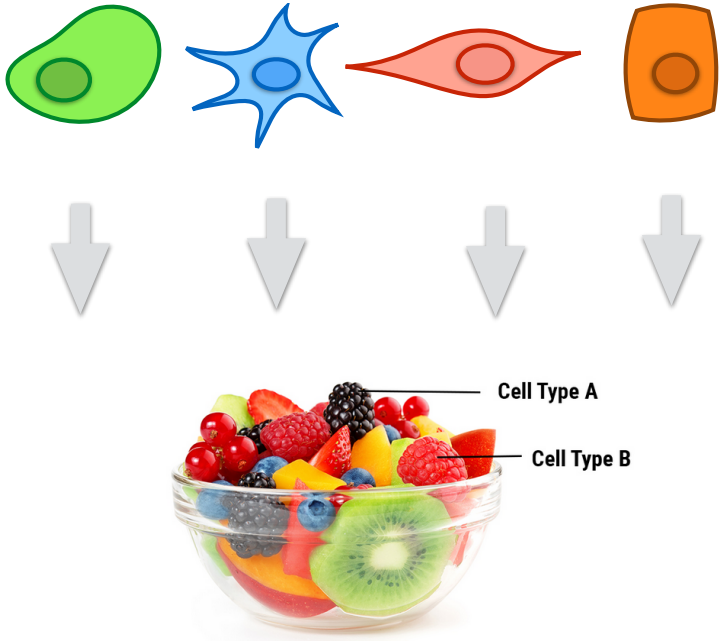
Single-cell sequencing

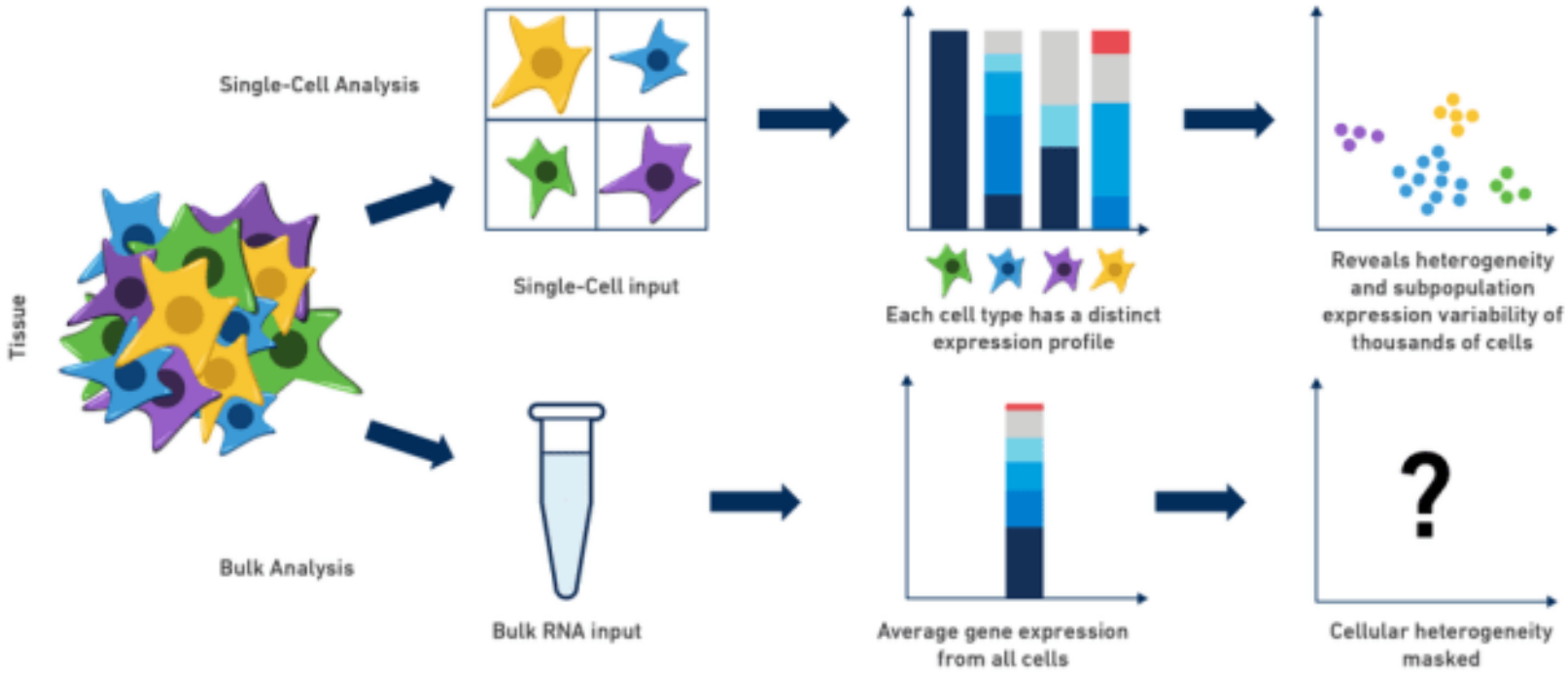
Bulk sequencing



Single-cell sequencing

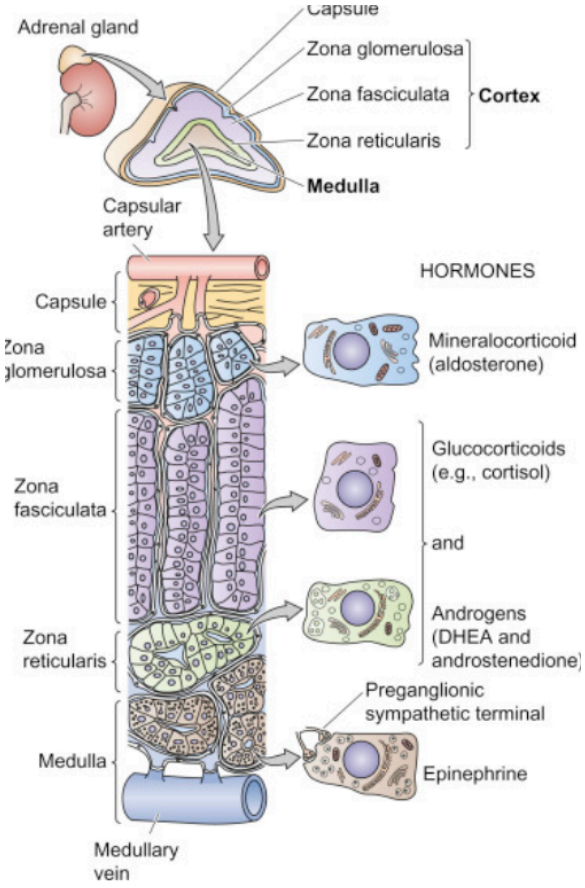
[Shalek & Regev, 2016]

Single-cell sequencing

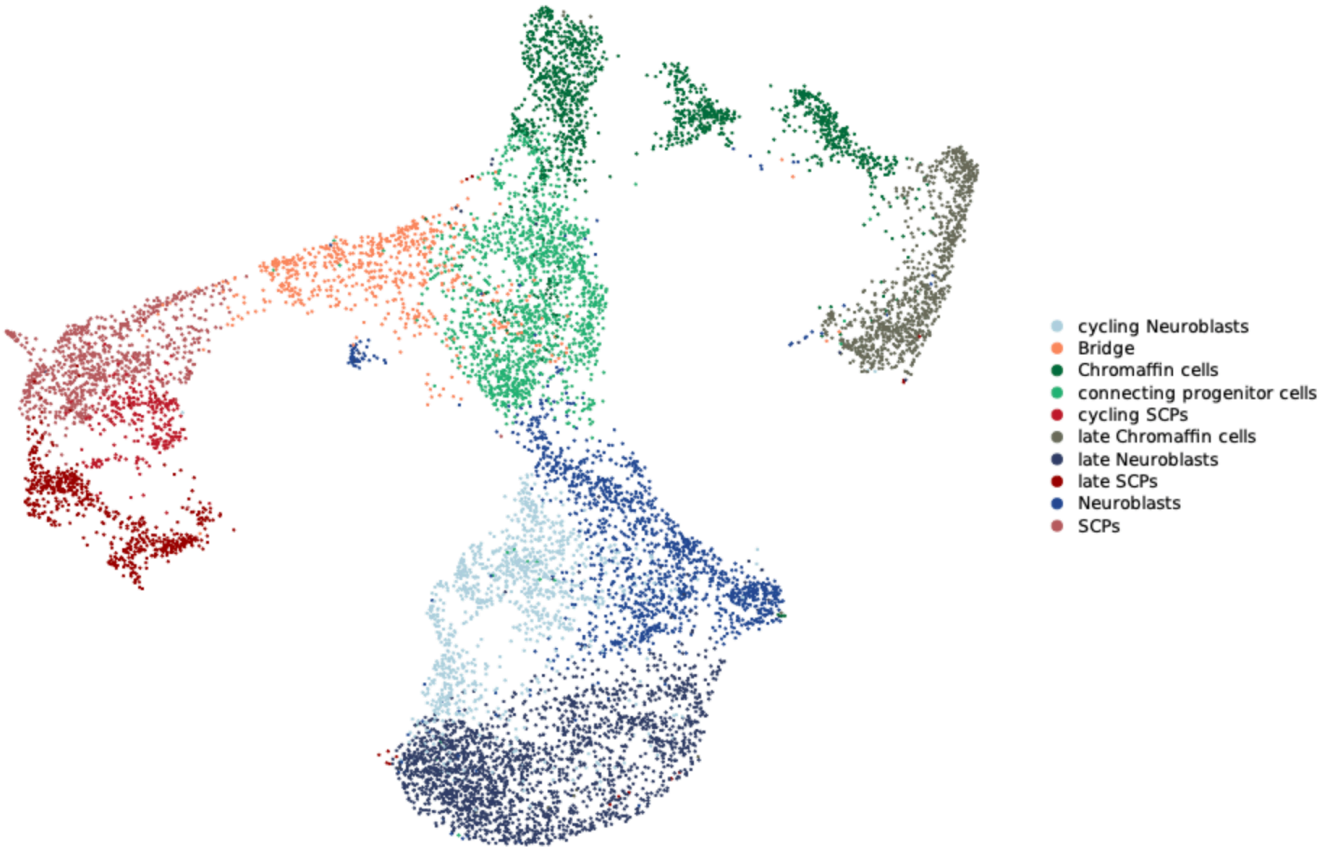


[10x genomics]

Adrenal medulla



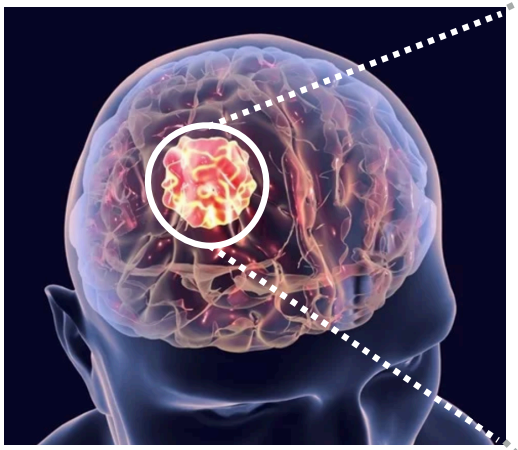
single-cell RNA-seq of
developping adrenal medulla

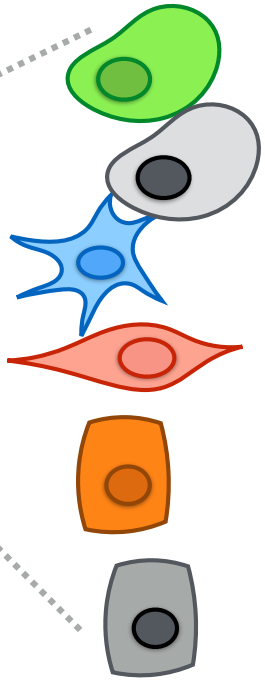


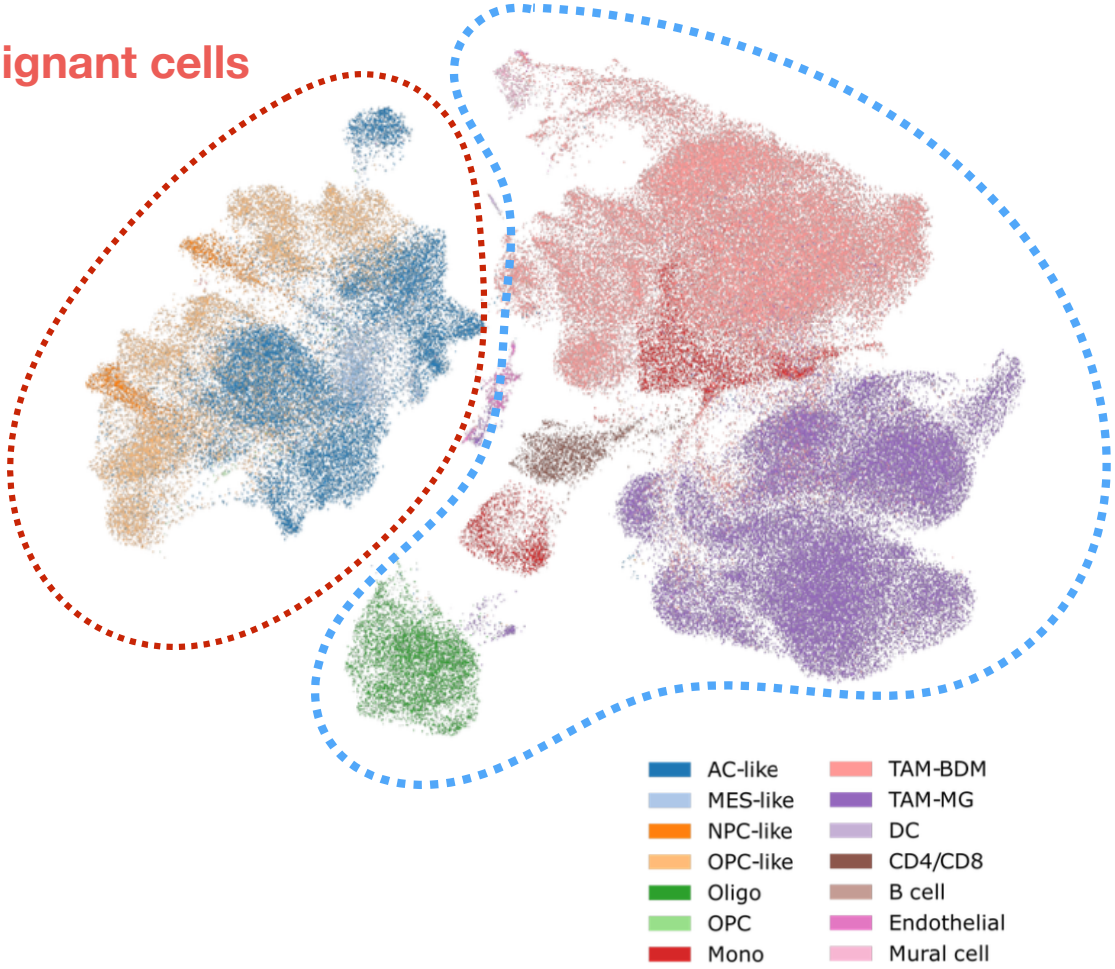
[Jansky et al., Nature Genetics 2021]

