Applications of HMMs in Epigenomics

Yuzhen Ye
School of Informatics and Computing
Indiana University, Bloomington

Spring 2017

Contents

- Background: chromatin structure & DNA methylation
- Epigenomic projects
  - ENCODE
  - modENCODE
  - Human epigenome atlas
- Techniques
  - ChIP-Seq
  - MeDIP (MeDIP-chip & MeDIP-seq)
- Two applications
  - MeDIP-HMM (second-order HMM)
  - ChromHMM (multivariate HMM)

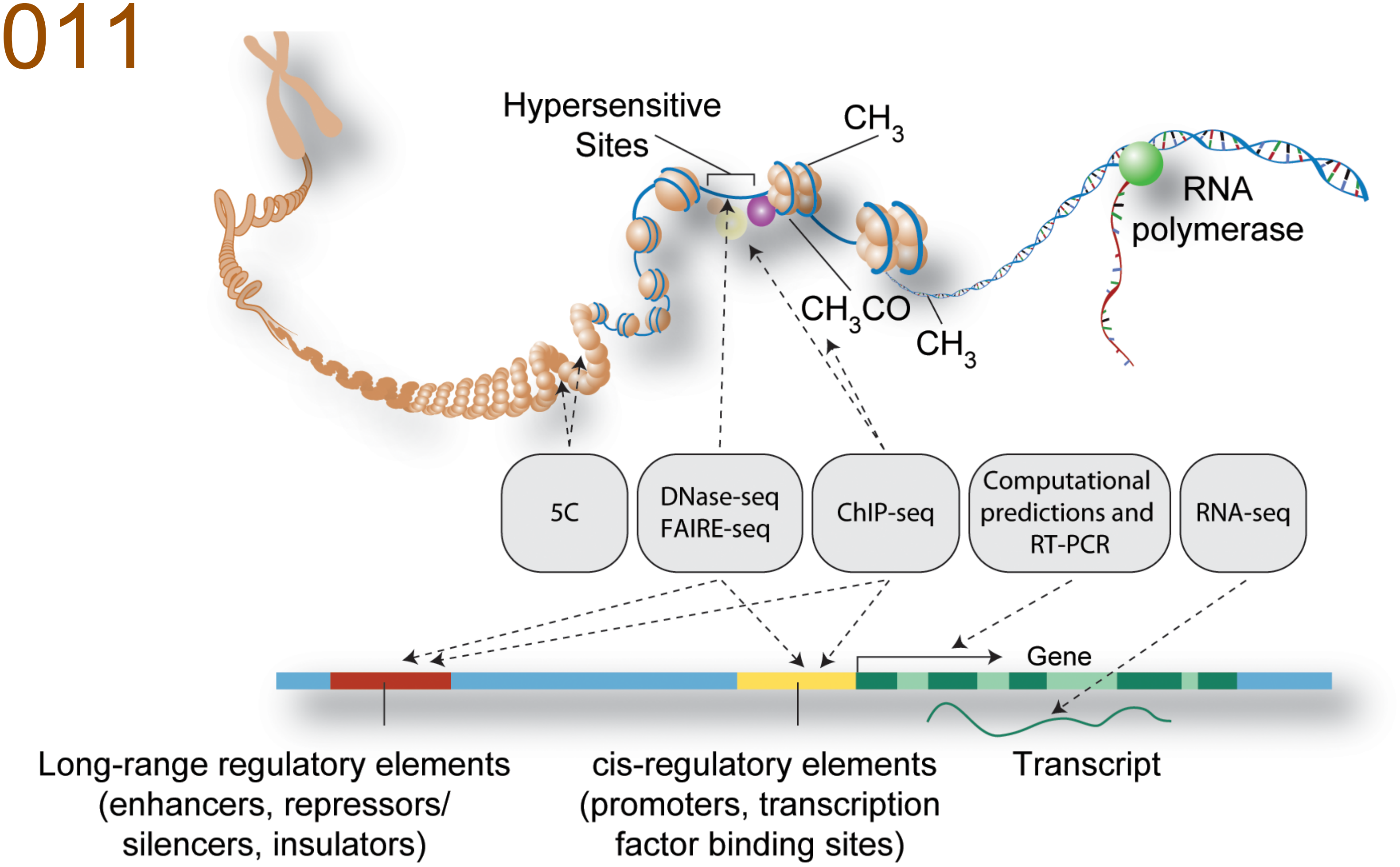
Background

- Genomic DNA is packaged into a complex molecular structure known as chromatin. This structure mediates the interaction between the genome and all types of regulatory and transcriptional molecules.
- In vertebrate genomes, methylation at position 5 of the cytosine in CpG dinucleotides is a heritable “epigenetic” mark that has been connected with both **transcriptional silencing and imprinting**
  - **Ref: DNA methylation patterns and epigenetic memory (***Genes & Dev. 2002. 16: 6-21* **)**

ENCODE

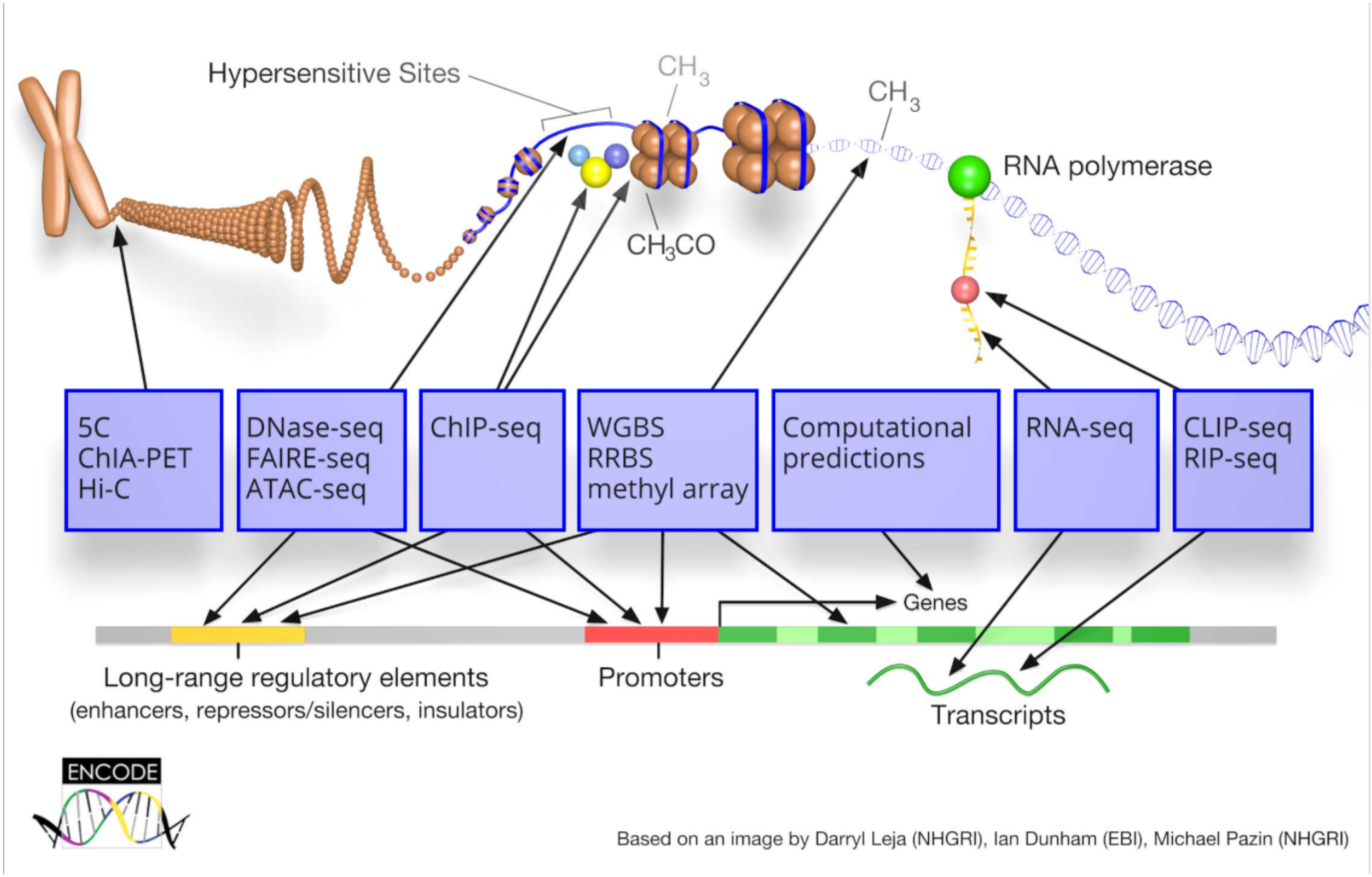
- Encyclopedia of DNA Elements
  - “The ENCODE Consortium is integrating multiple technologies and approaches in a collective effort to discover and define the functional elements encoded in the human genome, including **genes, transcripts, and transcriptional regulatory regions, together with their attendant chromatin states and DNA methylation patterns**. ”
  - Ref: A User's Guide to the Encyclopedia of DNA Elements (ENCODE) (PLoS Biology, 2011)
- Initial phase launched in 2003—1% of the human genome
  - Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project (Nature, June 13, 2007)

