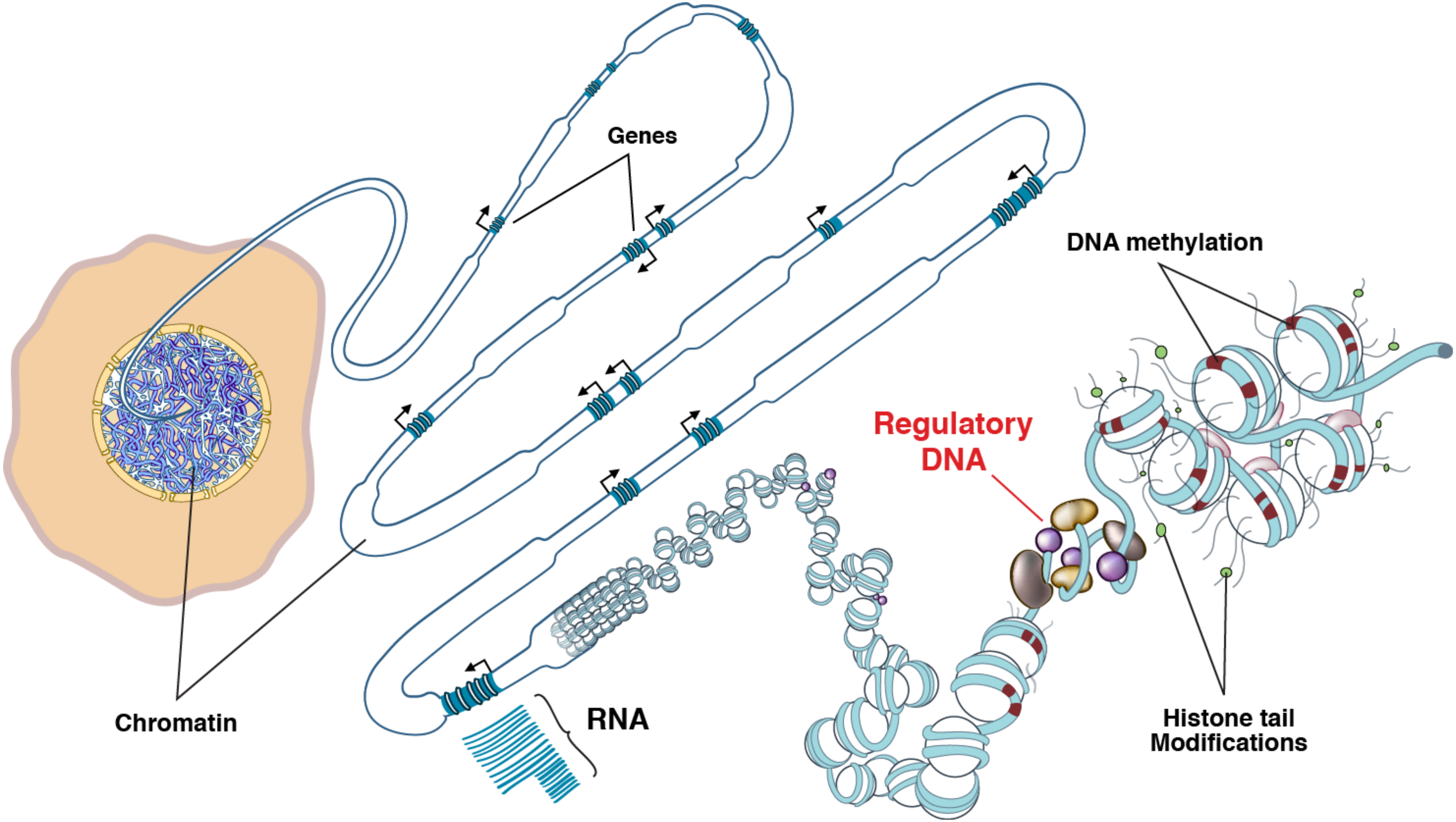
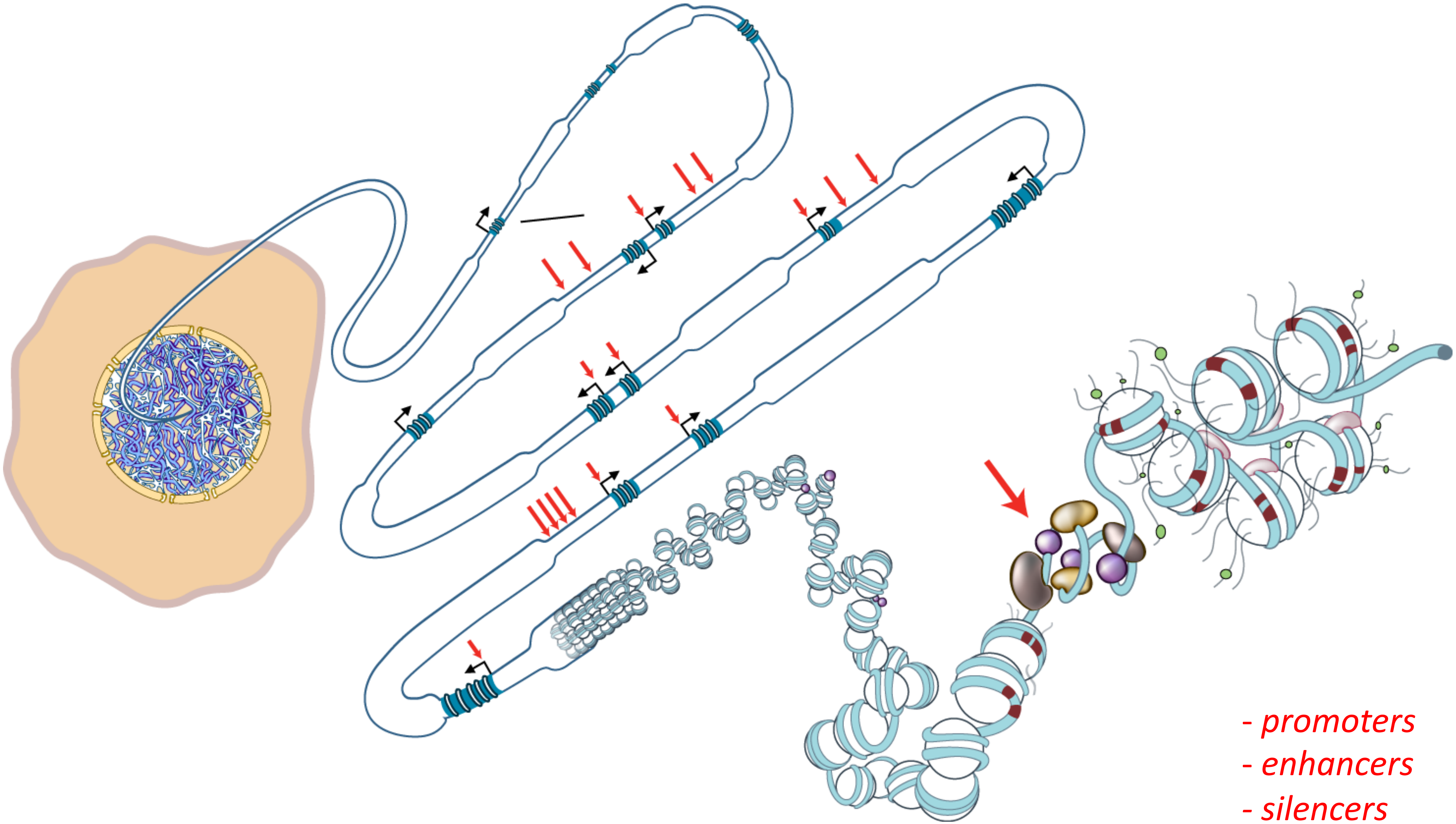
Integrative Analysis of Human and Mouse Regulomes

John A. Stamatoyannopoulos, M.D.
Depts. of Genome Sciences & Medicine
University of Washington

Living genome features



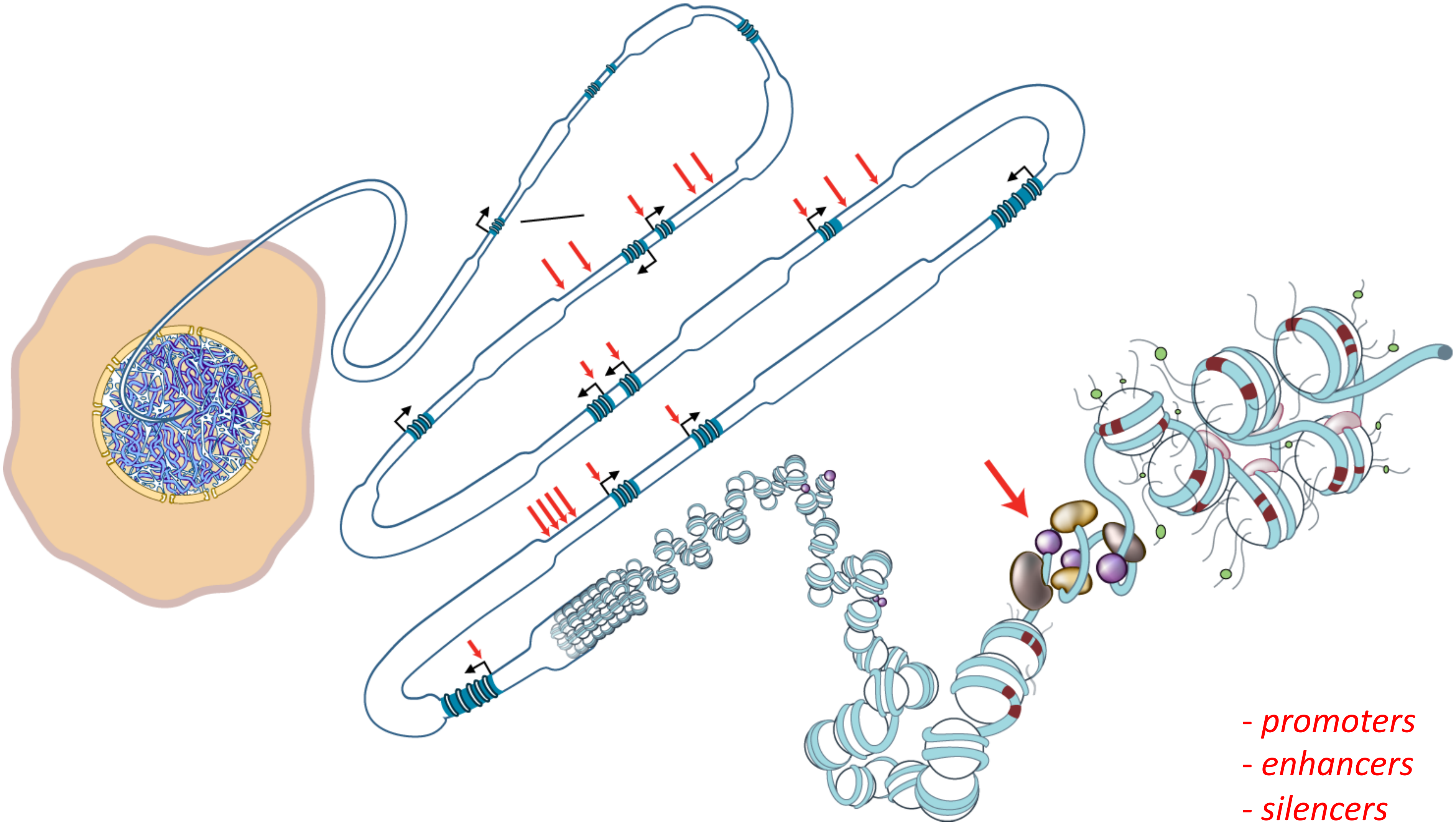
Functional elements



- *promoters*
- *enhancers*
- *silencers*
- *insulators*
- *etc.*

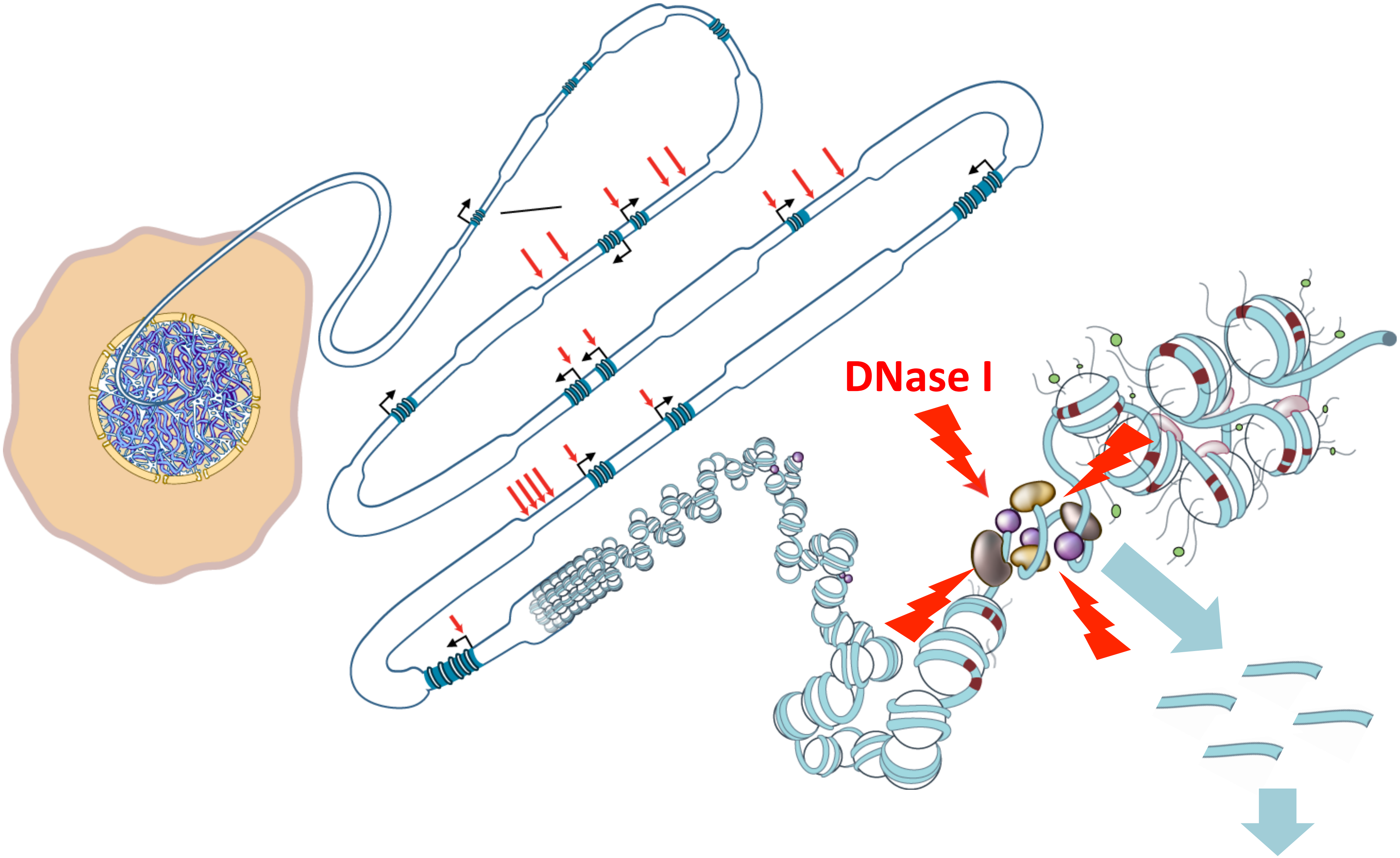
Mapping human regulatory DNA

Functional elements



- *promoters*
- *enhancers*
- *silencers*
- *insulators*
- *etc.*

Mapping regulatory DNA using nucleases (DNase I)

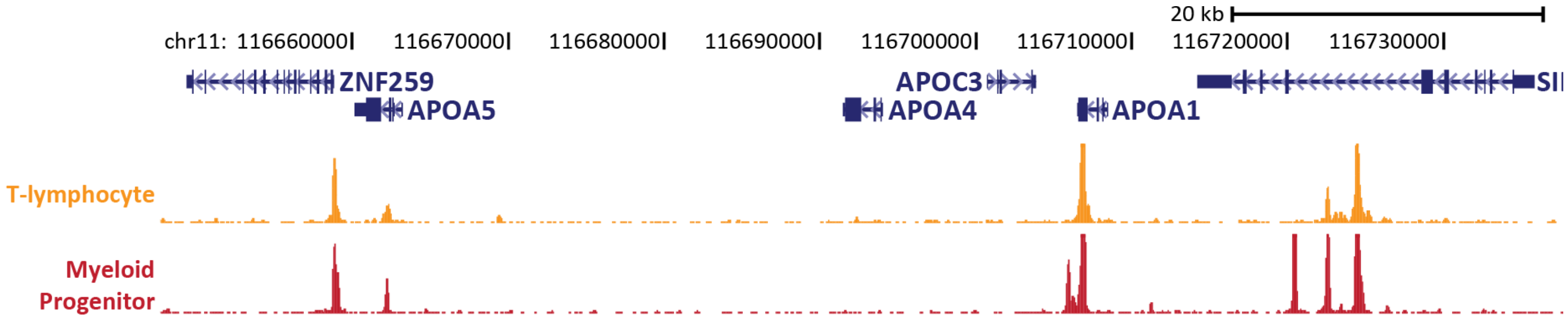


DNase I

Sequencing

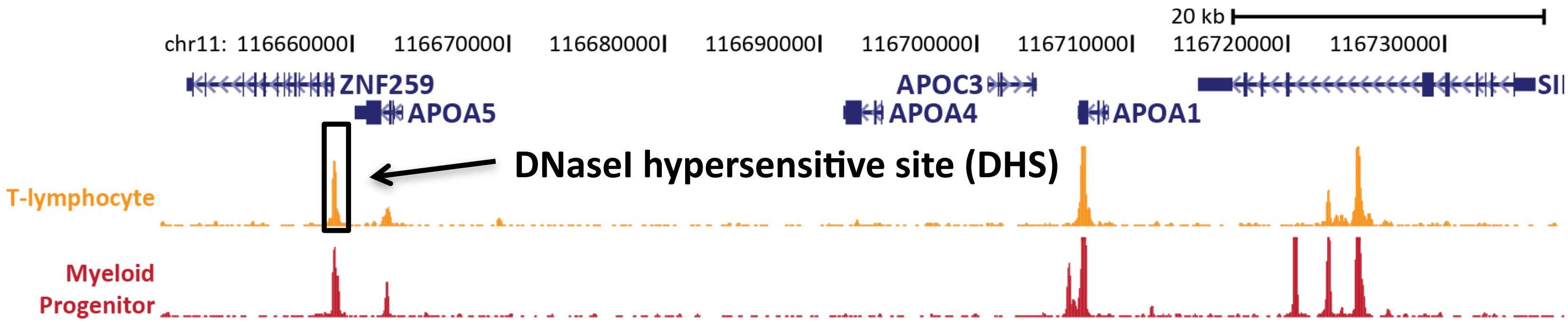
Mapping regulatory DNA

DNase I hypersensitive sites precisely mark regulatory DNA



Mapping regulatory DNA

DNase I hypersensitive sites precisely mark regulatory DNA

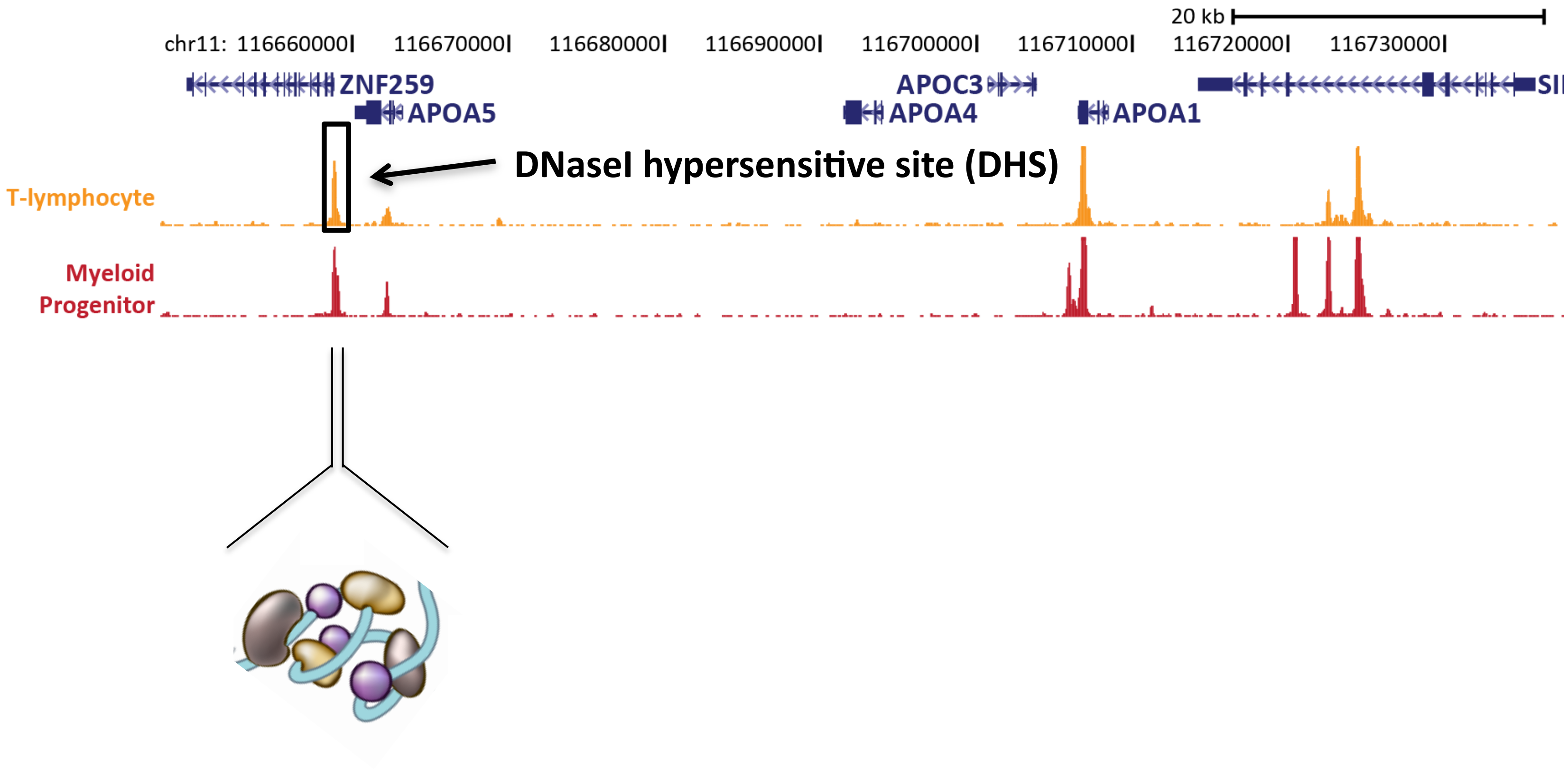


DNaseI hypersensitive site (DHS)