Tumor heterogeneity

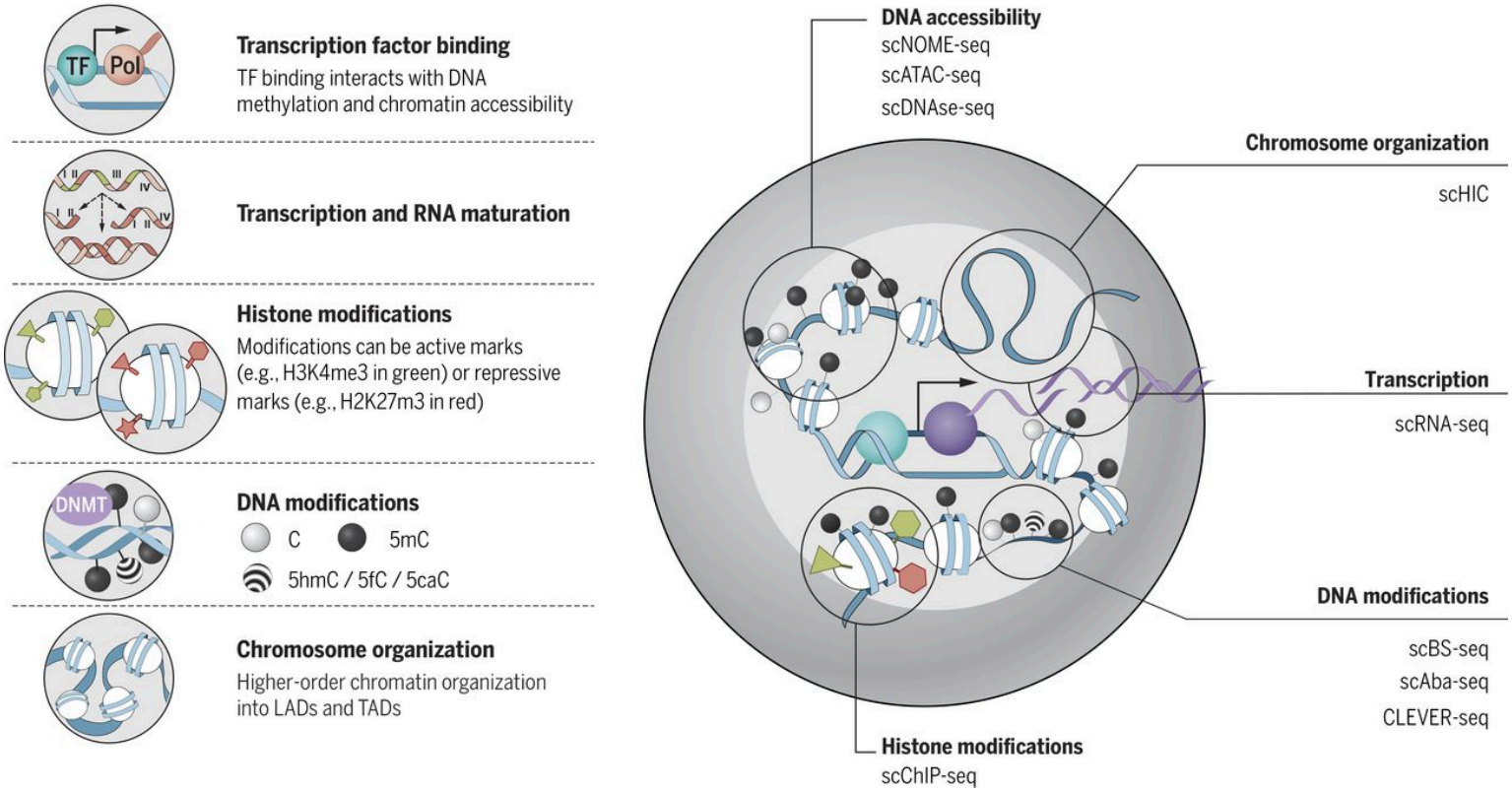
Glioblastoma





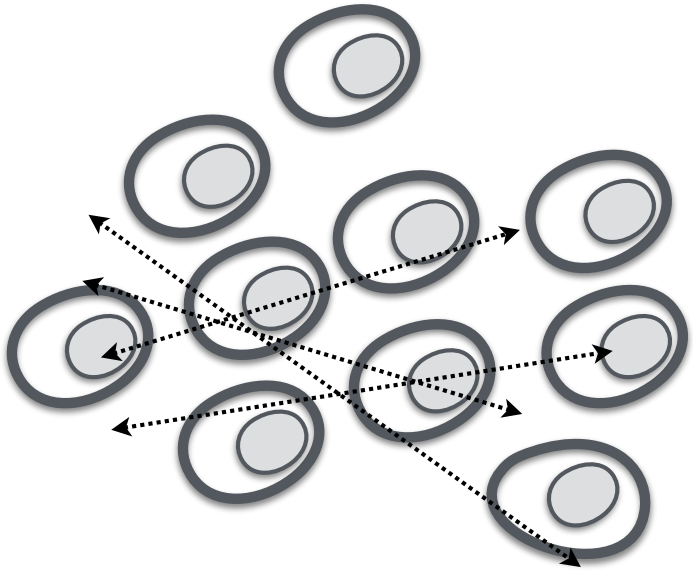


Single-cell regulatory genomics



[Kelsey et al., Science (2017)]

Single-cell multi-omics



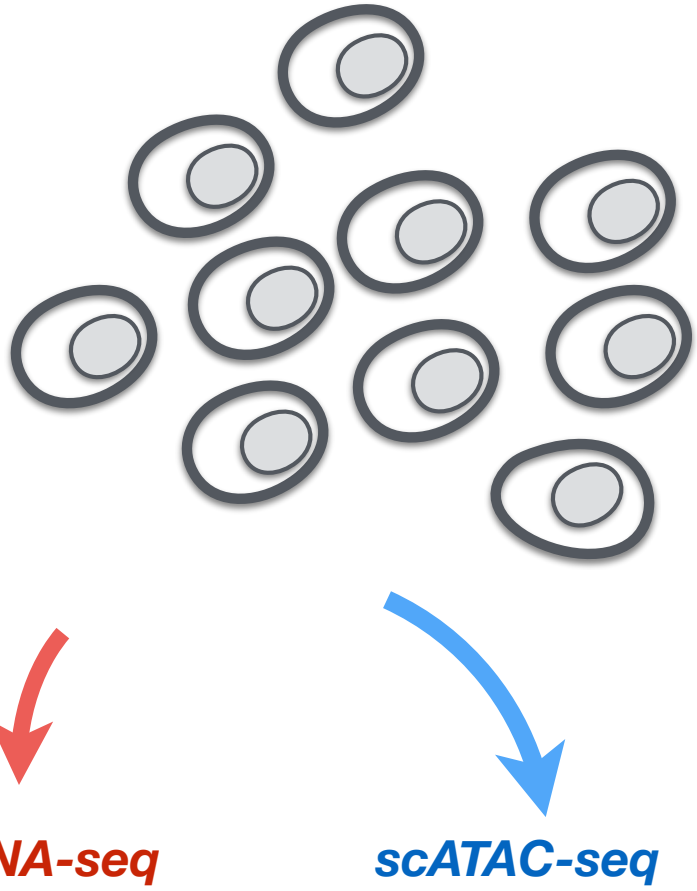
scRNA-seq

transfer learning
find best mapping between cells from different assays

scATAC-seq

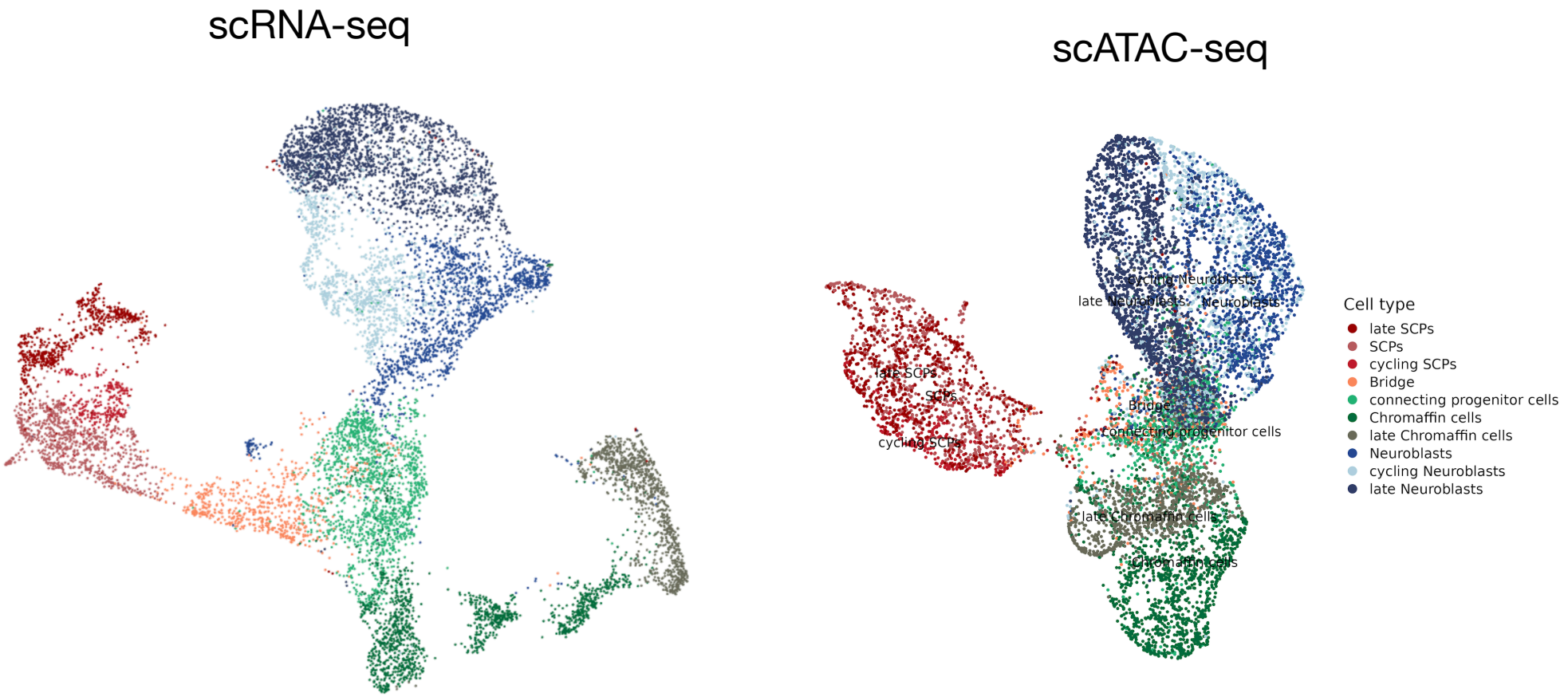
IPMB
Institut für Pharmazie und Molekulare Biotechnologie
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Single-cell multi-omics



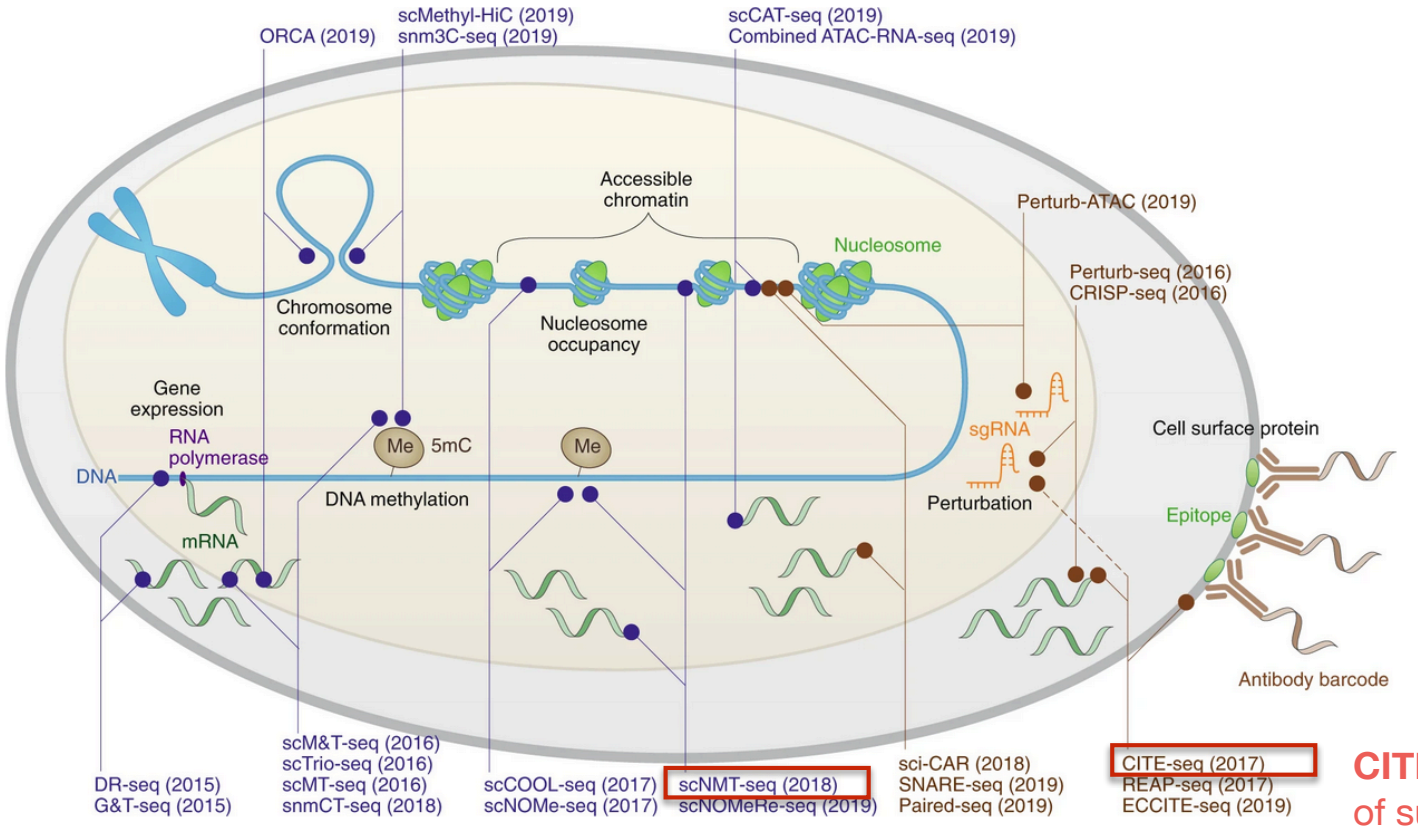
Accessibility & expression
[Cao et al. 2018]
[Clark et al. 2017]
[scCAT (Liu et al. 2019)]

single-cell Multiome: ATAC / Expression



cluster structure is slightly different between scATAC and scRNA!

