An introduction to ChIP-seq & ATAC-seq analysis

Carl Herrmann / Ashwini Sharma

Health Data Science Unit
Medical Faculty Heidelberg & BioQuant
carl.herrmann@uni-heidelberg.de

14-15 December 2020

Who are we ??

Medizinische Fakultät Heidelberg

Carl Herrmann

- Lecturer Heidelberg University
- Group leader *Biomedical Genomics* (BMG) @ Health Data Science Unit - Medical Faculty & BioQuant
- Interested in
  1. understanding transcription regulation in development and disease;
  2. developing computation/statistical methods for data integration
- mathematician → engineer → theoretical physicist → bioinformatician → ...
- proud father of four daughters

@CarlMHerrmann

Ashwini K. Sharma

- Computational Biologist
- Postdoc in the BMG group
- Interested in applying integrative genomics based approaches using various statistical and computational methods towards understanding tumour biology and other diseases

https://ashwini-kr-sharma.github.io/

Thank you Andres Quintero for technical support!

Goals

Medizinische Fakultät Heidelberg

- **What we will cover**
  - why are we doing ChIP-seq/ATAC-seq at all?
  - what are the main steps of the bioinformatics workflow?
  - how can we distinguish a good from a bad dataset (QC!) ?
  - which tools are available for each step of the analysis?
- **What we will NOT cover** (yet...)
  - (some) gory details ....
  - alternative ChIP-seq protocols (cut&run / cut&tag / ...)
  - DNA methylation / RNA-seq / whatever-seq
  - single-cell ATAC-seq / single-cell whatever-seq
- **After this course, you will be able to**
  - **perform some of the analysis yourself**
  - **talk without shame to your favorite bioinformatician**

Schedule

Medizinische Fakultät Heidelberg

Day 1 : ChIP-seq analysis

- **10am - 11am**
  General introduction on experimental and computational concepts
- **11am - 12.30am**
  First steps in the bioinformatics workflow (lectures + hands-on)
  - read QC / trimming / alignment
- **1.30pm - 5.30pm**
  Next steps in the bioinformatics workflow (lectures + hands on)
  - peak calling
  - peak annotation
  - signal tracks
  - IGV visualization

Day 2 : ATAC-seq analysis

- **10am - 12.30pm**
  First steps in the bioinformatics workflow (lectures + hands-on)
  - read QC / trimming / alignment
  - peak calling / peak annotation
- **1.30pm - 5.30pm**
  ATAC-seq specific part
  - QC
  - Footprinting
  - Integration with ChIP-seq
  - ...

Understanding gene regulation

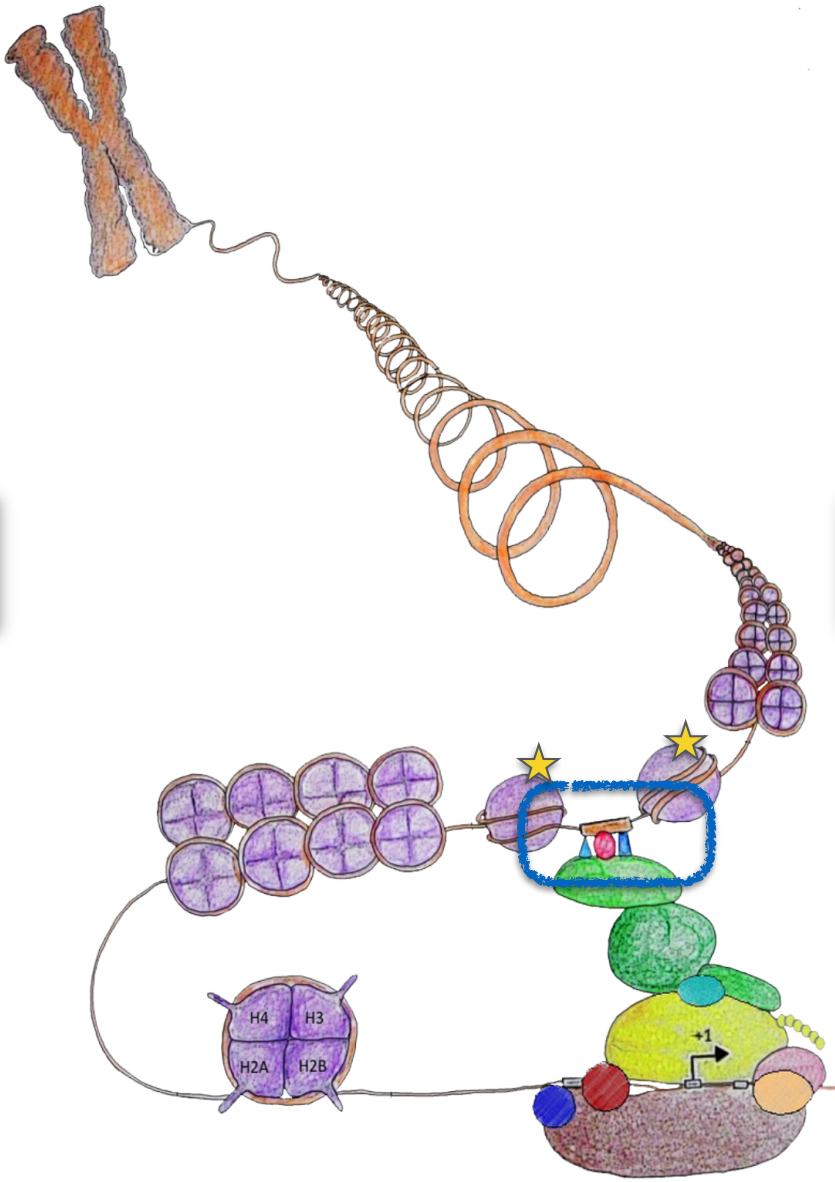
Transcriptional regulation

Medizinische Fakultät Heidelberg

2. chromatin structure, epigenetic

1. Binding of site specific transcription factors

3. three-dimensional DNA looping



4. Readout: gene expression

[Elodie Darbo]

Experimental methods

- **Genetic component:**
- **Chromatin structure and epigenetic**
- **Three-dimensional DNA looping**
- **Readout: gene expression**

Experimental methods

- **Sequence component:**
  → ChIP-seq: transcription factor binding sites

- **Chromatin structure and epigenetic**
  → ChIP-seq : post-translational histone modifications
  → whole genome bisulfite sequencing, arrays : DNA methylation
  → ATAC-seq, DNAse-seq, FAIRE-seq : open chromatin region

- **Three-dimensional DNA looping**
  → 3C/4C/Hi-C : interacting chromatin regions

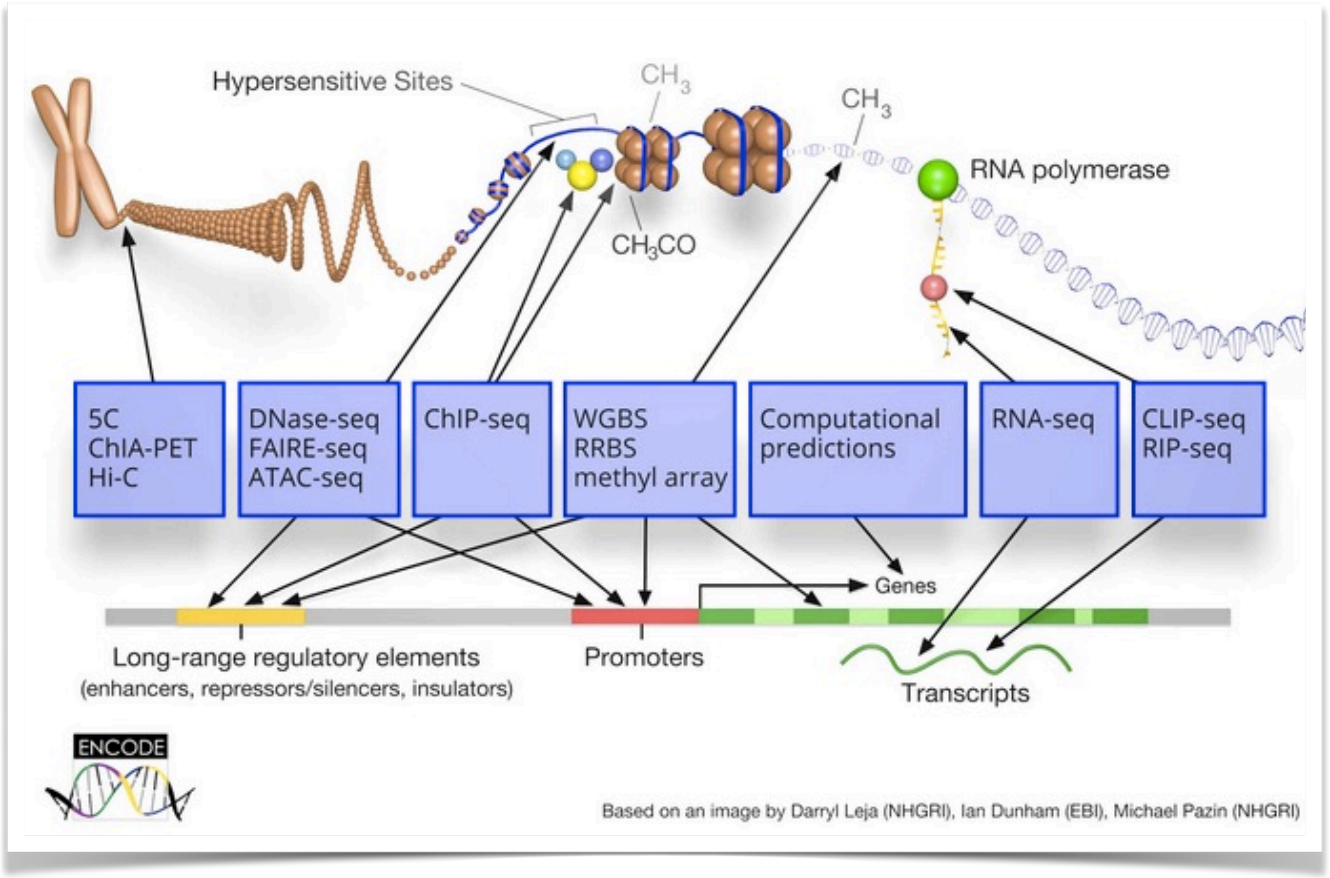
- **Readout: gene expression**
  → RNA-seq : expression of transcribed elements