Regulatory DNA

Mapping regulatory DNA

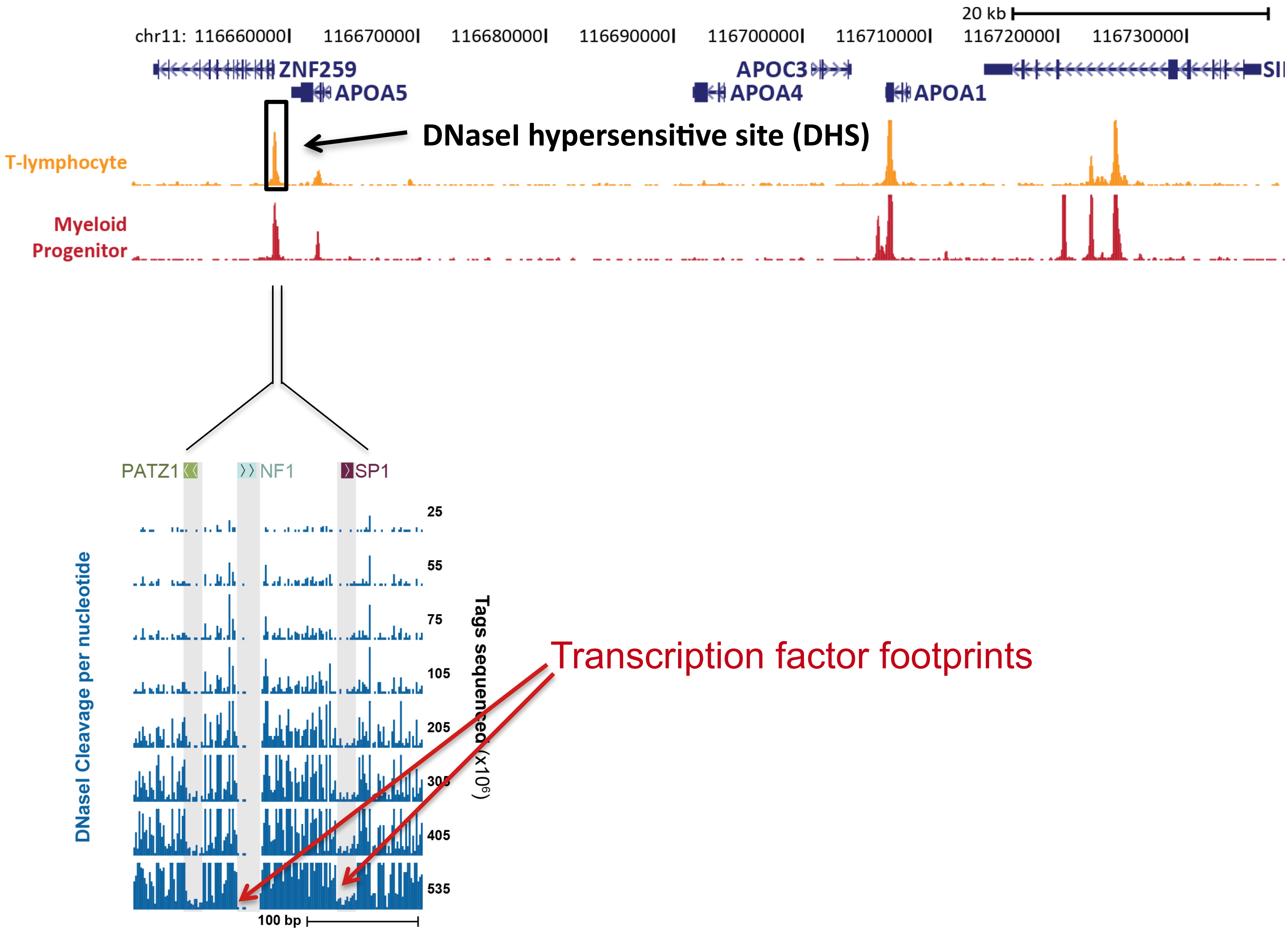
DNase I hypersensitive sites precisely mark regulatory DNA



DNaseI hypersensitive site (DHS)

Mapping regulatory DNA

DNase I hypersensitive sites precisely mark regulatory DNA

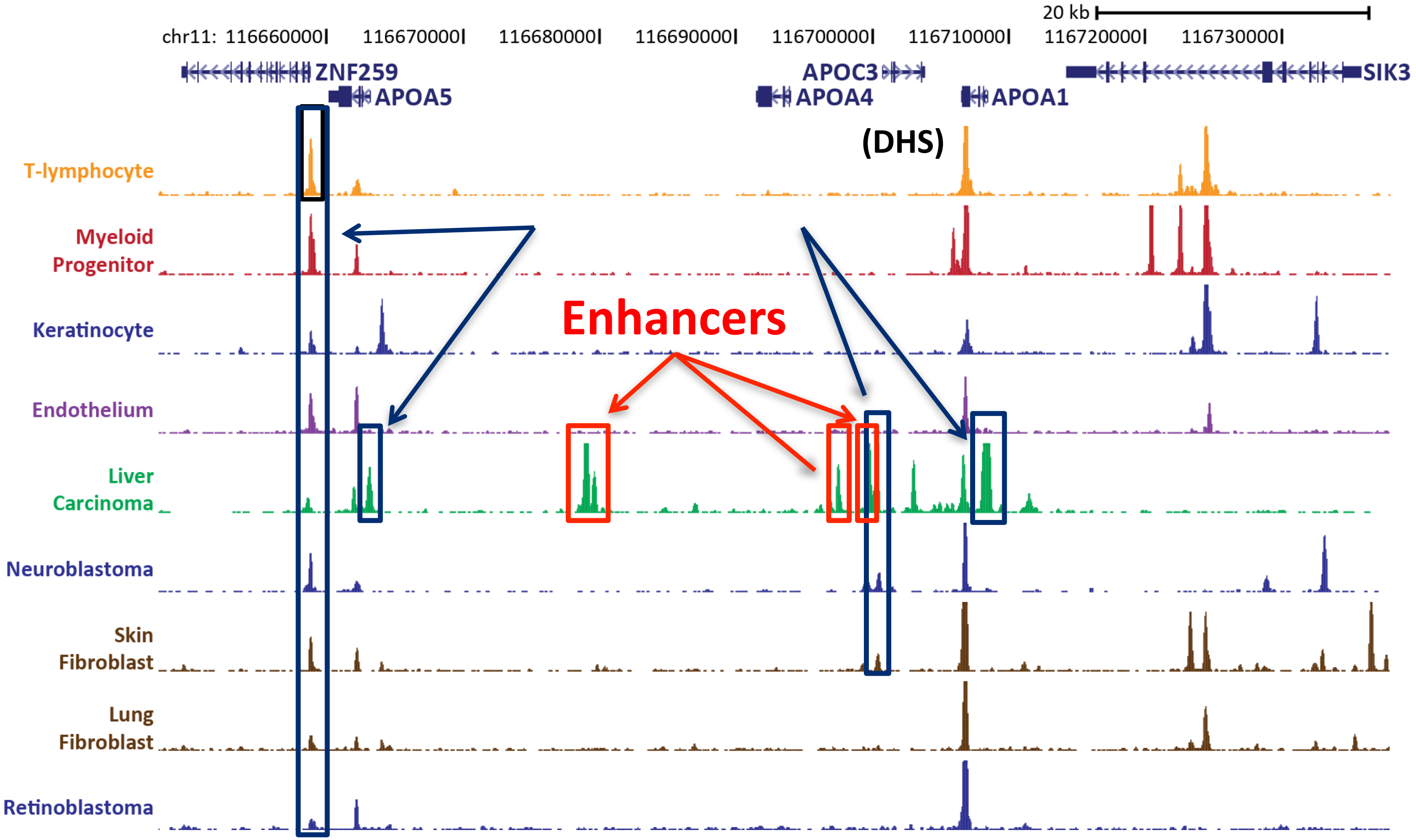


DNaseI hypersensitive site (DHS)

Transcription factor footprints

Mapping regulatory DNA

DNase I hypersensitive sites precisely mark regulatory DNA

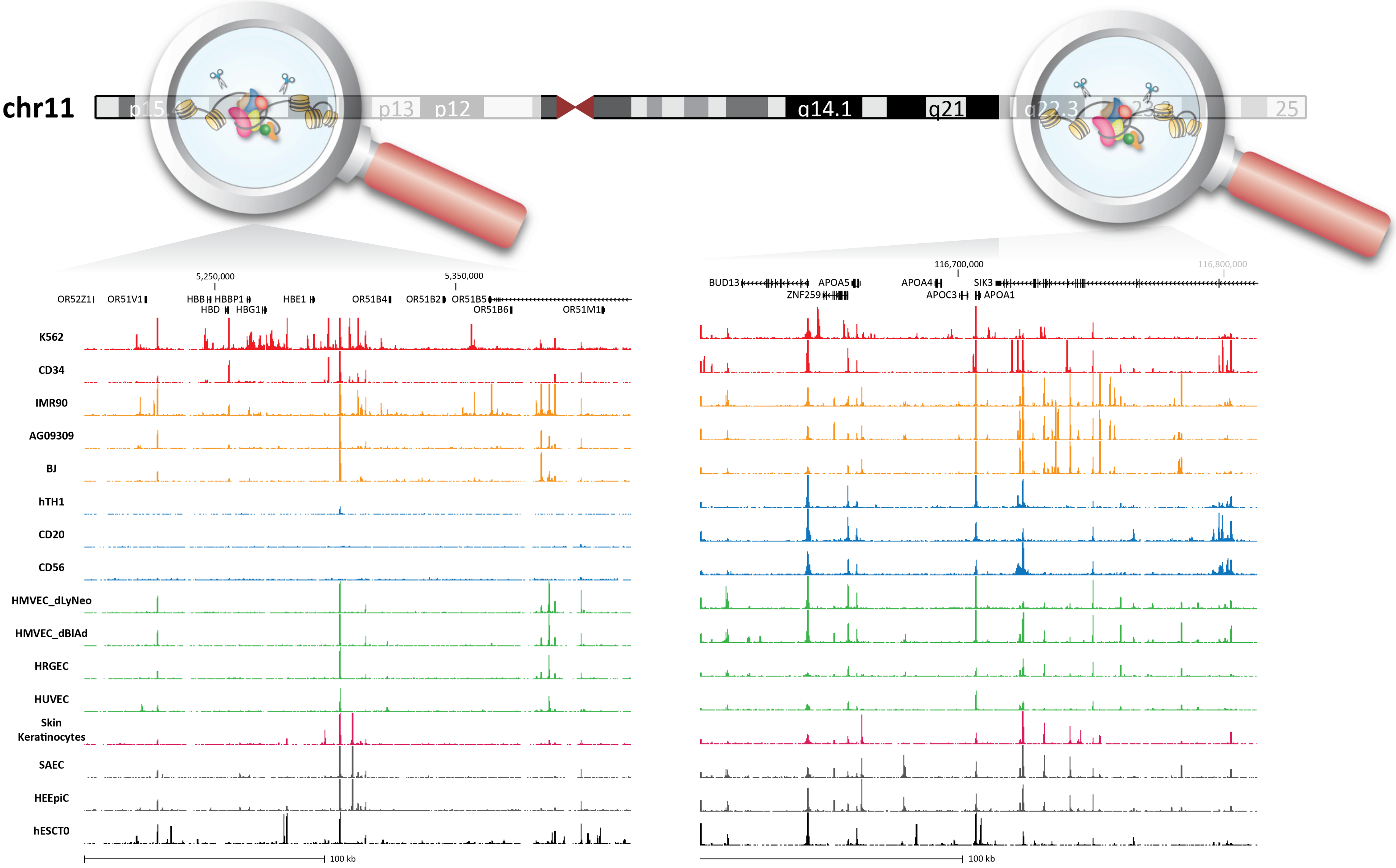


~100,000 – 250,000 elements per cell type (0.5-1.5% of genome)

Thurman et al, Nature, 2012

>400 cell/tissue types and developmental states studied to date

>95% from primary cells and tissues



Mapping the human regulatory genome c. 2015

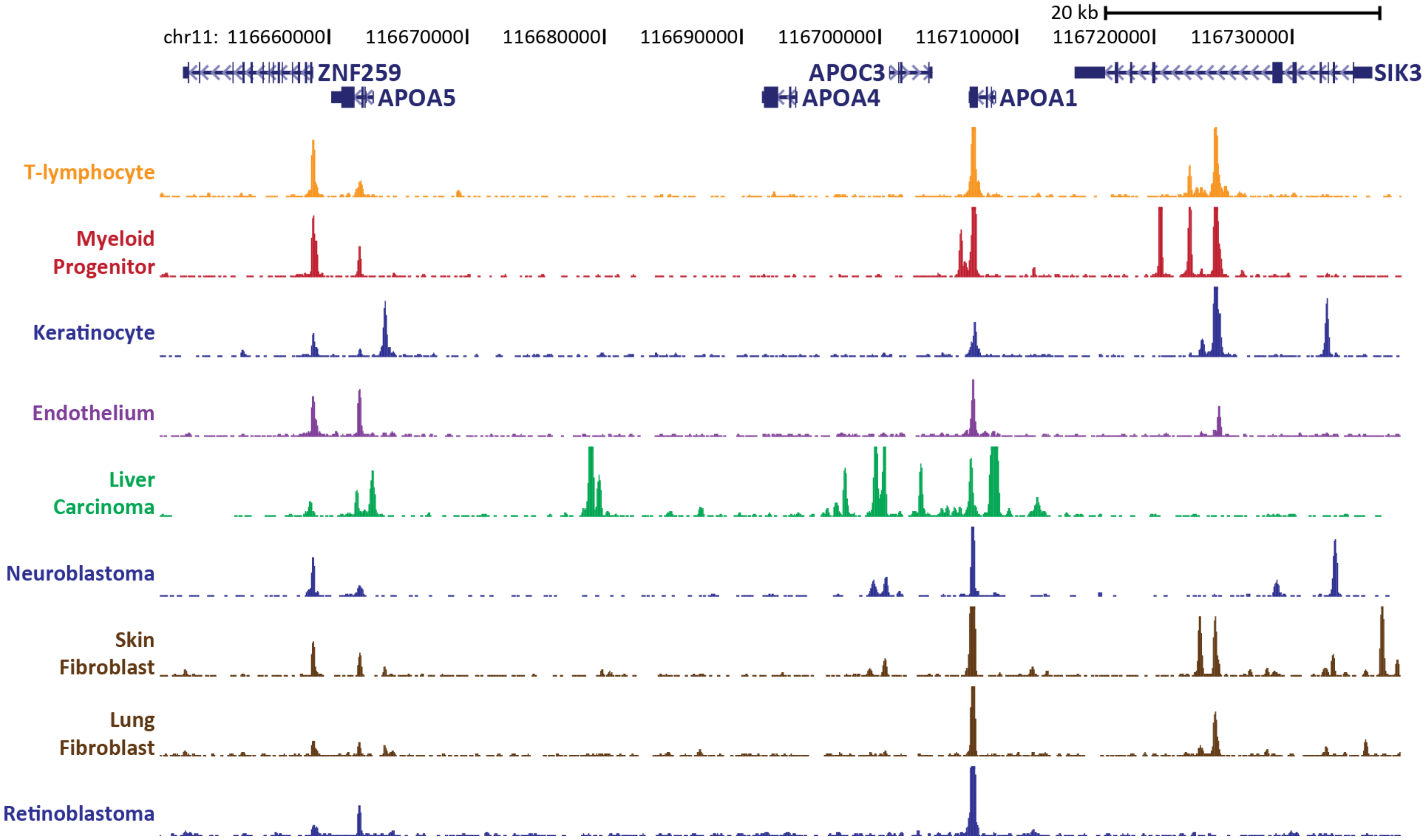
The human genome encodes at least 4 million DNaseI hypersensitive sites

→ *Virtually all (>>99%) are tissue/lineage or cell type-selective elements*

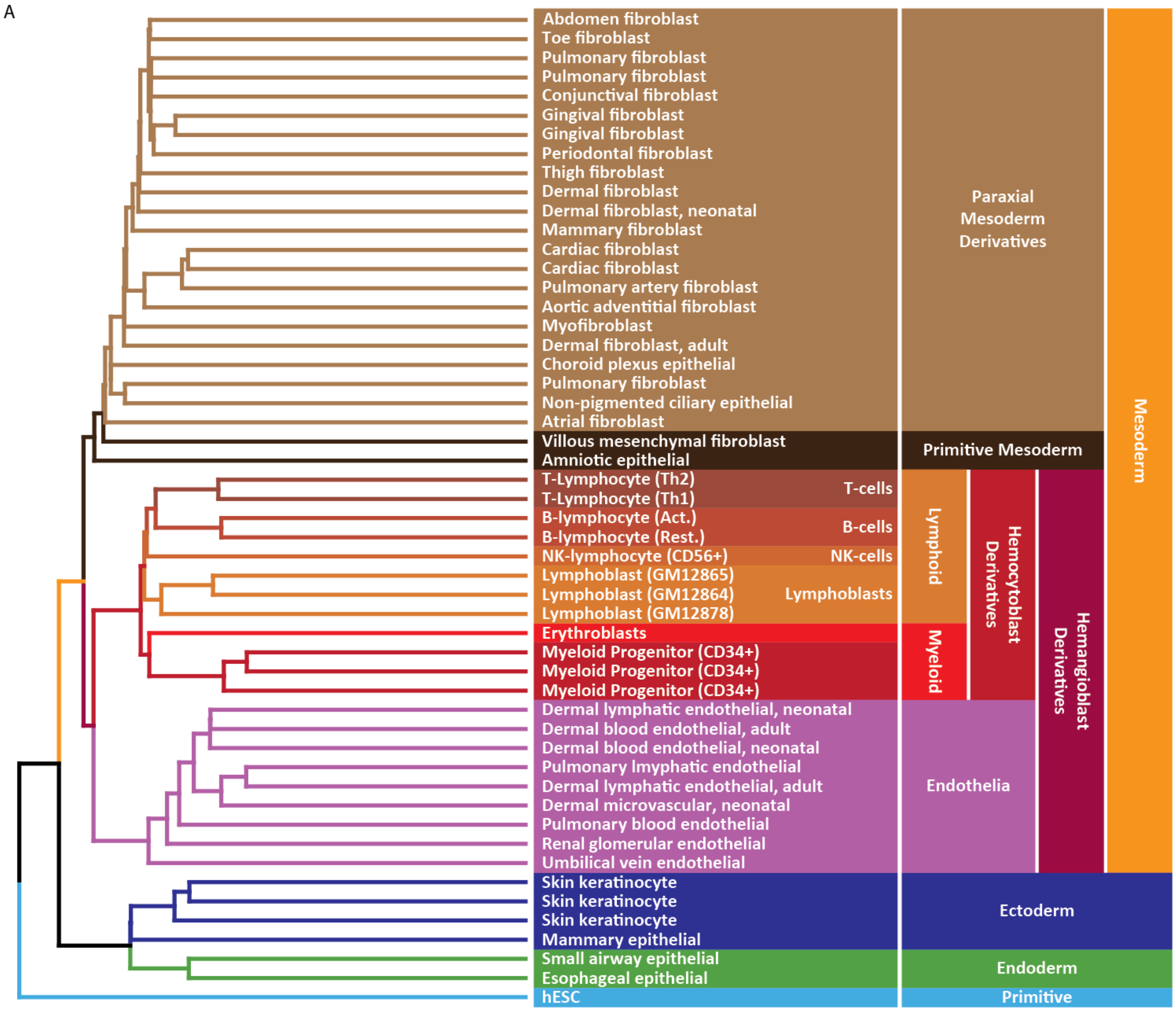
→ *>95% of these are distal non-promoter elements*

→ *~50% of DHSs are 'memory sites' – persistent marks of prior cell states*

What other information is encoded in regulatory DNA patterns?

