# Epigenomics data resources and the genome browser

Oana Ursu

# Motivation

- 2001: First draft of the human genome
- ~20 000 genes covering 1-3% of the genome

First printout of the human genome

Image from http://en.wikipedia.org/wiki/Human_Genome_Project#mediaviewer/File:Wellcome_genome_bookcase.png

**What about the remaining 97-99% of the genome?**

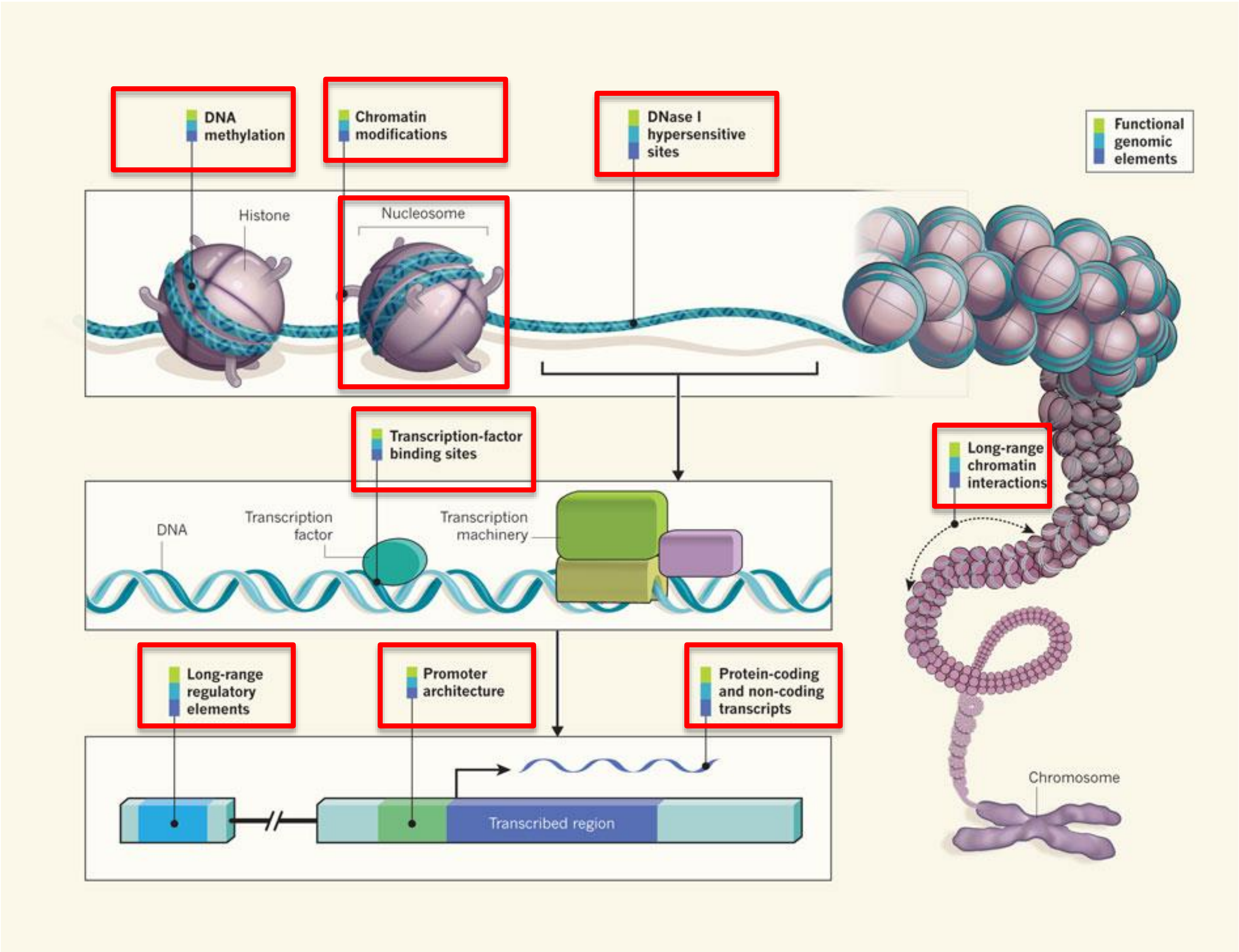
**ENCODE**
(Encyclopedia Of DNA Elements) set out to answer this question

# ENCODE: An Encyclopedia Of DNA Elements

- Goal: Complete catalog of all functional elements
  - Protein-coding genes: heavily studied, sequence-based, transcription
  - Non-coding DNA: much less studied, some motifs, diverse assays
- Dimensions for catalog completion
  - Genome-wide: systems-level view
  - Cell types: hundreds of human tissues and cell types
  - Dynamics: time, conditions, stimulation, environment, response

- Pilot phase 2003-: Small-scale targeted experiments in 1% of genome (30Mb)
  - Single gene, single pathway, few TFs, few cell-types, **tiling array-based**
- Scale-up 2007-: 100-fold increase in scale (3Gb), more assays, tech dev
  - Big change: RNA-seq, ChIP-seq, DNase-seq, **next-gen seq technologies**
  - Game changer: complete view, integration possible, networks and circuits
- Build-up 2012-: Further increases in all dimensions
  - Deeper sequencing, more assays, more conditions, more TFs.
  - More validation

# Diversity of assayed biochemical events


DNA methylation
Chromatin modifications
DNase I hypersensitive sites
Functional genomic elements
Histone
Nucleosome
Transcription-factor binding sites
Long-range chromatin interactions
DNA
Transcription factor
Transcription machinery
Long-range regulatory elements
Promoter architecture
Protein-coding and non-coding transcripts
Transcribed region
Chromosome

