6.047/6.878/HST.507
Computational Biology: Genomes, Networks, Evolution

Lecture 11 - Epigenomics

read mapping – peak calling – multivariate HMMs

Module III: Epigenomics and gene regulation

- Computational Foundations
  - L10: Gibbs Sampling: between EM and Viterbi training
  - L11: Rapid linear-time sub-string matching
  - L11: Multivariate HMMs
  - L12: Post-transcriptional regulation
- Biological frontiers:
  - L10: Regulatory motif discovery, TF binding
  - L11: Epigenomics, chromatin states, differentiation
  - L12: Post-transcriptional regulation

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

(Future: Chromatin states to interpret disease-associated variants)

One Genome – Many Cell Types

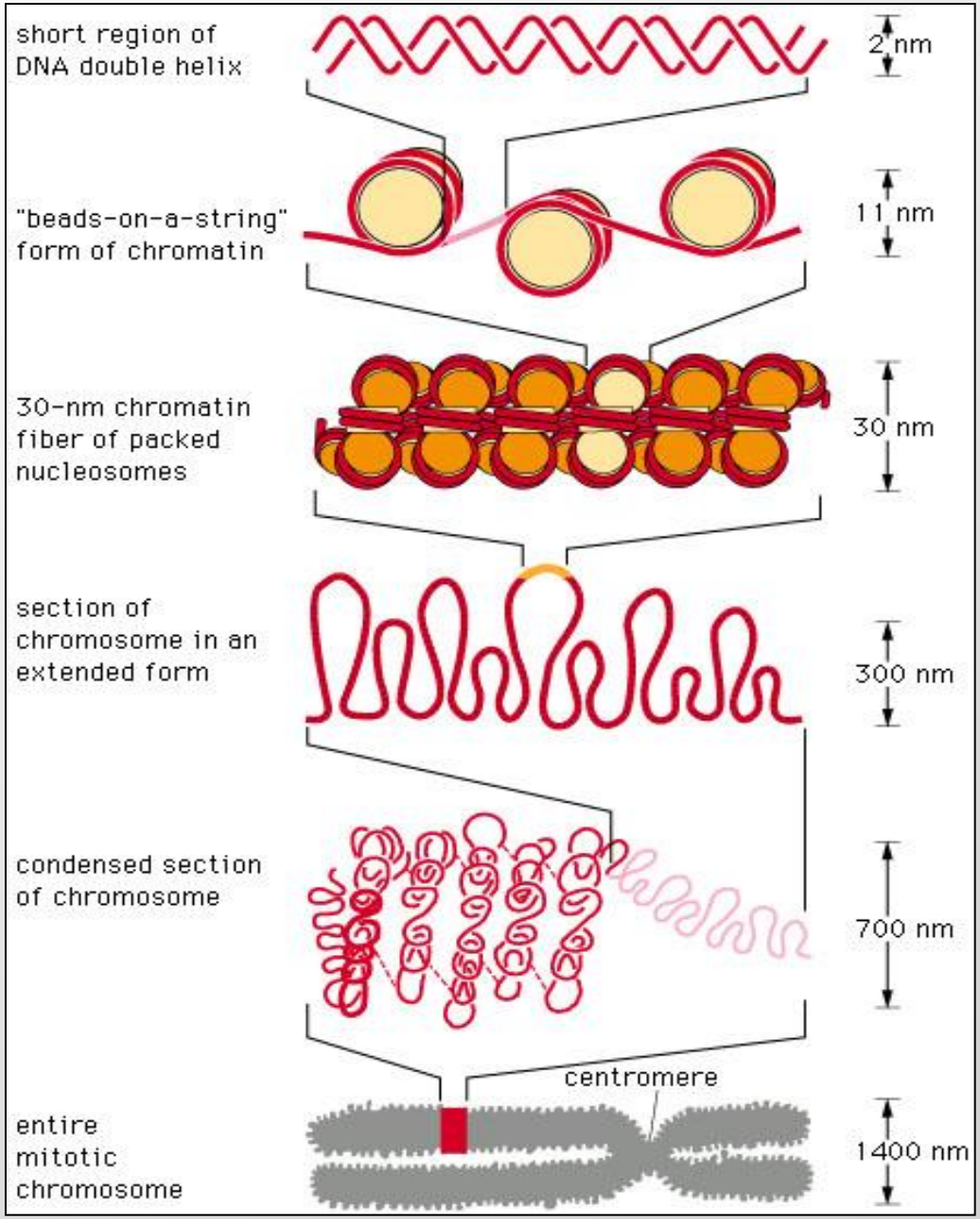
ACCAGTTACGACGGTCA
GGGTACTGATACCCCAA
ACCGTTGACCGCATTTA
CAGACGGGGTTTGGGTT
TTGCCCCACACAGGTAC
GTTAGCTACTGGTTTAG
CAATTTACCGTTACAAC
GTTTACAGGGTTACGGT
TGGGATTTGAAAAAAAG
TTTGAGTTGGTTTTTTC
ACGGTAGAACGTACCGT
TACCAGTA

Images of skin, heart, a red blood cell, and a human brain removed due to copyright restrictions.

Image Source wikipedia

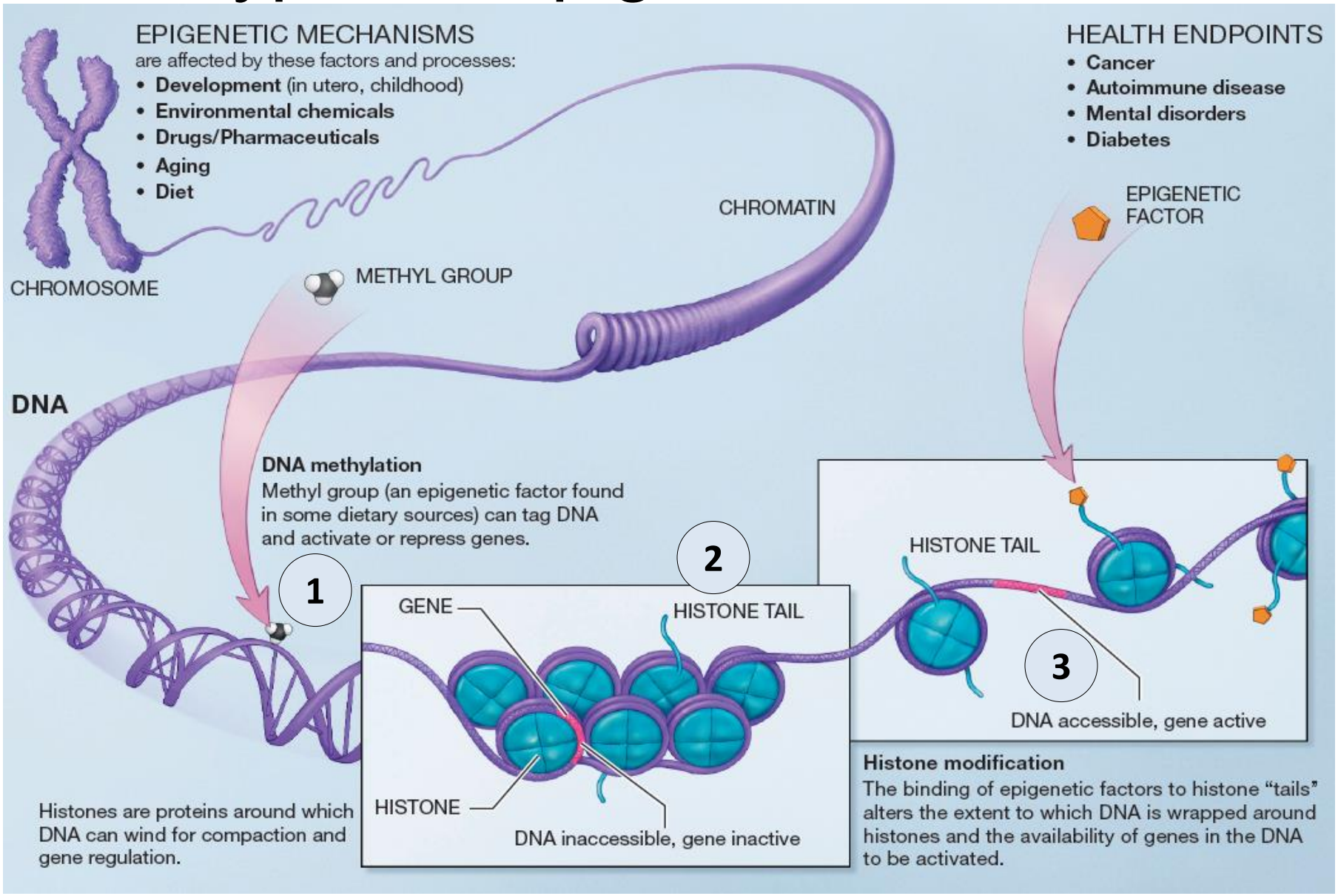
DNA packaging

- Why packaging
  - DNA is very long
  - Cell is very small
- Compression
  - Chromosome is 50,000 times shorter than extended DNA
- Using the DNA
  - Before a piece of DNA is used for anything, this compact structure must open locally
- Now emerging:
  - Role of accessibility
  - State in chromatin itself
  - Role of 3D interactions



Source: Alberts, Bruce et al. "Molecular Biology of the Cell." (2014).

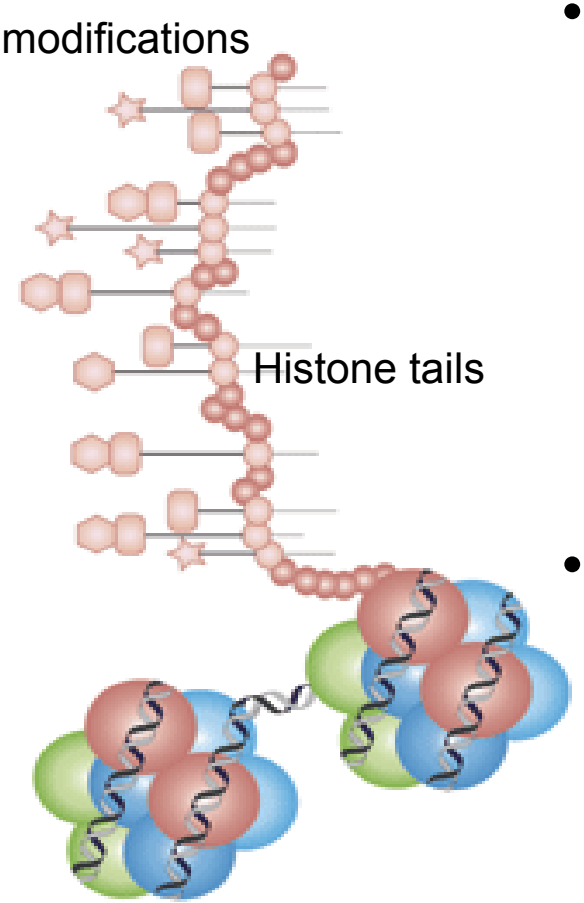
Three types of epigenetic modifications



Courtesy of National Institutes of Health. Image in the public domain.

Image source: http://nihroadmap.nih.gov/epigenomics/

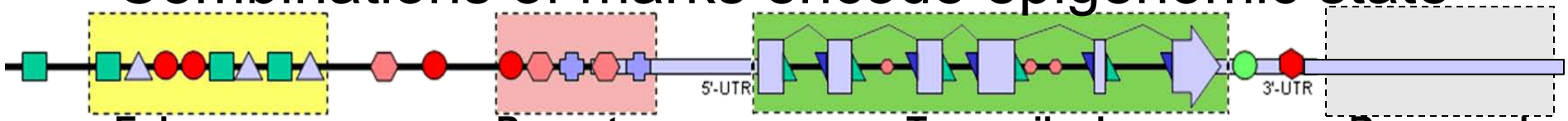
100s of histone tail modifications



DNA wrapped around histone proteins

- 100+ different histone modifications
  - Histone protein → H3/H4/H2A/H2B
  - AA residue → Lysine4(K4)/K36…
  - Chemical modification → Met/Pho/Ubi
  - Number → Me-Me-Me(me3)
  - Shorthand: H3K4me3, H2BK5ac
- In addition:
  - DNA modifications
  - Methyl-C in CpG / Methyl-Adenosine
  - Nucleosome positioning
  - DNA accessibility
- The constant struggle of gene regulation
  - TF/histone/nucleo/GFs/Chrom compete

Combinations of marks encode epigenomic state

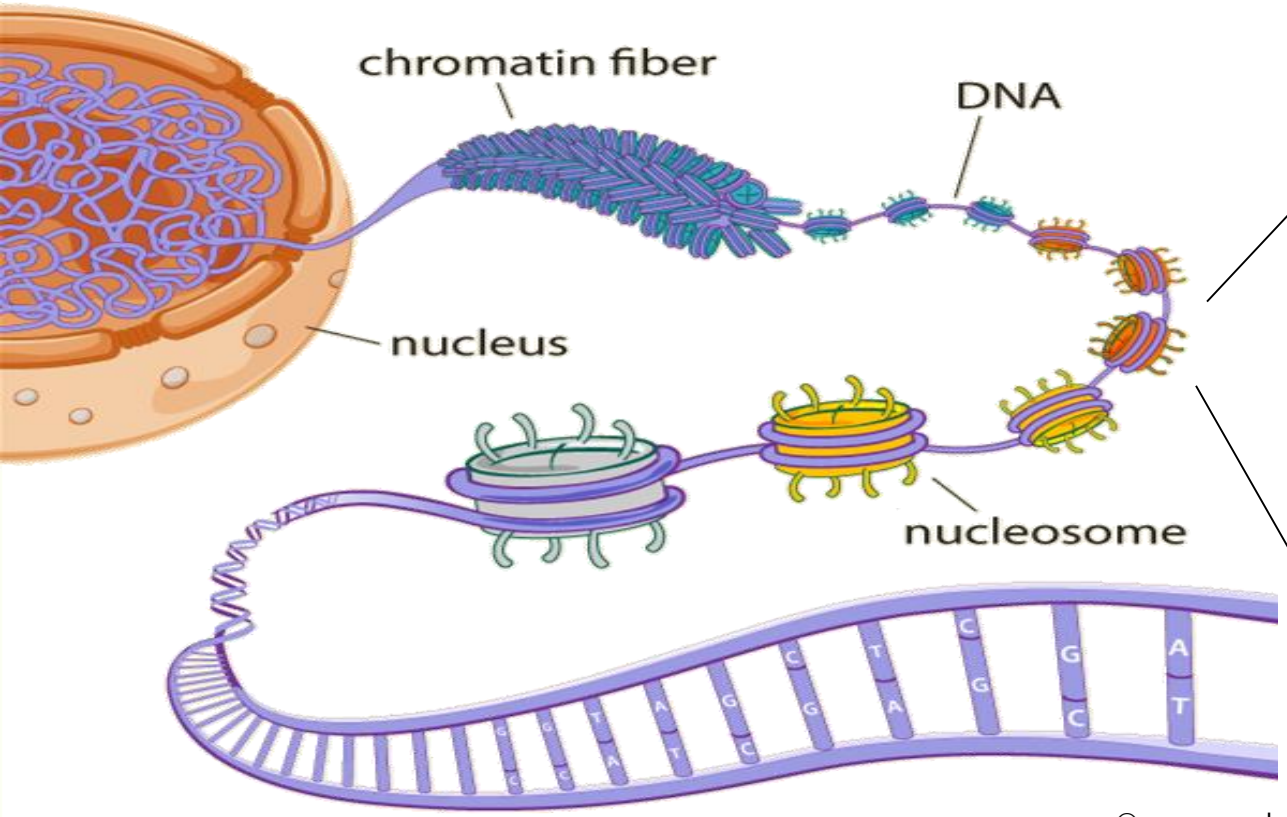


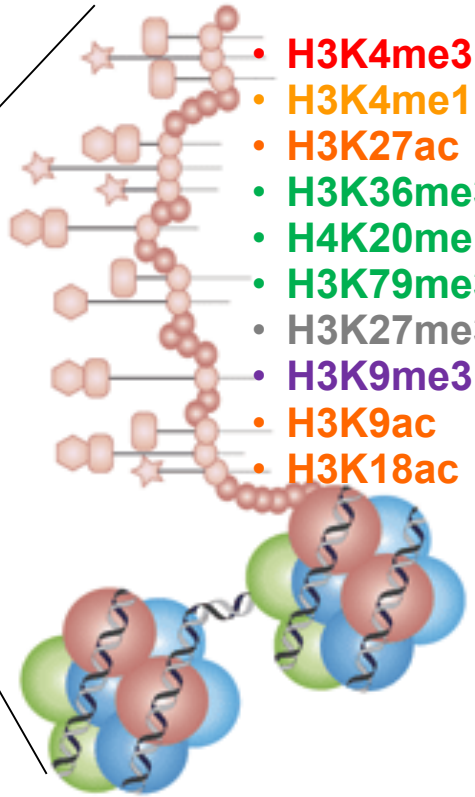
Enhancers
- H3K4me1
- H3K27ac
- DNase

Promoters
- H3K4me3
- H3K9ac
- DNase

Transcribed
- H3K36me3
- H3K79me2
- H4K20me1

Repressed
- H3K9me3
- H3K27me3
- DNAmethyl

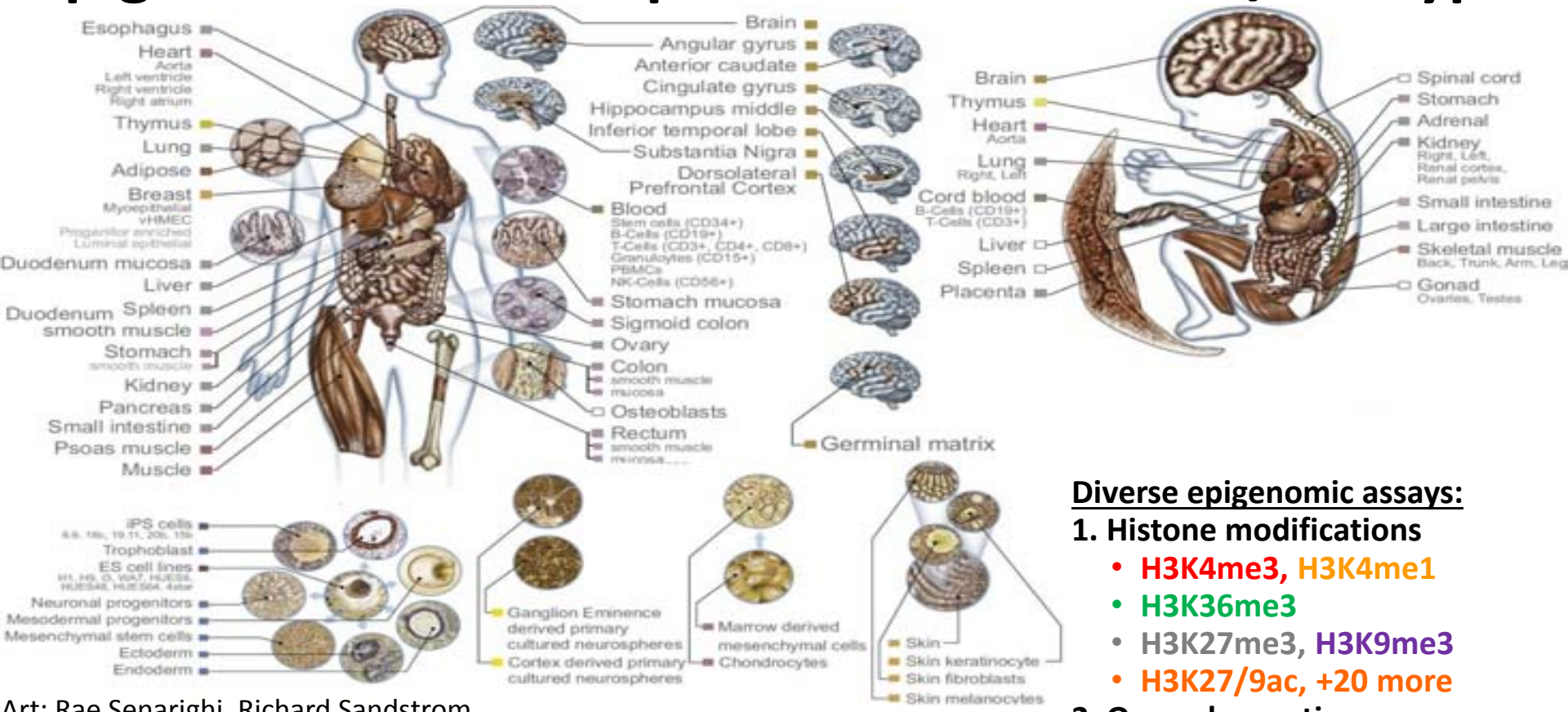




Courtesy of Broad Communications. Used with permission.

- 100s of known modifications, many new still emerging
- Systematic mapping using ChIP-, Bisulfite-, DNase-Seq

Epigenomics Roadmap across 100+ tissues/cell types



Art: Rae Senarighi, Richard Sandstrom

Courtesy of NIH Roadmap Epigenomics Mapping Consortium. Used with permission.

Diverse epigenomic assays:

1. **Histone modifications**
   - **H3K4me3, H3K4me1**
   - **H3K36me3**
   - **H3K27me3, H3K9me3**
   - **H3K27/9ac, +20 more**
2. **Open chromatin**:
   - **DNase**
3. **DNA methylation**:
   - **WGBS, RRBS, MRE/MeDIP**
4. **Gene expression**
   - **RNA-seq, Exon Arrays**

Diverse tissues and cells:

1. **Adult tissues and cells** (brain, muscle, heart, digestive, skin, adipose, lung, blood…)
2. **Fetal tissues** (brain, skeletal muscle, heart, digestive, lung, cord blood…)
3. **ES cells, iPS, differentiated cells** (meso/endo/ectoderm, neural, mesench, trophobl)

Ongoing epigenomic mapping projects

- Mapping multiple modifications
- In multiple cell types
- In multiple individuals
- In multiple species
- In multiple conditions
- With multiple antibodies
- Across the whole genome

- First wave published
- Lots more in pipeline
- Time for analysis!

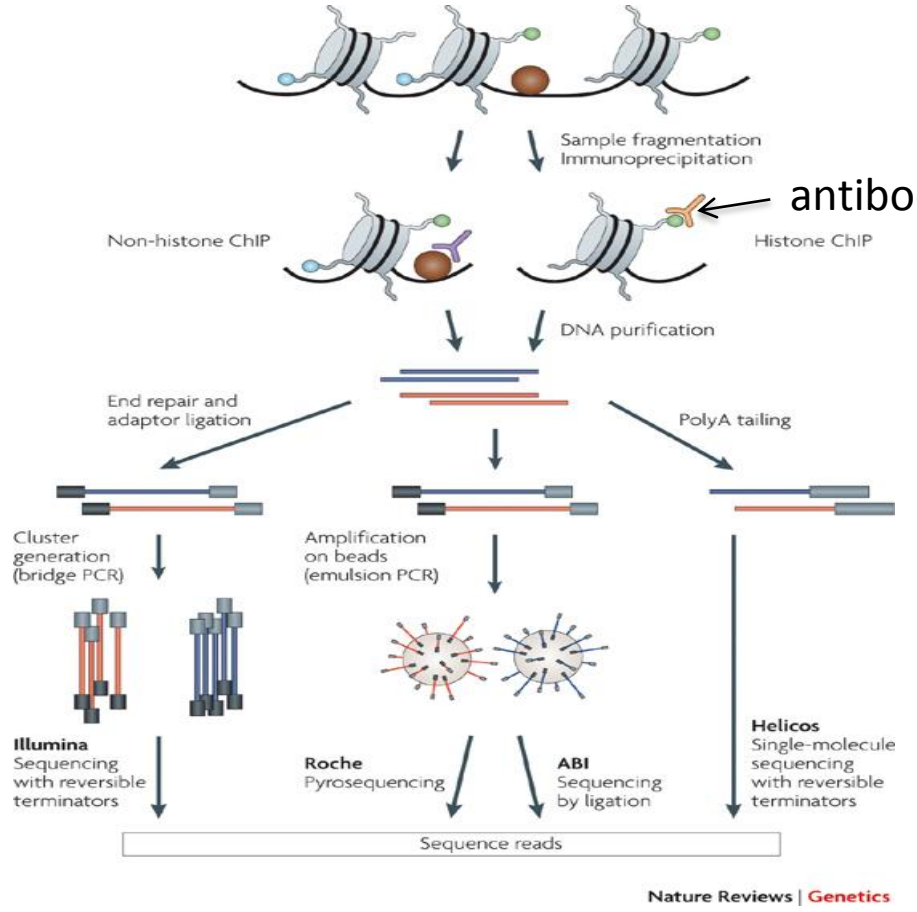
Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

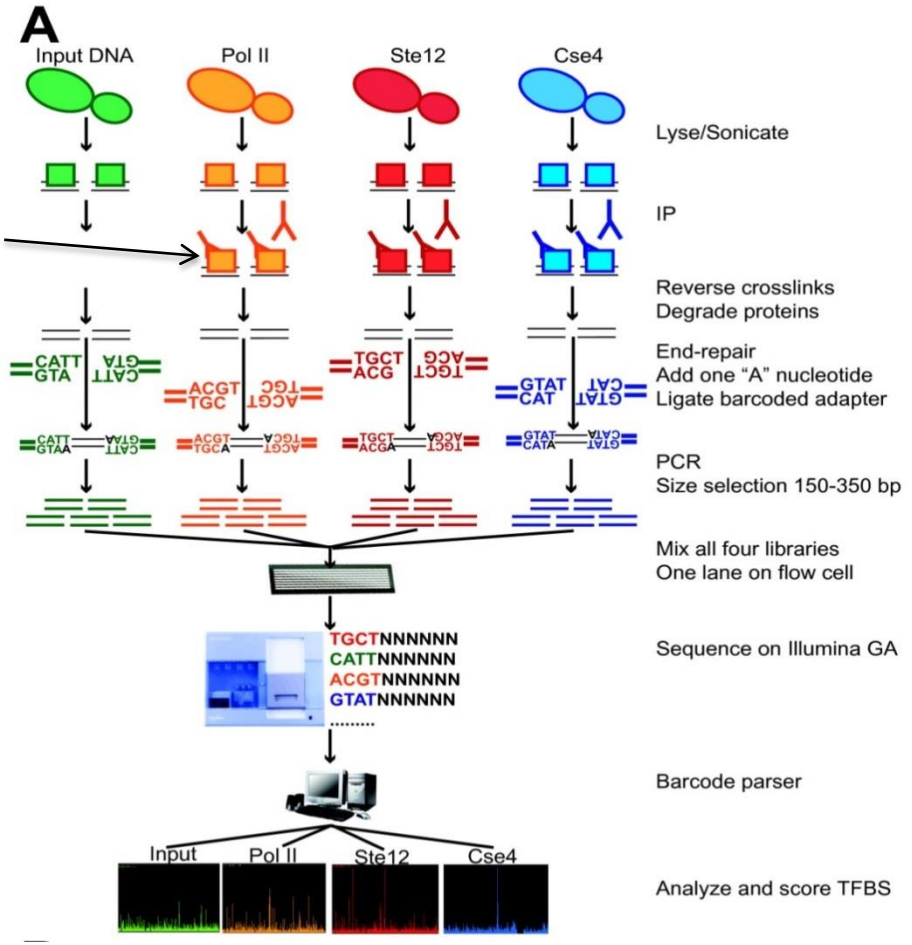
(Future: Chromatin states to interpret disease-associated variants)

ChIP-seq review

(Chromatin immunoprecipitation followed by sequencing)



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Park, Peter J. "ChIP–seq: advantages and challenges of a maturing technology." Nature Reviews Genetics 10, no. 10 (2009): 669-680.



Source: Lefrançois, Philippe et al. "Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing." BMC genomics 10, no. 1 (2009): 1.

Bar-coded multiplexed sequencing

ChIP-chip and ChIP-Seq technology overview

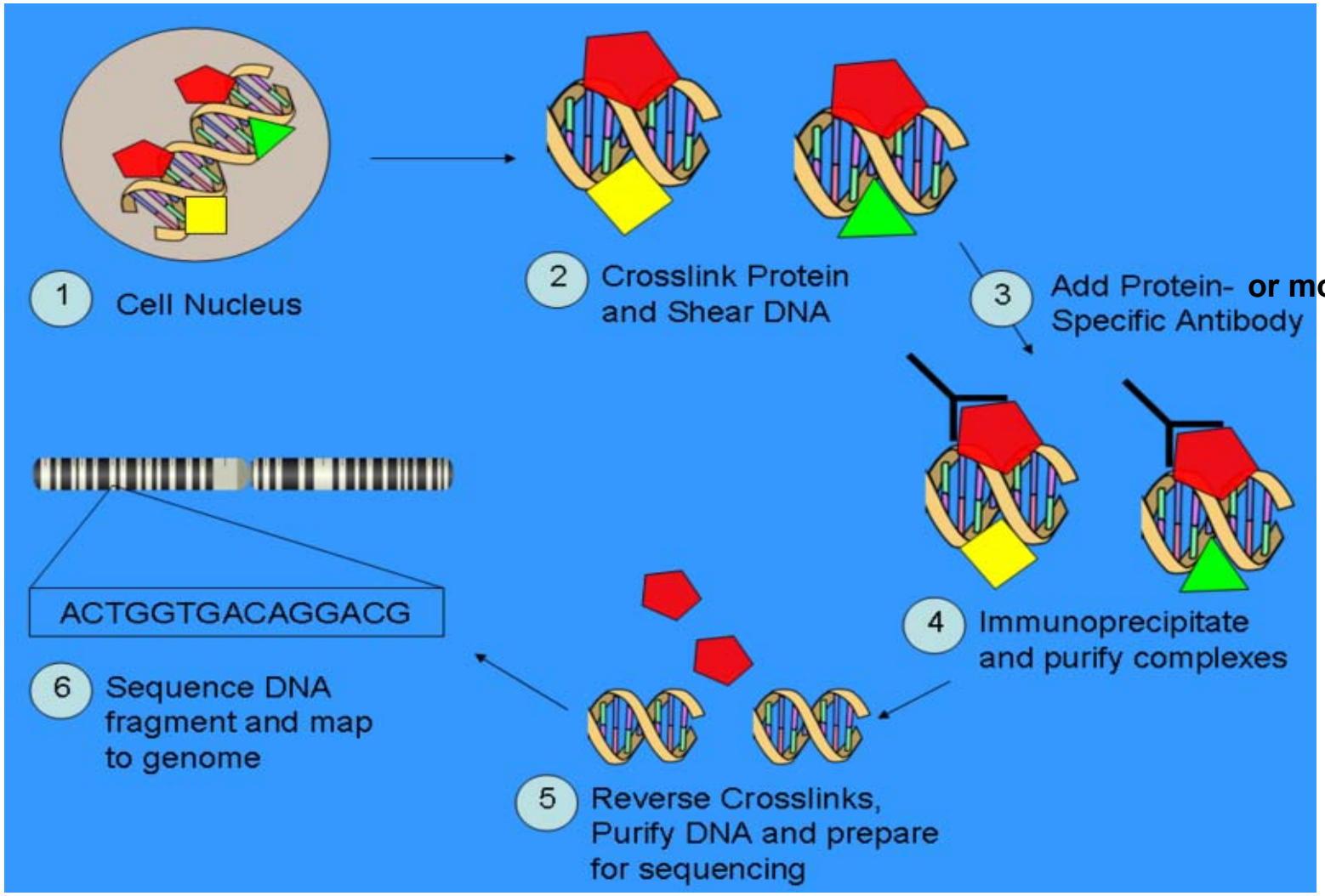


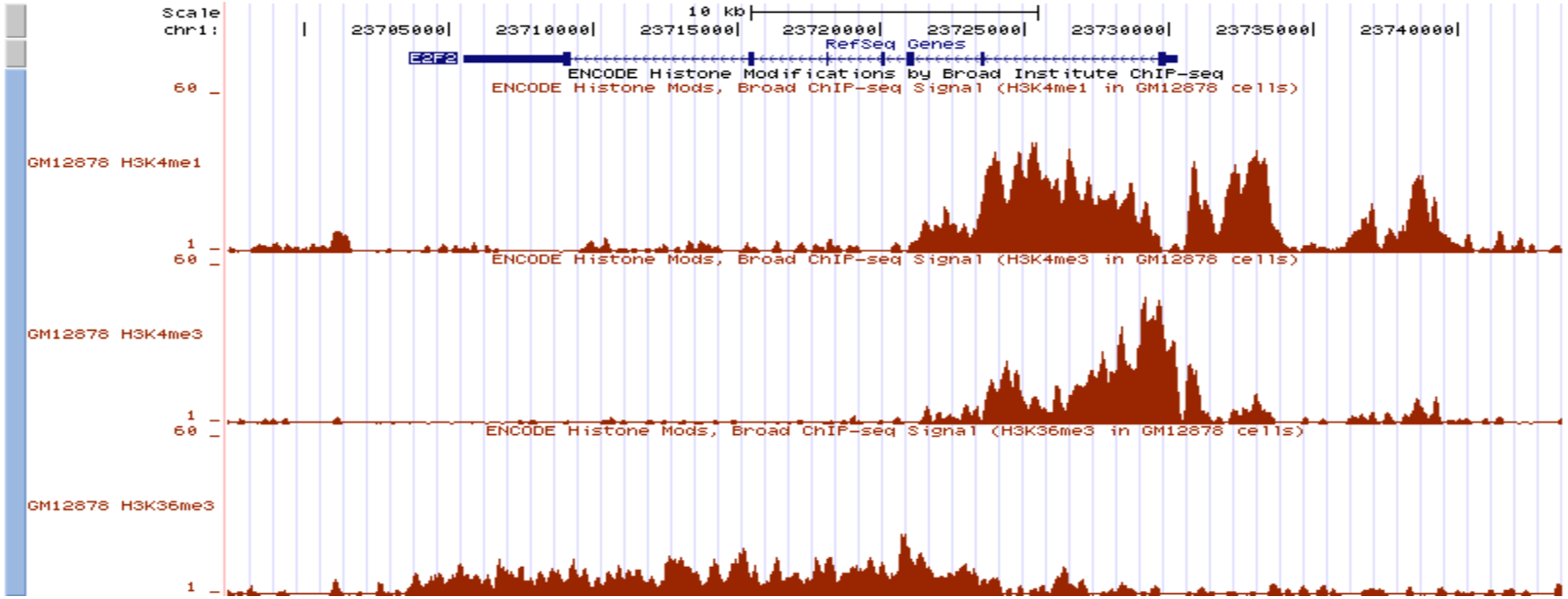
Image adapted from Wikipedia

Modification-specific antibodies → Chromatin Immuno-Precipitation

followed by: ChIP-chip: array hybridization

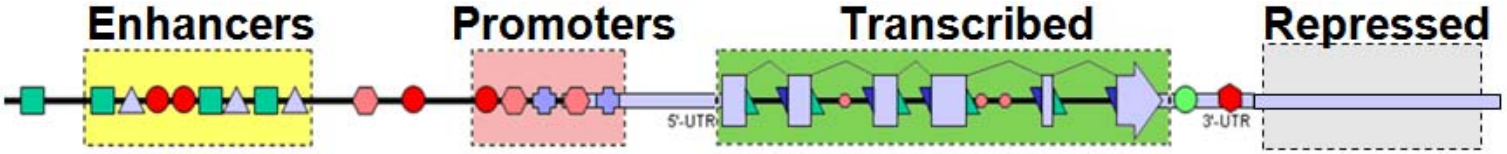
ChIP-Seq: Massively Parallel Next-gen Sequencing

ChIP-Seq Histone Modifications: What the raw data looks like



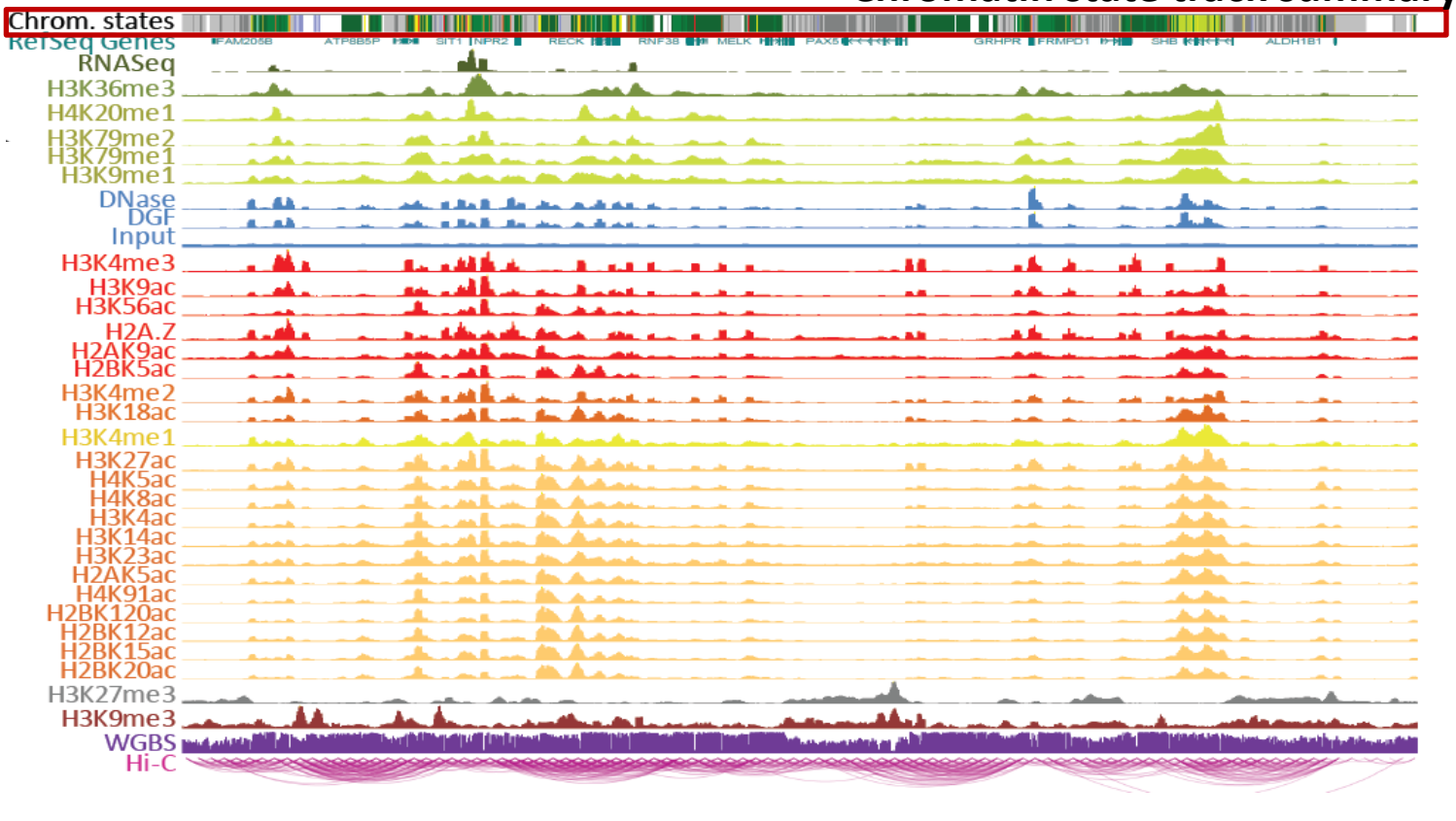
- Each sequence tag is 30 base pairs long
- Tags are mapped to unique positions in the ~3 billion base reference genome
- Number of reads depends on sequencing depth. Typically on the order of 10 million mapped reads.

Summarize multiple marks into chromatin states



Chromatin state track summary

30+ epigenomics marks



WashU Epigenome Browser

ChromHMM: multi-variate hidden Markov model

Mapping millions of short reads to the genome

Traditional Hashing Schemes

Burrows-Wheeler Transform (BWT)

Mapping Reads to the Genome

- Assign reads to best matching location in reference genome
- 10,000,000s of reads, ~30 bases long
- Example: CAGGGCTGATTGAGGACATTCATCACG
- Allow mismatches: sequencing errors, or SNPs
- Algorithmic and memory efficiency is critical

17

How would you do it:

- L2: Sequence alignment: O(m*n)
- L3: Hashing / BLAST: O(m+n)
  - Solution until 2008 (e.g. MAQ, Li et al, GR 2008)
- Other advanced algorithms:
  - Linear-time string matching: O(m+n). L3 addendum
  - Suffix trees and suffix arrays: O(m). L13 addendum
- Challenge: memory requirements
  - Hash table, suffix tree/array require O(m*n) space
- Today: Burrows-Wheeler transformation O(m)
  - Ultrafast/memory efficient. New norm since 2009.
  - Introduced in: Bowtie (Langmead GB 2009).

18

Second Generation Mappers have Leveraged the Burrows Wheeler Transformation

Software

Open Access

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

Correspondence: Ben Langmead. Email: langmead@cs.umd.edu

Published: 4 March 2009

Genome Biology 2009, **10**:R25 (doi:10.1186/gb-2009-10-3-r25)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/3/R25

Received: 21 October 2008
Revised: 19 December 2008
Accepted: 4 March 2009

Abstract

Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. Multiple processor cores can be used simultaneously to achieve even greater alignment speeds. Bowtie is open source http://bowtie.cbcb.umd.edu.

"…35 times faster than Maq and 300 times faster than SOAP under the same conditions"

BIOINFORMATICS **APPLICATIONS NOTE** Vol. 24 no. 5 2008, pages 713–714 doi:10.1093/bioinformatics/btn025

Sequence analysis

SOAP: short oligonucleotide alignment program

Ruiqiang Li[1,2], Yingrui Li[1], Karsten Kristiansen[2] and Jun Wang[1,2,*]

[1]Beijing Genomics Institute at Shenzhen, Shenzhen 518083, China and [2]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, DK-5230, Denmark

Received on November 10, 2007; revised on December 20, 2007; accepted on January 14, 2008

Advance Access publication January 28, 2008

Associate Editor: Keith Crandall

ABSTRACT

Summary: We have developed a program SOAP for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences. The program is designed to handle the huge amounts of short reads generated by parallel sequencing using the new generation Illumina-Solexa sequencing technology. SOAP is compatible with numerous applications, including single-read or pair-end resequencing, small RNA discovery and mRNA tag sequence mapping. SOAP is a command-driven program, which supports multi-threaded parallel computing, and has a batch module for multiple query sets.