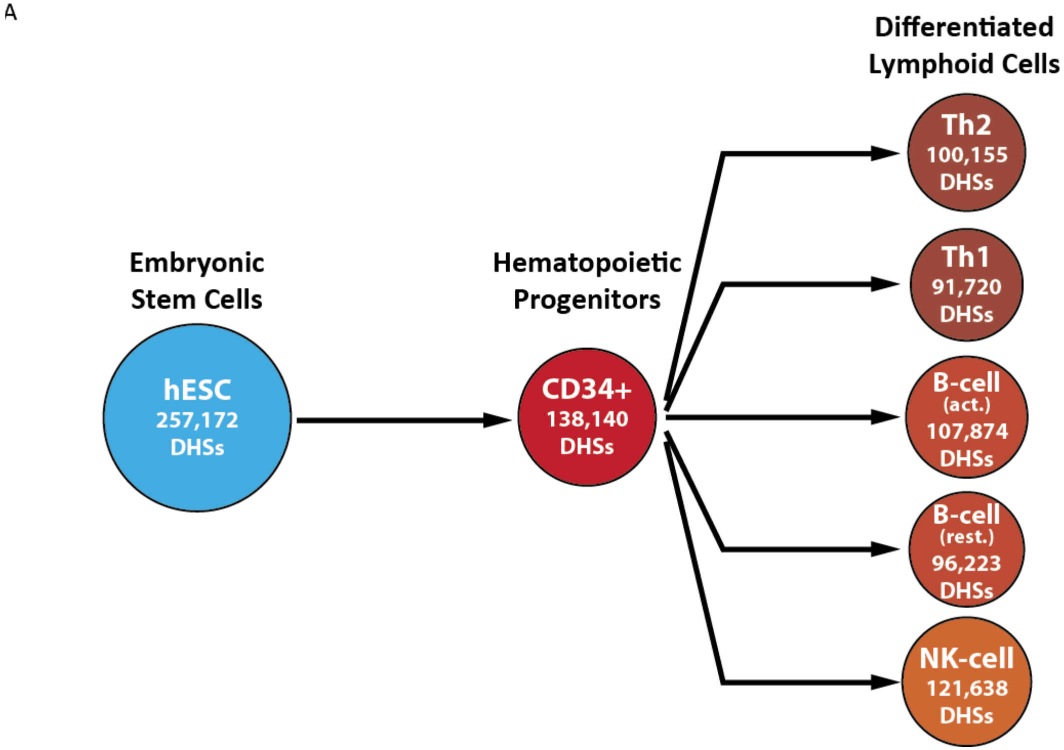


DHS patterns in fully differentiated cells encode memories of prior cell fate decisions



Stergachis et al, Cell, 2013

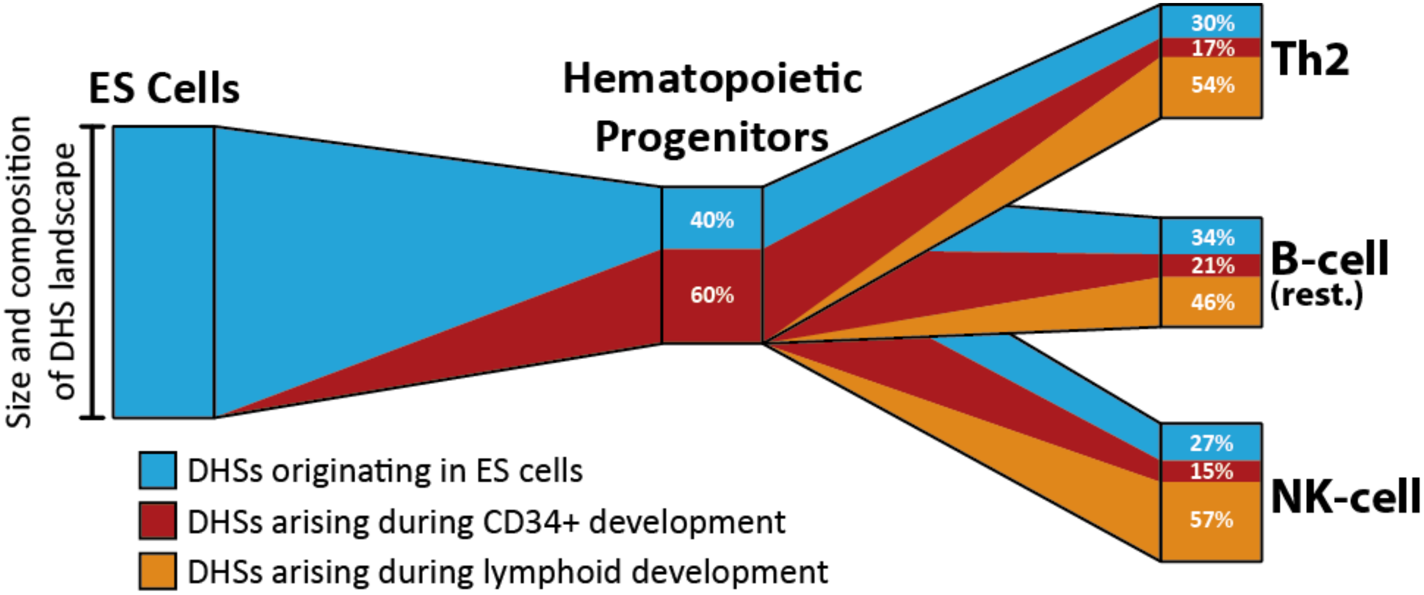
Extensive forward propagation of regulatory information during cell differentiation



Stergachis et al, Cell, 2013

~1/2 of DHSs in definitive cells are 'memory sites' that encode information about prior cell states

Systematic persistence of DHSs during differentiation



Stergachis *et al*, *Cell* 2013

Mapping the human regulatory genome c. 2015

The human genome encodes at least 4 million DNaseI hypersensitive sites

- → *Virtually all (>>99%) are tissue/lineage or cell type-selective elements*
- → *>95% of these are distal non-promoter elements*
- → *~50% of DHSs are 'memory sites' – persistent marks of prior cell states*
- → *Nearly 1 million elements can be linked with likely target genes by co-activation across cell types*
- → *Individual cell types have hundreds to thousands of DHSs that are completely unique for that cell type*

The genome encodes at least 20 million regulatory factor recognition sites

- → *Each cell type likely encodes ~2-5 million transcription factor footprints*
- → *The average cell type utilizes a recognition 'lexicon' of ~2-300 'words*
- → *We are closing in on a complete recognition lexicon for human TFs*

Mapping the human regulatory genome c. 2015

The human genome encodes at least 4 million DNaseI hypersensitive sites

We have little idea what most of these do

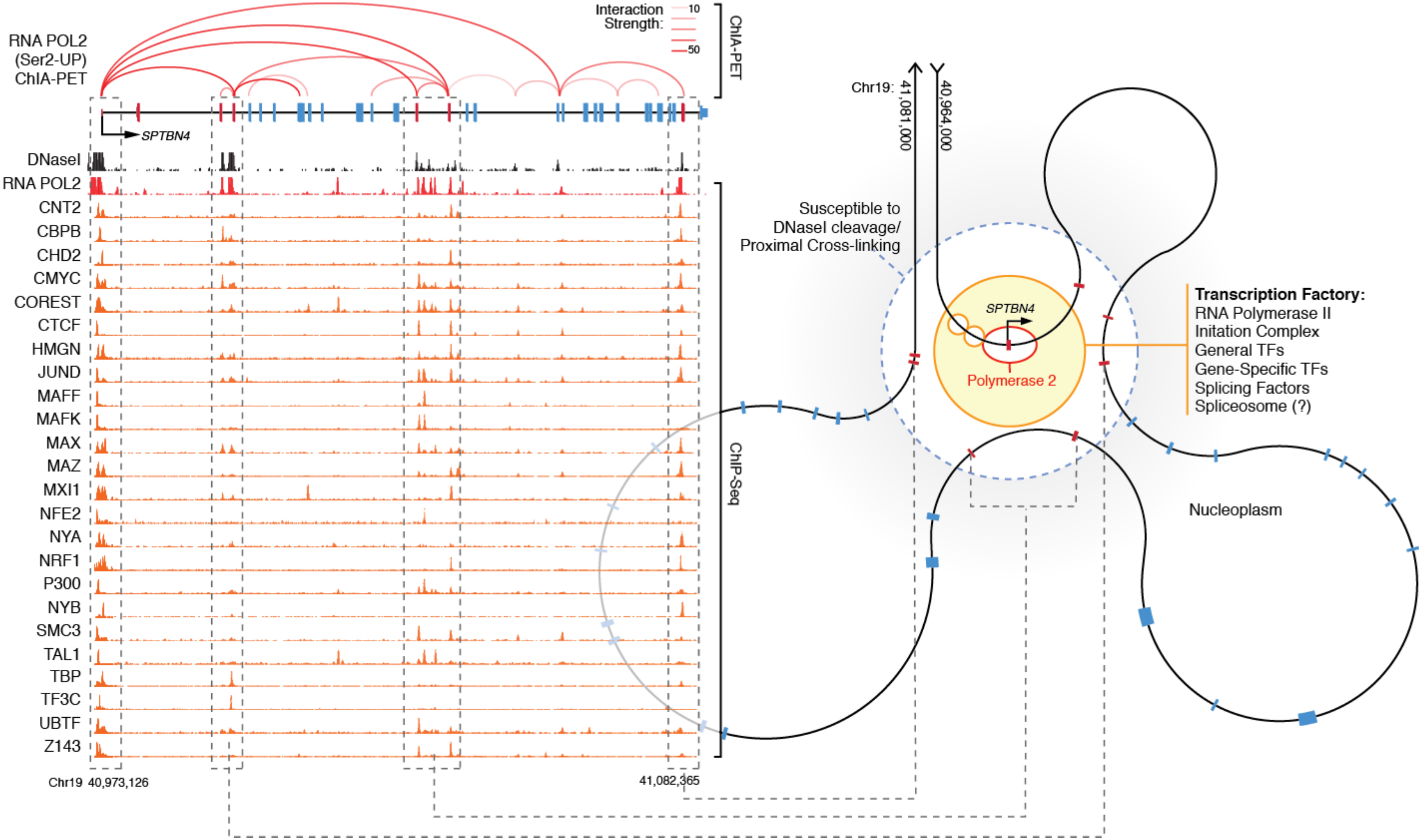
The genome encodes at least 20 million regulatory factor recognition sites

→ *Each cell type likely encodes ~2-5 million transcription factor footprints*

→ *Each cell type utilizes a recognition 'lexicon' of ~2-300 'words*

→ *We are closing in on a complete recognition lexicon for human TFs*

<u>Not just ‘enhancers’</u>: Most regulatory regions likely encode novel and complex activities that will take some time to sort out



Mercer et al., Nature Genetics 2013

Mapping the human regulatory genome c. 2015

The human genome encodes at least 4 million DNaseI hypersensitive sites

We have little idea what most of these do

The genome encodes at least 20 million regulatory factor recognition sites

→ *Each cell type likely encodes ~2-5 million transcription factor footprints*

→ *Each cell type utilizes a recognition 'lexicon' of ~2-300 'words*

→ *We are closing in on a complete recognition lexicon for human TFs*

Mapping the human regulatory genome c. 2015

The human genome encodes at least 4 million DNaseI hypersensitive sites

We have little idea what most of these do

The genome encodes at least 20 million regulatory factor recognition sites

Every regulatory region is built differently, and every TF must do its job (and cooperate with other TFs) in its local context.

Origins

How did the regulatory genome arise?

Part I:

Mouse and human regulatory regions

Evolutionary dynamics of regulatory DNA regions

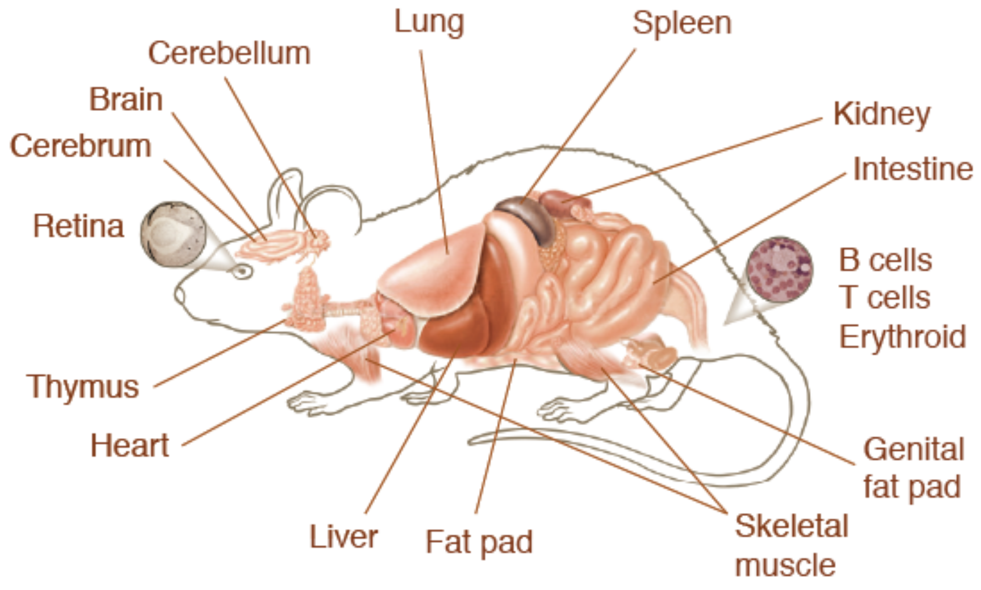
Part II:

Transcription factors and networks

Conservation of *trans* vs. *cis* regulatory circuitry

Creating comprehensive maps of mouse regulatory DNA marked by DNase I hypersensitive sites (DHSs)

- 44 cells/tissues studied
- **1.3 million** distinct DHSs
- Avg. 150,000 per cell/ tissue type



Primitive cells / tissues:
ES cells, limb, embyronic mesoderm

Integrative comparison with ~3 million DHSs from 230 human cell/tissue types

Comparative analysis of mouse and human regulatory DNA

Align sequence to human genome through pair-wise alignment

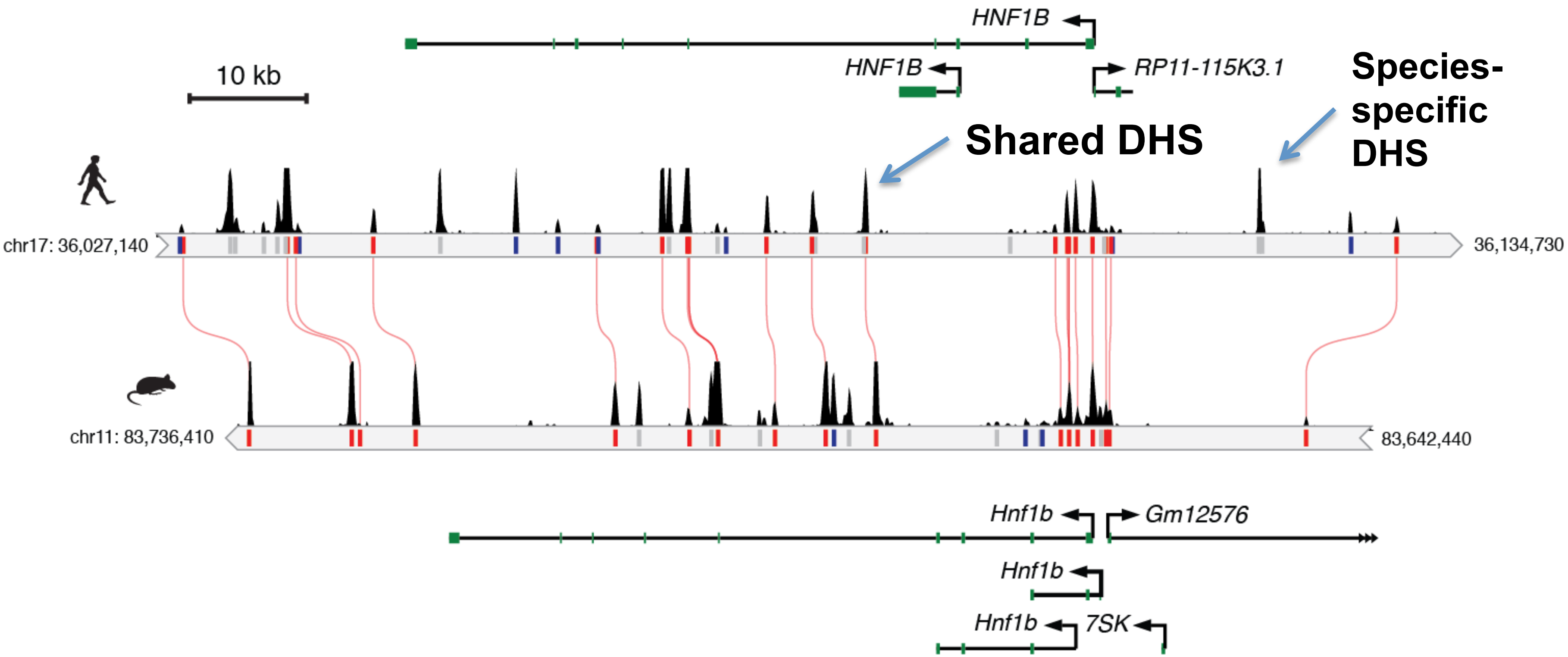
→

Overlap aligned segments with human DHSs (any cell type)

Comparative analysis of mouse and human regulatory DNA

Align sequence to human genome through pair-wise alignment

Overlap aligned segments with human DHSs (any cell type)



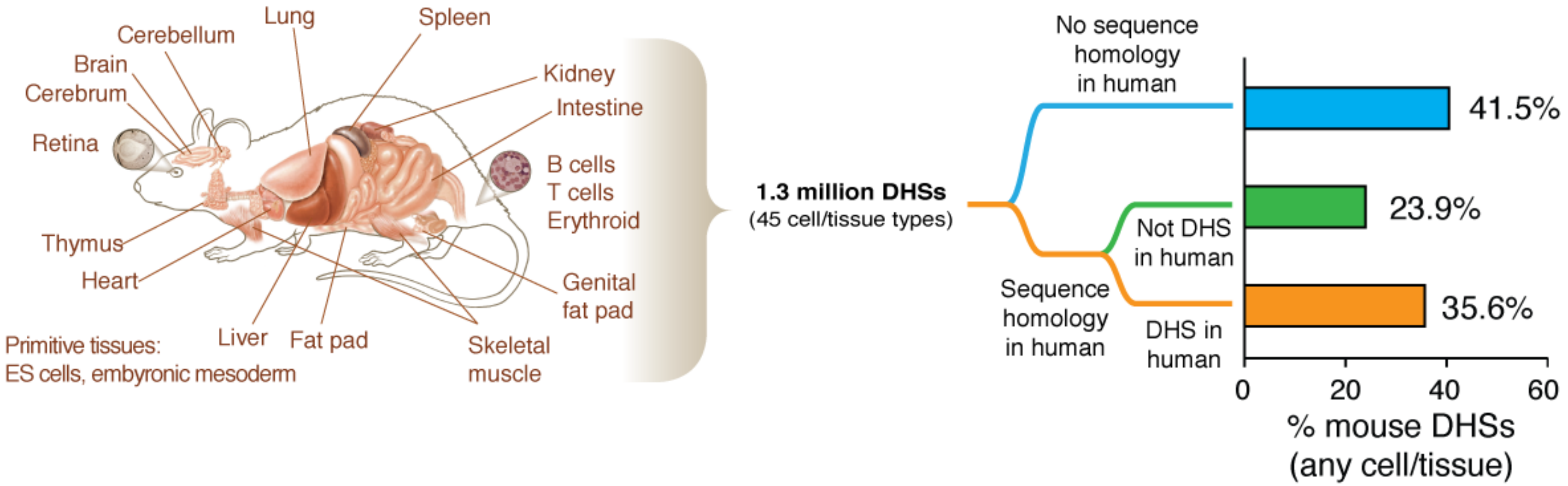
Shared DHS

Species-specific DHS

Comparative analysis of mouse and human regulatory DNA

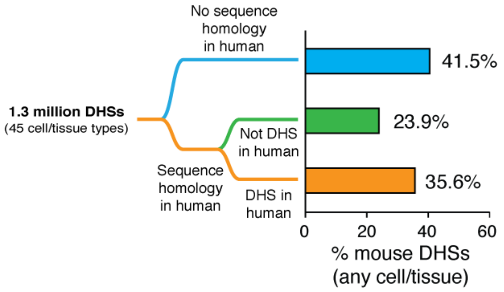
Align sequence to human genome through pair-wise alignment

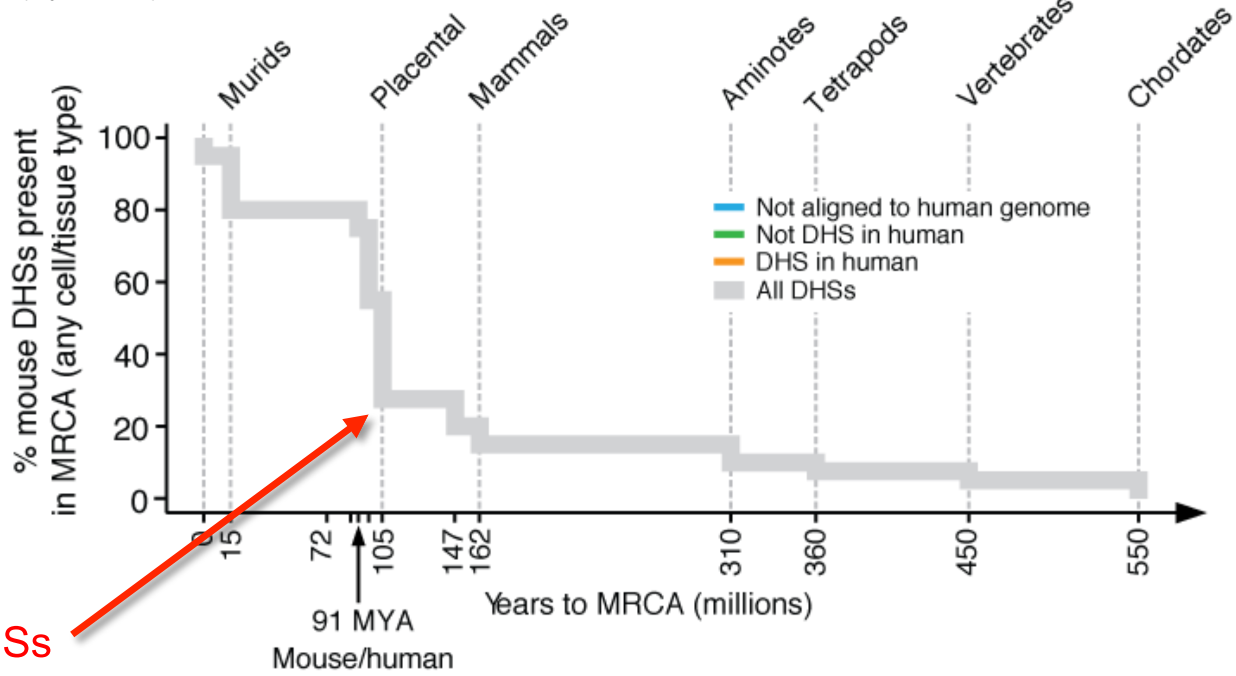
Overlap aligned segments with human DHSs (any cell type)