RNA-seq
CAGE-seq
Exon Arrays

TF ChIP-seq

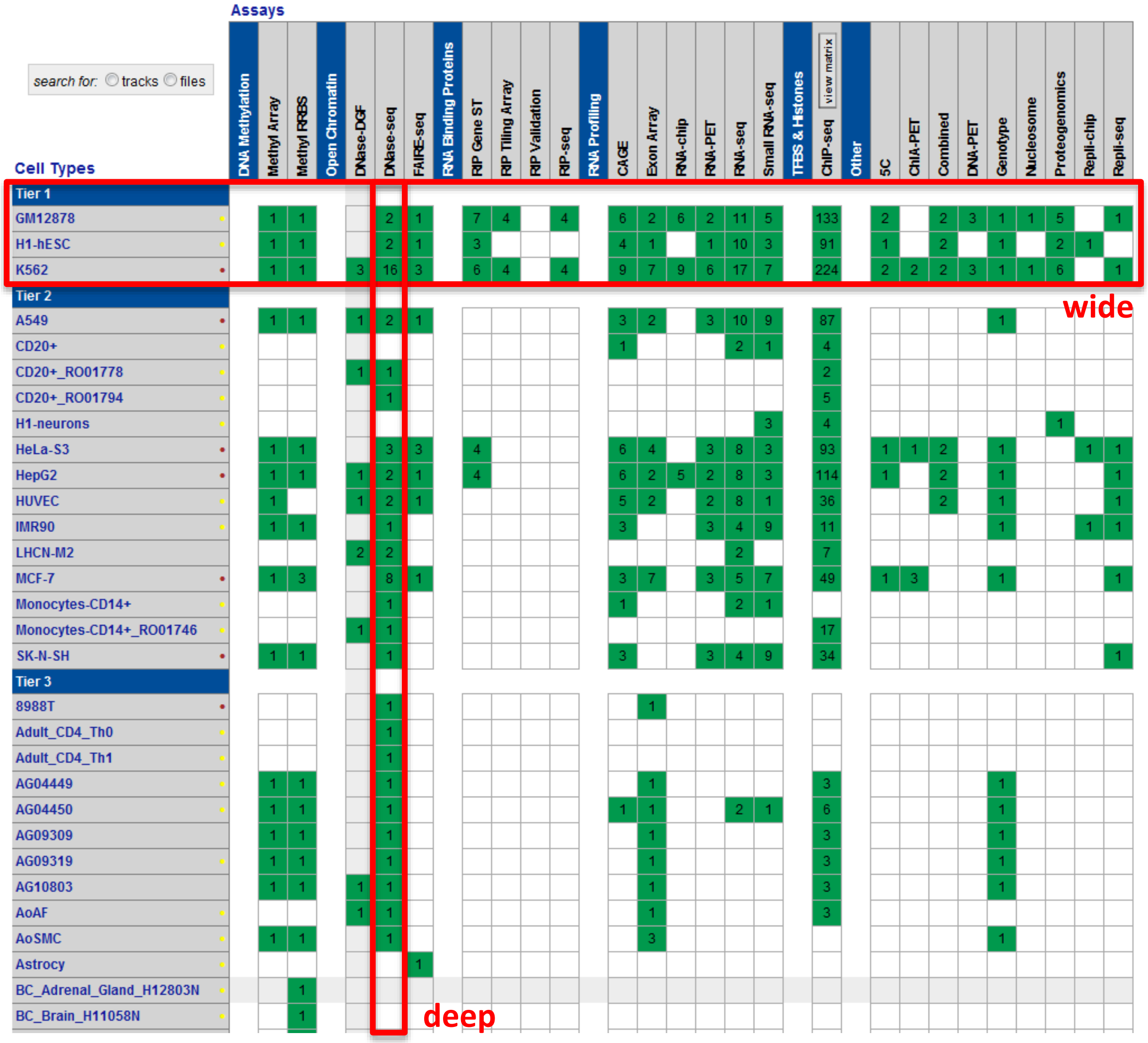
Chromatin ChIP-seq

DNase-seq
FAIRE-seq

Methyl RRBS
Methyl Arrays

3C, 4C, 5C
ChIA-PET
HiC

# ENCODE data at a glance



| Cell Types | DNA Methylation | | Open Chromatin | | | RNA Binding Proteins | | | | RNA Profiling | | | | | | TFBS & Histones | Other | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Methyl Array | Methyl RRBS | DNase-DGF | DNase-seq | FAIRE-seq | RIP Gene ST | RIP Tiling Array | RIP Validation | RIP-seq | CAGE | Exon Array | RNA-chip | RNA-PET | RNA-seq | Small RNA-seq | ChIP-seq | 5C | ChIA-PET | Combined | DNA-PET | Genotype | Nucleosome | Proteogenomics | Repli-chip | Repli-seq |
| Tier 1 | | | | | | | | | | | | | | | | | | | | | | | | | |
| GM12878 | 1 | 1 | | 2 | 1 | 7 | 4 | | 4 | 6 | 2 | 6 | 2 | 11 | 5 | 133 | 2 | | 2 | 3 | 1 | 1 | 5 | | 1 |
| H1-hESC | 1 | 1 | | 2 | 1 | 3 | | | | 4 | 1 | | 1 | 10 | 3 | 91 | 1 | | 2 | | 1 | | 2 | 1 | |
| K562 | 1 | 1 | 3 | 16 | 3 | 6 | 4 | | 4 | 9 | 7 | 9 | 6 | 17 | 7 | 224 | 2 | 2 | 2 | 3 | 1 | 1 | 6 | | 1 |
| Tier 2 | | | | | | | | | | | | | | | | | | | | | | | | | |
| A549 | 1 | 1 | 1 | 2 | 1 | | | | | 3 | 2 | | 3 | 10 | 9 | 87 | | | | | 1 | | | | |
| CD20+ | | | | | | | | | | 1 | | | | 2 | 1 | 4 | | | | | | | | | |
| CD20+_RO01778 | | | 1 | 1 | | | | | | | | | | | | 2 | | | | | | | | | |
| CD20+_RO01794 | | | | 1 | | | | | | | | | | | | 5 | | | | | | | | | |
| H1-neurons | | | | | | | | | | | | | | | 3 | 4 | | | | | | | 1 | | |
| HeLa-S3 | 1 | 1 | | 3 | 3 | 4 | | | | 6 | 4 | | 3 | 8 | 3 | 93 | 1 | 1 | 2 | | 1 | | | 1 | 1 |
| HepG2 | 1 | 1 | 1 | 2 | 1 | 4 | | | | 6 | 2 | 5 | 2 | 8 | 3 | 114 | 1 | | 2 | | 1 | | | | 1 |
| HUVEC | 1 | | 1 | 2 | 1 | | | | | 5 | 2 | | 2 | 8 | 1 | 36 | | | 2 | | 1 | | | | 1 |
| IMR90 | 1 | 1 | | 1 | | | | | | 3 | | | 3 | 4 | 9 | 11 | | | | | 1 | | | 1 | 1 |
| LHCN-M2 | | | 2 | 2 | | | | | | | | | | 2 | | 7 | | | | | | | | | |
| MCF-7 | 1 | 3 | | 8 | 1 | | | | | 3 | 7 | | 3 | 5 | 7 | 49 | 1 | 3 | | | 1 | | | | 1 |
| Monocytes-CD14+ | | | | 1 | | | | | | 1 | | | | 2 | 1 | | | | | | | | | | |
| Monocytes-CD14+_RO01746 | | | 1 | 1 | | | | | | | | | | | | 17 | | | | | | | | | |
| SK-N-SH | 1 | 1 | | 1 | | | | | | 3 | | | 3 | 4 | 9 | 34 | | | | | | | | | 1 |
| Tier 3 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8988T | | | | 1 | | | | | | | 1 | | | | | | | | | | | | | | |
| Adult_CD4_Th0 | | | | 1 | | | | | | | | | | | | | | | | | | | | | |
| Adult_CD4_Th1 | | | | 1 | | | | | | | | | | | | | | | | | | | | | |
| AG04449 | 1 | 1 | | 1 | | | | | | | 1 | | | | | 3 | | | | | 1 | | | | |
| AG04450 | 1 | 1 | | 1 | | | | | | 1 | 1 | | | 2 | 1 | 6 | | | | | 1 | | | | |
| AG09309 | 1 | 1 | | 1 | | | | | | | 1 | | | | | 3 | | | | | 1 | | | | |
| AG09319 | 1 | 1 | | 1 | | | | | | | 1 | | | | | 3 | | | | | 1 | | | | |
| AG10803 | 1 | 1 | 1 | 1 | | | | | | | 1 | | | | | 3 | | | | | 1 | | | | |
| AoAF | | | 1 | 1 | | | | | | | 1 | | | | | 3 | | | | | | | | | |
| AoSMC | 1 | 1 | | 1 | | | | | | | 3 | | | | | | | | | | 1 | | | | |
| Astrocy | | | | | 1 | | | | | | | | | | | | | | | | | | | | |
| BC_Adrenal_Gland_H12803N | | 1 | | | | | | | | | | | | | | | | | | | | | | | |
| BC_Brain_H11058N | | 1 | | | | | | | | | | | | | | | | | | | | | | | |

Tier 1

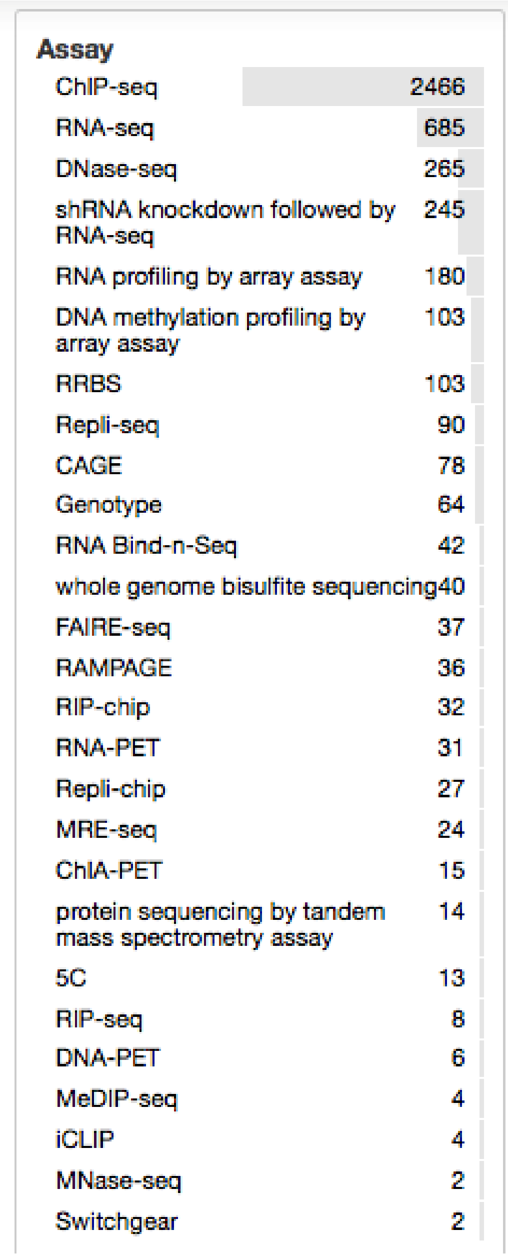
wide

deep

336 cell-types!

http://genome.ucsc.edu/ENCODE/cellTypes.html

http://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataSummaryHuman.html



| Assay | |
|---|---|
| ChIP-seq | 2466 |
| RNA-seq | 685 |
| DNase-seq | 265 |
| shRNA knockdown followed by RNA-seq | 245 |
| RNA profiling by array assay | 180 |
| DNA methylation profiling by array assay | 103 |
| RRBS | 103 |
| Repli-seq | 90 |
| CAGE | 78 |
| Genotype | 64 |
| RNA Bind-n-Seq | 42 |
| whole genome bisulfite sequencing | 40 |
| FAIRE-seq | 37 |
| RAMPAGE | 36 |
| RIP-chip | 32 |
| RNA-PET | 31 |
| Repli-chip | 27 |
| MRE-seq | 24 |
| ChIA-PET | 15 |
| protein sequencing by tandem mass spectrometry assay | 14 |
| 5C | 13 |
| RIP-seq | 8 |
| DNA-PET | 6 |
| MeDIP-seq | 4 |
| iCLIP | 4 |
| MNase-seq | 2 |
| Switchgear | 2 |

