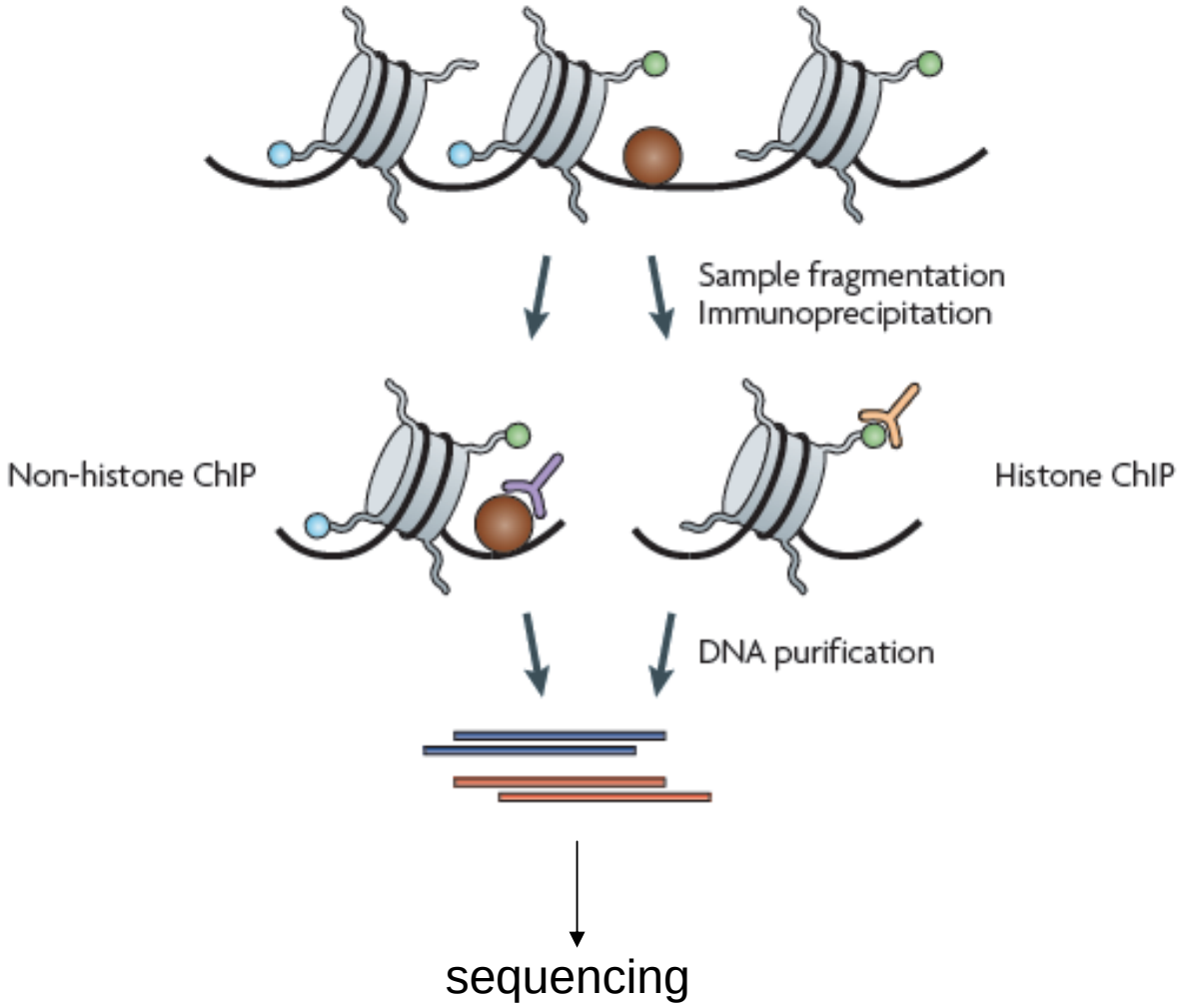
ChIP-seq

Chromatin immunoprecipitation followed by sequencing

Lecture outline

- Mapping short-read transcript sequences to a reference genome
- Finding ChIP-seq peaks using MACS
- Some results from the ENCODE Project