Pervasive turnover of regulatory DNA in placental mammals

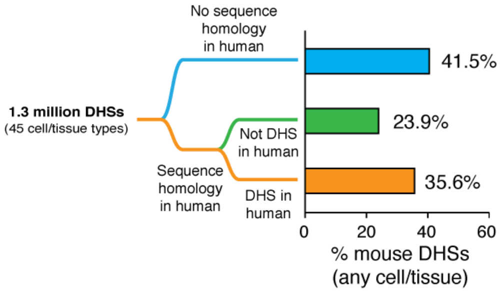
Pervasive turnover of *cis*-regulatory DNA during mammalian evolution

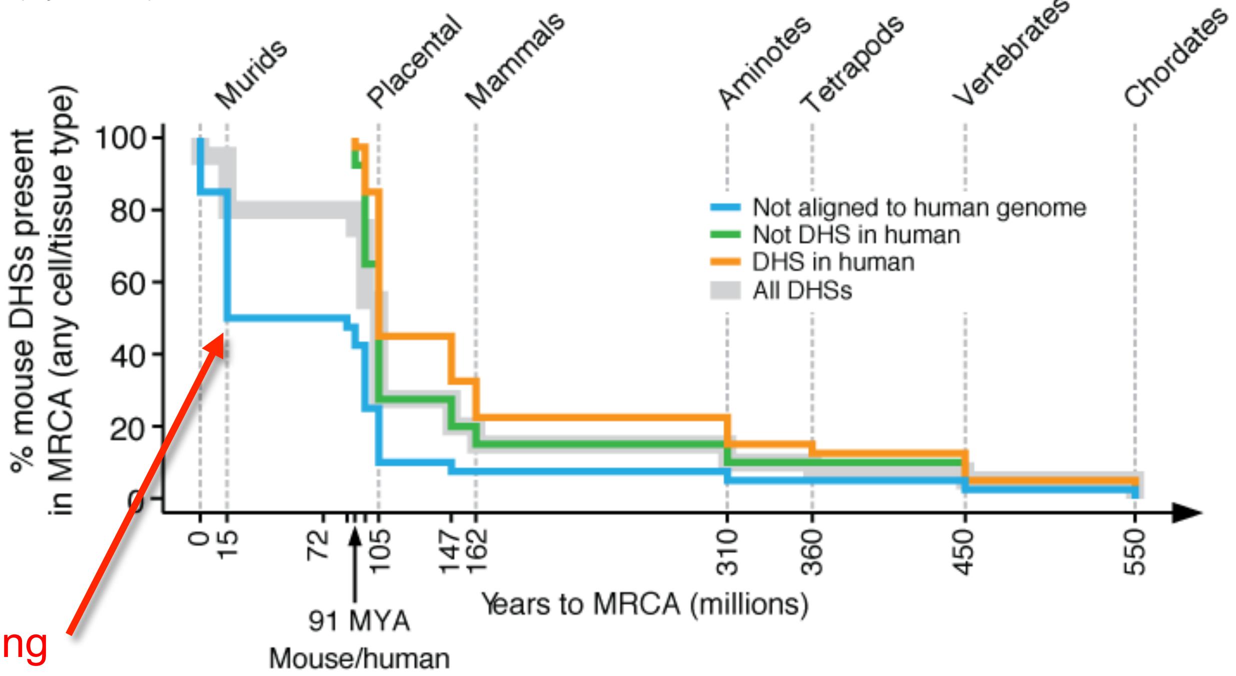




~75% of DHSs restricted to placental mammals

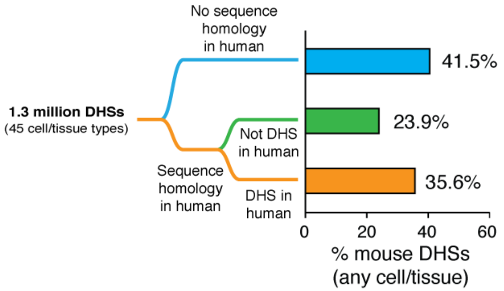
Pervasive turnover of *cis*-regulatory DNA during mammalian evolution

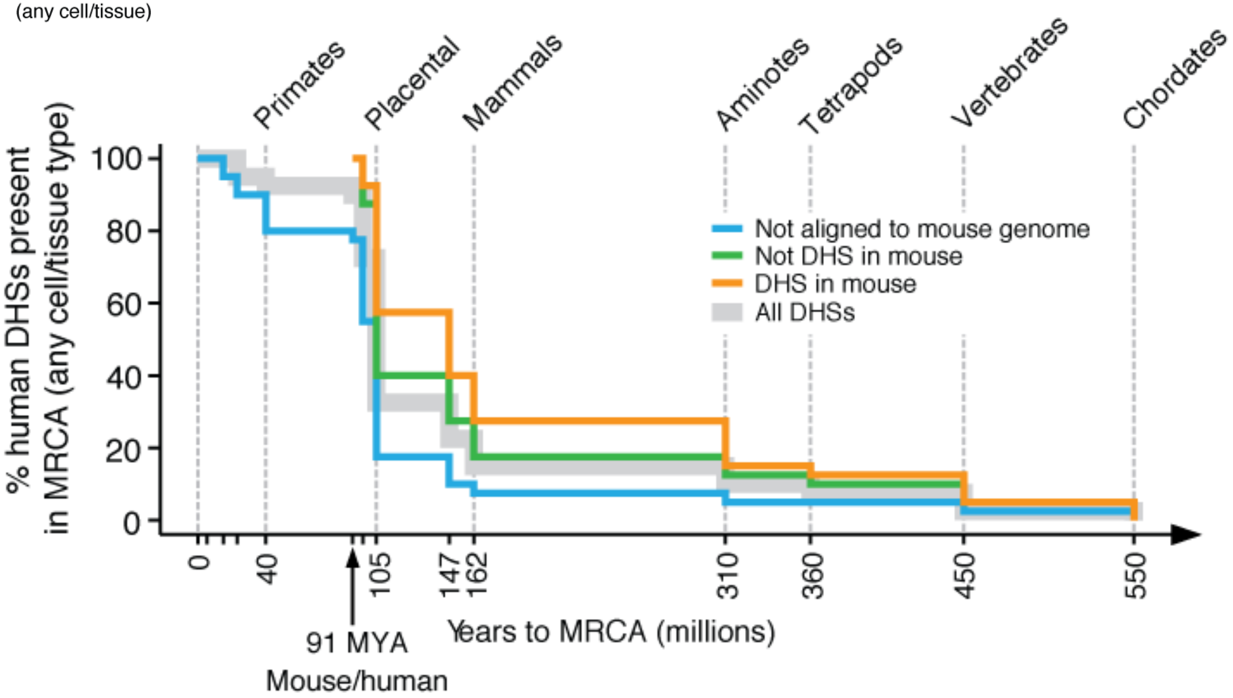




50% non-aligning DHSs restricted to murids

Pervasive turnover of *cis*-regulatory DNA during mammalian evolution





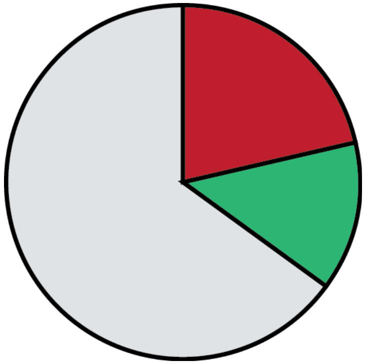
The vast majority of mouse and human regulatory DNA is Placental mammal-specific and has undergone rapid evolution

Evolutionary mechanism #1:

Functional repurposing of regulatory DNA

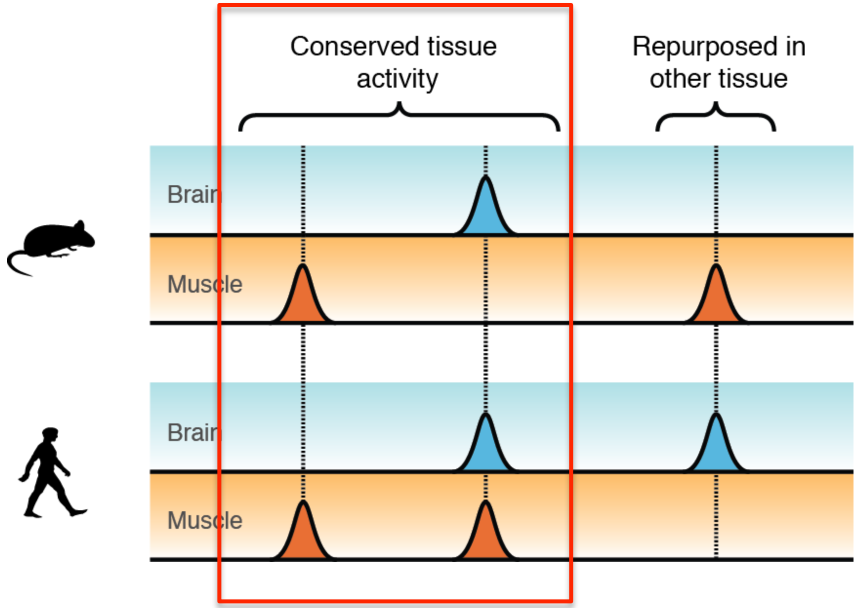
Extensive functional 'repurposing' of regulatory DNA

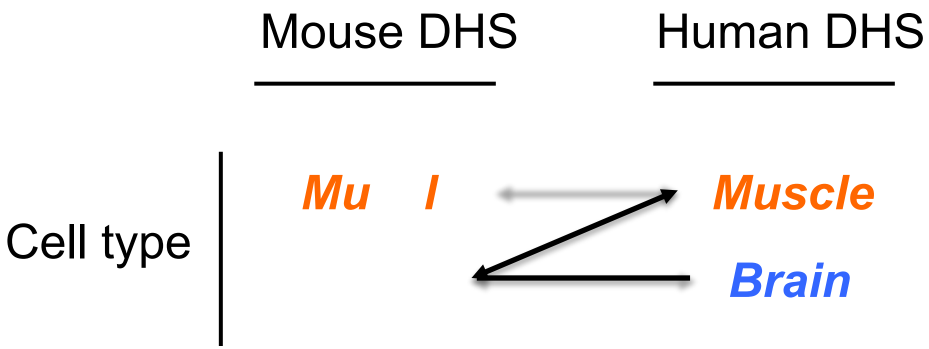
1.3 million mouse DHSs



21% same tissue

14% different tissue





Mouse DHS

Human DHS

Cell type

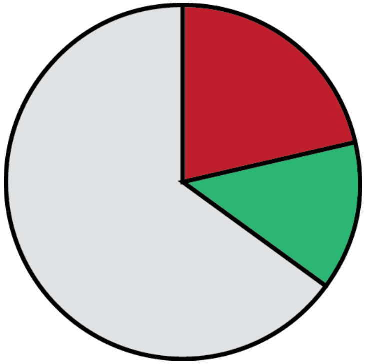
Mu l

Muscle

Brain

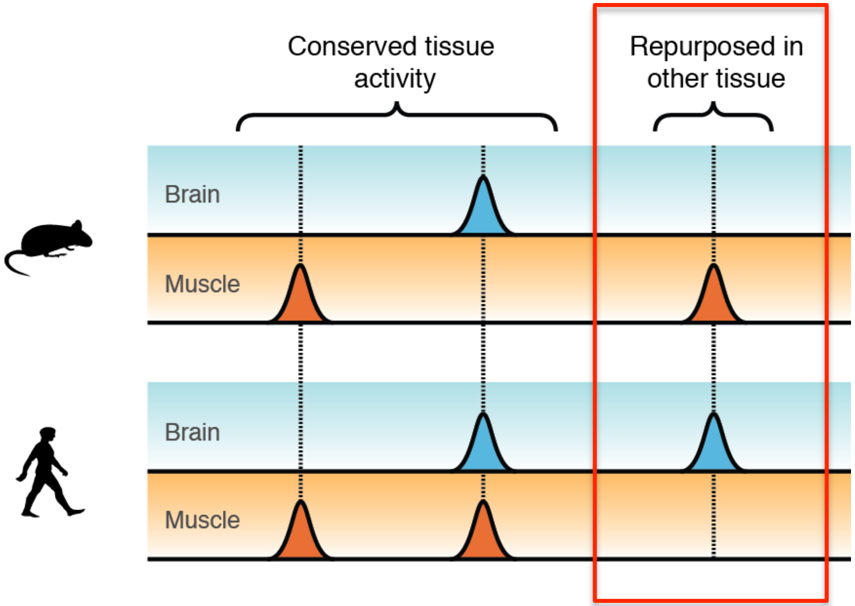
Extensive functional 'repurposing' of regulatory DNA

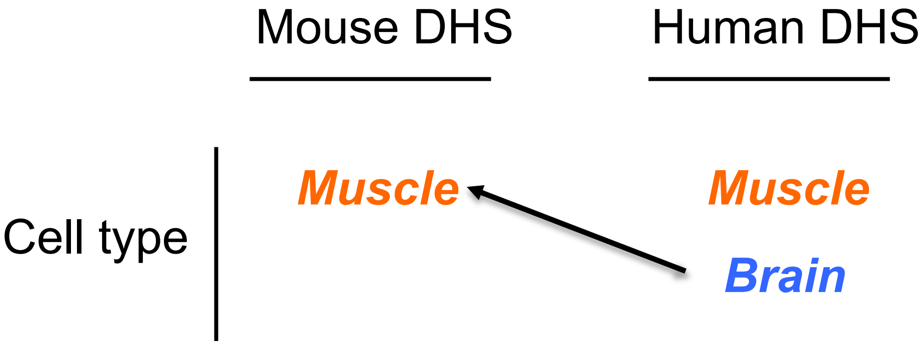
1.3 million mouse DHSs



21% same tissue

14% different tissue





Mouse DHS

Human DHS

Cell type

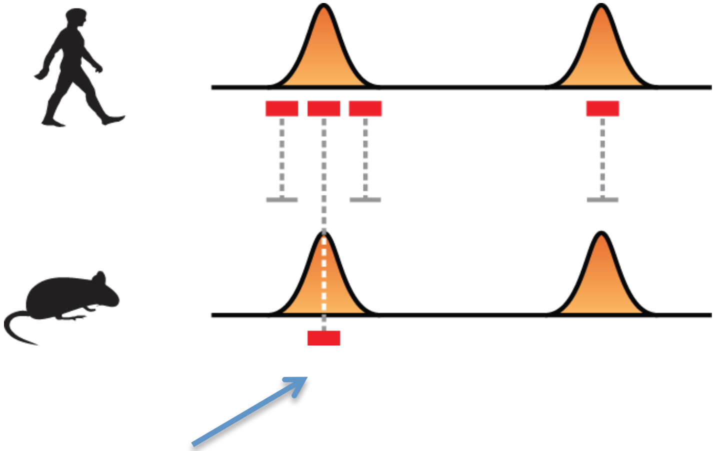
Muscle

Muscle

Brain

Simple but pervasive sequence changes underlie tissue repurposing

Mechanism for functional repurposing:
TF binding site turnover

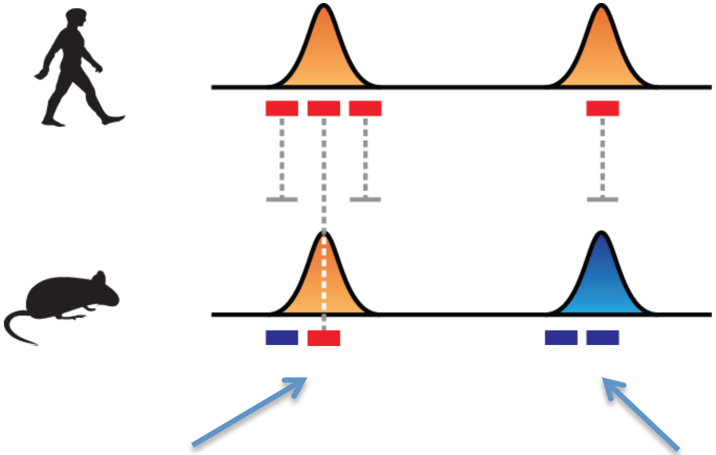


Conservation of TF binding site

Simple but pervasive sequence changes underlie tissue repurposing

Mechanism for functional repurposing:

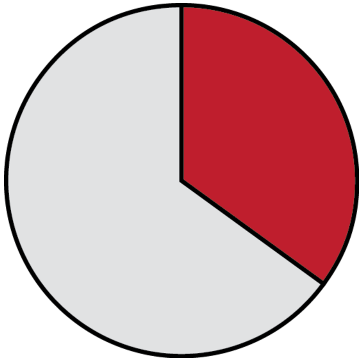
TF binding site turnover



Conservation of TF binding site

Novel binding site, novel function

1.3 million mouse DHSs

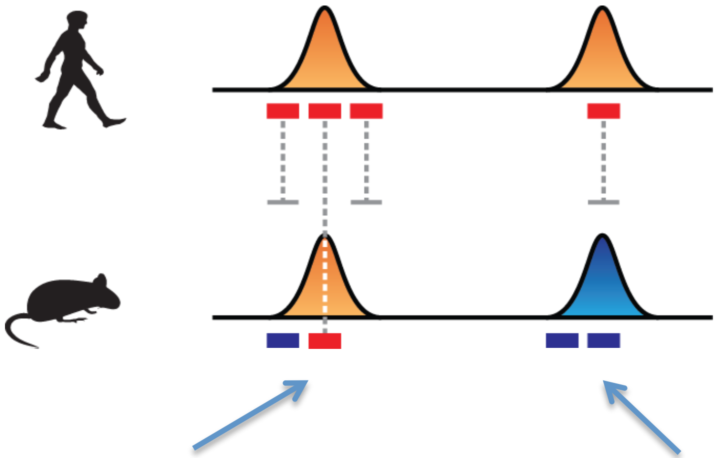


35% DHS in human (any cell type)

Simple but pervasive sequence changes underlie tissue repurposing

Mechanism for functional repurposing:

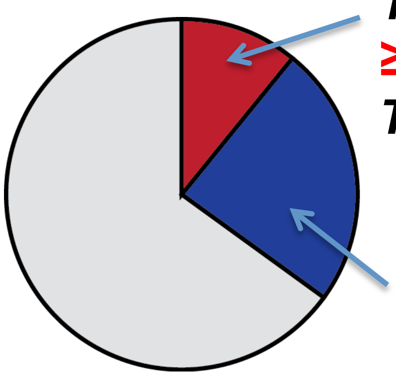
TF binding site turnover



Conservation of TF binding site

Novel binding site, novel function

1.3 million mouse DHSs

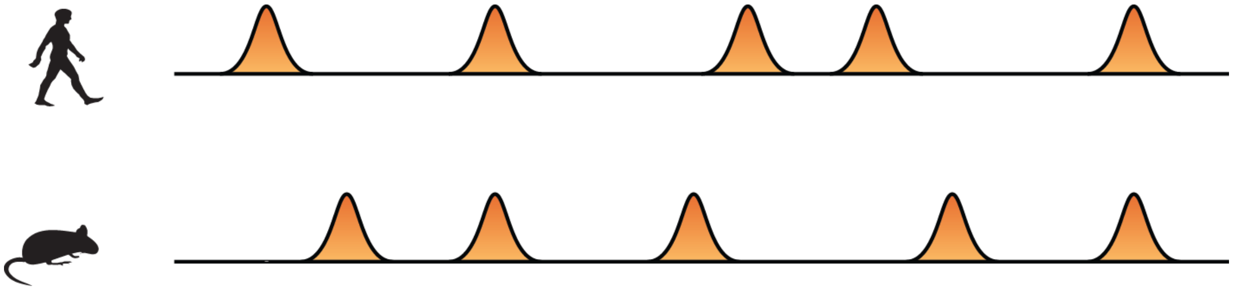


11%
≥1 conserved TF binding site

24%
No conserved TF binding sites

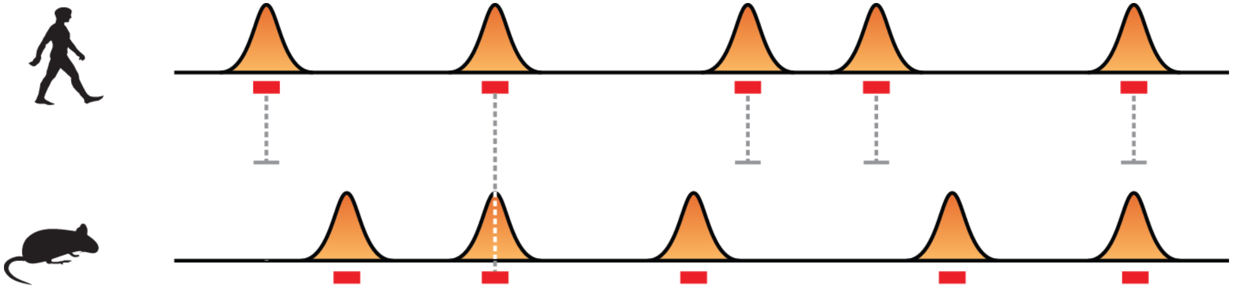
Conserved TF binding sites are significantly enriched in DHSs with conserved activity