# Outline

- **The ENCODE project: experiments, data, findings**
  - Genes and transcripts: RNAseq
  - Open chromatin: DNaseI-seq
  - DNA-binding proteins: ChIP-seq
  - Chromatin state: Histone ChIP-seq
  - Genome 3D: 3C
- **The genome browser**

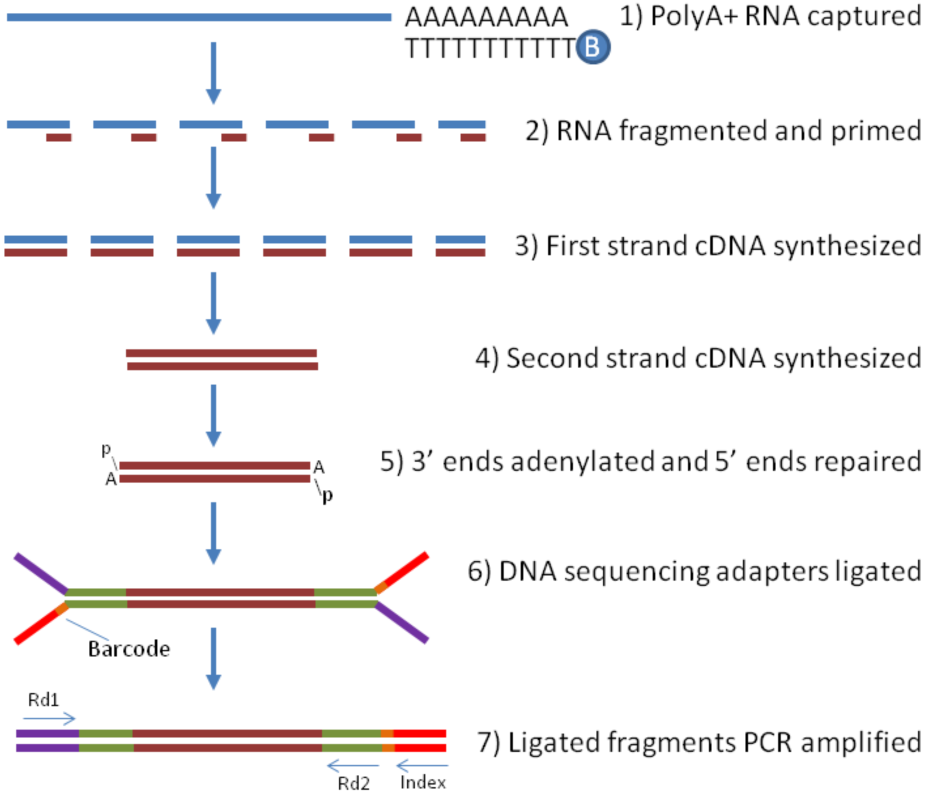
# Outline

- The ENCODE project: experiments, data, findings
  - **Genes and transcripts: RNAseq**
  - Open chromatin: DNaseI-seq
  - DNA-binding proteins: ChIP-seq
  - Chromatin state: Histone ChIP-seq
  - Genome 3D: 3C
- The genome browser

# Studying genes and transcripts: RNAseq

## RNA-seq workflow

Remove rRNA
(>97% of your sample is rRNA!)



1) PolyA+ RNA captured

2) RNA fragmented and primed

3) First strand cDNA synthesized

4) Second strand cDNA synthesized

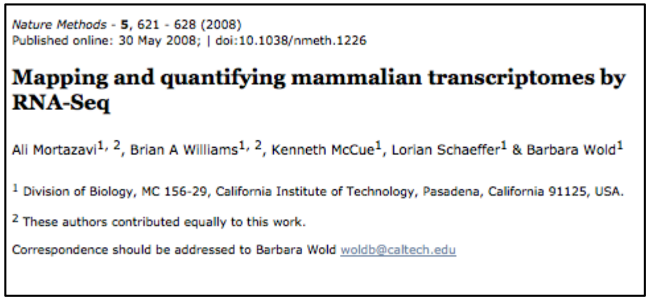
5) 3’ ends adenylated and 5’ ends repaired

6) DNA sequencing adapters ligated

7) Ligated fragments PCR amplified

8) Sequencing
9) Map reads to transcriptome

Figure adapted from Corney, 2014



*Nature Methods* - **5**, 621 - 628 (2008)
Published online: 30 May 2008; | doi:10.1038/nmeth.1226

**Mapping and quantifying mammalian transcriptomes by RNA-Seq**

Ali Mortazavi[1, 2], Brian A Williams[1, 2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]

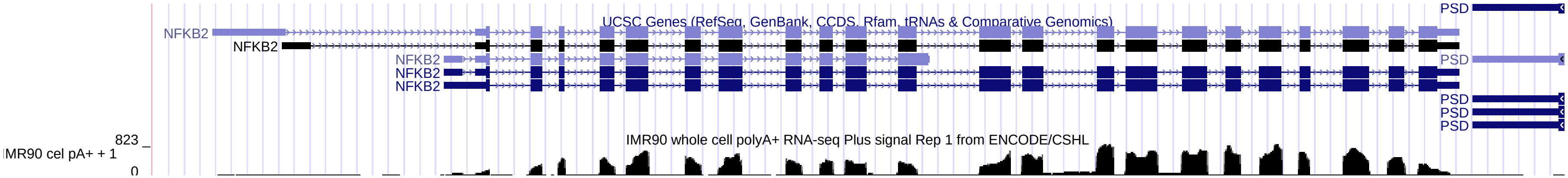
[1] Division of Biology, MC 156-29, California Institute of Technology, Pasadena, California 91125, USA.

[2] These authors contributed equally to this work.

Correspondence should be addressed to Barbara Wold woldb@caltech.edu

# Studying genes and transcripts: RNAseq

## Example


UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
NFKB2
NFKB2
NFKB2
NFKB2
NFKB2
PSD
PSD
PSD
PSD
PSD
823 _
IMR90 cel pA+ + 1
0
IMR90 whole cell polyA+ RNA-seq Plus signal Rep 1 from ENCODE/CSHL

## Goals

- Transcriptome assembly
- Gene expression quantification
- Splicing

# Studying genes and transcripts: RNAseq

## Main findings

- Pervasive transcription
  - “62% of genomic bases are reproducibly represented in sequenced long (>200 nucleotides) RNA molecules or GENCODE exons” (ENCODE, 2012)

Proportion of genomic bases included in a primary transcript, by number of technologies supporting the transcribed base

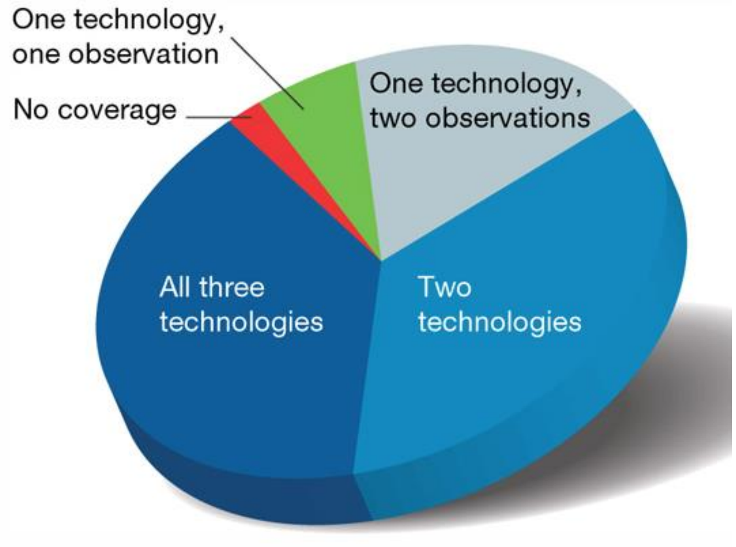


Figure from ENCODE 1 paper (1% of the genome)

ENCODE, 2007, Nature
ENCODE, 2012, Nature

# Studying genes and transcripts: RNAseq

## Main findings

- Pervasive transcription
  - “62% of genomic bases are reproducibly represented in sequenced long (>200 nucleotides) RNA molecules or GENCODE exons” (ENCODE, 2012)

- Many flavors of RNAs
  - ~8000 small RNAs, ~9000 lncRNAs
  - lncRNAs more cell-type restricted, lower expression levels (compared to protein-coding genes)