Availability: http://soap.genomics.org.cn

Contact: soap@genomics.org.cn

SOAP will allow either a certain number of mismatches or one continuous gap for aligning a read onto the reference sequence. The best hit of each read which has minimal number of mismatches or smaller gap will be reported. For multiple equal-best hits, the user can instruct the program to report all, or randomly report one, or disregard all of them. Since the typical read length is 25–50 bp, hits with too many mismatches are unreliable which are hard to distinguish with random matches. By default, the program will allow at most two mismatches. Between two haplotype genome sequences, occurrence of single nucleotide polymorphism is much higher than that of small insertions or deletions, so ungapped hits have precedence over gapped hits. For gapped alignment only one continuous gap with a size ranging from 1 to 3 bp is accepted, while no

BIOINFORMATICS **APPLICATIONS NOTE** Vol. 25 no. 15 2009, pages 1966–1967 doi:10.1093/bioinformatics/btp336

Sequence analysis

SOAP2: an improved ultrafast tool for short read alignment

Ruiqiang Li[1,2,†], Chang Yu[1,†], Yingrui Li[1], Tak-Wah Lam[3], Siu-Ming Yiu[3], Karsten Kristiansen[2] and Jun Wang[1,2,*]

[1]Beijing Genomics Institute at Shenzhen, Shenzhen, 518083, China, [2]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M, DK-5230, Denmark and [3]Department of Computer Science, University of Hong Kong, Hong Kong, China

Received on January 23, 2009; revised on April 27, 2009; accepted on May 24, 2009

Advance Access publication June 3, 2009

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: SOAP2 is a significantly improved version of the short oligonucleotide alignment program that both reduces computer memory usage and increases alignment speed at an unprecedented rate. We used a Burrows Wheeler Transformation (BWT) compression index to substitute the seed strategy for indexing the reference sequence in the main memory. We tested it on the whole human genome and found that this new algorithm reduced memory usage from 14.7 to 5.4 GB and improved alignment speed by 20–30 times. SOAP2 is compatible with both single- and paired-end reads. Additionally, this tool now supports multiple text and compressed file formats. A consensus builder has also been developed for consensus assembly and SNP detection from alignment of short reads on a reference genome.

Availability: http://soap.genomics.org.cn

Contact: soap@genomics.org.cn

variation map, will generate about 15 Tb of sequence using next-generation sequencing technologies. With even the fastest programs currently available, one would need ~1000 CPU months to align these short reads onto the human reference genome. Additionally, new methods are now needed to support longer reads as the existing methods were primarily designed for very short reads with typical lengths shorter than 50 bp. With improvements in sequencing chemistry and data processing algorithms, the Illumina Genome Analyzer can now generate up to 75–100 bp high-quality reads, and longer reads are expected in the near future.

Here, we have developed an improved version of SOAP, called SOAP2. The new program uses the Burrows Wheeler Transformation (BWT) compressed index instead of the seed algorithm that was used in the previous version for indexing the reference sequence in the main memory. Use of BWT substantially improved alignment speed; additionally, it significantly reduced memory usage.

Mapping short DNA sequencing reads and calling variants using mapping quality scores

Heng Li,[1] Jue Ruan,[2] and Richard Durbin[1,3]

[1]*The Wellcome Trust Sanger Institute, Hinxton CB10 1SA, United Kingdom;* [2]*Beijing Genomics Institute, Chinese Academy of Science, Beijing 100029, China*

New sequencing technologies promise a new era in the use of DNA sequence. However, some of these technologies produce very short reads, typically of a few tens of base pairs, and to use these reads effectively requires new algorithms and software. In particular, there is a major issue in efficiently aligning short reads to a reference genome and handling ambiguity or lack of accuracy in this alignment. Here we introduce the concept of *mapping quality*, a measure of the confidence that a read actually comes from the position it is aligned to by the mapping algorithm. We describe the software MAQ that can build assemblies by mapping shotgun short reads to a reference genome, using quality scores to derive genotype calls of the consensus sequence of a diploid genome, e.g., from a human sample. MAQ makes full use of mate-pair information and estimates the error probability of each read alignment. Error probabilities are also derived for the final genotype calls, using a Bayesian statistical model that incorporates the mapping qualities, error probabilities from the raw sequence quality scores, sampling of the two haplotypes, and an empirical model for correlated errors at a site. Both read mapping and genotype calling are evaluated on simulated data and real data. MAQ is accurate, efficient, versatile, and user-friendly. It is freely available at http://maq.sourceforge.net.

[Supplemental material is available online at www.genome.org. Short-read sequences have been deposited in the European Read Archive (ERA) under accession no. ERA000002 (ftp://ftp.era.ebi.ac.uk/ERA000002/).]

BIOINFORMATICS **ORIGINAL PAPER** Vol. 25 no. 14 2009, pages 1754–1760 doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009

Advance Access publication May 18, 2009

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The enormous amount of short reads generated by the new DNA sequencing technologies call for the development of fast and accurate read alignment programs. A first generation of hash table-based methods has been developed, including MAQ, which is accurate, feature rich and fast enough to align short reads from a single individual. However, MAQ does not support gapped alignment for single-end reads, which makes it unsuitable for alignment of longer reads where indels may occur frequently. The speed of MAQ is also a concern when the alignment is scaled up to the resequencing of hundreds of individuals.

Results: We implemented Burrows-Wheeler Alignment tool (BWA), a new read alignment package that is based on backward search with Burrows-Wheeler Transform (BWT), to efficiently align short sequencing reads against a large reference sequence such as the human genome, allowing mismatches and gaps. BWA supports both base space reads, e.g. from Illumina sequencing machines, and color space reads from AB SOLiD machines. Evaluations on both simulated and real data suggest that BWA is ~10–20× faster than MAQ, while achieving similar accuracy. In addition, BWA outputs alignment in the new standard SAM (Sequence Alignment/Map) format. Variant calling and other downstream analyses after the alignment can be achieved with the new SAMtools software package.

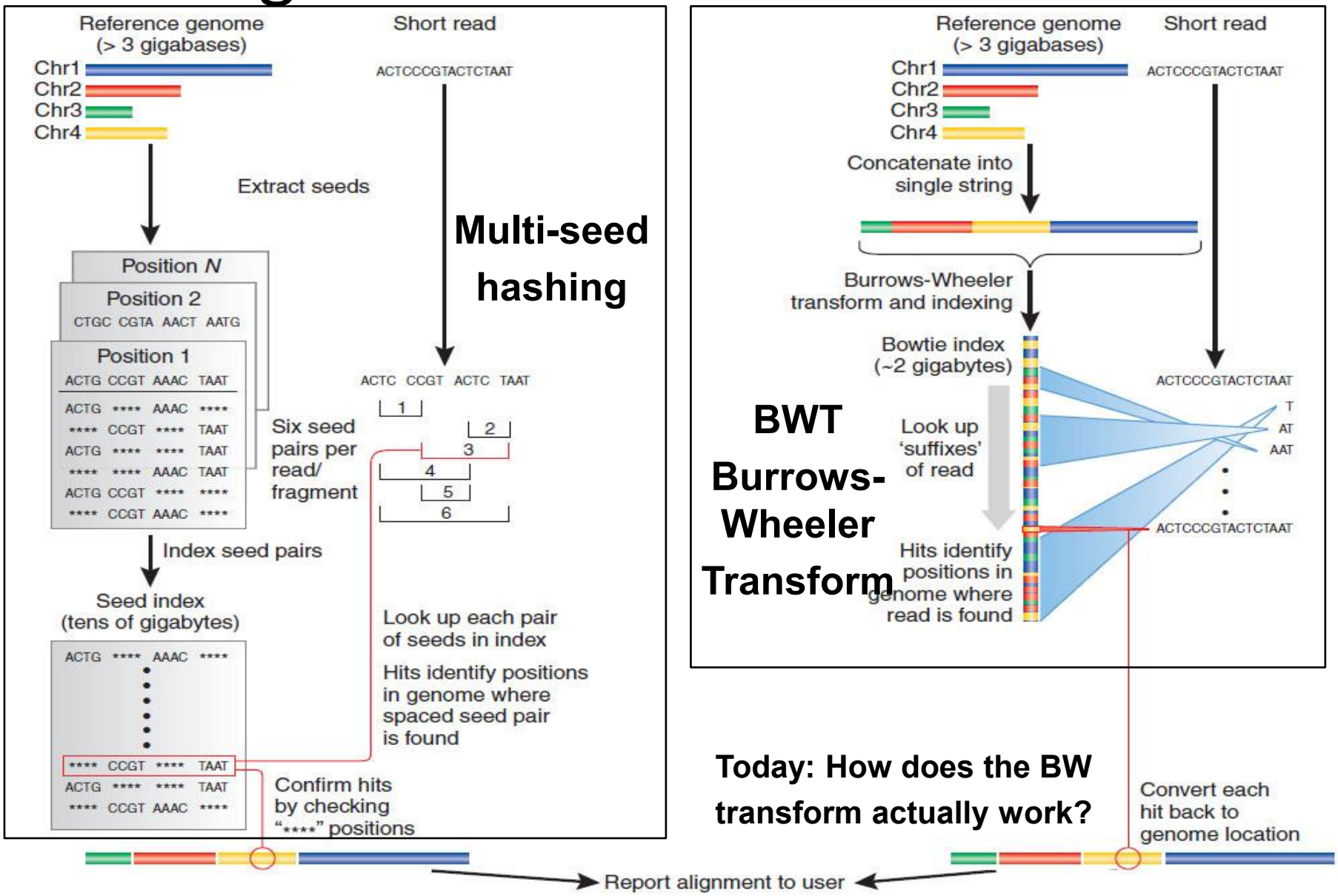
Availability: http://maq.sourceforge.net

Contact: rd@sanger.ac.uk

of scanning the whole genome when few reads are aligned. The second category of software, including SOAPv1 (Li *et al.*, 2008b), PASS (Campagna *et al.*, 2009), MOM (Eaves and Gao, 2009), ProbeMatch (Jung Kim *et al.*, 2009), NovoAlign (http://www.novocraft.com), ReSEQ (http://code.google.com/p/re-seq), Mosaik (http://bioinformatics.bc.edu/marthlab/Mosaik) and BFAST (http://genome.ucla.edu/bfast), hash the genome. These programs can be easily parallelized with multi-threading, but they usually require large memory to build an index for the human genome. In addition, the iterative strategy frequently introduced by these software may make their speed sensitive to the sequencing error rate. The third category includes slider (Malhis *et al.*, 2009) which does alignment by merge-sorting the reference subsequences and read sequences.

Recently, the theory on string matching using Burrows–Wheeler Transform (BWT) (Burrows and Wheeler, 1994) has drawn the attention of several groups, which has led to the development of SOAPv2 (http://soap.genomics.org.cn/), Bowtie (Langmead *et al.*, 2009) and BWA, our new aligner described in this article. Essentially, using backward search (Ferragina and Manzini, 2000; Lippert, 2005) with BWT, we are able to effectively mimic the top-down traversal on the prefix trie of the genome with relatively small memory footprint (Lam *et al.*, 2008) and to count the number of exact hits of a string of length m in $O(m)$ time independent of the size of the genome. For inexact search, BWA samples from the implicit prefix trie the distinct substrings that are less than k edit distance

Hashing vs. Burrows Wheeler Transform



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Trapnell, Cole and Steven L. Salzberg. "How to map billions of short reads onto genomes." Nature Biotechnology 27, no. 5 (2009): 455.

Burrows-Wheeler Transform (BWT)

http://www.hpl.hp.com/techreports/Compaq-DEC/SRC-RR-124.pdf

- Transform: ^BANANA@ INTO: BNN^AA@A

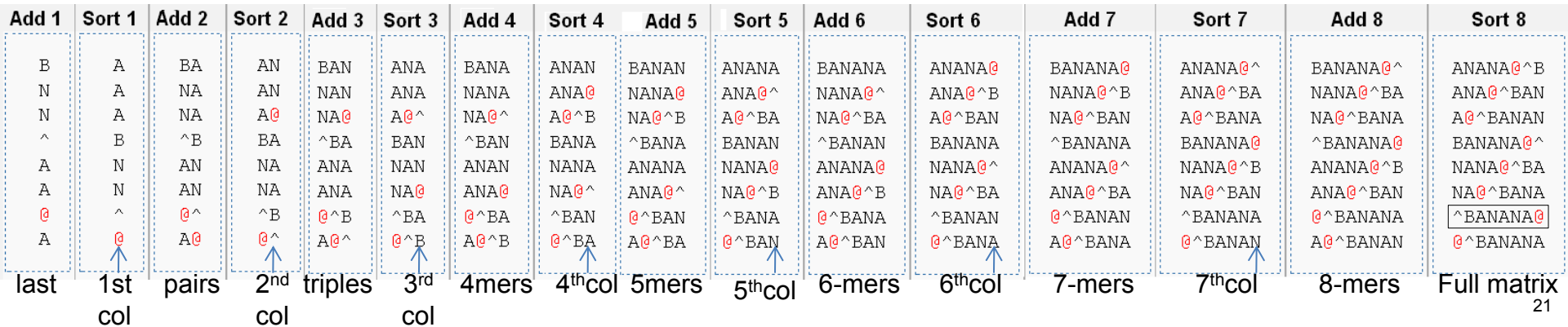
function BWT (*string* s)
    create a table, rows are all possible rotations of s
    sort rows alphabetically
    return (last column of the table)

All Rotations	Sorted List of Rotations	Output Last Column
^BANANA@	ANANA@^**B**	
@^BANANA	ANA@^BA**N**	
A@^BANAN	A@^BANA**N**	
NA@^BANA	BANANA@**^**	
ANA@^BAN	NANA@^B**A**	
NANA@^BA	NA@^BAN**A**	
ANANA@^B	^BANANA**@**	
BANANA@^	@^BANAN**A**	
^BANANA@		BNN^AA@A

- Reversible

function inverseBWT (*string* s)
    create empty table
    repeat length(s) **times**
        insert s as a column of table before first column of the table // first insert creates first column
        sort rows of the table alphabetically
    return (row that ends with the 'EOF' character)

Last column only suffices to reconstruct **entire matrix**, and thus recover **original string**



Searching for an Exact Match

e.g. Searching for OLIS

In MANOLISKELLIS

For simplicity (here):

- only exact matches
- Show entire matrix

In practice: only pointers

Algorithm 3 EXACTMATCH($P[1,p]$)

$c \Leftarrow P[p]$
$sp \Leftarrow C[c] + 1$
$ep \Leftarrow C[c+1] + 1$
$i \Leftarrow p - 1$
while $sp < ep$ and $i \geq 1$ **do**
    $c \Leftarrow P[i]$
    $sp \Leftarrow C[c] + \mathrm{Occ}(c, sp) + 1$
    $ep \Leftarrow C[c] + \mathrm{Occ}(c, ep) + 1$
    $i \Leftarrow i - 1$
end while
return sp, ep

P is the input substring

C[c] – is how many characters occur before c lexographically in the genome

Occ(c,k) is the number of occurrence of the character c before index *k* in the far right column

OLIS

```
1. $MANOLISKELLIS
2. ANOLISKELLIS$M
3. ELLIS$MANOLISK
4. IS$MANOLISKELL
5. ISKELLIS$MANOL
6. LIS$MANOLISKEL
7. LISKELLIS$MANO
8. LLIS$MANOLISKE
9. KELLIS$MANOLIS
10.MANOLISKELLIS$
11.NOLISKELLIS$MA
12.OLISKELLIS$MAN
13.S$MANOLISKELLI
14.SKELLIS$MANOLI
```

OLIS

```
1. $MANOLISKELLIS
2. ANOLISKELLIS$M
3. ELLIS$MANOLISK
4. IS$MANOLISKELL
5. ISKELLIS$MANOL
6. LIS$MANOLISKEL
7. LISKELLIS$MANO
8. LLIS$MANOLISKE
9. KELLIS$MANOLIS
10.MANOLISKELLIS$
11.NOLISKELLIS$MA
12.OLISKELLIS$MAN
13.S$MANOLISKELLI
14.SKELLIS$MANOLI
```

OLIS

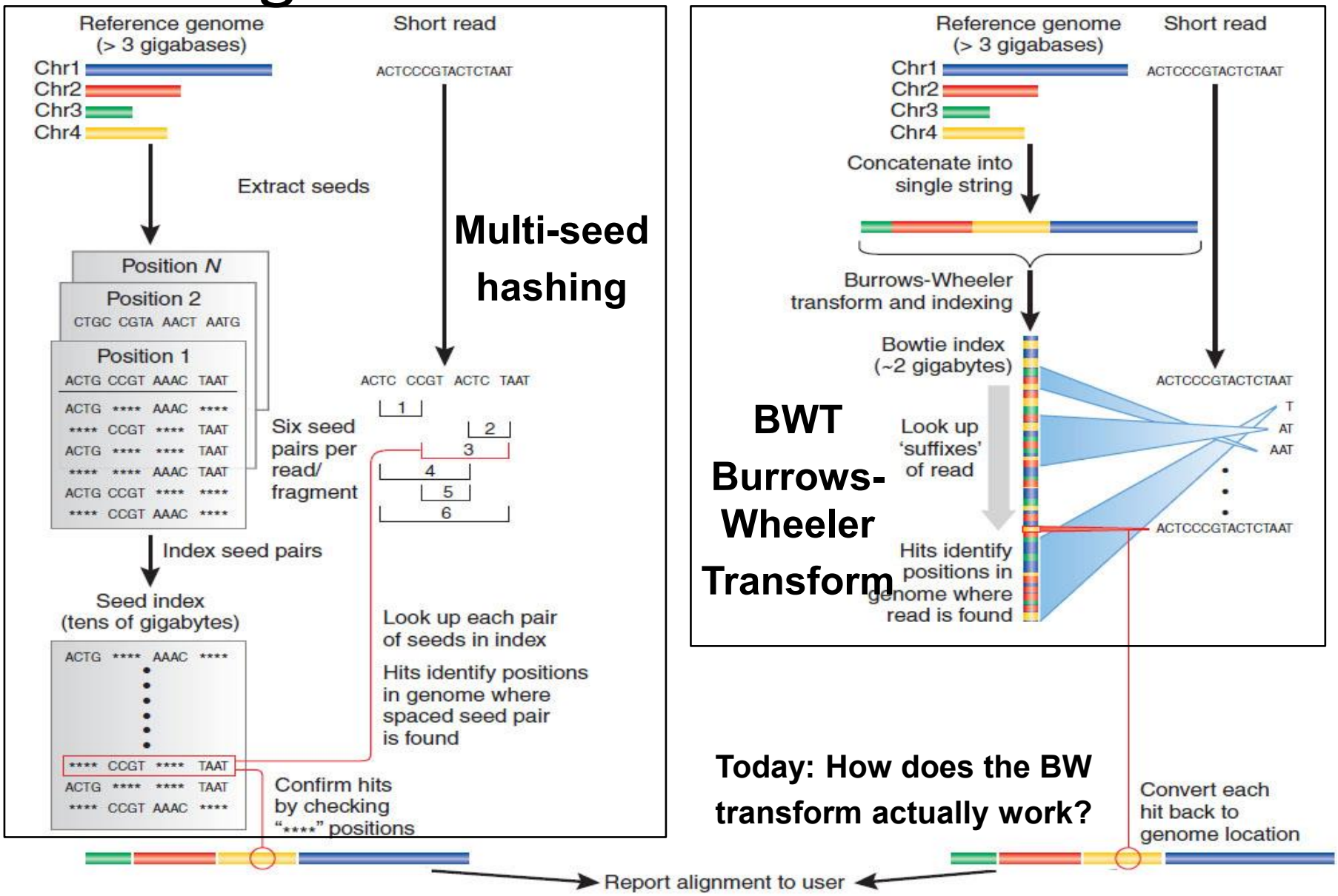
```
1. $MANOLISKELLIS
2. ANOLISKELLIS$M
3. ELLIS$MANOLISK
4. IS$MANOLISKELL
5. ISKELLIS$MANOL
6. LIS$MANOLISKEL
7. LISKELLIS$MANO
8. LLIS$MANOLISKE
9. KELLIS$MANOLIS
10.MANOLISKELLIS$
11.NOLISKELLIS$MA
12.OLISKELLIS$MAN
13.S$MANOLISKELLI
14.SKELLIS$MANOLI
```

OLIS

```
1. $MANOLISKELLIS
2. ANOLISKELLIS$M
3. ELLIS$MANOLISK
4. IS$MANOLISKELL
5. ISKELLIS$MANOL
6. LIS$MANOLISKEL
7. LISKELLIS$MANO
8. LLIS$MANOLISKE
9. KELLIS$MANOLIS
10.MANOLISKELLIS$
11.NOLISKELLIS$MA
12.OLISKELLIS$MAN
13.S$MANOLISKELLI
14.SKELLIS$MANOLI
```

Pseudocode from Langmead et al, 2009. Example by Jason Ernst.

Hashing vs. Burrows Wheeler Transform



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Trapnell, Cole and Steven L. Salzberg. "How to map billions of short reads onto genomes." Nature Biotechnology 27, no. 5 (2009): 455.

Key properties of Burrows-Wheeler Transform

- **Very little memory usage. Same as input (or less)**
  - Don't represent matrix, or strings, just pointers
  - Encode: Simply sort pointers. Decode: follow pointers
- **Original application: string compression (bZip2)**
  - Runs of letters compressed into (letter, runlength) pairs
- **Bioinformatics applications: substring searching**
  - Achieve similar run time as hash tables, suffix trees
  - But: very memory efficient ➔ practical speed gains
- **Mapping 100,000s of reads: only transform once**
  - Pre-process once; read counts in transformed space.
  - Reverse transform once, map counts to genome coords

Goals for today: Computational Epigenomics

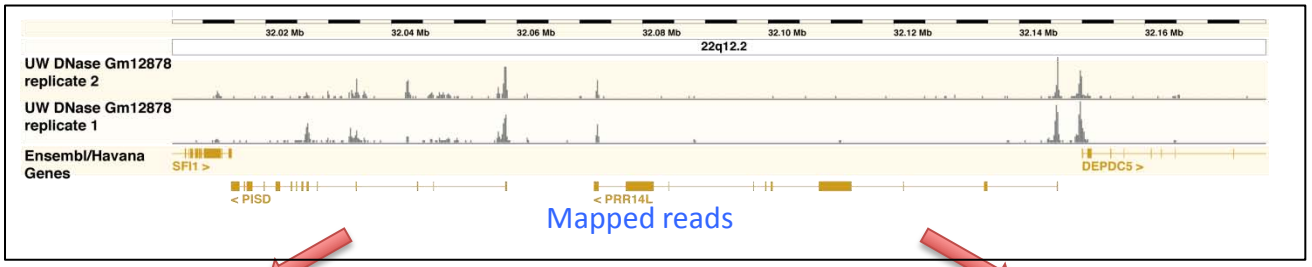
1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

(Future: Chromatin states to interpret disease-associated variants)

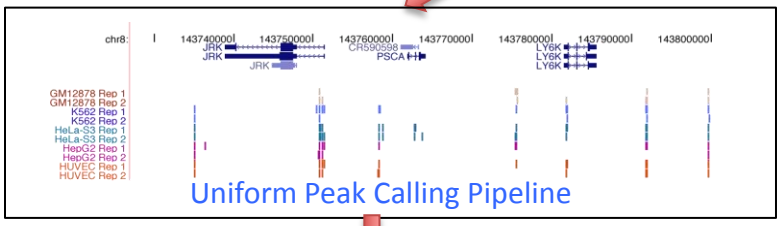
Quality control metrics

ChIP vs. Input DNA
Read quality
Mappability
Library complexity

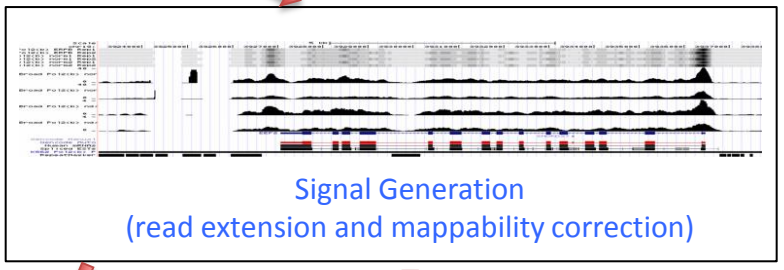
ENCODE uniform processing pipeline



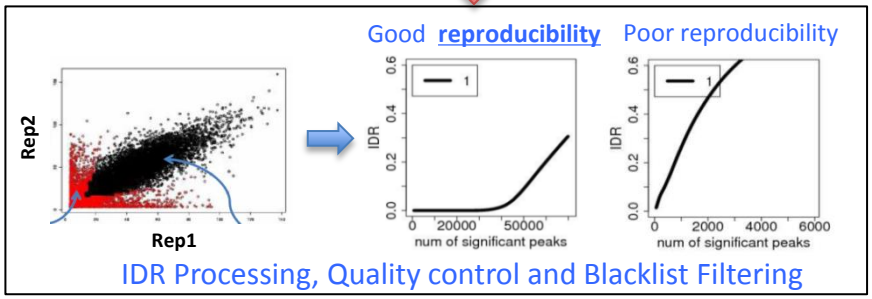
Mapped reads



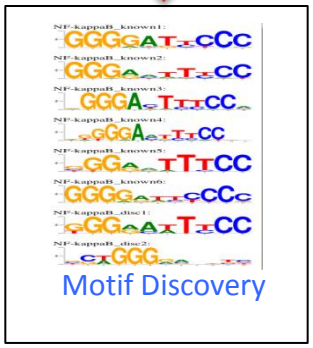
Uniform Peak Calling Pipeline



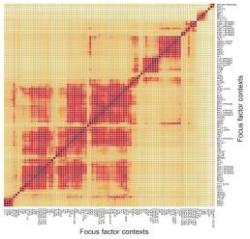
Signal Generation
(read extension and mappability correction)



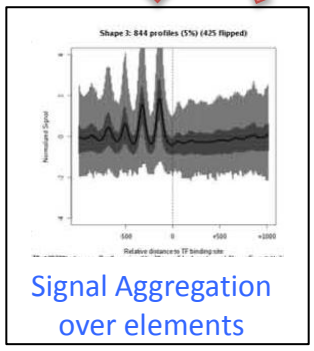
IDR Processing, Quality control and Blacklist Filtering



Motif Discovery

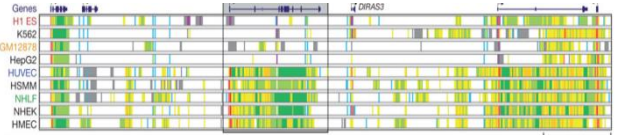


Co-association analysis

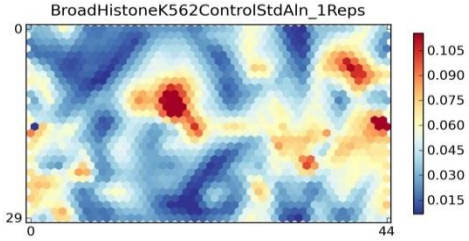


Signal Aggregation over elements

Segmentation

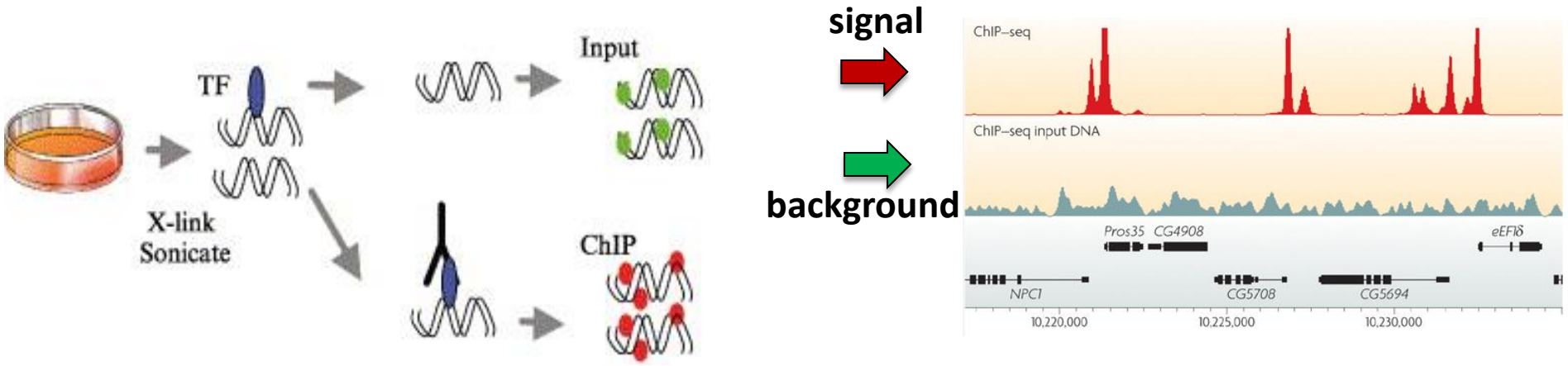


ChromHMM/Segway



Self Organising Maps

QC1: Use of input DNA as control dataset

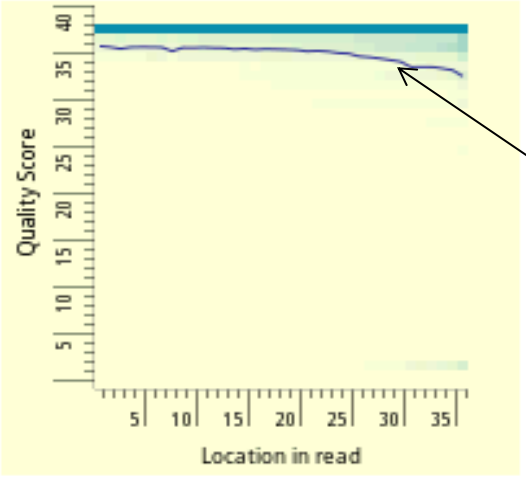


- **Challenge:**
  - Even without antibody: Reads are **not** uniformly scattered
- **Sources of bias in input dataset scatter:**
  - Non-uniform fragmentation of the genome
  - Open chromatin fragmented more easily than closed regions
  - Repetitive sequences over-collapsed in the assembled genome.
- **How to control for these biases:**
  - Remove portion of DNA sample before ChIP step
  - Carry out control experiment without an antibody (input DNA)
  - Fragment input DNA, sequence reads, map, use as background

QC2: Read-level sequencing quality score Q>10

Read quality histograms

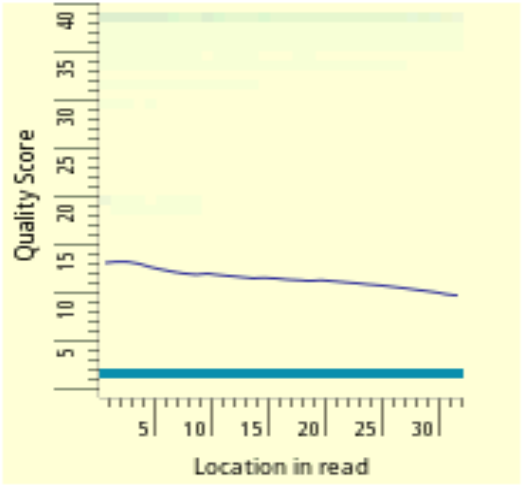
High quality reads



average base score per position

- Each column is a color-coded histogram
- Encodes fraction of all mapped reads that have base score Q (y-axis) at each position (x-axis)
- Darker blue = higher density

Low quality reads



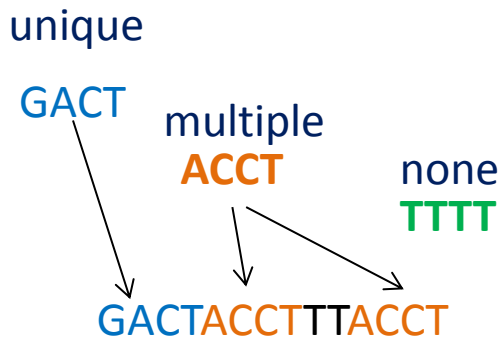
- Read quality tends to drop towards the ends of reads
- Low average per base score implies greater probabilty of mismappings.

- Typically, reject reads whose average score Q < 10

QC3: Fraction of short reads mapped >50%

Reads can map to:

- exactly one location (uniquely mapping)
- multiple locations (repetitive or multi-mapping)
- no locations (unmappable)



Dealing with multiply-mapping reads:

- Conservative approach: do not assign to any location
- Probabilistic approach: assign fractionally to all locations
- Sampling approach: pick one location at random, averages across many reads
- EM approach: map according to density, estimated from unambiguous reads
- Pair-end approach: use paired end read to resolve ambiguities in repeat reads

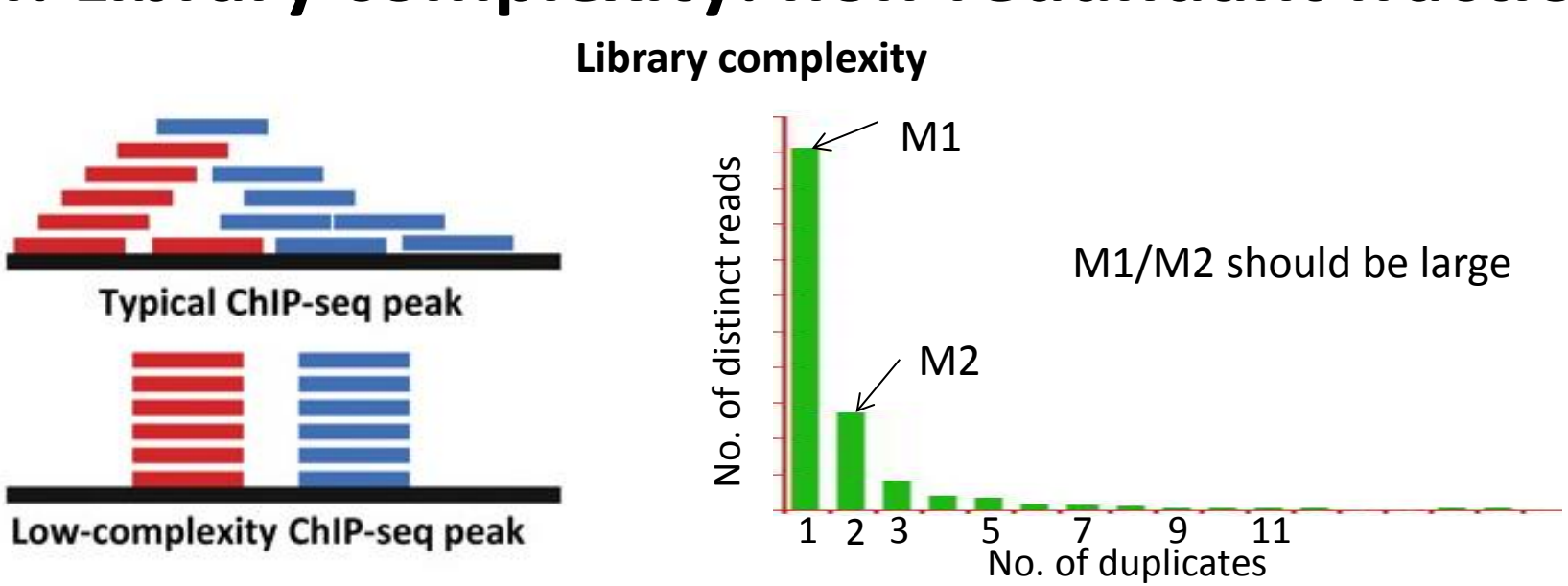
Absence of reads in a region could be due to:

- No assembly coverage in that region (e.g. peri-centromeric region)
- Too many reads mapping to this location (e.g. repetitive element)
- No activity observed in this location (e.g. inactive / quiescent / dead regions)

Dealing with mappability biases:

- 'Black-listed' regions, promiscuous across many datasets
- 'White-listed' regions, for which at least some dataset has unique reads
- Treat unmappable regions as missing data, distinguish from 'empty' regions

QC4: Library complexity: non-redundant fraction



How many distinct uniquely mapping read? How many duplicates?

If your sample does not contain sufficient DNA and/or you over-sequence, you will simply be repeatedly sequencing PCR duplicates of a restricted pool of distinct DNA fragments. This is known a **low-complexity library** and is not desirable.

- **Histogram of no. of duplicates**
- Non-redundant fraction (NRF) = $\frac{\text{No. of 'distinct' unique-mapping reads}}{\text{No. of unique-mapping reads}}$
- NRF should be > 0.8 when 10M < #reads < 80M unique-mapping reads

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

(Future: Chromatin states to interpret disease-associated variants)

Cross-correlation analysis

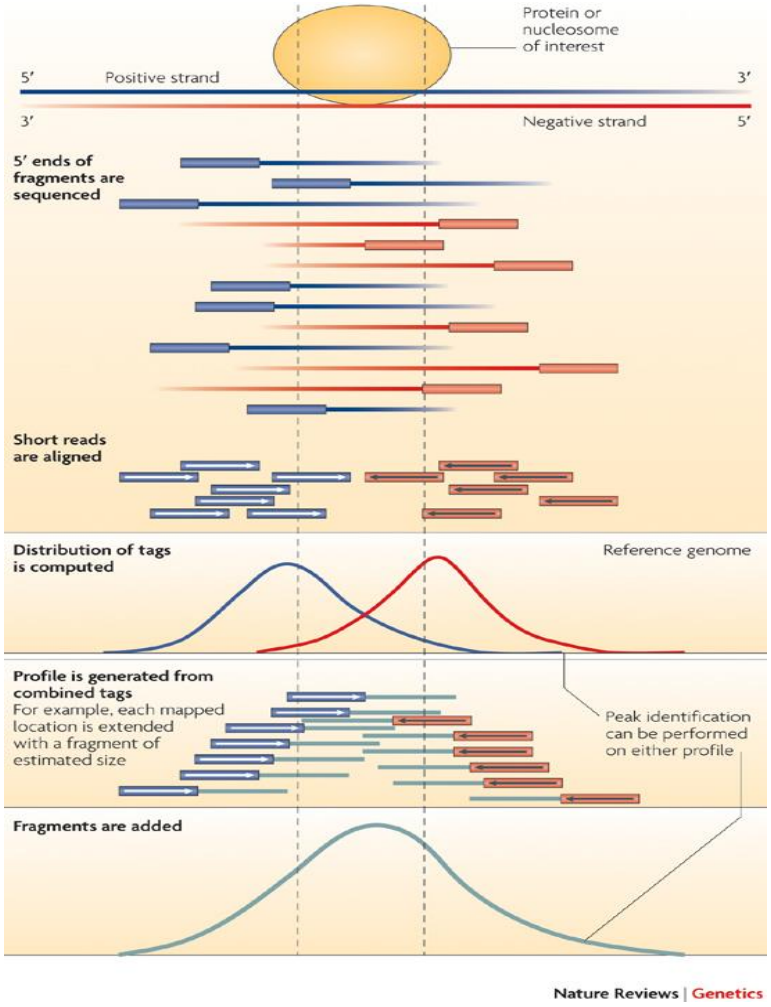
Exploiting forward and reverse reads

Fragment-length peak

Phantom read-length peak

ChIP-seq: exploiting forward and reverse reads

(Chromatin immunoprecipitation followed by sequencing)



Multiple IP fragments are obtained corresponding to each binding event

Ends of the fragments are sequenced i.e. "Short-reads/tags"

- Typically ~36 bp, 50 bp, 76 bp or 101 bp

Single-end (SE) sequencing

- Randomly sequence one of the ends of each fragment

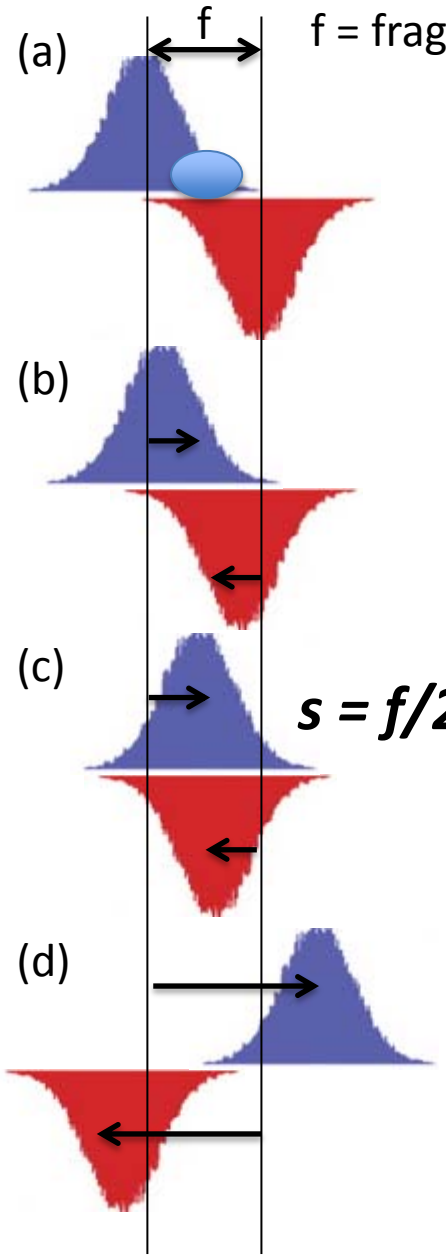
Paired-end (PE) sequencing

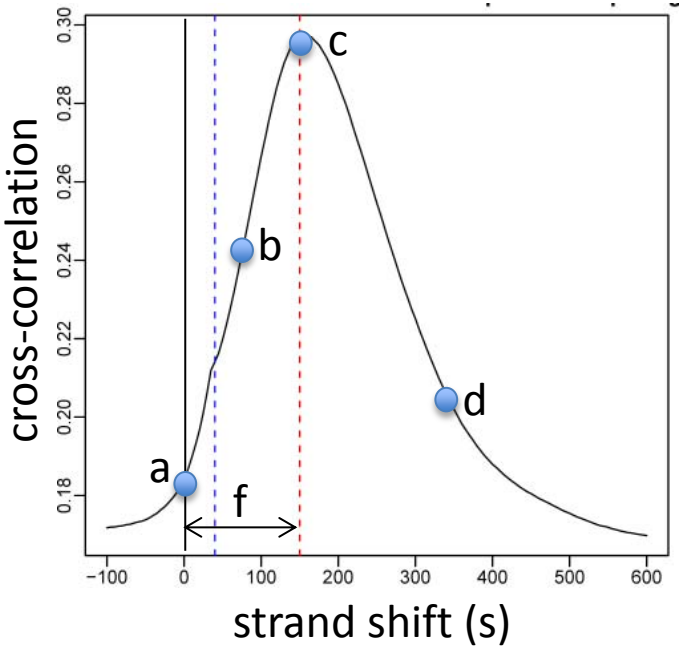
- sequence both ends of each fragment

Canonical "stranded mirror distribution of short-reads" after mapping reads to genome

- Heaps of reads on the **+ strand** and **– strand** separated by a distance ~= fragment length