Conservation of *cis*-regulatory architecture vs. individual elements



Tissue-specific DHS landscape

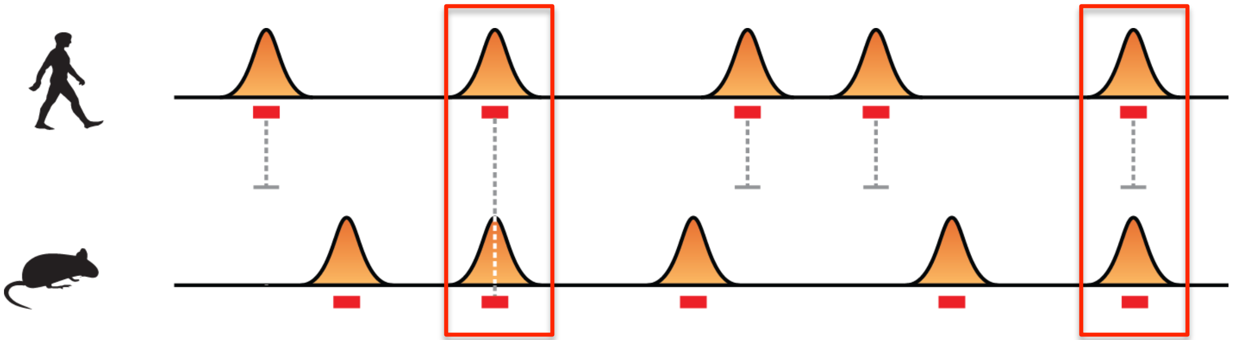
Conservation of *cis*-regulatory architecture vs. individual elements



Tissue-specific DHS landscape

- **21% of mouse DHS landscape is shared with a corresponding human tissue**

Conservation of *cis*-regulatory architecture vs. individual elements

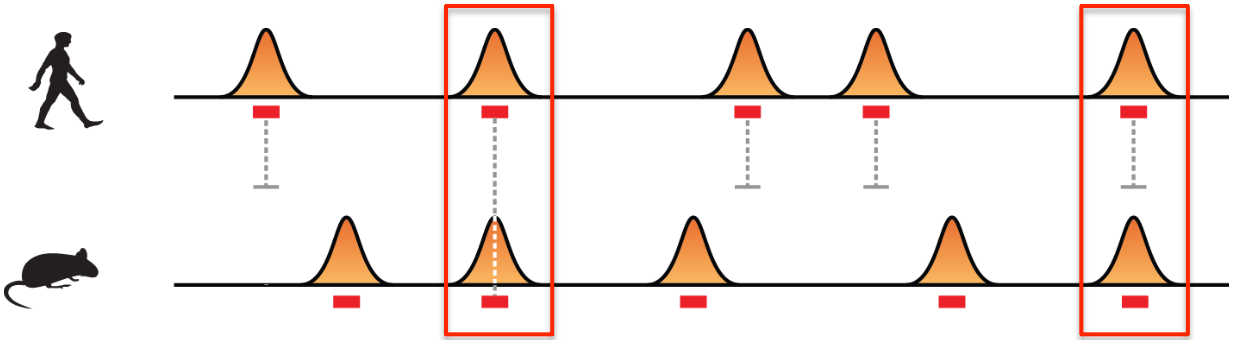


Tissue-specific DHS landscape

- 21% of mouse DHS landscape is shared with a corresponding human tissue

- 11% has a conserved TF binding site

Conservation of *cis*-regulatory architecture vs. individual elements



Tissue-specific DHS landscape

- **21% of mouse DHS landscape is shared with a corresponding human tissue**

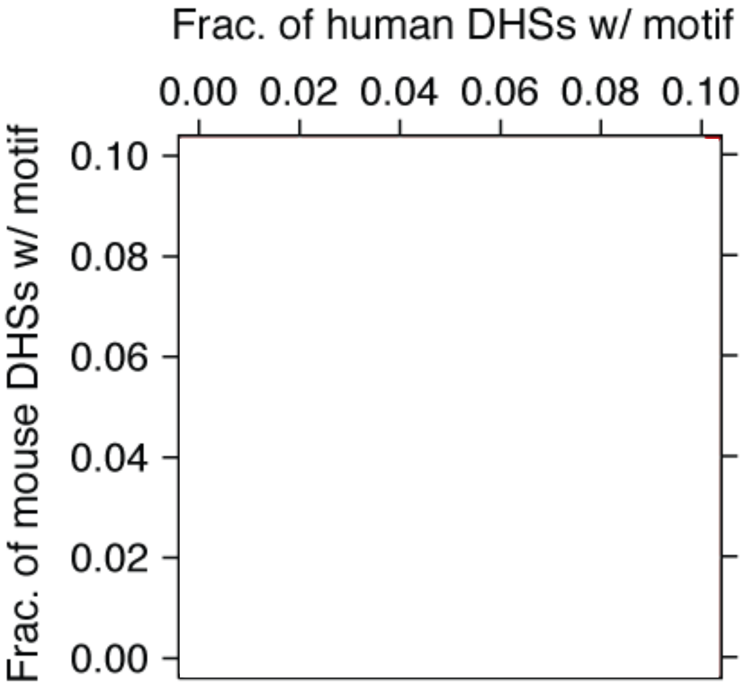
- **11% has a conserved TF binding site**

Given divergent regulatory landscapes, what is maintaining functional conservation in mouse and human?

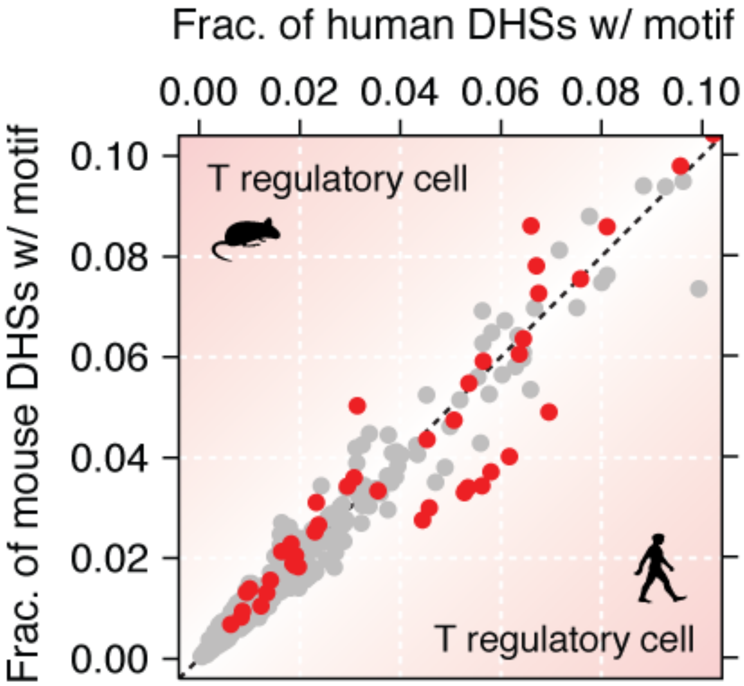
Evolutionary mechanism #2:

Conservation of global *cis*-regulatory ‘content’

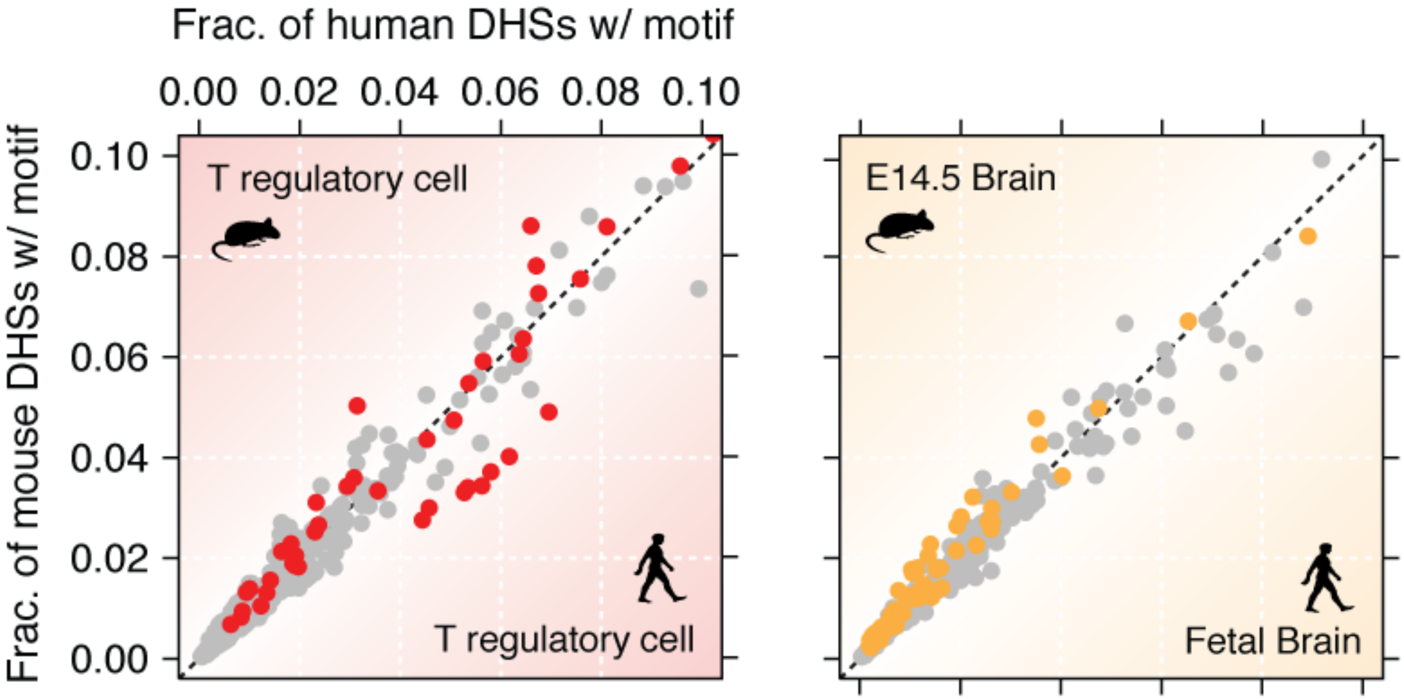
Conservation of overall *cis*-regulatory content



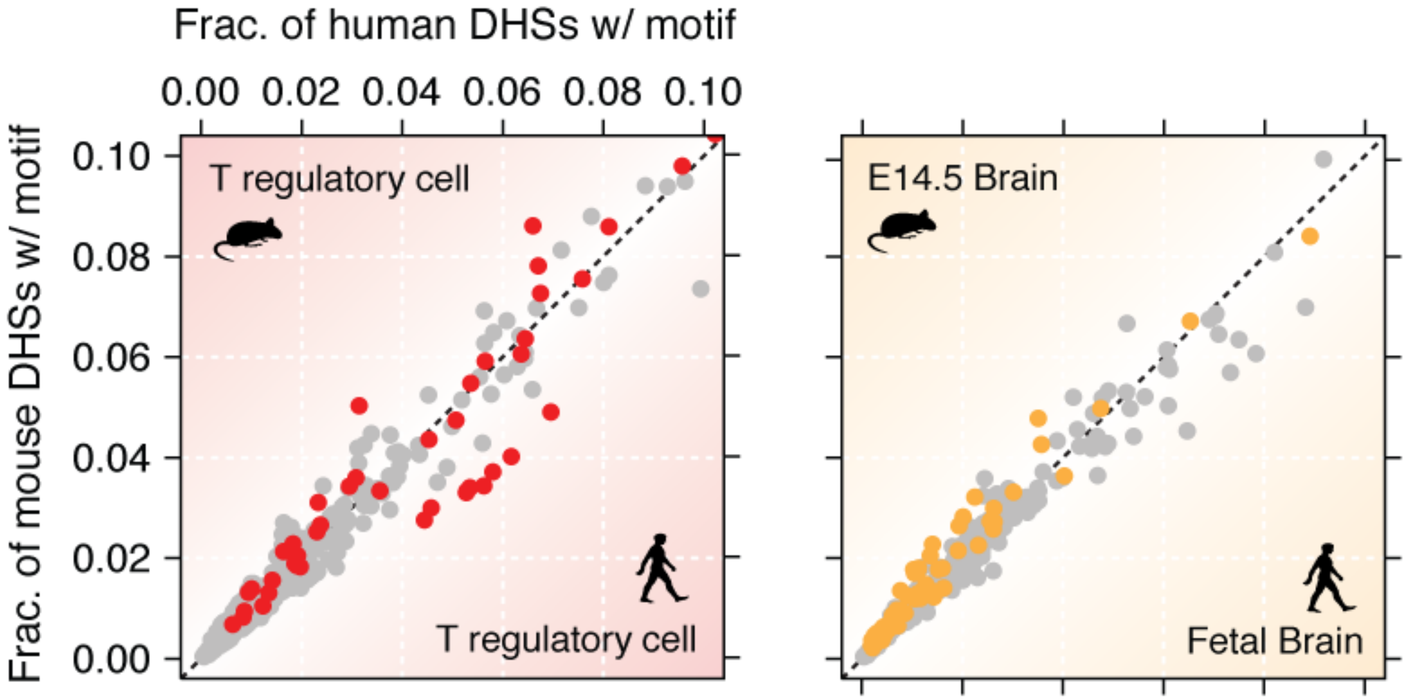
Conservation of overall *cis*-regulatory content



Conservation of overall *cis*-regulatory content



Conservation of overall *cis*-regulatory content



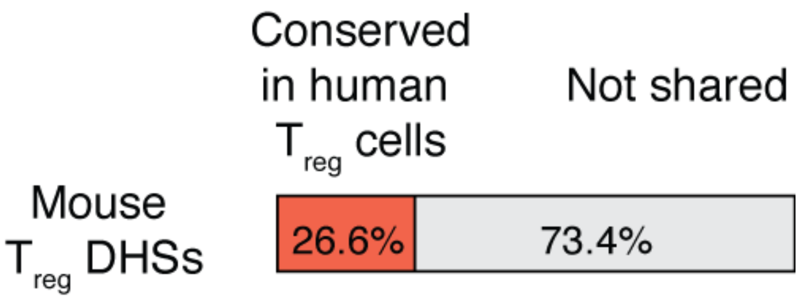
Despite poor conservation of individual binding sites,

the overall proportion of regulatory DNA 'real estate' available to each TF in each organism remains nearly constant

Rigid conservation of global TF recognition landscapes

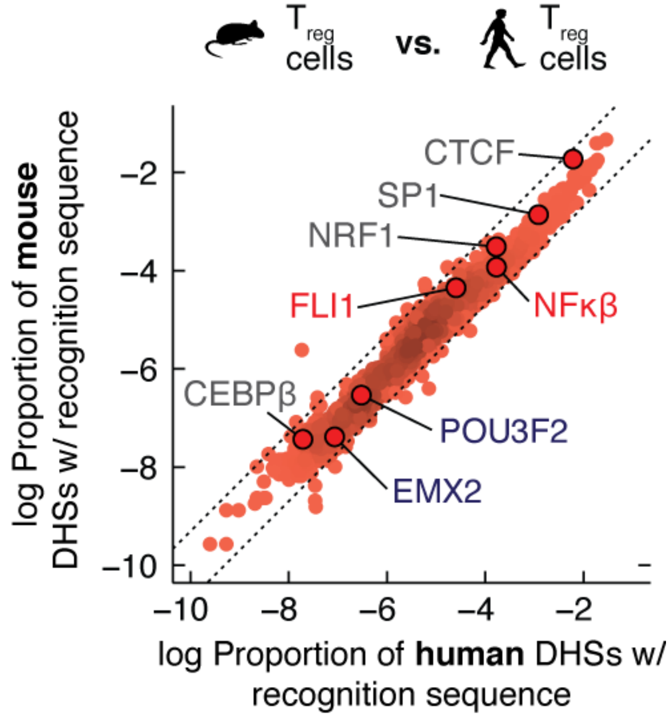
Every TF, every cell type → Different sequence targets, same occupancy fraction

Individual cis-regulatory elements



Many *cis*-regulatory architectures with convergent phenotypic outcomes

Content of cis-regulatory compartment



TF activity patterns are highly constrained (i.e., trans-environment)

Recap

The regulatory DNA landscape has undergone wholesale rewiring during the mouse-human interval

Humans and mice share a core mammalian regulon encoding cell identity and lineage programs

Regulatory DNA landscape evolution involves

- *Extensive repurposing of elements from one tissue context to another*
  - *Continuous ‘re-evolution’ on the same ancestral DNA template*
- *Strict conservation of the proportion of regulatory DNA encoding binding sites for each transcription factor*

Part I:

Mouse and human regulatory regions

Evolutionary dynamics of regulatory DNA regions

Part II:

Transcription factors and networks

Conservation of *trans* vs. *cis* regulatory circuitry

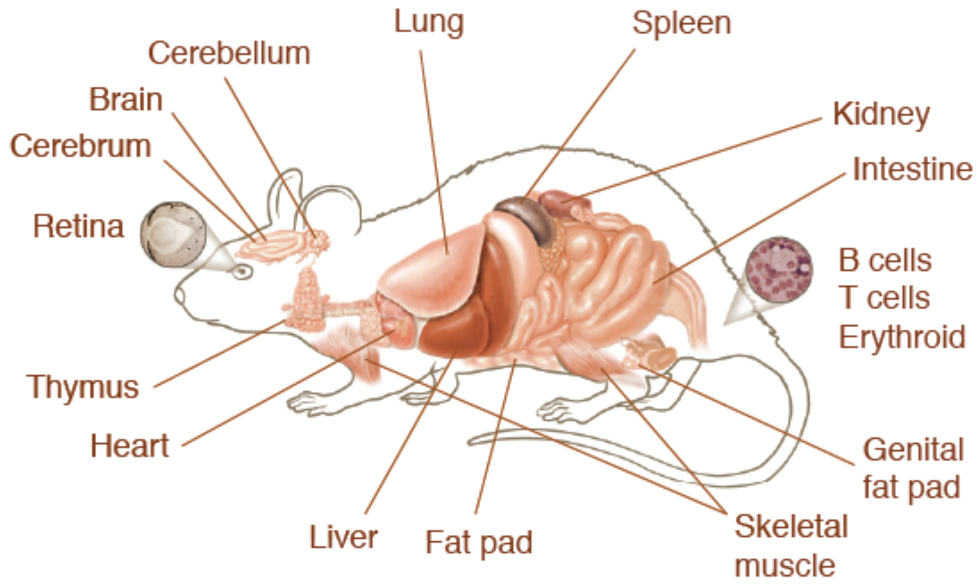
Footprinting the mouse genome

25 mouse cell
and tissue types

~323 million DNaseI cleavages
sequenced per cell type

~1 million DNaseI footprints
identified per cell type
(25.8 million total)

**8.6 million footprinted
DNA elements**



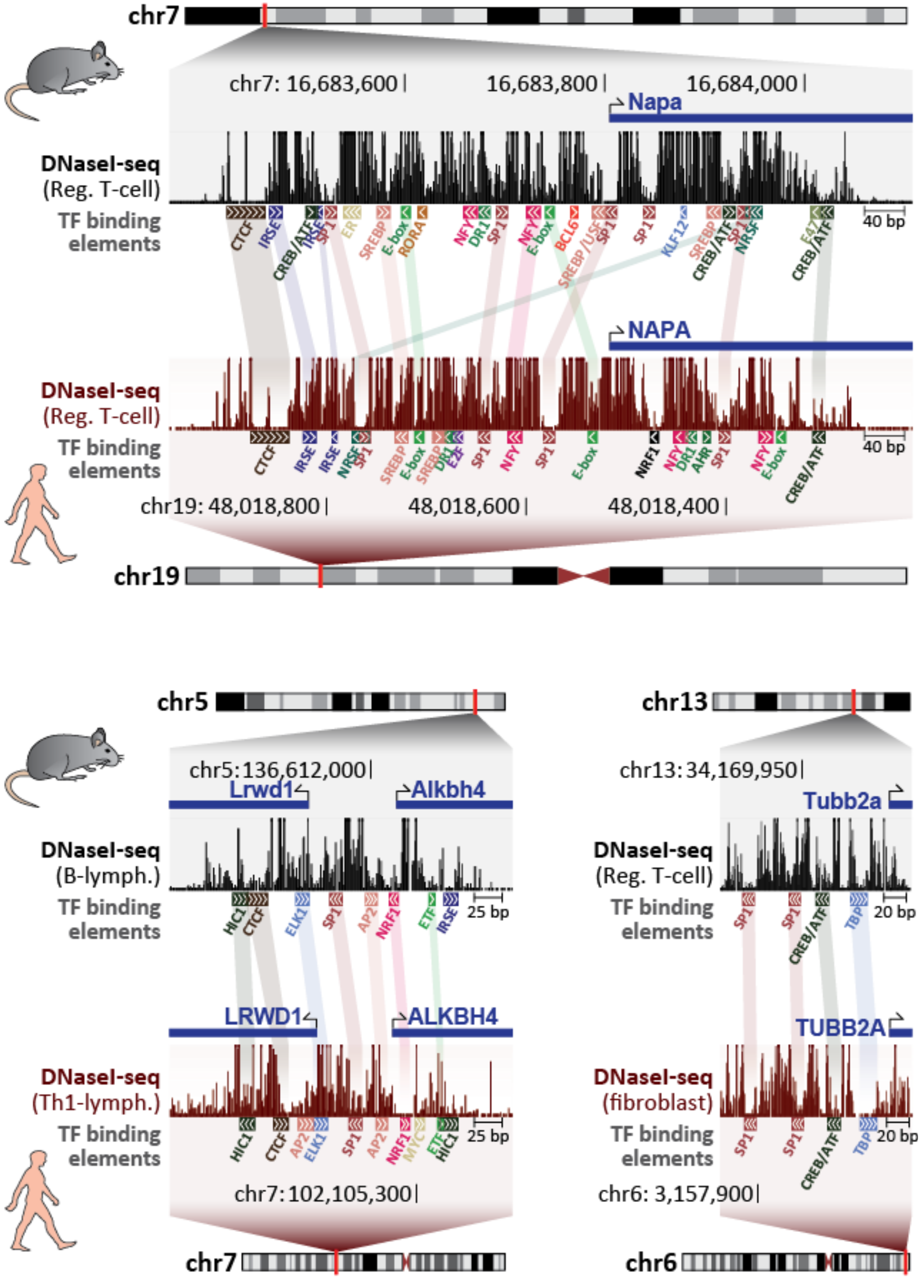
Primitive cells / tissues:
ES cells, limb, embyronic mesoderm