Adapted from Park, 2009

Steps in ChIP-seq experiment

- DNA crosslinked to proteins using formaldehyde
- Chromatin sheared through sonication or digested into small fragments (fragment size is important – see later)
- Antibody specific to the protein (or protein variant/modification) of interest is used to precipitate out the protein-bound DNA (antibodies attached to beads)
- Unlink the protein
- Purify and size-select the DNA
- Sequence (mostly using single-end Illumina sequencing)

Controls for ChIP experiments

- A peak of mapped reads may reflect local sequence specific biases, e.g. in fragmentation
- Important to compare numbers of mapped reads against a control
- Input DNA without IP or IP with no antibody or with a ‘mock’ non-specific antibody typically used
- A ‘peak’ with a greater proportion of mapped reads in immunoprecipitated sample compared to control indicates binding of the protein to DNA

Issues in ChIP-seq experiments

- GC content bias resulting from sonication
- Potentially high levels of PCR duplicates, particularly if sample not of high quality
- Experiments expensive – may be possible to validate the steps in the process prior to sequencing (e.g. generate smaller numbers of sequences or carry out quantitative PCR against known binding sites to confirm IP has worked correctly)

Steps in ChIP-seq analysis

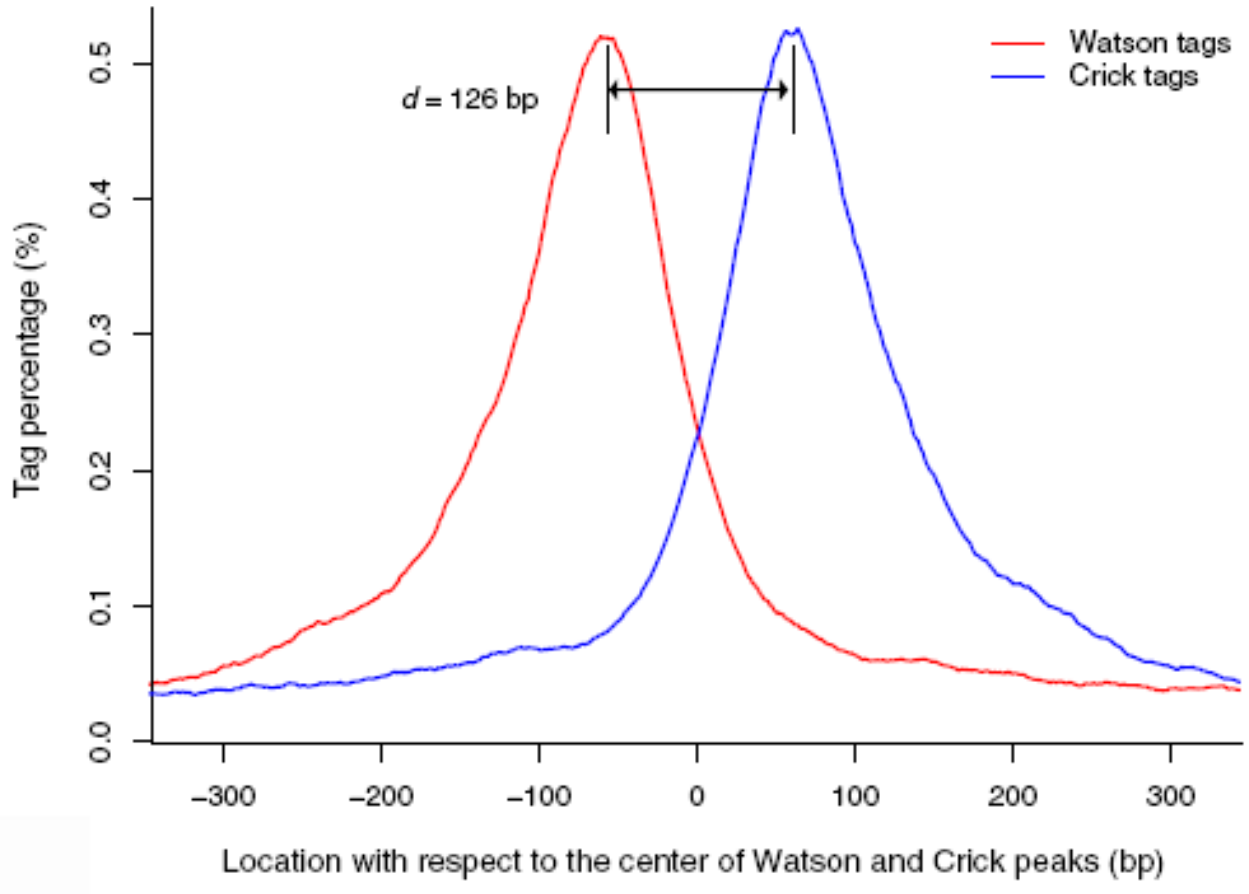
- Quality control
- Map short-reads to genome (without gaps)
- Quality control
- Identify peaks (‘peak calling’)
- Visualize results
- Bespoke analysis