Exploring the genome's activity

Medizinische Fakultät Heidelberg

- Large scale consortia (ENCODE, Roadmap, …) have systematically explored the **activity** of the genome using experimental assays

"***The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/ or chromatin-associated event in at least one cell type***.
99% is within 1.7kb of at least one of the biochemical events measured by ENCODE."



https://www.encodeproject.org/

https://www.encodeproject.org/matrix/?type=Experiment

ENCODE data

Medizinische Fakultät Heidelberg

Experiment Matrix

EXPERIMENTS

Clear Filters

Filter the experiments included in the matrix:

Showing 16278 results

Enter search term(s)

List | Report | Download | Visualize | { ; }

Assay type

Assay title

Search

TF ChIP-seq	3790
Histone ChIP-seq	3160
Control ChIP-seq	2338
DNase-seq	1192
scRNA-seq	1063
polyA plus RNA-seq	685
total RNA-seq	456
Mint-ChIP-seq	281
microRNA-seq	256

Status

Selected filters: released

released	16278
archived	1088
revoked	331

Perturbation

Selected filters: not perturbed

not perturbed	16278
perturbed	2166

Target category

Target of assay

ASSAY →

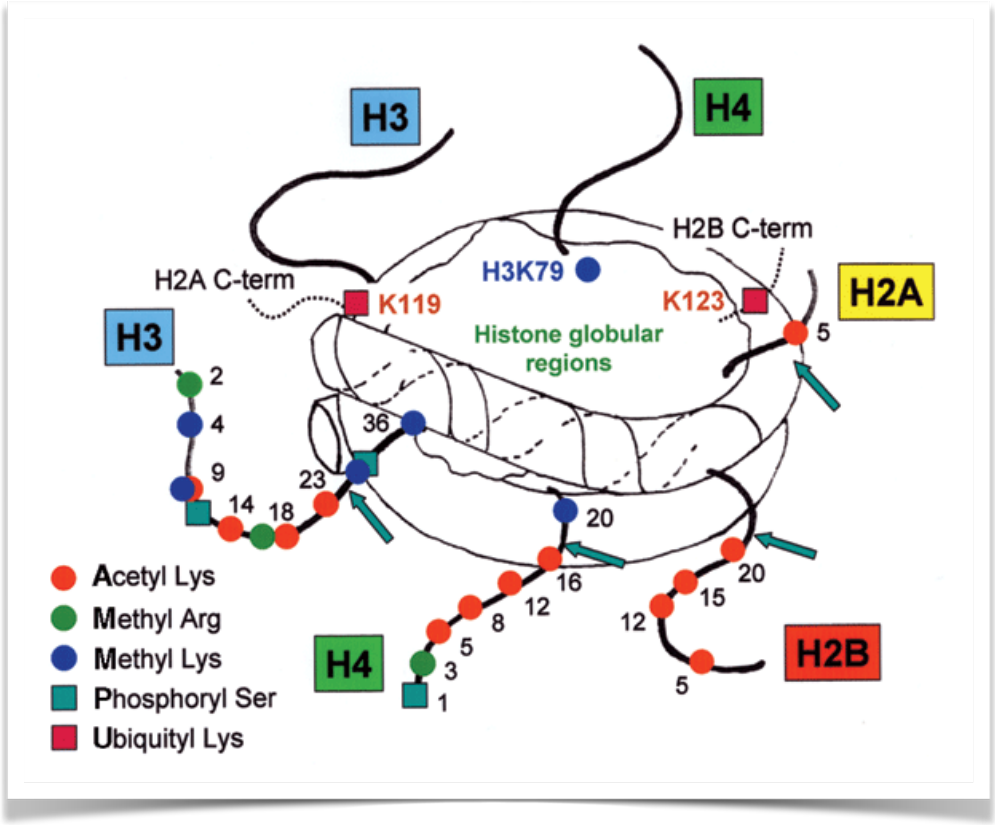
BIOSAMPLE ↓

	TF ChIP-seq	Histone ChIP-seq	Control ChIP-seq	DNase-seq	scRNA-seq	polyA plus RNA-seq	total RNA-seq	Mint-ChIP-seq	microRNA-seq	DNAme array	Control eCLIP	eCLIP	small RNA-seq
cell line	2361	655	606	170	2	143	84	16	26	87	232	223	110
K562	606	19	182	4		15	11		2	3	125	120	7
HepG2	640	15	73	2		11	5		2	3	107	103	3
GM12878	187	15	26	2	2	13	3		2	3			6
MCF-7	146	18	34	4		4			2	2			7
HEK293	198	6	35							2			
tissue	332	1792	567	569	14	397	143		179	122	2	2	67
liver	42	91	29	9		20	3		7	1			1
stomach	17	72	26	22		15	5		4	3			4
heart	5	79	15	21		16	3		9				1
spleen	18	58	23	6		11	4		3	4			4
lung	6	58	15	17		11	1		4	2			1
whole organisms	996		987			41	68						
whole organism	996		987			33	64						
carcass						8	4						
primary cell	66	502	124	436	8	82	154	233	36	38			24
T-cell		11	3	59		1		6					
macrophage						1	78						
activated CD4-positive, alpha-beta T cell				78									
CD14-positive monocyte	1	21	3	7		2		24					1
endothelial cell of umbilical vein	13	16	7	2		5				1			1

https://www.encodeproject.org

ChIP-seq for histone modifications

Medizinische Fakultät Heidelberg



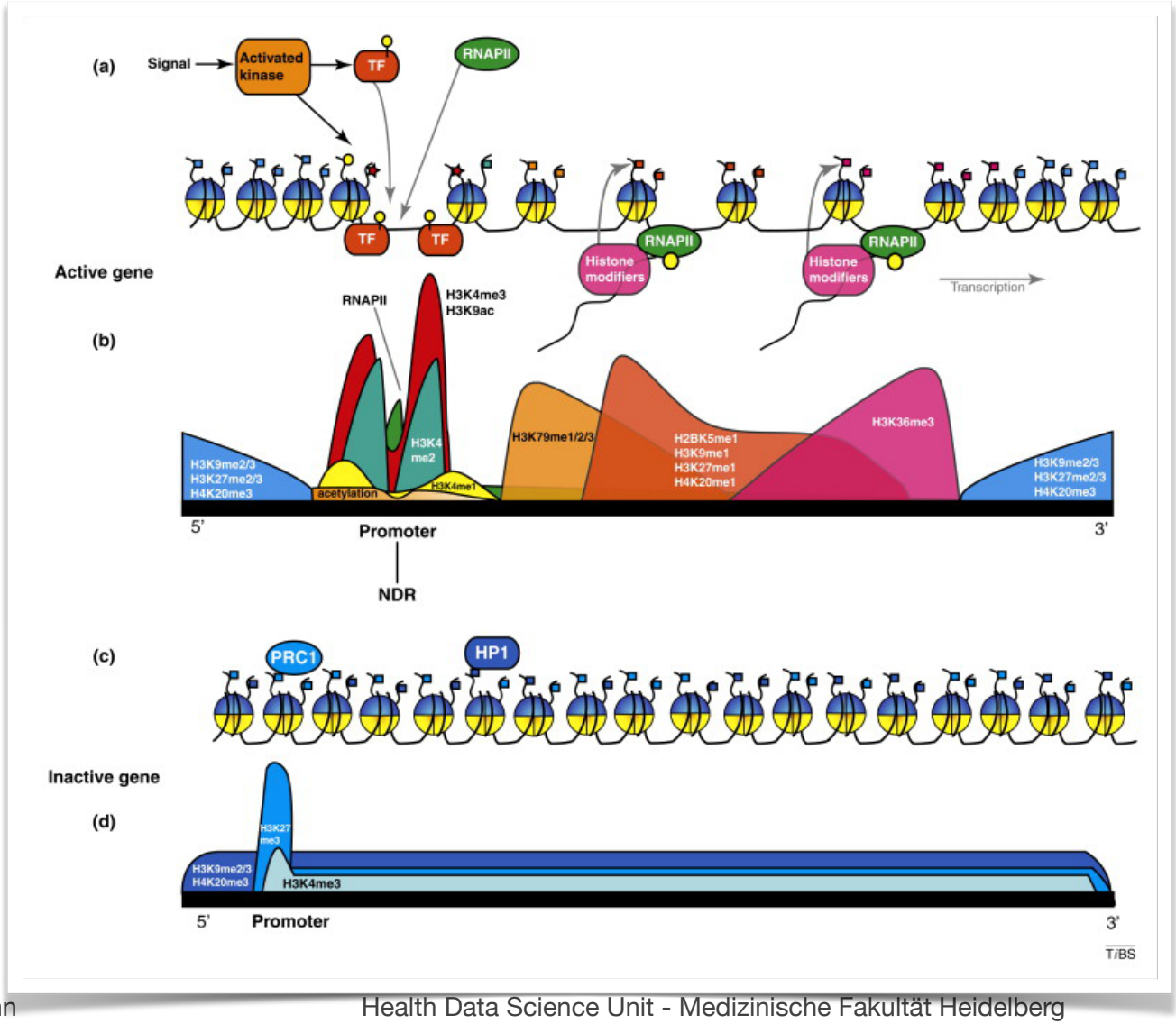
- histones are subject to **post-translational modifications** at their N-terminal tail
  - Lysine methylation
  - Lysine/arginine acetylation
  - Serine phosphorylation
  - ubiquitylation
- they **modify the physical properties of the DNA-nucleosome interactions**

