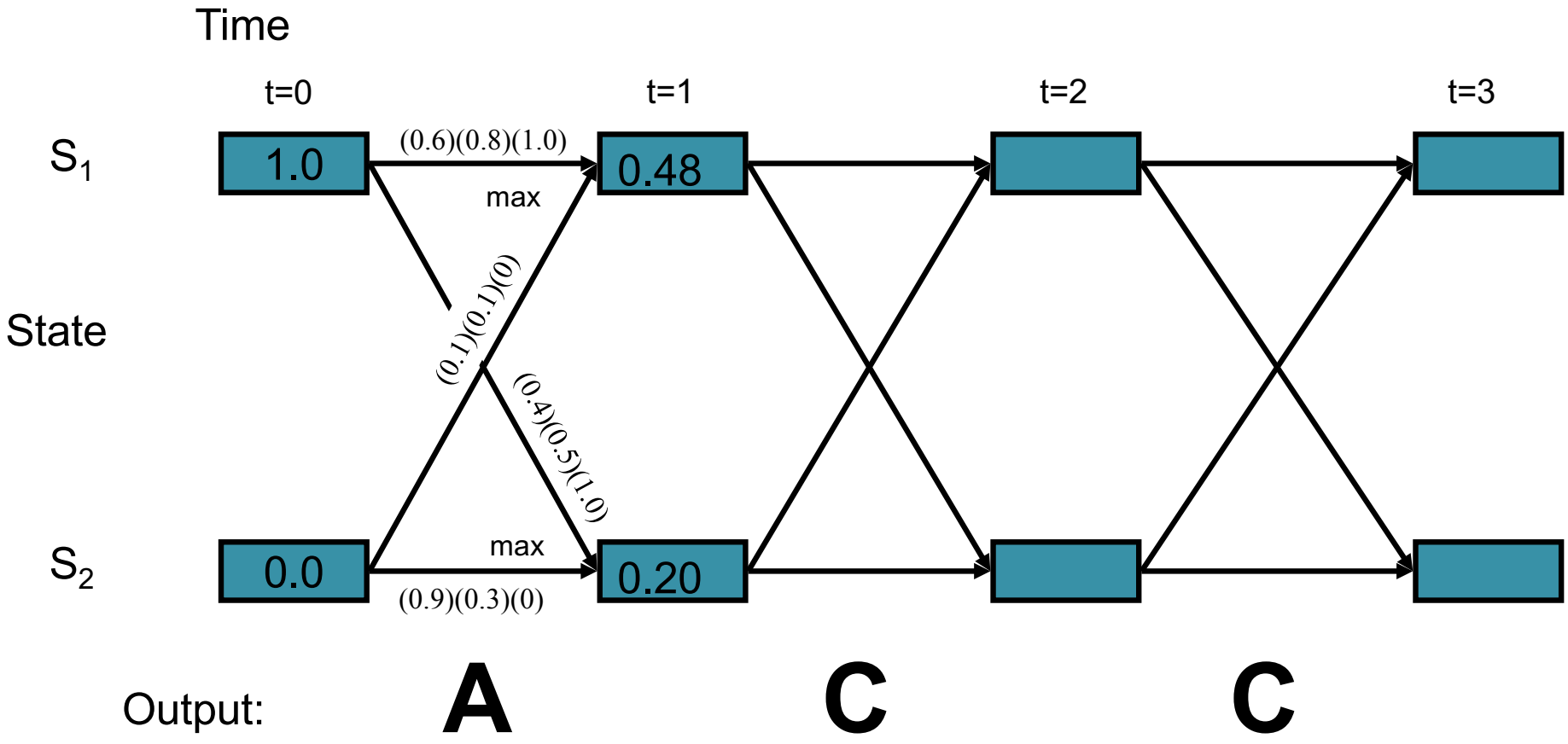
- To solve the decoding problem (find the most likely sequence of states), we evaluate the Viterbi algorithm

$$V_i(t) = \begin{cases} 0 & : \quad t = 0 \wedge i \neq S_I \\ 1 & : \quad t = 0 \wedge i = S_I \\ \max V_j(t-1) a_{ji} b_{ji}(y) & : \quad t > 0 \end{cases}$$