2011


Hypersensitive Sites
CH_3
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9(4): e1001046. doi:10.1371/journal.pbio.1001046
http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046

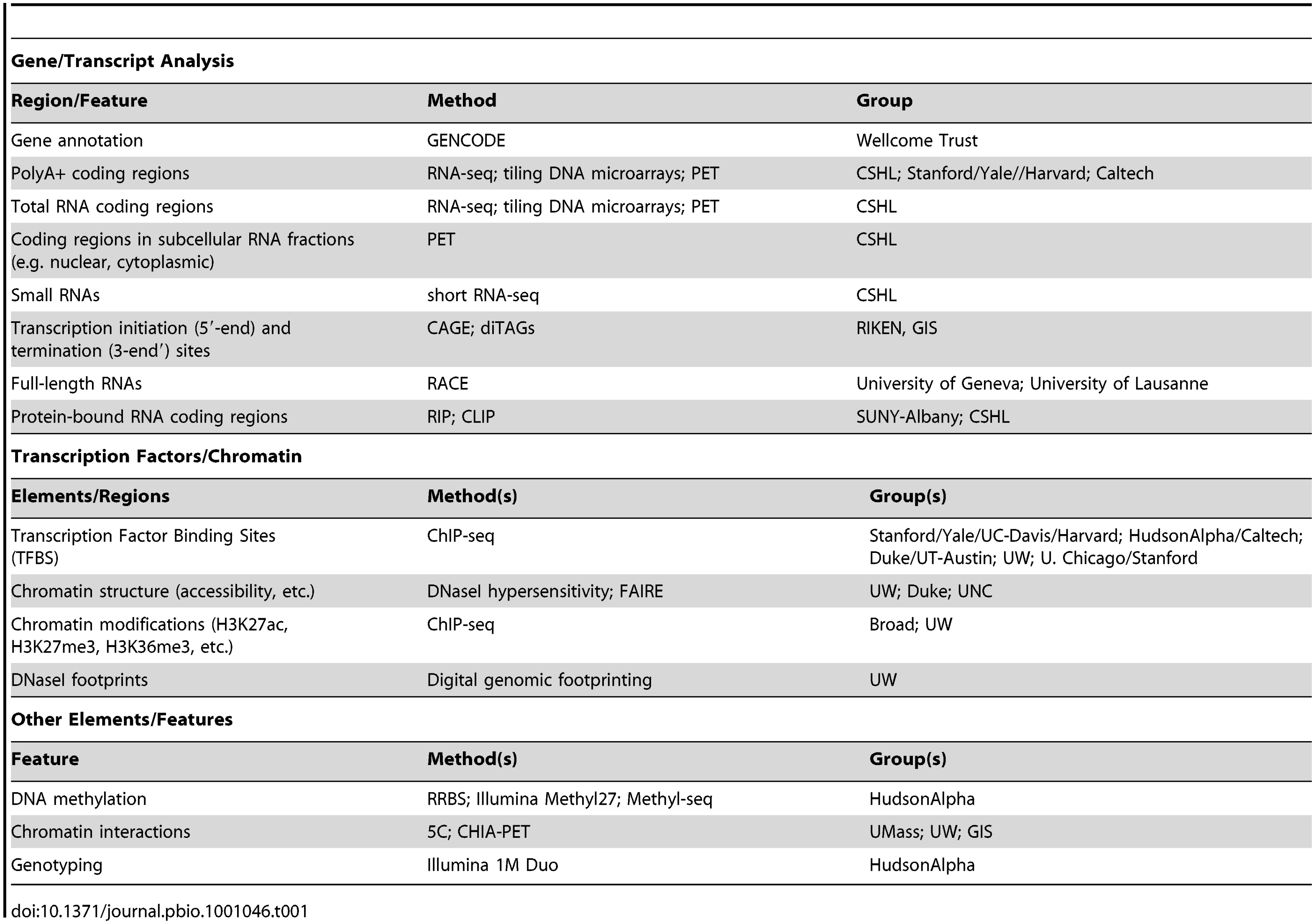
2017



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Ref: https://www.encodeproject.org

Table 1. Experimental assays used by the ENCODE Consortium.



Gene/Transcript Analysis

Region/Feature	Method	Group
Gene annotation	GENCODE	Wellcome Trust
PolyA+ coding regions	RNA-seq; tiling DNA microarrays; PET	CSHL; Stanford/Yale//Harvard; Caltech
Total RNA coding regions	RNA-seq; tiling DNA microarrays; PET	CSHL
Coding regions in subcellular RNA fractions (e.g. nuclear, cytoplasmic)	PET	CSHL
Small RNAs	short RNA-seq	CSHL
Transcription initiation (5′-end) and termination (3-end′) sites	CAGE; diTAGs	RIKEN, GIS
Full-length RNAs	RACE	University of Geneva; University of Lausanne
Protein-bound RNA coding regions	RIP; CLIP	SUNY-Albany; CSHL

Transcription Factors/Chromatin

Elements/Regions	Method(s)	Group(s)
Transcription Factor Binding Sites (TFBS)	ChIP-seq	Stanford/Yale/UC-Davis/Harvard; HudsonAlpha/Caltech; Duke/UT-Austin; UW; U. Chicago/Stanford
Chromatin structure (accessibility, etc.)	DNaseI hypersensitivity; FAIRE	UW; Duke; UNC
Chromatin modifications (H3K27ac, H3K27me3, H3K36me3, etc.)	ChIP-seq	Broad; UW
DNaseI footprints	Digital genomic footprinting	UW