nomenclature: H3K27ac = acetylation of lysine 27 on histone 3

Histone modifications

Medizinische Fakultät Heidelberg

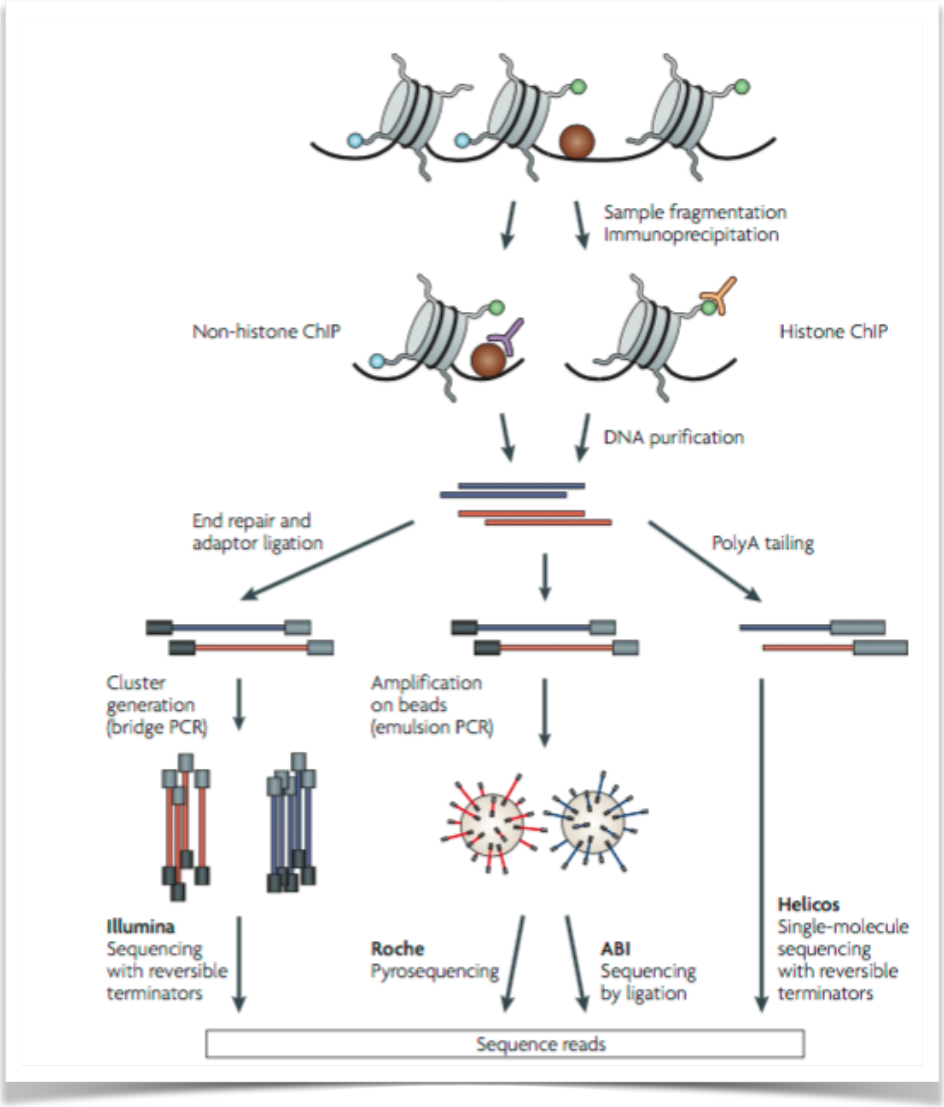
histone modifications are a good proxy of gene expression and presence of regulatory elements



active marks
→ open chromatin
H3K4me1; H3K4me3; H3K27ac

repressive marks
→ closed chromatin
H3K27me3; H3K9me3

Chromatin Immunoprecipitations



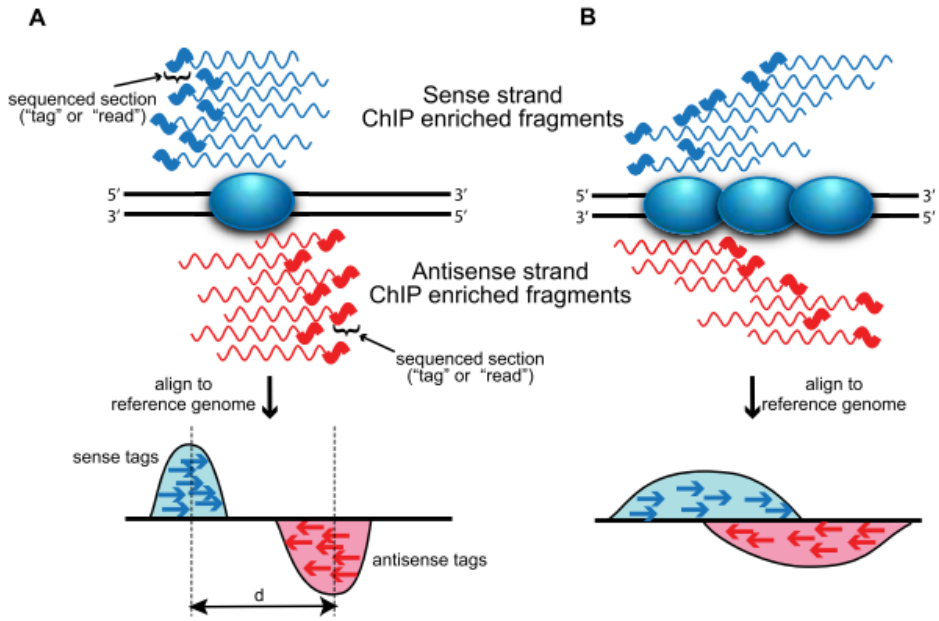
[Park, Nat.Rev. 2009]

- Chromatin immunoprecipitation (ChIP) yields **DNA fragments**, that are
  - bound by the protein of interest
  - marked by a specific chemical modification (acetylation, methylation,.)
- Identification of the fragments :
  - sequencing (ChIP-seq)
    → genome-wide
  - PCR/qPCR
    → targeted experiment
- Important aspect
  - Quality/Specificity of the antibody ?
  - DNA fragment (~200-300bp)
    → binding site (~10 bp) ?