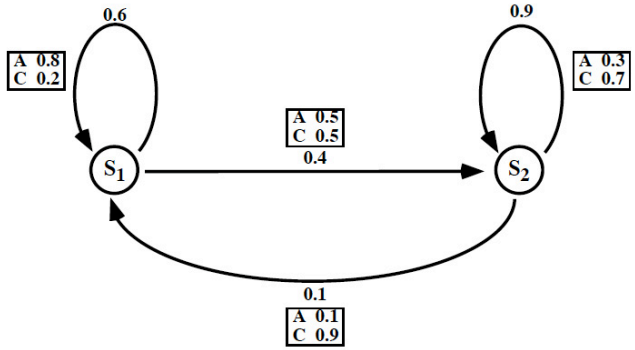
Where $V_i(t)$ is the probability that the HMM is in state i after generating the sequence $y_1, y_2, \ldots, y_t$, following the *most probable path* in the HMM

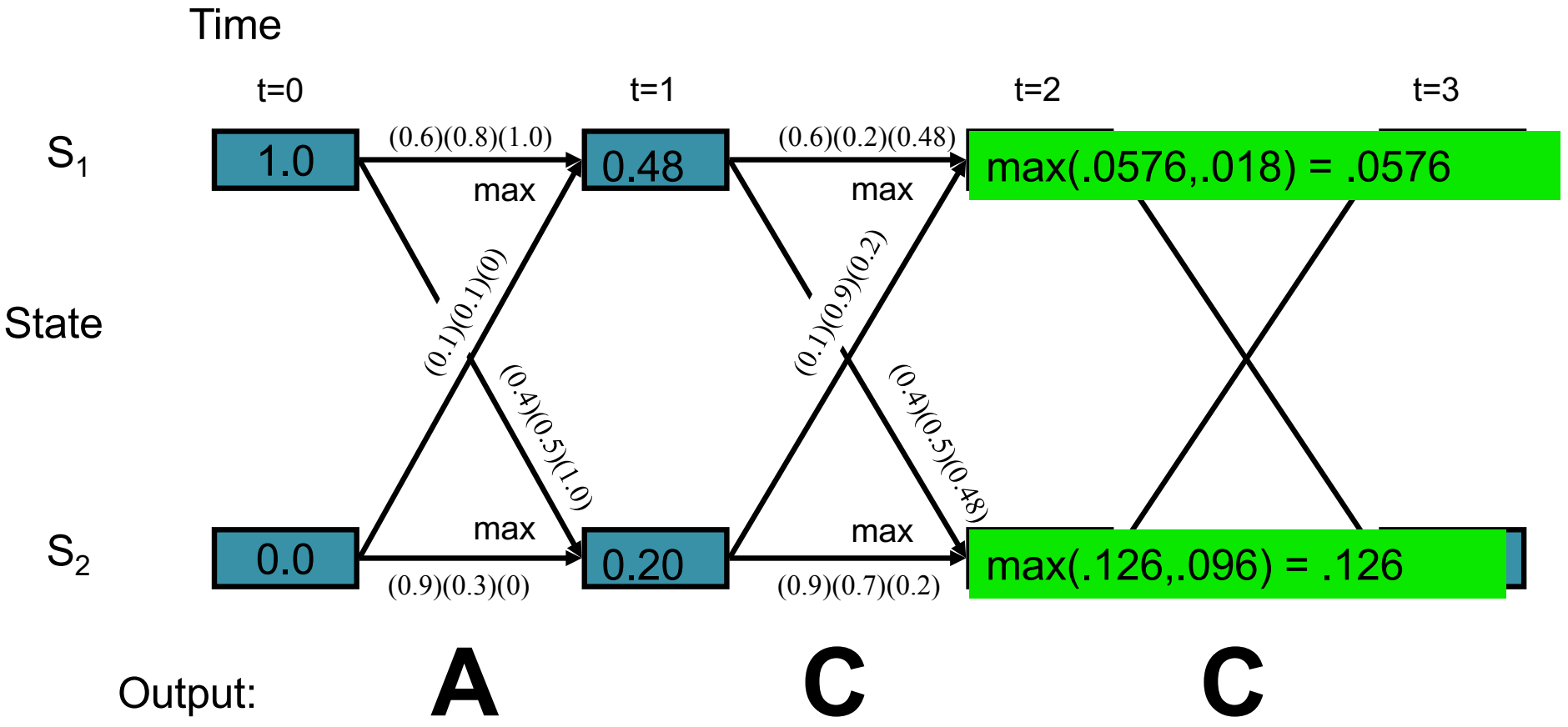
A trellis for the Viterbi Algorithm



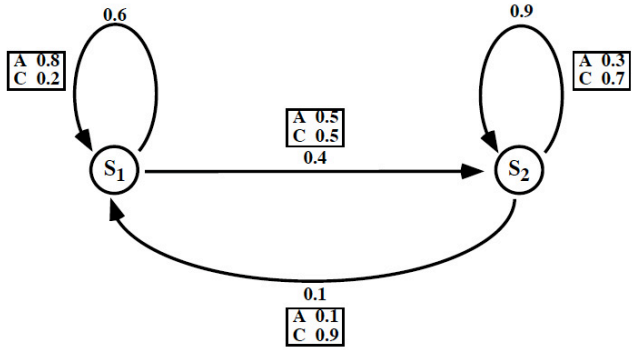


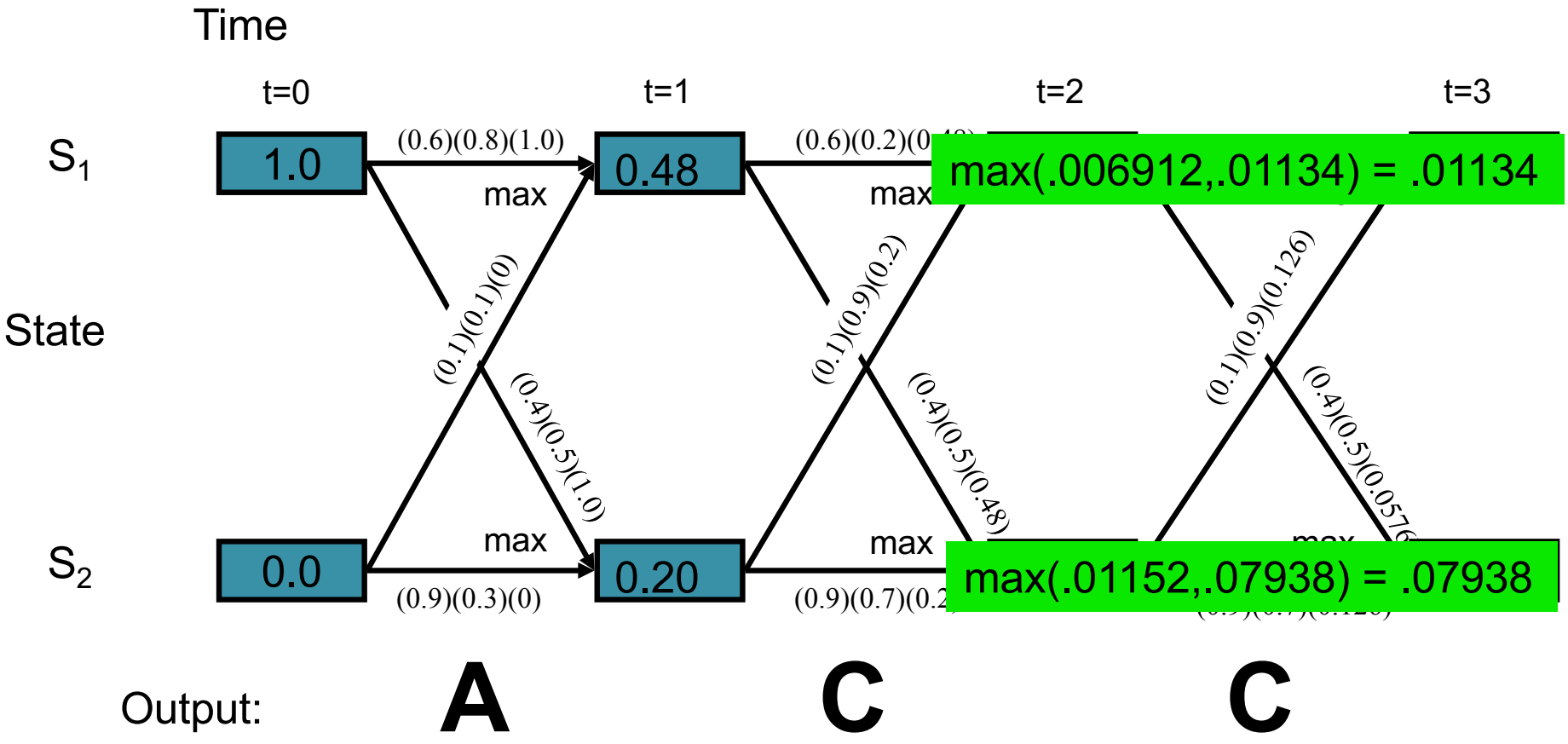
A trellis for the Viterbi Algorithm