Single-cell multi-omics



CITE-seq: identification of surface proteins + scRNA-seq

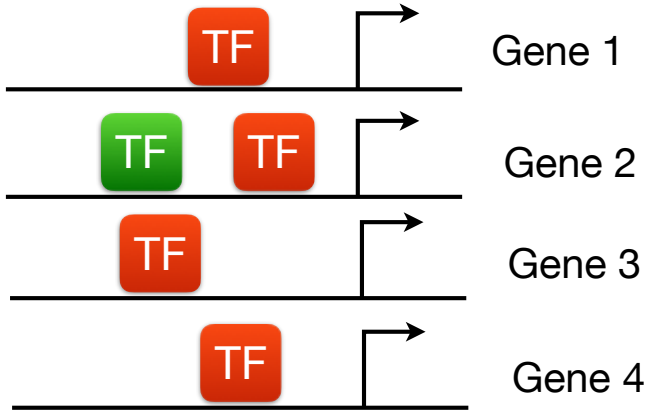
scNMT-seq: identification of DNA-methylation + accessible DNA

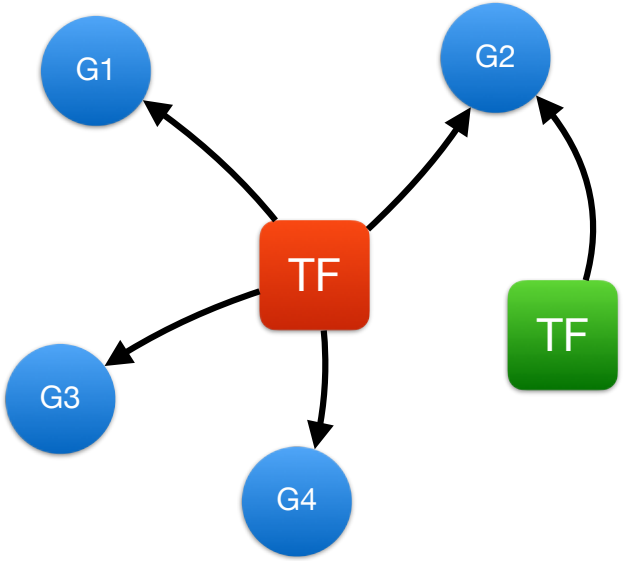
[Zhou et al., Nature Methods (2020)]

IPMB
Institut für Pharmazie und Molekulare Biotechnologie
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Gene regulatory networks (GRNs)

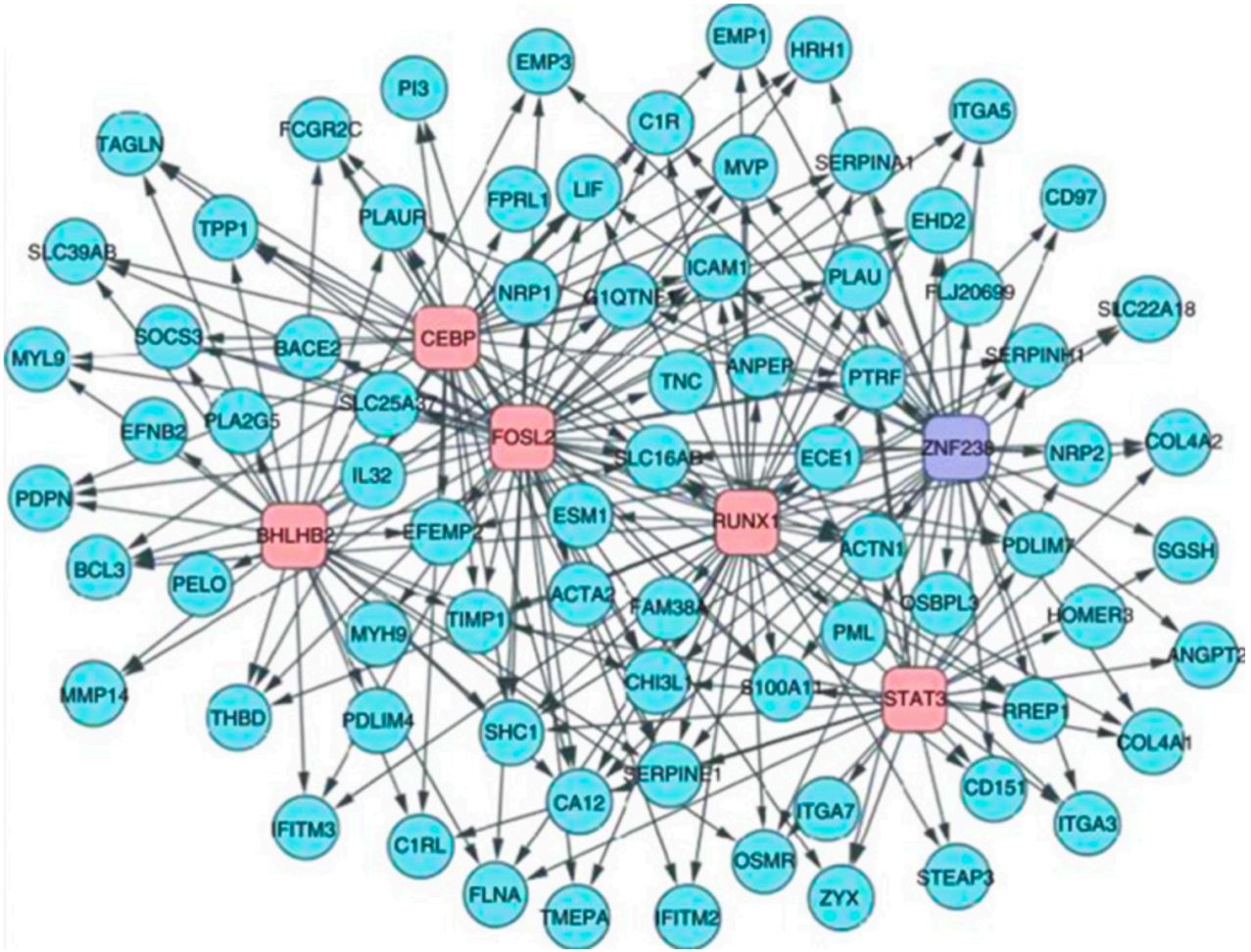
- Transcription factors (TFs) can regulate **multiple genes**
- Genes can be regulated through **multiple transcription factors**
- These TF → target interactions are **cell type and context specific**!
- All interactions form a **gene regulatory network (GRN)**





http://califano.c2b2.columbia.edu/modeling-cell-regulatory-networks

Gene regulatory networks (GRNs)

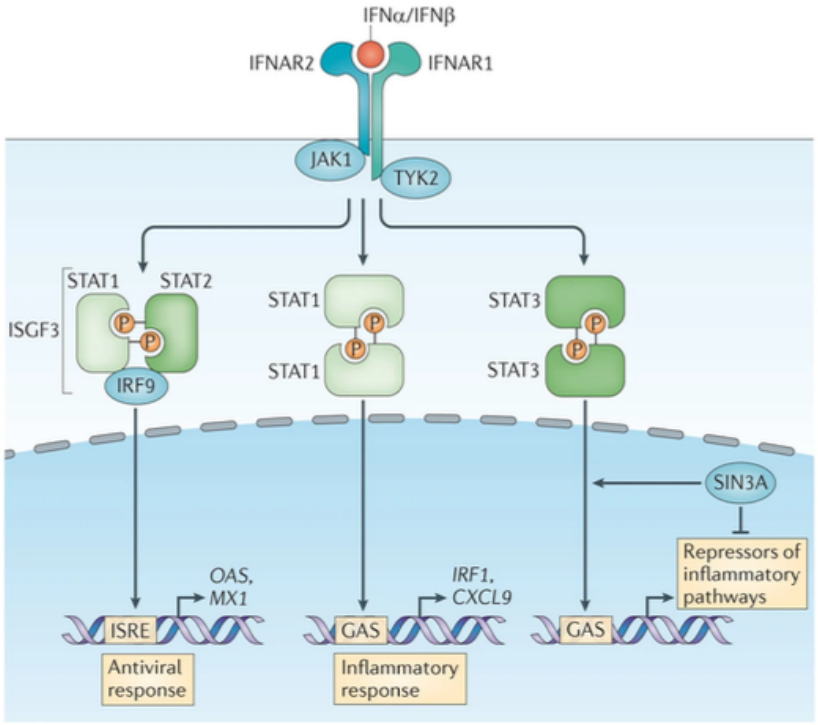


Gene regulatory network controlling the mesenchymal signature of high-grade glioma

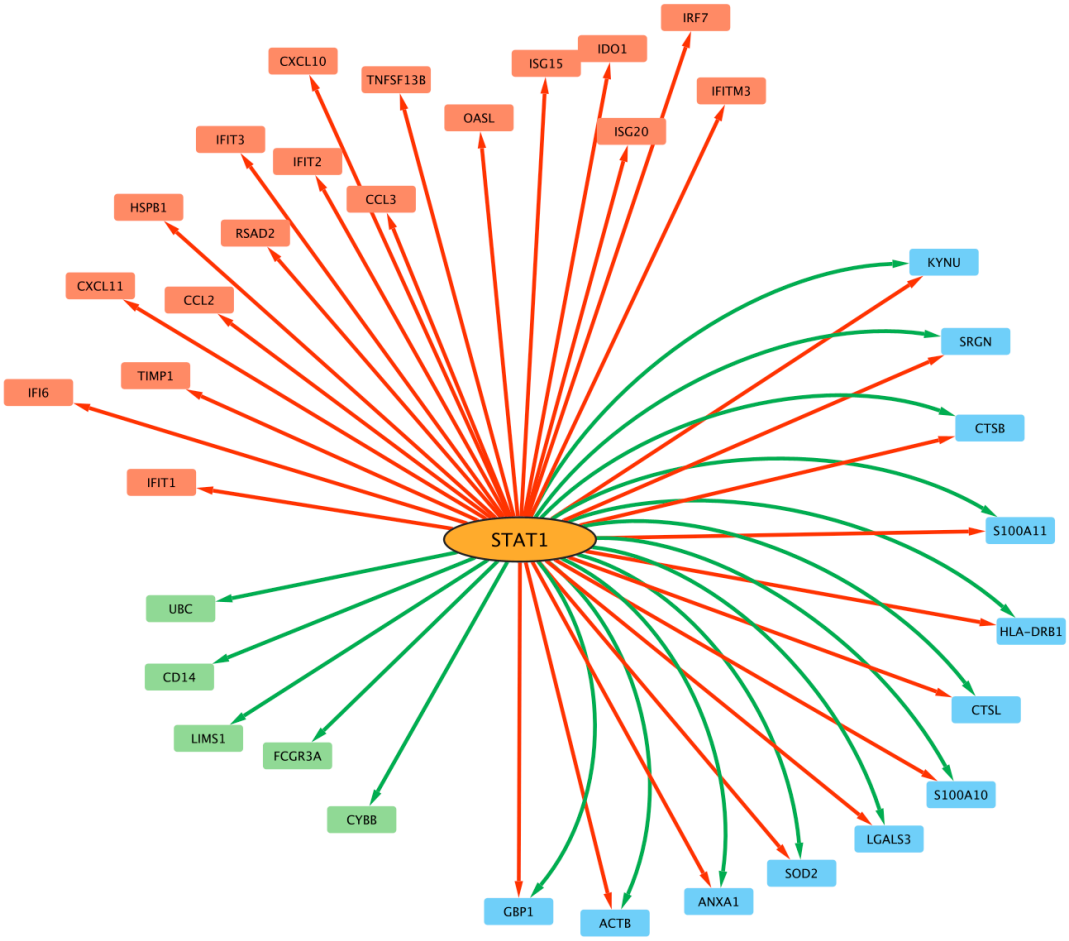
http://califano.c2b2.columbia.edu/modeling-cell-regulatory-networks

Context dependent GRNs

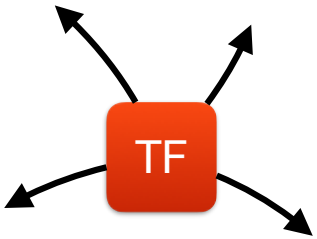
Effect of interferon treatment on cells



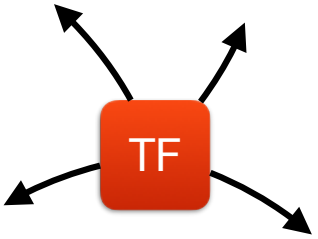
Interferon leads to rewiring of GRN



Computing TF activity



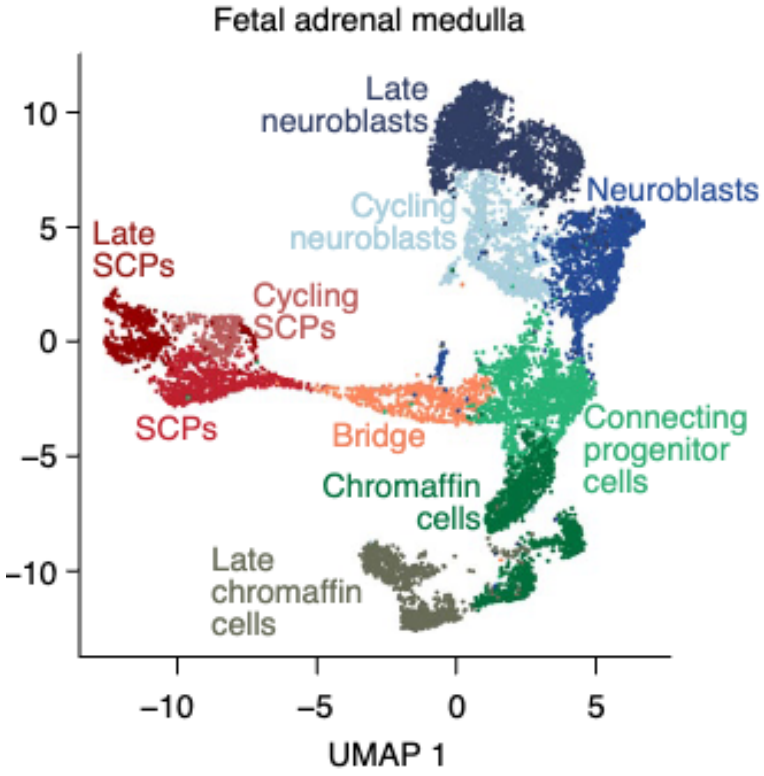
Target genes are lowly expressed
→ transcription factor is lowly active