Strand cross-correlation (CC) analysis



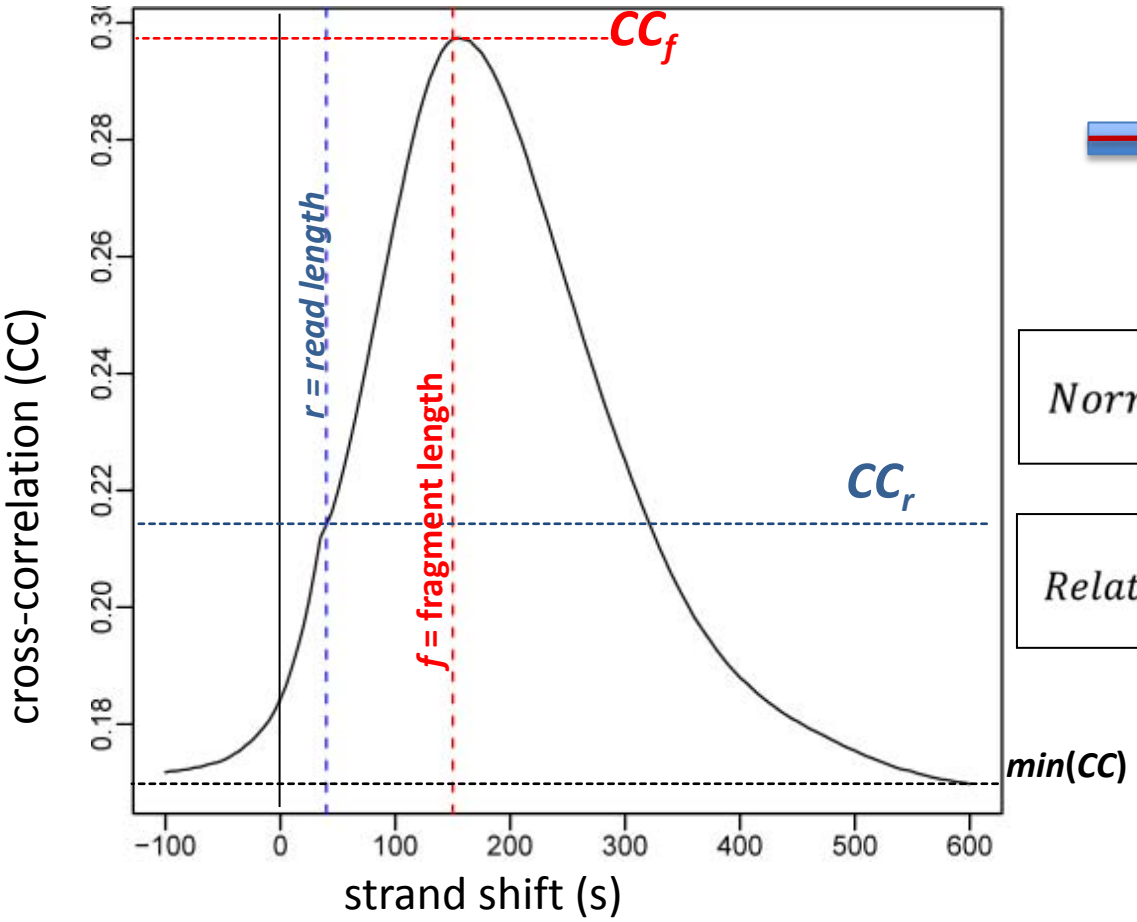


s = f/2 + f/2

1. Calculate forward and reverse strand signals
2. Shift both by specified offset towards each other
3. Calculate correlation of two signals at that shift
4. Correlation peaks at **fragment length** offset ***f***

f is the length at which ChIP DNA is fragmented

Cross-correlation at *read* vs. *fragment* length

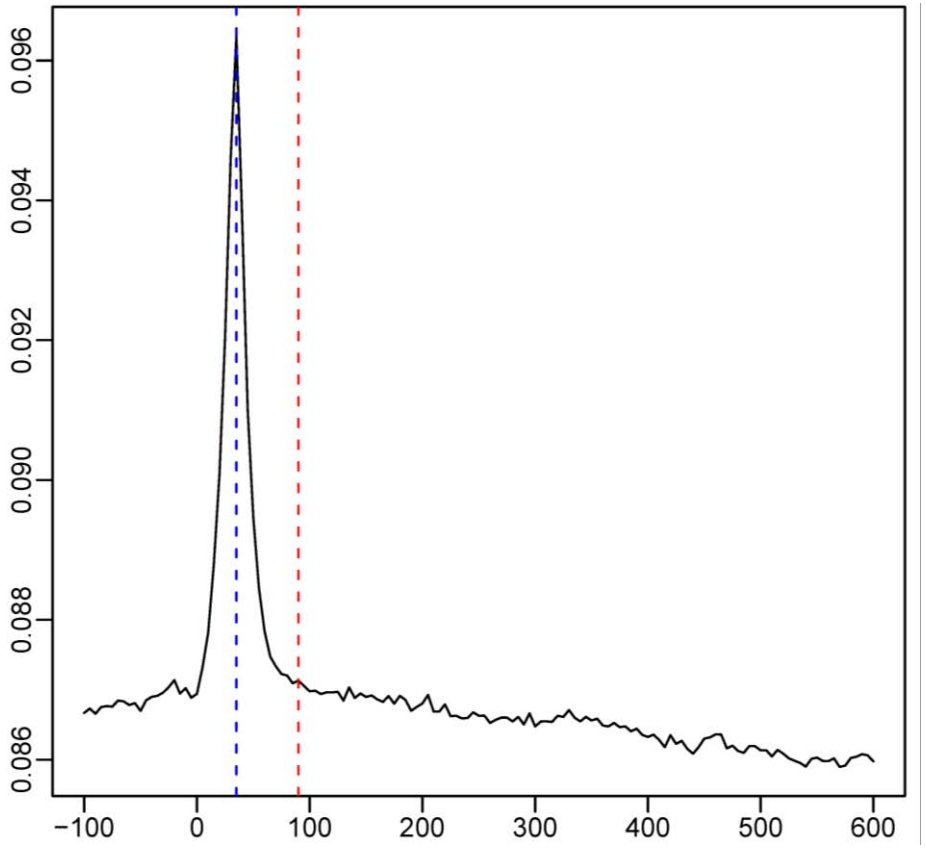


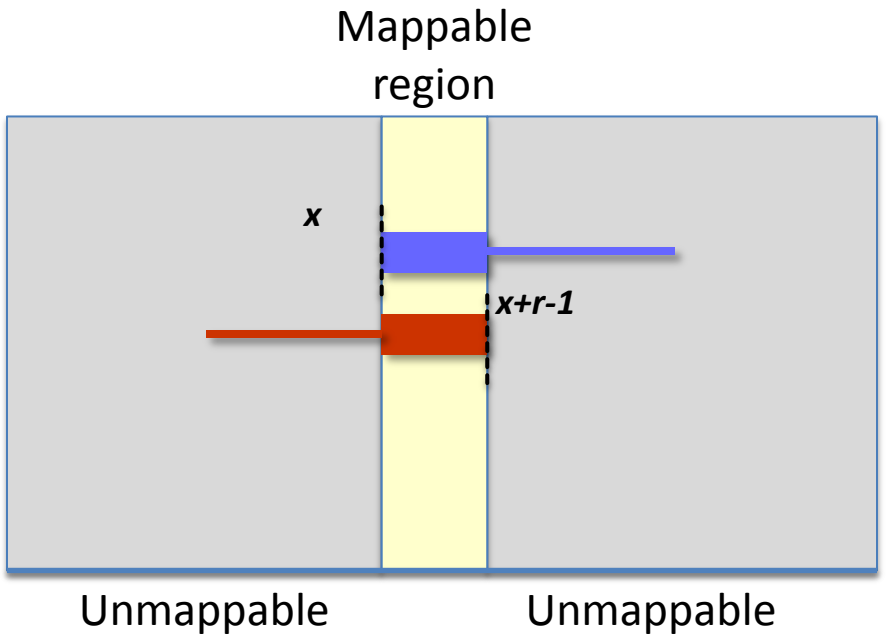
$$Normalized\ strand\ CC(NSC) = \frac{CC_f}{\min(CC)}$$

$$Relative\ strand\ CC(RSC) = \frac{(CC_f - \min(CC))}{(CC_r - \min(CC))}$$

- Sign of a good dataset:
  - High absolute cross-correlation at *fragment* length (NSC)
  - High *fragment* length CC relative to *read* length CC (RSC)

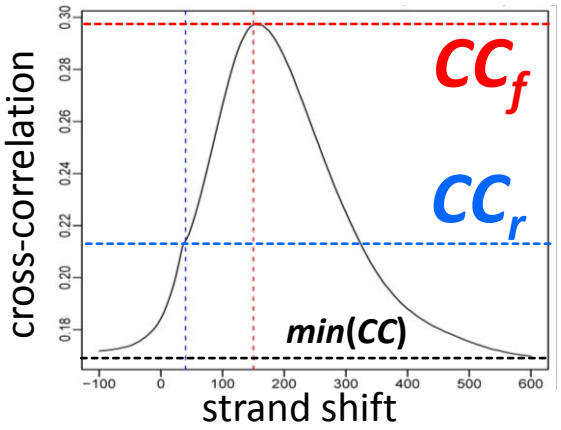
Where does *read* cross-correlation come from?



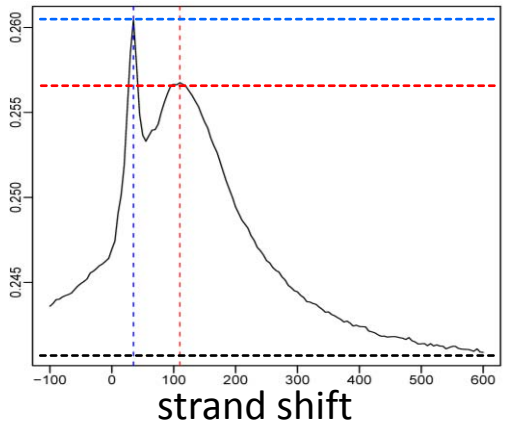


- Input dataset (no ChIP) shows 'phantom' peak at ***read*** length only
- Due to read mappability:
  - If position 'x' is uniquely mappable on + strand
  - Then position 'x+r-1' is uniquely mappable on – strand
- *Fragment*-length peak should always dominate the read-length peak

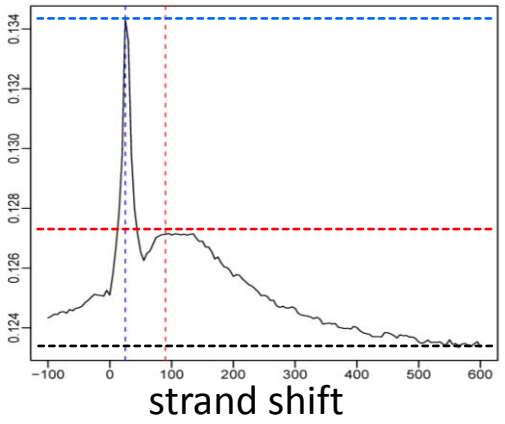
Example of good, medium, bad CC datasets



Highly quality



Medium quality



Low quality

$$Normalized\ strand\ CC(NSC) = \frac{CC_f}{\min(CC)}$$

$$Relative\ strand\ CC(RSC) = \frac{(CC_f - \min(CC))}{(CC_r - \min(CC))}$$

For highly enriched datasets, fragment length cross-correlation peak should be able to beat read-length phantom peak

RSC should be > 1

Goals for today: Computational Epigenomics

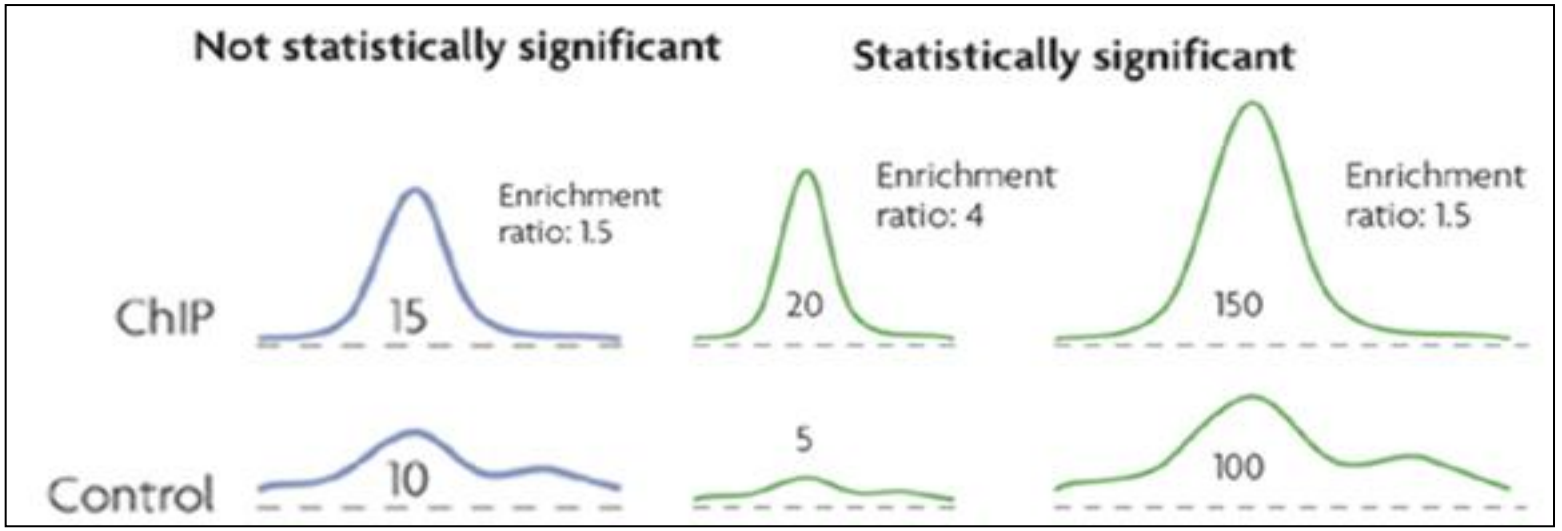
1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

(Future: Chromatin states to interpret disease-associated variants)

Peak Calling

Continuous signal → Intervals

Peak calling: detect regions of enrichment



Goal: Transform read counts into **normalized intensity signal**

Steps:

1. Estimate fragment-length *f* using strand cross-correlation analysis
2. Extend each read from 5' to 3' direction to fragment length *f*
3. Sum intensity for each base in 'extended reads' from both strands
4. Perform same operation on input-DNA control data (correct for sequencing depth differences)
5. Calculate enrichment ratio value for every position in the genome

Result: Enrichment fold difference for ChIP / control signal

Peak calling: identify discrete intervals

Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method exludes putative duplicated regions, no treatment of deletions

Source: Wilbanks, Elizabeth G. and Marc T. Facciotti. "Evaluation of algorithm performance in ChIP-seq peak detection." PLOS ONE 5, no. 7 (2010): e11471.

Peak calling thresholds

Poisson p-value thresholds

- Read count model: Locally-adjusted' Poisson distribution

$$P(count = x) = \frac{\lambda_{local}^{x} \exp(-\lambda_{local})}{x!}$$

- λ_{local} = max(λ_{BG}, [λ_{1k},] λ_{5k}, λ_{10k}) estimated from control data
  - Poisson *p*-value = $P(count \geq x)$
  - *q*-value : Multiple hypothesis correction

Peaks: Genomic locations that pass a user-defined *p*-value (e.g. 1e-5) or *q*-value (e.g. 0.01) threshold

Empirical False discovery rates

- Swap ChIP and input-DNA tracks
  - Recompute*p*-values
- At each *p*-value, eFDR = Number of control peaks / Number of ChIPpeaks
  - Use an FDR threshold to call peaks

Issues with peak calling thresholds

Cannot set a universal threshold for empirical FDRs and p-values

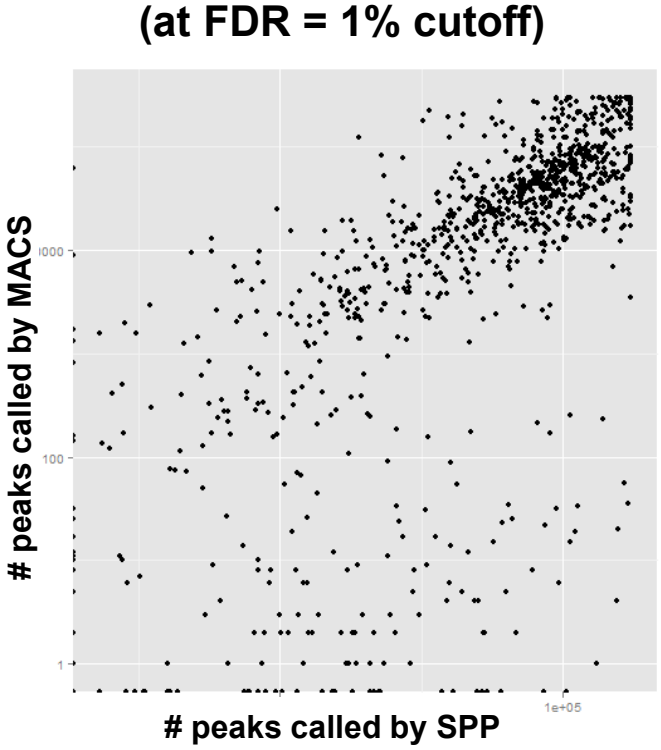
- Depends on ChIP and input sequencing depth
  - Depends on binding ubiquity of factor
  - Stronger antibodies get an advantage

FDRs quite unstable

- Small changes in threshold => massive changes in peak numbers

Difficult to compare results across peak callers with a fixed threshold

- Different methods to compute eFDR or q-values



Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

(Future: Chromatin states to interpret disease-associated variants)

Selecting meaningful peaks using reproducibility

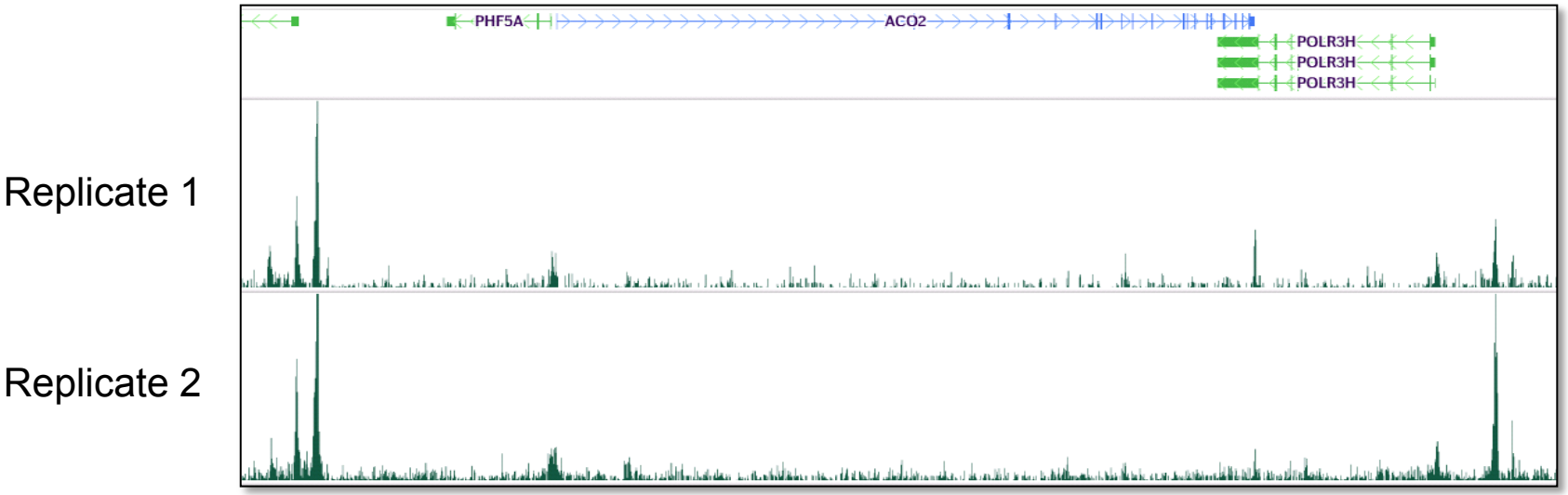
Use peak ranks in replicate experiments

IDR: Irreproducible Discovery Rate

http://anshul.kundaje.net/projects/idr

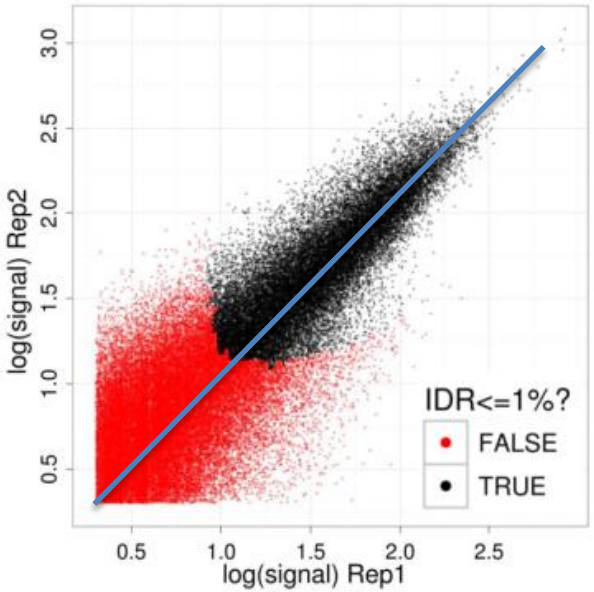
A. Kundaje, Q. Li , B. Brown, J. Rozowsky, S. Wilder, M. Gerstein, I. Dunham, E. Birney, P. Bickel

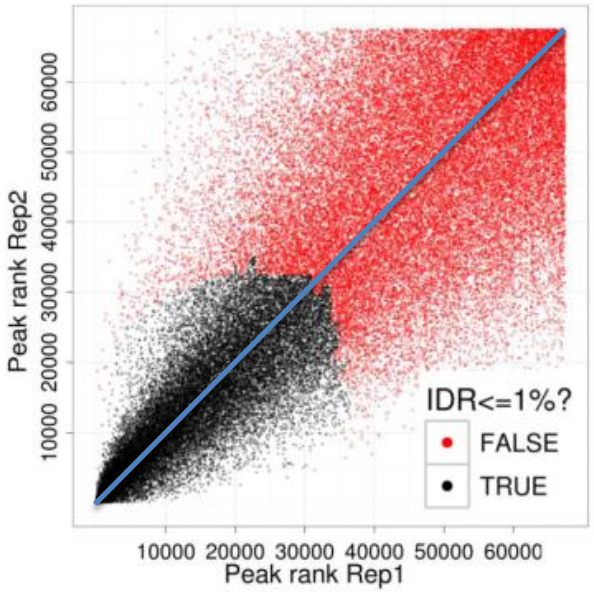
How to combine two replicates

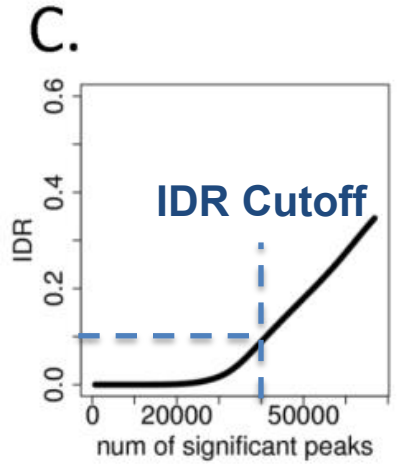


- Challenge:
  - Replicates show small differences in peak heights
  - Many peaks in common, but many are unique
- Problem with simple solutions:
  - Union: too lenient, keeps garbage from both
  - Intersection: too stringent, throws away good peaks
  - Sum: does not exploit independence of two datasets

IDR idea: Exploit peak rank similarity in replicates







- Key idea: True peaks will be highly ranked in both replicates
  - Keep going down rank list, until ranks are no longer correlated
    - This cutoff could be different for the two replicates
  - The actual peaks included may differ between replicates
    - Adaptively learn optimal peak calling threshold
  - **F**DR threshold of 10% ➔ 10% of peaks are false (widely used)
  - **I**DR threshold of 10% ➔ 10% of peaks are not reproducible

The IDR model: A two component mixture model

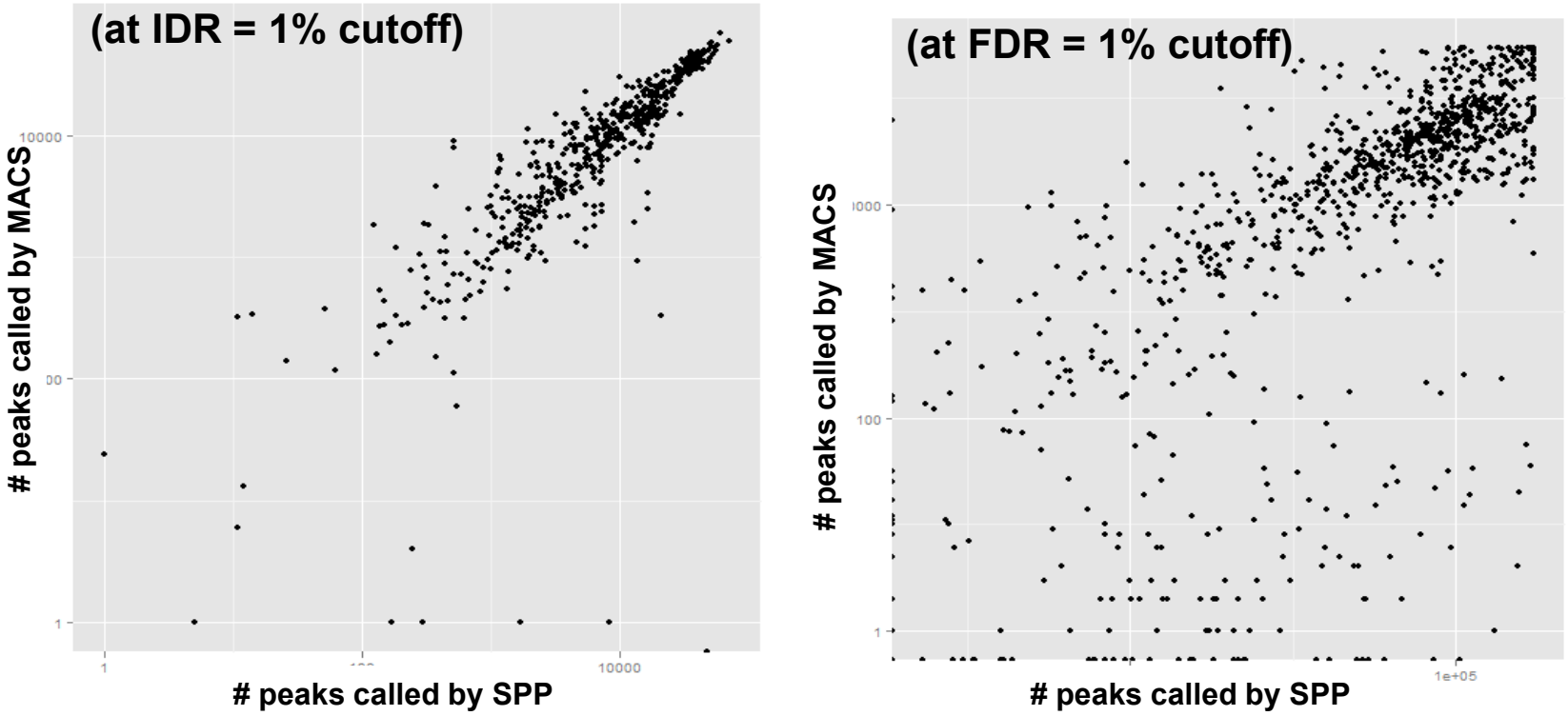
- Looking only at ranks means that the marginals are uniform, so all the information is encoded in the joint distribution.
- Model the joint distribution of ranks as though it came from a two component Gaussian mixture model:

$$(x, y) \sim pN(\mu, \mu, \sigma, \sigma, \rho) + (1 - p)N(0,0,1,1,0)$$

- This can be fit via an EM-like algorithm.

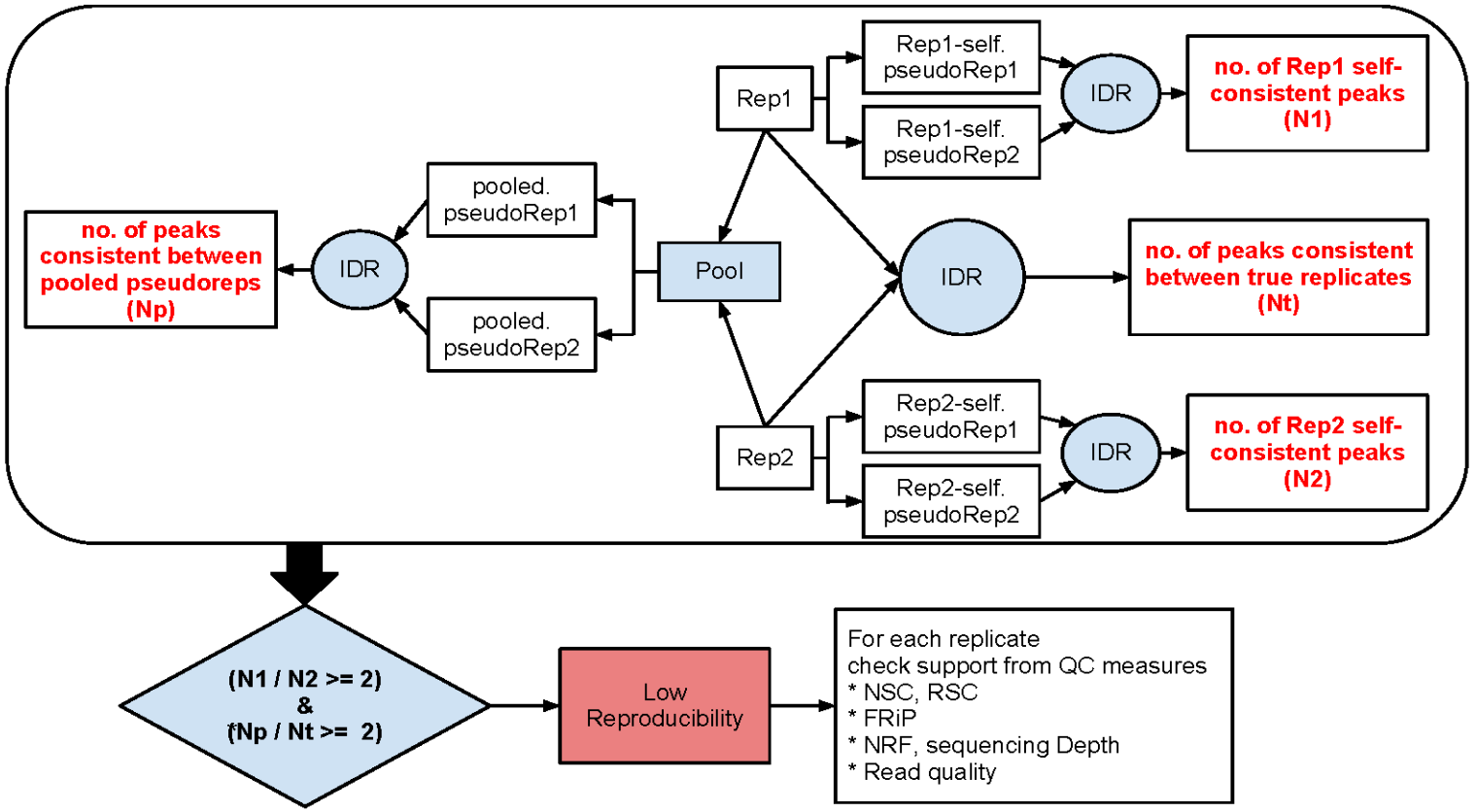
IDR leads to higher consistence between peak callers

IDR = Irreproducible Discovery Rate FDR = False Discovery Rate



- Compare number of peaks found by two different peak callers
- IDR thresholds are far more robust and comparable than FDR
- FDR only relies on enrichment over input, IDR exploits replicates

What if we don't have good replicates?

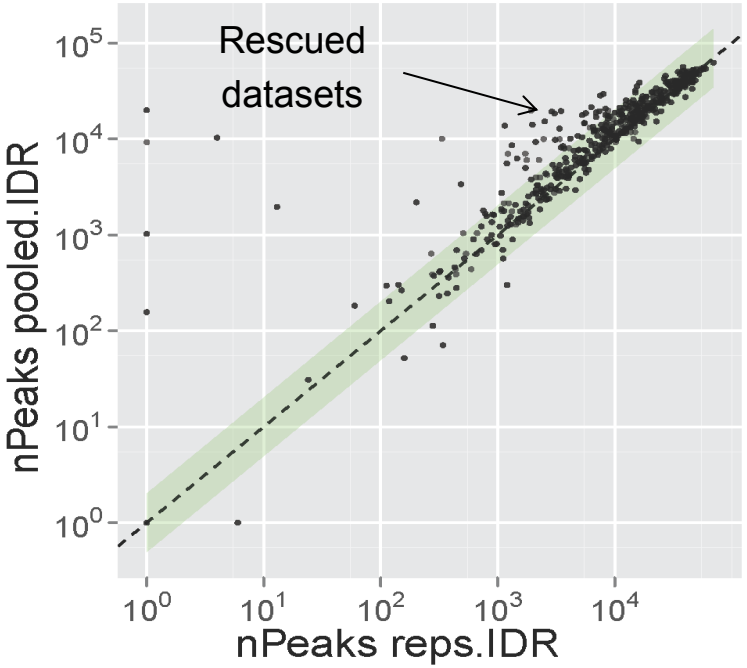


- IDR pipeline uses replicates when they are available
- IDR pipeline also evaluates each replicate individually
  - Pooling strategy to generate pseudo-replicates

➔ Can pin-point 'bad' replicates that may lead to low reproducibility

➔ Can estimate IDR thresholds when replicates are not available

Only one good replicate: Pseudo-replicates



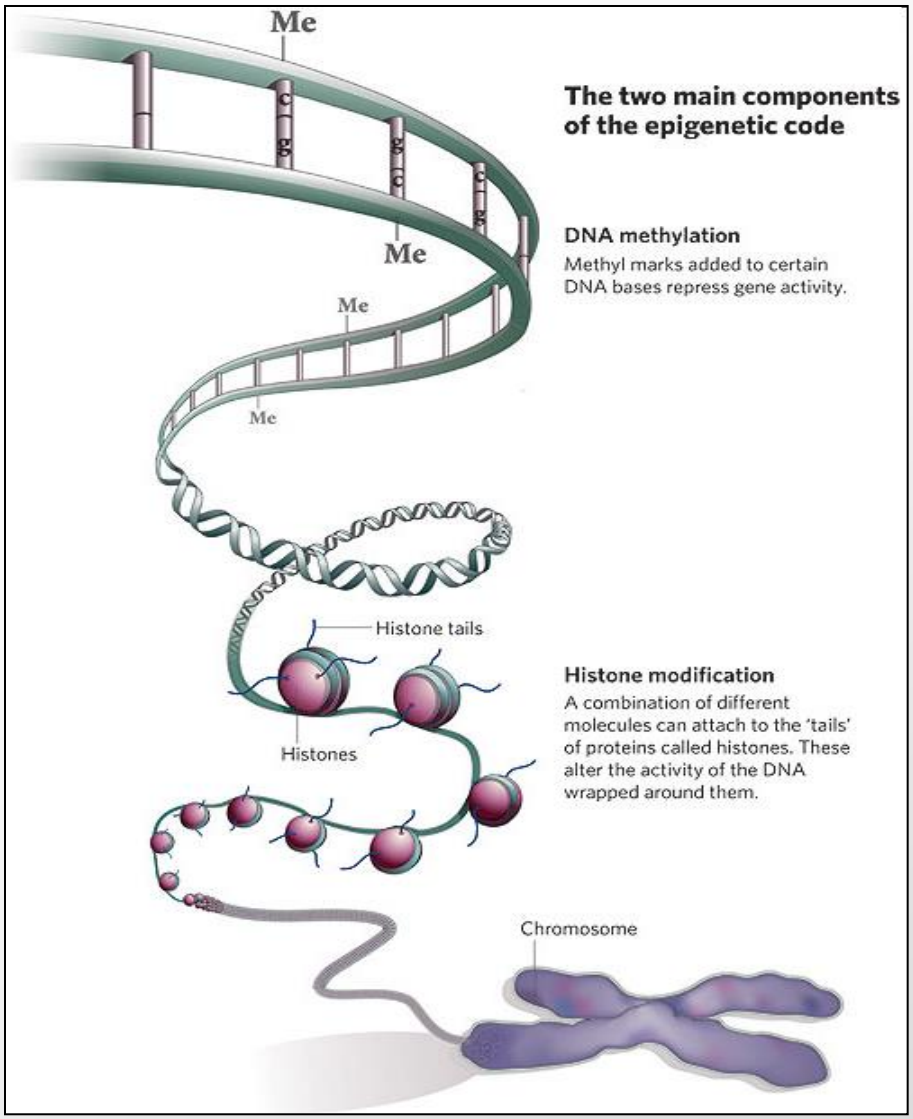
- IDR pipeline can be used to rescue datasets with only one good replicate (using pseudo-replicates)
- IDR pipeline can also be used to call optimal thresholds on a dataset with a single replicate (e.g. when there isn't enough material to perform multiple reps)

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

(Future: Chromatin states to interpret disease-associated variants)

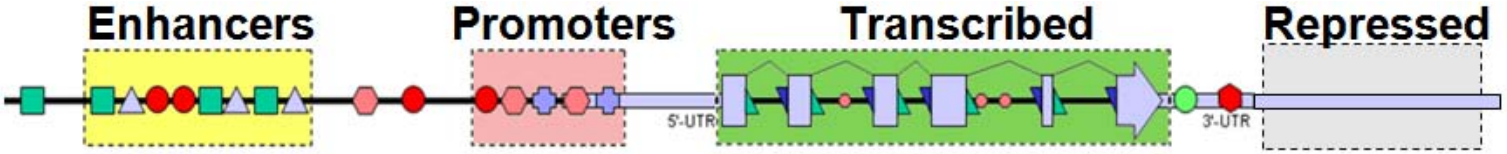
Chromatin signatures for genome annotation



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Qiu, Jane. "Epigenetics: Unfinished Symphony." Nature 441, no. 7090 (2006): 143-145.

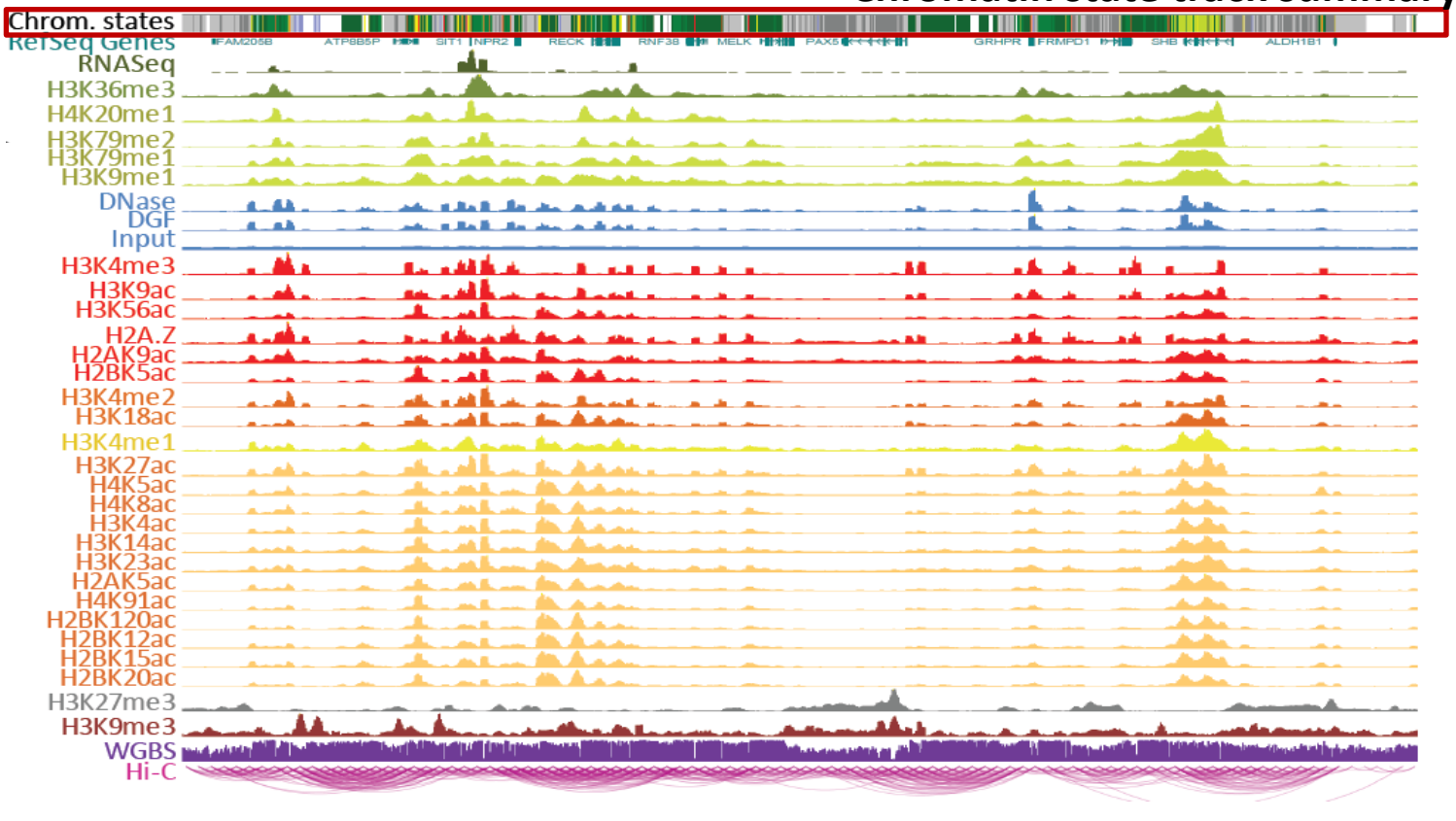
- **Challenges**
  - Dozens of marks
  - Complex combinatorics
  - Diversity and dynamics
- **Histone code hypothesis**
  - Distinct function for distinct combinations of marks?
  - Both additive and combinatorial effects
- **How do we find biologically relevant ones?**
  - Unsupervised approach
  - Probabilistic model
  - Explicit combinatorics

Summarize multiple marks into chromatin states



Chromatin state track summary

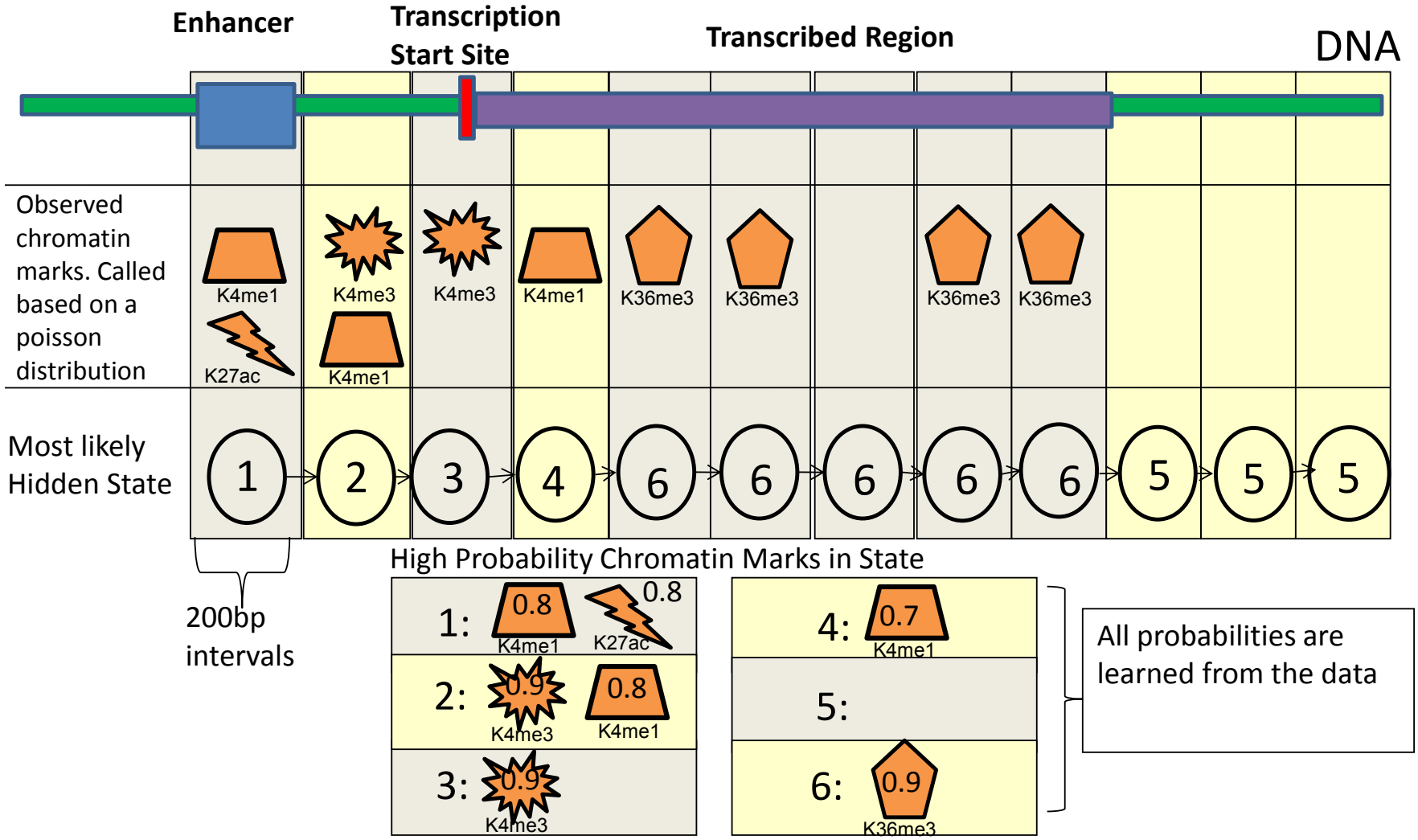
30+ epigenomics marks

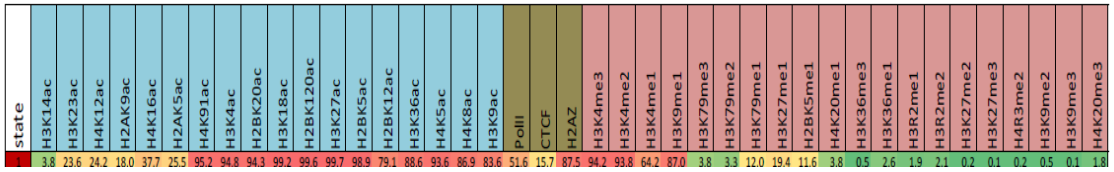


WashU Epigenome Browser

ChromHMM: multi-variate hidden Markov model

Multivariate HMM for Chromatin States





Ernst and Kellis
Nature Biotech 2010

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Ernst, Jason and Manolis Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature Biotechnology 28, no. 8 (2010): 817-825.