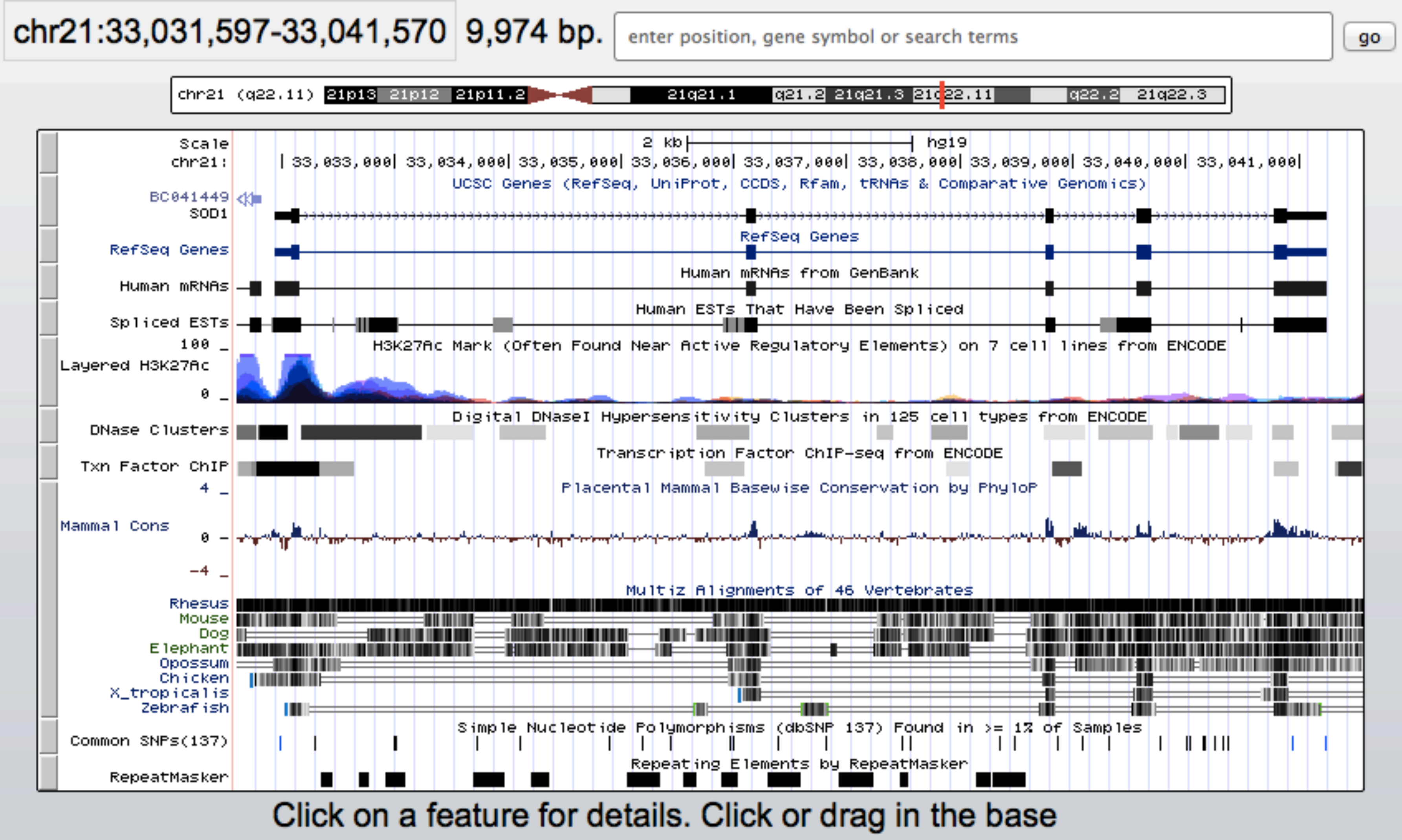
Other Elements/Features

Feature	Method(s)	Group(s)
DNA methylation	RRBS; Illumina Methyl27; Methyl-seq	HudsonAlpha
Chromatin interactions	5C; CHIA-PET	UMass; UW; GIS
Genotyping	Illumina 1M Duo	HudsonAlpha

doi:10.1371/journal.pbio.1001046.t001

The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9(4): e1001046. doi:10.1371/journal.pbio.1001046
http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046

ENCODE data



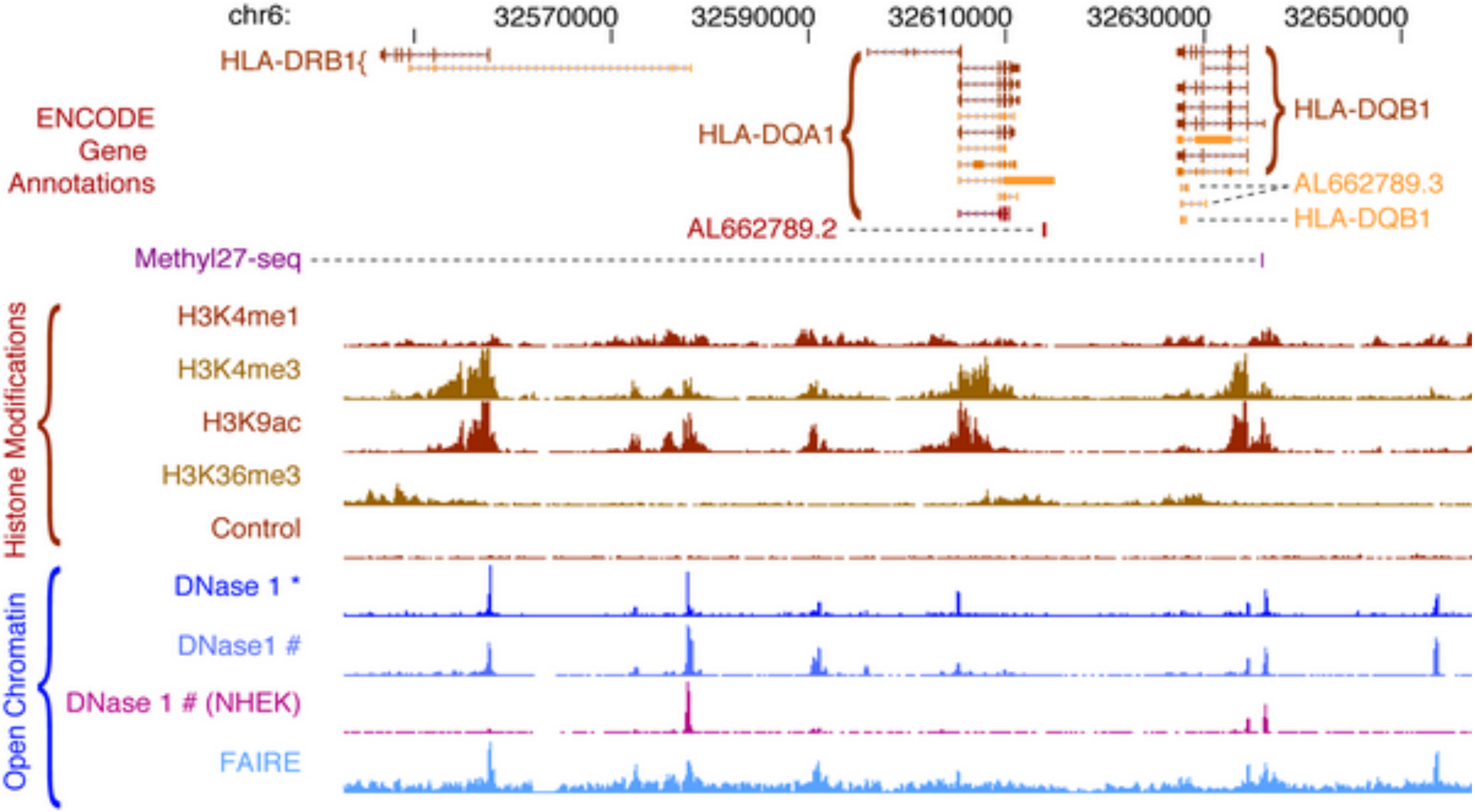
chr21:33,031,597-33,041,570 9,974 bp.

Click on a feature for details. Click or drag in the base

enter position, gene symbol or search terms
go
chr21 (q22.11) 21p13 21p12 21p11.2 21q21.1 q21.2 21q21.3 21q22.11 q22.2 21q22.3
Scale
chr21:
2 kb
hg19
33,033,000 33,034,000 33,035,000 33,036,000 33,037,000 33,038,000 33,039,000 33,040,000 33,041,000
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
BC041449
SOD1
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
4
Mammal Cons
0
-4
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Common SNPs(137)
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker

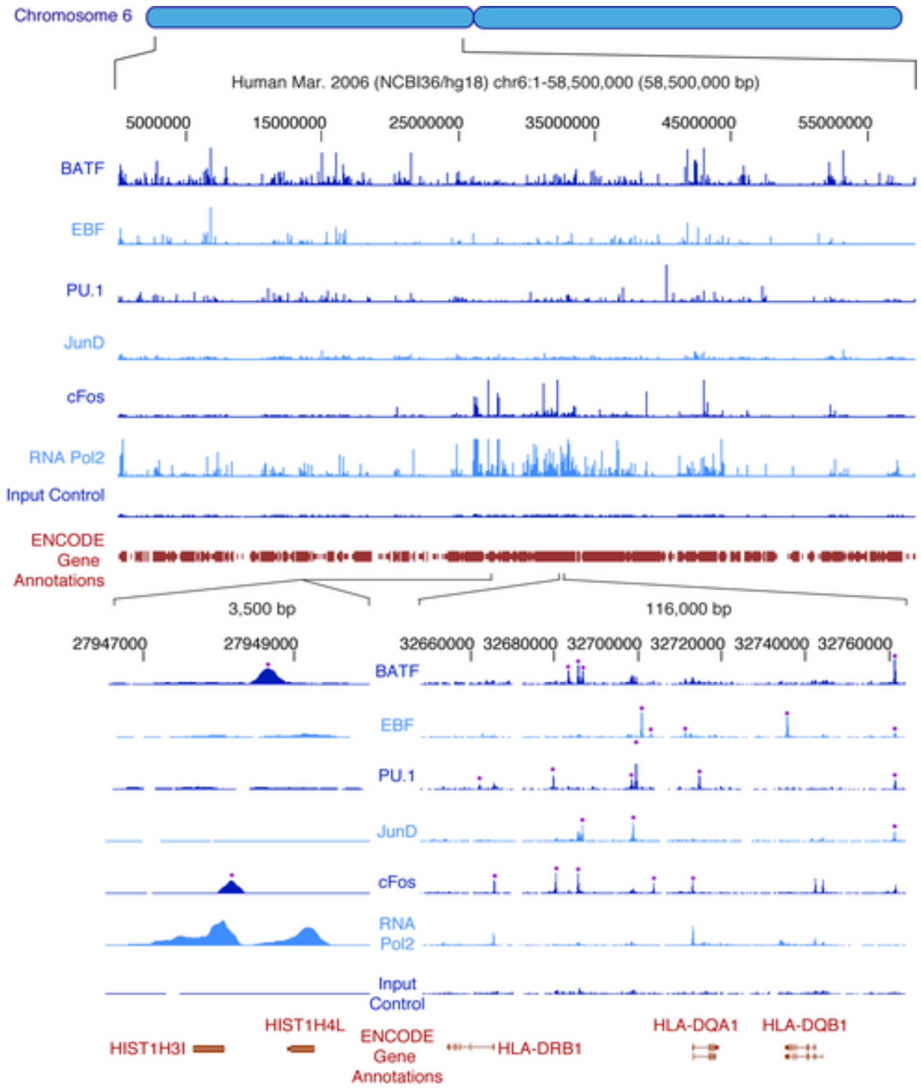
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

Figure 4. ENCODE chromatin annotations in the HLA locus.