ChIP-sequencing

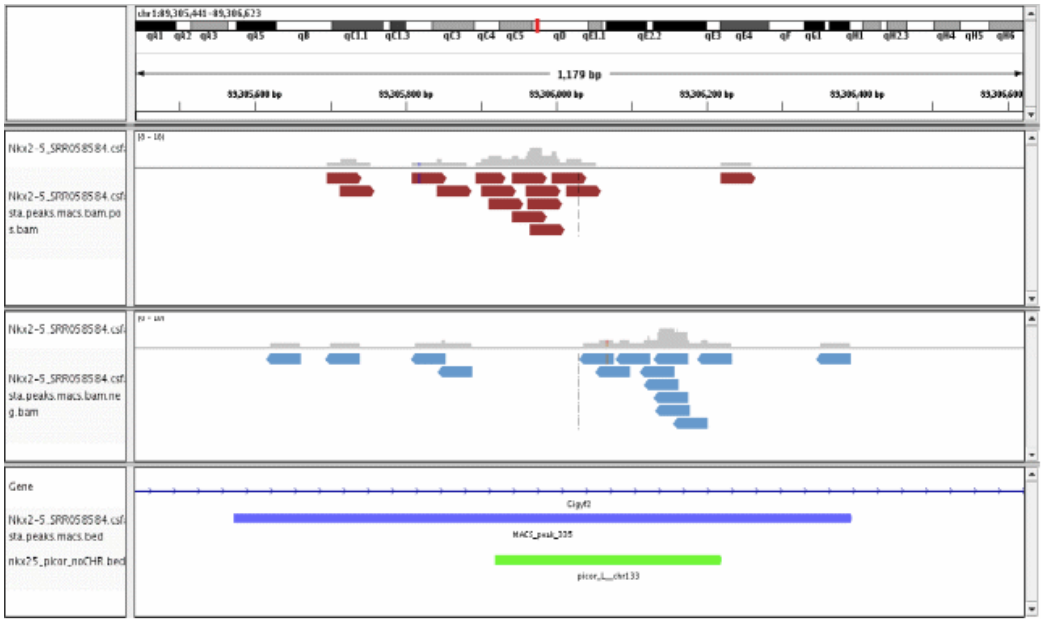
Medizinische Fakultät Heidelberg

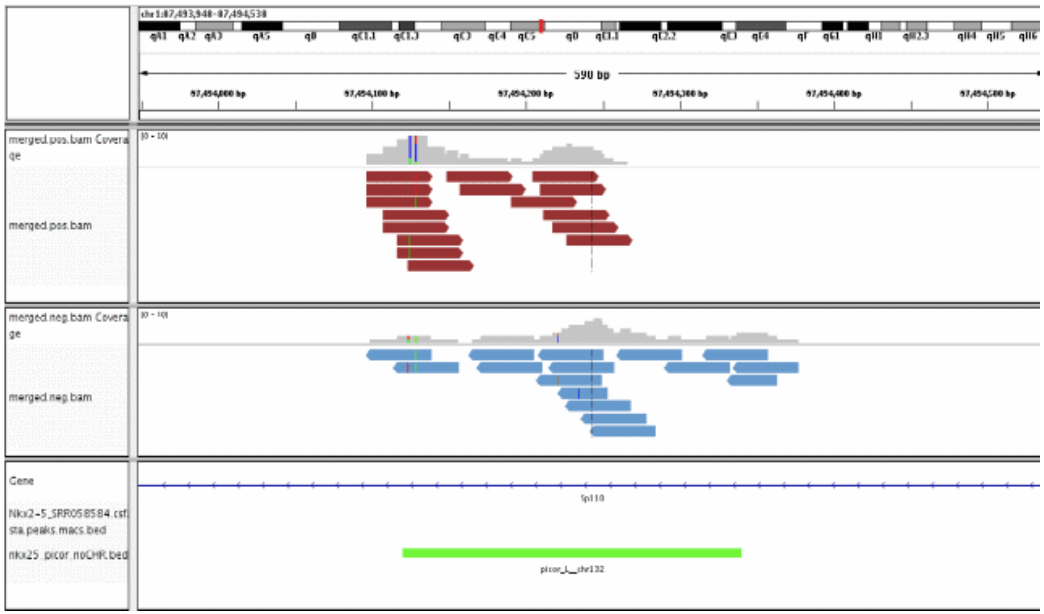
Sharp signal



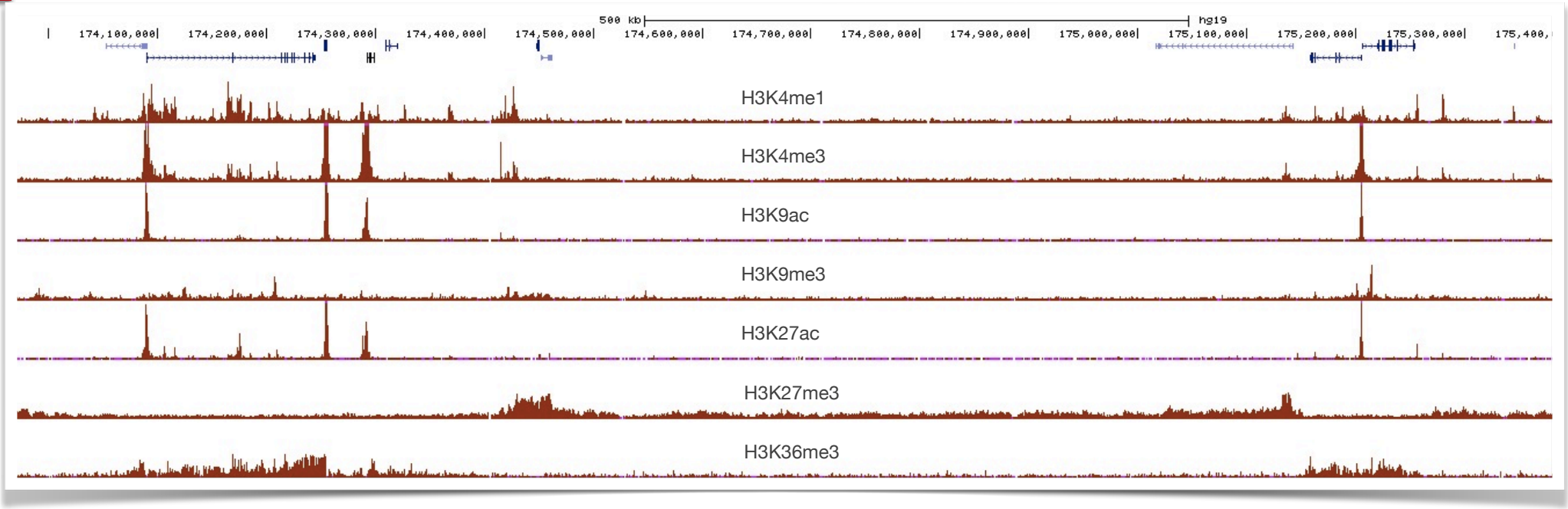
Broad signal

[Wilbanks & Faccioti PLoS One (2010)]]



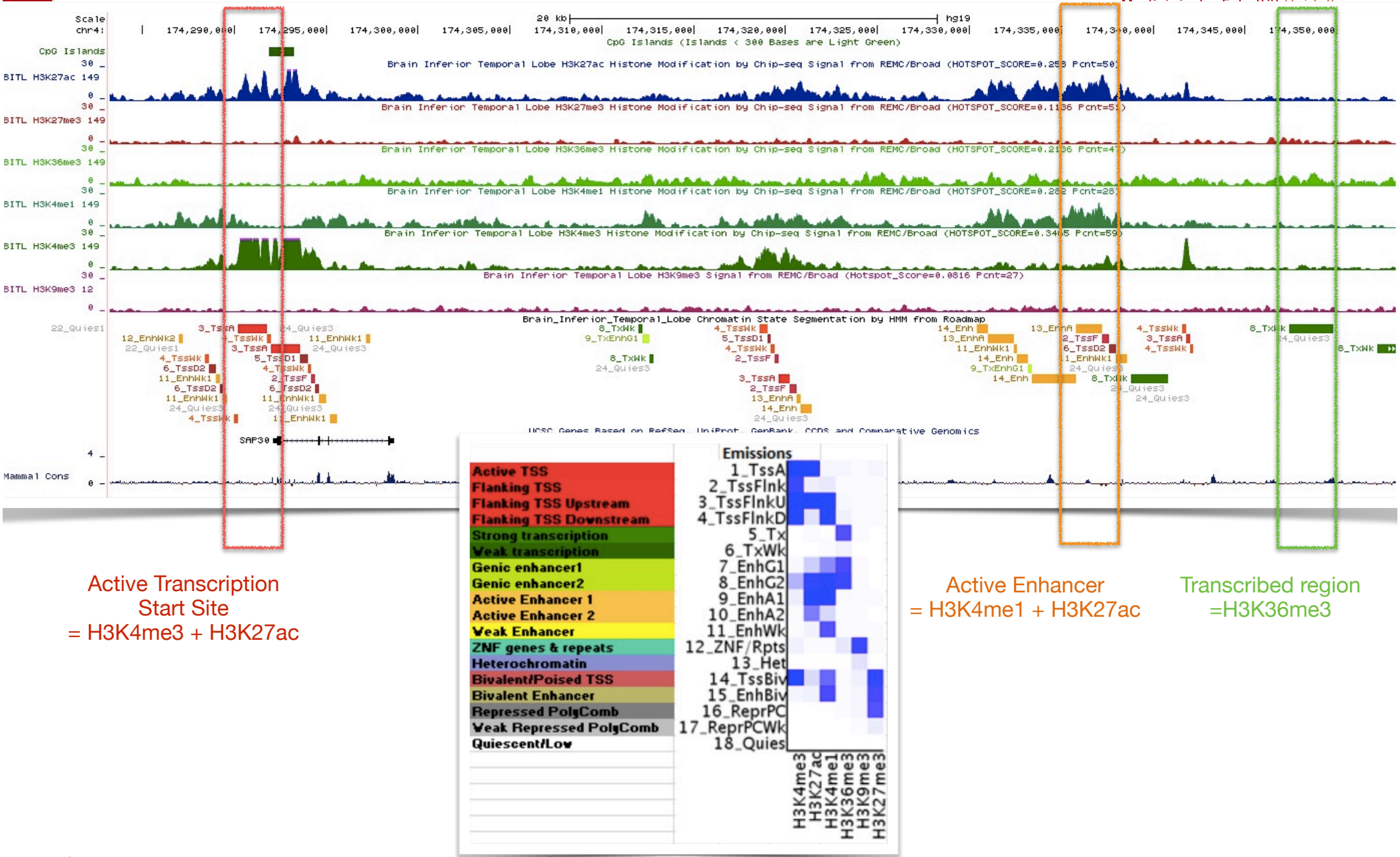


Histone modifications



- Histone marks have **distinct signal profiles**
  - sharp signal : narrow peaks of enrichment at specific loci (H3K4me3 = promoters, H3K27ac = enhancers,…)
  - broad signal : wide regions of enrichment (H3K36me3 = transcribed genes; H3K27me3 = repressed regions)

Chromatin states



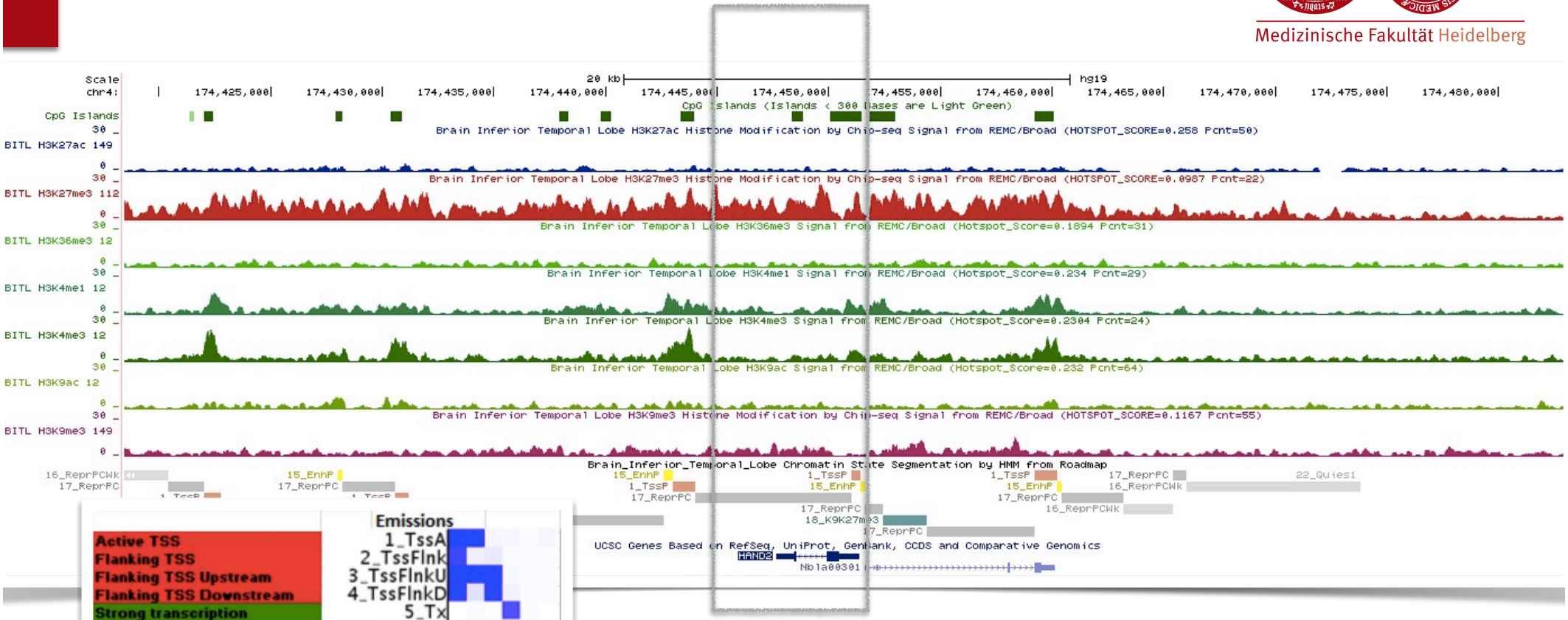
Active Transcription Start Site
= H3K4me3 + H3K27ac