Design Choice

- How to model the emission distribution
  - Model the signal directly
  - Locally binarize the data
- For *M* input marks each state *k* has a vector of $(p_{k1},\ldots,p_{kM})$ of parameters for independent Bernoulli random variables which determine the emission probability for an observed combination of marks

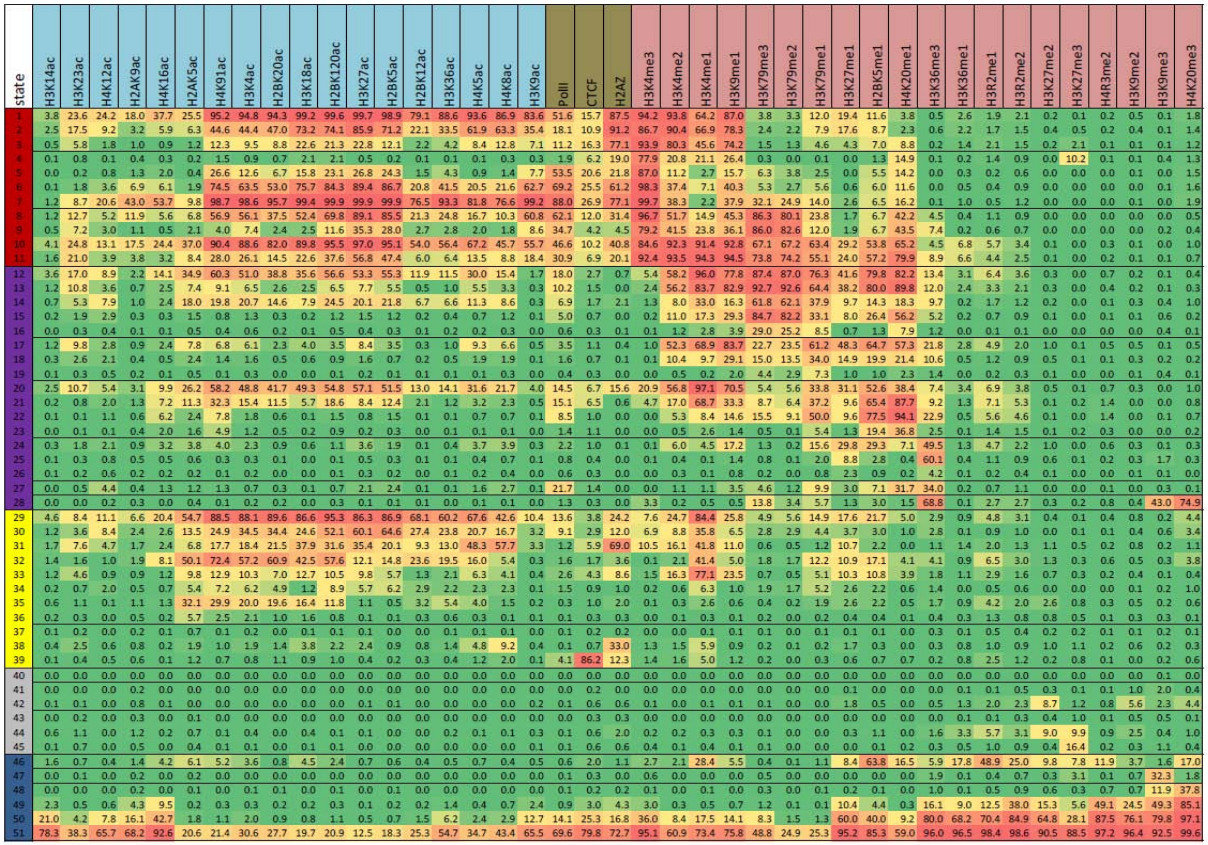
Data Binarization

- Leads to biologically interpretable models that can be robustly learned
- Let c_{ij} be the number of reads for mark *i.* mapping to bin *j.* λ_i be the average number of reads mapping to a bin for modification *i*. The input for feature *i* becomes *'1'* if

$$P(X>c_{ij})<10^{-4}$$

where *X* is a Poisson random variable with mean λ_i

Emission Parameter Matrix $e_k(\vec{x}_i)$



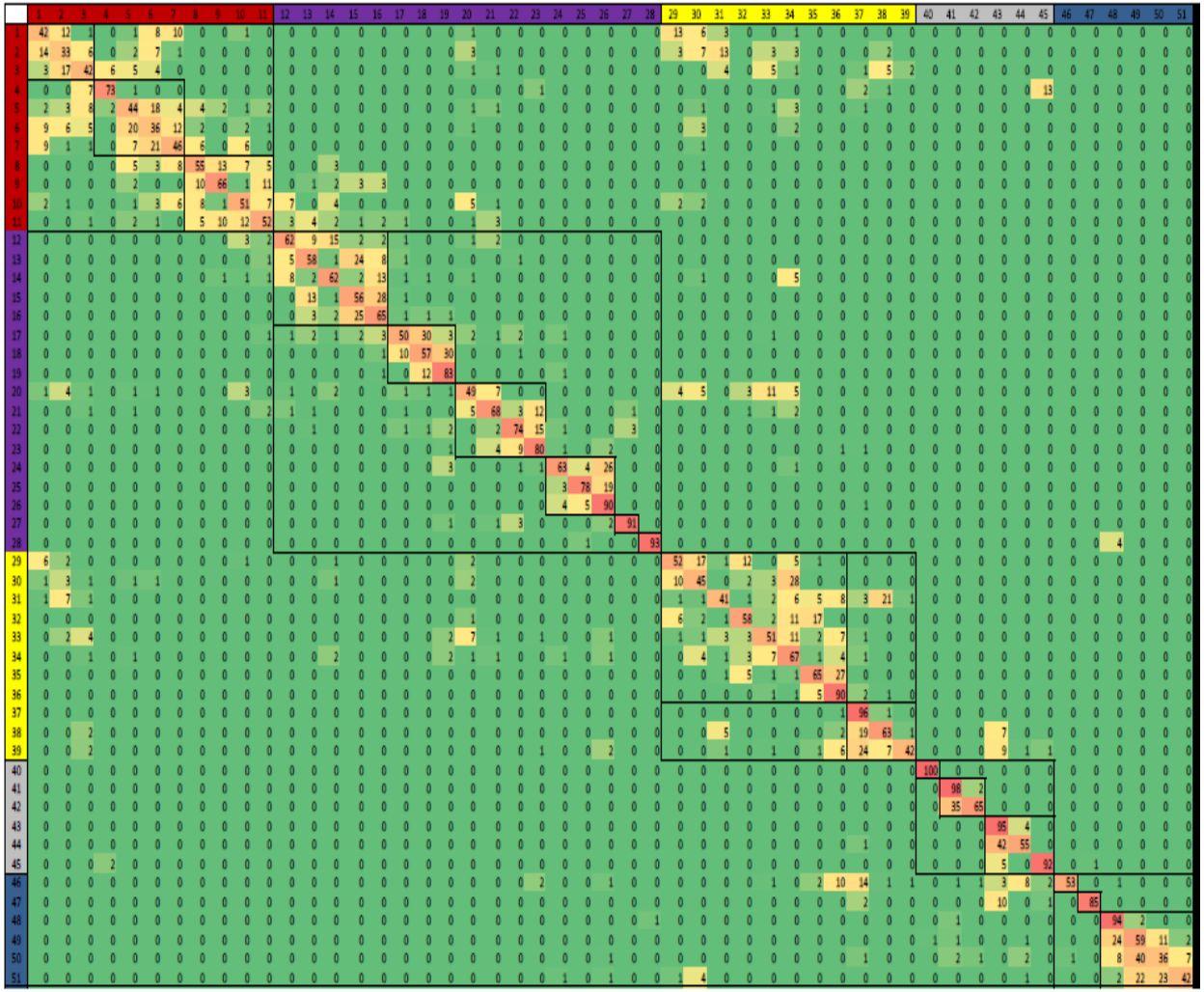
State	H3K14ac	H3K23ac	H4K12ac	H2AK9ac	H4K16ac	H2AK5ac	H4K91ac	H3K4ac	H2BK20ac	H3K18ac	H2BK120ac	H3K27ac	H2BK5ac	H2BK12ac	H3K36ac	H4K5ac	H4K8ac	H3K9ac	PolII	CTCF	H2AZ	H3K4me3	H3K4me2	H3K4me1	H3K9me1	H3K79me3	H3K79me2	H3K79me1	H3K27me1	H2BK5me1	H4K20me1	H3K36me3	H3K36me1	H3R2me1	H3R2me2	H3K27me2	H3K27me3	H4R3me2	H3K9me2	H3K9me3	H4K20me3
1	3.8	23.6	24.2	18.0	37.7	25.5	95.2	94.8	94.3	99.2	99.6	99.7	98.9	79.1	88.6	93.6	86.9	83.6	51.6	15.7	87.5	94.2	93.8	64.2	87.0	3.8	3.3	12.0	19.4	11.6	3.8	0.5	2.6	1.9	2.1	0.2	0.1	0.2	0.5	0.1	1.8
2	2.5	17.5	9.2	3.2	5.9	6.3	44.6	44.4	47.0	73.2	74.1	85.9	71.2	22.1	33.5	61.9	63.3	35.4	18.1	10.9	91.2	86.7	90.4	66.9	78.3	2.4	2.2	7.9	17.6	8.7	2.3	0.6	2.2	1.7	1.5	0.4	0.5	0.2	0.4	0.1	1.4
3	0.5	5.8	1.8	1.0	0.9	1.2	12.3	9.5	8.8	22.6	21.3	22.8	12.1	2.2	4.2	8.4	12.8	7.1	11.2	16.3	77.1	93.9	80.3	45.6	74.2	1.5	1.3	4.6	4.3	7.0	8.8	0.2	1.4	2.1	1.5	0.2	2.1	0.1	0.1	0.1	1.2
4	0.1	0.8	0.1	0.4	0.3	0.2	1.5	0.9	0.7	2.1	2.1	0.5	0.2	0.1	0.1	0.1	0.3	0.3	1.9	6.2	19.0	77.9	20.8	21.1	26.4	0.3	0.0	0.1	0.0	1.3	14.9	0.1	0.2	1.4	0.9	0.0	10.2	0.1	0.1	0.4	1.3
5	0.0	0.2	0.8	1.3	2.0	0.4	26.6	12.6	6.7	15.8	23.1	26.8	24.3	1.5	4.3	0.9	1.4	7.7	53.5	20.6	21.8	87.0	11.2	2.7	15.7	6.3	3.8	2.5	0.0	5.5	14.2	0.0	0.3	0.2	0.6	0.0	0.0	0.0	0.1	0.0	1.5
6	0.1	1.8	3.6	6.9	6.1	1.9	74.5	63.5	53.0	75.7	84.3	89.4	86.7	20.8	41.5	20.5	21.6	62.7	69.2	25.5	61.2	98.3	37.4	7.1	40.3	5.3	2.7	5.6	0.6	6.0	11.6	0.0	0.5	0.4	0.9	0.0	0.0	0.0	0.0	0.1	1.6
7	1.2	8.7	20.6	43.0	53.7	9.8	98.7	98.6	95.7	99.4	99.9	99.9	99.9	76.5	93.3	81.8	76.6	99.2	88.0	26.9	77.1	99.7	38.3	2.2	37.9	32.1	24.9	14.0	2.6	6.5	16.2	0.1	1.0	0.5	1.2	0.0	0.0	0.0	0.1	0.0	1.9
8	1.2	12.7	5.2	11.9	5.6	6.8	56.9	56.1	37.5	52.4	69.8	89.1	85.5	21.3	24.8	16.7	10.3	60.8	62.1	12.0	31.4	96.7	51.7	14.9	45.3	86.3	80.1	23.8	1.7	6.7	42.2	4.5	0.4	1.1	0.9	0.0	0.0	0.0	0.0	0.0	0.5
9	0.5	7.2	3.0	1.1	0.5	2.1	4.0	7.4	2.4	2.5	11.6	35.3	28.0	2.7	2.8	2.0	1.8	8.6	34.7	4.2	4.5	79.2	41.5	23.8	36.1	86.0	82.6	12.0	1.9	6.7	43.5	7.4	0.2	0.6	0.7	0.0	0.0	0.0	0.0	0.2	0.4
10	4.1	24.8	13.1	17.5	24.4	37.0	90.4	88.6	82.0	89.8	95.5	97.0	95.1	54.0	56.4	67.2	45.7	55.7	46.6	10.2	40.8	84.6	92.3	91.4	92.8	67.1	67.2	63.4	29.2	53.8	65.2	4.5	6.8	5.7	3.4	0.1	0.0	0.3	0.1	0.0	1.0
11	1.6	21.0	3.9	3.8	3.2	8.4	28.0	26.1	14.5	22.6	37.6	56.8	47.4	6.0	6.4	13.5	8.8	18.4	30.9	6.9	20.1	92.4	93.5	94.3	94.5	73.8	74.2	55.1	24.0	57.2	79.9	8.9	6.6	4.4	2.5	0.1	0.0	0.2	0.1	0.1	0.7
12	3.6	17.0	8.9	2.2	14.1	34.9	60.3	51.0	38.8	35.6	56.6	53.3	55.3	11.9	11.5	30.0	15.4	1.7	18.0	2.7	0.7	5.4	58.2	96.0	77.8	87.4	87.0	76.3	41.6	79.8	82.2	13.4	3.1	6.4	3.6	0.3	0.0	0.7	0.2	0.1	0.4
13	1.2	10.8	3.6	0.7	2.5	7.4	9.1	6.5	2.6	2.5	6.5	7.7	5.5	0.5	1.0	5.5	3.3	0.3	10.2	1.5	0.0	2.4	56.2	83.7	82.9	92.7	92.6	64.4	38.2	80.0	89.8	12.0	2.4	3.3	2.1	0.3	0.0	0.4	0.2	0.1	0.3
14	0.7	5.3	7.9	1.0	2.4	18.0	19.8	20.7	14.6	7.9	24.5	20.1	21.8	6.7	6.6	11.3	8.6	0.3	6.9	1.7	2.1	1.3	8.0	33.0	16.3	61.8	62.1	37.9	9.7	14.3	18.3	9.7	0.2	1.7	1.2	0.2	0.0	0.1	0.3	0.4	1.0
15	0.2	1.9	2.9	0.3	0.3	1.5	0.8	1.3	0.3	0.2	1.2	1.5	1.2	0.2	0.4	0.7	1.2	0.1	5.0	0.7	0.0	0.2	11.0	17.3	29.3	84.7	82.2	33.1	8.0	26.4	56.2	5.2	0.2	0.7	0.9	0.1	0.0	0.1	0.1	0.6	0.2
16	0.0	0.3	0.4	0.1	0.1	0.5	0.4	0.6	0.2	0.1	0.5	0.4	0.3	0.1	0.2	0.2	0.3	0.0	0.6	0.3	0.1	0.1	1.2	2.8	3.9	29.0	25.2	8.5	0.7	1.3	7.9	1.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.4	0.1
17	1.2	9.8	2.8	0.9	2.4	7.8	6.8	6.1	2.3	4.0	3.5	8.4	3.5	0.3	1.0	9.3	6.6	0.5	3.5	1.1	0.4	1.0	52.3	68.9	83.7	22.7	23.5	61.2	48.3	64.7	57.3	21.8	2.8	4.9	2.0	1.0	0.1	0.5	0.5	0.1	0.5
18	0.3	2.6	2.1	0.4	0.5	2.4	1.4	1.6	0.5	0.6	0.9	1.6	0.7	0.2	0.5	1.9	1.9	0.1	1.6	0.7	0.1	0.1	10.4	9.7	29.1	15.0	13.5	34.0	14.9	19.9	21.4	10.6	0.5	1.2	0.9	0.5	0.1	0.1	0.3	0.2	0.2
19	0.1	0.3	0.5	0.2	0.1	0.5	0.1	0.3	0.0	0.0	0.1	0.2	0.1	0.1	0.1	0.1	0.3	0.0	0.4	0.3	0.0	0.0	0.5	0.2	2.0	4.4	2.9	7.3	1.0	1.0	2.3	1.4	0.0	0.2	0.3	0.1	0.0	0.0	0.1	0.4	0.1
20	2.5	10.7	5.4	3.1	9.9	26.2	58.2	48.8	41.7	49.3	54.8	57.1	51.5	13.0	14.1	31.6	21.7	4.0	14.5	6.7	15.6	20.9	56.8	97.1	70.5	5.4	5.6	33.8	31.1	52.6	38.4	7.4	3.4	6.9	3.8	0.5	0.1	0.7	0.3	0.0	1.0
21	0.2	0.8	2.0	1.3	7.2	11.3	32.3	15.4	11.5	5.7	18.6	8.4	12.4	2.1	1.2	3.2	2.3	0.5	15.1	6.5	0.6	4.7	17.0	68.7	33.3	8.7	6.4	37.2	9.6	65.4	87.7	9.2	1.3	7.1	5.3	0.1	0.2	1.4	0.0	0.0	0.8
22	0.1	0.1	1.1	0.6	6.2	2.4	7.8	1.8	0.6	0.1	1.5	0.8	1.5	0.1	0.1	0.7	0.7	0.1	8.5	1.0	0.0	0.0	5.3	8.4	14.6	15.5	9.1	50.0	9.6	77.5	94.1	22.9	0.5	5.6	4.6	0.1	0.0	1.4	0.0	0.1	0.7
23	0.0	0.1	0.1	0.4	2.0	1.6	4.9	1.2	0.5	0.2	0.9	0.2	0.3	0.0	0.1	0.1	0.1	0.0	1.4	1.1	0.0	0.0	0.5	2.6	1.4	0.5	0.1	5.4	1.3	19.4	36.8	2.5	0.1	1.4	1.5	0.1	0.2	0.3	0.0	0.0	0.2
24	0.3	1.8	2.1	0.9	3.2	3.8	4.0	2.3	0.9	0.6	1.1	3.6	1.9	0.1	0.4	3.7	3.9	0.3	2.2	1.0	0.1	0.1	6.0	4.5	17.2	1.3	0.2	15.6	29.8	29.3	7.1	49.5	1.3	4.7	2.2	1.0	0.0	0.6	0.3	0.1	0.3
25	0.1	0.3	0.8	0.5	0.5	0.6	0.3	0.3	0.1	0.0	0.1	0.5	0.3	0.1	0.1	0.4	0.7	0.1	0.8	0.4	0.0	0.1	0.4	0.1	1.4	0.8	0.1	2.0	8.8	2.8	0.4	60.1	0.4	1.1	0.9	0.6	0.1	0.2	0.3	1.7	0.3
26	0.1	0.2	0.6	0.2	0.2	0.2	0.1	0.2	0.0	0.0	0.1	0.3	0.2	0.0	0.1	0.2	0.4	0.0	0.6	0.3	0.0	0.0	0.3	0.1	0.8	0.2	0.0	0.8	2.3	0.9	0.2	4.2	0.1	0.2	0.4	0.1	0.0	0.0	0.1	0.1	0.0
27	0.0	0.5	4.4	0.4	1.3	1.2	1.3	0.7	0.3	0.1	0.7	2.1	2.4	0.1	0.1	1.6	2.7	0.1	21.7	1.4	0.0	0.0	1.1	1.1	3.5	4.6	1.2	9.9	3.0	7.1	31.7	34.0	0.2	0.7	1.1	0.0	0.0	0.1	0.0	0.3	0.1
28	0.0	0.0	0.2	0.3	0.0	0.4	0.1	0.2	0.2	0.0	0.3	0.1	0.1	0.1	0.0	0.1	0.1	0.0	1.3	0.3	0.0	3.3	0.2	0.5	0.5	13.8	3.4	5.7	1.3	3.0	1.5	68.8	0.1	2.7	2.7	0.3	0.2	0.8	0.4	43.0	74.9
29	4.6	8.4	11.1	6.6	20.4	54.7	88.5	88.1	89.6	86.6	95.3	86.3	86.9	68.1	60.2	67.6	42.6	10.4	13.6	3.8	24.2	7.6	24.7	84.4	25.8	4.9	5.6	14.9	17.6	21.7	5.0	2.9	0.9	4.8	3.1	0.4	0.1	0.4	0.8	0.2	4.4
30	1.2	3.6	8.4	2.4	2.6	13.5	24.9	34.5	34.4	24.6	52.1	60.1	64.6	27.4	23.8	20.7	16.7	3.2	9.1	2.9	12.0	6.9	8.8	35.8	6.5	2.8	2.9	4.4	3.7	3.0	1.0	2.8	0.1	0.9	1.0	0.0	0.1	0.1	0.4	0.6	3.4
31	1.7	7.6	4.7	1.7	2.4	6.8	17.7	18.4	21.5	37.9	31.6	35.4	20.1	9.3	13.0	48.3	57.7	3.3	1.2	5.9	69.0	10.5	16.1	41.8	11.0	0.6	0.5	1.2	10.7	2.2	0.0	1.1	1.4	2.0	1.3	1.1	0.5	0.2	0.8	0.2	1.1
32	1.4	1.6	1.0	1.9	8.1	50.1	72.4	57.2	60.9	42.5	57.6	12.1	14.8	23.6	19.5	16.0	5.4	0.3	1.6	1.7	3.6	0.1	2.1	41.4	5.0	1.8	1.7	12.2	10.9	17.1	4.1	4.1	0.9	6.5	3.0	1.3	0.3	0.6	0.5	0.3	3.8
33	1.2	4.6	0.9	0.9	1.2	9.8	12.9	10.3	7.0	12.7	10.5	9.8	5.7	1.3	2.1	6.3	4.1	0.4	2.6	4.3	8.6	1.5	16.3	77.1	23.5	0.7	0.5	5.1	10.3	10.8	3.9	1.8	1.1	2.9	1.6	0.7	0.3	0.2	0.4	0.1	0.4
34	0.2	0.7	2.0	0.5	0.7	5.4	7.2	6.2	4.9	1.2	8.9	5.7	6.2	2.9	2.2	2.3	2.3	0.1	1.5	0.9	1.0	0.2	0.6	6.3	1.0	1.9	1.7	5.2	2.6	2.2	0.6	1.4	0.0	0.5	0.6	0.0	0.0	0.0	0.1	0.2	1.0
35	0.6	1.1	0.1	1.1	1.3	32.1	29.9	20.0	19.6	16.4	11.8	1.1	0.5	3.2	5.4	4.0	1.5	0.2	0.3	1.0	2.0	0.1	0.3	2.6	0.6	0.4	0.2	1.9	2.6	2.2	0.5	1.7	0.9	4.2	2.0	2.6	0.8	0.3	0.5	0.2	0.6
36	0.2	0.3	0.0	0.5	0.2	5.7	2.5	2.1	1.0	1.6	0.8	0.1	0.1	0.3	0.6	0.3	0.1	0.1	0.1	0.3	0.3	0.0	0.1	0.3	0.1	0.2	0.0	0.2	0.4	0.4	0.1	0.4	0.3	1.3	0.8	0.8	0.5	0.1	0.3	0.3	0.1
37	0.1	0.2	0.0	0.2	0.1	0.7	0.1	0.2	0.0	0.1	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.0	0.3	0.1	0.5	0.4	0.2	0.2	0.1	0.1	0.2	0.1
38	0.4	2.5	0.6	0.8	0.2	1.9	1.0	1.9	1.4	3.8	2.2	2.4	0.9	0.8	1.4	4.8	9.2	0.4	0.1	0.7	33.0	1.3	1.5	5.9	0.9	0.2	0.1	0.2	1.7	0.3	0.0	0.3	0.8	1.0	0.9	1.0	1.1	0.2	0.6	0.2	0.3
39	0.1	0.4	0.5	0.6	0.1	1.2	0.7	0.8	1.1	0.9	1.0	0.4	0.2	0.3	0.4	1.2	2.0	0.1	4.1	86.2	12.3	1.4	1.6	5.0	1.2	0.2	0.0	0.3	0.6	0.7	0.7	0.2	0.8	2.5	1.2	0.2	0.8	0.1	0.0	0.2	0.6
40	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
41	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.1	0.5	0.3	0.1	0.1	1.0	2.0	0.4
42	0.1	0.1	0.0	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.2	0.1	0.6	0.6	0.1	0.0	0.1	0.1	0.1	0.0	0.0	0.0	1.8	0.5	0.0	0.5	1.3	2.0	2.3	8.7	1.2	0.8	5.6	2.3	4.4
43	0.0	0.2	0.0	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.3	0.4	1.0	0.1	0.5	0.5	0.1
44	0.6	1.1	0.0	1.2	0.2	0.7	0.1	0.4	0.0	0.4	0.1	0.1	0.0	0.0	0.2	0.1	0.1	0.3	0.1	0.6	2.0	0.2	0.2	0.3	0.3	0.1	0.0	0.0	0.3	1.1	0.0	1.6	3.3	5.7	3.1	9.0	9.9	0.9	2.5	0.4	1.0
45	0.1	0.7	0.0	0.5	0.0	0.4	0.1	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.6	0.6	0.4	0.1	0.4	0.1	0.1	0.0	0.0	0.0	0.1	0.2	0.3	0.5	1.0	0.9	0.4	16.4	0.2	0.3	1.1	0.4
46	1.6	0.7	0.4	1.4	4.2	6.1	5.2	3.6	0.8	4.5	2.4	0.7	0.6	0.4	0.5	0.7	0.4	0.5	0.6	2.0	1.1	2.7	2.1	28.4	5.5	0.4	0.1	1.1	8.4	63.8	16.5	5.9	17.8	48.9	25.0	9.8	7.8	11.9	3.7	1.6	17.0
47	0.0	0.1	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.0	0.6	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	1.9	0.1	0.4	0.7	0.3	3.1	0.1	0.7	32.3	1.8
48	0.0	0.0	0.0	0.2	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.2	0.1	0.0	0.2	0.2	0.1	0.0	0.1	0.0	0.3	0.0	0.1	0.2	0.1	0.0	1.0	0.1	0.5	0.9	0.6	0.3	0.7	0.7	11.9	37.8
49	2.3	0.5	0.6	4.3	9.5	0.2	0.3	0.3	0.2	0.2	0.3	0.1	0.2	0.2	1.4	0.4	0.7	2.4	0.9	3.0	4.3	3.0	0.3	0.5	0.7	1.2	0.1	0.1	10.4	4.4	0.3	16.1	9.0	12.5	38.0	15.3	5.6	49.1	24.5	49.3	85.1
50	21.0	4.2	7.8	16.1	42.7	1.8	1.1	2.0	0.9	0.8	1.1	0.5	0.7	1.5	6.2	2.4	2.9	12.7	14.1	25.3	16.8	36.0	8.4	17.5	14.1	8.3	1.5	1.3	60.0	40.0	9.2	80.0	68.2	70.4	84.9	64.8	28.1	87.5	76.1	79.8	97.1
51	78.3	38.3	65.7	68.2	92.6	20.6	21.4	30.6	27.7	19.7	20.9	12.5	18.3	25.3	54.7	34.7	43.4	65.5	69.6	79.8	72.7	95.1	60.9	73.4	75.8	48.8	24.9	25.3	95.2	85.3	59.0	96.0	96.5	98.4	98.6	90.5	88.5	97.2	96.4	92.5	99.6

- Multi-variate HMM emits vector of values, not just one value
- Can emit real values (SegWay) or binary presence /absence values (ChromHMM)
- Use to learn mark combinations

Ernst and Kellis
Nature Biotech 2010

Source: Ernst, Jason and Manolis Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature Biotechnology 28, no. 8 (2010): 817-825.

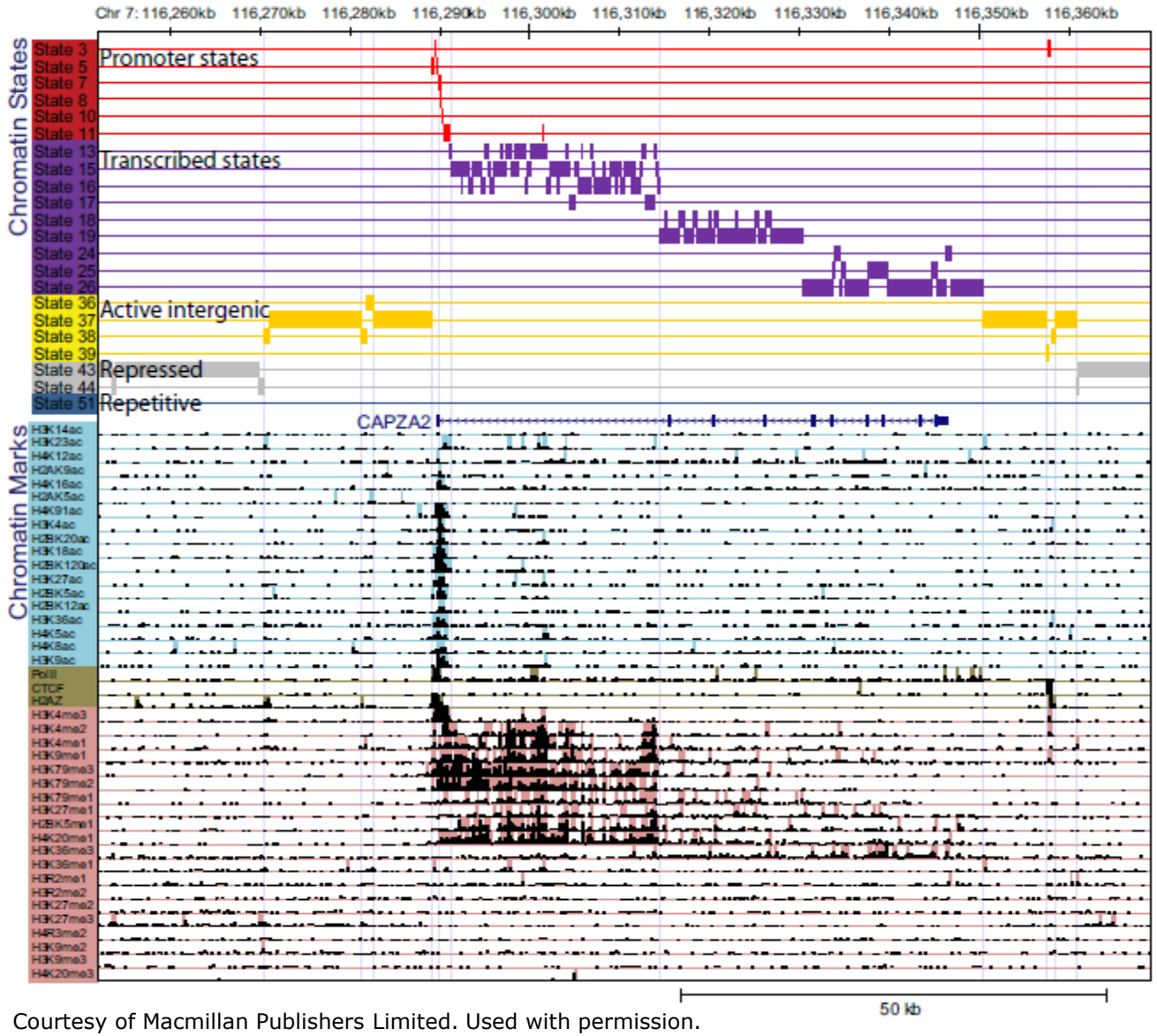
Transition matrix a_{kl}



- Learns spatial relationships between neighboring states
- Reveals distinct sub-groups of states
- Reveals transitions between different groups

Source: Ernst, Jason and Manolis Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature Biotechnology 28, no. 8 (2010): 817-825.

Example Chromatin State Annotation

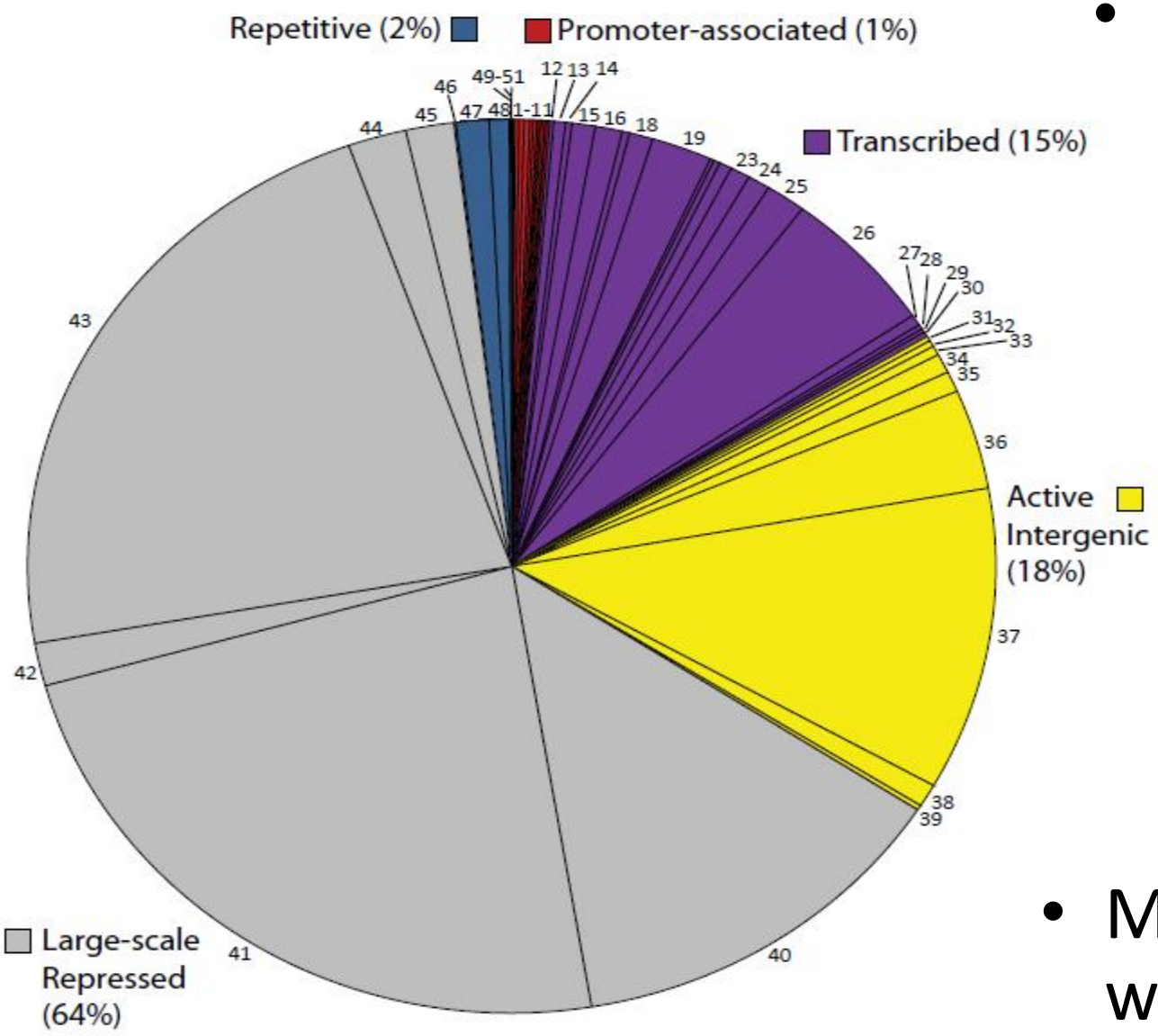


Courtesy of Macmillan Publishers Limited. Used with permission.

Source: Ernst, Jason and Manolis Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature Biotechnology 28, no. 8 (2010): 817-825.

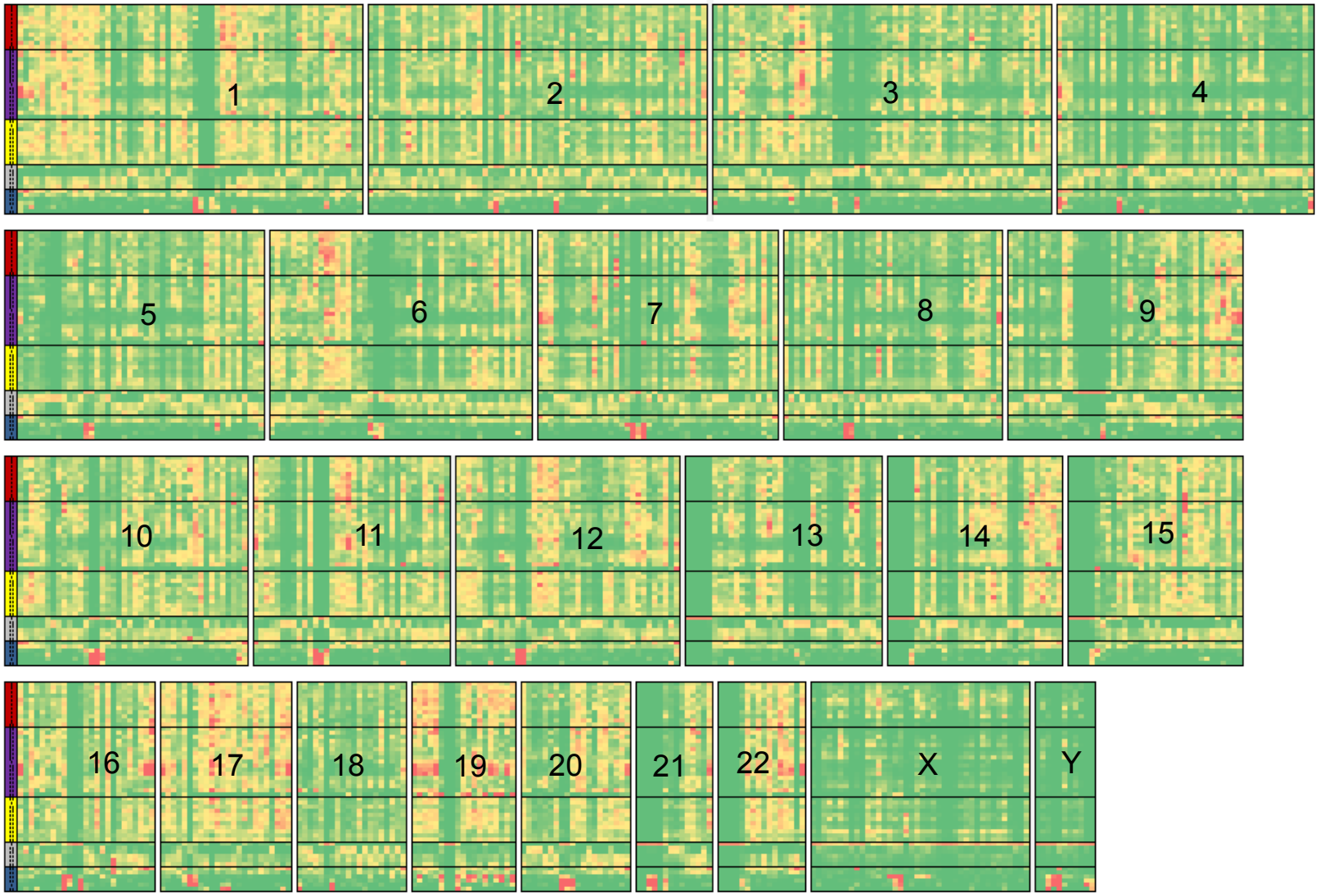
- Use Baum Welch to learn hidden states and their annotations
- Learned states correspond to known functional elements
- *De novo* discovery of major types of chromatin

Model complexity matches that of genome



- Handful of repressed states capture vast majority of genome
  - Only 1% of genome split in 14 promoter states
- Modeling power well distributed where needed

Apply genome wide to classify chromatin states *de novo*



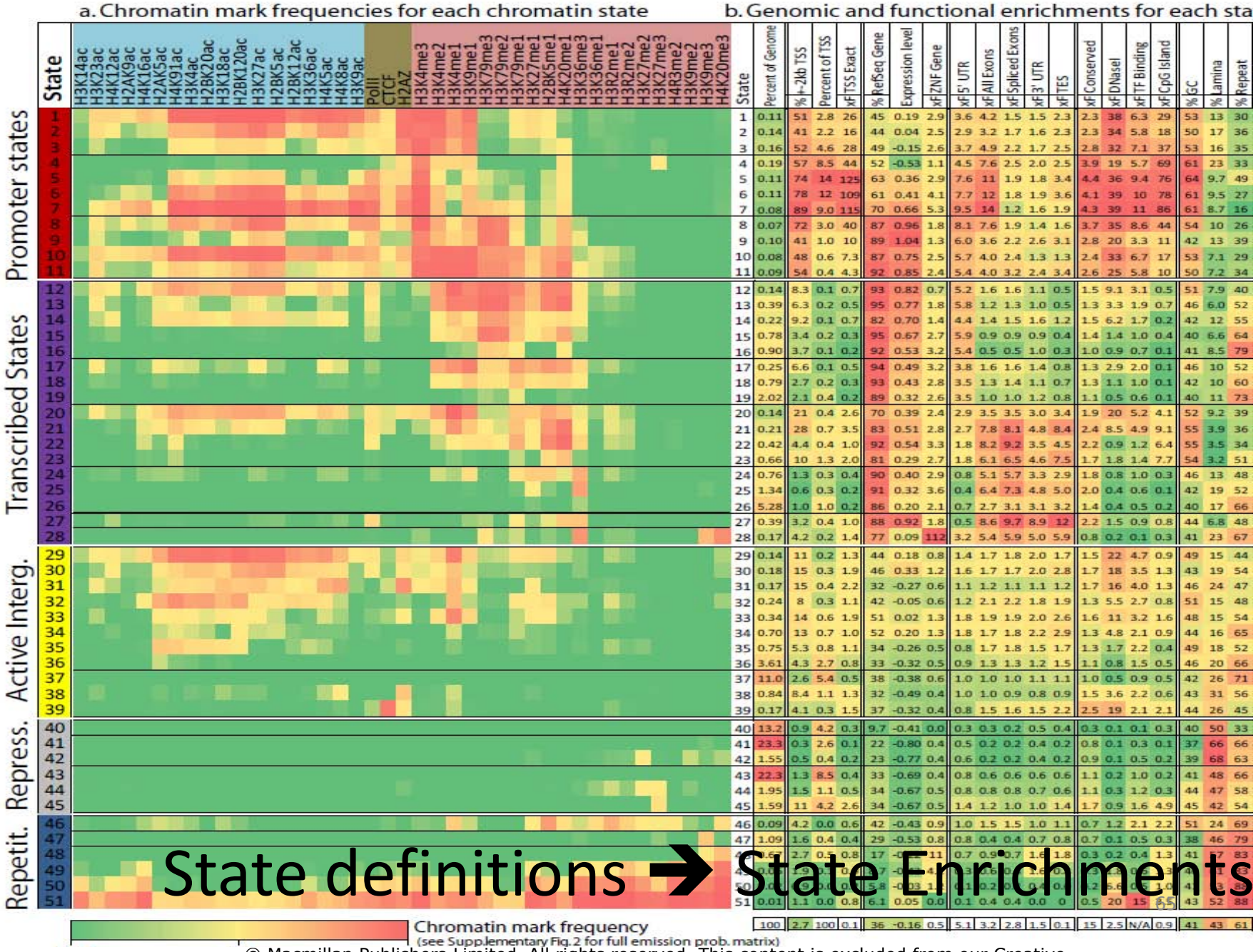
Source: Ernst, Jason and Manolis Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature Biotechnology 28, no. 8 (2010): 817-825.