The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9(4): e1001046. doi:10.1371/journal.pbio.1001046
http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046

Figure 5. Occupancy of transcription factors and RNA polymerase 2 on human chromosome 6p as determined by ChIP-seq.

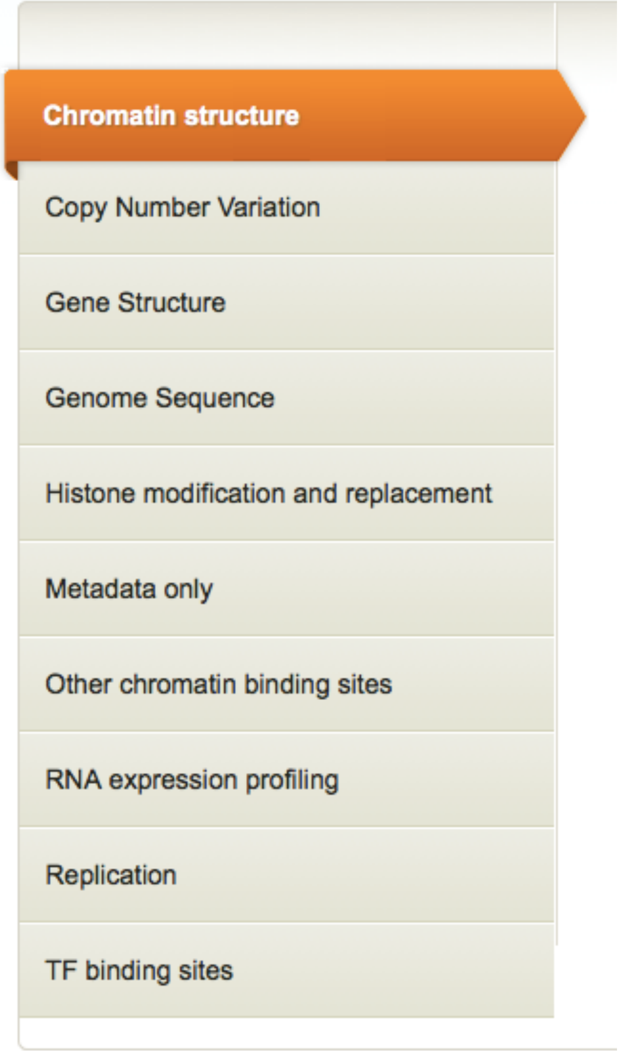


The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9(4): e1001046. doi:10.1371/journal.pbio.1001046
http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046

modENCODE

http://www.modencode.org/

“ The modENCODE Project will try to identify all of the sequence-based functional elements in the *Caenorhabditis elegans* and *Drosophila melanogaster* genomes.”



Human epigenome atlas

- Successive releases of the Atlas will provide progressively more detailed insights into locus-specific epigenomic states, including histone marks and DNA methylation marks across specific tissues and cell types, developmental stages, physiological conditions, genotypes, and disease states.

CHIP-seq

- By combining chromatin immunoprecipitation (ChIP) assays with sequencing, ChIP sequencing (ChIP-Seq) is a powerful method for identifying genome-wide DNA binding sites for transcription factors and other proteins.
- Following ChIP protocols, DNA-bound protein is immunoprecipitated using a specific antibody.
- The bound DNA is then coprecipitated, purified, and sequenced.

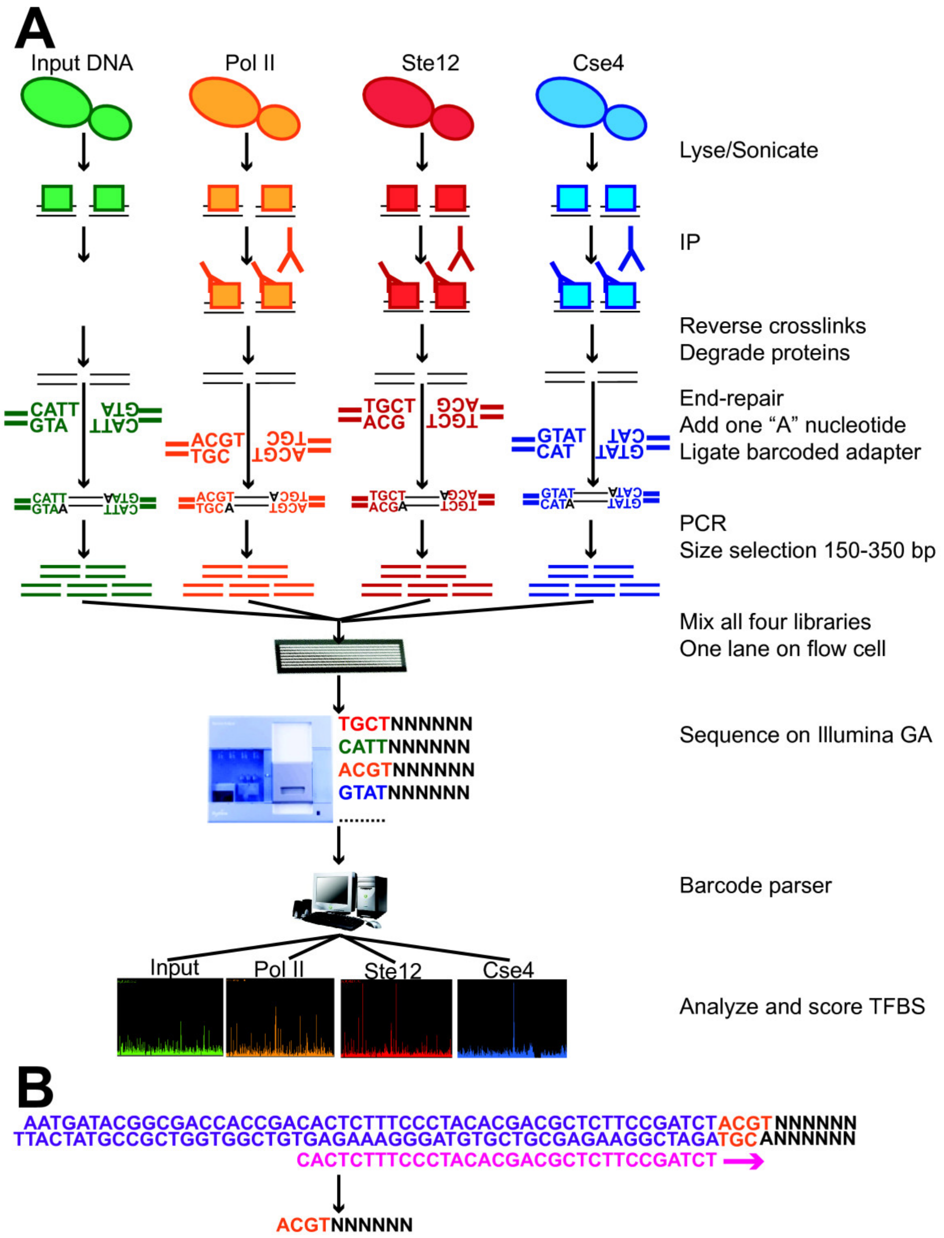
https://www.illumina.com/techniques/sequencing/dna-sequencing/chip-seq.html