Model-based Analysis of ChIP-Seq (MACS)

Without Control

- Set fragment size (*bandwidth*) & fold enrichment threshold (*mfold*)
- Slide windows of size 2 X *bandwidth* along genome
- Identify windows with *mfold* enrichment of probes compared to genomic average

Bimodal distribution of reads expected around binding sites



Zhang et al. 2008

Model-based Analysis of ChIP-Seq (MACS)

With Control/Input

- Use Poisson process to model number of reads in a window

- ‘Dynamic Poisson distribution’ i.e. rate parameter varies, reflecting differences in the input read densities across the genome

- λ_{local} = max (λ_{BG}, λ_{1k}, λ_{5k}, λ_{10k})

- Look for ‘peaks’ – windows with an excess of reads, given the expectation, based on λ_{local} ($p < 10^{-5}$)

MACS false discovery rate estimation

- Performs sample swap (swaps contol and ChIP samples)

- Estimates peaks of control over ChIP

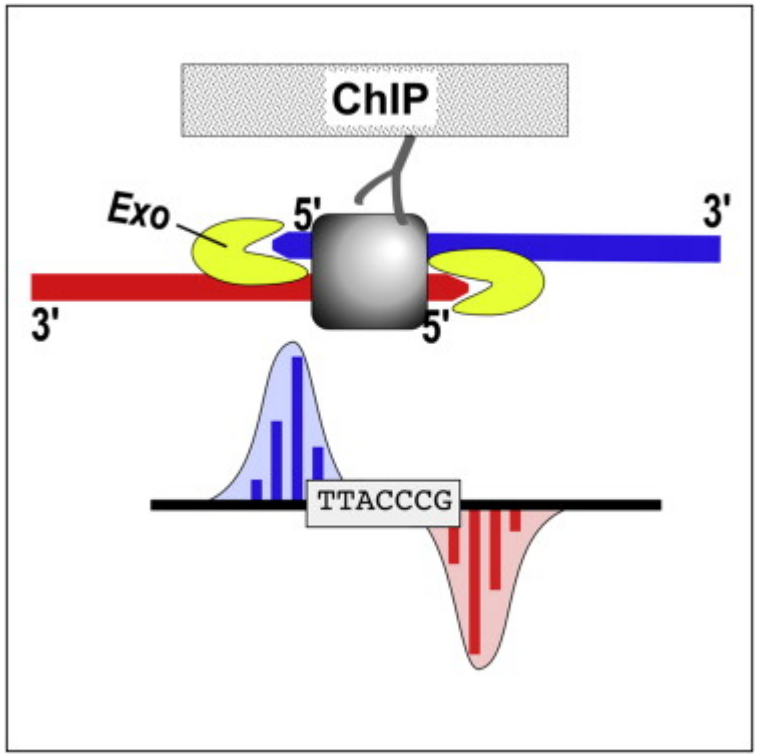
- False discovery rate estimated as number of control peaks, divided by number of ChIP peaks

MACS can also be used to identify differential binding, by treating the ChIP reads from one of the samples as the control.