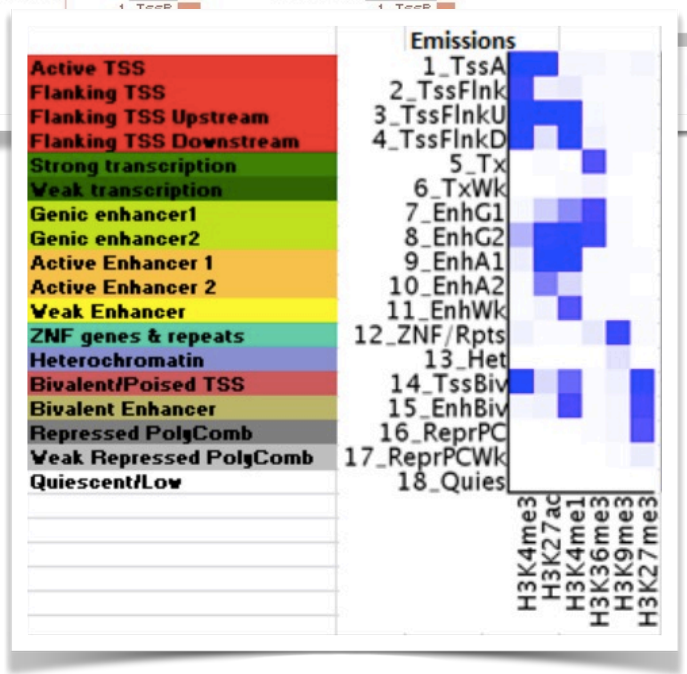
Active Enhancer
= H3K4me1 + H3K27ac

Transcribed region
=H3K36me3

Chromatin states

Medizinische Fakultät Heidelberg

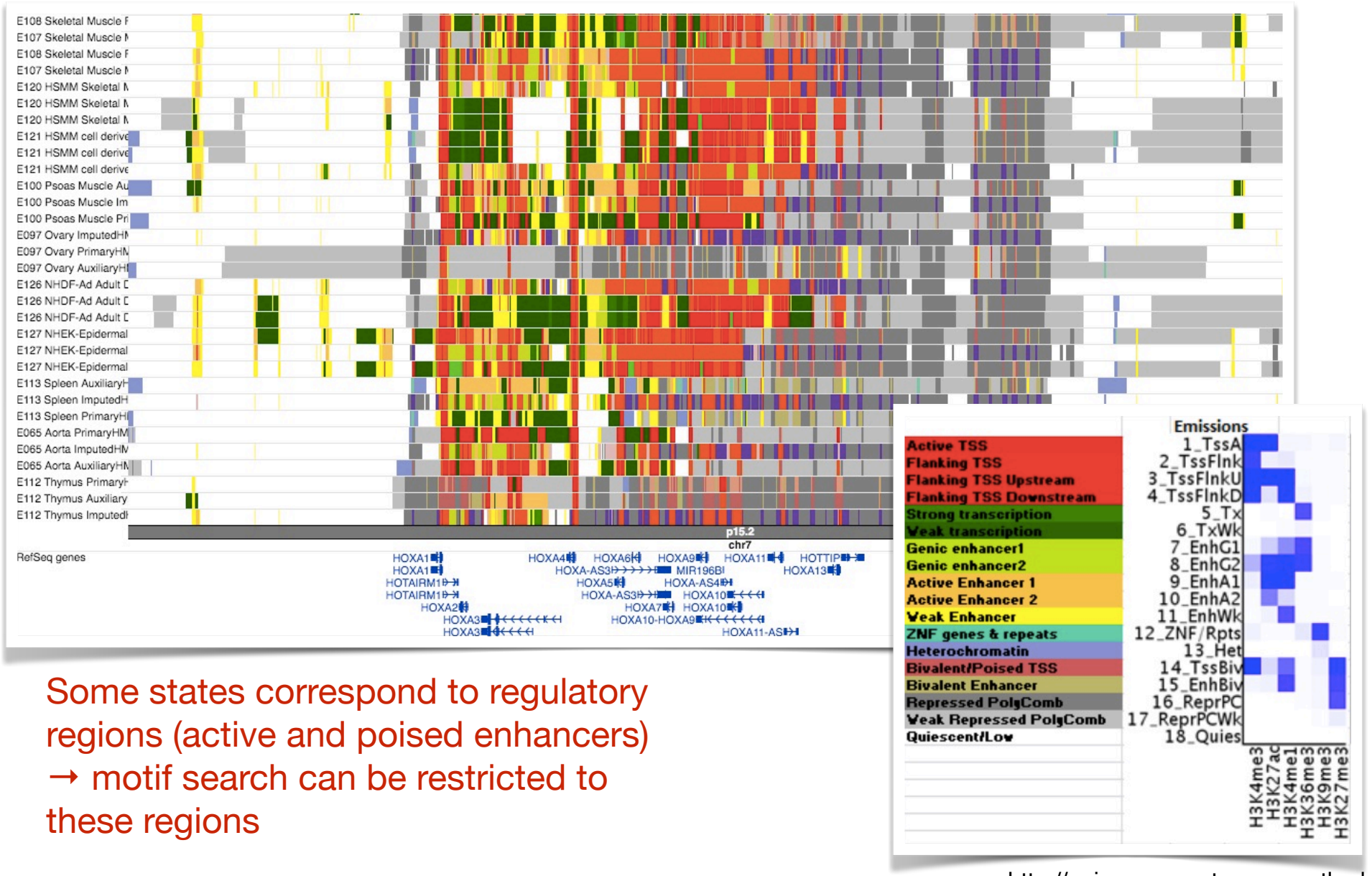




Repressed Polycomb
= H3K27me3

Roadmap chromatin segmentation in different human adult tissues

Medizinische Fakultät Heidelberg

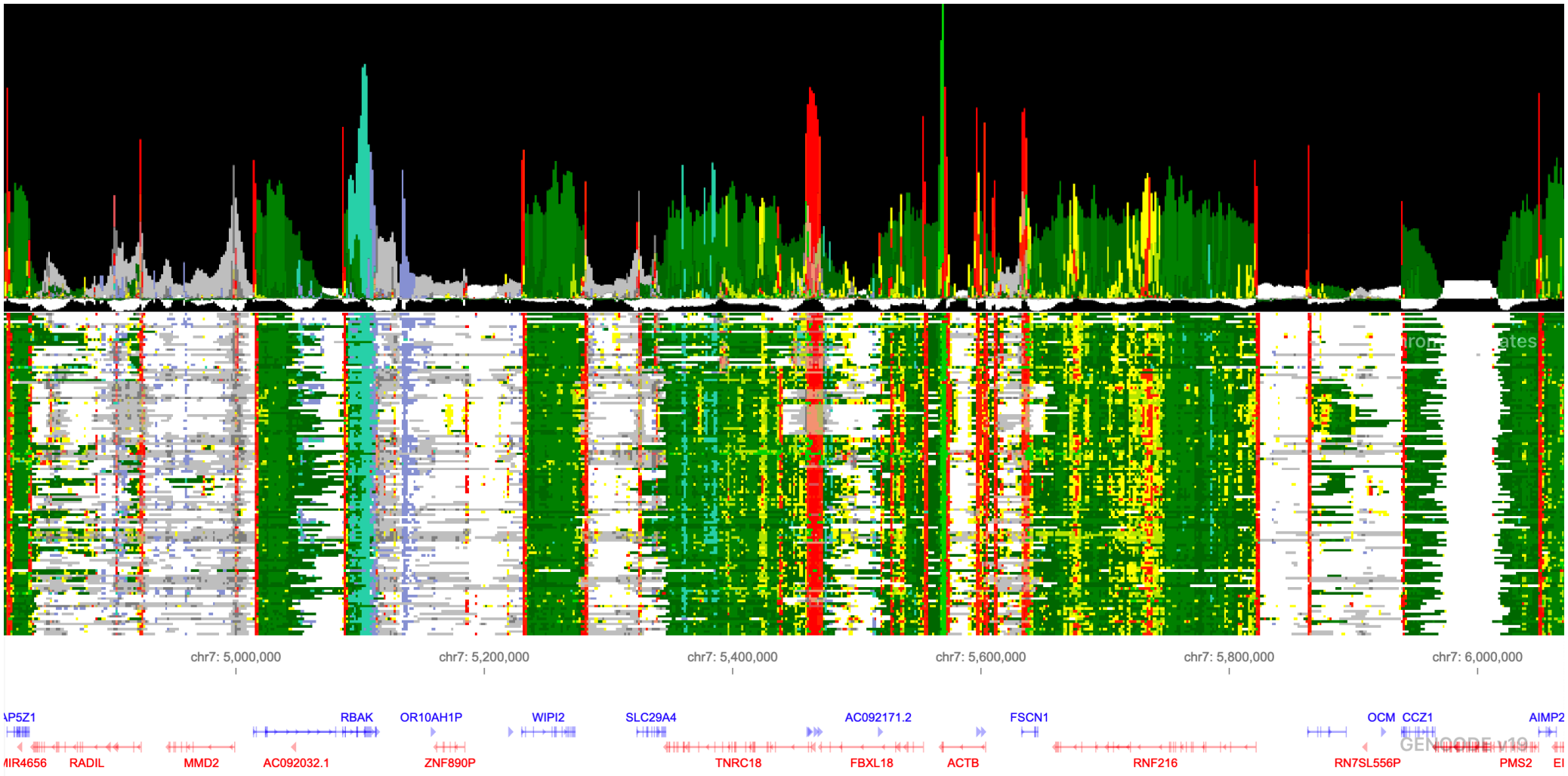


Some states correspond to regulatory regions (active and poised enhancers)
→ motif search can be restricted to these regions