ChIP-seq

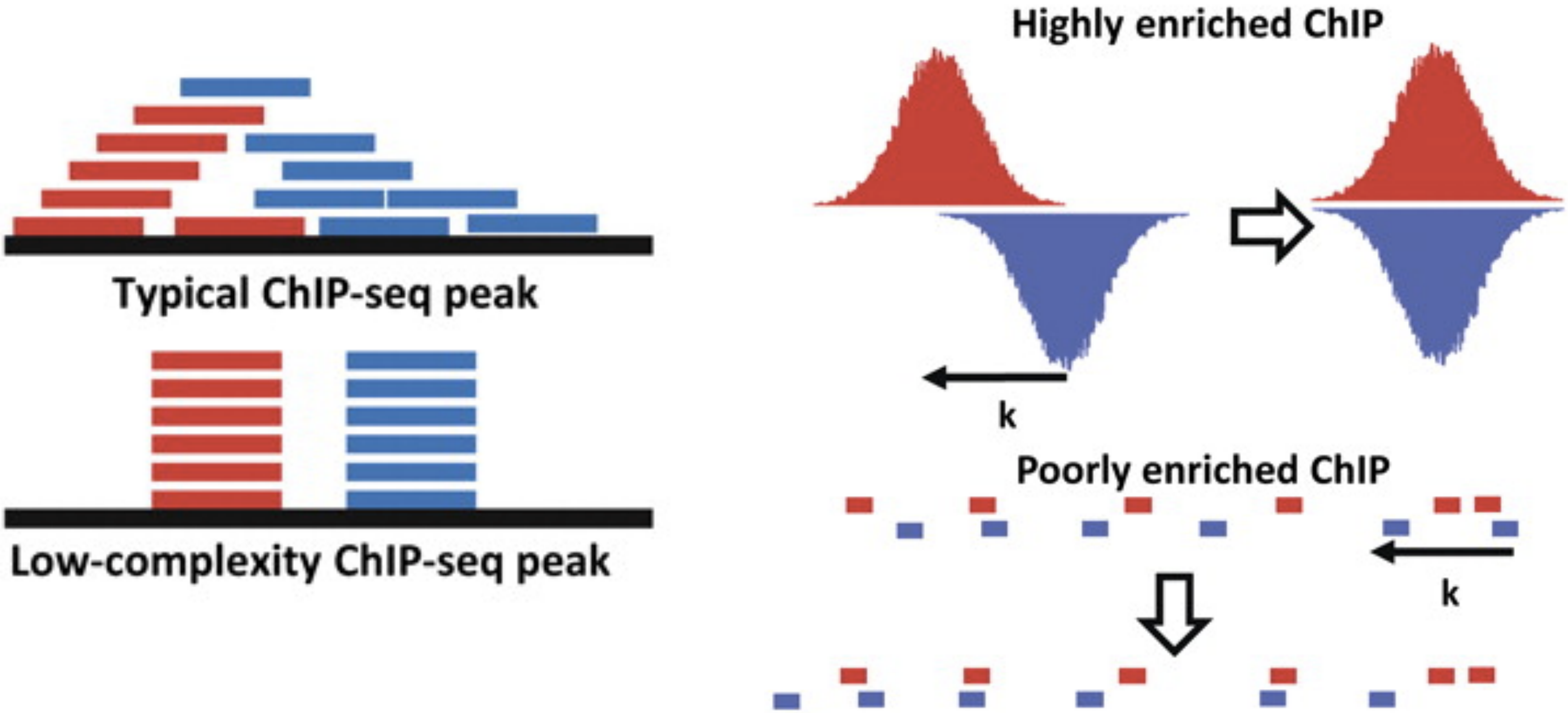
- Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing (ChIP-seq) has become a valuable and widely used approach for mapping the genomic location of **transcription-factor binding and histone modifications** in living cells.
  - Genome-Wide Mapping of in Vivo Protein-DNA Interactions (Science, 2007); 1946 binding sites of the Neuron-restrictive silencer factor (NRSF) were mapped at ~50bp resolution
- There are considerable differences in how these experiments are conducted, how the results are scored and evaluated for quality, and how the data and metadata are archived for public use.
  - Genome Res. 2012 Sep;22(9):1813-31

Barcoded ChIP-seq

Efficient yeast ChIP-seq using multiplex short-read DNA sequencing (*BMC Genomics* 2009, **10**:37)



CHIP-seq: peak detection



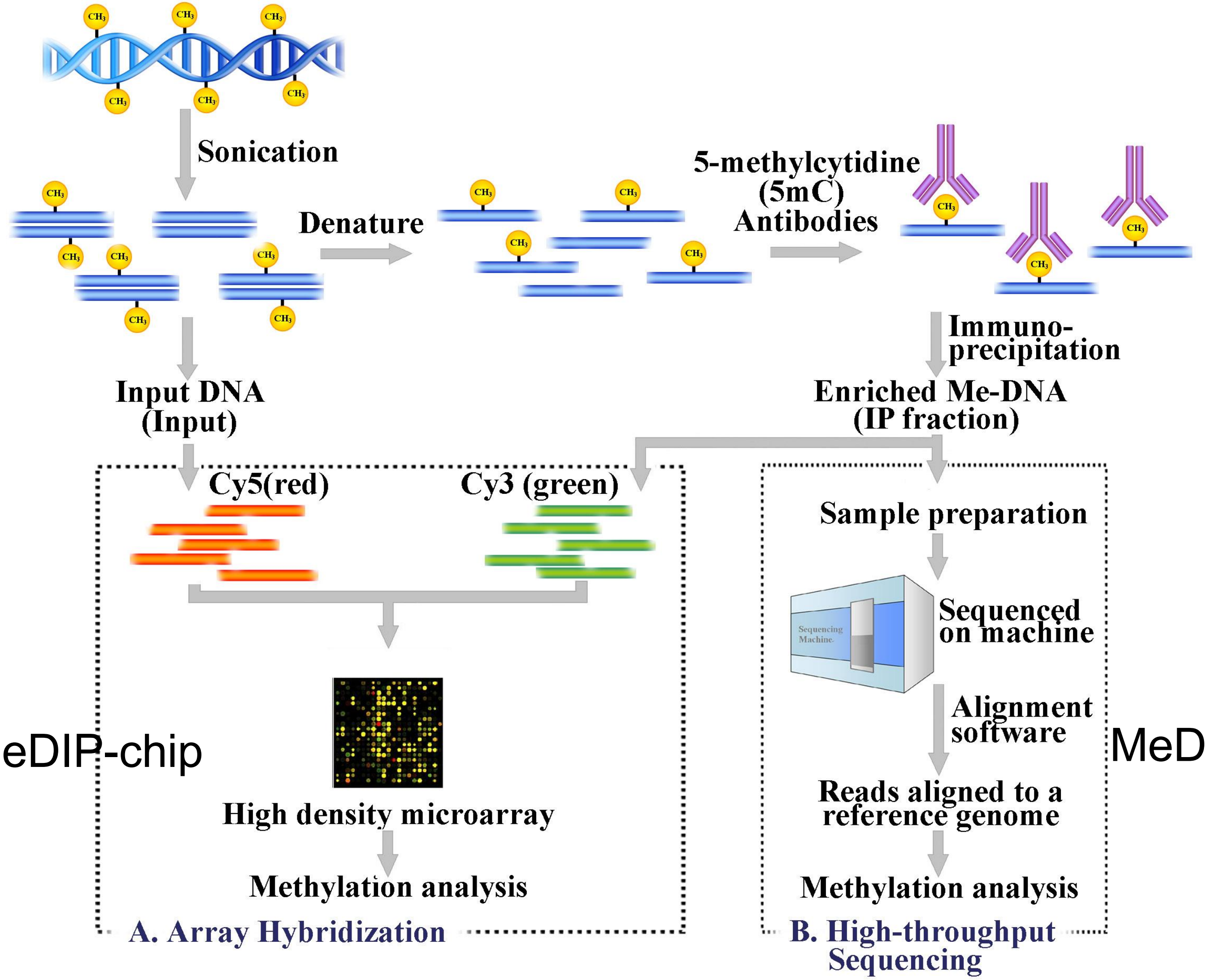
DNase-seq

- DNase digestion followed by sequencing.
- DNase I hypersensitive sites (DHS), short regions of chromatin that are highly sensitive to cleavage by DNase I, typically occur in nucleosome free (nucleosome-depleted) regions as a result of transcription factor binding.
- **DNA sequence motif analysis** on DHS data was proposed as a method for discovering the binding sites of multiple transcription factors in a single experiment.
- DNase-seq profile resemble to some extent the data from ChIP-seq, with important differences (Front. Genet., 31 October 2012)

Genome-wide DNA methylation profiling

- Restriction enzyme-based methods
  - Use one or more enzymes that will restrict DNA only if it is unmethylated (e.g. HpaII or NotI), or methylated (e.g. McrBC).
  - Limited to the analysis of CpG sites located within the enzyme recognition site(s).
- Bisulfite-conversion based approaches
  - Unmethylated cytosines are converted to uracil; offer single CpG resolution; the gold standard for DNA methylation analysis
  - Con: reduction of sequence complexity following bisulfite conversion (Bi-chip) & Bi-seq approach is expensive.
  - Align BS-treated reads to a reference genome
- Immunoprecipitation-based methods
  - Use either 5-methylcytosine-specific antibodies (MeDIP) or methyl-binding domain proteins, to enrich for the methylated (or unmethylated) fraction of the genome.