## Outstanding questions

- lncRNA functions
- Gene regulation through the act of transcription, not the transcript

ENCODE, 2007, Nature
ENCODE, 2012, Nature

# Studying genes and transcripts: RNAseq

**RNAseq experiment variants**

- RNA selection: polyA selection
- Location: cytoplasmic, nuclear
- Read length: short reads, long-read RNAseq (PacBio, Moleculo)
- More: CAGE-seq (TSS), Ribo-seq (translated transcripts)

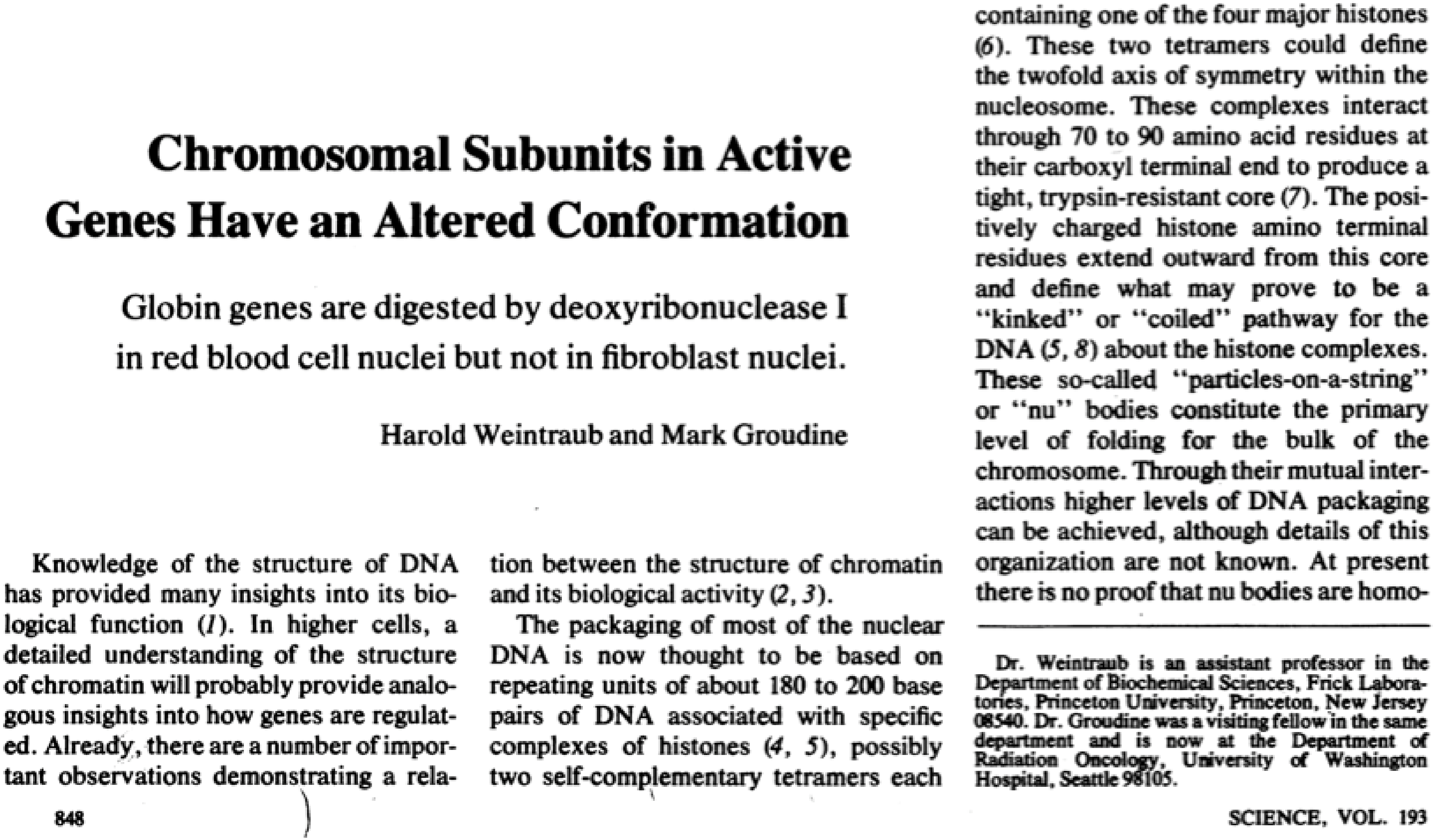
# Outline

- The ENCODE project: experiments, data, findings
  - **Genes and transcripts: RNAseq**
  - Open chromatin: DNaseI-seq
  - DNA-binding proteins: ChIP-seq
  - Chromatin state: Histone ChIP-seq
  - Genome 3D: 3C
- The genome browser

# Outline

- **The ENCODE project: experiments, data, findings**
  - **Genes and transcripts: RNAseq**
  - **Open chromatin: DNaseI-seq**
  - DNA-binding proteins: ChIP-seq
  - Chromatin state: Histone ChIP-seq
  - Genome 3D: 3C
- **The genome browser**

# Studying open chromatin: DNaseI-seq



## Chromosomal Subunits in Active Genes Have an Altered Conformation

Globin genes are digested by deoxyribonuclease I in red blood cell nuclei but not in fibroblast nuclei.

Harold Weintraub and Mark Groudine

Knowledge of the structure of DNA has provided many insights into its biological function (1). In higher cells, a detailed understanding of the structure of chromatin will probably provide analogous insights into how genes are regulated. Already, there are a number of important observations demonstrating a relation between the structure of chromatin and its biological activity (2, 3).

The packaging of most of the nuclear DNA is now thought to be based on repeating units of about 180 to 200 base pairs of DNA associated with specific complexes of histones (4, 5), possibly two self-complementary tetramers each containing one of the four major histones (6). These two tetramers could define the twofold axis of symmetry within the nucleosome. These complexes interact through 70 to 90 amino acid residues at their carboxyl terminal end to produce a tight, trypsin-resistant core (7). The positively charged histone amino terminal residues extend outward from this core and define what may prove to be a "kinked" or "coiled" pathway for the DNA (5, 8) about the histone complexes. These so-called "particles-on-a-string" or "nu" bodies constitute the primary level of folding for the bulk of the chromosome. Through their mutual interactions higher levels of DNA packaging can be achieved, although details of this organization are not known. At present there is no proof that nu bodies are homo-

Dr. Weintraub is an assistant professor in the Department of Biochemical Sciences, Frick Laboratories, Princeton University, Princeton, New Jersey 08540. Dr. Groudine was a visiting fellow in the same department and is now at the Department of Radiation Oncology, University of Washington Hospital, Seattle 98105.