Now what? Interpret these states biologically

Goals for today: Computational Epigenomics

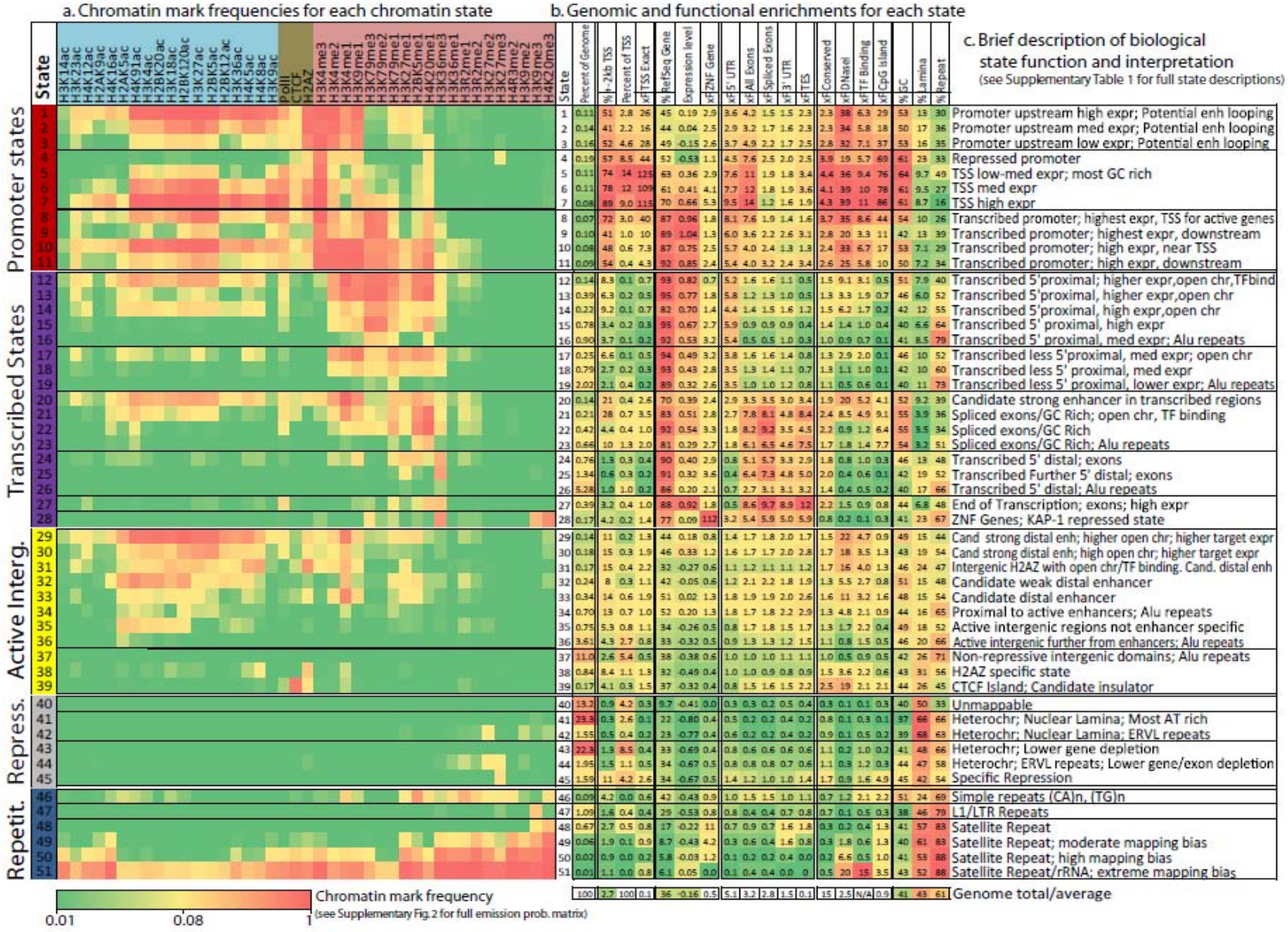
1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Chromatin state characterization: Functional/positional enrichment
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

(Future: Chromatin states to interpret disease-associated variants)



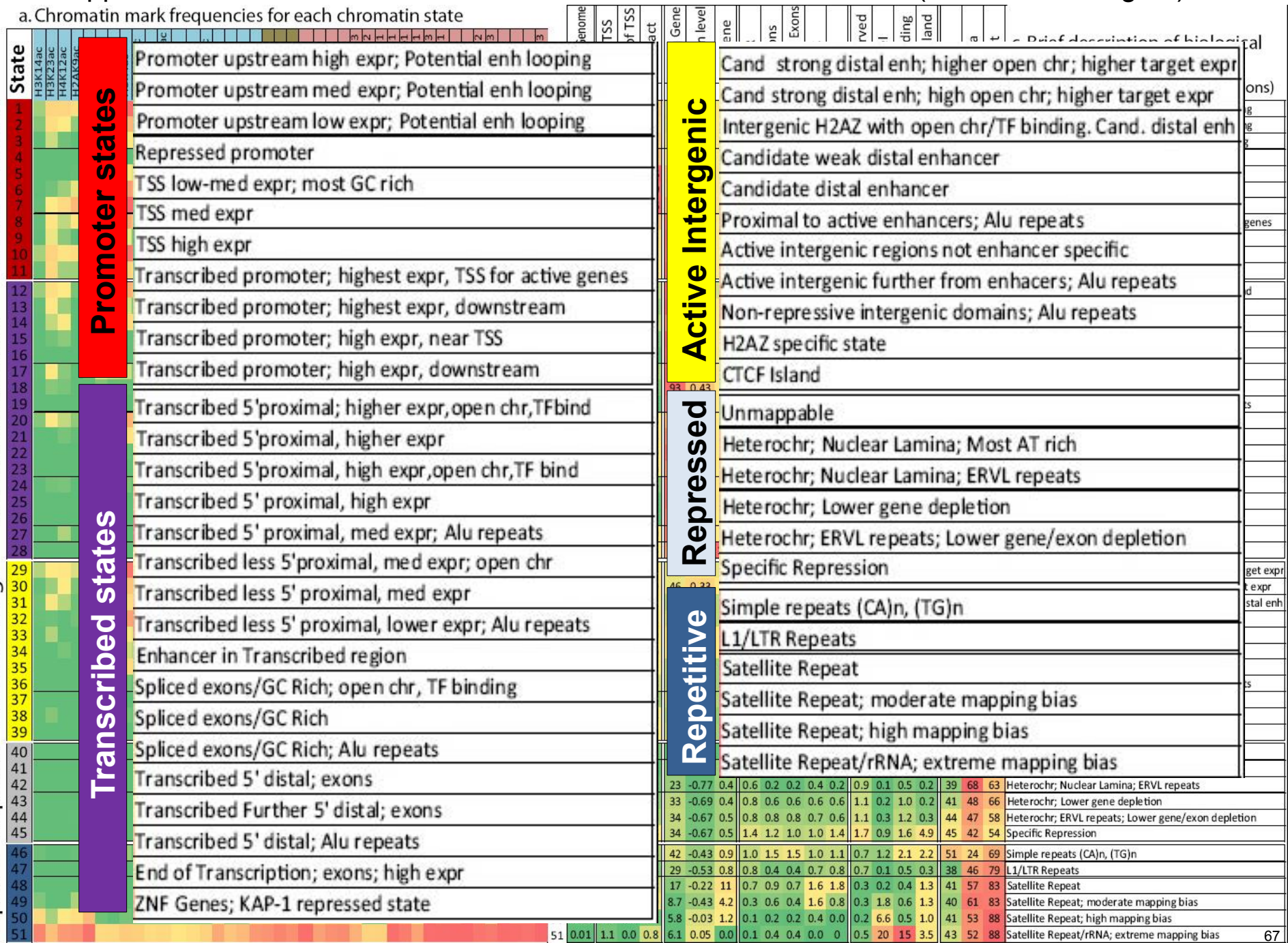
Source: Ernst, Jason and Manolis Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature Biotechnology 28, no. 8 (2010): 817-825.

Functional enrichments enable annotation of 51 distinct states



Source: Ernst, Jason and Manolis Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature Biotechnology 28, no. 8 (2010): 817-825.

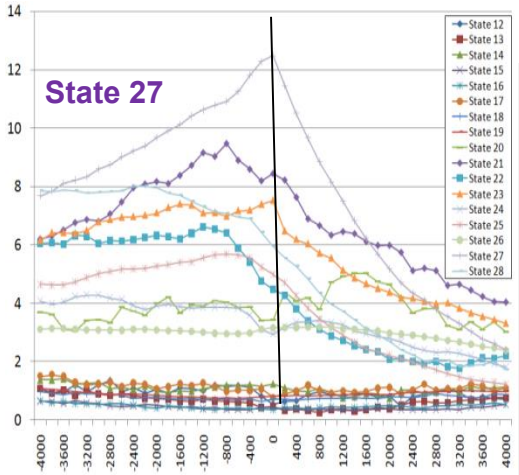
Application of ChromHMM to 41 chromatin marks in CD4+ T-cells (Barski'07, Wang'08)



Functional properties of discovered chromatin states

GO Category	State 3	State 4	State 5	State 6	State 7	State 8
Cell Cycle Phase	2.10 ($2x10^{-7}$)	0.57 (1)	1.61 (0.001)	1.45 (1)	1.15 (1)	1.51 (1)
Embryonic Development	1.24 (1)	2.82 ($9x10^{-23}$)	1.07 (1)	0.85 (1)	0.54 (1)	1.00 (1)
Chromatin	1.20 (1)	0.48 (1)	2.2 ($1.4x10^{-7}$)	1.64 (1)	0.85 (1)	0.85 (1)
Response to DNA Damage Stimulus	1.20 (1)	0.35 (1)	1.55 (0.074)	2.13 ($6.5x10^{-11}$)	1.97 ($1.0x10^{-4}$)	0.84 (1)
RNA Processing	0.49 (1)	0.26 (1)	1.31 (1)	1.91 ($4.2x10^{-11}$)	2.64 ($8.7x10^{-24}$)	2.45 ($3.0x10^{-4}$)
T cell Activation	0.77 (1)	0.88 (1)	1.27 (1)	0.70 (1)	0.79 (1)	4.72 ($2x10^{-7}$)

Promoter state ➔ gene GO function

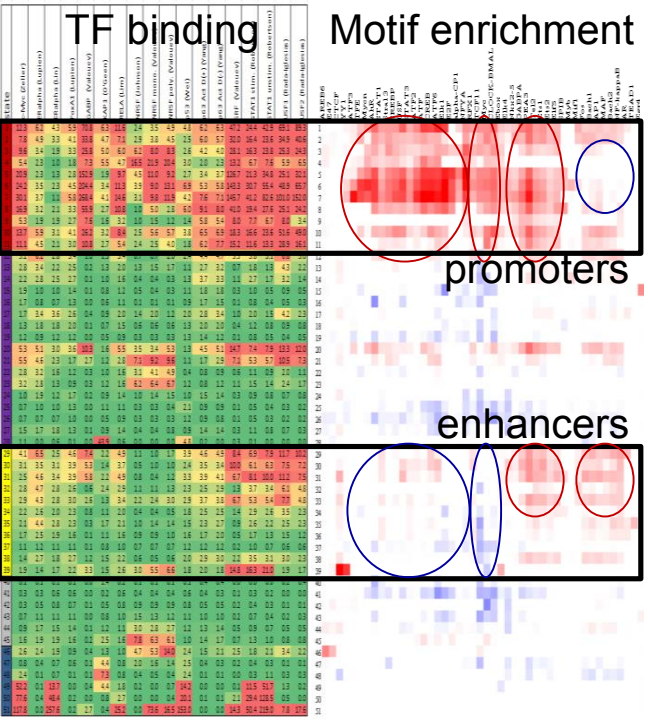


Transcription End State

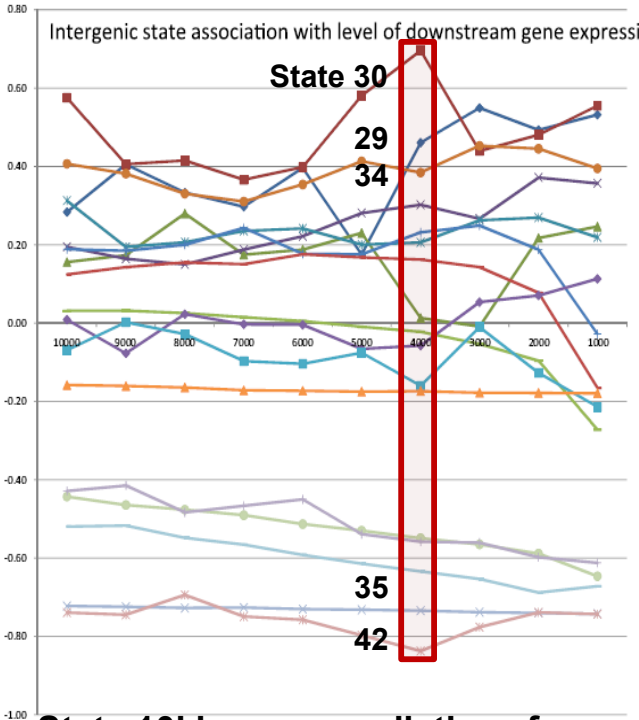
State 28: 112-fold ZNF enrich

"The achievement of the repressed state by wild-type KAP1 involves decreased recruitment of **RNA polymerase II**, reduced levels of histone **H3 K9 acteylation** and **H3K4 methylation**, an increase in **histone occupancy**, enrichment of **trimethyl histone H3K9**, **H3K36**, and **histone H4K20** …" MCB 2006.

ZNF repressed state recovery



Promoter vs. enhancer regulation



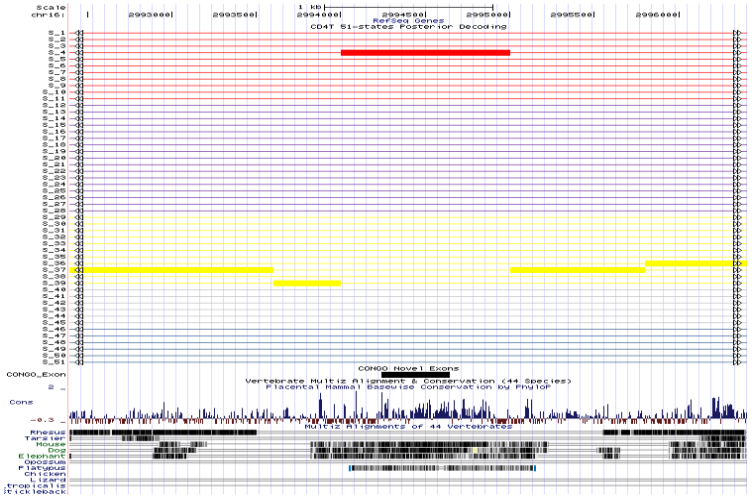
State 10kb away predictive of expr.

state	stalk	variable heterocrhomatic	acrocentric	gneg	gpos25	gpos50	gpos75	gpos100
40	7.6	6.6	6.1	0.6	0.4	0.5	0.3	0.5
41	0.0	0.2	0.4	0.5	0.4	0.9	1.7	2.5
42	0.0	0.2	0.3	0.5	0.5	1.1	1.7	2.2
43	0.0	0.1	0.1	1.1	1.2	1.3	1.1	0.7
44	0.0	0.1	0.1	1.2	1.3	1.3	1.0	0.6
45	0.0	0.1	0.1	1.3	1.6	1.3	0.8	0.4
46	0.0	0.1	0.2	1.7	1.7	0.7	0.4	0.2
47	0.0	0.2	0.1	1.2	1.3	1.3	0.9	0.6
48	0.0	3.2	6.2	0.8	2.2	0.2	0.5	0.4
49	0.0	3.6	11.2	0.5	1.8	0.2	0.4	0.1
50	0.0	4.7	12.0	0.6	0.6	0.2	0.2	0.1
51	0.0	4.4	12.7	0.5	1.4	0.2	0.2	0.1
% Overall	0.6	3.9	3.6	42.1	6.8	13.6	13.1	16.2

Distinct types of repression
- Chrom bands / HDAC resp
- Repeat family / composition

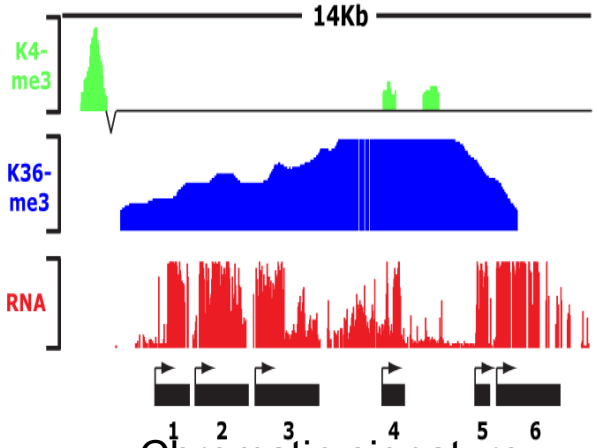
Applications to genome annotation

New protein-coding genes

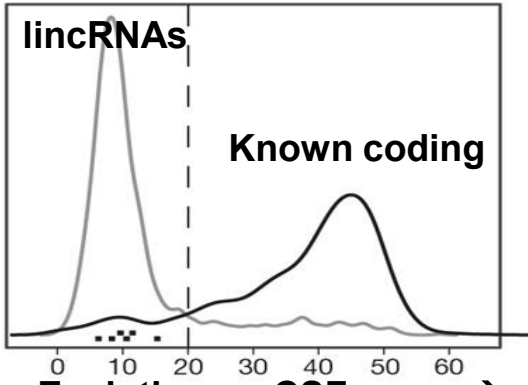


In promoter(short)/low-expr states

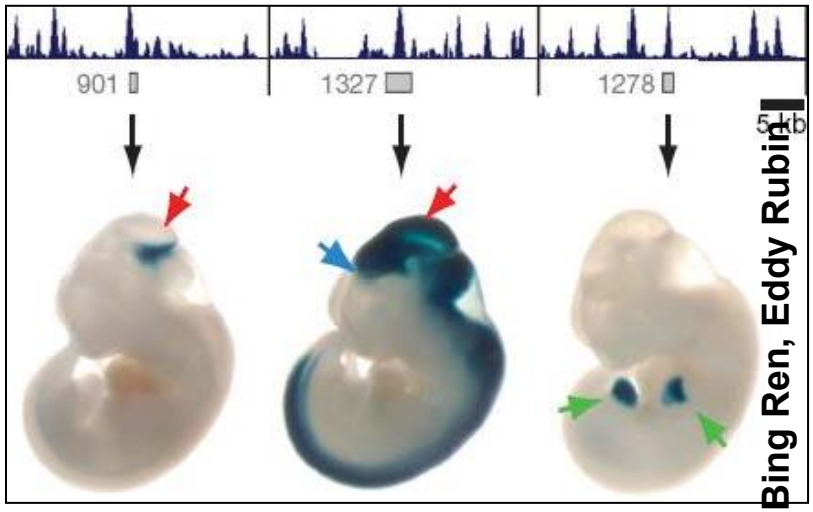
Long intergenic non-coding RNAs/lincRNAs



Chromatin signature: promoter / transcribed

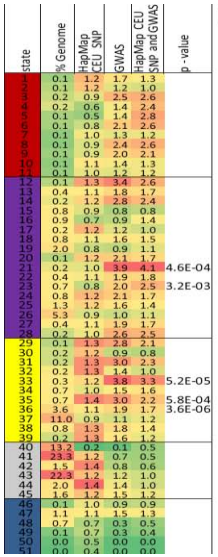


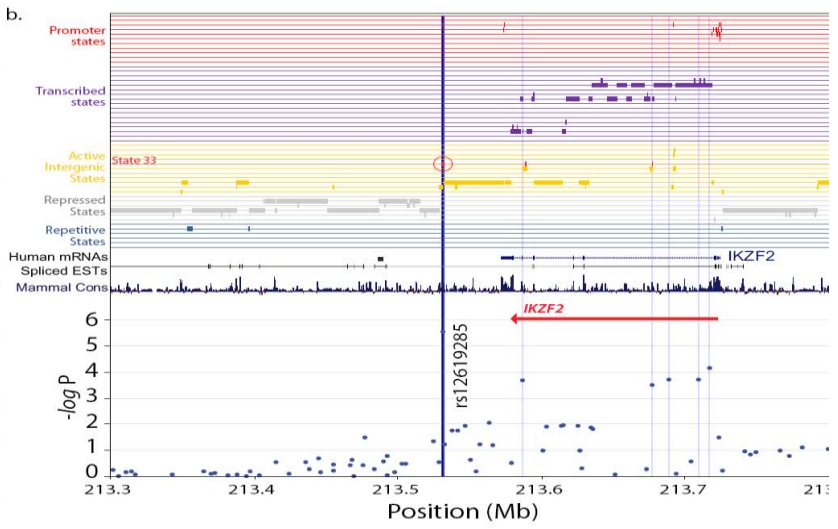
Evolutionary signature: not protein-coding



Bing Ren, Eddy Rubin

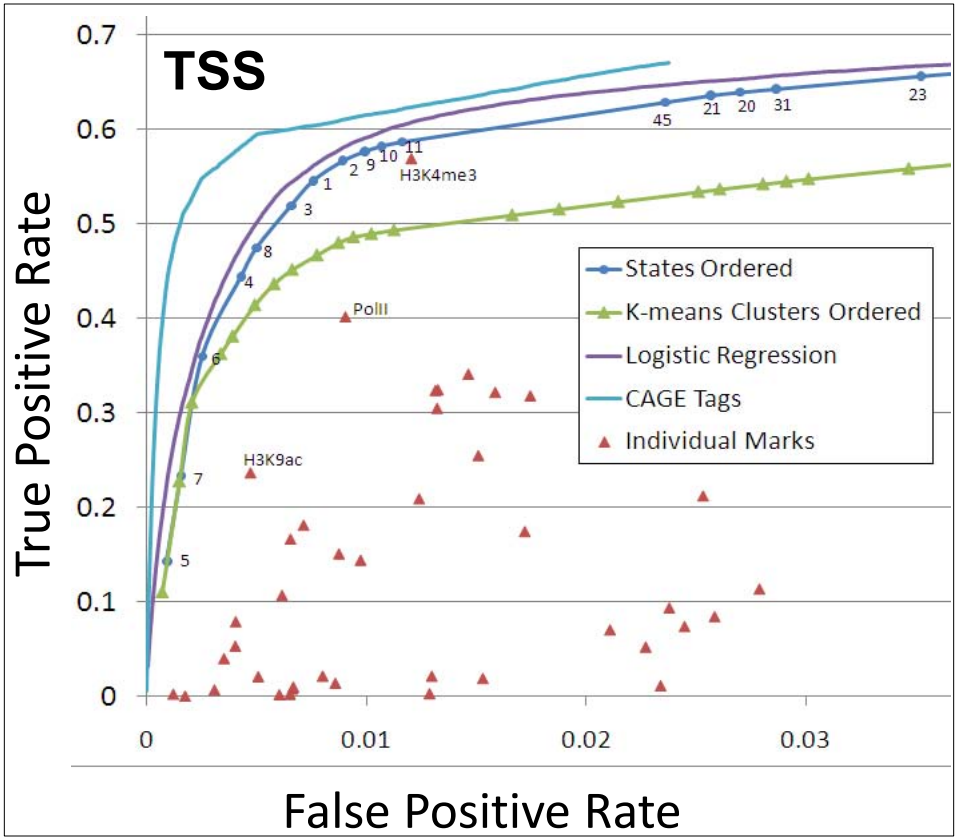
New developmental enhancer regions

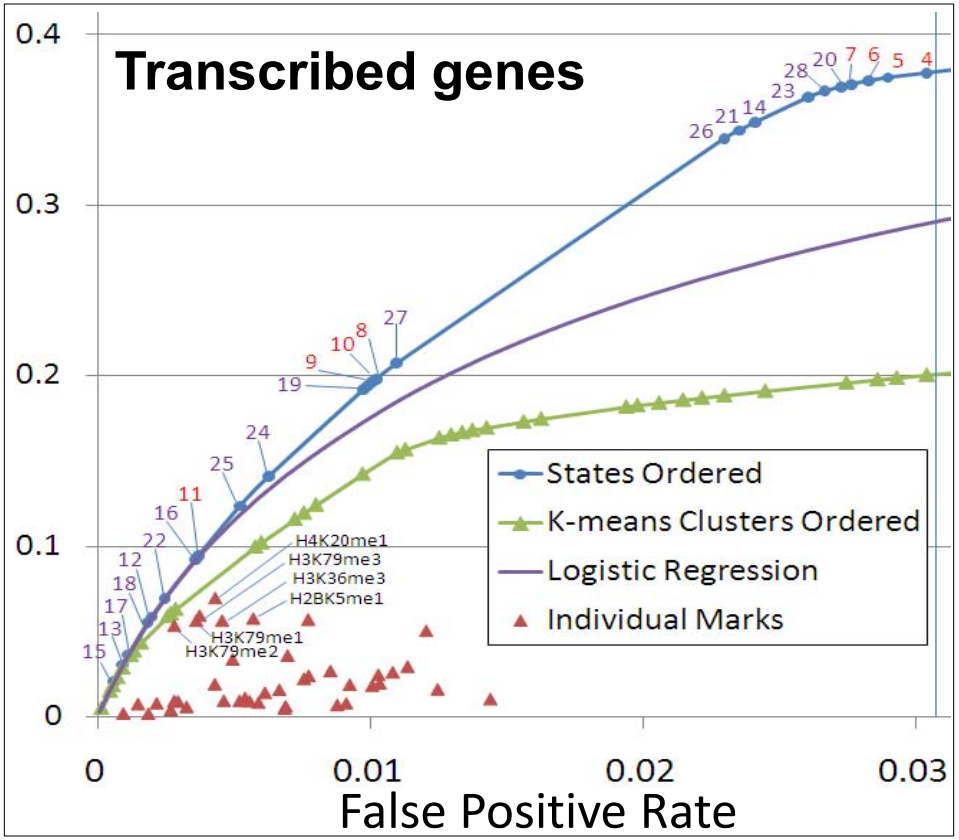




Assign candidate functions to intergenic SNPs from genome-wide association studies

Discovery power for promoters, transcripts





- Significantly outperforms single-marks
- Similar power to supervised learning approach
- CAGE experiments give possible upper bound

Goals for today: Computational Epigenomics

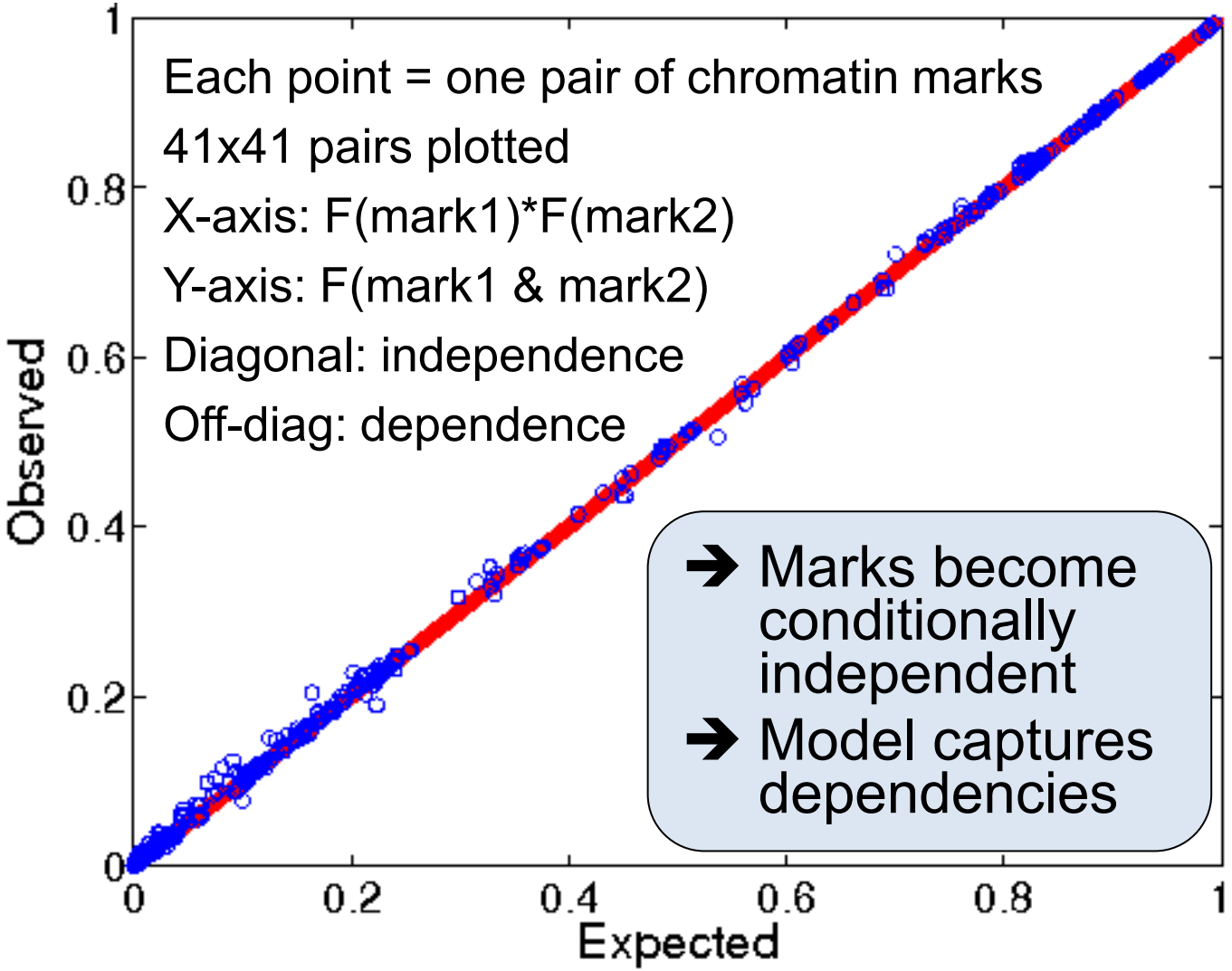
1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Capturing dependencies. State-conditional mark independence
   - Selecting the number of states, selecting number of marks
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

(Future: Chromatin states to interpret disease-associated variants)

State-conditional mark independence

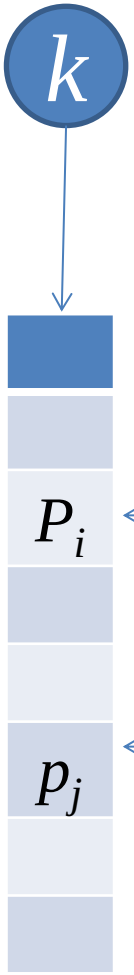
Do hidden states actually capture dependencies between marks?

Pairwise Expected vs. Observed Mark Co-Occurence



➔ Marks become conditionally independent

➔ Model captures dependencies



p_i emission prob for mark i

$q_{i,j}$ freq w/ which marks i and j co-occur

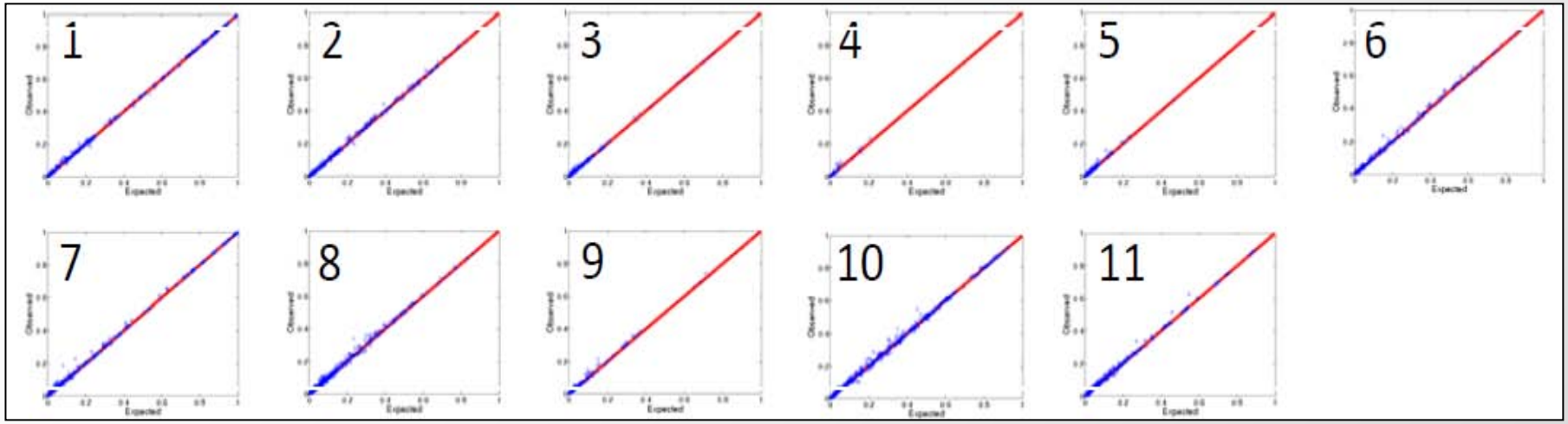
Test each pair of chromatin marks

$$q_{i,j} \stackrel{?}{=} p_i * p_j$$

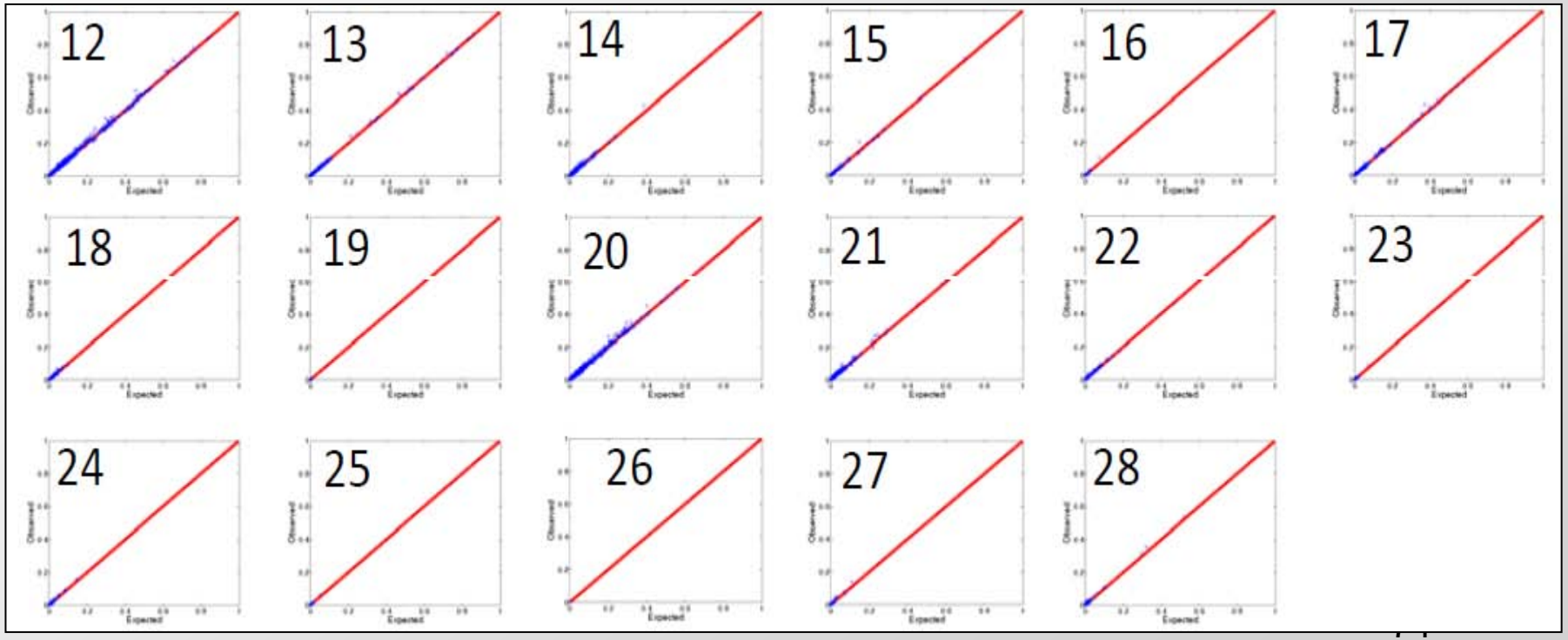
Multi-variate HMM emits entire vector of marks at a time
Model assumes mark independence *conditional* upon state
In fact, it specifically seeks to *capture* these dependencies

Test conditional independence for each state

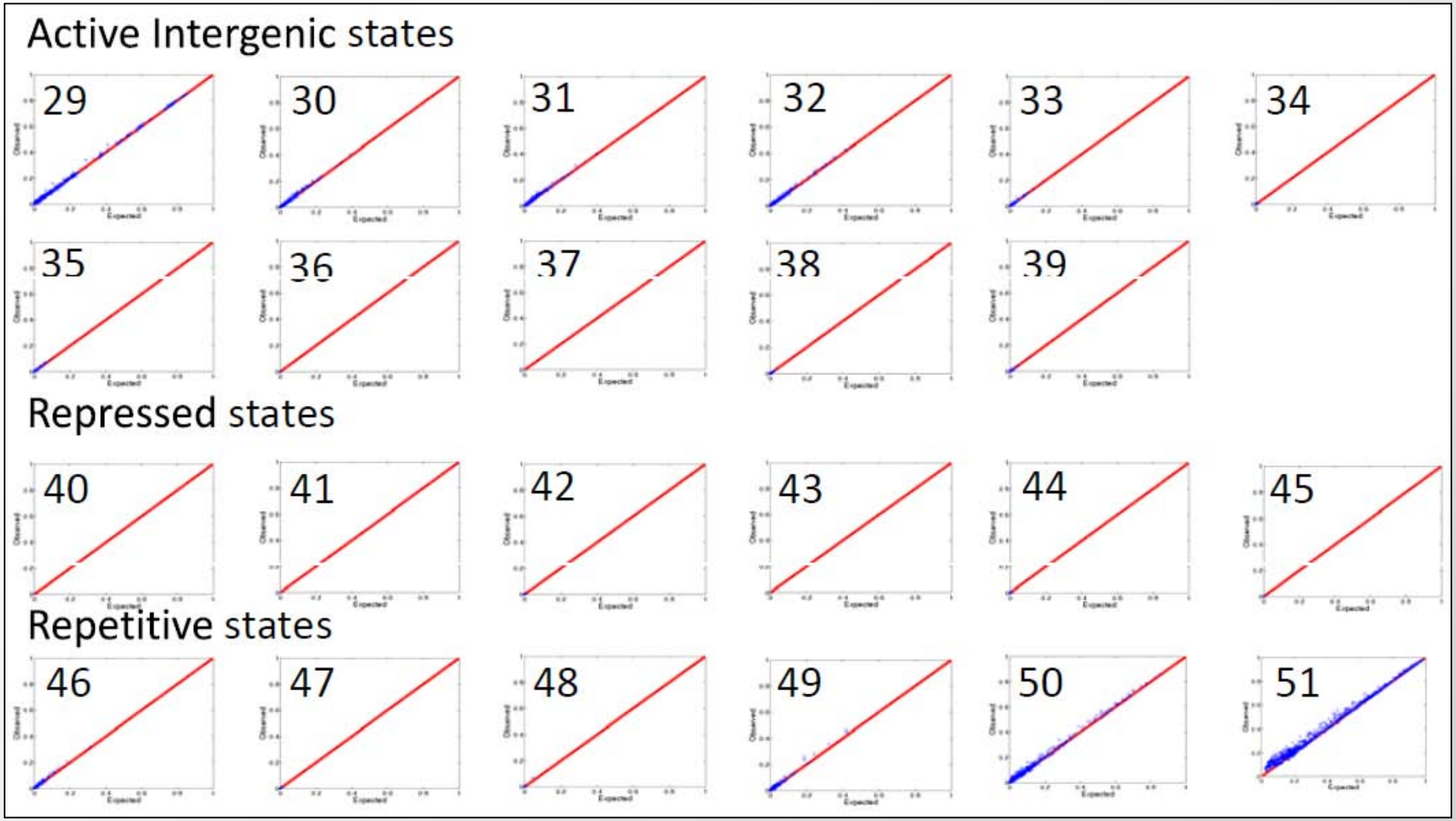
Promoter states



Transcribed states

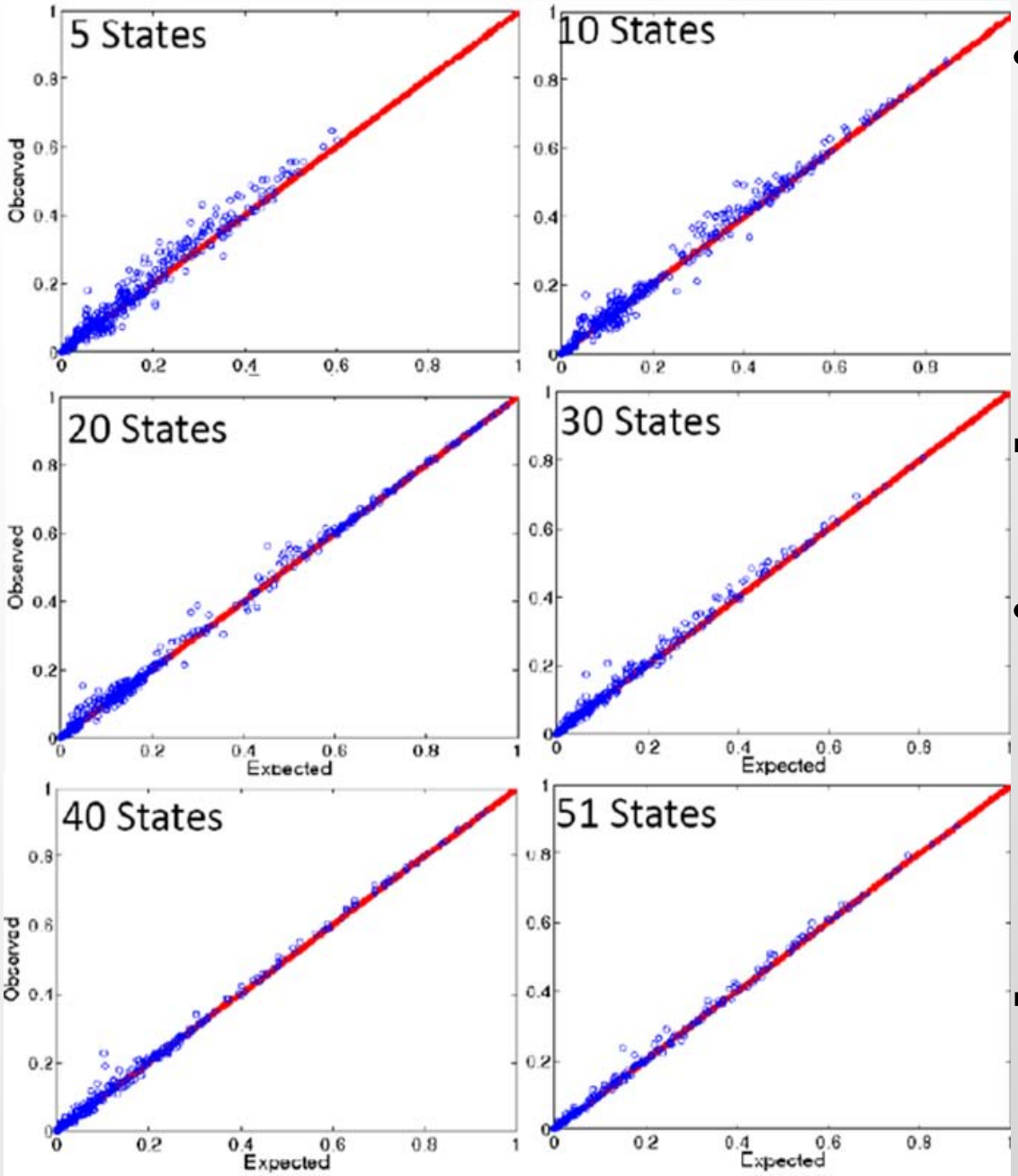


Non-independence reveals cases of model violation



- **Repetitive states show more dependencies**
- **Conditional independence does not hold**