Methylation analysis by DNA immunoprecipitation (MeDIP)



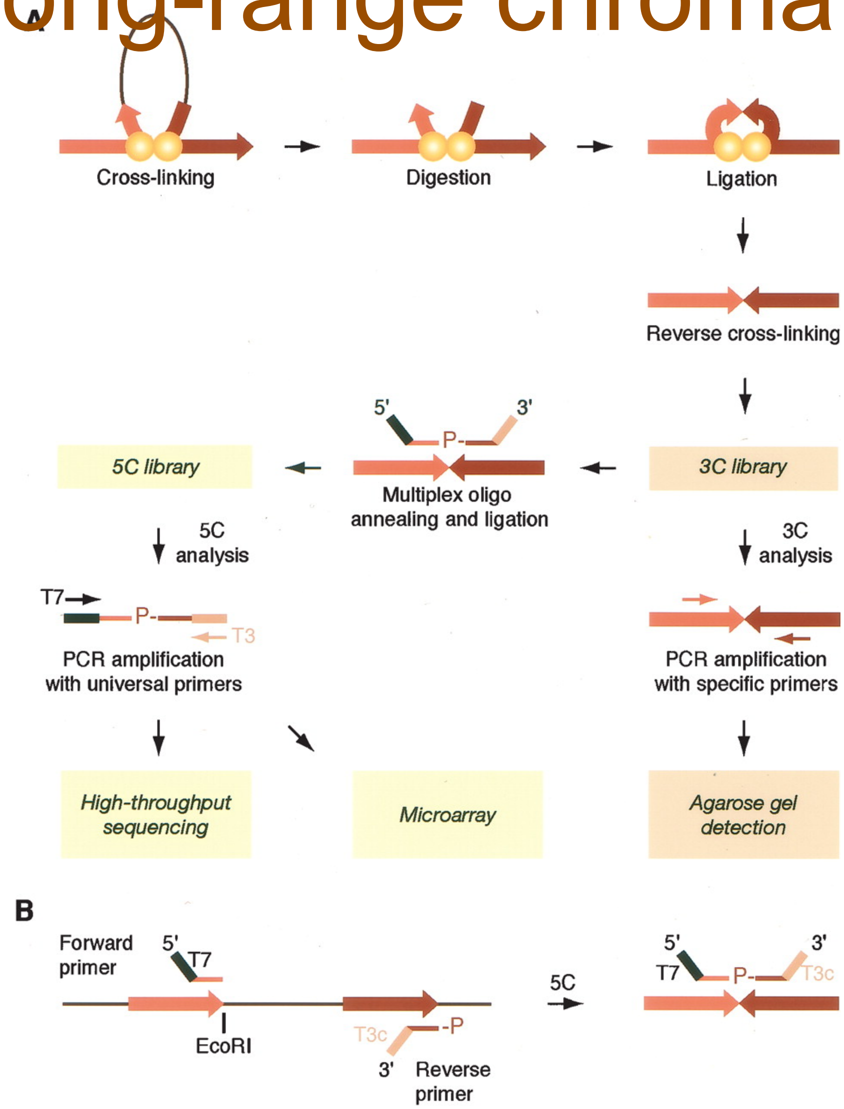
MeDIP-chip

MeDIP-seq

MeDIP can be coupled with microarray or high-throughput sequencing

Image from wikipedia

Long-range chromatin interaction



Dostie J et al. Genome Res. 2006;16:1299-1309

Long-range Chromatin interactions:
Chromosome Conformation Capture Carbon Copy (5C)

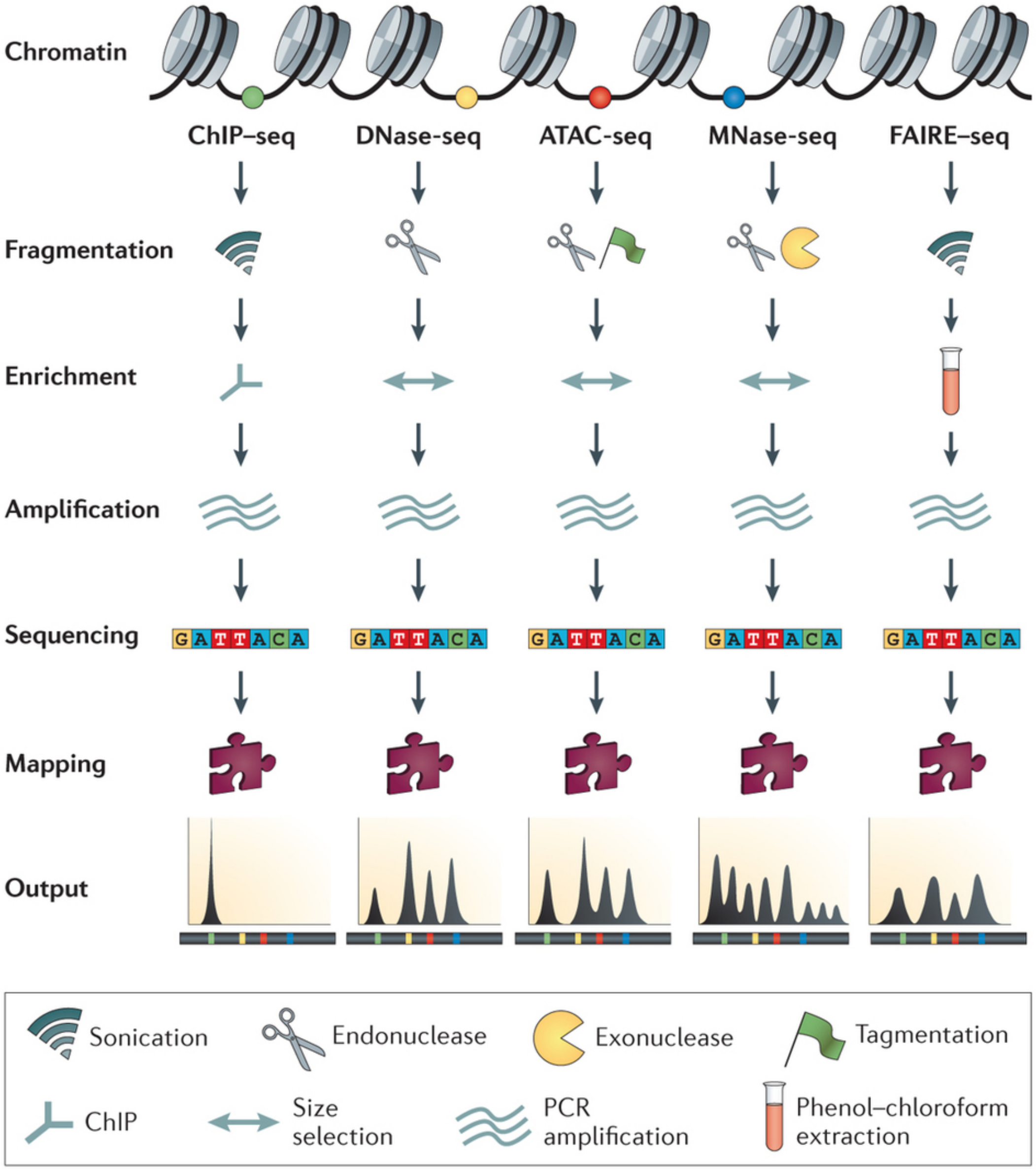
Comparison of chromatin profiling experiments

Complementary chromatin profiling experiments reveals different aspects of chromatin structure:

- ChIP–seq reveals binding sites of specific transcription factors (TFs);
- DNase-seq, ATAC-seq and FAIRE–seq reveal regions of open chromatin;
- MNase-seq identifies well-positioned nucleosomes.

These experiments differ in the enrichment method

- In ChIP–seq, specific antibodies are used to extract DNA fragments that are bound to the target protein.
- In DNase-seq, chromatin is lightly digested by the DNase I endonuclease. Size selection is used to enrich for fragments that are produced in regions of chromatin where the DNA is highly sensitive to DNase I attack.
- ATAC-seq uses an engineered Tn5 transposase to cleave DNA and to integrate primer DNA sequences into the cleaved genomic DNA (tagmentation).
- Micrococcal nuclease (MNase) is an endo–exonuclease that processively digests DNA until an obstruction, such as a nucleosome, is reached.
- In FAIRE–seq, formaldehyde is used to crosslink chromatin, and phenol–chloroform is used to isolate sheared DNA.