25 mouse cell and tissue types

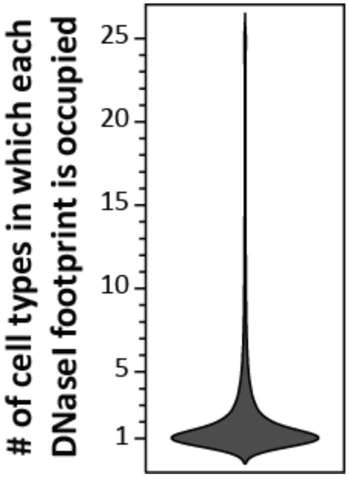
~323 million DNaseI cleavages sequenced per cell type

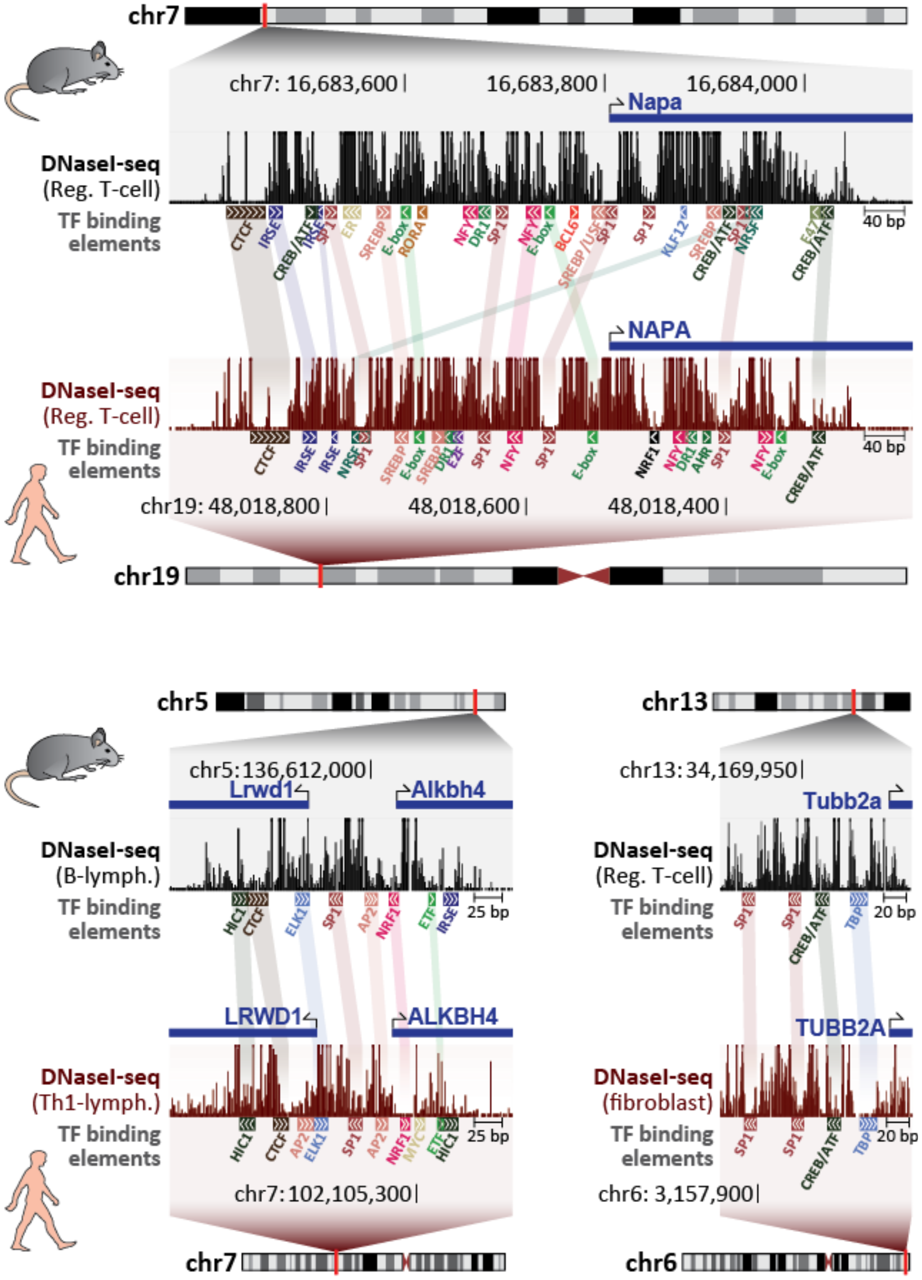
~1 million DNaseI footprints identified per cell type (25.8 million total)

8.6 million footprinted DNA elements

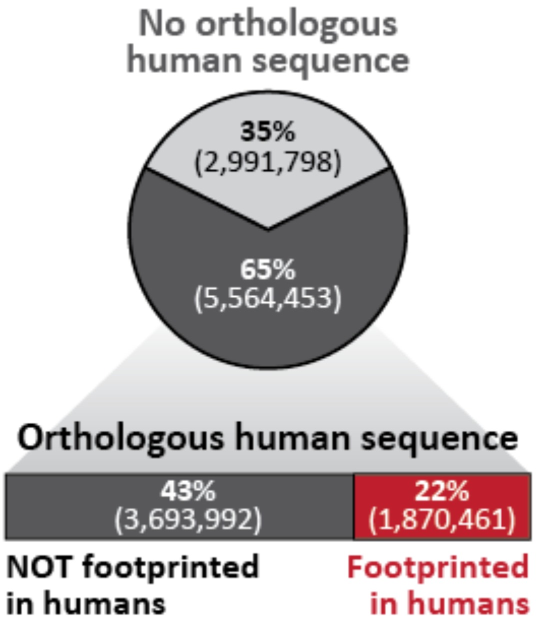


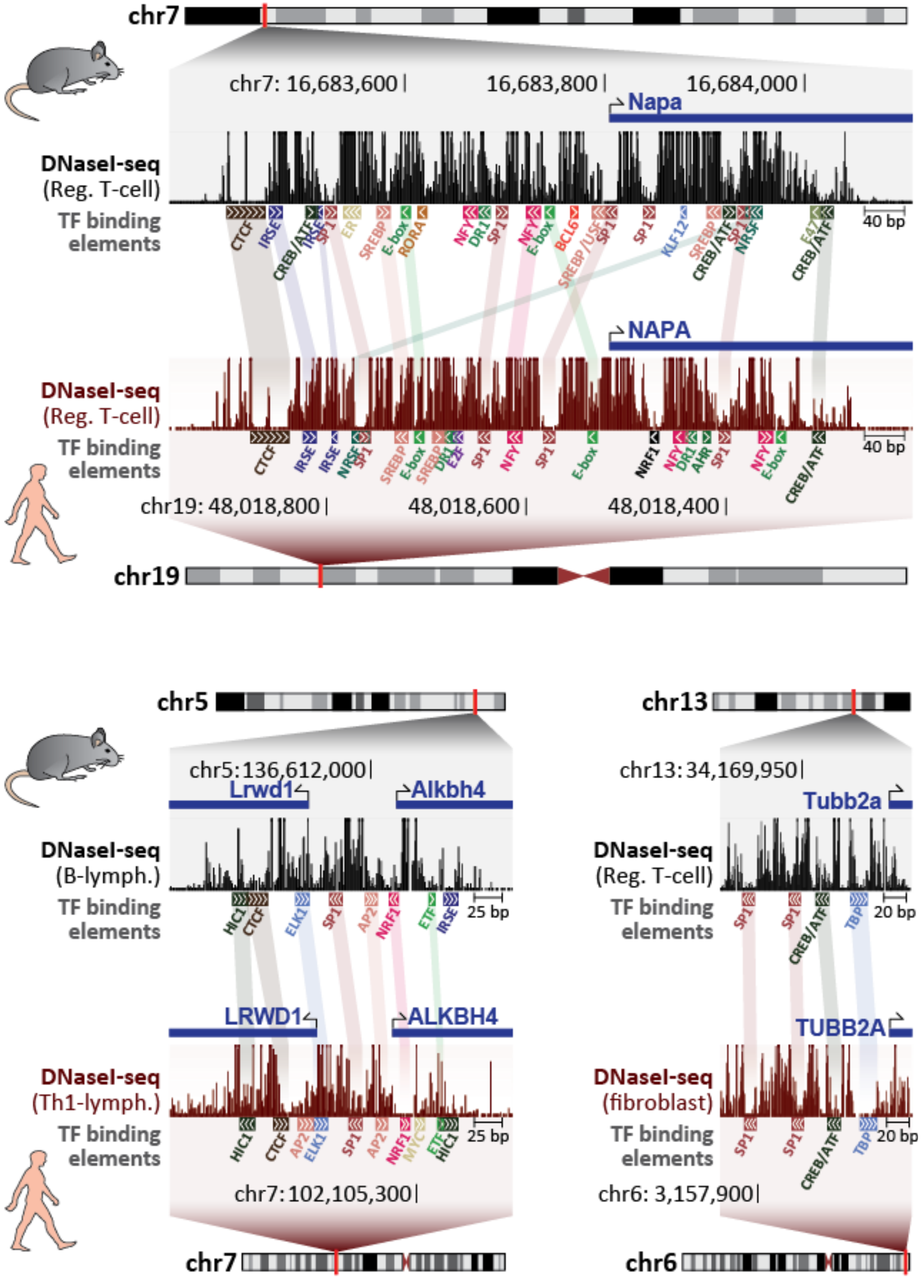
Cell-selectivity of Mouse DNaseI footprints





Sequence and functional constraint of mouse DNaseI footprints





Conservation of TF recognition repertoires

Deriving a mouse *cis*-regulatory lexicon

25.8 million mouse DNaseI footprints

↓

Database independent, *de novo* motif discovery

↓

604 unique motif models

Deriving a mouse *cis*-regulatory lexicon

25.8 million mouse DNaseI footprints

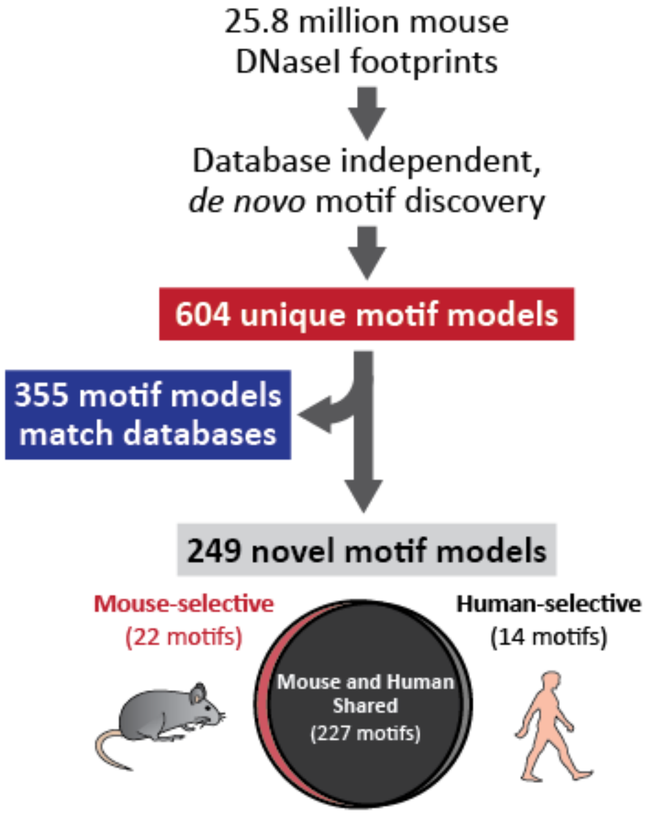
Database independent, *de novo* motif discovery

604 unique motif models

355 motif models match databases

249 novel motif models

Deriving a mouse *cis*-regulatory lexicon



25.8 million mouse
DNaseI footprints

Database independent,
de novo motif discovery

604 unique motif models

**355 motif models
match databases**

249 novel motif models

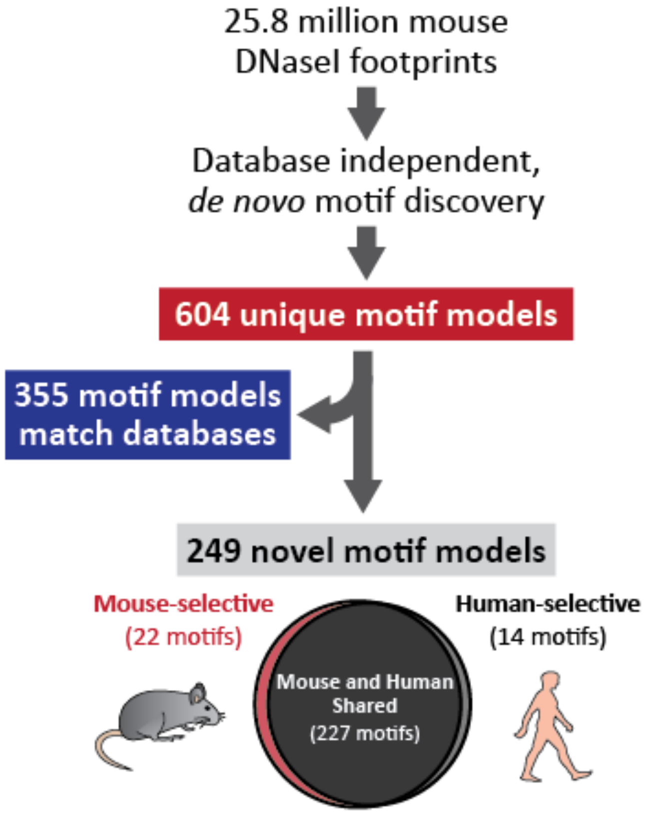
Mouse-selective
(22 motifs)

**Mouse and Human
Shared**
(227 motifs)

Human-selective
(14 motifs)

Human footprint-derived motifs from Neph, Vierstra *et al.* *Nature* 2012

Mouse-specific motifs are largely selective for ES cells



25.8 million mouse DNaseI footprints

Database independent, *de novo* motif discovery

604 unique motif models

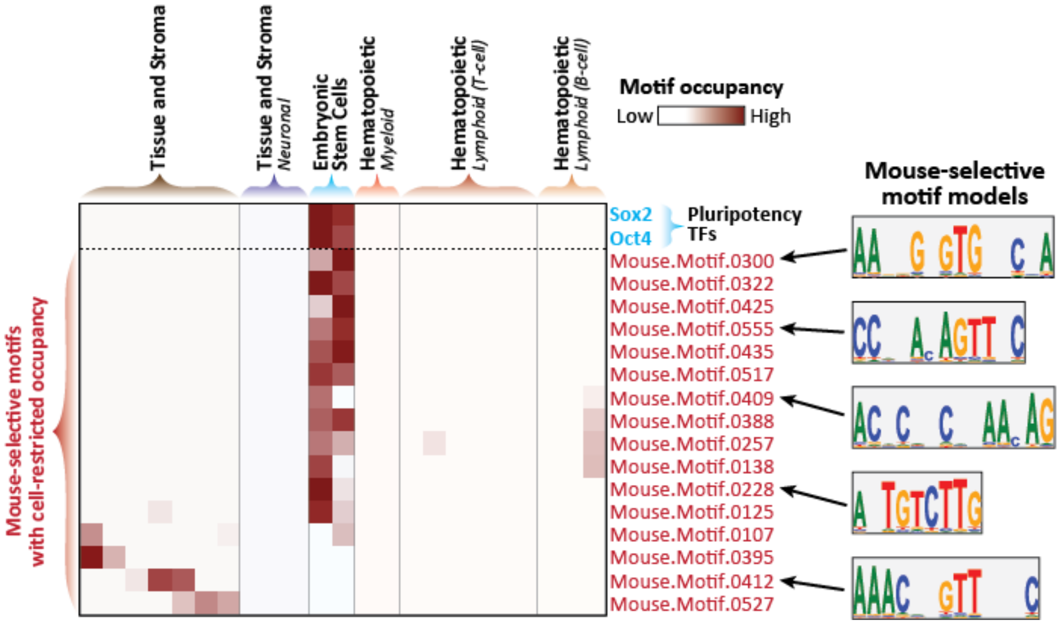
355 motif models match databases

249 novel motif models

Mouse-selective (22 motifs)

Mouse and Human Shared (227 motifs)

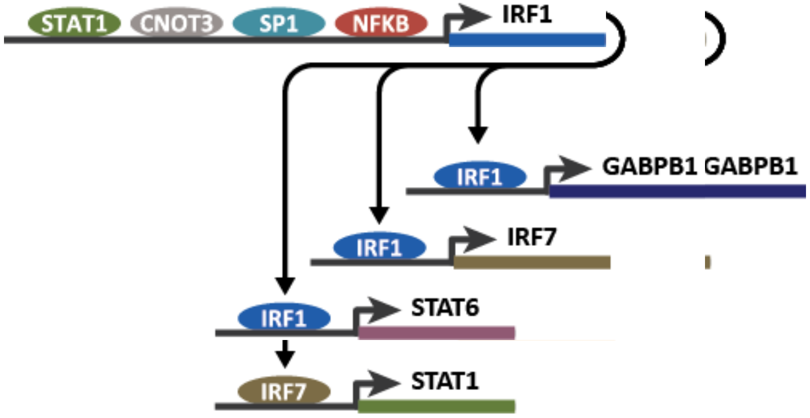
Human-selective (14 motifs)

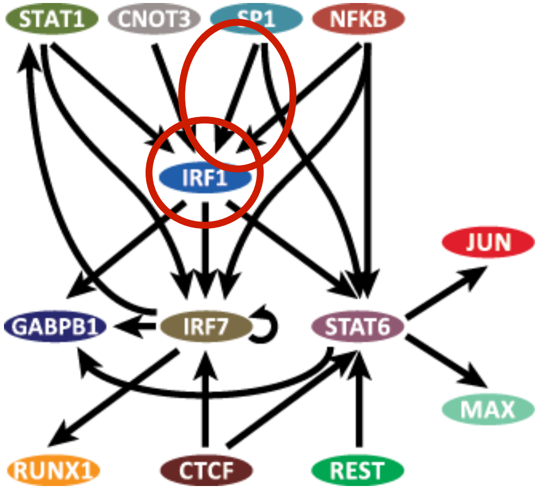


Human footprint-derived motifs from Neph, Vierstra *et al.* *Nature* 2012

Conservation of *trans* regulatory circuitry

Building direct TF networks using TF footprints

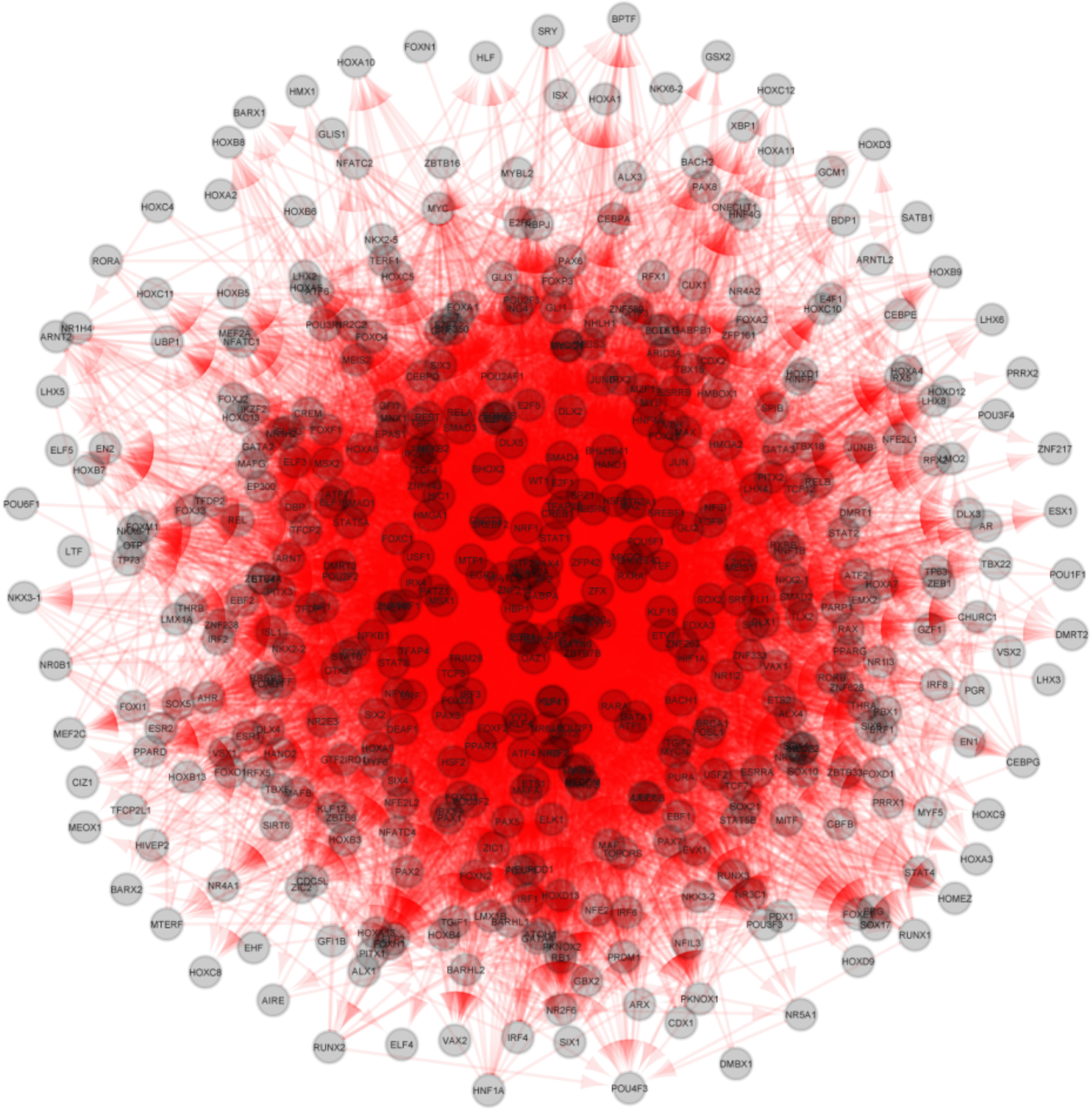




Node: **Transcription factor**

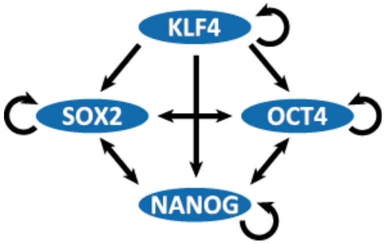
Edge: **Regulatory interaction between 2 TFs**