Recent advance: ChIP-exo

- Developed by Rhee & Pugh, *Cell*, 2011
- Uses an exonuclease to trim DNA bound to the protein to a precise distance from the cross-linking site
- Can identify binding sites of proteins to base-pair resolution

ChIP-exo variant (nucleotide-level resolution)

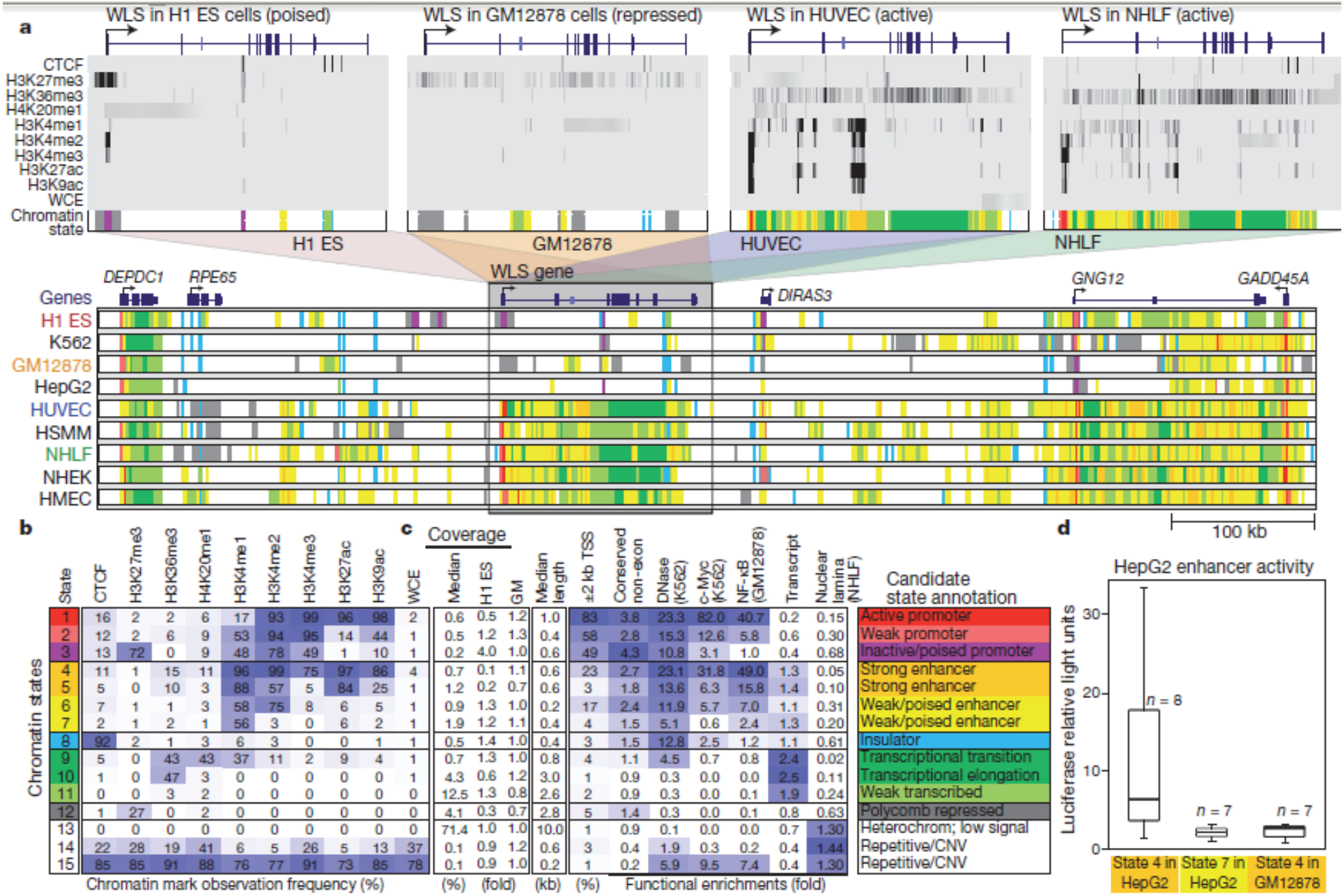


Source: Rhee & Pugh, *Cell,* 2011

The ENCODE project

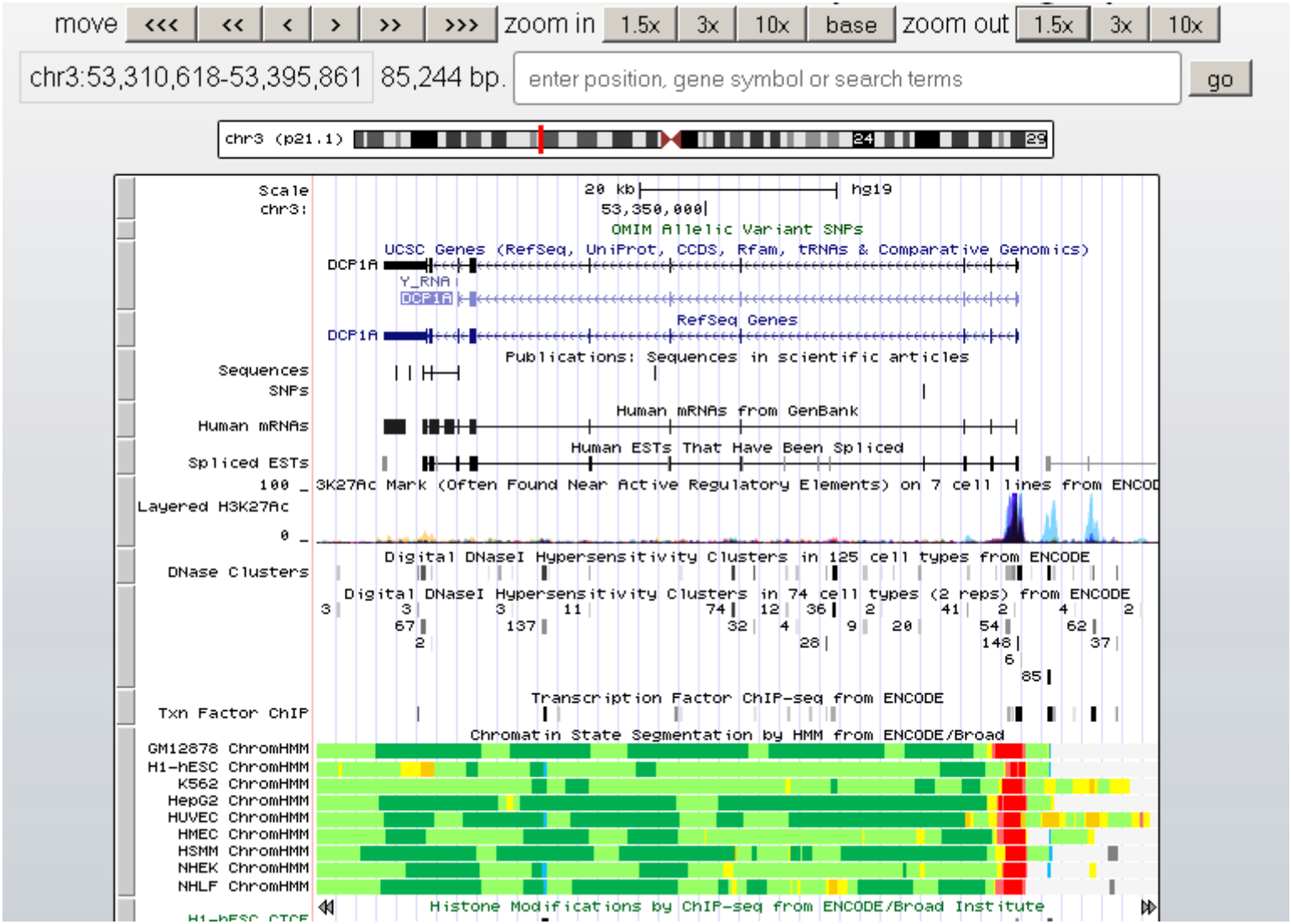
Ernst *et al. Nature* 2011, Mapping and analysis of chromatin state dynamics in nine human cell types