As more states are added, dependencies captured



- With only 5 states in HMM, not enough power to distinguish different properties

➔ Dependencies remain

- As model complexity increases, states learned become more precise

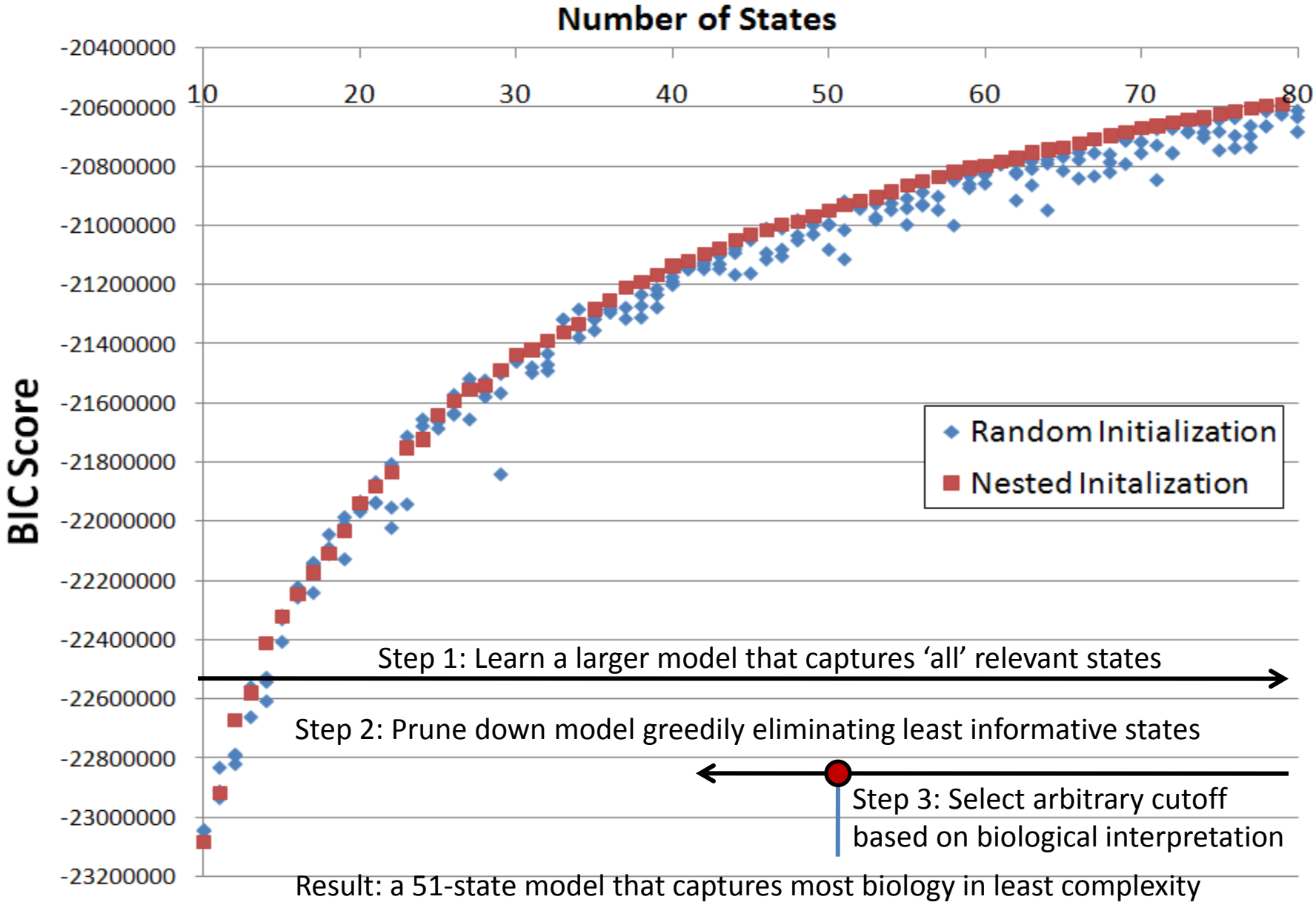
➔ Dependencies captured

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Capturing dependencies. State-conditional mark independence
   - Selecting the number of states, selecting number of marks
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

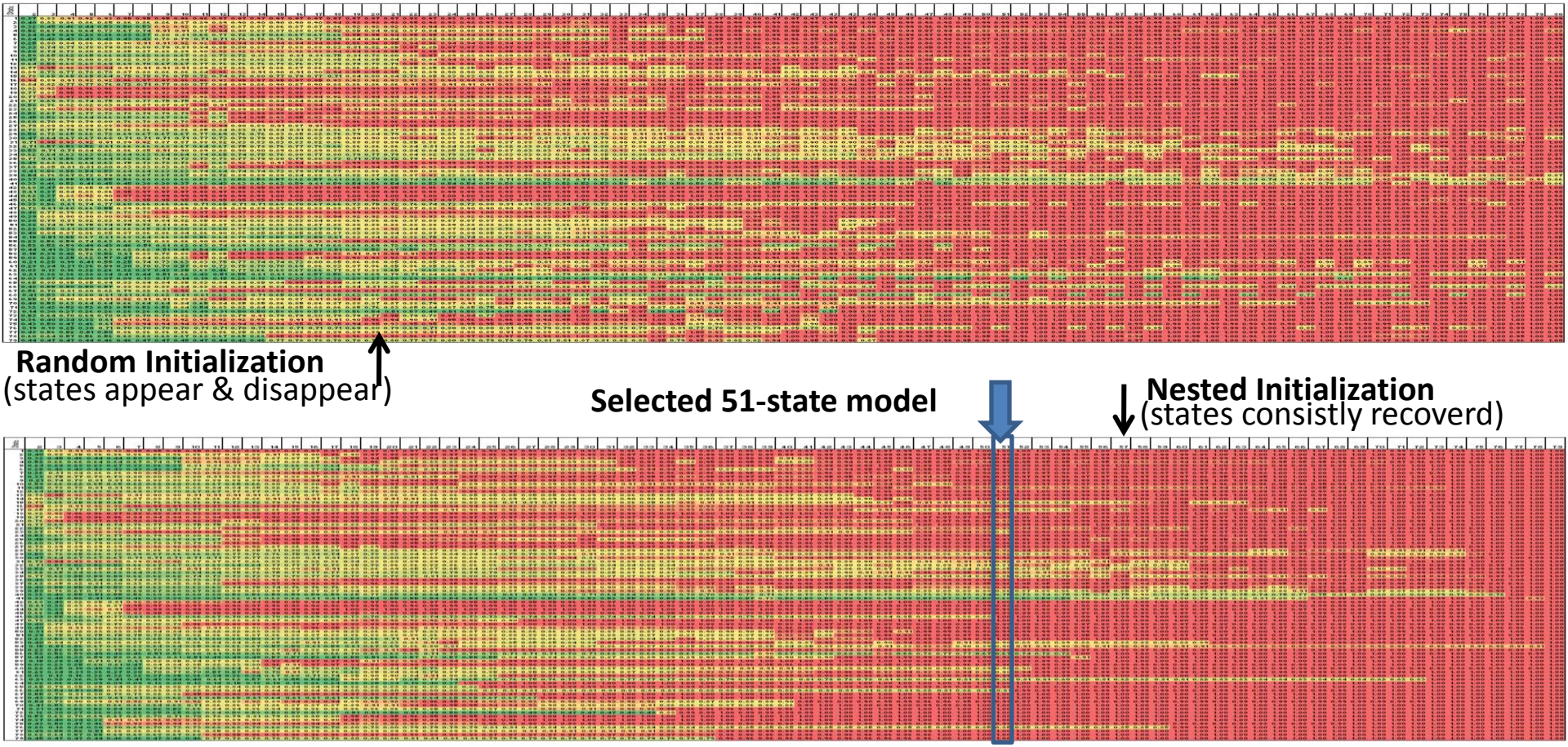
(Future: Chromatin states to interpret disease-associated variants)

Comparison of BIC Score vs. Number of States for Random and Nested Initialization



- Standard model selection criteria fail due to genome complexity: more states always preferred
- Instead: Start w/complex model, keep informative states, prune redundant states. Pick cutoff

Recovery of 79-state model in random vs. nested initialization



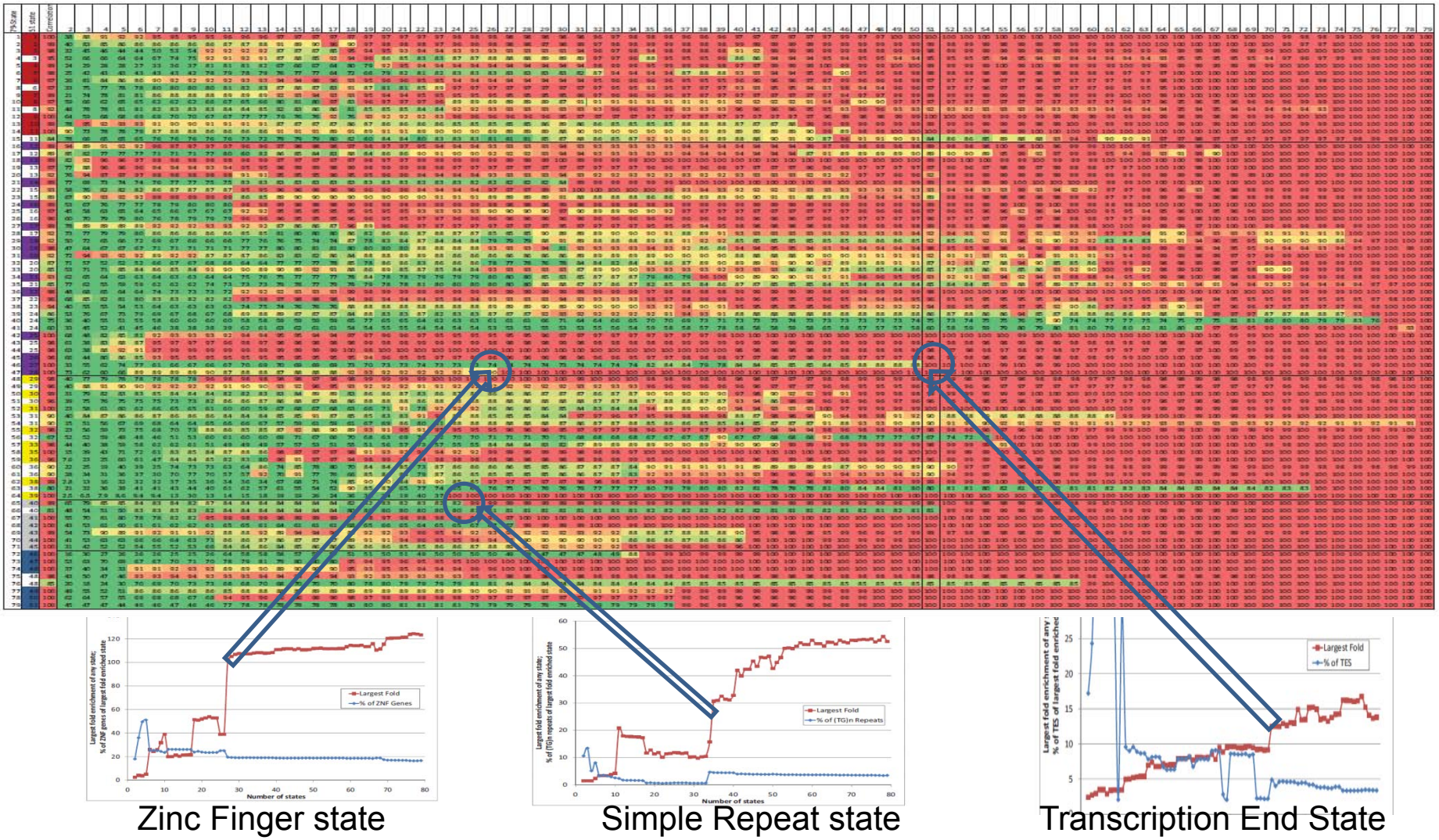
Nested initialization approach:

- **First pass:** learn models of increasing complexity
- **Second pass:** form nested set of emission parameter initializations by greedily removing states from best BIC model found

Nested models criteria:

- Maximize sum of correlation of emission vectors with nested model
- Models learned in parallel

Functional recovery with increasing numbers of states



Zinc Finger state | Simple Repeat state | Transcription End State

- Red: Maximum fold functional enrichment for corresponding biological category
- Blue: Percent of that functional category that overlaps regions annotated to this state
- Top plot: Correlation of emission parameter vector for that state to closest state

Chromatin state recovery with increasing numbers of marks

Which states are well-recovered?

Increasing numbers of marks (greedy)

State Inferred with all 41 marks

Recovery of states with increasing number of marks

Precisely what mistakes are made?

(for a given subset of 11 ENCODE marks)

State Inferred with subset of marks

State Inferred with all 41 marks

State confusion matrix with 11 ENCODE marks

Source: Ernst, Jason and Manolis Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature Biotechnology 28, no. 8 (2010): 817-825.

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

(Future: Chromatin states to interpret disease-associated variants)

ENCODE: Study nine marks in nine human cell lines

9 marks

H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
H3K27me3
H4K20me1
H3K36me3
CTCF
+WCE
+RNA

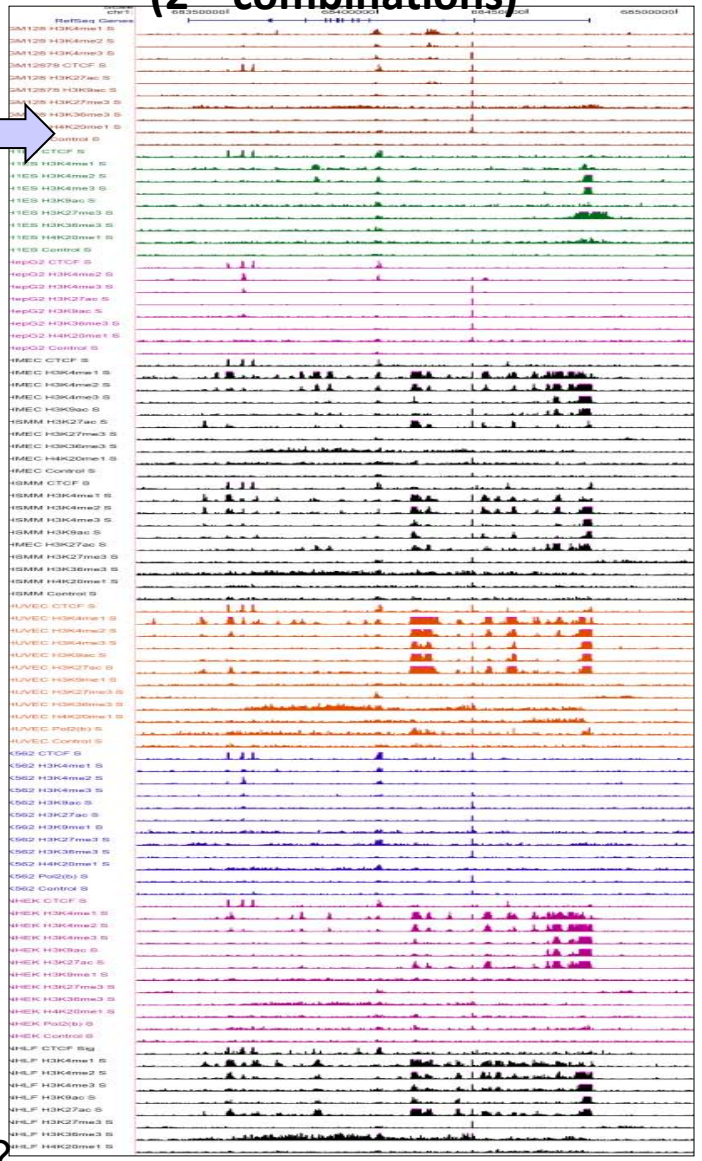
X

9 human cell types

HUVEC	Umbilical vein endothelial
NHEK	Keratinocytes
GM12878	Lymphoblastoid
K562	Myelogenous leukemia
HepG2	Liver carcinoma
NHLF	Normal human lung fibroblast
HMEC	Mammary epithelial cell
HSMM	Skeletal muscle myoblasts
H1	Embryonic

Brad Bernstein ENCODE Chromatin Group

81 Chromatin Mark Tracks (2^{81} combinations)



b.

State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE
1	16	2	2	6	17	93	99	96	98	2
2	12	2	6	9	53	94	95	14	44	1
3	13	72	0	9	48	78	49	1	10	1
4	11	1	15	11	96	99	75	97	86	4
5	5	0	10	3	88	57	5	84	25	1
6	7	1	1	3	58	75	8	6	5	1
7	2	1	2	1	56	3	0	6	2	1
8	92	2	1	3	6	3	0	0	1	1
9	5	0	43	43	37	11	2	9	4	1
10	1	0	47	3	0	0	0	0	0	1
11	0	0	3	2	0	0	0	0	0	0
12	1	27	0	2	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	22	28	19	41	6	5	26	5	13	37
15	85	85	91	88	76	77	91	73	85	78

Chromatin States

Chromatin Mark Observation Frequency (%)

How to learn single set of chromatin states?

Ernst et al, *Nature* 2011

Solution 1: Learn independent models and cluster

state	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE
1	13.2	72.0	0.2	9.1	47.9	77.8	49.5	1.3	10.2	0.7
2	11.9	1.9	6.1	9.0	52.7	93.7	95.0	14.1	44.1	0.9
3	16.4	1.5	2.4	5.5	17.0	92.6	99.0	95.7	98.1	1.9
4	11.4	0.6	14.5	11.3	96.3	99.3	75.1	97.2	85.7	3.7
5	5.3	0.2	9.5	2.6	88.1	56.8	5.3	84.4	24.9	1.5
6	6.7	0.9	1.0	3.2	58.3	74.7	8.4	5.8	5.4	0.8
7	1.6	0.6	1.6	1.3	56.5	2.7	0.4	5.9	1.6	0.6
8	91.5	1.8	0.9	2.8	6.3	3.3	0.4	0.5	1.0	0.8
9	4.6	0.3	43.2	43.1	36.5	11.5	1.9	9.1	3.9	1.3
10	1.2	0.1	47.2	2.7	0.4	0.0	0.1	0.3	0.3	0.5
11	0.4	0.1	2.7	1.7	0.2	0.1	0.1	0.2	0.3	0.4
12	0.9	26.8	0.0	2.1	0.4	0.1	0.1	0.1	0.1	0.4
13	0.2	0.4	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.1
14	21.9	27.9	19.1	41.0	5.7	4.8	25.9	5.3	13.1	37.5
15	85.2	85.1	91.5	88.4	76.2	76.9	90.8	72.7	85.0	78.3

Promoter

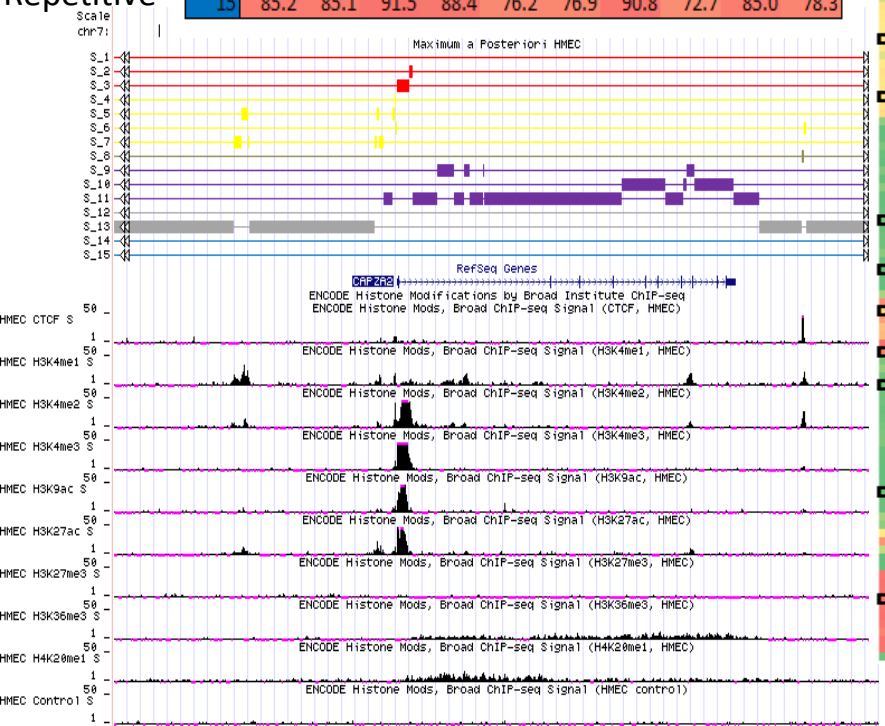
Candidate enhancer

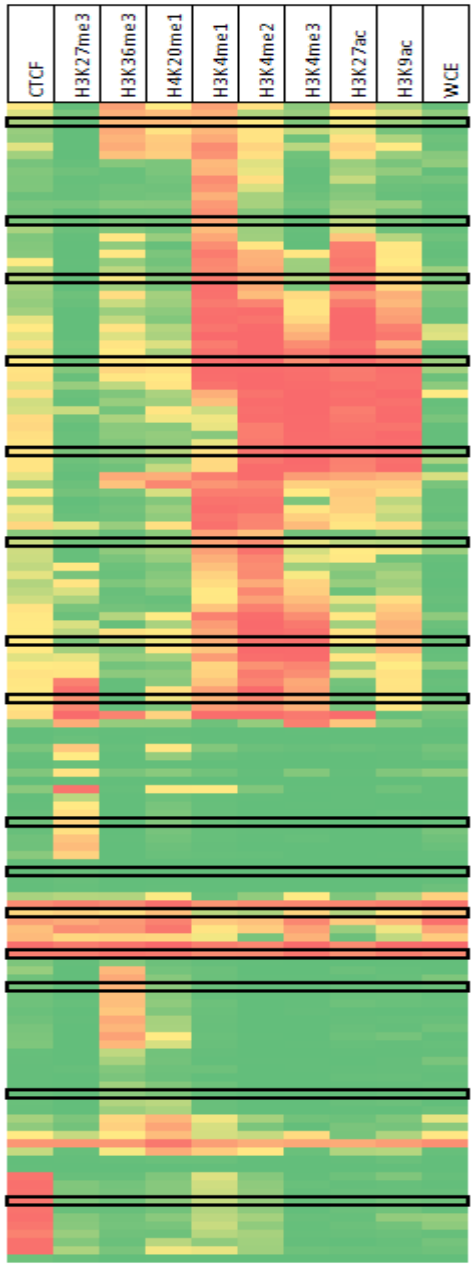
Insulator

Transcribed

Repressive

Repetitive





Basic approach:

a) Train a k-state model in each cell type independently

b) Cluster models learned independently

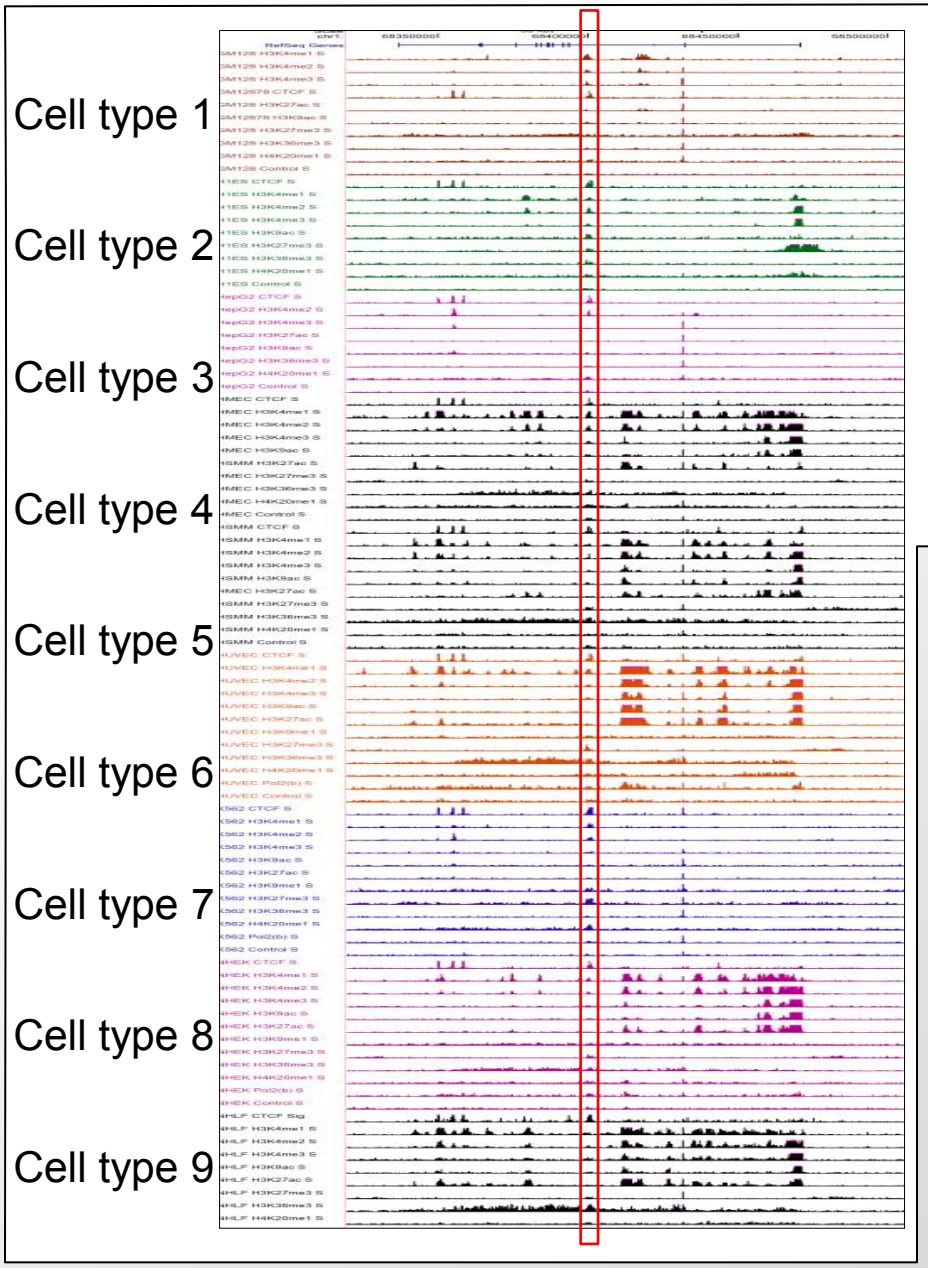
c) Merge clusters and re-apply to each cell type

How to cluster

a) Using emission probability matrix: most similar definitions

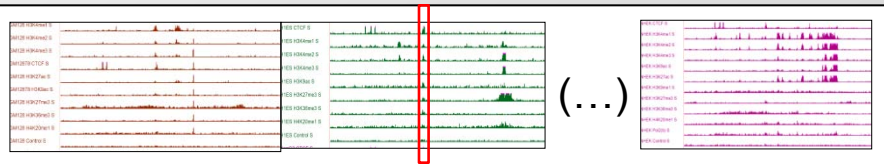
b) Using genome annotation: posterior probability decoding

Joint learning of states across multiple cell types



Solution 2: Stacking

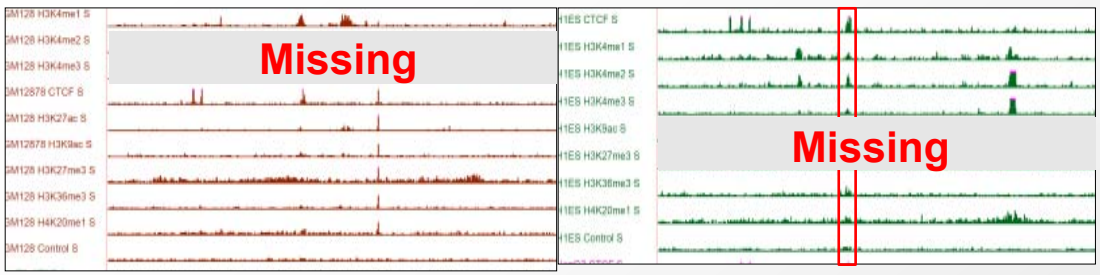
- Learns each combination of activity as a separate state
- Ex: ES-specific enhancers: enhancer marks in ES, no marks in other cell types



Solution 3: Concatenation

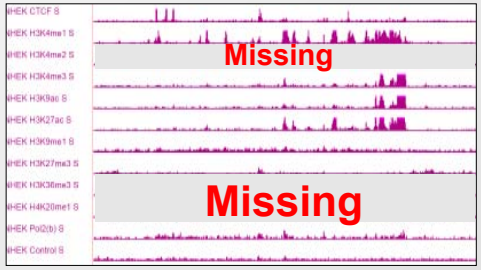
- Requires that profiled marks are the same (or treat as missing data)
- Ensures common state definitions across cell types

Joint learning with different subsets of marks (Solution 3)



Cell type 1 Cell type 2

(…)



Cell type 9

Option (a) Treat missing tracks as missing data

- EM framework allows for unspecified data points
- As long as pairwise relationship observed in some cell type

Option (b) Chromatin mark imputation

- Explicitly predict max-likelihood chromatin track for missing data
- Less powerful if ultimate goal is chromatin state learning

ENCODE: Study nine marks in nine human cell lines

9 marks

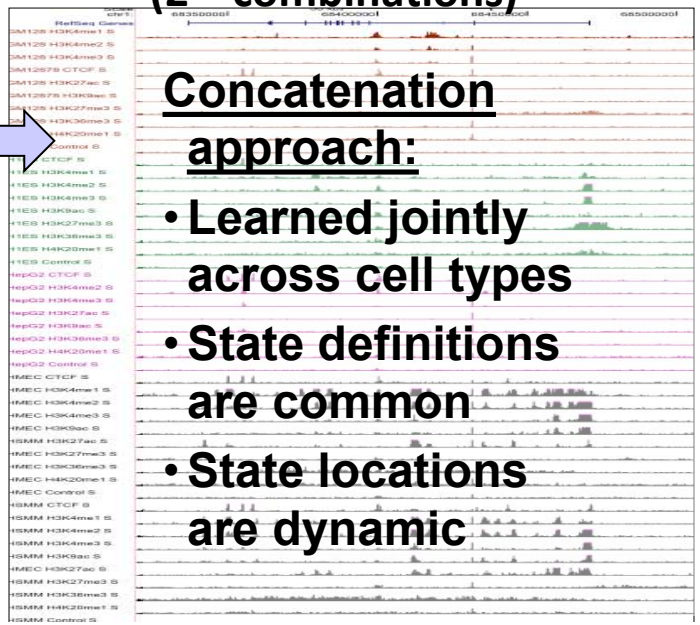
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
H3K27me3
H4K20me1
H3K36me3
CTCF
+WCE
+RNA

X

9 human cell types

HUVEC	Umbilical vein endothelial
NHEK	Keratinocytes
GM12878	Lymphoblastoid
K562	Myelogenous leukemia
HepG2	Liver carcinoma
NHLF	Normal human lung fibroblast
HMEC	Mammary epithelial cell
HSMM	Skeletal muscle myoblasts
H1	Embryonic

81 Chromatin Mark Tracks (2^{81} combinations)



Concatenation approach:

- Learned jointly across cell types
- State definitions are common
- State locations are dynamic

b.

Brad Bernstein ENCODE Chromatin Group

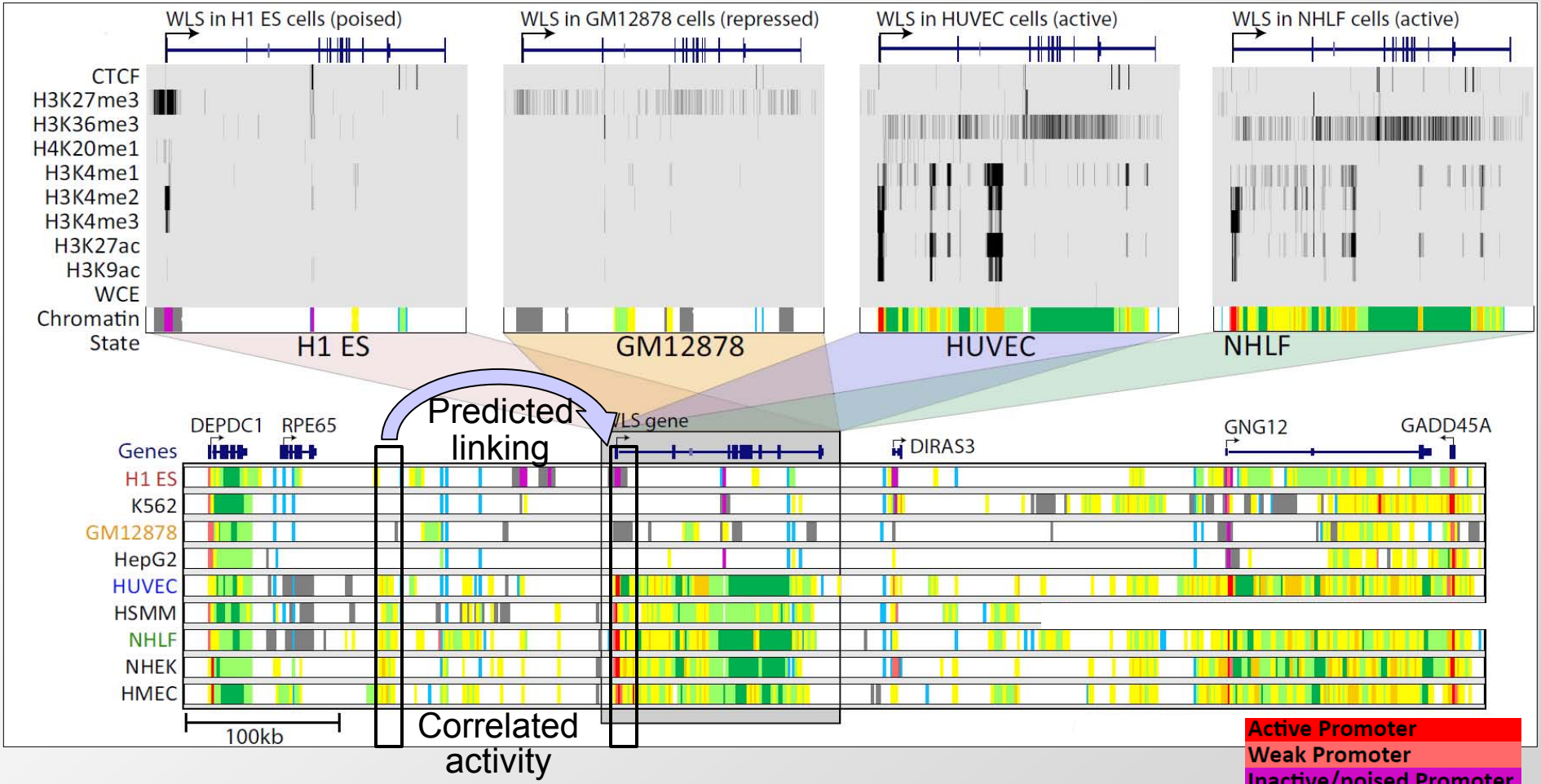
State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE
1	16	2	2	6	17	93	99	96	98	2
2	12	2	6	9	53	94	95	14	44	1
3	13	72	0	9	48	78	49	1	10	1
4	11	1	15	11	96	99	75	97	86	4
5	5	0	10	3	88	57	5	84	25	1
6	7	1	1	3	58	75	8	6	5	1
7	2	1	2	1	56	3	0	6	2	1
8	92	2	1	3	6	3	0	0	1	1
9	5	0	43	43	37	11	2	9	4	1
10	1	0	47	3	0	0	0	0	0	1
11	0	0	3	2	0	0	0	0	0	0
12	1	27	0	2	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0
14	22	28	19	41	6	5	26	5	13	37
15	85	85	91	88	76	77	91	73	85	78

Chromatin States

Chromatin Mark Observation Frequency (%)

Coverage: Median	Coverage: H1 ES	Coverage: GM	Median Length	+/-2kb TSS	Conserved non-exon	DNase (K562)	C-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear Lamina (NHLF)	Candidate state annotation
0.6	0.5	1.2	1.0	83	3.8	23.3	82.0	40.7	0.2	0.15	Active Promoter
0.5	1.2	1.3	0.4	58	2.8	15.3	12.6	5.8	0.6	0.30	Weak Promoter
0.2	4.0	1.0	0.6	49	4.3	10.8	3.1	1.0	0.4	0.68	Inactive/poised Promoter
0.7	0.1	1.1	0.6	23	2.7	23.1	31.8	49.0	1.3	0.05	Strong enhancer
1.2	0.2	0.7	0.6	3	1.8	13.6	6.3	15.8	1.4	0.10	Strong enhancer
0.9	1.3	1.0	0.2	17	2.4	11.9	5.7	7.0	1.1	0.31	Weak/poised enhancer
1.9	1.2	1.1	0.4	4	1.5	5.1	0.6	2.4	1.3	0.20	Weak/poised enhancer
0.5	1.4	1.0	0.4	3	1.5	12.8	2.5	1.2	1.1	0.61	Insulator
0.7	1.3	1.0	0.8	4	1.1	4.5	0.7	0.8	2.4	0.02	Transcriptional transition
4.3	0.6	1.2	3.0	1	0.9	0.3	0.0	0.0	2.5	0.11	Transcriptional elongation
12.5	1.3	0.8	2.6	2	0.9	0.3	0.0	0.1	1.9	0.24	Weak transcribed
4.1	0.3	0.7	2.8	5	1.4	0.3	0.0	0.1	0.8	0.63	Polycomb-repressed
71.4	1.0	1.0	10.0	1	0.9	0.1	0.0	0.0	0.7	1.30	Heterochrom; low signal
0.1	0.9	1.2	0.6	3	0.4	1.9	0.3	0.2	0.4	1.44	Repetitive/CNV
0.1	0.9	1.0	0.2	1	0.2	5.9	9.5	7.4	0.4	1.30	Repetitive/CNV
(%)	(fold)		(kb)	(%)	Functional enrichments (fold)						

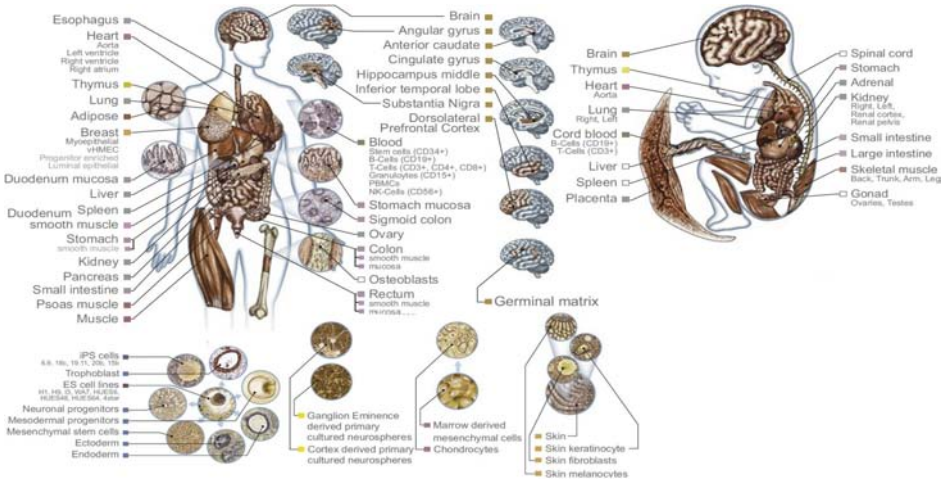
Chromatin states dynamics across nine cell types



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

Epigenomic mapping across 100+ tissues/cell types

Diverse tissues and cells



Courtesy of NIH Roadmap Epigenomics Mapping Consortium. Used with permission.

Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood…)

Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood…)

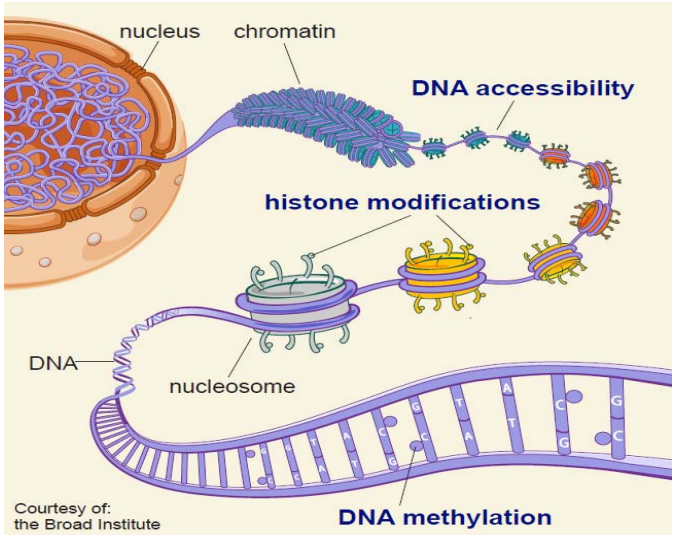
ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench…)

ROADMAP epigenomics PROJECT

ENCODE

X

Diverse epigenomic assays



Courtesy of Broad Communications. Used with permission.

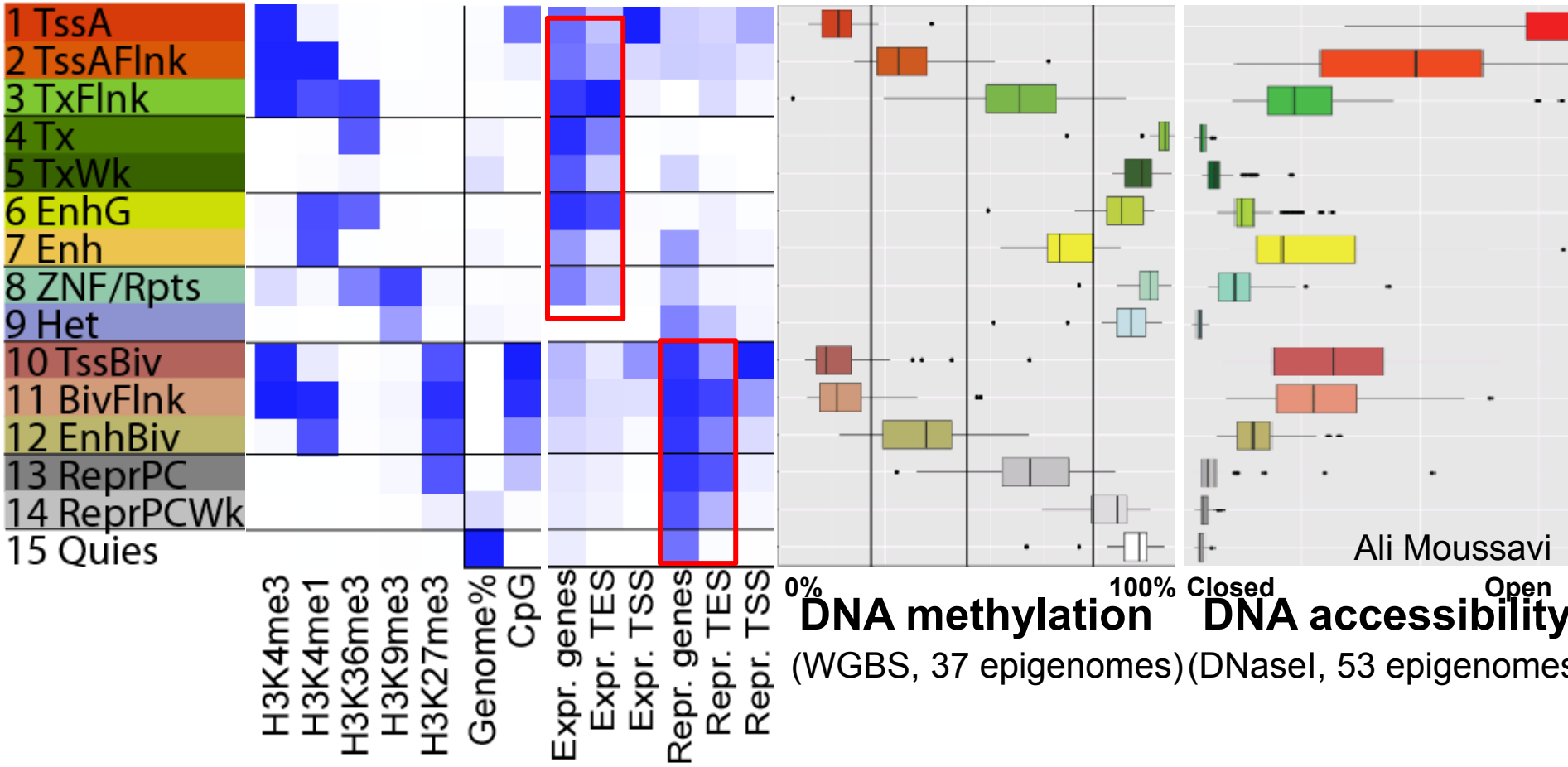
Histone modifications
- **H3K4me3, H3K4me1, H3K36me3**
- **H3K27me3, H3K9me3, H3K27/9ac**
- **+20 more**

Open chromatin:
- **DNA accessibility**

DNA methylation:
- **WGBS, RRBS, MRE/MeDIP**

Gene expression
- **RNA-seq, Exon Arrays**

States show distinct mCpG, DNase, Tx, Ac profiles



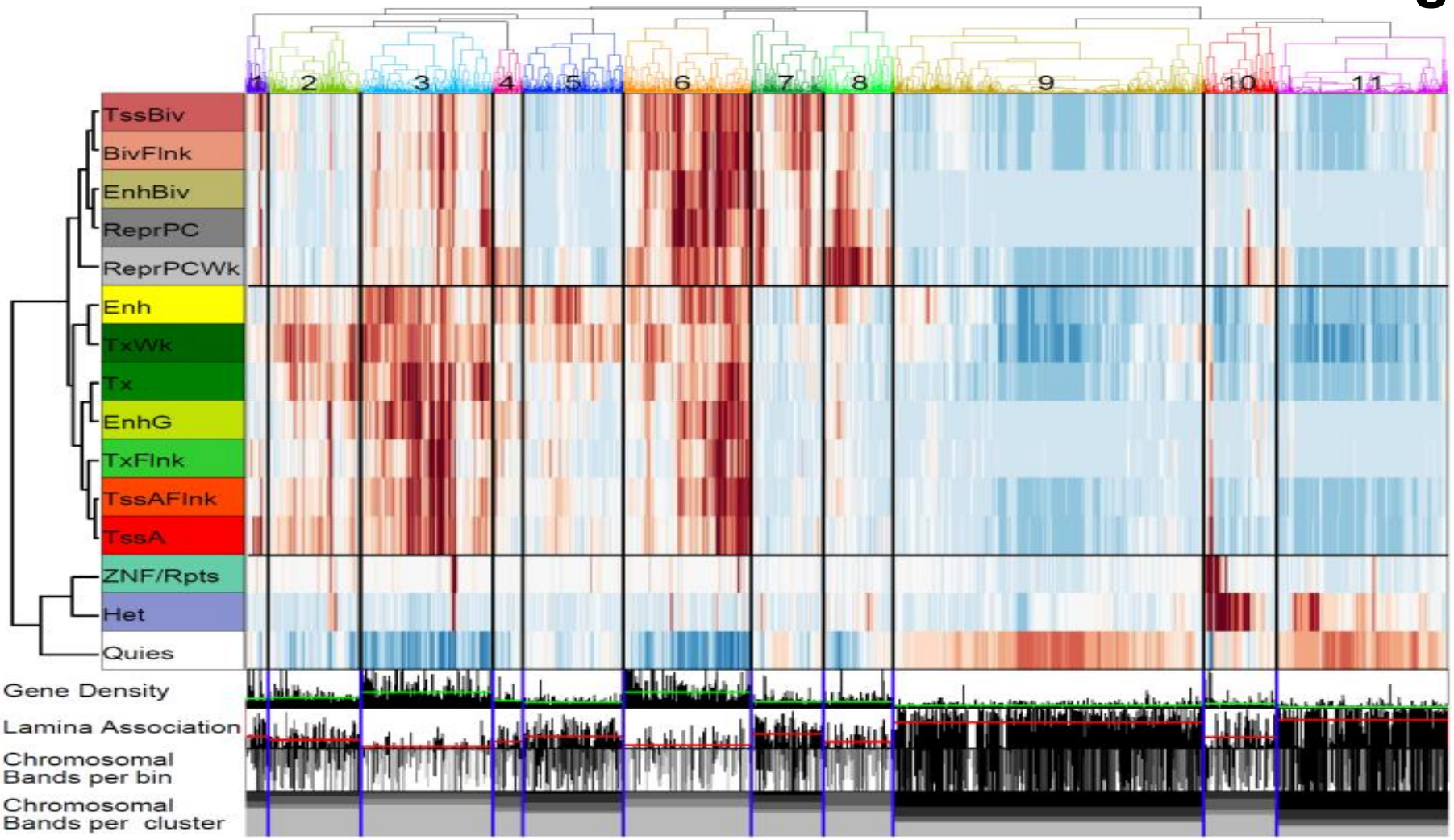
TssA vs. **TssBiv**: diff. activity, both open, both unmethylated!

Enh vs. **ReprPC**: diff. activity, both intermediate DNase/Methyl

Tx: Methylated, closed, actively transcribed

➔ Distinct modes of repression: **H3K27me3** vs. **DNAme** vs. **Het**

Chromosomal ‘domains’ from chromatin state usage

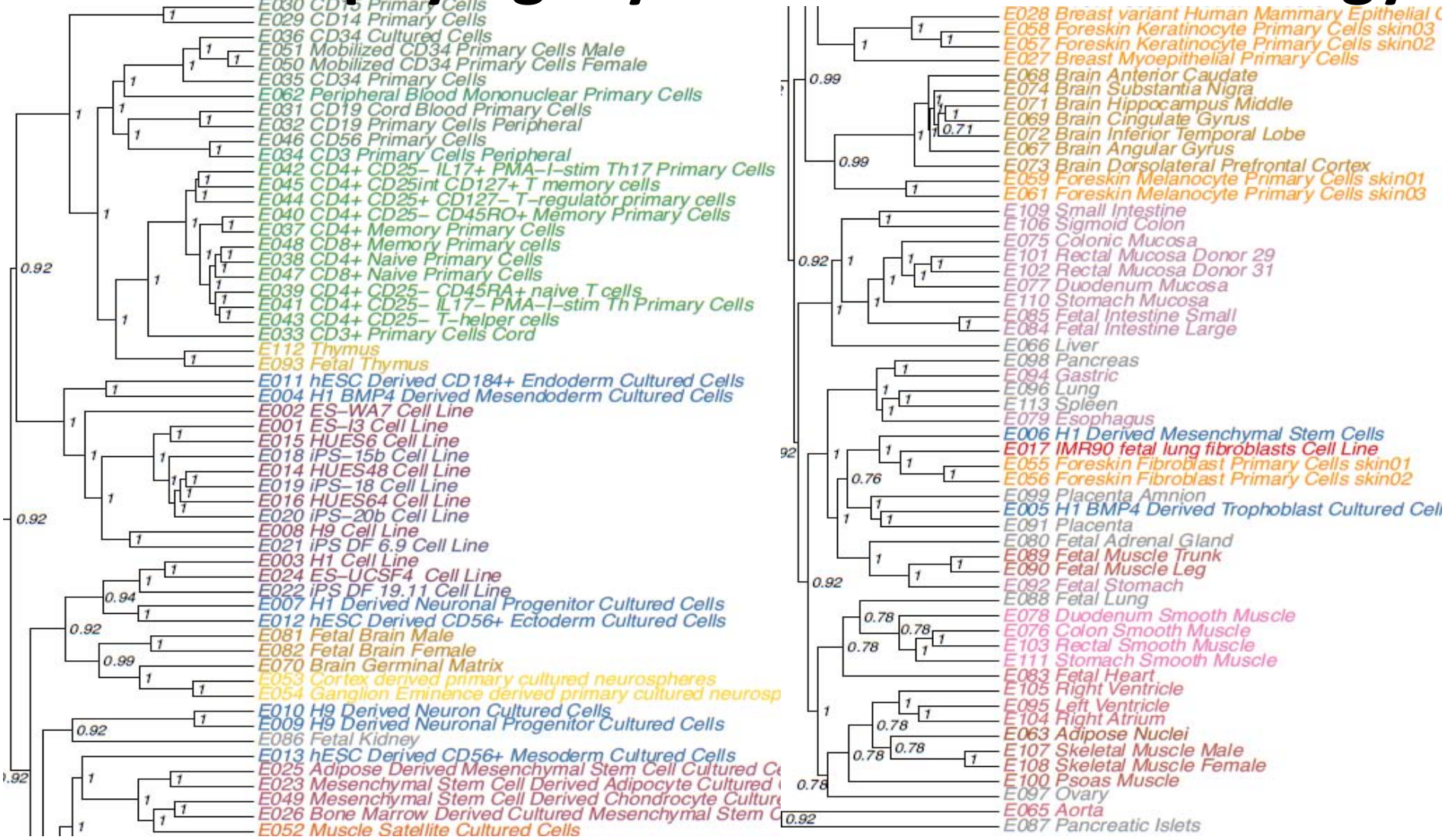


Source: Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes." Nature 518, no. 7539 (2015): 317-330.

Misha Bilenky

- **State usage ➔ gene density, lamina, cytogenetic bands**
- **Quies/ZNF/het | gene rich/poor, each active/repressed**

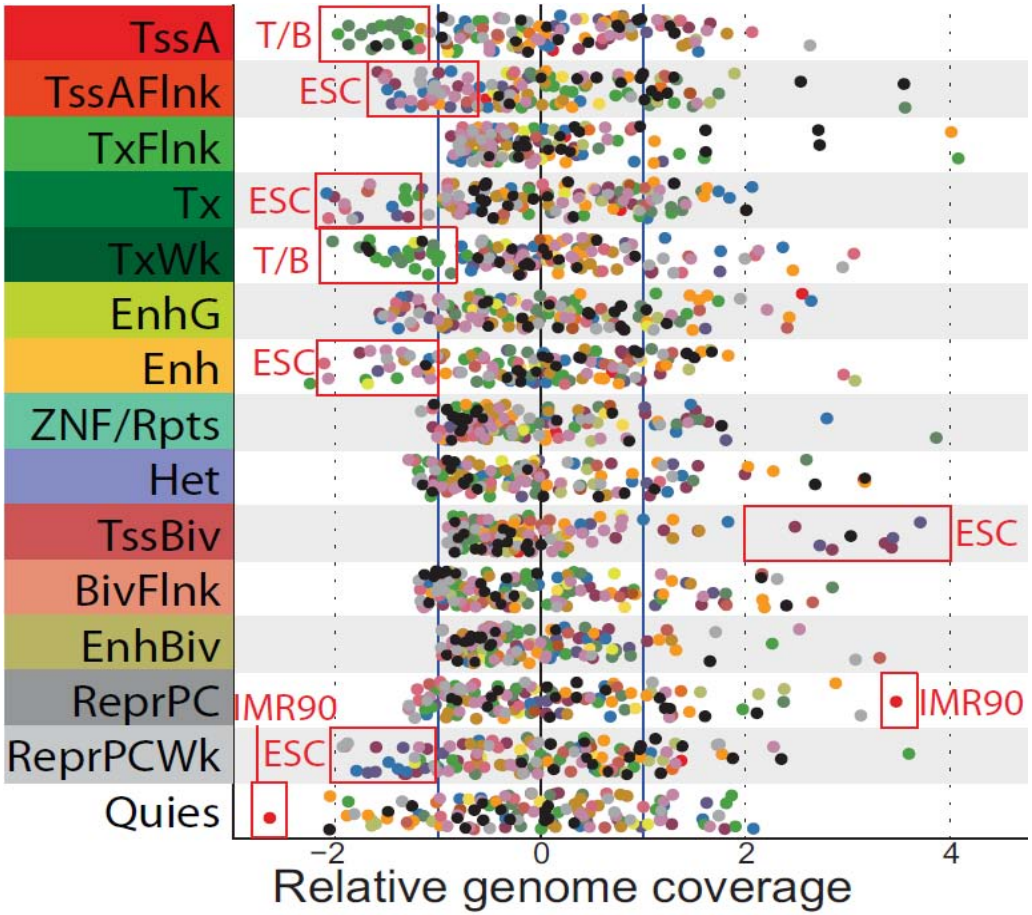
H3K4me1 phylogeny reveals common biology

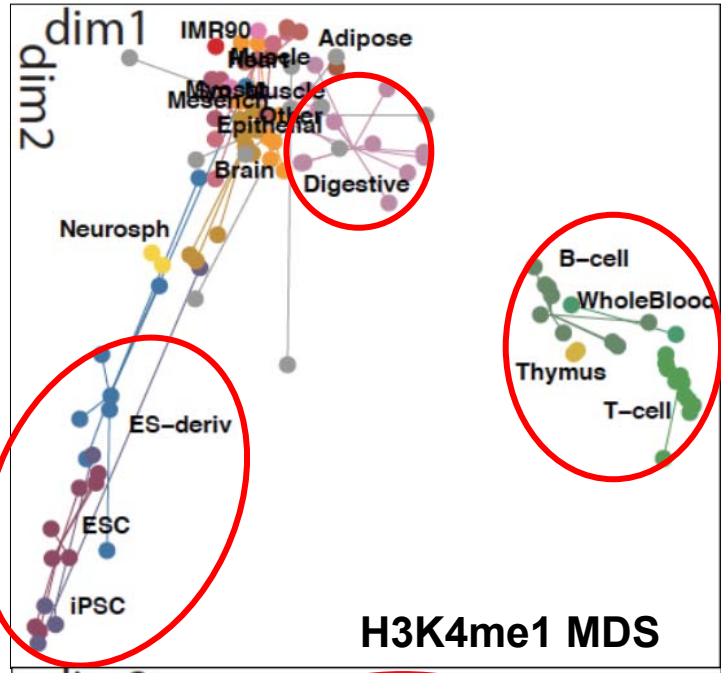


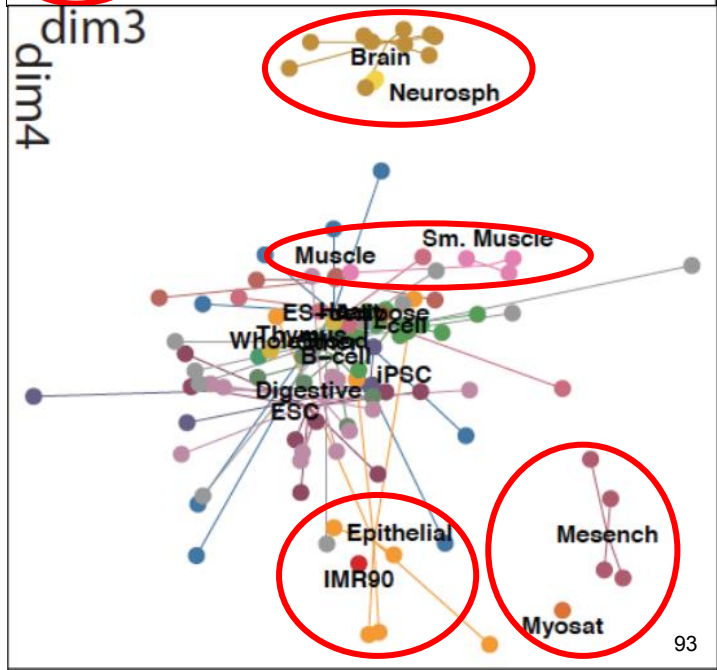
Source: Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes." Nature 518, no. 7539 (2015): 317-330.

- Grouping of ES, immune, brain, muscle, heart, smooth muscle, fetal

Cells/Tissues at extremes of epigenomic variation







Source: Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes." Nature 518, no. 7539 (2015): 317-330.

- **ES/Immune/IMR90 most extreme**
- **ES: ↑Biv, ↓Enh/Tx/TssFlnk/PCwk**
- **Immune: ↓TssA, ↓TxWk**
- **IMR90: ↑ReprPC, ↓Quies**

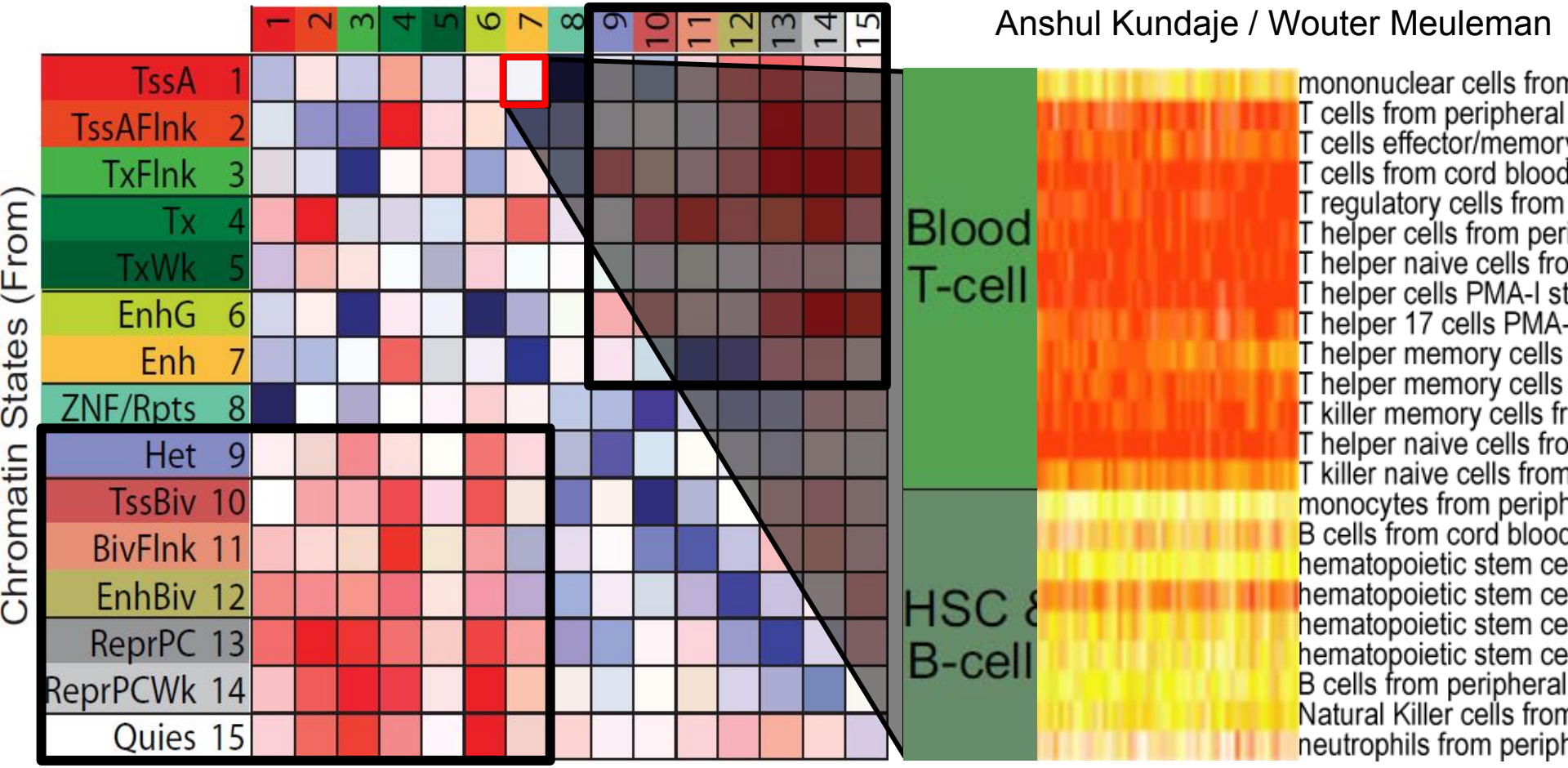
Misha Bilenky, Wouter Meuleman

Chromatin state annotations across 127 epigenomes

Source: Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes." Nature 518, no. 7539 (2015): 317-330.

Reveal epigenomic variability: enh/prom/tx/repr/het

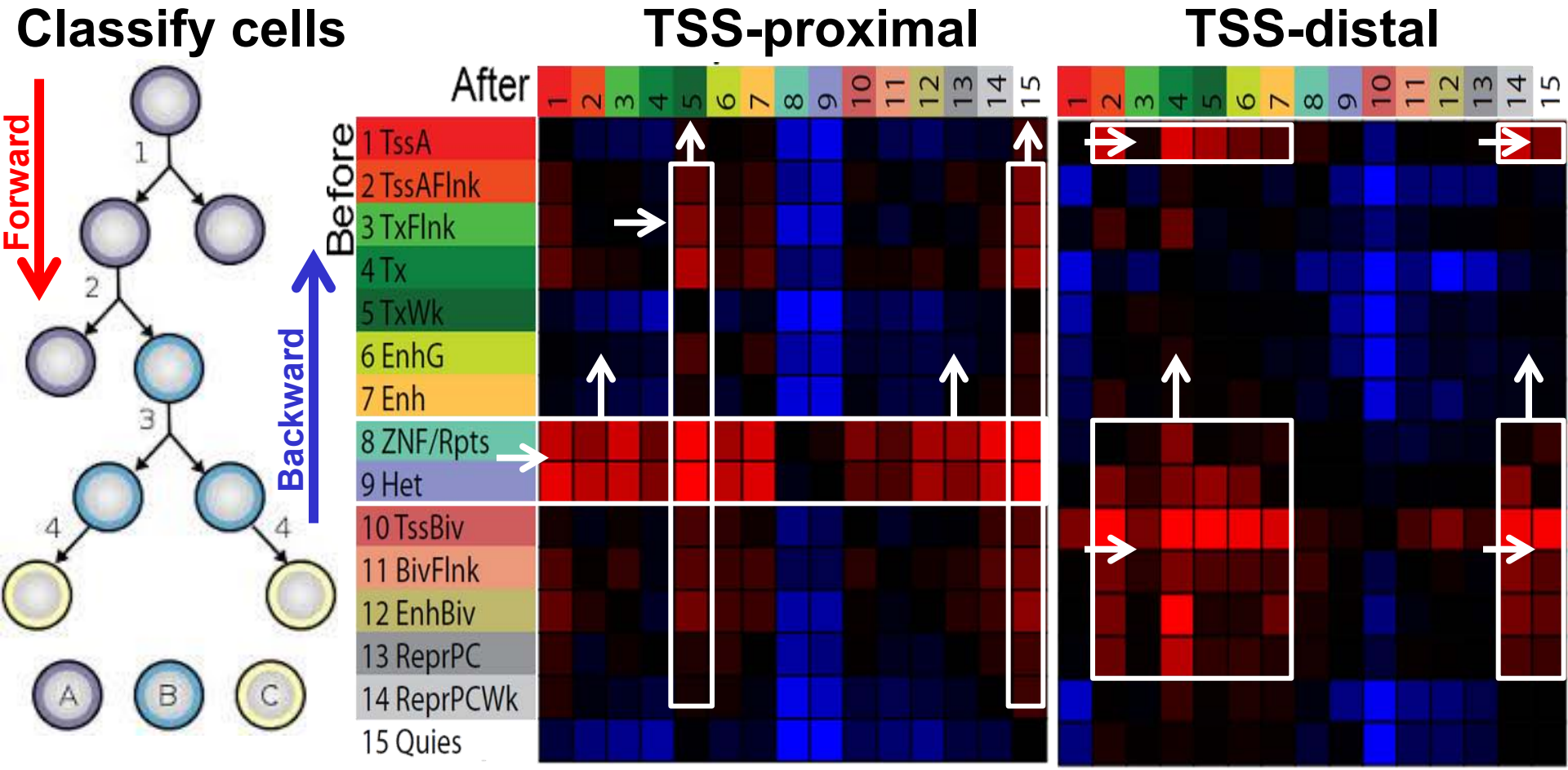
State switching: active/inactive, mostly keep identity



Source: Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes." Nature 518, no. 7539 (2015): 317-330.

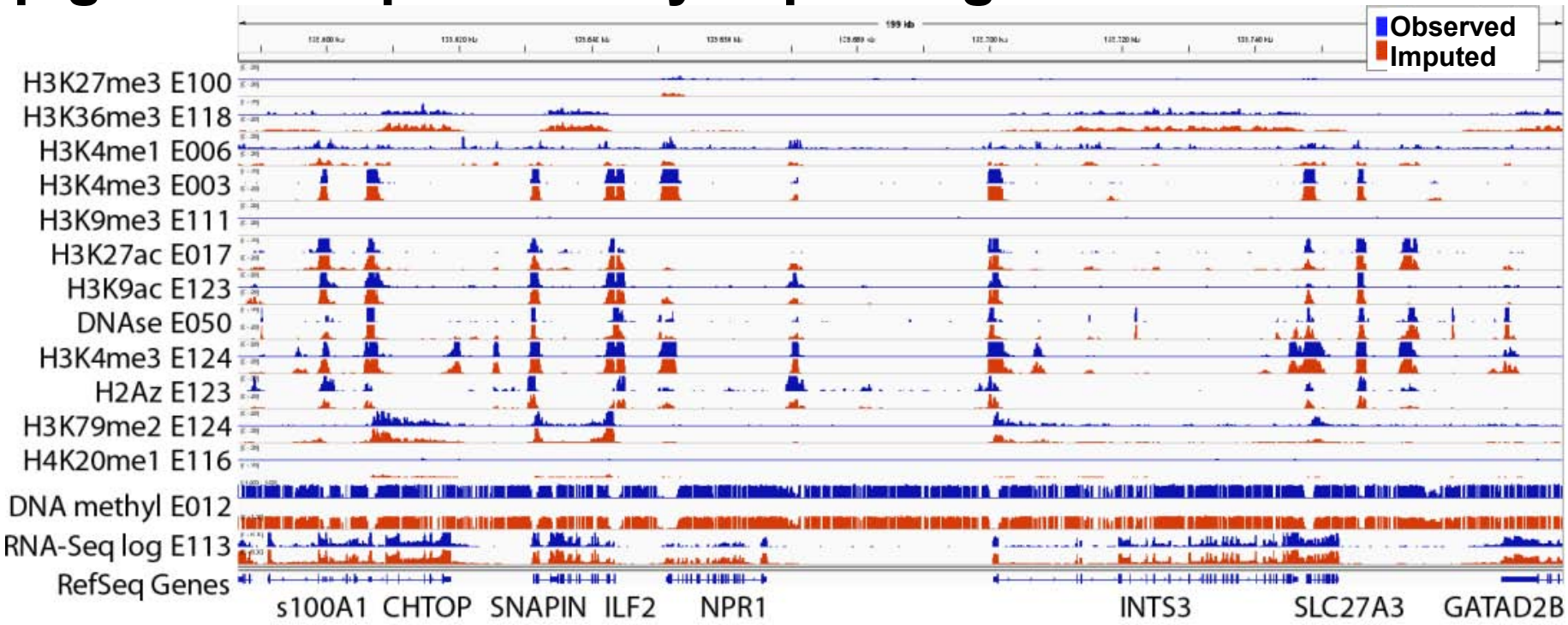
- **Most variable: Enhancers. Least: TssA/Tx/Quies**
- **State switching: Active (1-7)⇔Inactive (10-15)**
- **Exception: Dyadic regions: enhancer⇔promoter**

Chromatin state changes during differentiation

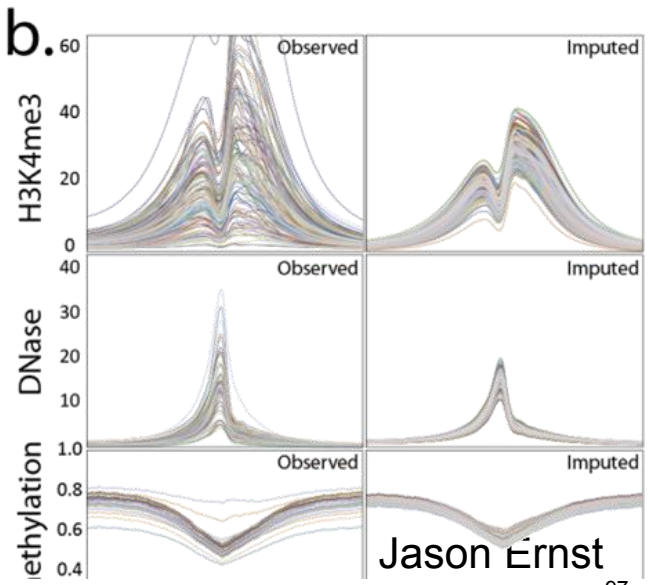


- **Epigenomic features can predict directionality:** AUC 78%
- **TSS-proximal**: (1) Loss of Het/ZNF. (2) Gain of TxWk, Quies
- **TSS-distal**: Bivalent, PCrepressed ➔ Enhancer, Tx, TssFlnk

Epigenome imputation by exploiting mark correlations



- **Two types of features**
  - Other marks + context in same tissue
  - Same mark in ‘closest’ tissues
- **Impute missing datasets**
  - Predict DNase, marks @ 25bp res
  - Predict RNA-Seq @ 25 bp res
  - Predict DNA methylation @ 1bp res



Goals for today: Computational Epigenomics

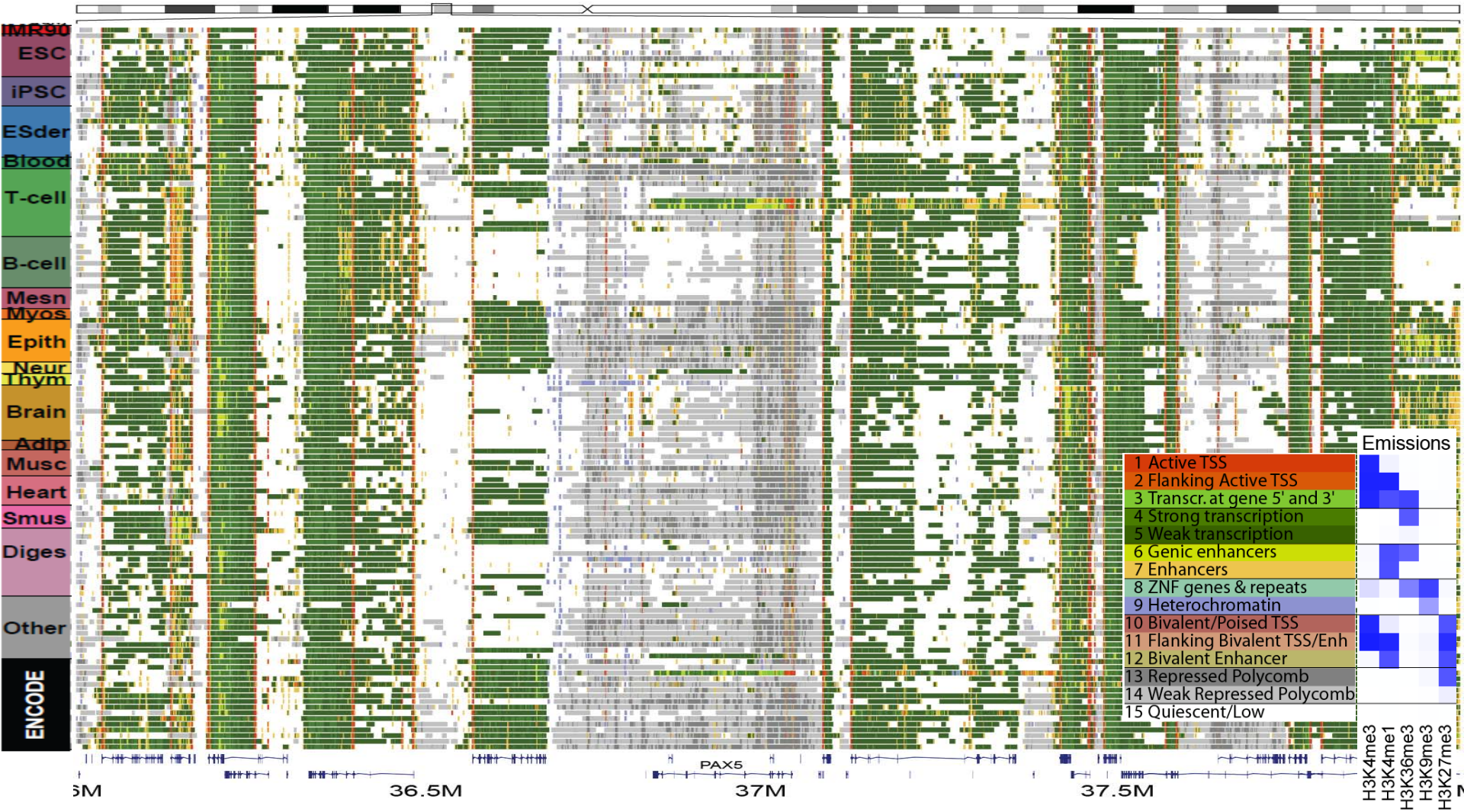
1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

(Future: Chromatin states to interpret disease-associated variants)

5. Correlation-based links of enhancer networks

Regulators → Enhancers → Target genes

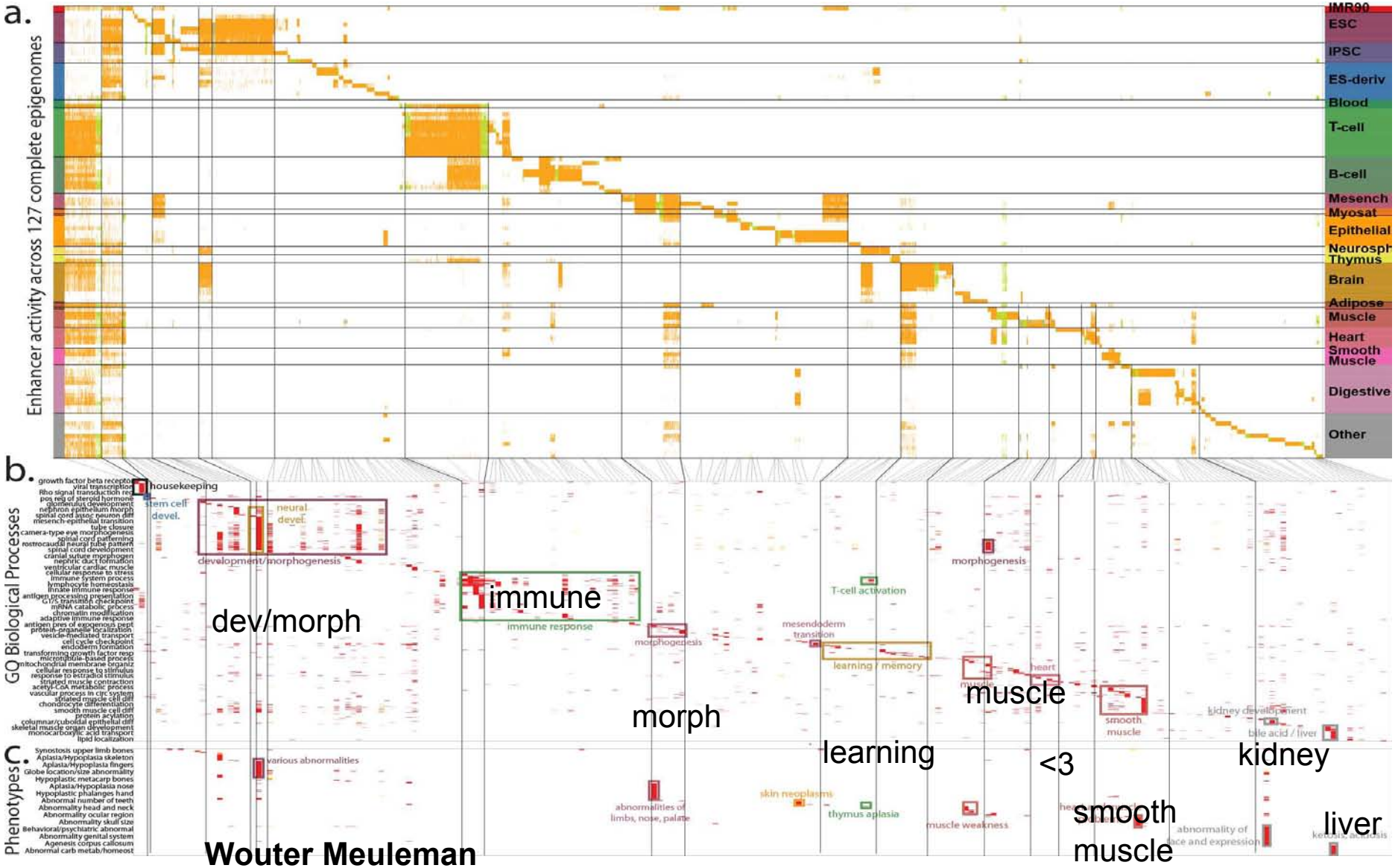
Chromatin state annotations across 127 epigenomes



Source: Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes." Nature 518, no. 7539 (2015): 317-330.

Reveal epigenomic variability: enh/prom/tx/repr/het

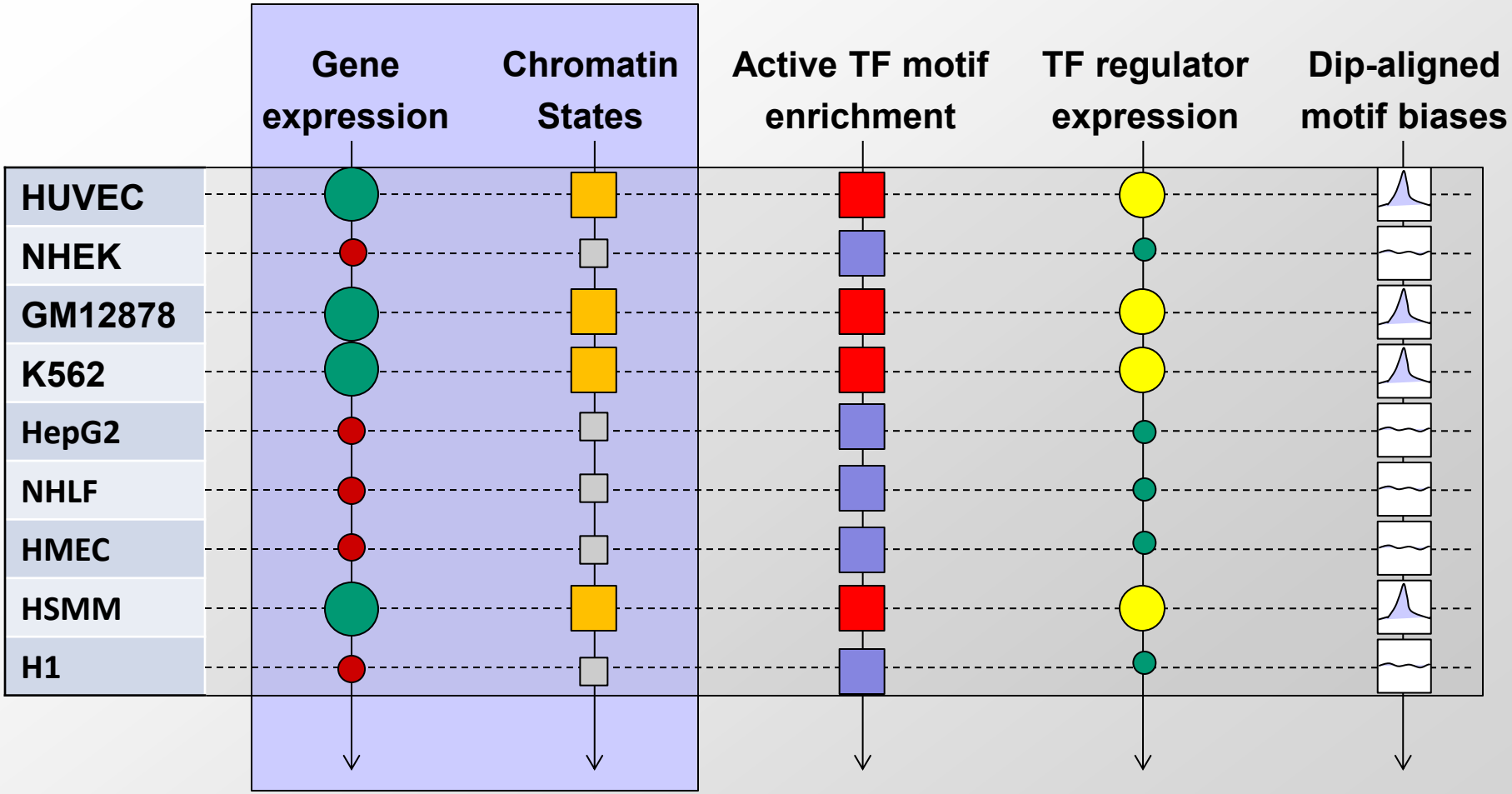
2.3M enhancer regions ⇔ only ~200 activity patterns



Wouter Meuleman

Source: Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes." Nature 518, no. 7539 (2015): 317-330.