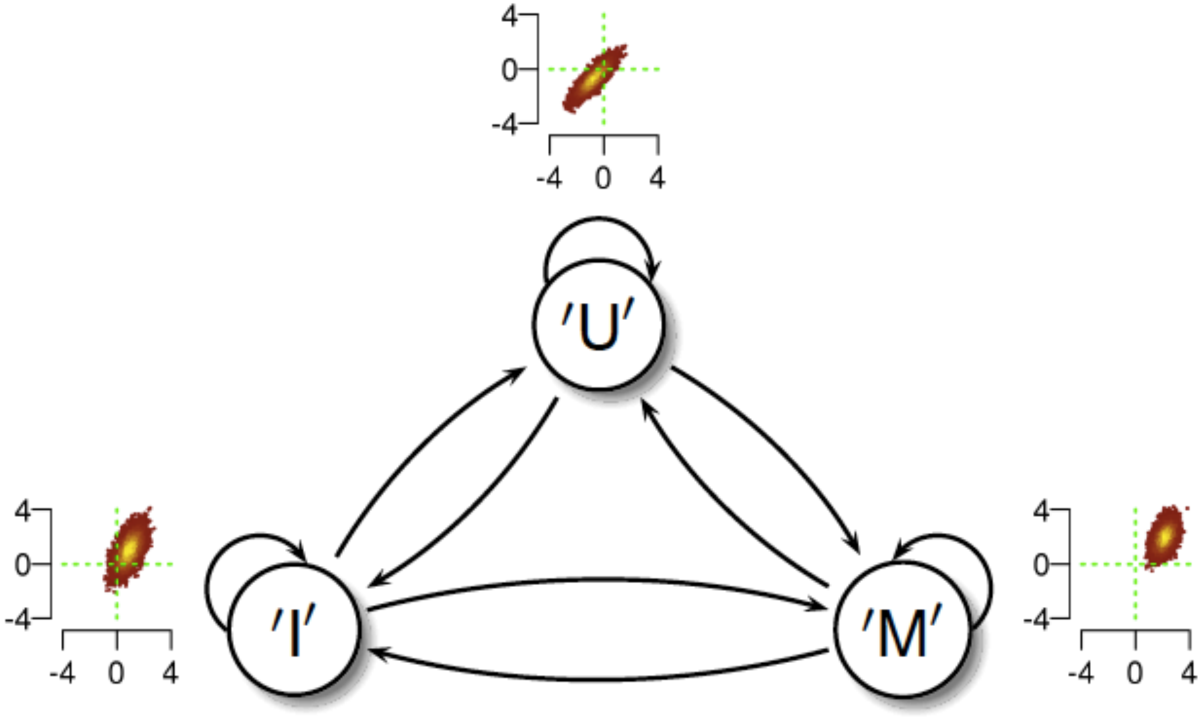
Ref: http://www.nature.com/nrg/journal/v15/n11/fig_tab/nrg3788_F1.html

Nature Reviews | Genetics

A HMM application for the inference of DNA methylation

- **MeDIP-HMM: genome-wide identification of distinct DNA methylation states from high-density tiling arrays**
- MeDIP-HMM utilizes a higher-order state-transition process improving modeling of spatial dependencies between chromosomal regions
- Enables a differentiation between **unmethylated, methylated and highly methylated** genomic regions.
- Training algorithm: a Bayesian Baum-Welch algorithm integrating prior knowledge on methylation levels.
- Application of MeDIP-HMM to the analysis of the Arabidopsis root methylome and systematically investigate the benefit of using higher-order HMMs.
- *Bioinformatics (2012) doi: 10.1093*

MeDIP-HMM: three-state architecture



Second-order HMM

Multivariate Gaussian Emission Distribution:

$$b_i(\vec{o}) := \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_i)}} \exp\left(-\frac{1}{2}(\vec{o}-\vec{\mu}_i)\cdot\Sigma_i^{-1}\cdot(\vec{o}-\vec{\mu}_i)^T\right)$$

Chromatin-state decoding

- **Automated mapping of large-scale chromatin structure in ENCODE**
  - *Bioinformatics (2008) 24 (17): 1911-1916.*

- **ChromHMM: automating chromatin-state discovery and characterization**
  - Nature Methods 9, 215–216 (2012)

Integrative annotation of chromatin elements from ENCODE data

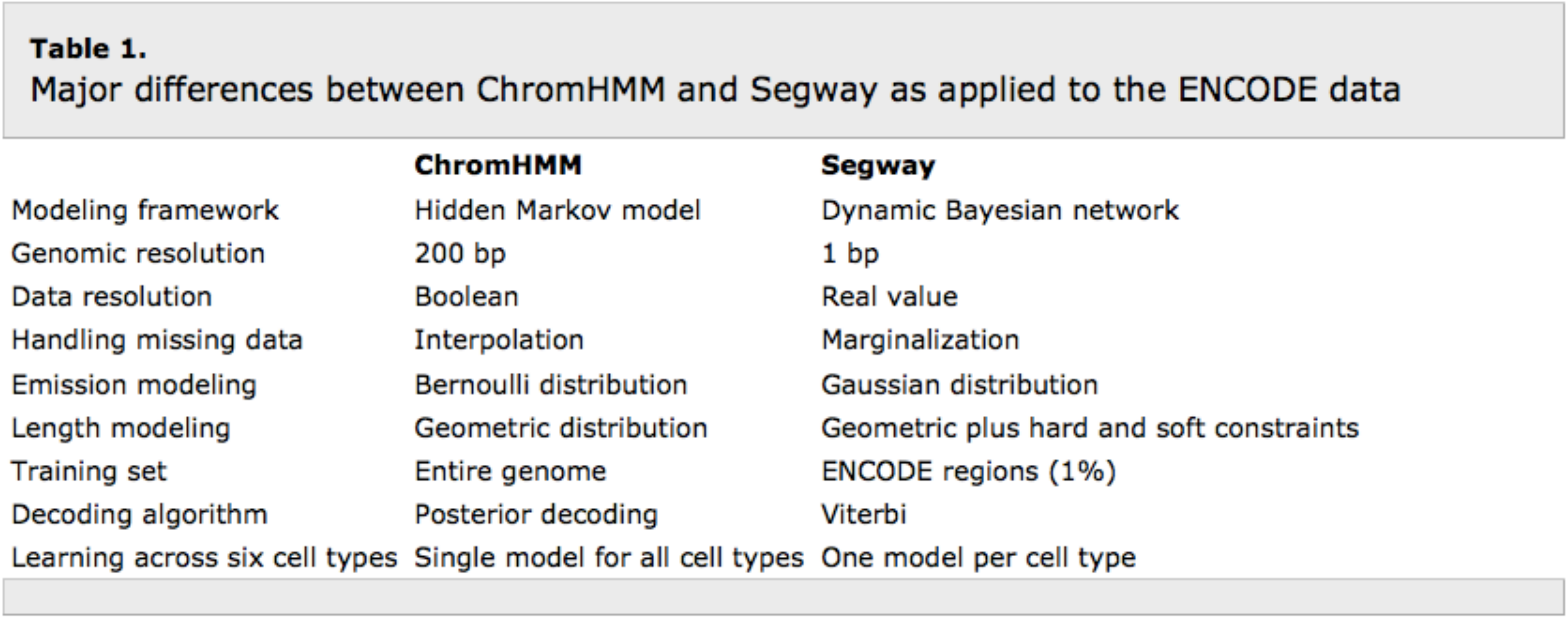
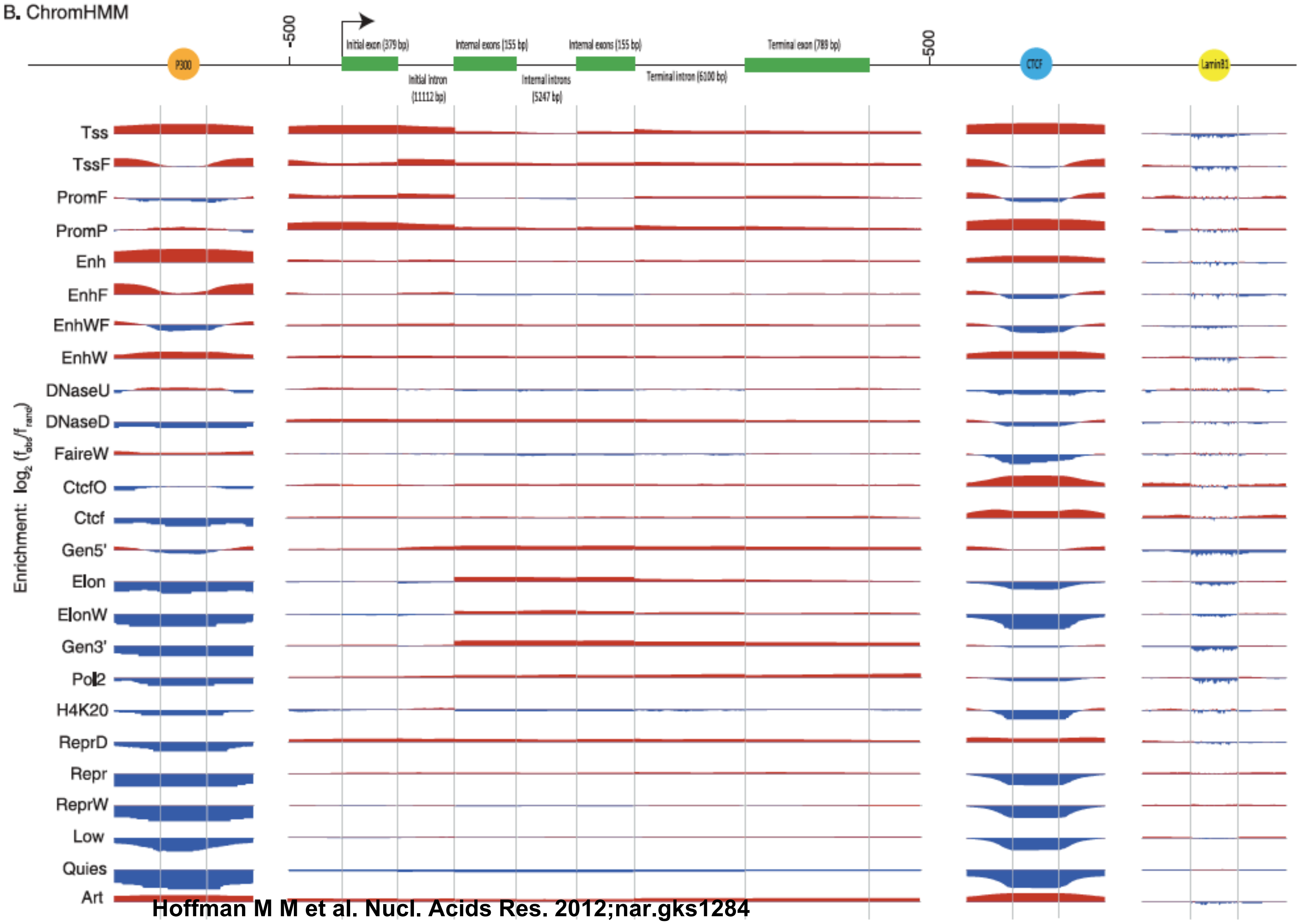