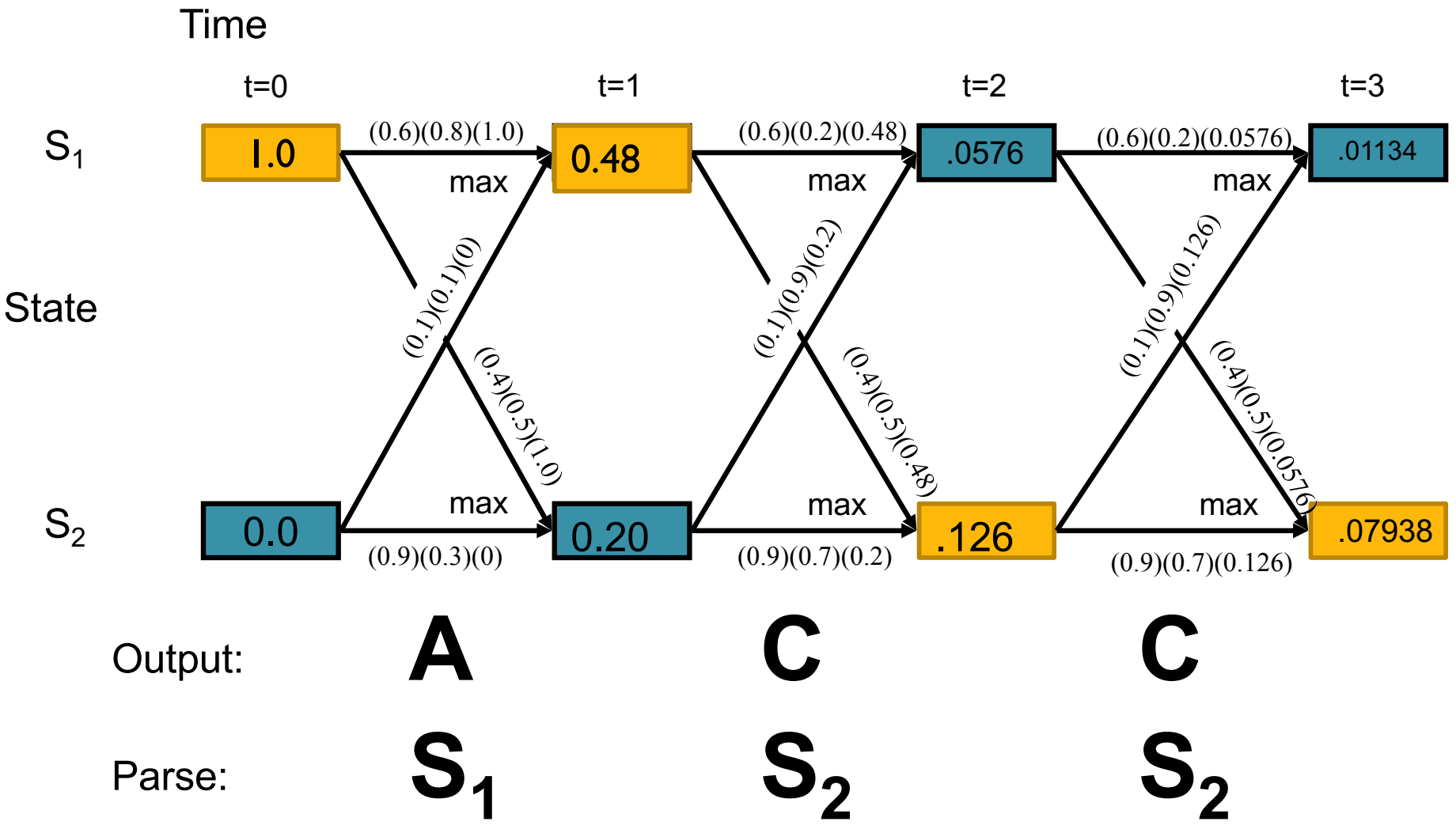


A trellis for the Viterbi Algorithm





A trellis for the Viterbi Algorithm



Three classic HMM problems

3. **Learning**: given a model and a set of observed sequences, how do we set the model's parameters so that it has a high probability of generating those sequences?

- This is perhaps the most important, and most difficult problem.

- A solution to this problem allows us to determine all the probabilities in an HMMs by using an ensemble of training data

Learning in HMMs: The E-M algorithm

- The learning algorithm is called "Expectation-Maximization" or E-M
  - Also called the Forward-Backward algorithm
  - Also called the Baum-Welch algorithm

- In order to learn the parameters in an "empty" HMM, we need:
  - The topology of the HMM
  - Data - the more the better

➔ Covered in Data Analysis Class