Neph, Stergachis *et al.* *Cell* 2012

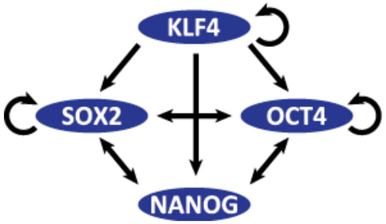

BPTF
SRY
GSX2
HLF
FOXN1
HOXA10
ISX
HOXA1
NKX6-2
HOXC12
HMX1
BARX1
GLIS1
XBP1
HOXB8
HOXD3
HOXA11
NFATC2
ZBTB16
MYBL2
ALX3
BACH2
GCM1
HOXA2
PAX8
HOXC4
HOXB6
MYC
E2F4
RBPJ
CEBPA
ONECUT1
HNF4G
BDP1
SATB1
NKX2-5
TERF1
PAX6
ARNTL2
RORA
RFX1
HOXB9
HOXC5
GLI3
FOXP3
CUX1
NR4A2
E4F1
HOXC11
HOXB5
LHX2
HOXA5
ATF6
POU5F1
NR4A1
GLI1
FOXA1
HOXC10
CEBPE
LHX6
ZNF589
NR1H4
ARNT2
UBP1
MEF2A
NFATC1
POU3F1
NR2C2
FOXO4
NHLH1
BCL6
GABPB1
FOXA2
ZFP161
MEIS2
MEIS3
ARID3A
LHX5
SIX3
CEBPG
POU2AF1
JUN
TBX15
CDX2
HOXD1
HOXA4
PRRX2
HOXD12
LHX8
FOXJ2
IKZF2
HOXC13
CREM
REST
RELA
E2F5
DLX2
ESRRB
HMBOX1
SPIB
POU3F4
NFE2L1
ZNF217
ELF5
EN2
HOXB7
GATA3
MAFG
FOXF1
SMAD3
DLX5
SHOX2
HNF1B
MAX
HMGA2
GATA1
TBX18
JUNB
RFX3
MSX2
POU6F1
ELF3
EP300
HOXA9
TCF4
BHLHE41
HAND1
SMAD4
WT1
E2F1
RELB
PITX2
DMRT1
LHX4
ESX1
DLX3
AR
LTF
FOXM1
OTP
TP73
TFDP2
FOXJ3
DBP
REL
SMAD1
STAT5A
HMGA1
CREB1
NRF1
STAT1
SREBF1
NFIB
GLI2
STAT2
TBX22
TP63
ZEB1
POU1F1
NKX3-1
ARNT
ZBTB4
FOXC1
USF1
MTF1
ZFP42
ATF2
MYOD
HOXA7
EMX2
CHURC1
LMX1A
THRB
EBF2
IRX4
ZFX
SOX2
FLI1
PARP1
GZF1
DMRT2
ZNF238
IRF2
ISL1
NKX2-2
NFKB1
CDX1
ATBF1
ETV1
HNF1A
RAX
VSX2
PPARG
ZNF333
VAX1
NR1I3
LHX3
NR0B1
STAT6
STAT3
TCF3
NR1I2
RORB
IRF8
PGR
ZNF628
THRA
PBX1
FOXI1
SOX5
AHR
MYF6
BACH1
RARA
ETS2
ESR2
ESR1
DLX4
NR2E3
PAX5
SIX2
DEAF1
MEF2C
PPARD
NR3C1
HSF2
PPARA
ATF4
EN1
CEBPG
CIZ1
HOXB13
VSX1
HAND2
GTF2IRD1
SOX10
ZBTB33
FOXD1
FOXO1
RFX5
TBX5
MAFB
SIX4
MYCN
USF2
ESRRA
TCF7
PRRX1
MYF5
HOXC9
TFCP2L1
KLF12
ZBTB6
NFE2L2
SIRT6
MITF
CBFB
MEOX1
HIVEP2
HOXB3
NFATC4
PAX4
ELK1
EBF1
SOX21
STAT5B
HOXA3
PAX2
ZIC1
MAF
PAX7
EVX1
RUNX3
STAT4
HOMEZ
BARX2
NR4A1
CDC5L
ZIC2
FOXN2
NEUROD1
IRF1
HOXD13
NKX3-2
MTERF
PDX1
ERG
FOXP1
SOX17
RUNX1
POU3F3
NFE2
IRF6
EHF
GFI1B
HOXA13
FOXH1
HOXB4
TGIF1
LMX1B
BARHL1
ATOH1
NFIL3
PRDM1
HOXD9
HOXC8
PITX1
ALX1
PKNOX2
RB1
BARHL2
GBX2
AIRE
NR2F6
ARX
PKNOX1
NR5A1
CDX1
RUNX2
ELF4
VAX2
IRF4
SIX1
DMBX1
HNF1A
POU4F3

Direct TF footprint-derived networks accurately recapitulate known TF network relationships

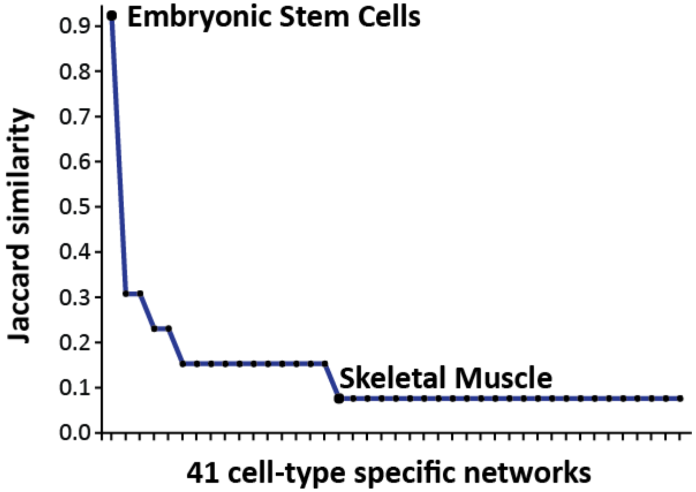
Mouse Embryonic Stem Cell Network



Human Embryonic Stem Cell Network



Kim et al., 2008

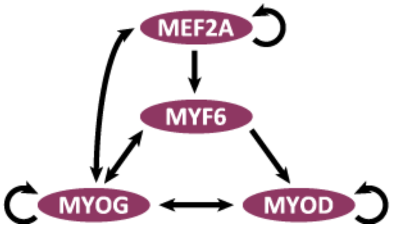


Jaccard similarity index

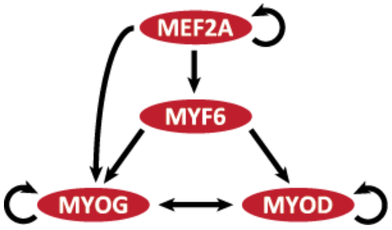
1 = Identical networks

0 = Completely different networks

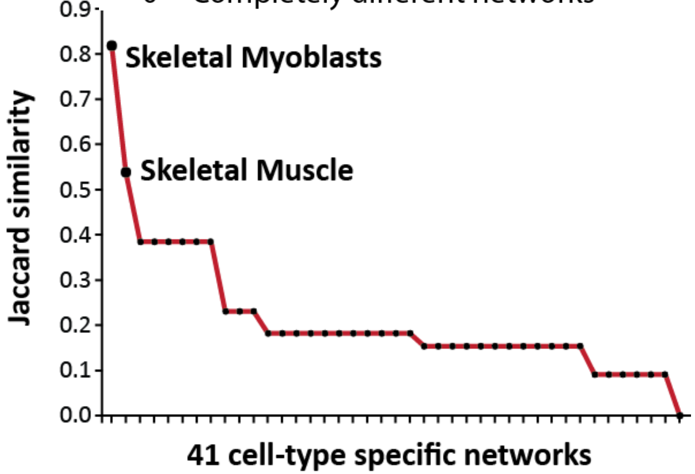
Skeletal muscle developmental Network



Skeletal Muscle Myoblast Network

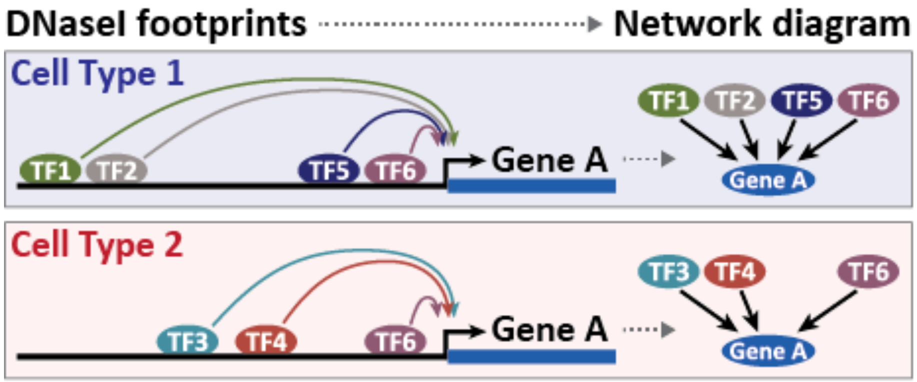


Naidu *et al.* 1995; Yun & Wold 1996; Ramachandran *et al.* 2008

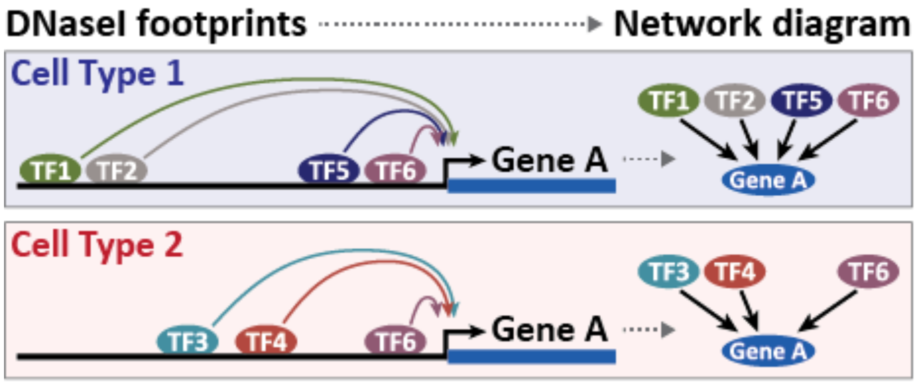


Neph, Stergachis *et al.* *Cell* 2012

TF-to-TF connections are cell-selective

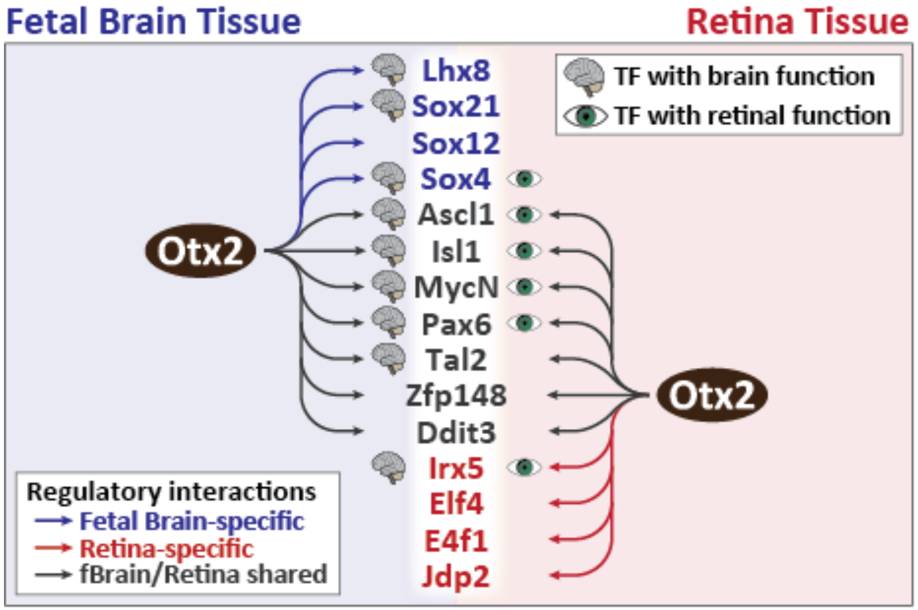


TF-to-TF connections are cell-selective


DNaseI footprints → Network diagram

Cell Type 1
TF1 TF2 TF5 TF6
Gene A
TF1 TF2 TF5 TF6
Gene A

Cell Type 2
TF3 TF4 TF6
Gene A
TF3 TF4 TF6
Gene A

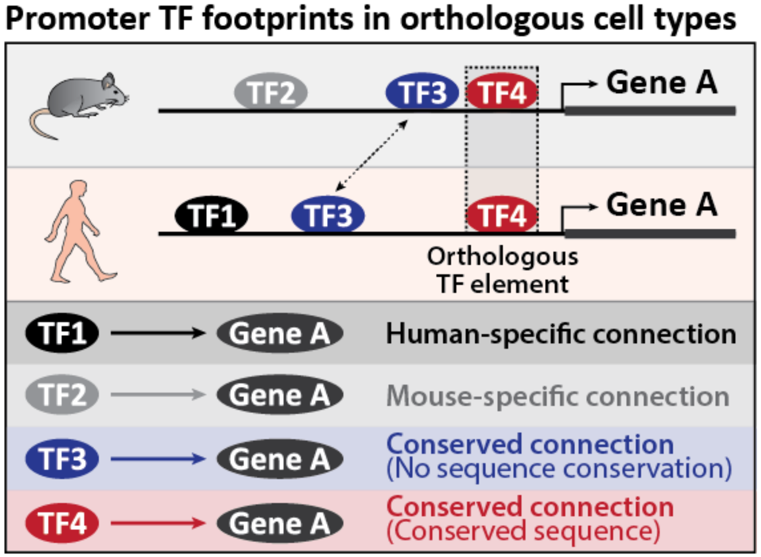

Fetal Brain Tissue
Retina Tissue

🧠 TF with brain function
👁 TF with retinal function

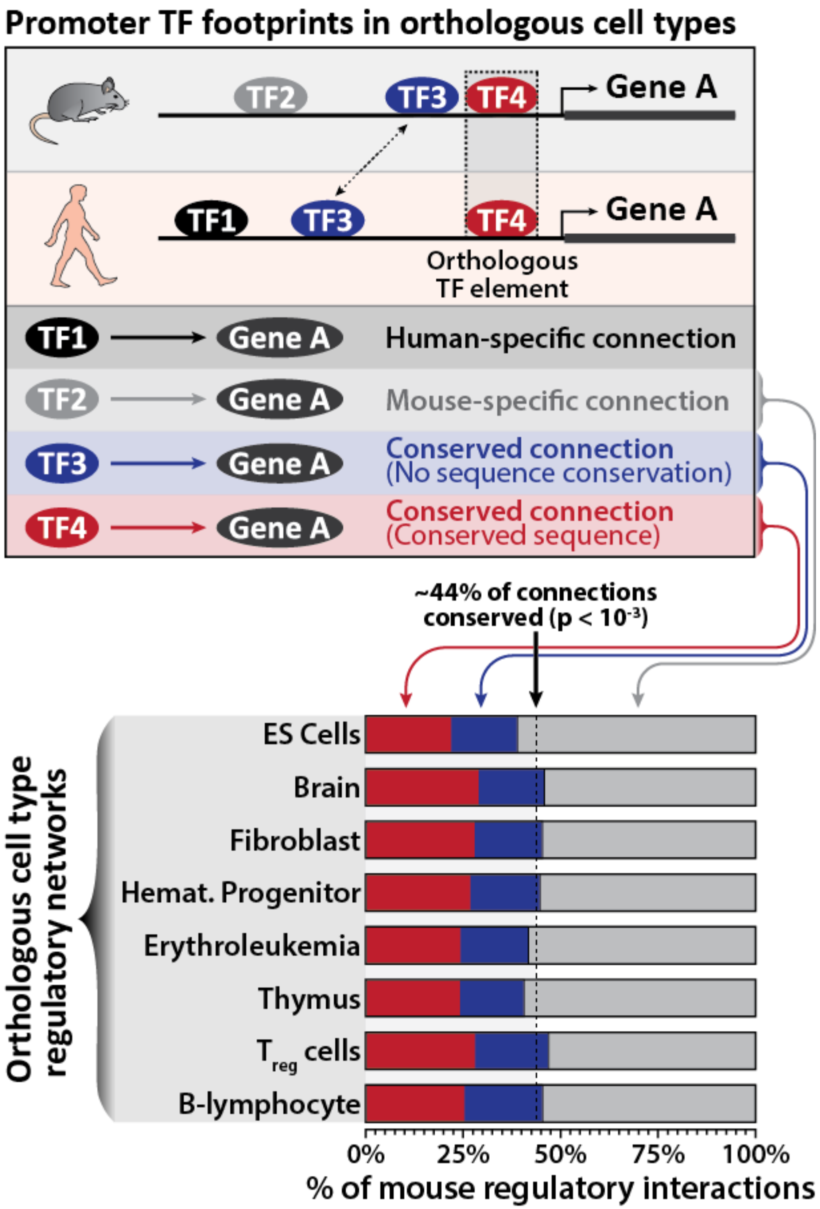
Otx2
Lhx8
Sox21
Sox12
Sox4
Ascl1
Isl1
MycN
Pax6
Tal2
Zfp148
Ddit3
Irx5
Elf4
E4f1
Jdp2
Otx2

Regulatory interactions
→ Fetal Brain-specific
→ Retina-specific
→ fBrain/Retina shared