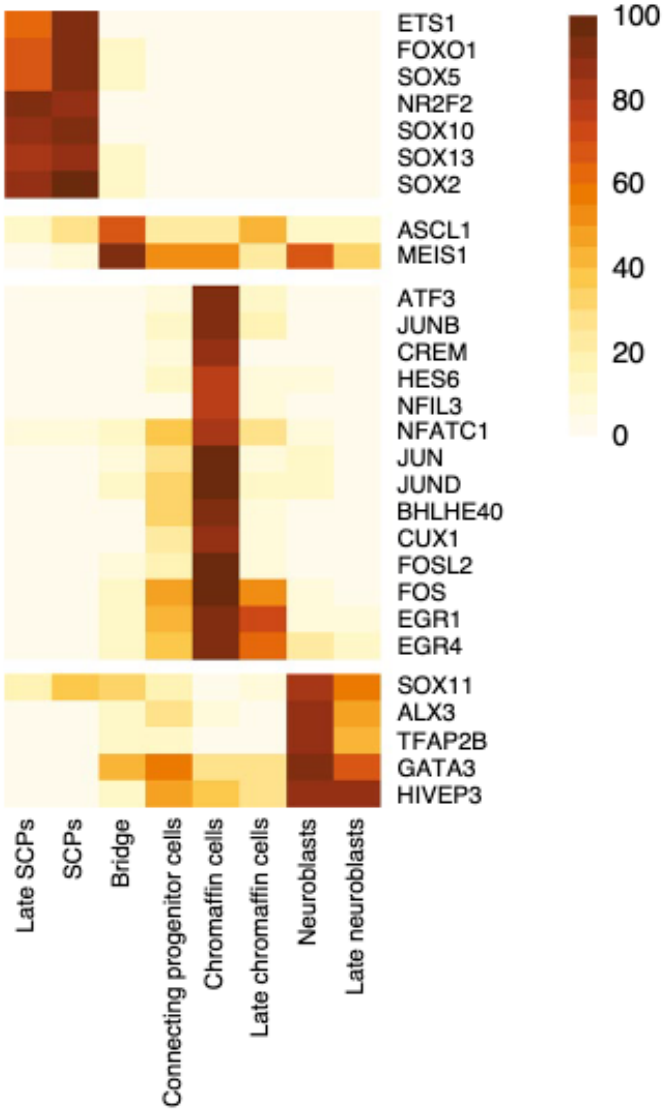


Target genes are highly expressed
→ transcription factor is highly active

- Transcription factor activity can be defined through the expression of the target genes of a TF
- This is a proxy for the protein activity (e.g. phosphorylation state of the transcription factor)

Computing TF activity



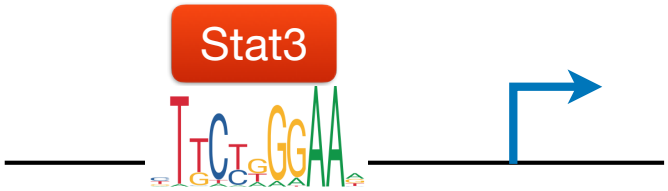


Transcription factor activity highlights cell type specific transcription factors → master regulators

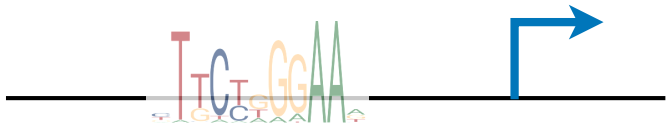
3. reconstructing gene regulatory networks

Carl Herrmann

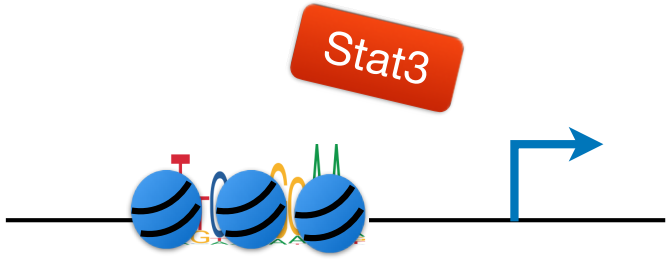
Reconstructing GRNs



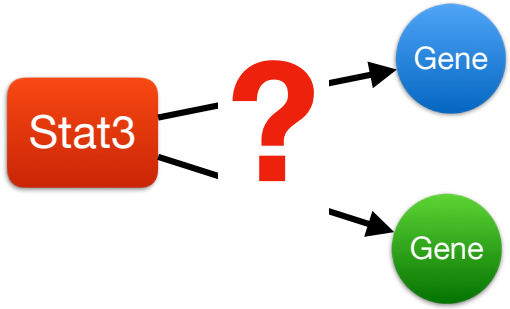
real binding site



non functional binding site

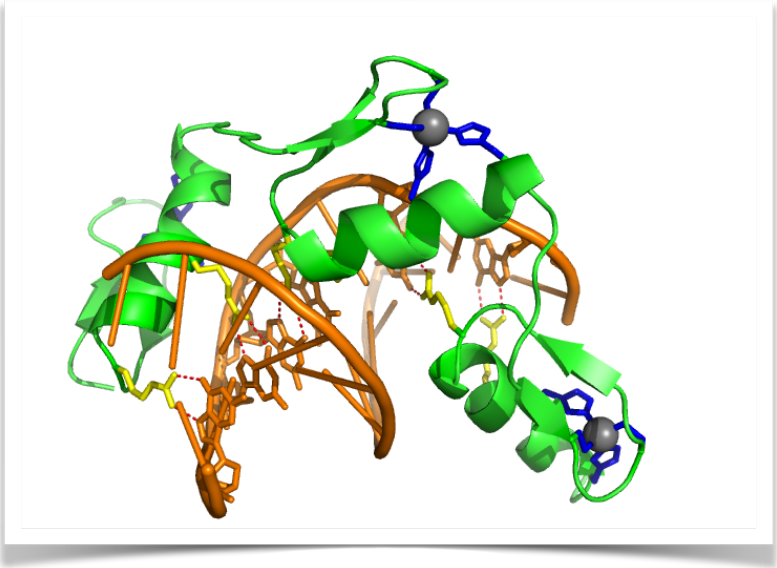


non accessible binding site



distal binding site

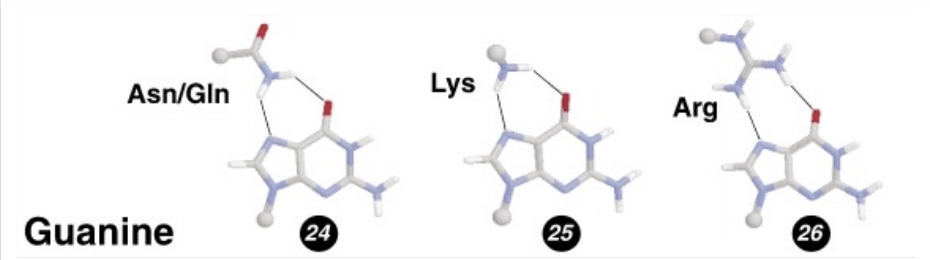
Protein DNA interactions



- majority of protein-DNA interactions for TF occur through a **alpha-helix** fitting into the major groove (=DNA binding domain)
- **hydrogen bonds** with specific bases
- stabilization of the protein-DNA complex is ensured by additional structures (helix, beta-sheet) via **van der Walls** interactions

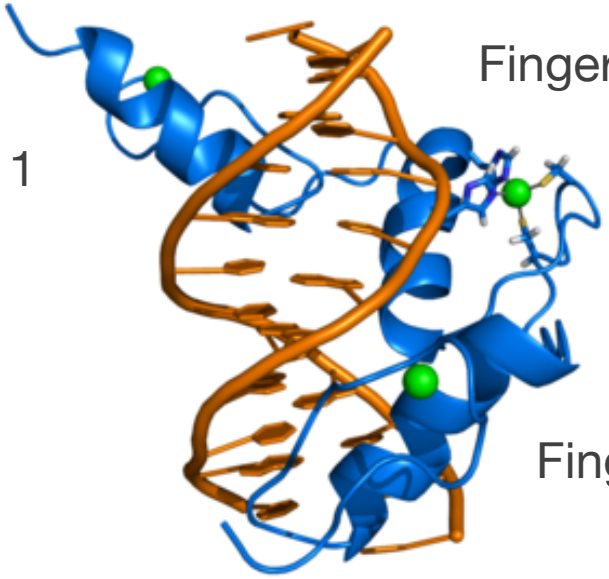
Amino acids	Mode of interaction	Recognised base
Hydrogen bond		
[ARG, LYS]	Multiple-donor	G/complex
[HIS]	Multiple-donor (bifurcate)	G
[SER]	Multiple-donor (bifurcate)	G
	Acceptor+donor	complex
[ASN, GLN]	Acceptor+donor	A/complex
[ASP, GLU]	Multiple-acceptor	complex
van der Waals contacts		
[PHE, PRO]	Ring-stacking	A, T
[THR]	Methyl contact	T
[GLY, ALA, VAL,	-	many (non-
LEU, ISO, TYR]	-	specific)
No base contact		
[CYS, MET, TRP]	-	-

[Luscombe et al., NAR (2001)]

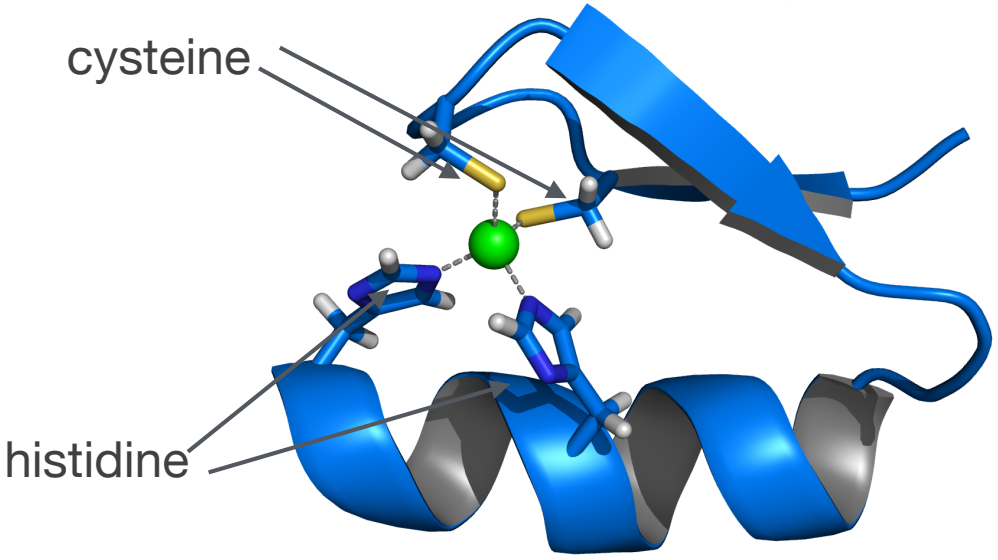


[Cheng et al., JMB (2003)]

Structural family: Zinc coordinating

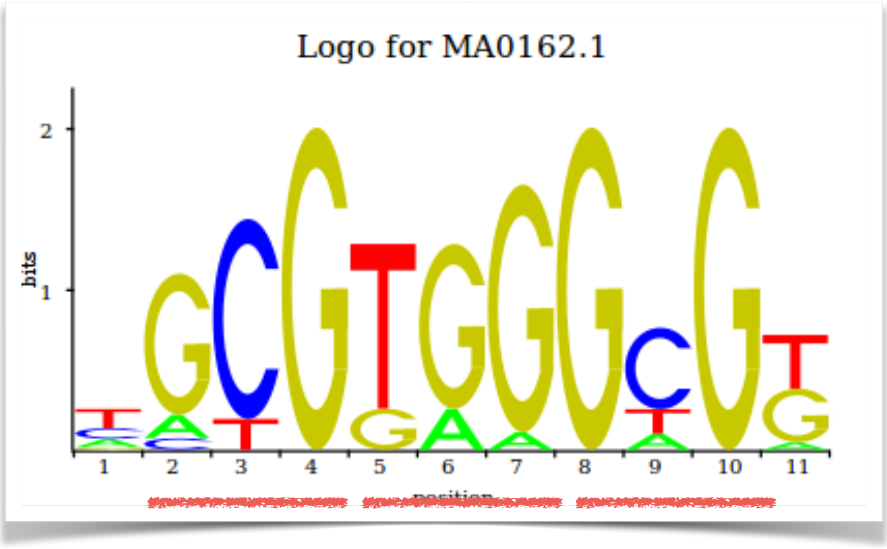


Egr1/Zif268



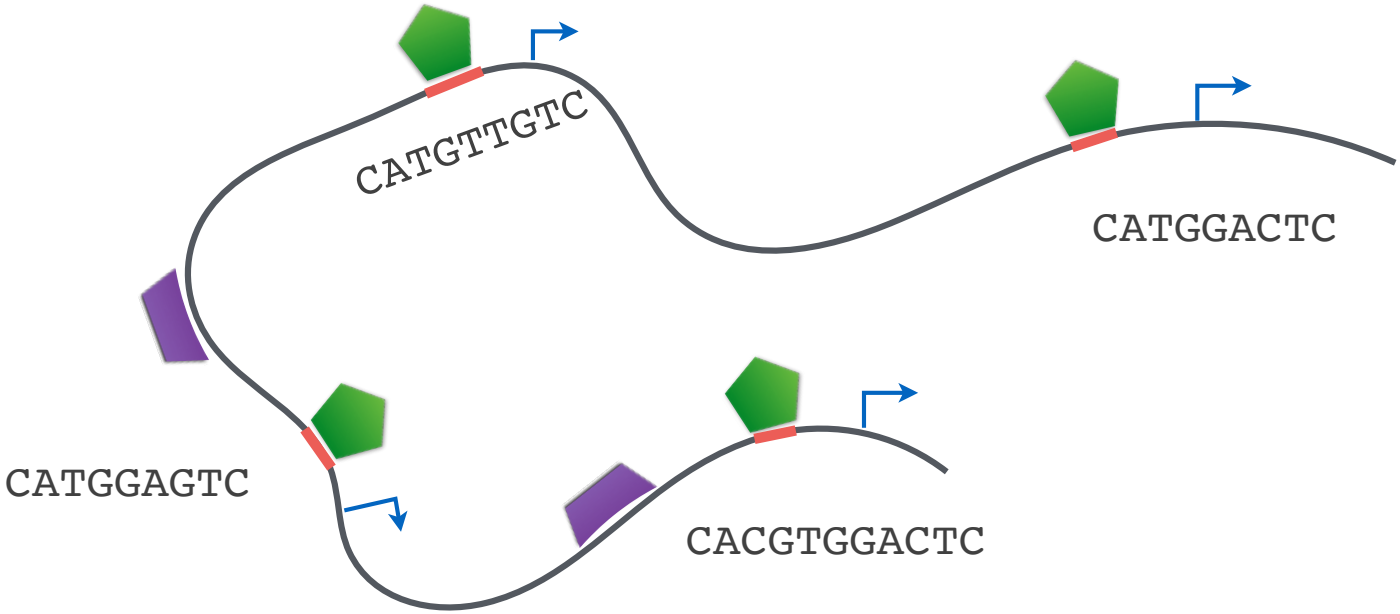
Cys2His2 Fold (“Zinc finger”)