http://epigenomegateway.wustl.edu

EpiLogos

Medizinische Fakultät Heidelberg

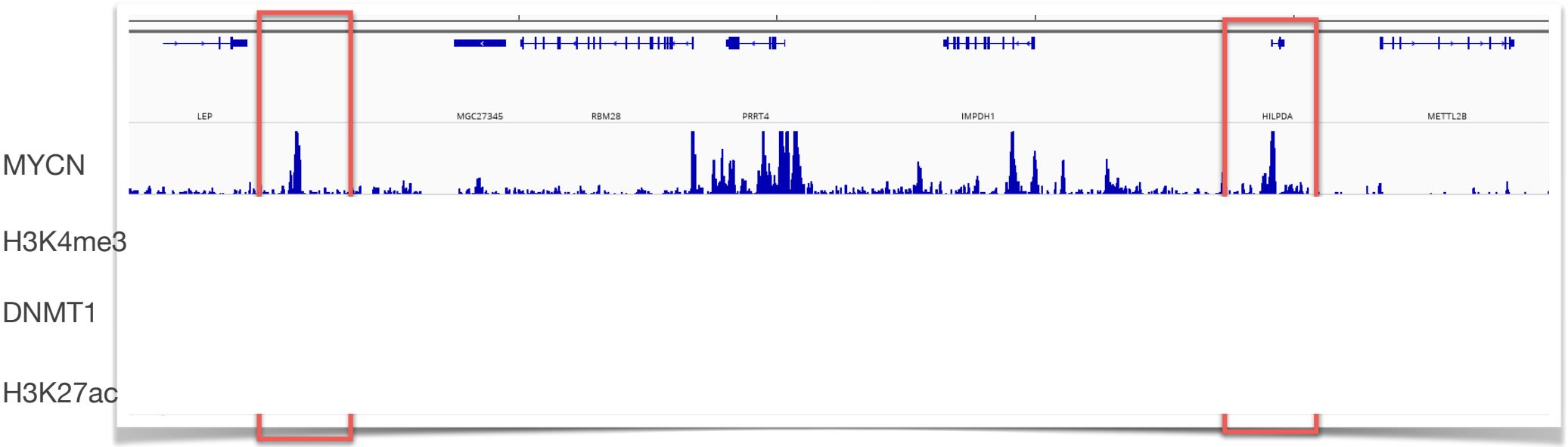


[W. Meuleman]

https://epilogos.altius.org

Example of ChIP-seq signal for transcription factors / DNA-binding proteins

Medizinische Fakultät Heidelberg



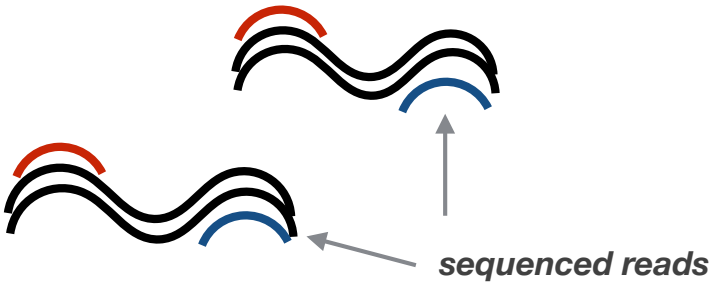
distal
MYCN peak
(away from gene promoters)