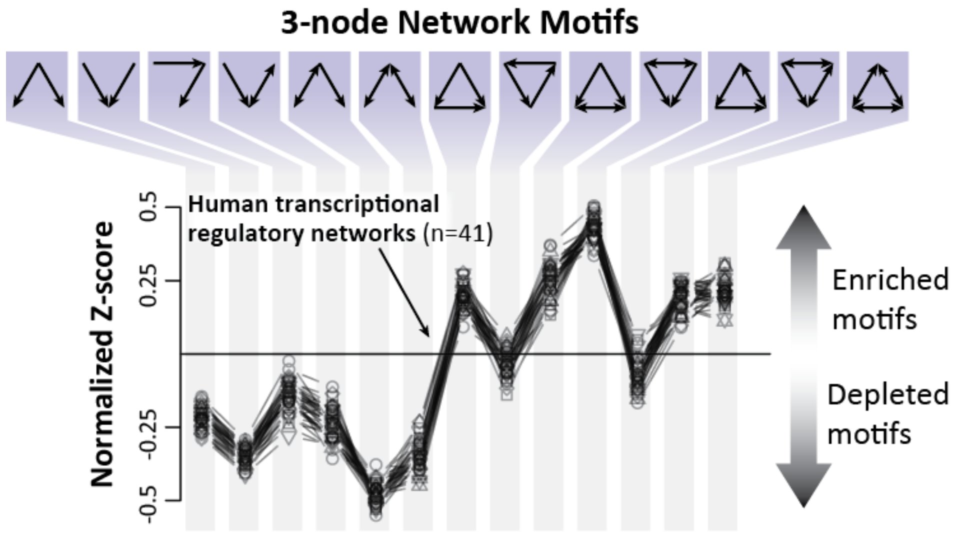
Conservation of TF-to-TF connections



Conservation of TF-to-TF connections



Conservation of global TF network architecture

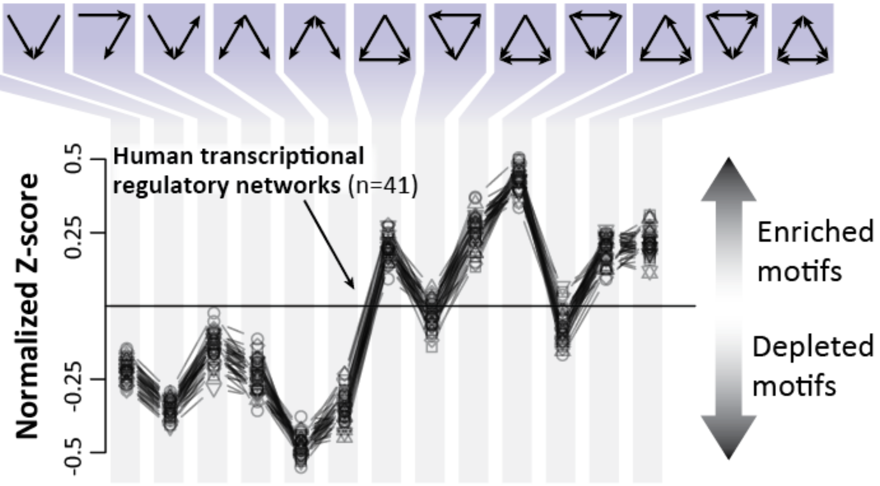


3-node Network Motifs

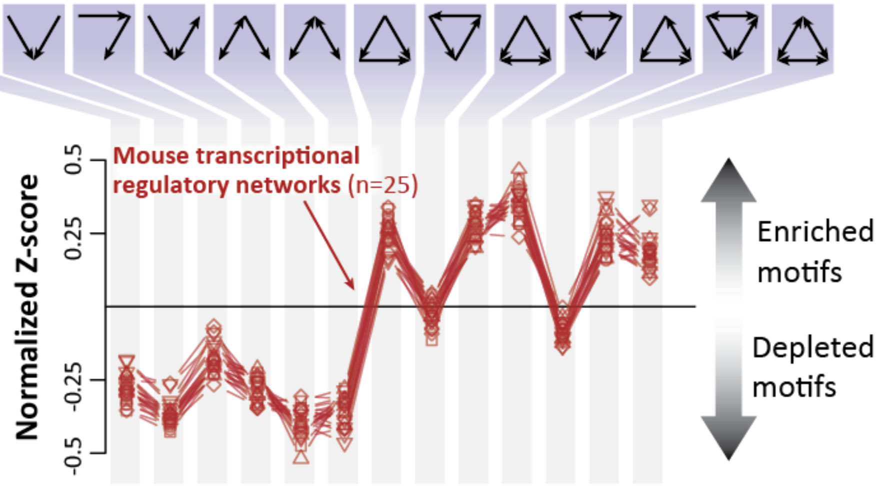
Neph, Stergachis *et al.* *Cell* 2012

Conservation of global TF network architecture

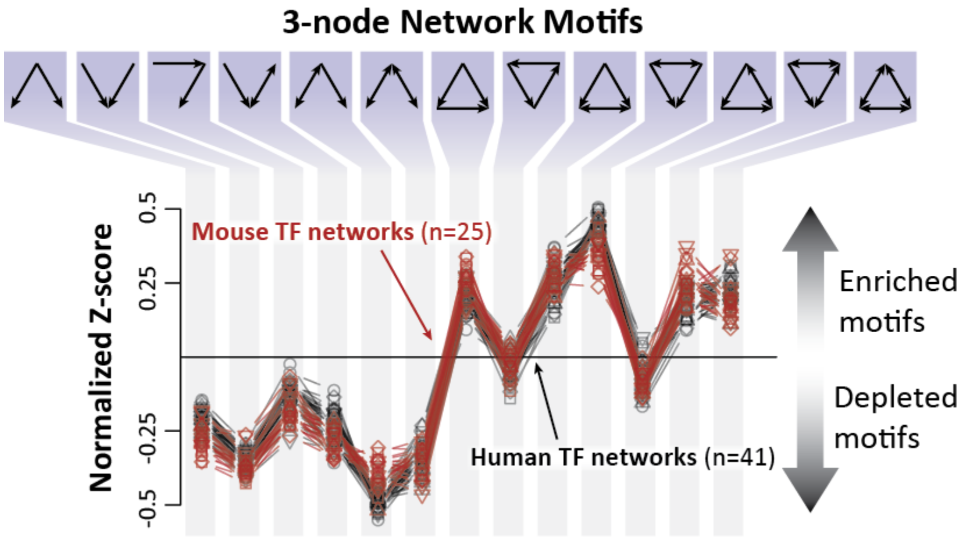
3-node Network Motifs



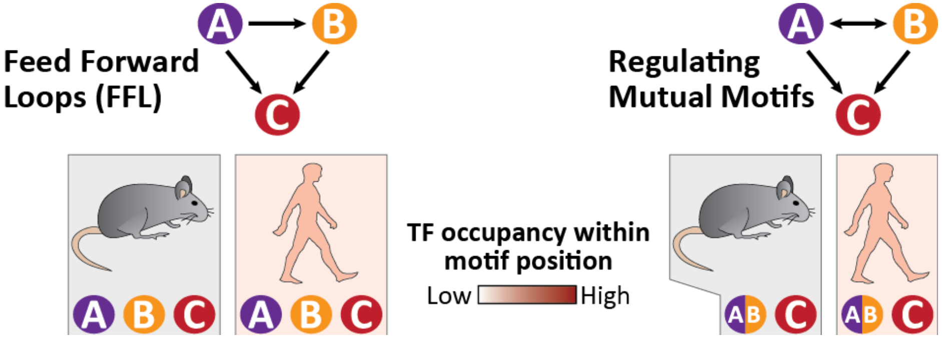
3-node Network Motifs



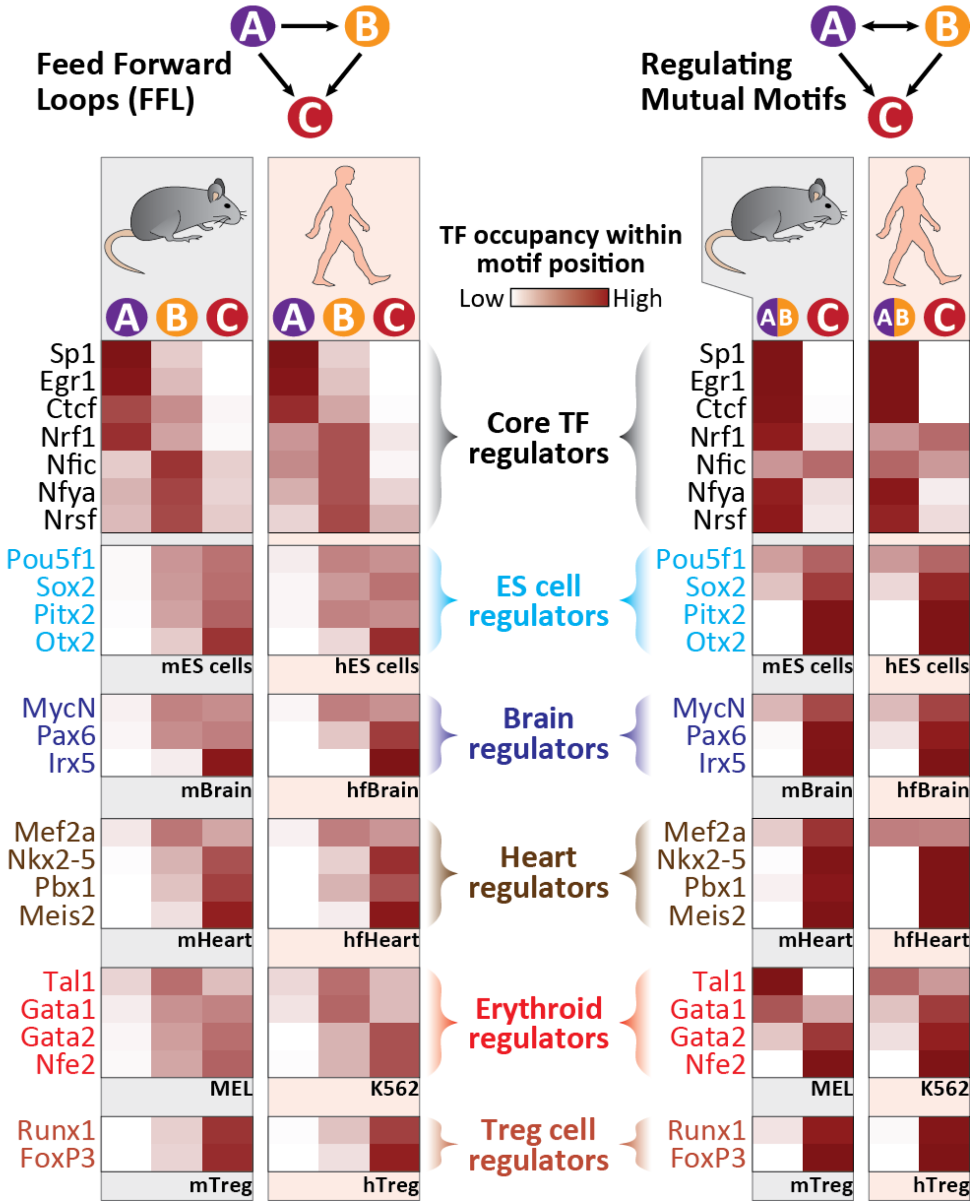
Conservation of global TF network architecture



Conservation of fine network architecture



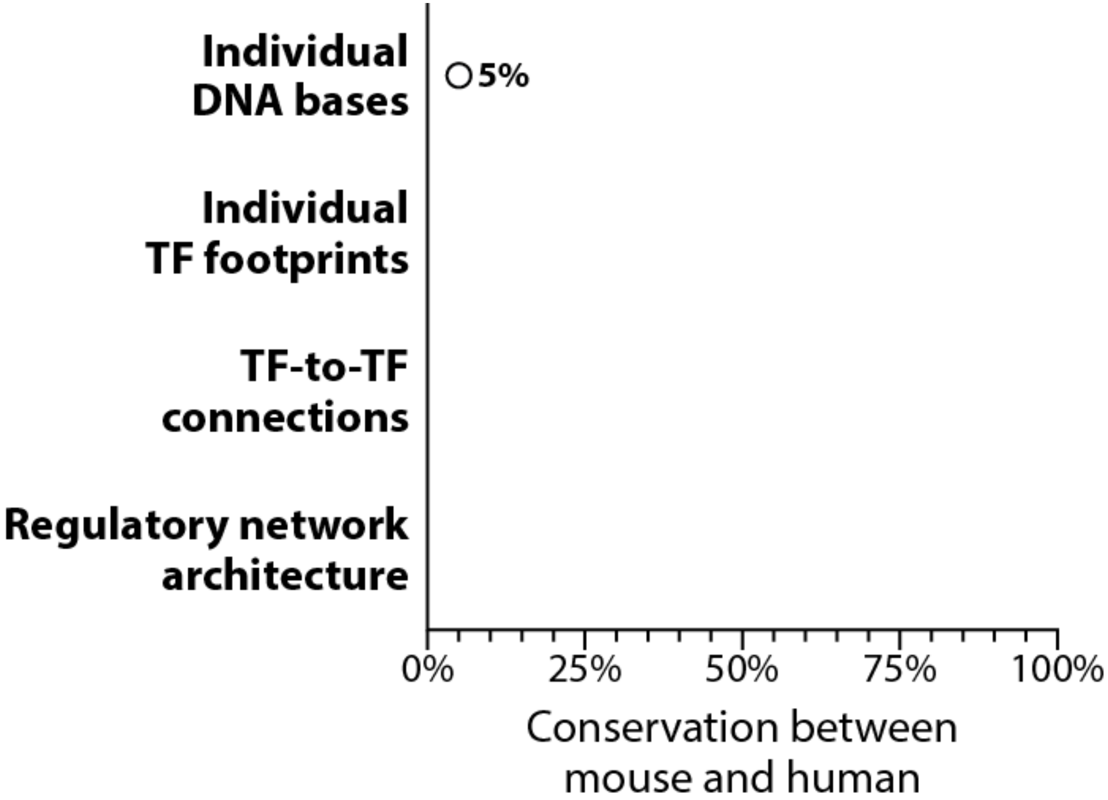
Conservation of fine network architecture



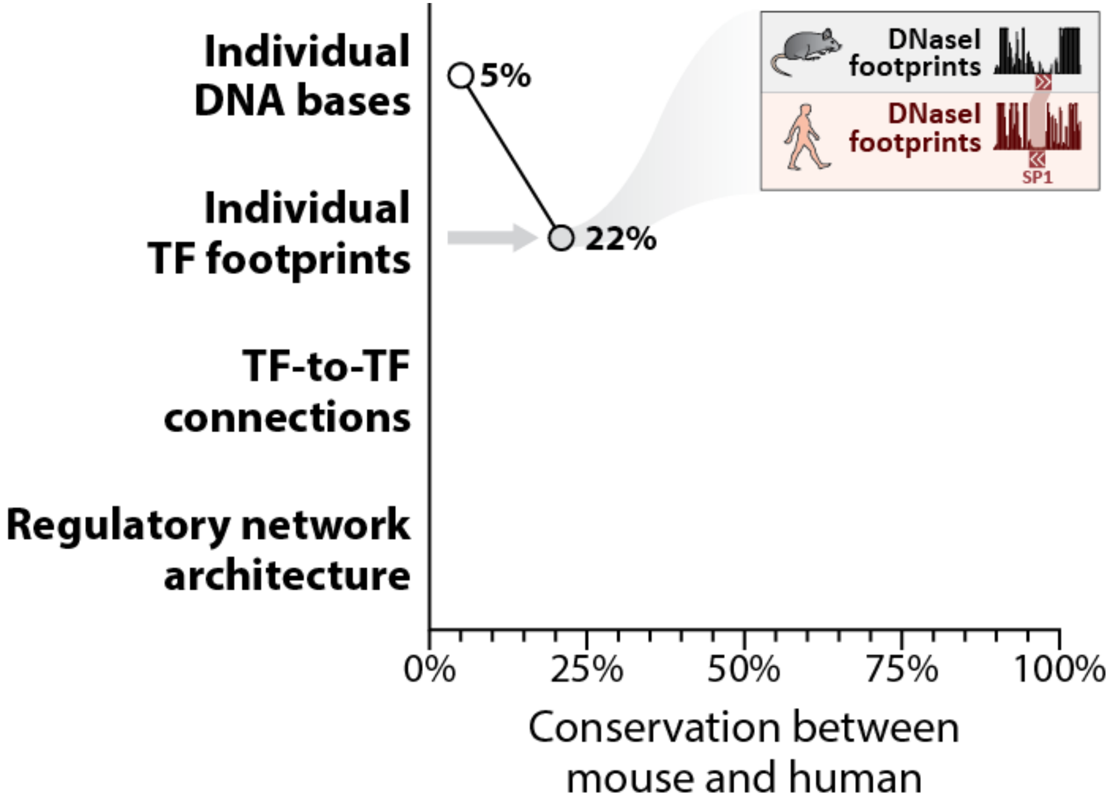
Stepping away from the genome:

Where evolution is really acting

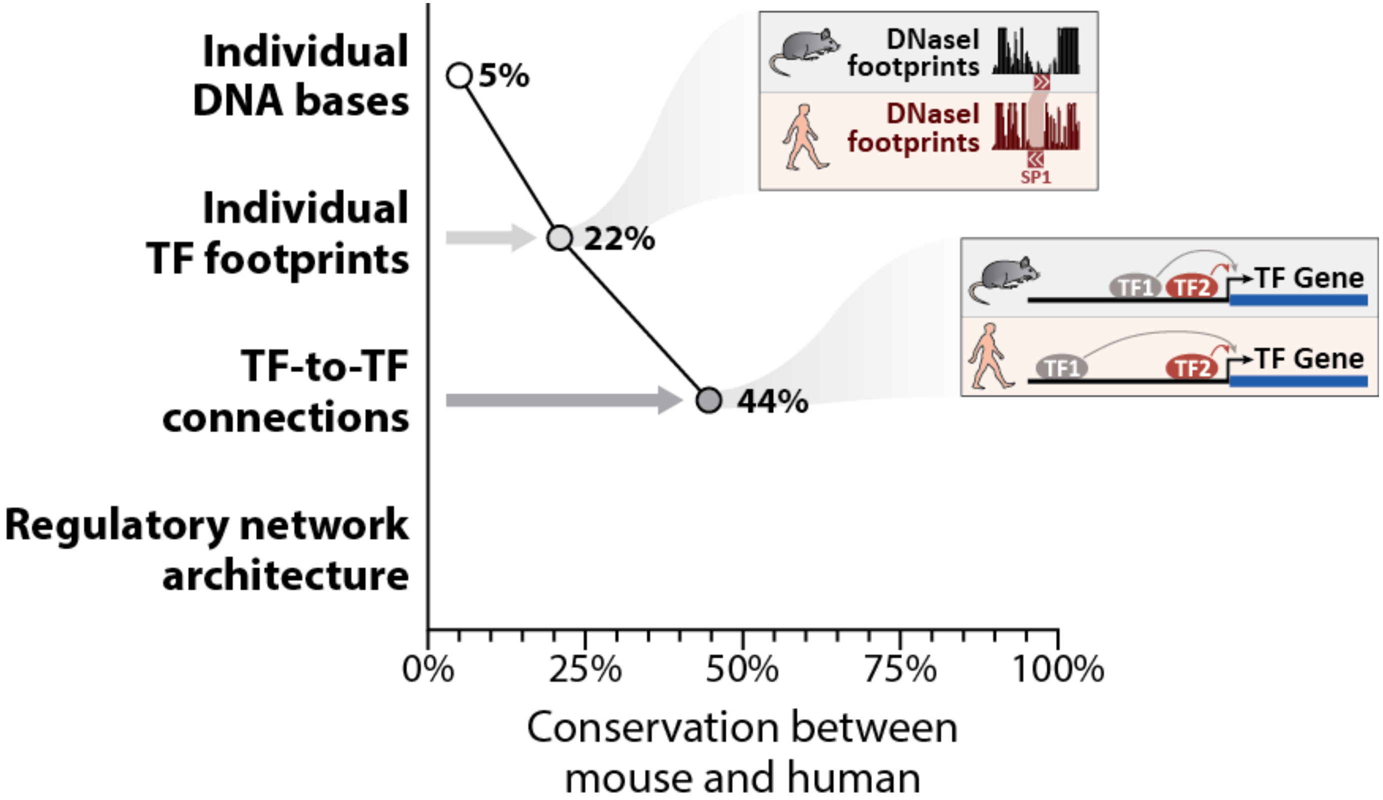
Evolutionary engineering of a regulatory genome



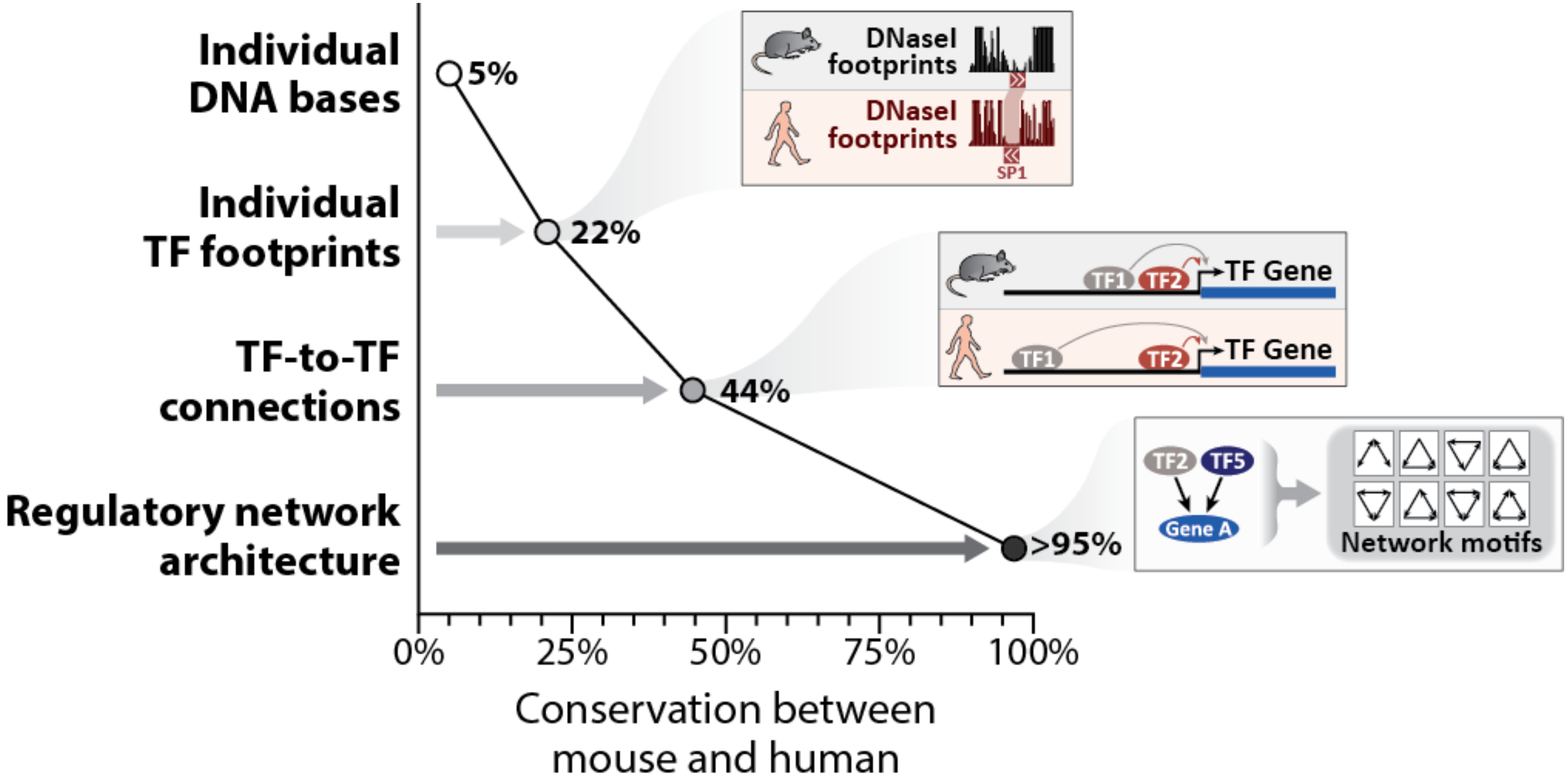
Evolutionary engineering of a regulatory genome



Evolutionary engineering of a regulatory genome



Evolutionary engineering of a regulatory genome



Acknowledgements - Mouse

Key experiments/analyses
Jeff Vierstra, Andrew Stergachis, Shane Neph
Eric Rynes, Matthew Maurano, Eric Haugen,
Richard Sandstrom , Richard Humbert,
Bob Thurman, Alex Reynolds,
Benjamin Vernot, Wenqing Fu

Mouse DNaseI pipeline group
Miaohua Zhang, Rachael Byron, Peter Sabo,
Theresa Cantwell, Kristen Lee
Shinny Vong, Vaughan Roach, Erica Giste
Sandra Stehling-Sun

Collaborators
Mark Groudine (FHCRC)
M.A. Bender (FHCRC)
Piper Treuting (UW Comparative Med)
Elhanan Borenstein (UW Genome Sciences)
Joshua Akey (UW Genome Sciences)
Shelly Heimfeld (FHCRC)
Mouse ENCODE Consortium

Funding: NHGRI (MouseENCODE)