→ one of the most common family of transcription factors in mammalians



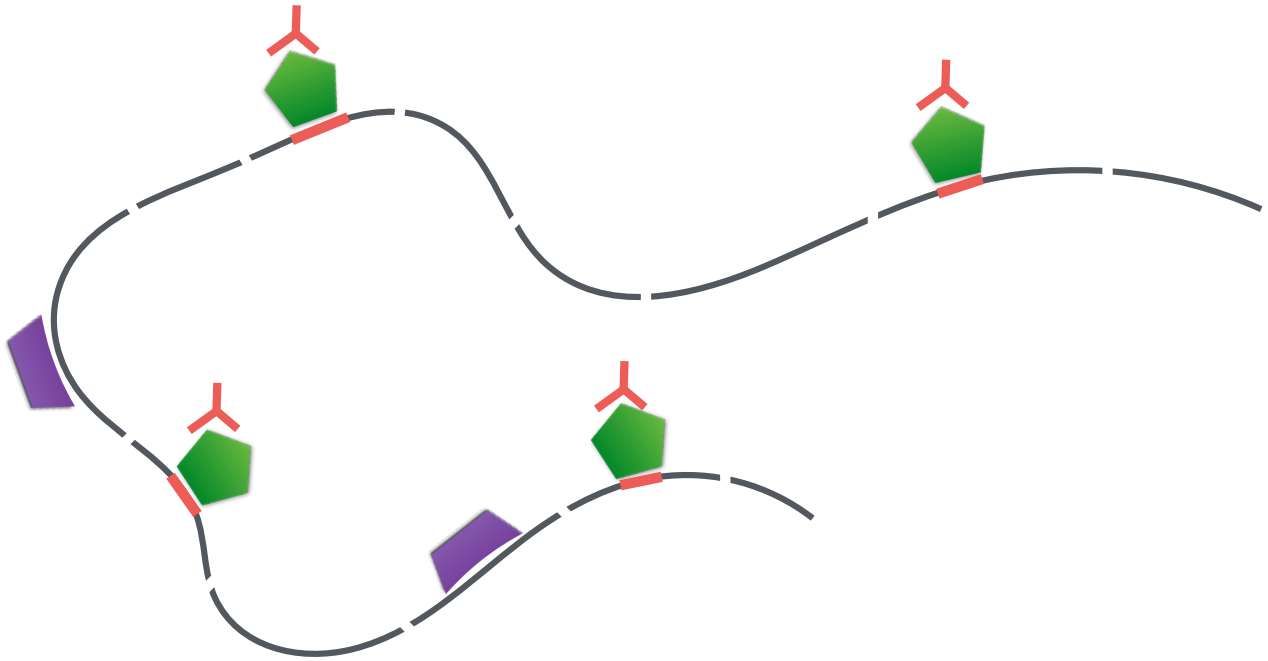
Finger 1 Finger 2 Finger 3

Transcription factors



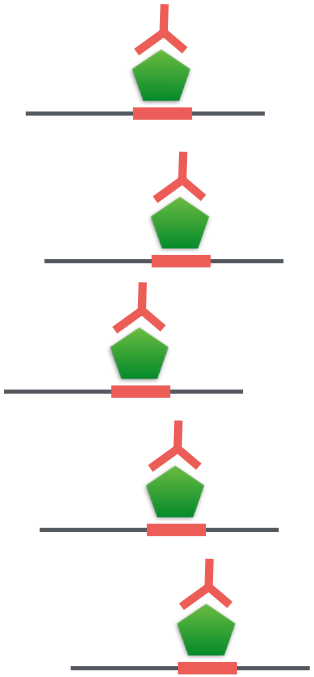
Identification of binding sites using ChIP-seq

1. Binding with transcription factor specific antibody
2. Sonication or fragmentation of the chromatin

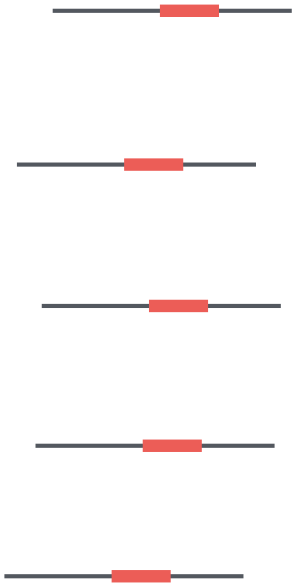


Identification of binding sites using ChIP-seq

3. Pull-down



4. DNA purification



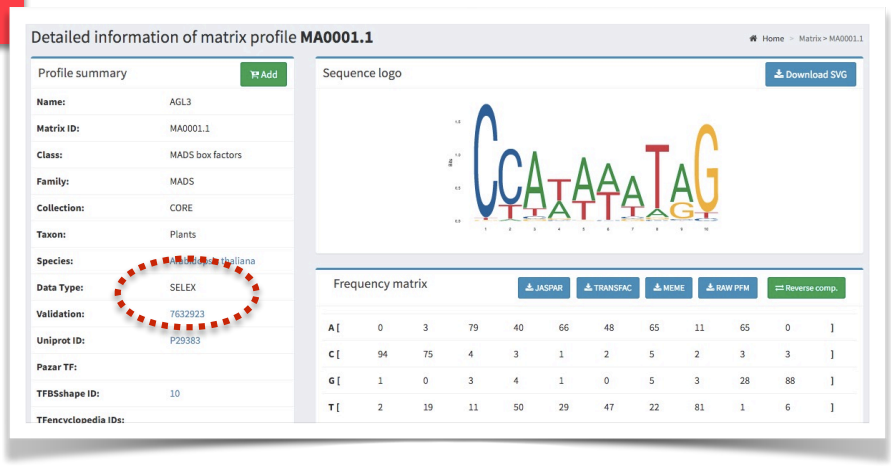
5. Sequencing

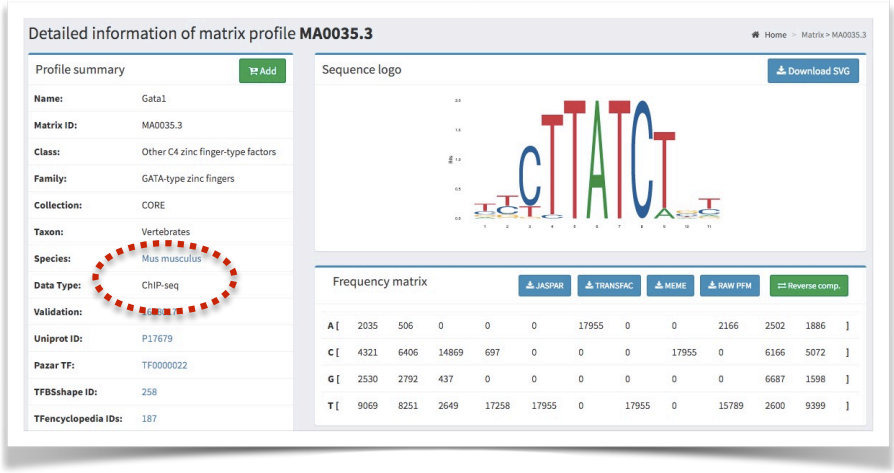
ChIP-seq sequences

ChIP-seq:

real binding site is hidden in much longer sequence

Different sources





Sequences for model MA0001.1

Site	Occurences
acaaCCATATATAGtagccactgaat	1
ccaccCCATATATAGtagtgcgggtggtg	1
CCATAAATAGataggcagactgtcgctgt	1
gtaaacataCCATAAATAGga	1
ttcaagaaactgCCATAAATAGcgat	1
tagaggtttttgtgCCATAAATAGgt	1
ccCCATAAATAGgaatatcgggatga	1
ttgccattaatagattataCCATATATGG	1
tatcaacaacgataccaacCCATATATGG	1
tttCCAAATATAGaaggtgtggaaag	1
tCCAAATATAGtaaaatcgagtcgcggat	1
gactggggcCCAAATATAGcatgttc	1
atcattagcttttacttaCCATAAATGG	1
attcttttgCCATAAATGGtaactcg	1
CCATAAATGGcaagtctgtcgaataacgg	1
cCCATAAATGGcagggtattagcacg	1
CCAAAAATAGatatgtgtcgtaacagctt	1
CCAAAAATAGggggacaatggaagtgggg	1
CCAAAAATAGgccagacgtgtttacaacg	1
CCAAAAATAGttaaataatgtcatacatt	1
ctacaccttCCAAAAATAGtaatct	1
ttgCCAAATATGGggttagagtgttc	1
gtctttaCCAAAAATGGtgatcctgt	1
ttgCCAAAAATGGagcgtttaccaat	1
atccaCCATTTATAGaaagttcaggaggc	1

ChIP-seq: real binding site is hidden in much longer sequence
→ lower resolution

Sequences for model MA0035.3

Site	Occurences
atcattttcttattatcttttttgtccttttgtatgtgcggtgtaaaatTTCTTATCTGTaaatggcgcaggctcttcctcttttttctactggtcttttggaagaaaaag	1
gcaaaacaatcttgacaccatgtaggcgatttatgcctctaagtcttggtTTCTTATCTGTgaaatgggggaggaaaatgcttgcagtcggagcctgaccacaaactgact	1
gttgatgatgatgatgttgtctagagaaattttaacttttaaaaagcactTCCTTATCTTTtacatttttctgctgtgaactaggtatttcttatctgtgtgttcctcccc	1
ttaaaaagcacttccttatctttttacatttttctgctgtgaactaggtatTTCTTATCTGTgtgttcctccccatccctggcttctggtgttttaccccatatctaagatag	1
catcagcaagcagttcctctccctcattctgcaaatgtgtgagttctgctTTCTTATCTGTctcttatcacttcacagcagtgtgaccctgaaccagttccctgtaaaaga	1
agttcctctctccctcattctgcaaatgtgtgagttctgctttcttatctgtCTCTTATCACTtcacagcagtgtgaccctgaaccagttccctgtaaaagactagggaggat	1
gctgctcagaccaagaatgggagaagtggggctgggagacaaagaaattctcTTCTTATCTGTgtaggatacacatgtctgagaggcagagatcagatgtcctgccctctgat	1
aaacggggctacaagagcaagactcatacacactttgattgcacaactccTTCTTATCTGTtcagcacagctgacctgccgagggagtcaccatagtggttggctgtggcc	1
gtaaactacaaacaaaacaaaatggccatccggcttctttcttccttcggTTCTTATCTGTgacctaagctggcttgcctggaatttactaaatagactaggtcagccgtg	1
ggacagggagagaccagagaacaccacactaaatgaattctgccagacatTTCTTATCTGTggctcacactctttatctatccataaagtcatgcacaaacaaaggacaaa	1
acaccacactaaatgaattctgccagacatttcttatctgtggctcacacTCTTTATCTATccataaagtcatgcacaaacaaaggacaaaaatggagaagtgaaactgtg	1
gtccacaaaccttgagctttcacgctgtgtataaactctgcattggacttTTCTTATCTGTcatgcccagggcaaggtaggactcttatcaactatgaacactagcatgac	1
gtgtagctcaacataaggacacagcacaggaagttggctcatgtttgataTTCTTATCTGTtgaaagttatgttgggaggacagtactattaagataaccacaagagcaag	1
tcttggaaattatctttcctgtcacagaatagagcatcataataacgtgtTTCTTATCTGTcctgtccaccttcagcaagaaacctcagaaagtatgttttctattgcta	1
aggctacagaacttgcatgtcactggcggtggtaaataatcatctcccgtTTCTTATCTGTctgtgggtttcctctgtggttcagtgacggggcttgtcagtcaggtgaga	1
actcatgcccagcaggcacgctgctgtctgcgtcttttcctcgataacagTTCTTATCTGTgttgtgcaatagctgtgatataatttgtagcagaatgggaggttgcataac	1
ctccccagcagcttgccttaaactaggtaagaactttccattcttaagcaTTCTTATCTGTggaggctttaggggtacaccctaacaacatctcttctggacacacctaa	1
atccgttgagtggaagcatgctctggtaaattgcattacctcagccccaTTCTTATCTGTgtgtcacacttgtcttataggaaataagcagtgtacagagaaatgtgccc	1
aataacaccttactgtggtcagtatttattgtctacatgagcctaagacgTTCTTATCTGTggtcccacagtgtgaaatcctttgaaacttgcattgcccaagagagtgct	1
aggcttccatggcagtccatcgctgctgagggaggagttactctgttctcTTCTTATCTGTaatacacaaacatgagttaaaatgagagctaagctgcaaaggacaacaaa	1
agccccatttcagagtatactccatatcccctcagcattagaaaaaaatcTTCTTATCTGTaactctaattgctcacaagccaccaacattttgccttatctaactggctt	1
aggtttcagctctacacttcctgctcggtacctagttcgaggcctccattTTCTTATCTGTaaaactggtctgctgaggcagtttgtcaagtcatcagcggagtcttctct	1
gagctattagcatctttgactaagcaccacaaggcagctgctgcttctgcTTCTTATCTGTcttgttttaaaatgatgactcccctcaggaagacagaatctgagtctttg	1
atctctactctacagataagaaagccggcctagataggttaagtgccttgCCCTTATCTACagtttccagcttgcaatacacaggatctgacaagcatattaattcttcga	1
agctggaaactgtagataagggcaaggcacttaacctatctaggccggctTTCTTATCTGTagagtagagattaaaatttcattcccattcttcctgggagtcagcattag	1
aactgcagagcttatgtcagacactctcatttaaaccacaacagccaattTACTTATCTGCcttctaaggacctagtgcagaacagataagaaggaacaaaagtagctcag	1

Predicting binding sites in sequences

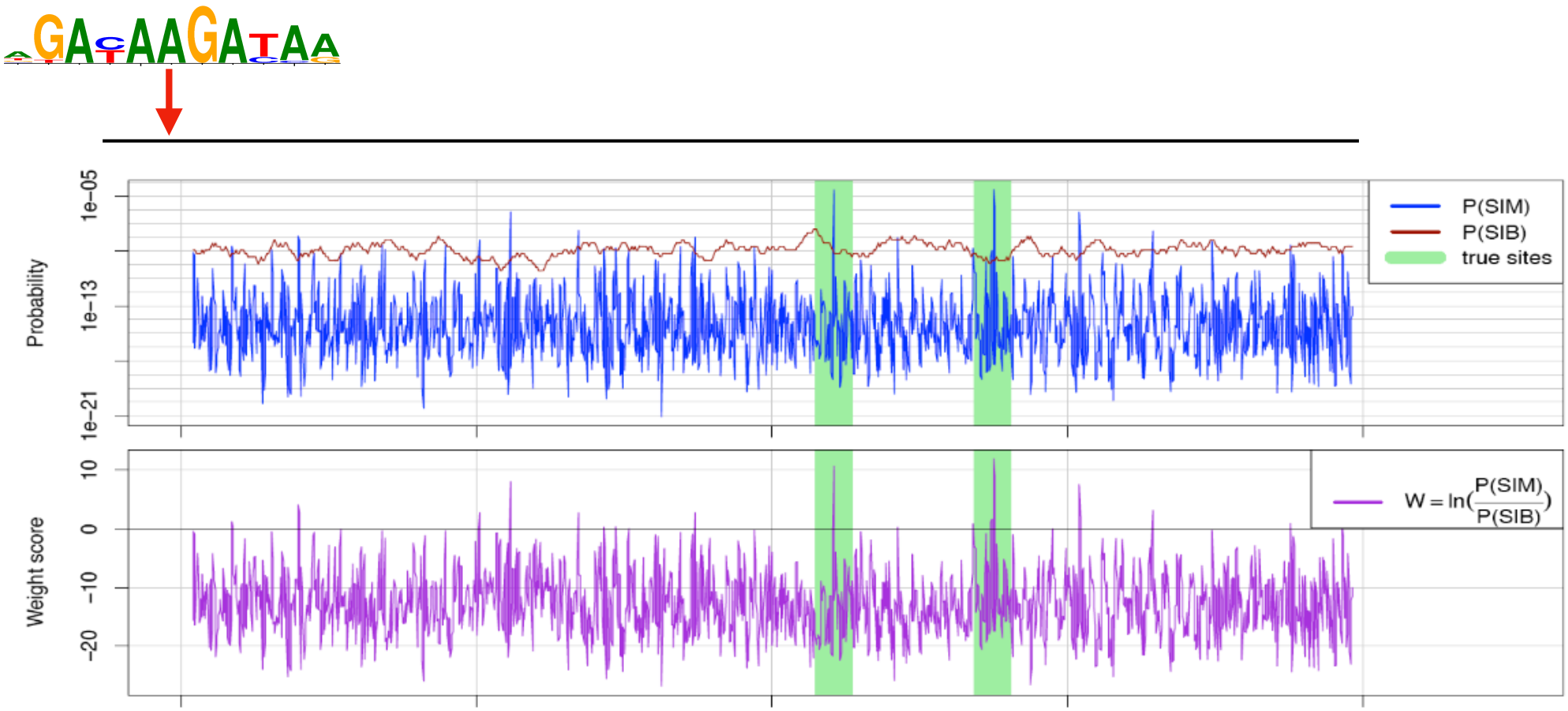
a|	0.55	0.02	**0.95**	0.02	**0.95**	0.95	0.02	**0.95**	0.02	0.88	0.75
c|	0.08	0.02	0.02	**0.48**	0.02	**0.02**	0.02	0.02	**0.22**	**0.08**	0.02
g|	0.15	**0.88**	0.02	0.02	0.02	0.02	**0.95**	0.02	0.02	0.02	**0.22**
t|	**0.22**	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02