Introducing multi-cell activity profiles

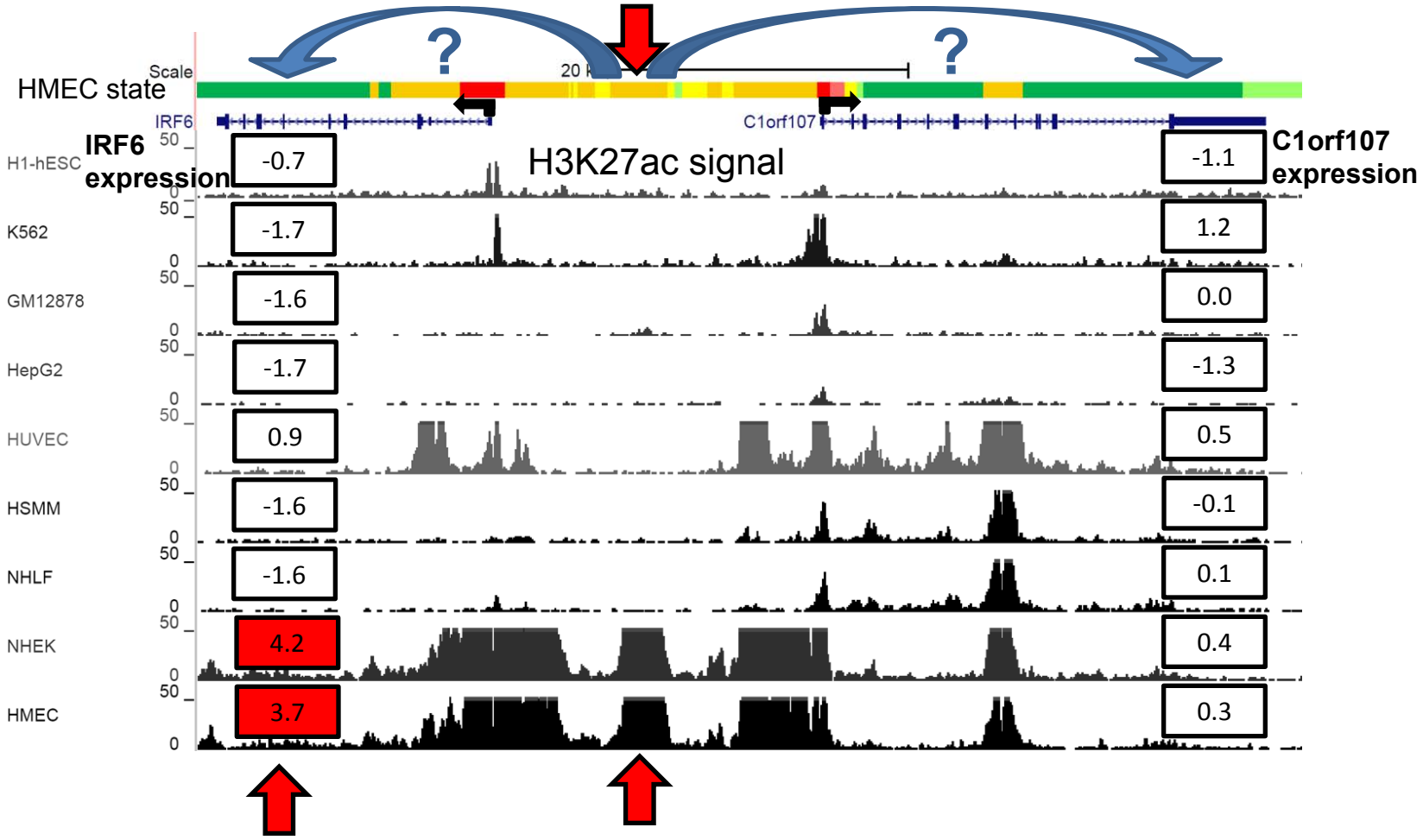


Link enhancers to target genes

ON
OFF
Active enhancer
Repressed
Motif enrichment
Motif depletion
TF On
TF Off
Motif aligned
Flat profile

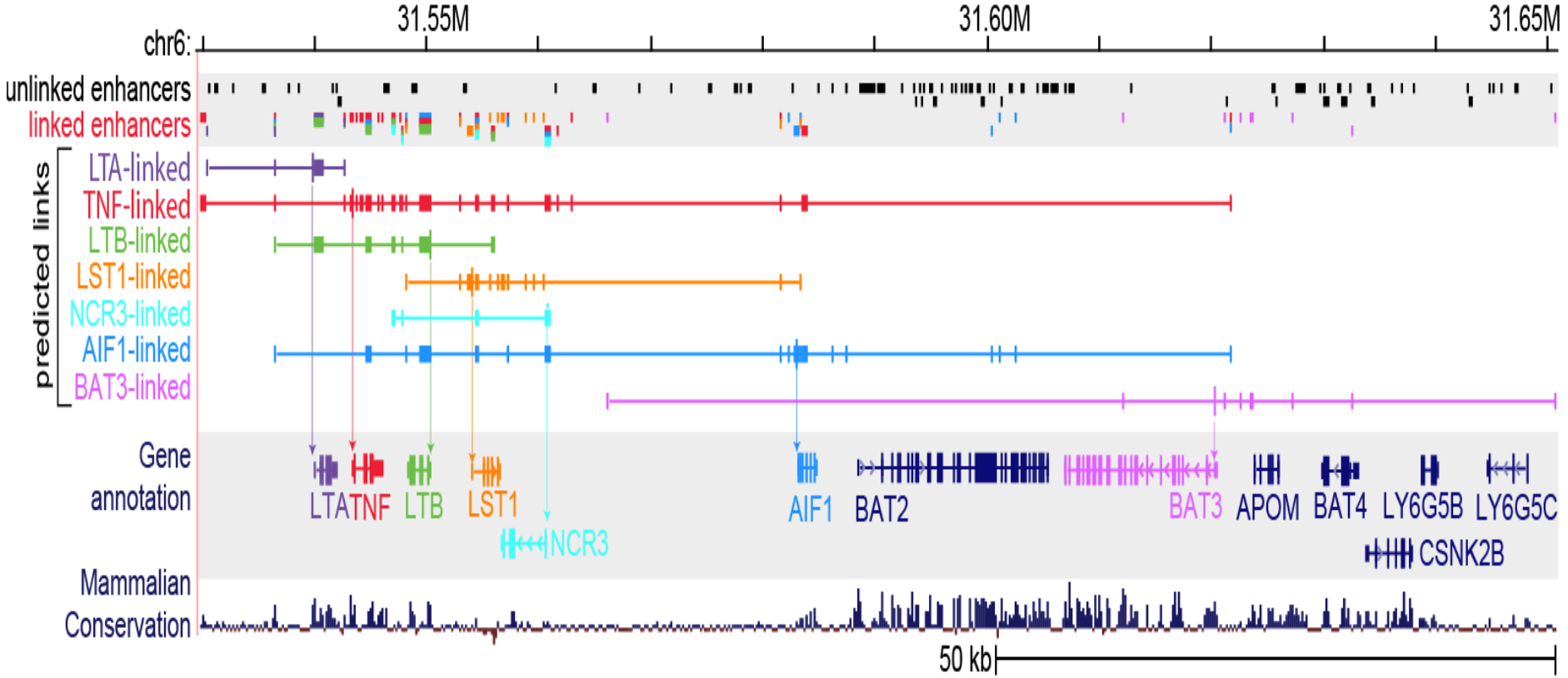
Activity-based linking of enhancers to target genes

Finding correct target of enhancer in divergently transcribed genes



Compute correlations between gene expression levels and enhancer associated histone modification signals

Visualizing 10,000s predicted enhancer-gene links

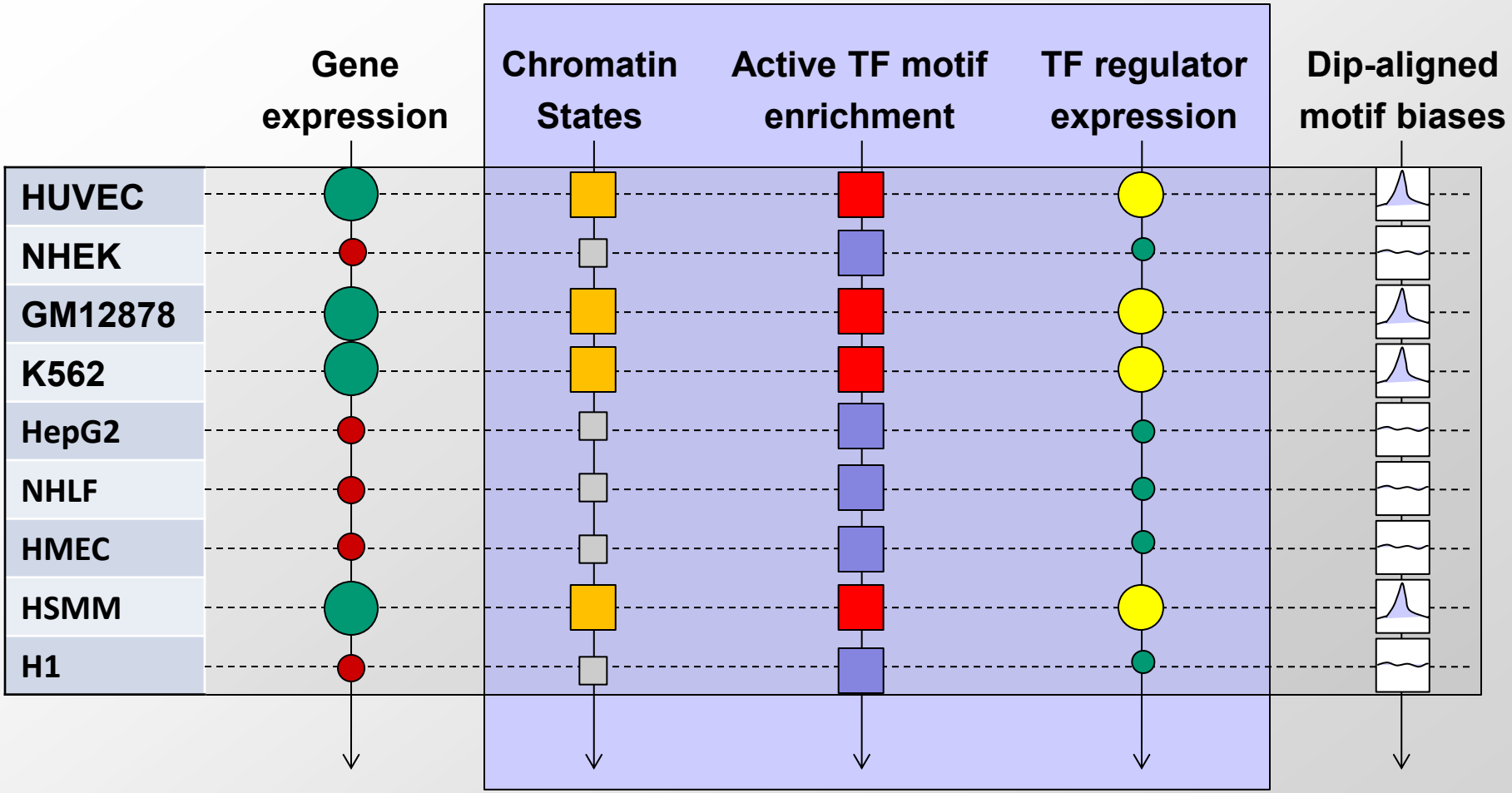


- Overlapping regulatory units, both few and many
- Both upstream and downstream elements linked
- Enhancers correlate with sequence constraint

Chromatin dynamics: linking enhancer networks

TFs → enhancers → target genes

Introducing multi-cell activity profiles

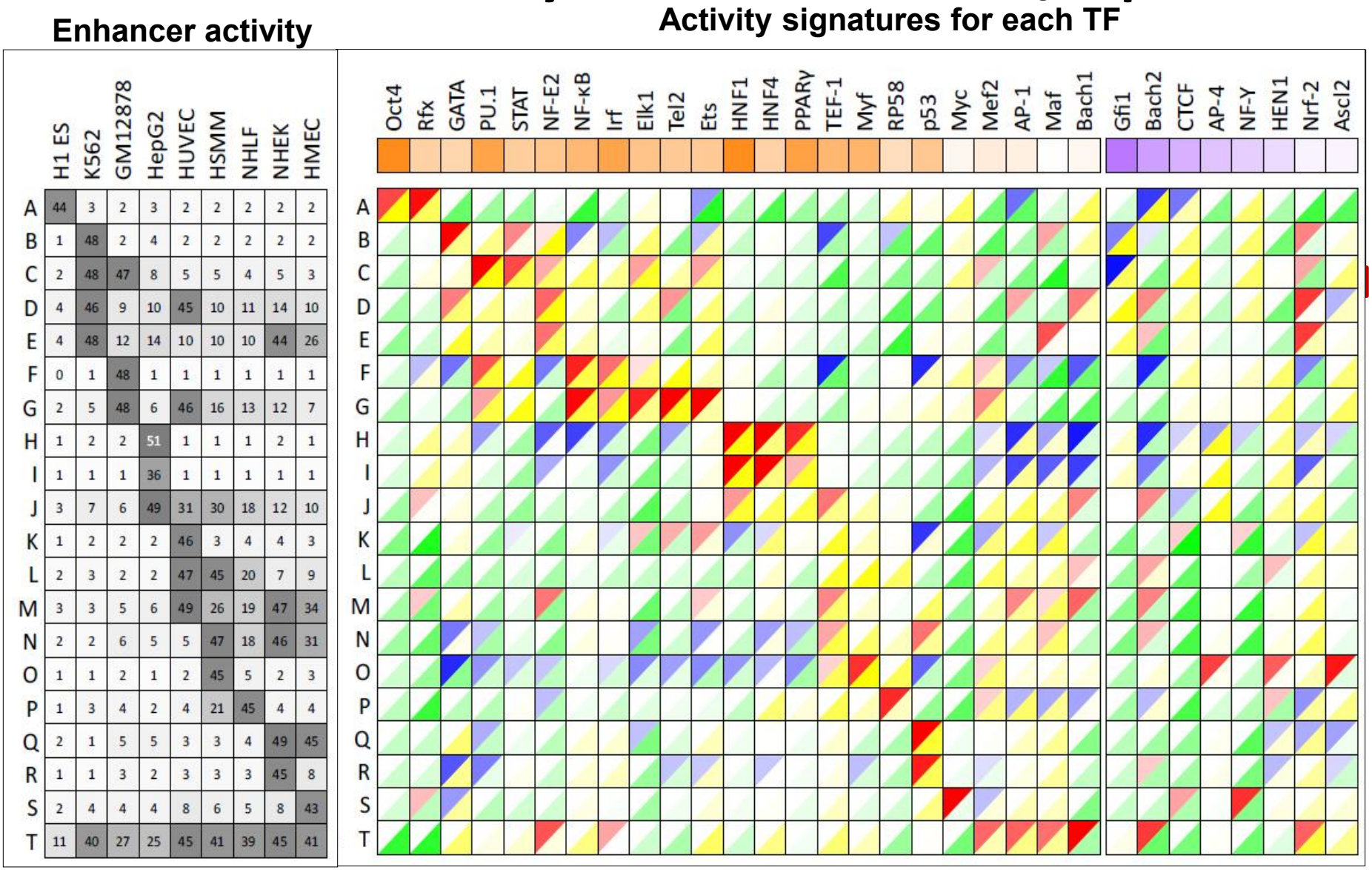


Link TFs to target enhancers

Predict activators vs. repressors

ON
OFF
Active enhancer
Repressed
Motif enrichment
Motif depletion
TF On
TF Off
Motif aligned
Flat profile

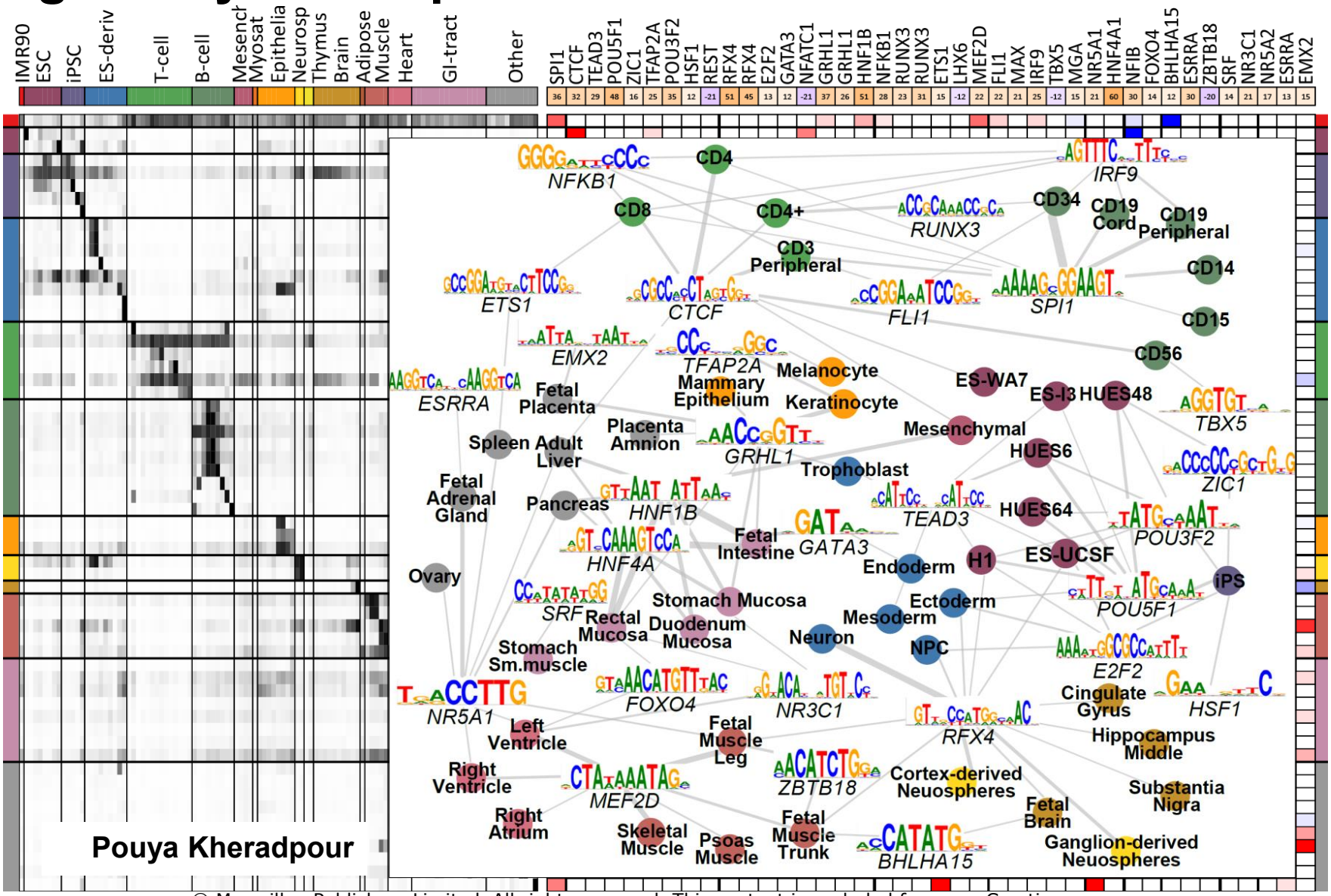
Coordinated activity reveals activators/repressors



Source: Ernst, Jason et al. "Mapping and analysis of chromatin state dynamics in nine human cell types." Nature 473, no.7345 (2011): 43-49.

- Enhancer networks: Regulator → enhancer → target gene

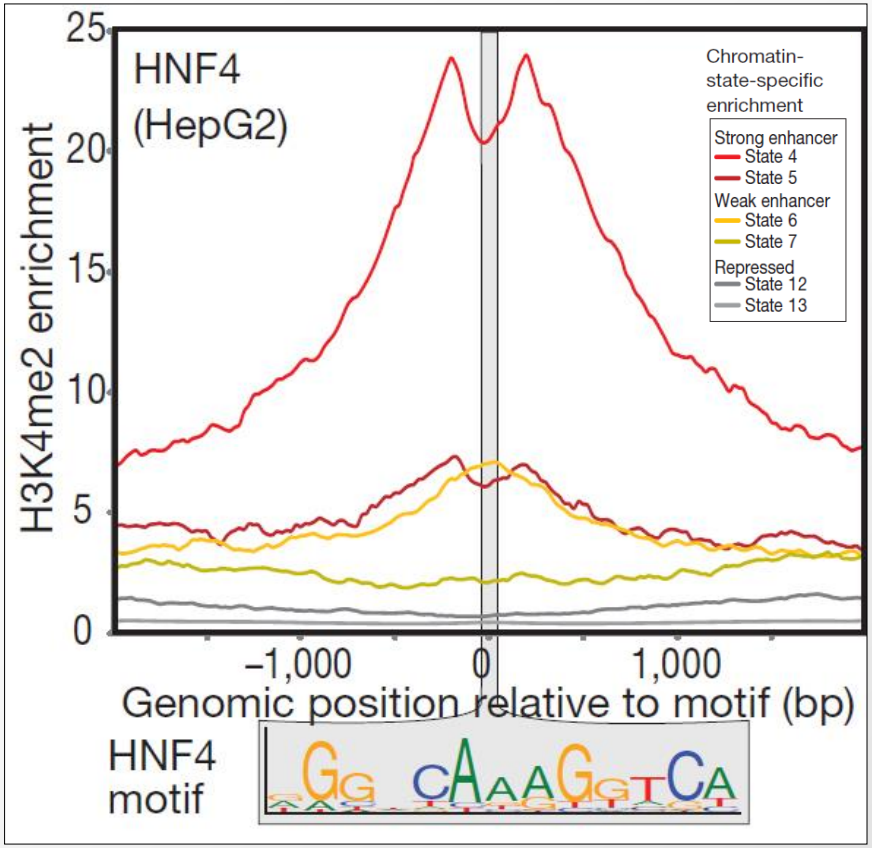
Regulatory motifs predicted to drive enhancer modules

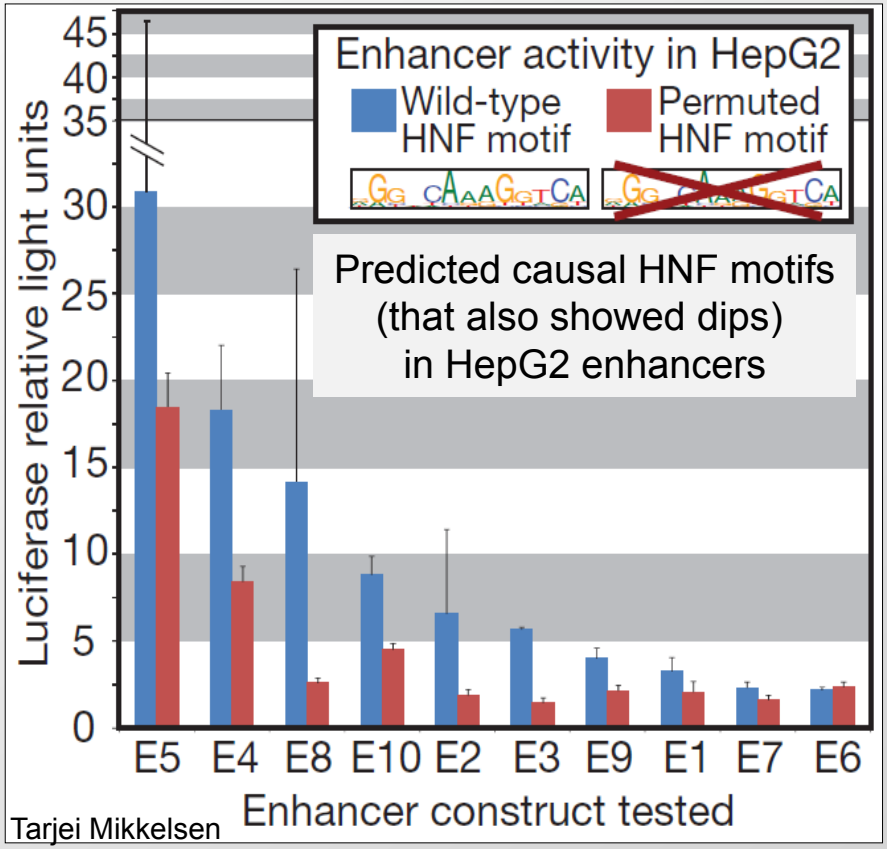


Source: Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes." Nature 518, no. 7539 (2015): 317-330.

- **Activator and repressor motifs consistent with tissues**

Causal motifs supported by dips & enhancer assays





Source: Ernst, Jason et al. "Mapping and analysis of chromatin state dynamics in nine human cell types." Nature 473, no.7345 (2011): 43-49.

Dip evidence of TF binding (nucleosome displacement)

Enhancer activity halved by single-motif disruption

➔ Motifs bound by TF, contribute to enhancers

Goals for today: Computational Epigenomics

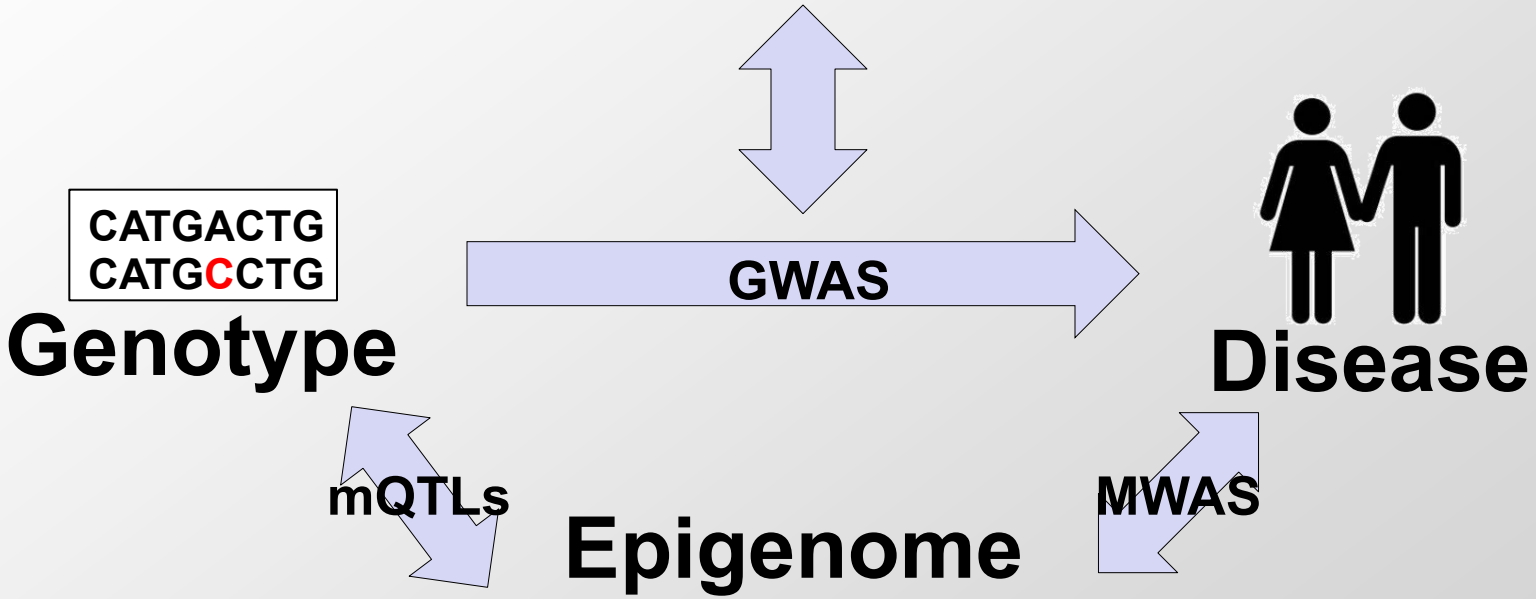
1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - A multi-variate HMM for chromatin combinatorics
   - Promoter, transcribed, intergenic, repressed, repetitive states
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

(Future: Chromatin states to interpret disease-associated variants)

Interpreting disease-association signals

Interpret variants using reference states

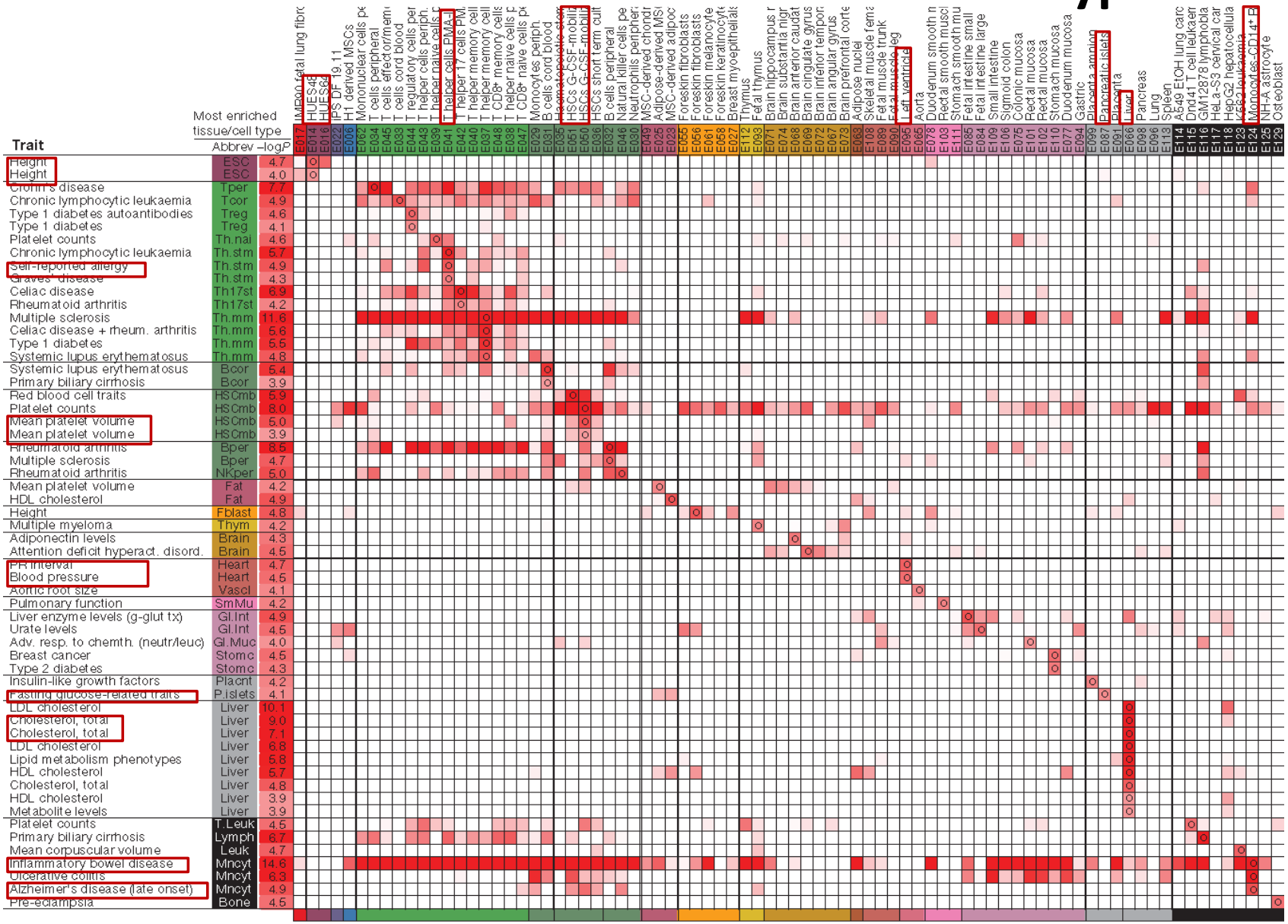
- Chromatin states: Enhancers, promoters, motifs
- Enrichment in individual loci, across 1000s of SNPs in T1D



Epigenome changes in disease

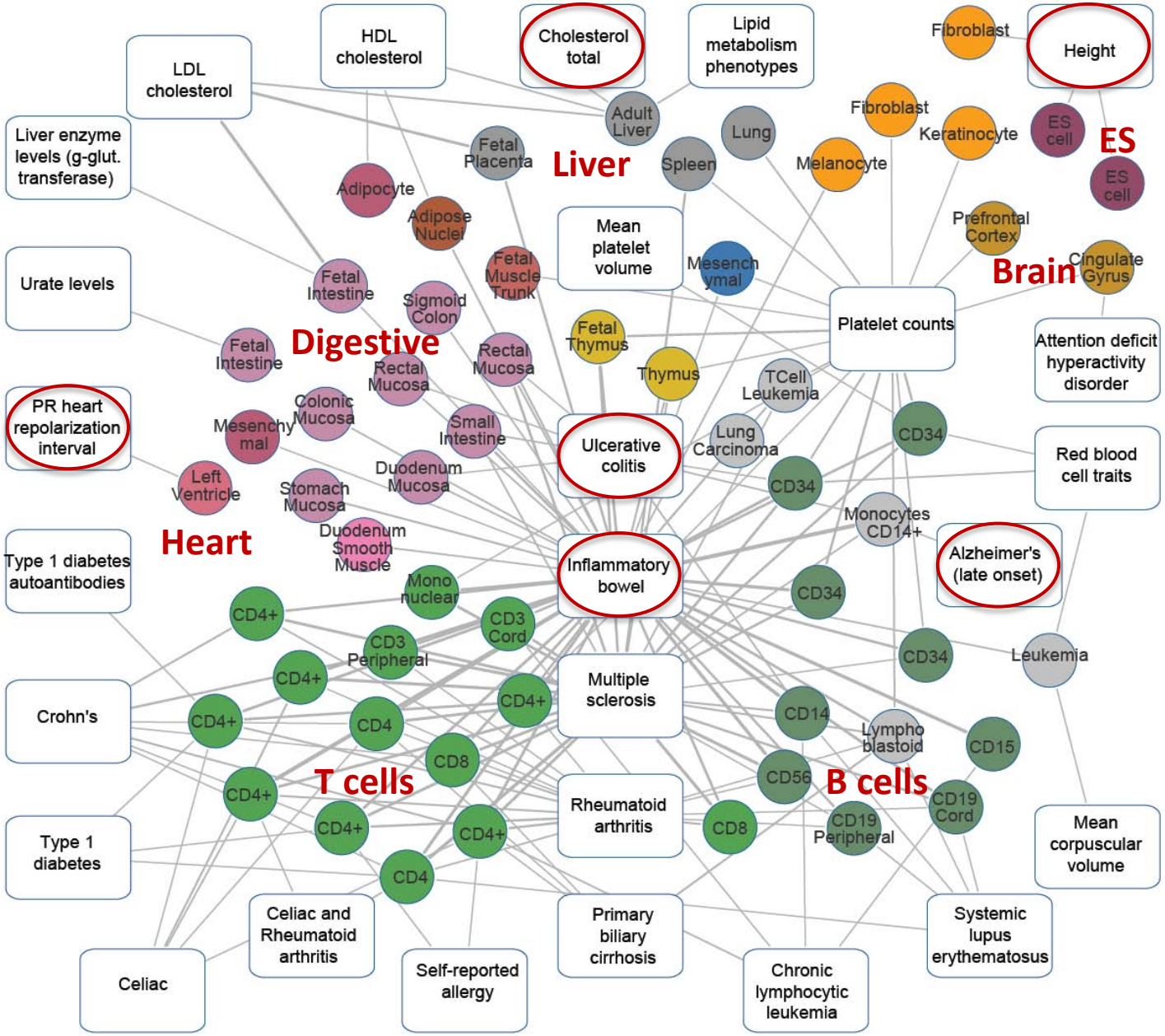
- Molecular phenotypic changes in patients vs. controls
- Small variation in brain methylomes, mostly genotype-driven
- 1000s of brain-specific enhancers increase methylation in Alzheimer's

GWAS hits in enhancers of relevant cell types



Source: Roadmap Epigenomics Consortium et al. "Integrative analysis of 111 reference human epigenomes." Nature 518, no. 7539 (2015): 317-330.

Linking traits to their relevant cell/tissue types



HaploReg: systematic mining of GWAS variants

Query SNP: rs4684847 and variants with r^2 >= 0.8

pos (hg19)	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL tissues	Motifs changed	Drivers disrupted	GENCODE genes	dbSNP func annot
chr3:12329783	chr3:12288284	0.95	0.97	rs17036160	C	T	0.01	0.06	0.04	0.12		24 organs	7 organs	4 organs			4 altered motifs		PPARG	intronic
chr3:12336507	chr3:12295008	0.95	0.97	rs11709077	G	A	0.01	0.07	0.04	0.12		LNG	9 organs	15 organs			4 altered motifs		PPARG	intronic
chr3:12344730	chr3:12303231	0.94	0.97	rs11712037	C	G	0.01	0.08	0.04	0.12			6 organs	BLD			AP-1,TCF11::MafG		PPARG	intronic
chr3:12351521	chr3:12310022	0.95	0.97	rs35000407	T	G	0.01	0.07	0.04	0.12		LNG	5 organs				Smad		PPARG	intronic
chr3:12360884	chr3:12319385	0.95	0.97	rs150732434	TG	T	0.01	0.07	0.04	0.12		FAT	7 organs	MUS,VAS	CFOS		Hdx,Sox,TATA		PPARG	intronic
chr3:12365308	chr3:12323809	0.95	0.97	rs13083375	G	T	0.01	0.07	0.04	0.12		BLD	BLD, FAT				Homez,Sox,YY1		PPARG	intronic
chr3:12369401	chr3:12327902	0.95	0.97	rs13064760	C	T	0.01	0.07	0.04	0.12			7 organs				9 altered motifs		PPARG	intronic
chr3:12375956	chr3:12334457	0.95	0.97	rs2012444	C	T	0.01	0.07	0.04	0.12			SKIN, FAT, BLD				7 altered motifs		PPARG	intronic
chr3:12383265	chr3:12341766	0.96	0.99	rs13085211	G	A	0.18	0.10	0.04	0.12			FAT, SKIN				NRSF		PPARG	intronic
chr3:12383714	chr3:12342215	0.96	0.99	rs7638903	G	A	0.18	0.10	0.04	0.12			6 organs	CRVX					PPARG	intronic
chr3:12385828	chr3:12344329	0.95	1	rs11128603	A	G	0.18	0.10	0.04	0.12			CRVX				RXRA		PPARG	intronic
chr3:12386337	chr3:12344838	1	1	rs4684847	C	T	0.01	0.07	0.04	0.12			6 organs						PPARG	intronic
chr3:12388409	chr3:12346910	0.99	1	rs7610055	G	A	0.17	0.09	0.04	0.12			BLD				4 altered motifs		PPARG	intronic
chr3:12389313	chr3:12347814	0.99	1	rs17036326	A	G	0.17	0.09	0.04	0.12			FAT, BL						PPARG	intronic
chr3:12390484	chr3:12348985	0.99	1	rs17036328	T	C	0.17	0.09	0.04	0.12			FAT, CR						PPARG	intronic
chr3:12391207	chr3:12349708	0.99	1	rs6802898	C	T	0.61	0.15	0.04	0.12			FAT, BL						PPARG	intronic
chr3:12391583	chr3:12350084	0.99	1	rs2197423	G	A	0.17	0.09	0.04	0.12		FAT, LIV	8 organ						PPARG	intronic
chr3:12391813	chr3:12350314	0.99	1	rs7647481	G	A	0.17	0.09	0.04	0.12		4 organs	9 organ						PPARG	intronic
chr3:12392272	chr3:12350773	0.99	1	rs7649970	C	T	0.17	0.09	0.04	0.12		5 organs	9 organ						PPARG	intronic
chr3:12393125	chr3:12351626	1	1	rs1801282	C	G	0.01	0.07	0.04	0.12		FAT, LIV	9 organ						PPARG	missense
chr3:12393682	chr3:12352183	0.99	1	rs17036342	A	G	0.17	0.09	0.04	0.12		FAT	9 organ						PPARG	intronic
chr3:12394840	chr3:12353341	0.99	1	rs1899951	C	T	0.61	0.15	0.04	0.12		FAT	9 organs				Mef2		PPARG	intronic
chr3:12395645	chr3:12354146	0.99	1	rs4684848	G	A	0.61	0.15	0.04	0.12		FAT, BLD	9 organs	ADRL,GI,CRVX	5 bound proteins				PPARG	intronic
chr3:12396845	chr3:12355346	0.93	1	rs4135250	A	G	0.17	0.09	0.04	0.13			4 organs	PLCNT					PPARG	intronic
chr3:12396913	chr3:12355414	0.98	1	rs71304101	G	A	0.01	0.07	0.04	0.12			4 organs	PLCNT			Crx,NF-E2		PPARG	intronic
chr3:12396955	chr3:12355456	0.96	1	rs2881654	G	A	0.61	0.15	0.04	0.12			4 organs				7 altered motifs		PPARG	intronic

Adipose_Derived_Mesenchymal_Stem_Cell_Cultured_Cells, CD4+_CD25-_IL17+_PMA-Ionomcyin_stimulated_Th17_Primary_Cells, Muscle_Satellite_Cultured_Cells, Penis_Foreskin_Fibroblast_Primary_Cells_skin01, Penis_Foreskin_Fibroblast_Primary_Cells_skin02, Penis_Foreskin_Keratinocyte_Primary_Cells_skin02, Penis_Foreskin_Keratinocyte_Primary_Cells_skin03, A549_EtOH_0.02pct_Lung_Carcinoma, HeLa-S3_Cervical_Carcinoma, NHEK-Epidermal_Keratinocytes

Source: Ward, Lucas D. and Manolis Kellis. "HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants." Nucleic Acids Research 40, no. D1 (2012): D930-D934.

- **Start with any list of SNPs or select a GWA study**
  - Mine ENCODE and Roadmap epigenomics data for hits
  - Hundreds of assays, dozens of cells, conservation, motifs
  - Report significant overlaps and link to info/browser
- **Try it out: http://compbio.mit.edu/HaploReg** Ward, Kellis NAR 2011

6.047 / 6.878 / HST.507 Computational Biology
Fall 2015