# Studying open chromatin: DNaseI-seq

## DNaseI-seq workflow

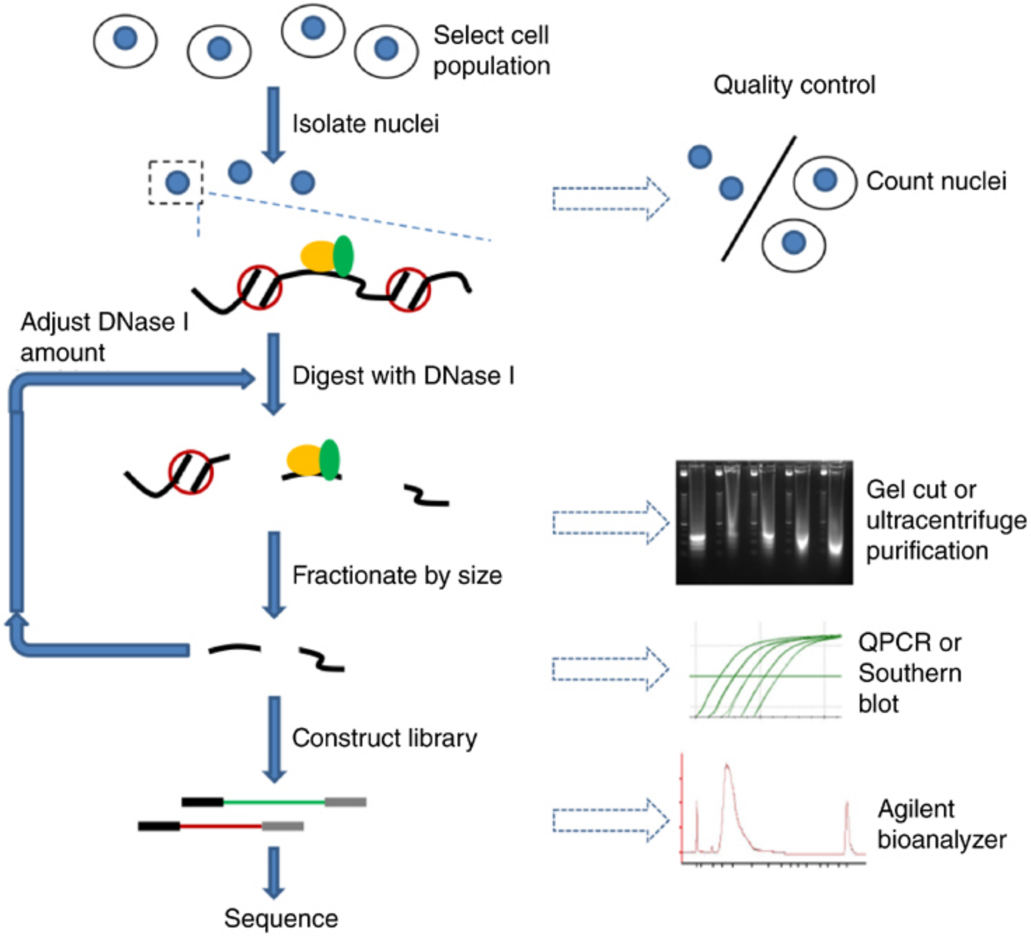
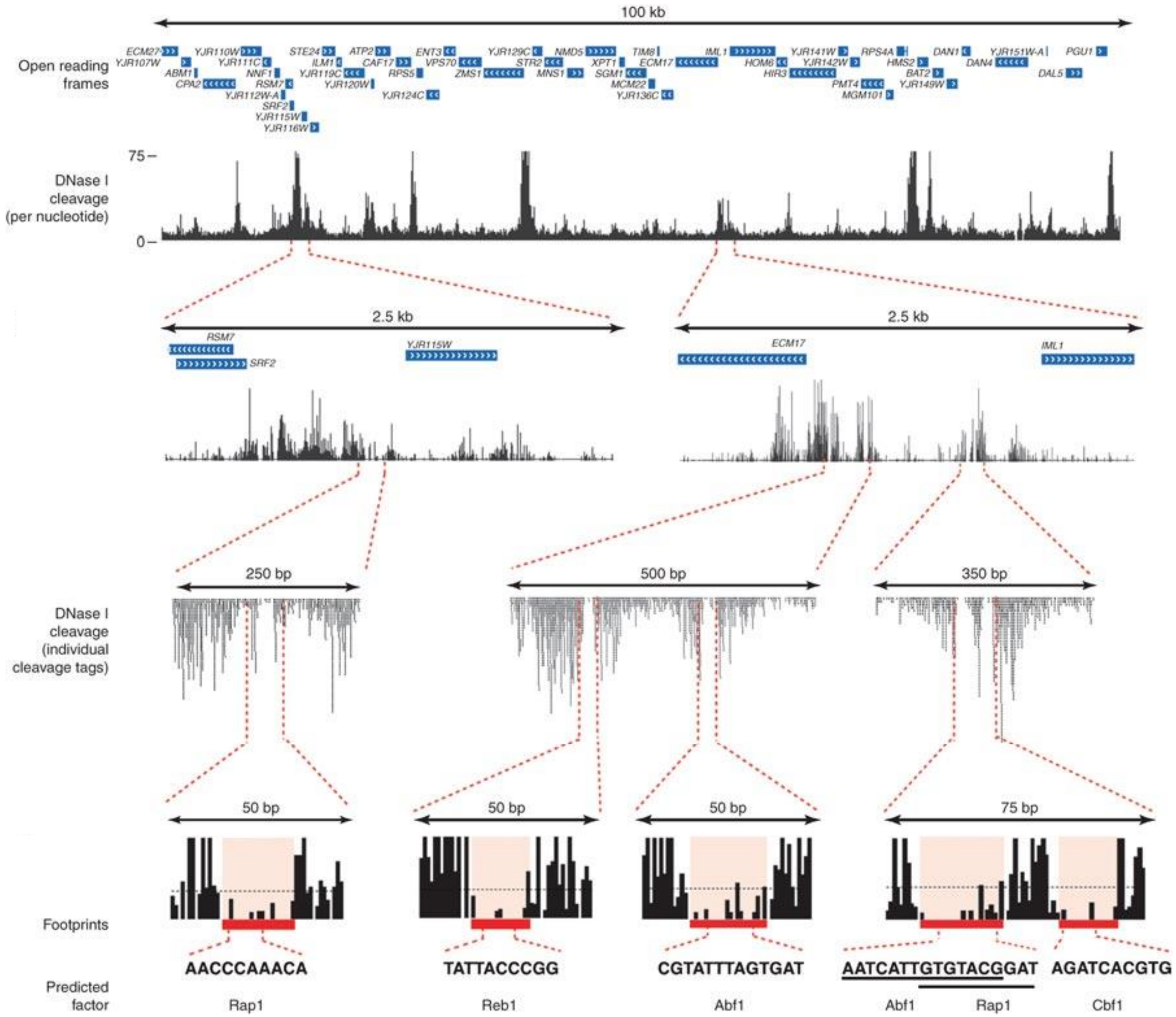


Figure from Zeng et al., Nature, 2012

This is hard!

# Studying open chromatin: DNaseI-seq



## Goals

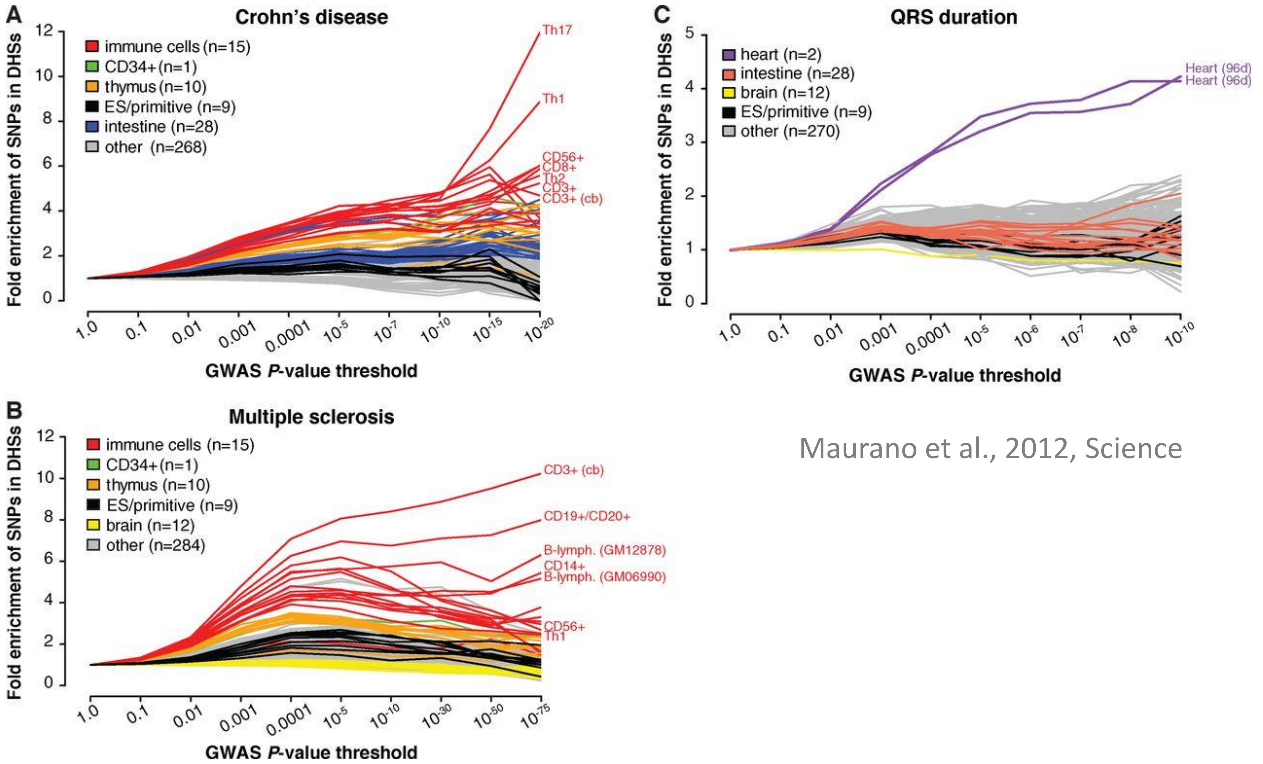
- Open chromatin
- TF footprinting

Figure from Hesselberth et al., Nature, 2009

# Studying open chromatin: DNaseI-seq

## Main findings

- DNaseI-seq sites at TSS, at enhancers, at protein-bound regions
- DNaseI-seq very cell-type specific (modules)
- Using DNaseI sites, can match disease with most likely affected cell type (because disease mutations fall in cell-type specific DNaseI sites)