T G A C A C G A C C G

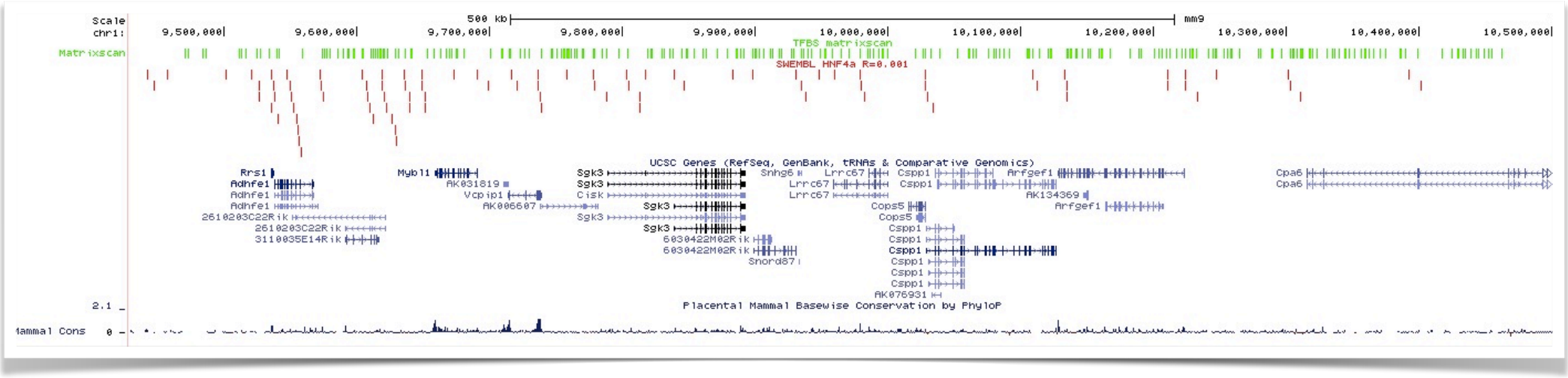
p(S|M) = 0.22 * 0.88 * 0.95 * 0.48 * 0.95 * 0.02 * 0.95 * 0.95 * 0.22 * 0.08 * 0.22

= **4.5e-6**

Predicting binding sites in sequences



Predicting TFBS on real sequences



- Predicting HNF4a on a 1 Mb portion of Mouse chromosome 1
- **259 predicted TFBS using the TFBS motif**
- **72 real binding events HNF4a ChIP-seq peaks (red)**

→ Many false positives / false negatives

Improving TFBS predictions

TFBS prediction suffers from a high degree of **false-positive** and **false-negative** predictions

by TFs *in vitro*. In fact the methods do detect potential binding sites, albeit not necessarily those of functional importance. By most accounts, the three orders of magnitude difference between true and false predictions is intolerable, resulting in what we choose to term the FUTILITY THEOREM — that essentially all predicted TFBSs will have no functional role. Fortunately, there are biologically motivated approaches to overcome this 1000-fold excess of false predictions.

[Wasserman & Sandelin, Nat.Rev.Gen (2004)]

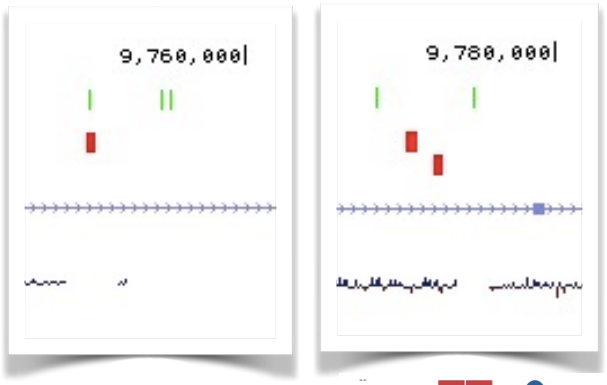
4. improving predictions

Problem of sequence-only predictions

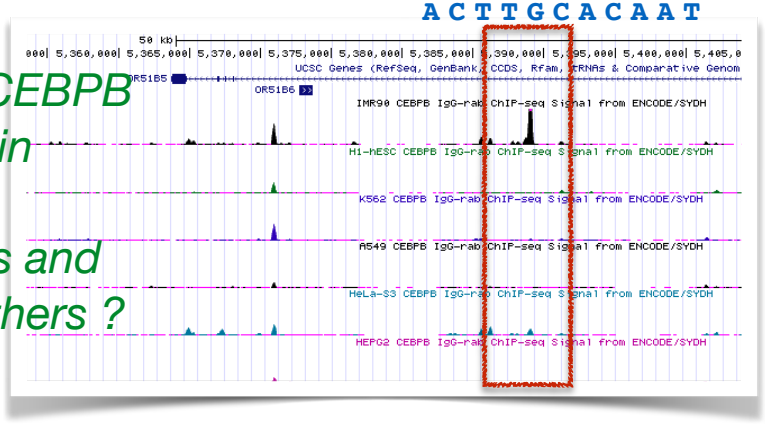
- **Large number of false-positive/ false-negative**
  the sequence looks like a binding site, but the TF is not binding!

- **Cellular/tissue-context not taken into account**
  a TF might bind in one tissue, but not in another (but the sequence is the same...)

predictions
real binding



Why is CEBPB binding in some cell-lines and not in others ?



Can we reduce/optimize the search space for regulatory elements?

IPMB
Institut für Pharmazie und Molekulare Biotechnologie
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Phylogenetic footprinting

- Tagle et al. (1988) : study of the promoter of globin genes in vertebrates identifies **conserved regulatory elements**

Phylogenetic footprinting

The pattern of mutations that have occurred during evolution is an excellent indicator of functional constraints. Genomes continually undergo mutations, but the outcome of each mutation depends on its phenotypic effect. Mutations that are deleterious are generally eliminated by natural selection, whereas mutations that have no phenotypic effect (neutral mutations) or that are only slightly deleterious can be randomly fixed in the population (genetic drift). The consequence of this is that mutations accumulate much faster at nonfunctional DNA bases than at functionally constrained base positions. Hence, if one detects a sequence that has remained highly conserved during evolution, then it probably means that this sequence is functional (but the reverse proposal is not true: a sequence can be functional albeit nonconserved). Tagle *et al.* [31] proposed the term 'phylogenetic footprinting' to describe the phylogenetic comparisons that reveal

onal elements in homologous

hylogenetic footprinting is

h shows the comparison of

. This comparison shows that

rs) of divergence (450 Myrs

ete elements in noncoding

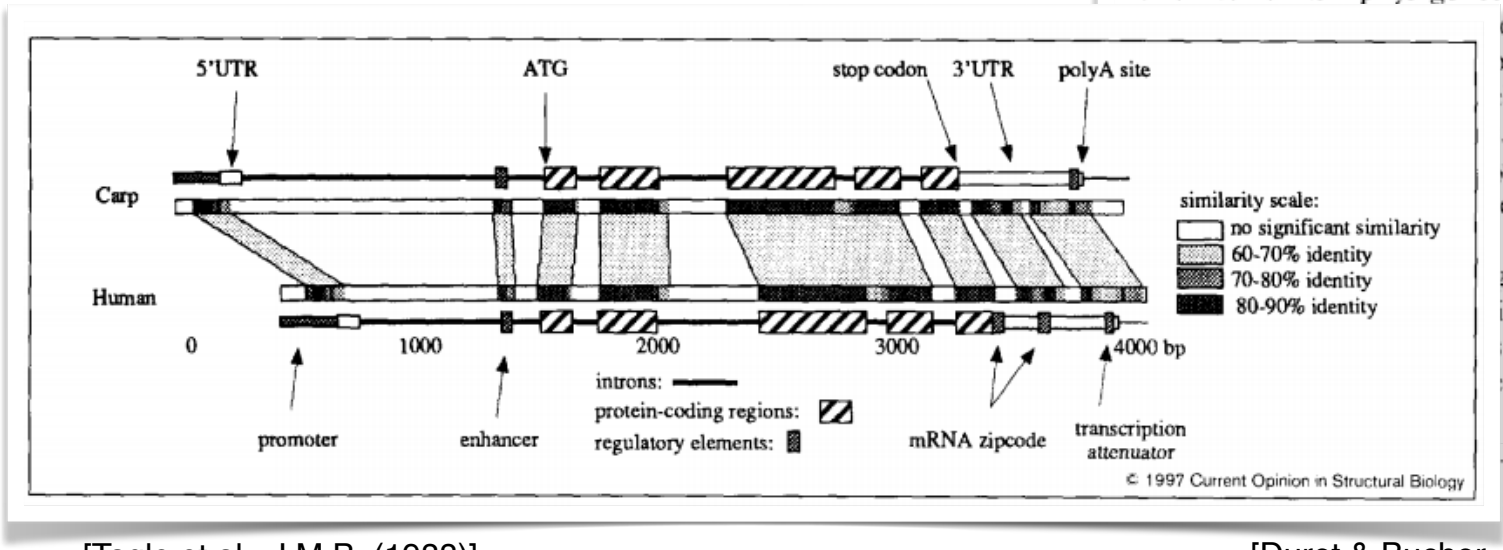
conserved. Indeed, these

gions correspond to essential

nvolved in transcription and

(Fig. 1). Thus, the simple

quences can reveal essential

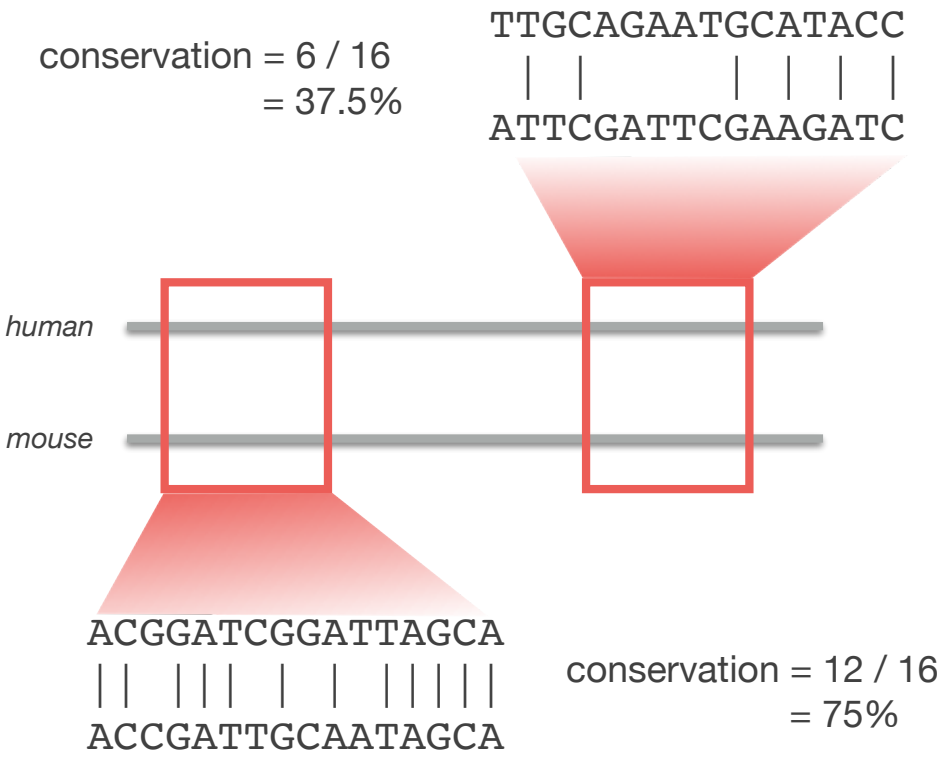


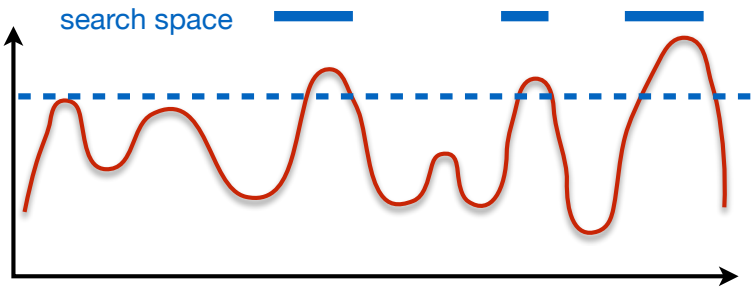
[Tagle et al., J.M.B. (1988)]

[Duret & Bucher, Curr.Op.Str.Biol. (1997)]

Phylogenetic footprinting

- Starting point : alignment of **2 orthologous regions** (e.g. promoter of orthologous genes)
- Compute the **conservation** inside a sliding window (number of conserved positions divided by length)
- TFBS search using PWM (fixed threshold)
- **Only TFBS inside highly conserved regions** are retained !
- *Choice of organisms to be compared is crucial !*