MYCN peak and H3K4me3 enrichment at promoter

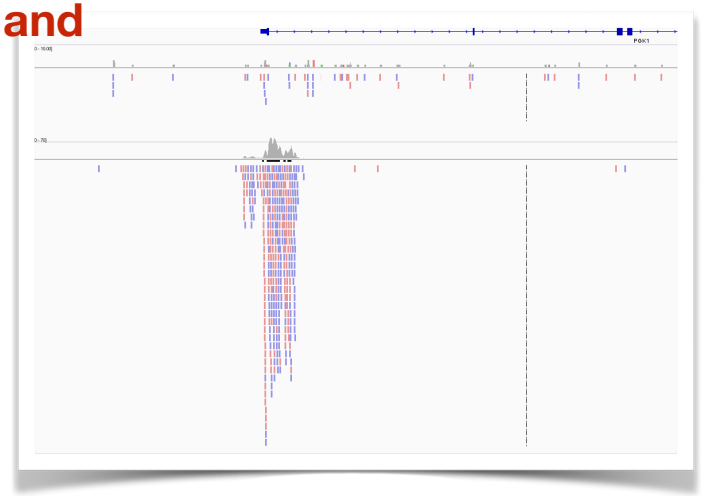
Principle of ChIP-seq

Sonication and immunoprecipitation

Sequencing of IP'd fragments



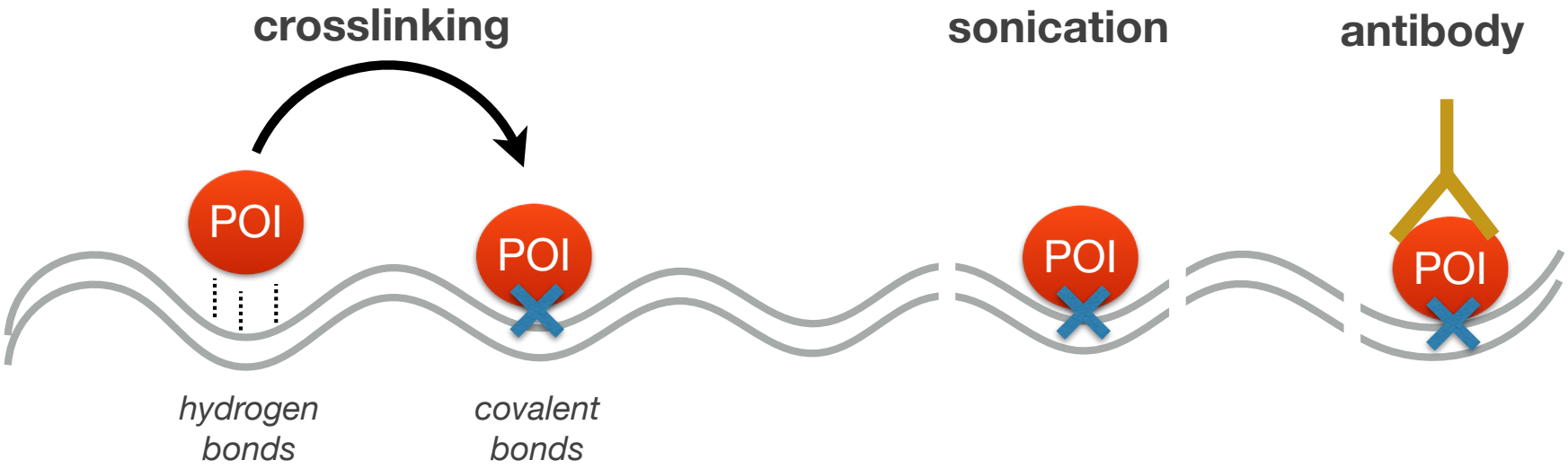
Alignment and analysis



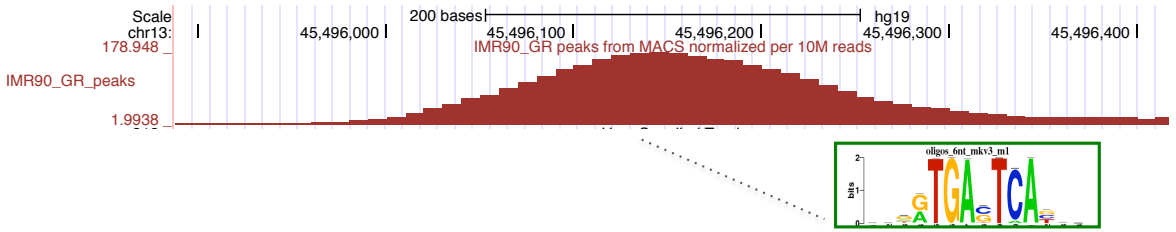
ChIP-seq

Medizinische Fakultät Heidelberg

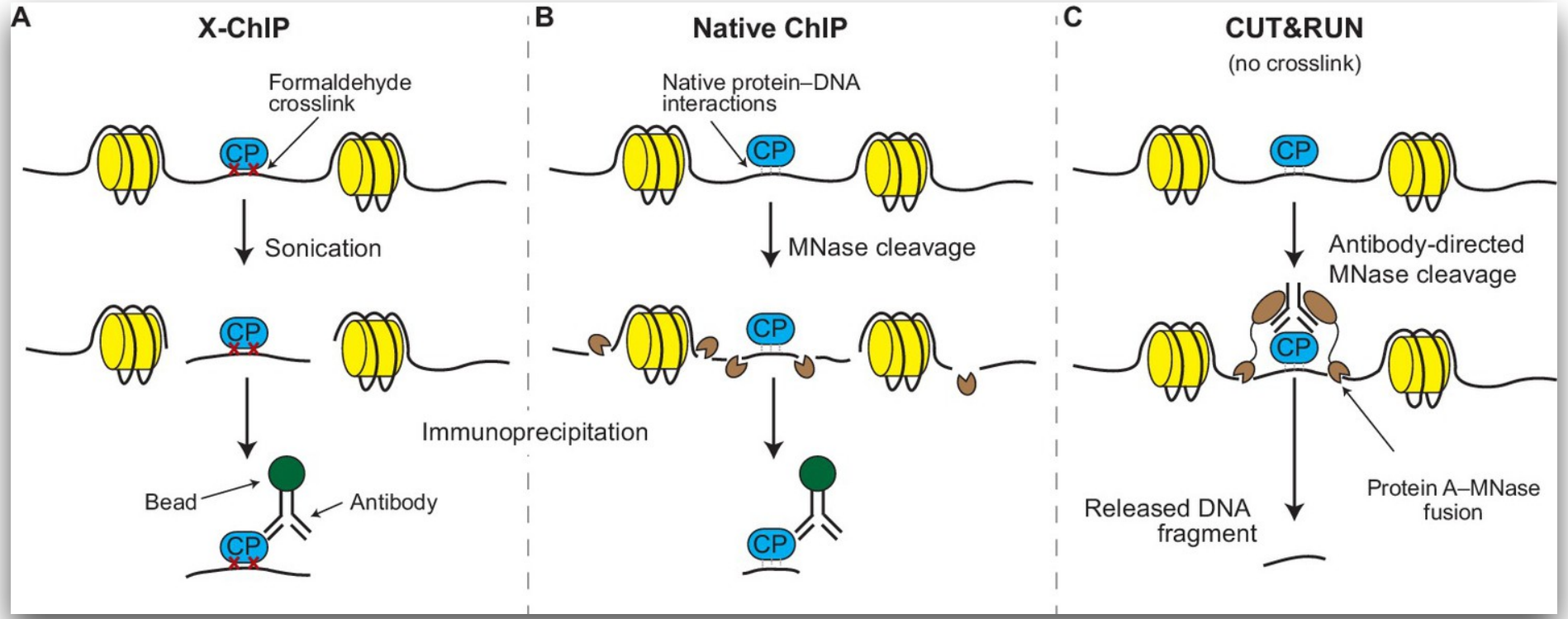
- Crucial steps in conventional ChIP-seq
  - cross-linking of the protein to the DNA
  - sonication of the cross-linked chromatin
  - antibody



low resolution (200-300bp) background noise due to crosslinking



Alternative protocols



standard protocol with crosslinking

no crosslinking
cleavage using MNase

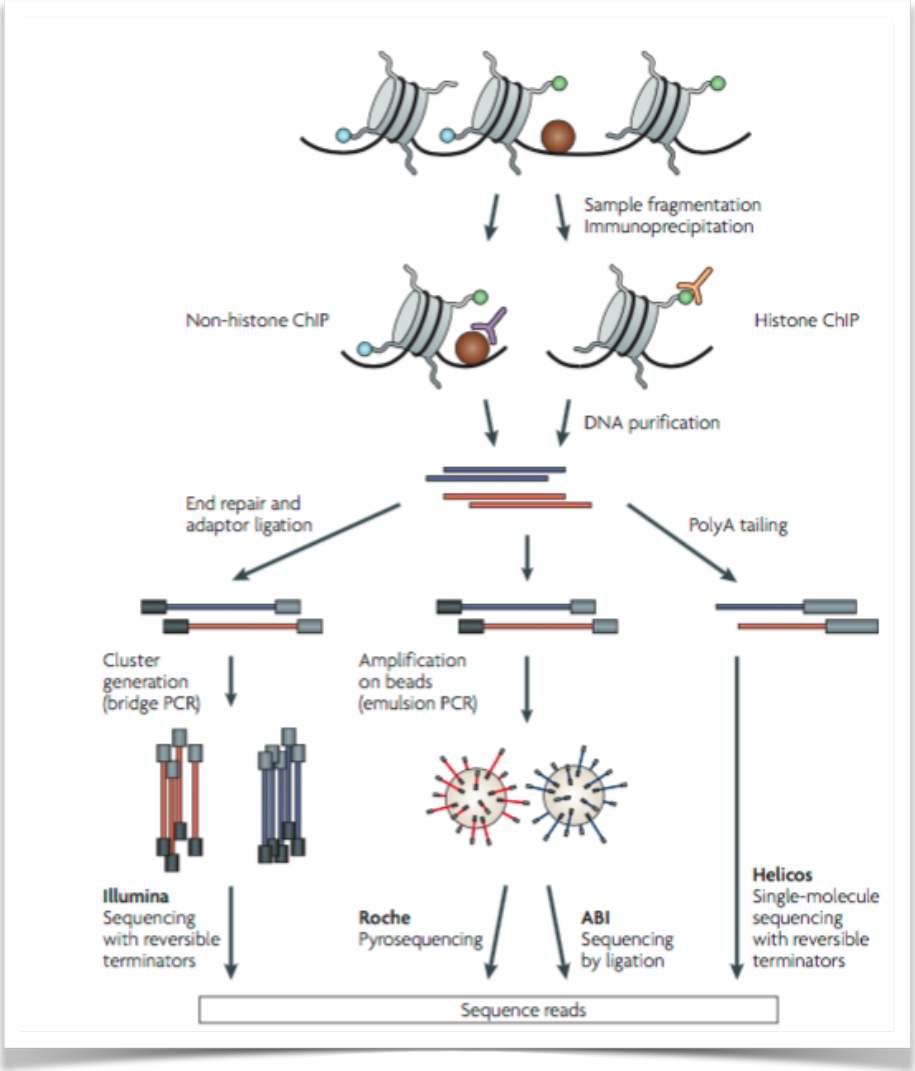
no crosslinking
MNase guided through tethering to antibody

[He & Bonasio, Elife 2017]

Controls in ChIP-seq

Medizinische Fakultät Heidelberg

- **Input DNA**: controls biases due to chromatin fragmentation (natively open regions,...)
- **Unspecific IgG** (mock-IP): controls for unspecific IP enrichment
- **H3 antibody** (for histone ChIP-seq): controls for the presence of histones

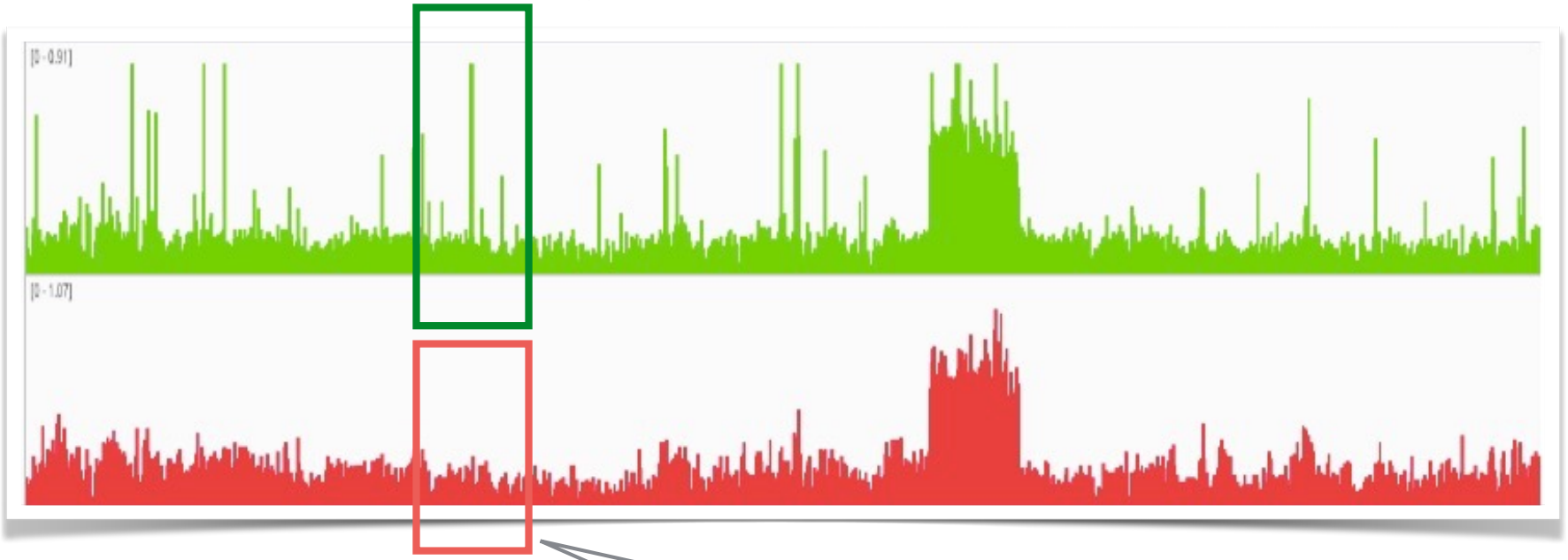


Fundamental question in ChIP-seq analysis

Medizinische Fakultät Heidelberg

Do we have more signal here …

Signal (“treatment")



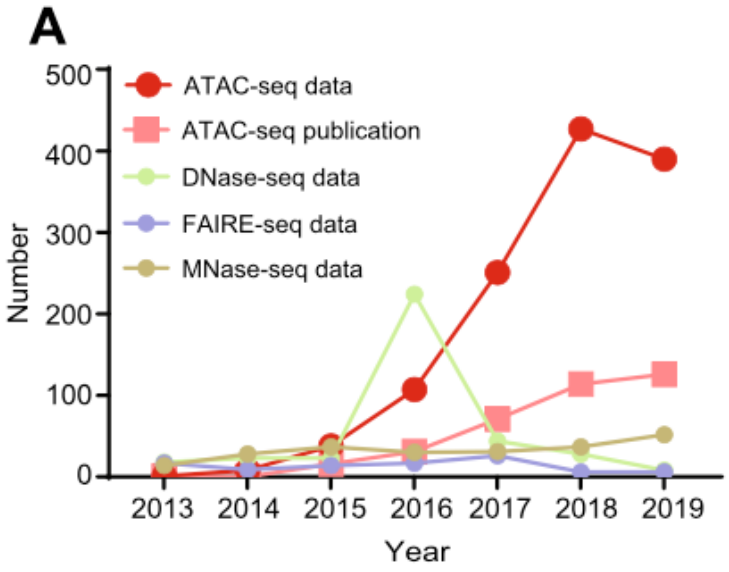
Background (“input”)

…than here ?