Maurano et al., 2012, Science

# Outline

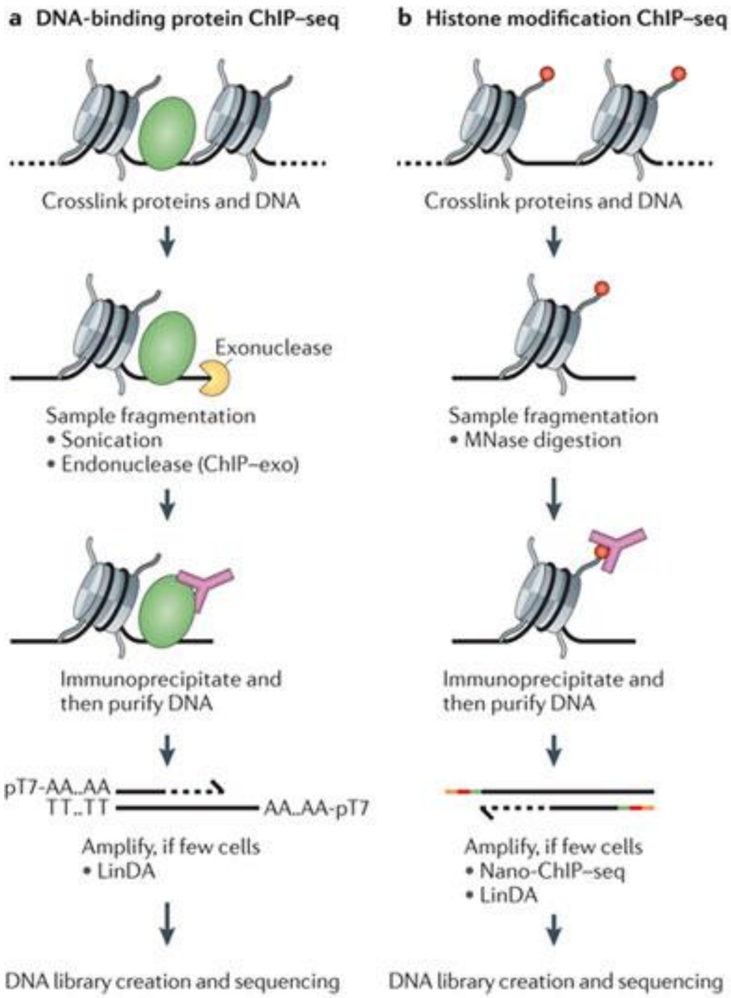
- The ENCODE project: experiments, data, findings
  - **Genes and transcripts: RNAseq**
  - **Open chromatin: DNaseI-seq**
  - DNA-binding proteins: ChIP-seq
  - Chromatin state: Histone ChIP-seq
  - Genome 3D: 3C
- The genome browser

# Outline

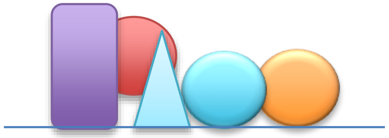
- **The ENCODE project: experiments, data, findings**
  - **Genes and transcripts: RNAseq**
  - **Open chromatin: DNaseI-seq**
  - **DNA-binding proteins: ChIP-seq**
  - Chromatin state: Histone ChIP-seq
  - Genome 3D: 3C
- **The genome browser**

# Studying DNA-binding proteins: ChIP-seq

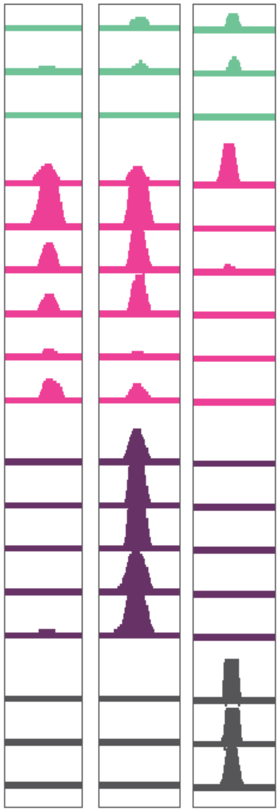
## ChIP-seq workflow


a DNA-binding protein ChIP–seq
b Histone modification ChIP–seq
Crosslink proteins and DNA
Crosslink proteins and DNA
Exonuclease
Sample fragmentation
• Sonication
• Endonuclease (ChIP–exo)
Sample fragmentation
• MNase digestion
Immunoprecipitate and then purify DNA
Immunoprecipitate and then purify DNA
pT7-AA..AA
TT..TT
AA..AA-pT7
Amplify, if few cells
• LinDA
Amplify, if few cells
• Nano-ChIP–seq
• LinDA
DNA library creation and sequencing
DNA library creation and sequencing

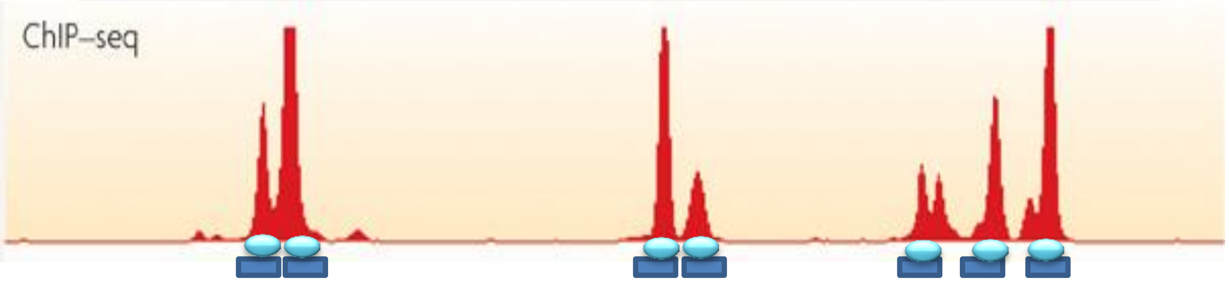
# Studying DNA-binding proteins: ChIP-seq



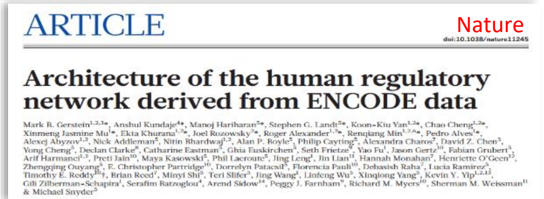
**100s of binding maps** of different regulatory proteins!



Genome-wide protein-DNA binding maps
**CONTINUOUS**



Signal Peaks (potential binding sites)
**DISCRETE**



ARTICLE

Nature

doi:10.1038/nature11245

Architecture of the human regulatory network derived from ENCODE data

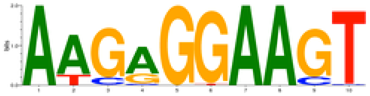
Mark B. Gerstein[1,2,3*], Anshul Kundaje[4*], Manoj Hariharan[5*], Stephen G. Landt[5*], Koon-Kiu Yan[1,2*], Chao Cheng[1,2*], Xinmeng Jasmine Mu[1*], Ekta Khurana[1,2*], Joel Rozowsky[2*], Roger Alexander[1,2*], Renqiang Min[1,2,6*], Pedro Alves[1*], Alexej Abyzov[1,2], Nick Addleman[5], Nitin Bhardwaj[1,2], Alan P. Boyle[5], Philip Cayting[5], Alexandra Charos[7], David Z. Chen[3], Yong Cheng[5], Declan Clarke[8], Catharine Eastman[5], Ghia Euskirchen[5], Seth Frietze[9], Yao Fu[1], Jason Gertz[10], Fabian Grubert[5], Arif Harmanci[1,2], Preti Jain[10], Maya Kasowski[5], Phil Lacroute[5], Jing Leng[1], Jin Lian[11], Hannah Monahan[7], Henriette O'Geen[12], Zhengqing Ouyang[5], E. Christopher Partridge[10], Dorrelyn Patacsil[5], Florencia Pauli[10], Debasish Raha[7], Lucia Ramirez[5], Timothy E. Reddy[10]†, Brian Reed[7], Minyi Shi[5], Teri Slifer[5], Jing Wang[1], Linfeng Wu[5], Xinqiong Yang[5], Kevin Y. Yip[1,2,13], Gili Zilberman-Schapira[1], Serafim Batzoglou[4], Arend Sidow[14], Peggy J. Farnham[9], Richard M. Myers[10], Sherman M. Weissman[11] & Michael Snyder[5]

# Studying DNA-binding proteins: ChIP-seq

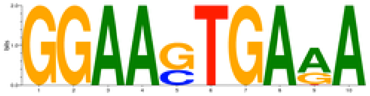
## Main findings

- High quality TF binding motifs

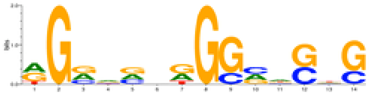
SPI1_disc1:



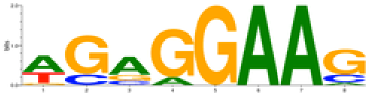
SPI1_disc2:



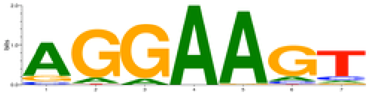
SPI1_disc3:



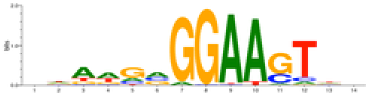
SPI1_known1:



SPI1_known2:



SPI1_known3:



Kheradpour et al., 2012, Genome Research

# Outline

- The ENCODE project: experiments, data, findings
  - **Genes and transcripts: RNAseq**
  - **Open chromatin: DNaseI-seq**
  - **DNA-binding proteins: ChIP-seq**
  - Chromatin state: Histone ChIP-seq
  - Genome 3D: 3C
- The genome browser

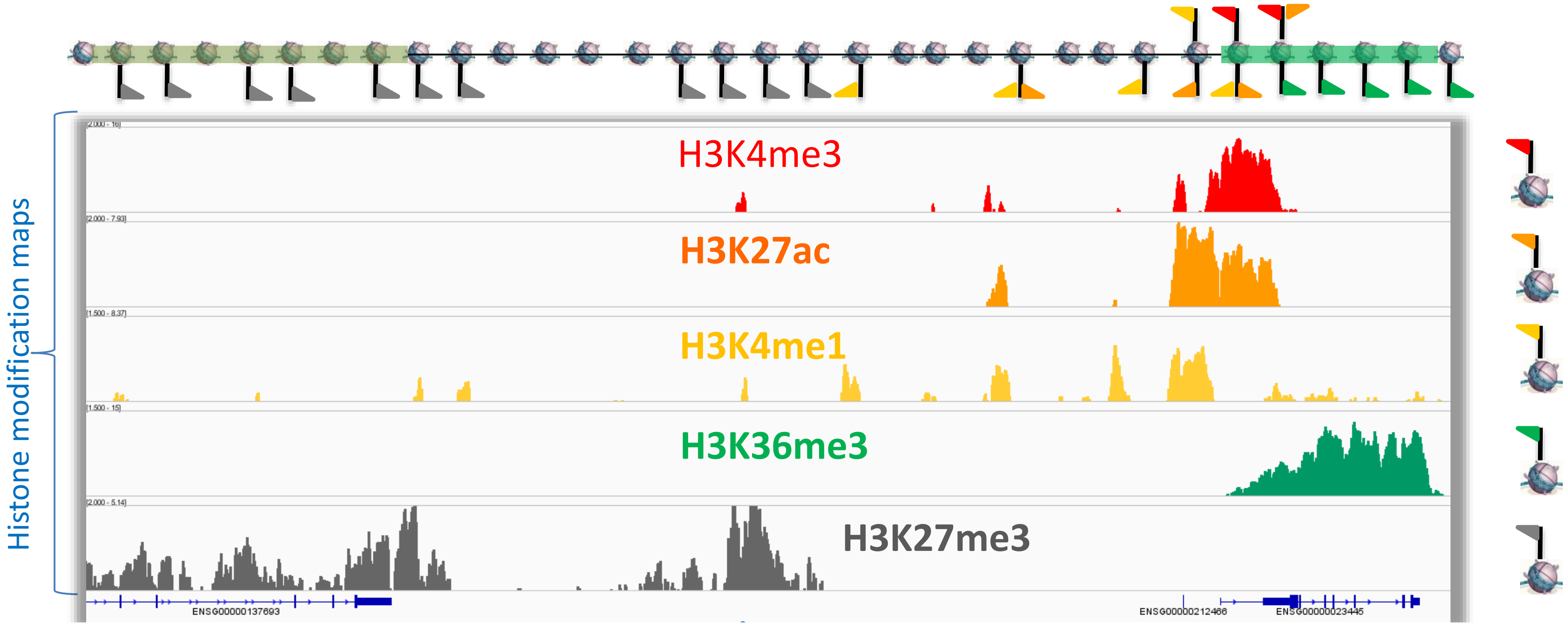
# Outline

- The ENCODE project: experiments, data, findings
  - **Genes and transcripts: RNAseq**
  - **Open chromatin: DNaseI-seq**
  - **DNA-binding proteins: ChIP-seq**
  - **Chromatin state: Histone ChIP-seq**
  - Genome 3D: 3C
- The genome browser

# Studying chromatin state: ChromHMM

## ChromHMM workflow

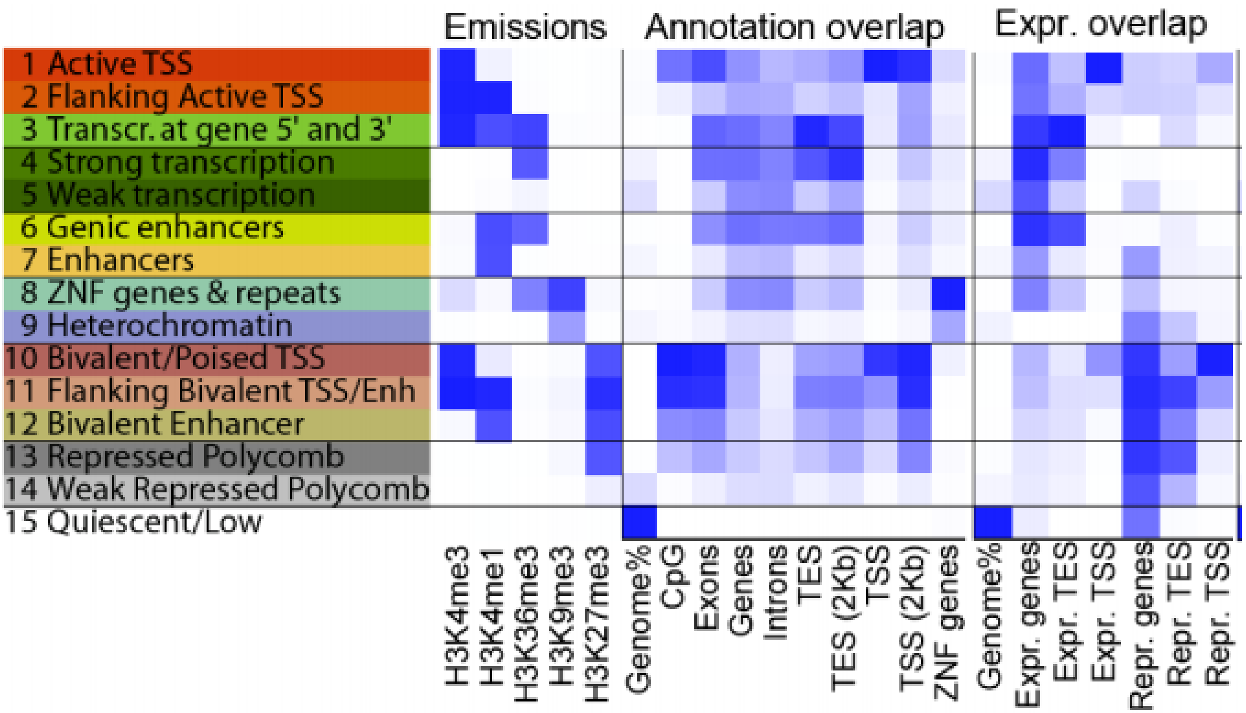
- Find functional elements from histone marks using Hidden Markov Models