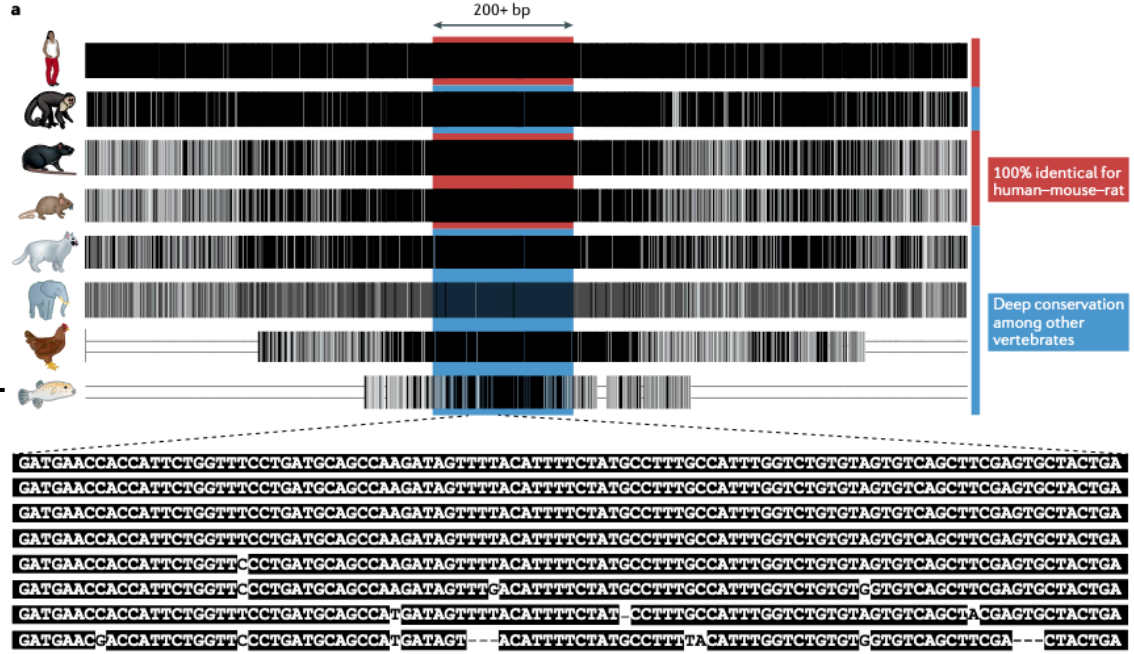


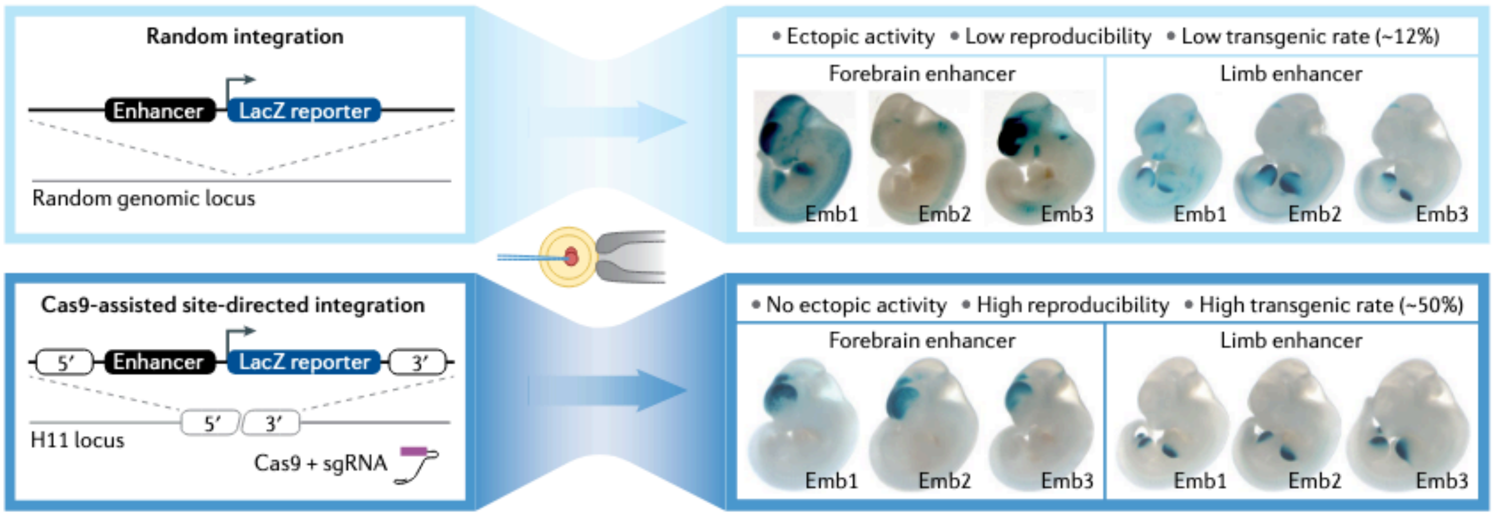


Deeply conserved non-coding elements

Ultra-conserved elements
perfect conservation over 200 bp between human and mouse/rat

~450 million years



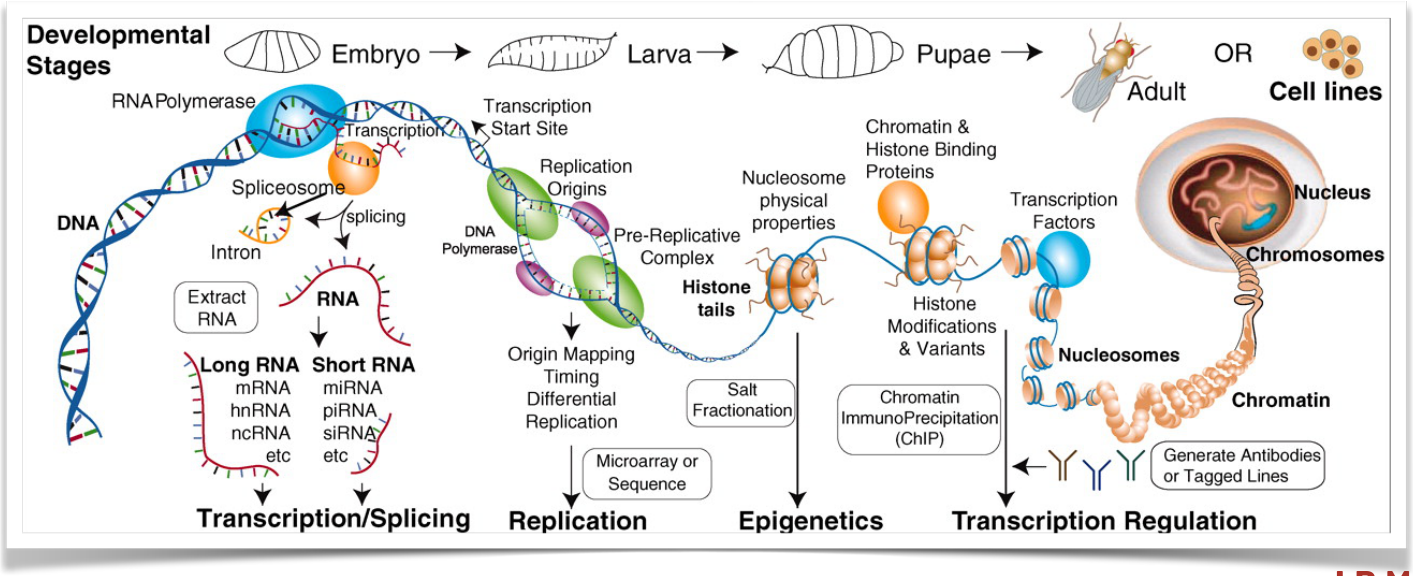


[Snetkova, Nature Rev.Gen. (2020)]

Important chromatin components

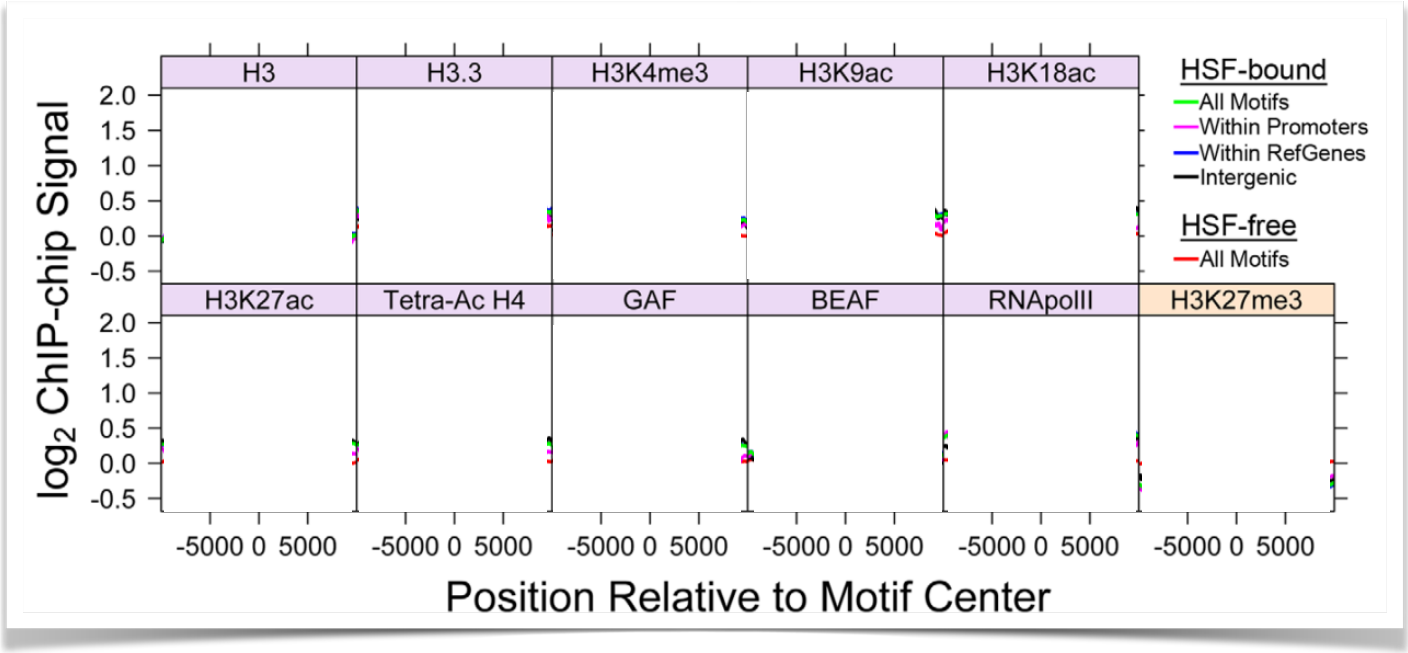
Some chromatin features are associated with open/accessible regions

- Methylation of histone 3 Lysine 4 (**H3K4me1**)
- acetylation of histone 3 Lysine (**H3K9ac, H3K27ac**,...)
- Presence of **Pol II** binding at active enhancers
- Presence of modified form of H3 → **H3.3**
- DNA accessibility (**DNAse hypersensitive** sites; **ATAC-seq**)



Motifs are not always binding events

- Compare **in-silico** TFBS to **in-vivo** binding event using ChIP data
- some bona fide motifs are not bound in-vivo : why ?
- example in Drosophila : heat-shock factor (HSF)
  - 464 ChIP peaks containing a HSF-motif ($p < 0.001$)
  - 708 unbound motifs (with $p < 5e\text{-}6$)



Bound sites have:

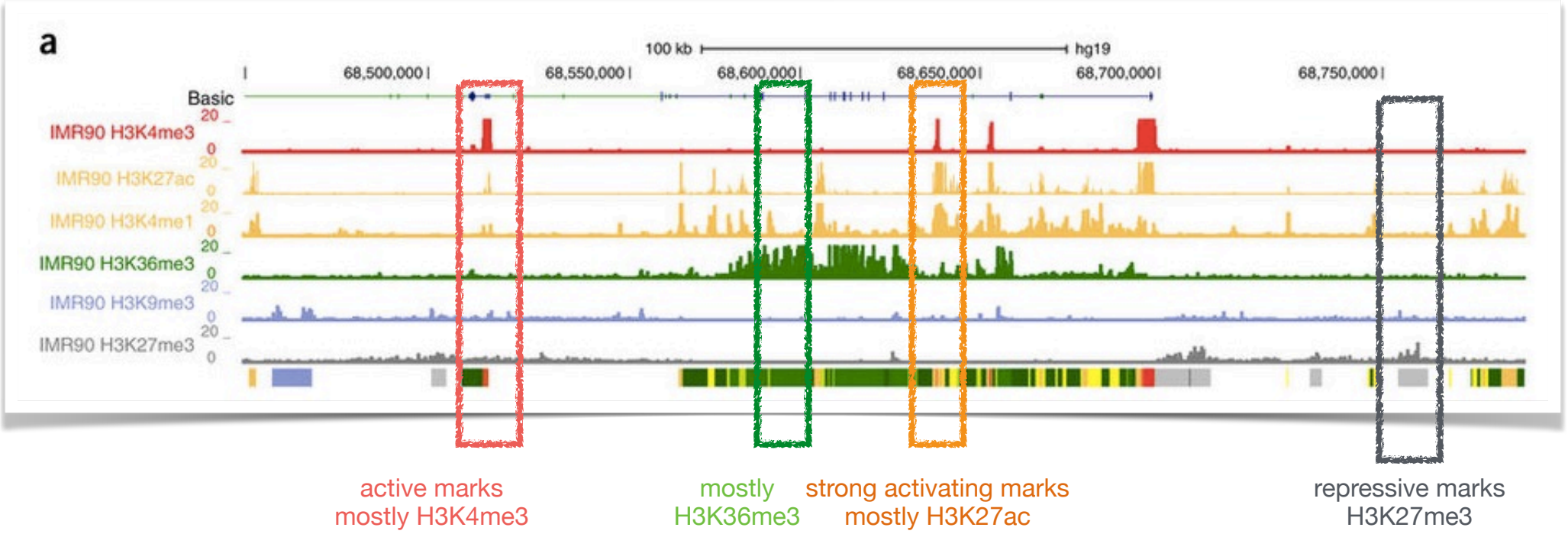
- **high levels of lysine acetylation**
- **high levels of polII binding**
- **low levels of H3K27me3 (repressive mark related to polycomb repression)**

[Guertin et al., PLoS Gen. (2010)]

Histone code

Mark	Interpretation
H3K4me1	activating mark; found at promoters and enhancers
H3K4me3	mark of active and open gene promoters
H3K27ac	mark of active promoters and enhancers
H3K36me3	mark of actively transcribed genes
H3K9me3	mark of closed heterochromatin
H3K27me3	polycomb associated mark → repressed chromatin