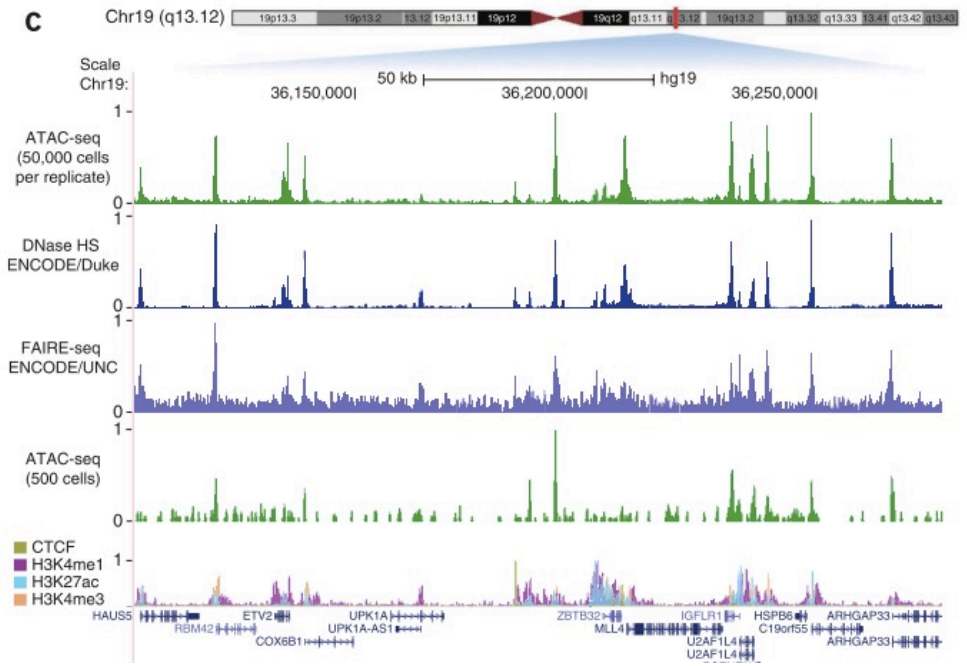
ATAC-seq: finding open regions

Medizinische Fakultät Heidelberg

- Several experimental methods available to identify open chromatin regions



[From reads to insight: a hitchhiker's guide to ATAC-seq data analysis; Yan et al. Genome Biology 2020]



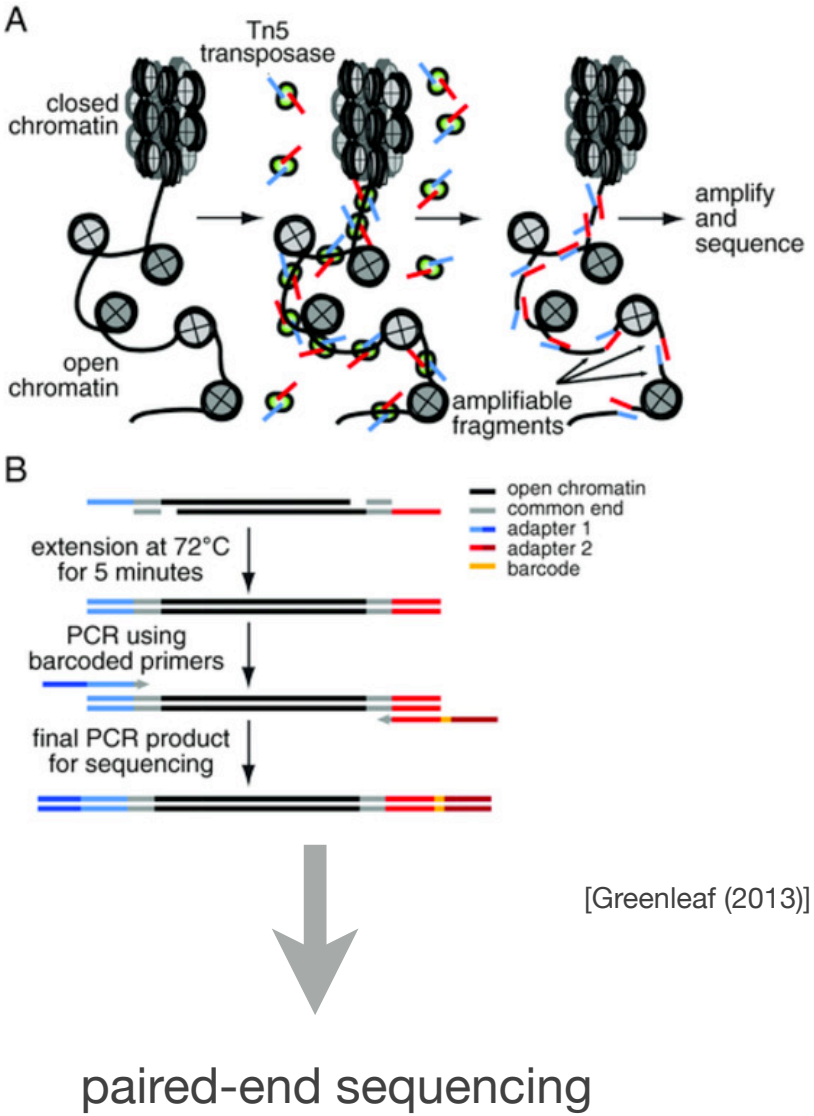
- Lower input material required
- Simpler protocol with less steps
- Comparable sensitivity/specificity compared to DNAse-seq

[Greenleaf (2013)]

ATAC-seq : finding open regions

Medizinische Fakultät Heidelberg

- **ATAC-seq**: using Tn5 transposase prepared with sequencing primers
- requires a small number of input material (~10,000 cells)
- identification of open chromatin regions (peaks)
- **There is no control in ATAC-seq experiments (unlike ChIP-seq)**



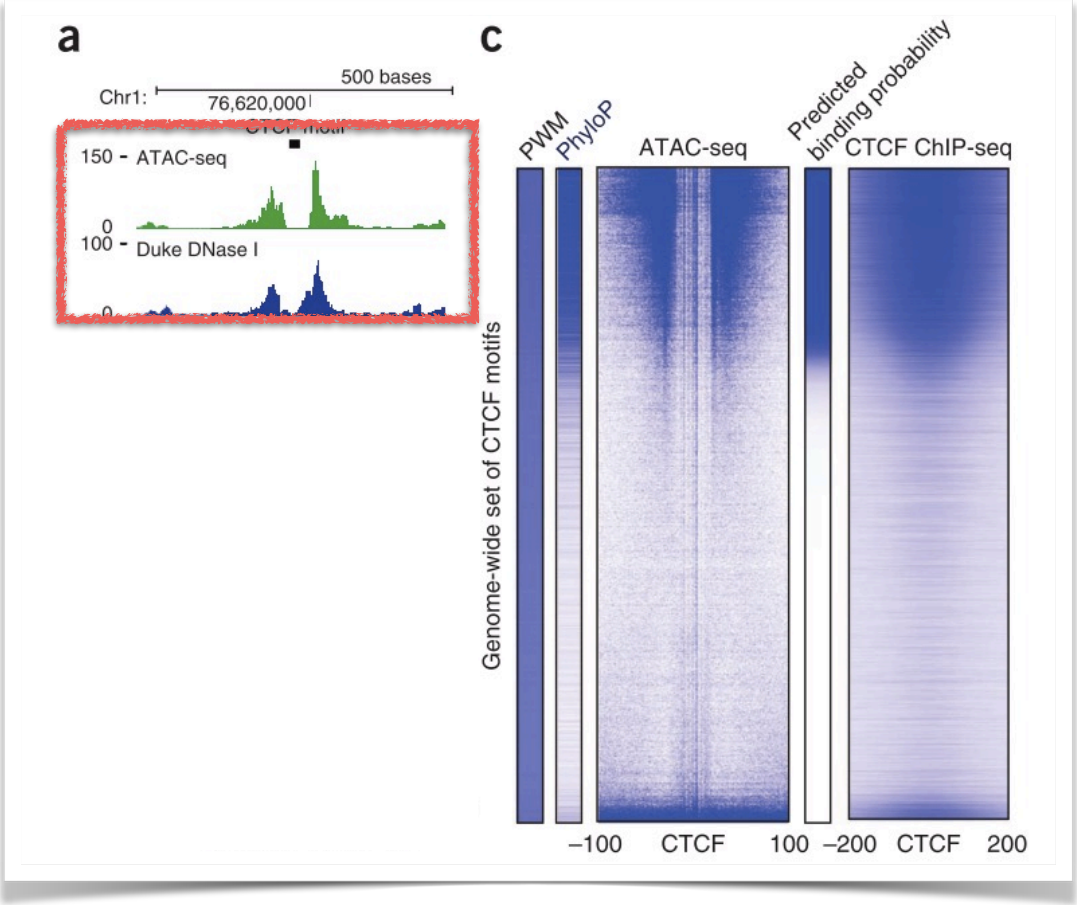
[Greenleaf (2013)]

paired-end sequencing

Footprinting

Medizinische Fakultät Heidelberg

- From open regions to transcription factor binding sites
  → **footprinting**
- Zooming into the peaks (open regions) : valleys of undigested / un-transposed DNA
  → **TF binding sites (TFBS)**
- binding sequence can be identified with base-pair resolution



[Greenleaf (2013)]