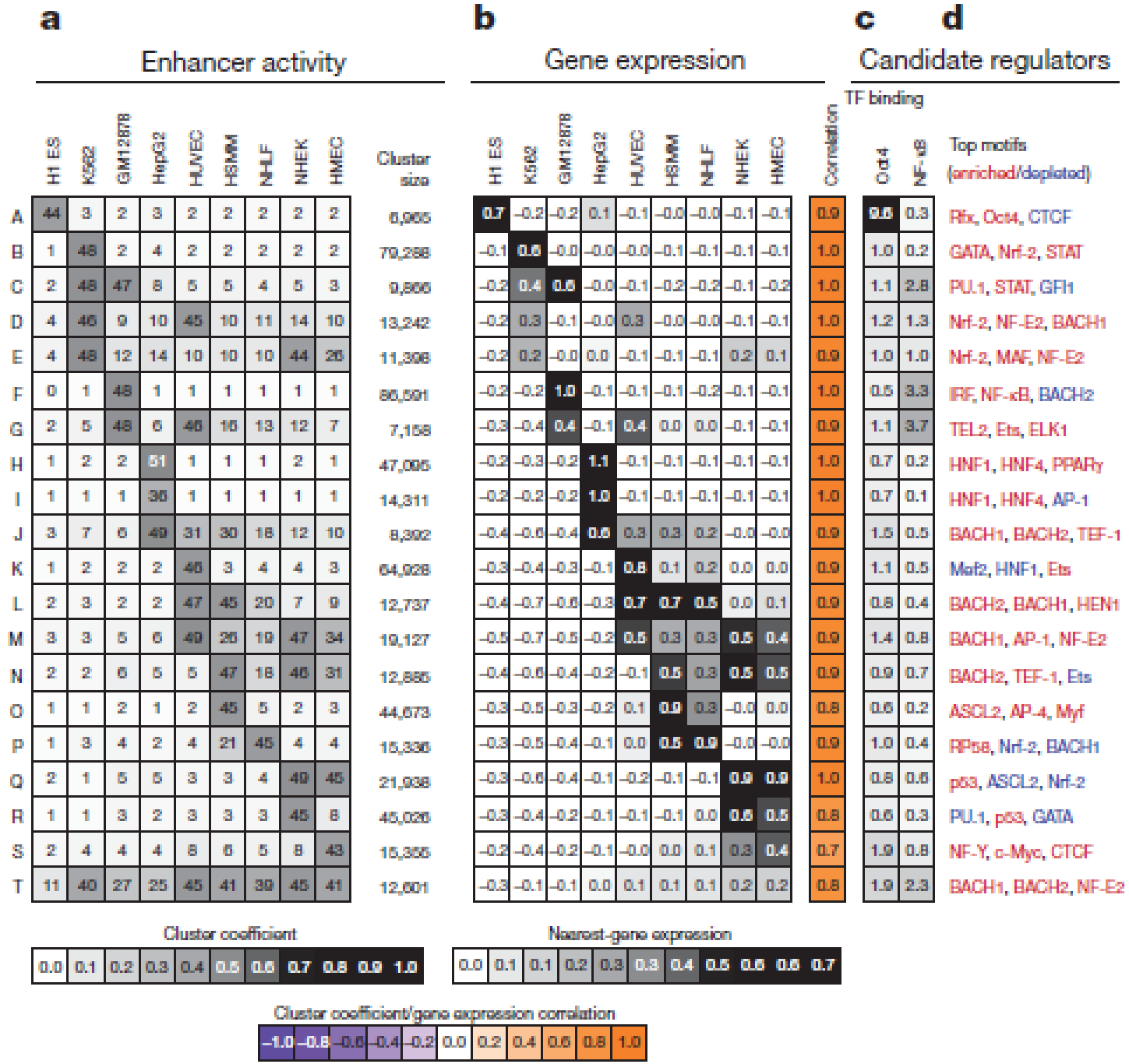
- Results of ChIP-seq profiling of 9 cell types
- Developed methods to learn chromatin state associated with distinct features using Hidden Markov Models (ChromHMM)
- Profiled all for 9 key chromatin marks (e.g. H3K4me3 associated with promoters; H3K4me1 associated with enhancers etc.)
- Partitioned genome into states using Hidden Markov Models



Kellis *et al., Nature*

ENCODE histone modification data via UCSC genome browser:





Kellis *et al., Nature*