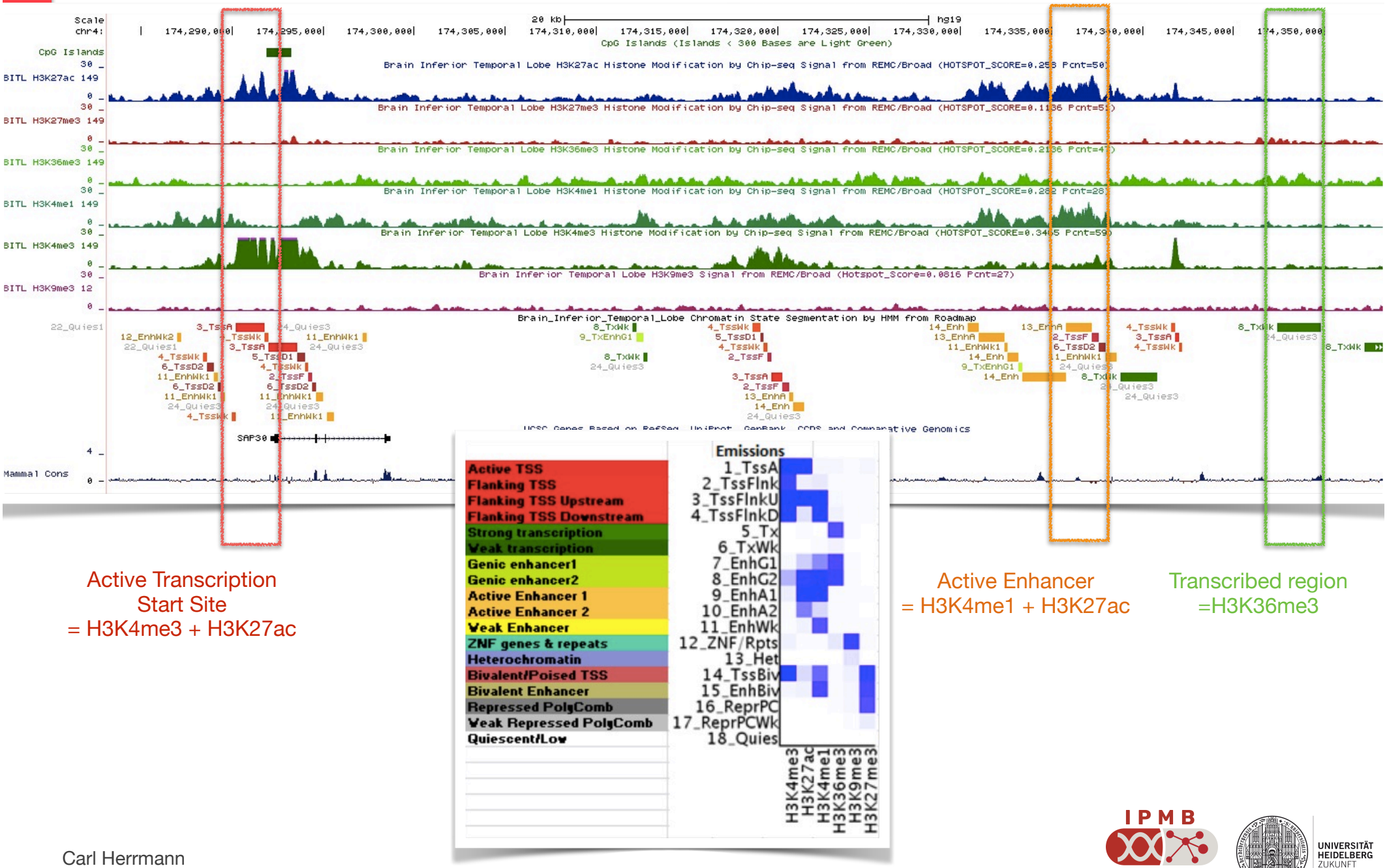
Histone code



Histone modifications appear to occur in specific combinations related to functional impact → **combinatorial chromatin states**

How can we define/annotate these **chromatin states** ?
→ **Hidden Markov model**

Chromatin states



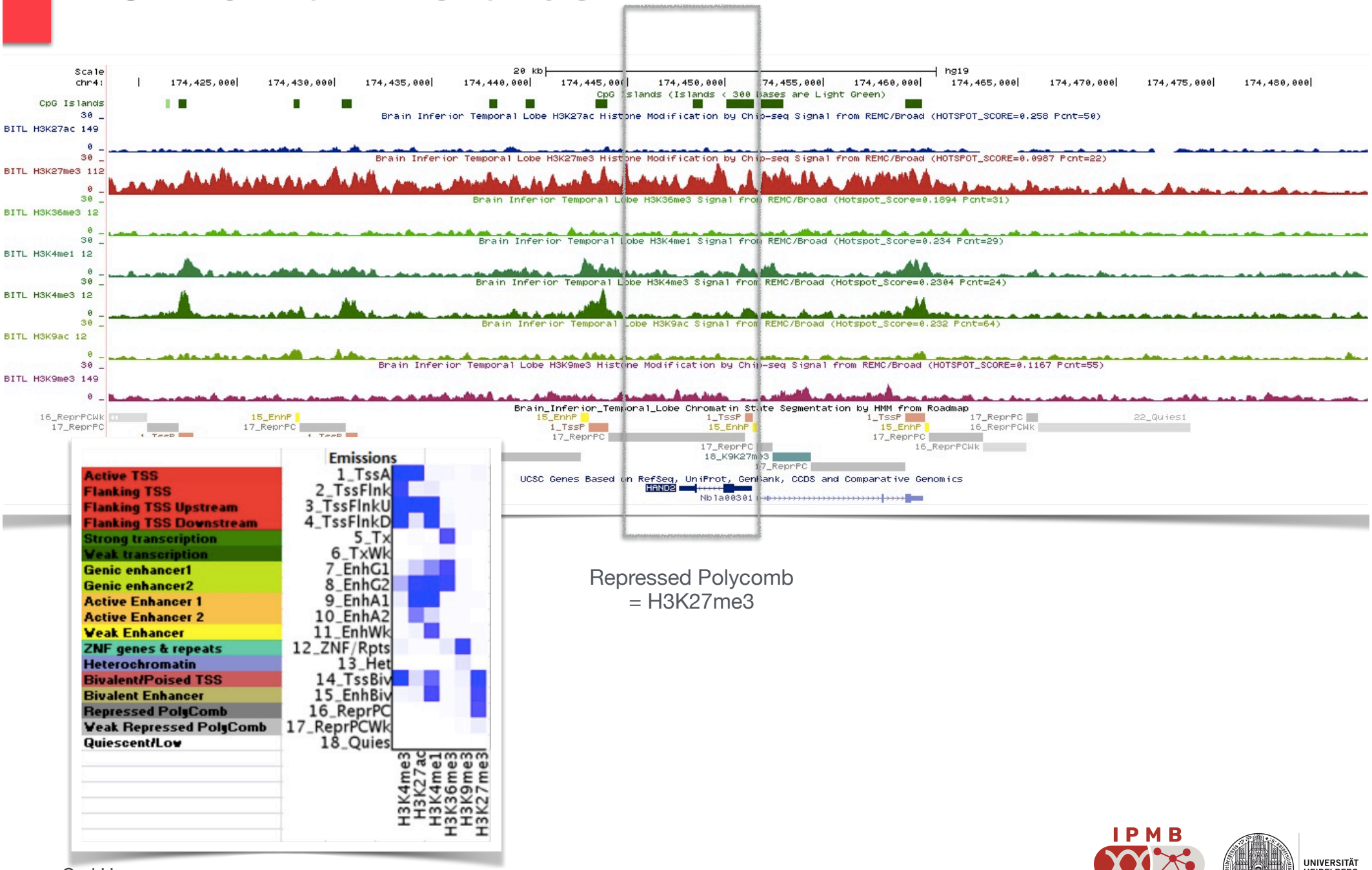
Active Transcription Start Site
= H3K4me3 + H3K27ac

Active Enhancer
= H3K4me1 + H3K27ac

Transcribed region
=H3K36me3

IPMB
Institut für Pharmazie und Molekulare Biotechnologie
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

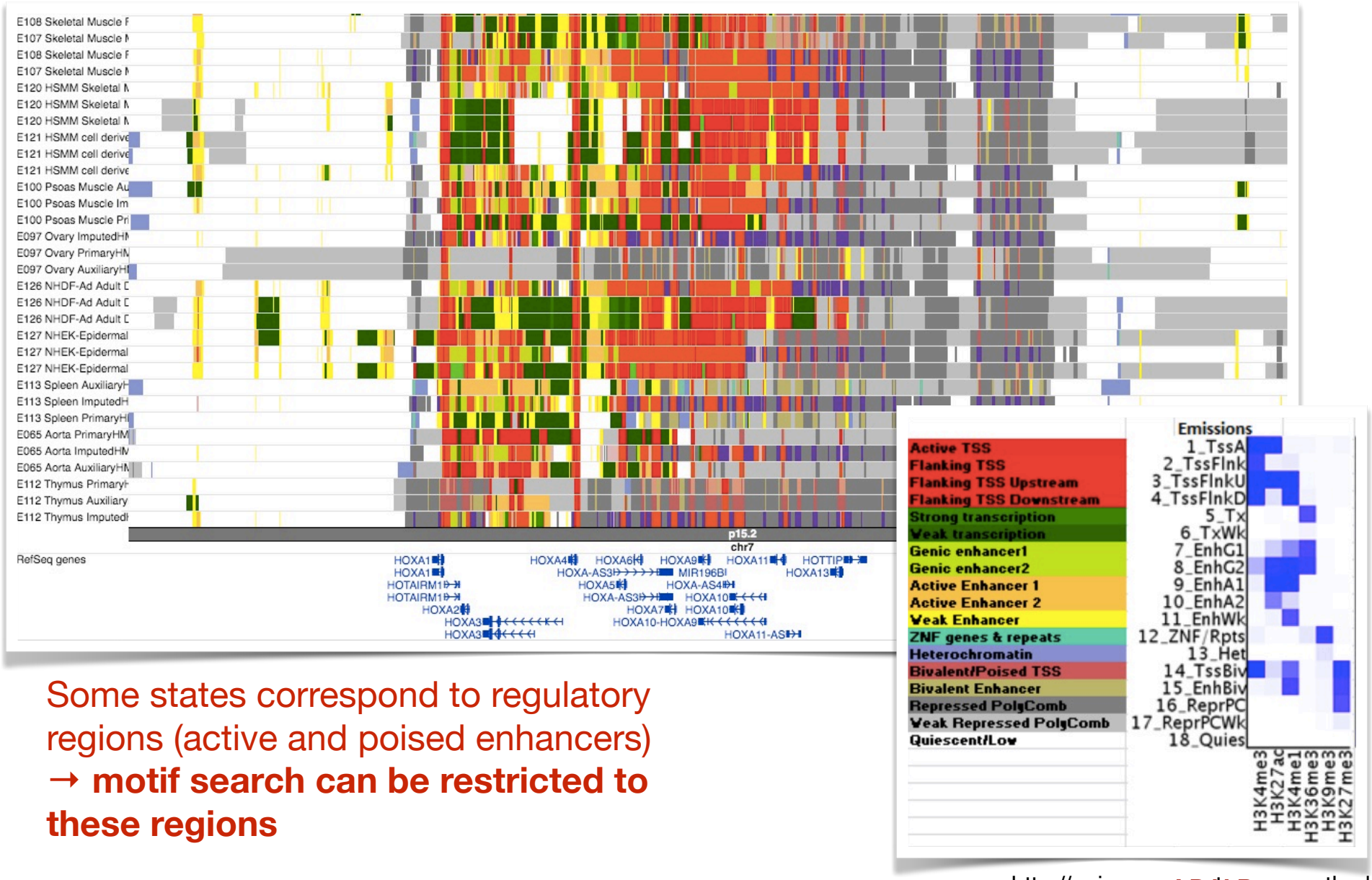
Chromatin states



Repressed Polycomb
= H3K27me3

Carl Herrmann

Roadmap chromatin segmentation in different human adult tissues

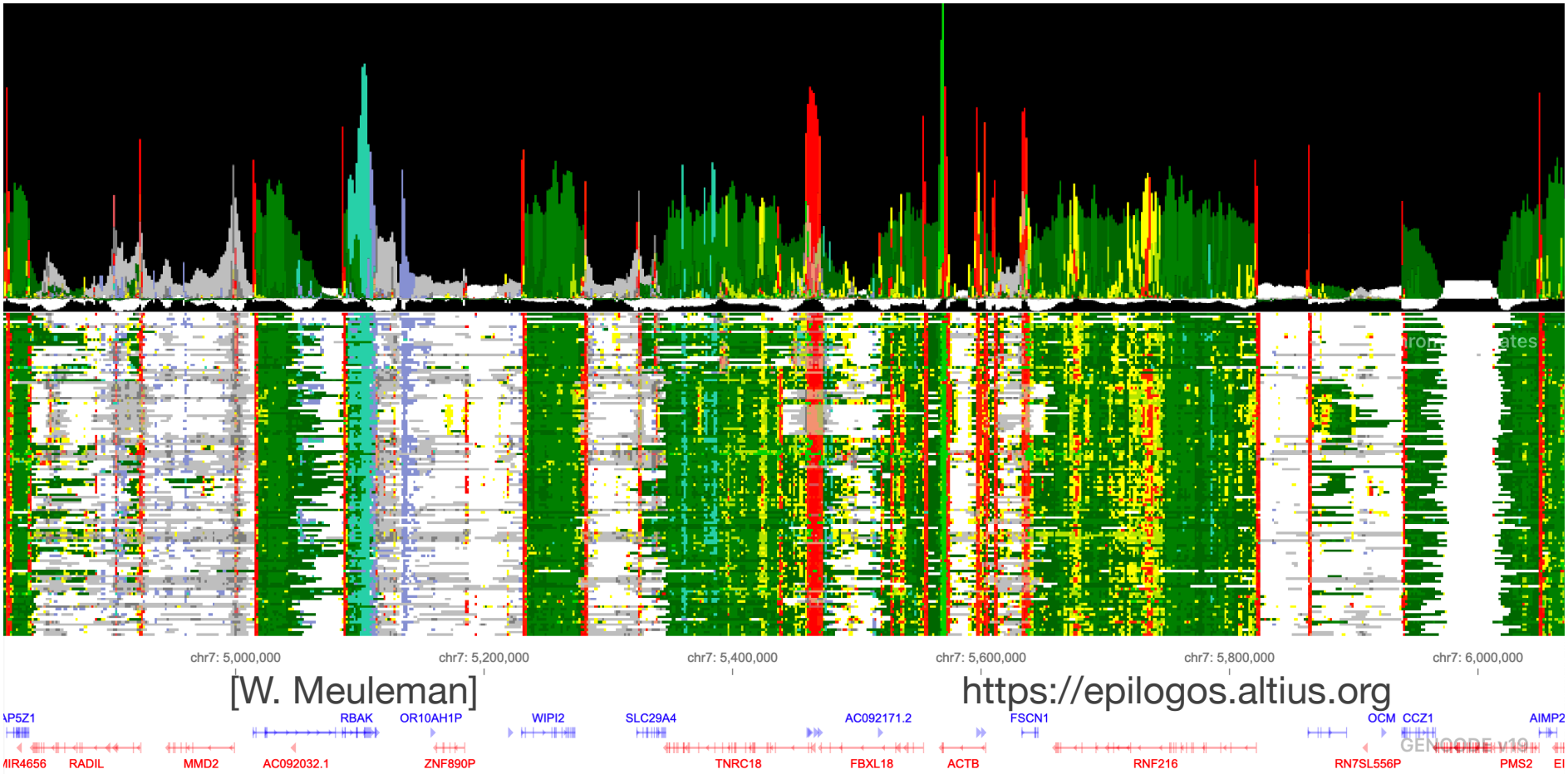


Some states correspond to regulatory regions (active and poised enhancers)
→ **motif search can be restricted to these regions**

http://epigenomegateway.wustl.edu

Institut für Pharmazie und Molekulare Biotechnologie
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

EpiLogos



[W. Meuleman] https://epilogos.altius.org

IPMB
Institut für Pharmazie und Molekulare Biotechnologie
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Conclusions

- Transcription regulation is a **complex process** with an interplay of **multiple components**
- **Transcription factors** play a central role, usually in combination with other TF inside **enhancers**
- Tissue / context specificity of the activity of regulatory elements is given by the **cell-specific chromatin state**: open/accessible or closed/compact
- **Many data types available** to build integrative models of regulatory activity
- **Single-cell genomics** is becoming the new challenge in regulatory genomics