Table 1.
Major differences between ChromHMM and Segway as applied to the ENCODE data

	ChromHMM	**Segway**
Modeling framework	Hidden Markov model	Dynamic Bayesian network
Genomic resolution	200 bp	1 bp
Data resolution	Boolean	Real value
Handling missing data	Interpolation	Marginalization
Emission modeling	Bernoulli distribution	Gaussian distribution
Length modeling	Geometric distribution	Geometric plus hard and soft constraints
Training set	Entire genome	ENCODE regions (1%)
Decoding algorithm	Posterior decoding	Viterbi
Learning across six cell types	Single model for all cell types	One model per cell type

Ref: *Nucl. Acids Res. (2013) 41 (2): 827-841.*

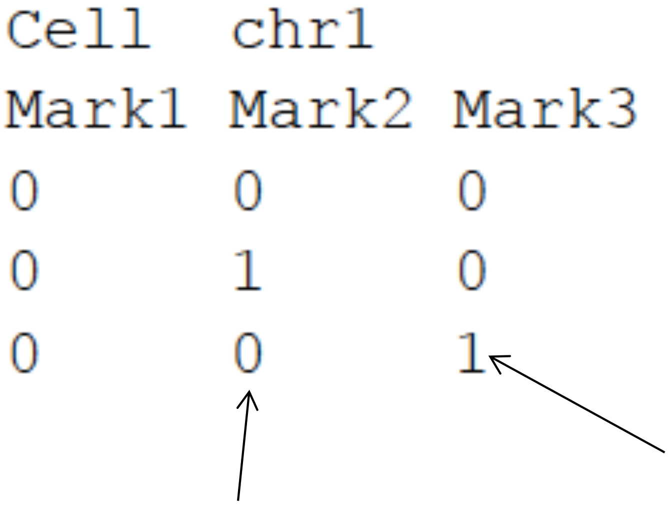
Fib 1b. Segmentation results from ChromHMM



Hoffman M M et al. Nucl. Acids Res. 2012;nar.gks1284

ChromHMM is a *multivariate* HMM

- ChromHMM uses a multivariate HMM that explicitly models the combination of marks

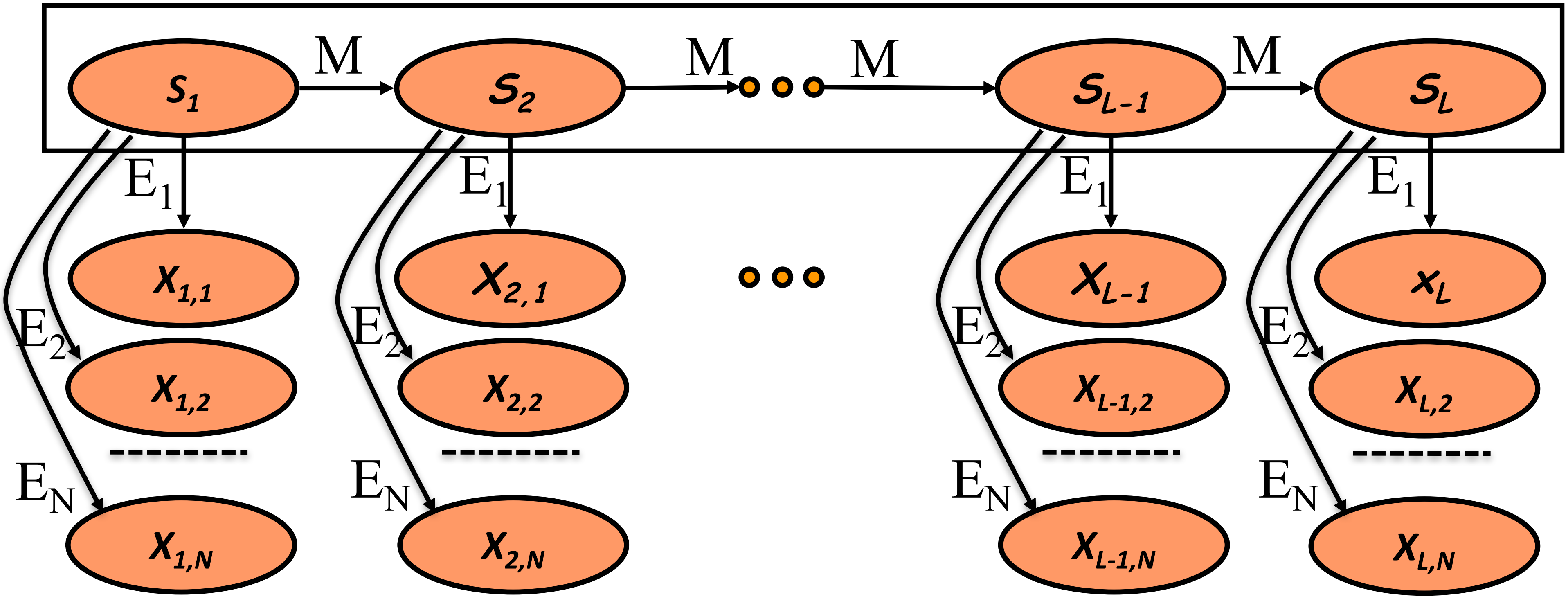


```
Cell   chr1
Mark1  Mark2  Mark3
0      0      0
0      1      0
0      0      1
```

absent

present

Multivariate HMM



Multivariate HMM (formal definition)

- A multivariate HMM θ has
  - **N** sets of observation symbols, each for one given observation sequence n (n=1, 2, …, N)
  - A set of hidden states
  - Transition probabilities a_{ij}, for any pair of hidden states i and j
  - Initial probabilities $B_j=a_{0j}$ for any hidden states j
  - **N** sets of emission probabilities $e^n{}_s(x_n)$ for the observation symbol being emitted in the *n*th observation sequence from the hidden state *s*.

Multivariate HMM

- Given N observation sequences of the same length L, $X=\{(x_{1,1}\ldots x_{1,L}), \ldots,(x_{N,1}\ldots x_{N,L})\}$ and the hidden state sequence $S=(s_1\ldots s_L)$, the full probability from the multivariate HMM is,

$$P(S,X\mid\theta)=\prod_{j=1}^{L}\left[a_{s_{j-1}s_j}\prod_{n=1}^{N}e_{s_j}\left(x_{n,j}\right)\right]$$

Let $e_{s_i}\left(x_{n,1},\ldots,x_{n,j}\right)=\prod_{n=1}^{N}e_{s_i}\left(x_{n,j}\right)$, the multivariate HMM can be reduced to conventional HMM, except the observation symbol becomes a vector $(x_{n,1}\ldots x_{n,j})$ at position j. The same algorithms for model inference (Viterbi and forward/backward) and learning can be used.