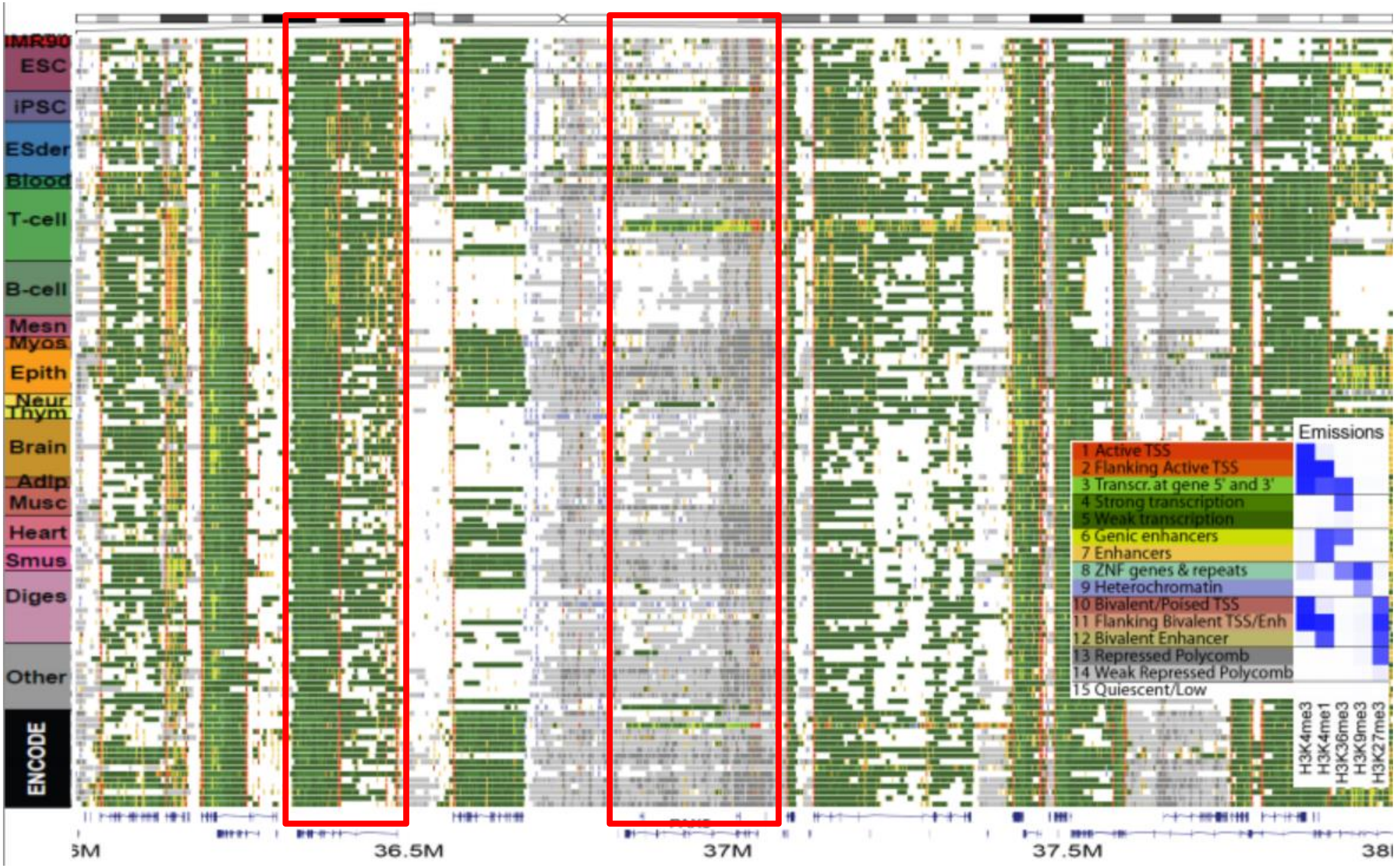
# Studying chromatin state: ChromHMM

## ChromHMM workflow

- Find functional elements from histone marks using Hidden Markov Models
- Chromatin states annotated using known genomic features



# The dynamic chromatin state



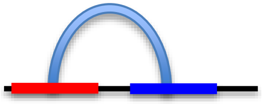
# Outline

- The ENCODE project: experiments, data, findings
  - **Genes and transcripts: RNAseq**
  - **Open chromatin: DNaseI-seq**
  - **DNA-binding proteins: ChIP-seq**
  - **Chromatin state: Histone ChIP-seq**
  - Genome 3D: 3C
- The genome browser

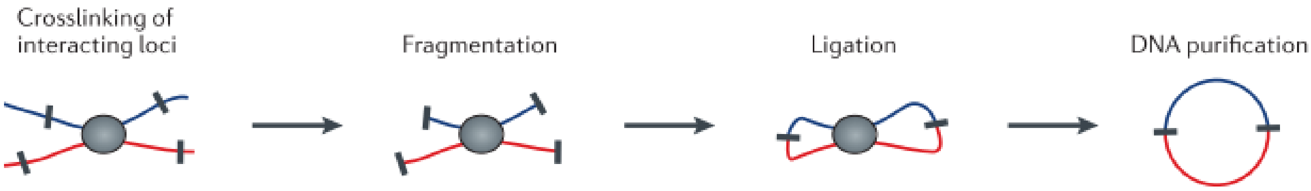
# Outline

- The ENCODE project: experiments, data, findings
  - Genes and transcripts: RNAseq
  - Open chromatin: DNaseI-seq
  - DNA-binding proteins: ChIP-seq
  - Chromatin state: Histone ChIP-seq
  - Genome 3D: 3C
- The genome browser

# Studying the genome 3D: 3C



New!
DNaseHiC, CaptureC



Dekker et al. 2013

# Studying the genome 3D: 3C

## Main findings

Figure adapted from Lieberman-Aiden et al. 2009 and Noordermeer et al. 2013

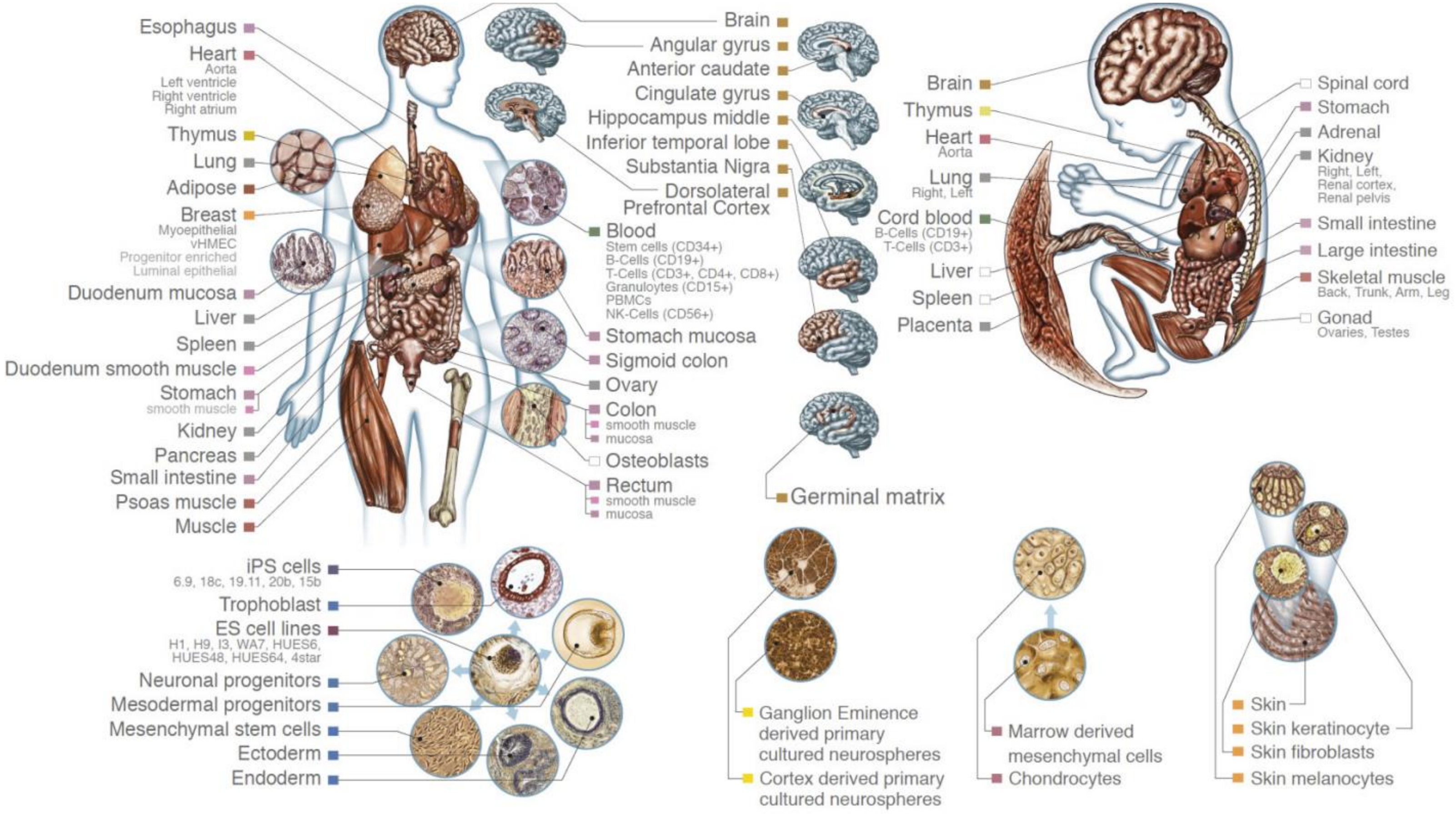
# Beyond ENCODE

**> 150 Cell-Types/Tissues**

- 6 histone marks (Histone ChIP-seq)
- Open chromatin (DNase-seq)
- DNA methylation (WGBS, RRBS)
- Gene expression (RNA-seq)

## Roadmap Epigenomics Project

- Primary tissues



http://roadmapepigenomics.org

# Beyond ENCODE

**Roadmap Epigenomics Project** http://roadmapepigenomics.org

- Primary tissues: chromatin state, open chromatin, expression

**BLUEPRINT**

**GTEx (Genotype-Tissue Expression)** **https://www.gtexportal.org/home/**

- Genetic variation