# DNA ZOO ATAC PAPERS DICUSSION

2022/07/29

# CURRENT DNAZOO DATA SUMMARY

# ATAC-seq data summary:

- **120** species (**118** + human and mouse)
- Libraries generated for **118**
- Sequenced for **77**
- Aligned against official release: **16**
- Aligned against draft/pre-release: **23**
- Any alignment: **39**

| Subclass / Infraclass | order | # species |
|---|---|---|
| **Monotremata** | Monotremata | |
| **Metatheria** | Didelphimorphia | 1 |
| | Paucituberculata | |
| | Microbiotheria | |
| | Dasyuromorphia | |
| | Peramelemorphia | |
| | Notoryctemorphia | |
| | Diprotodontia | 3 |
| **Eutheria** | Macroscelidea | |
| | Afrosoricida | |
| | Tubulidentata | |
| | Proboscidea | 0 |
| | Hyracoidea | |
| | Sirenia | 0 |
| | Cingulata | |
| | Pilosa | 0 |
| | Scandentia | 1 |
| | Dermoptera | |
| | Primates | 21 |
| | Rodentia | 39 |
| | Lagomorpha | 6 |
| | Erinaceomorpha | |
| | Soricomorpha | |
| | Chiroptera | 13 |
| | Pholidota | |
| | Carnivora | 26 |
| | Artiodactyla | 6 |
| | Perissodactyla | 2 |
| | Total | 118 |

# PREVIOUS ENCODE WORK

ARTICLE

OPEN

doi:10.1038/nature13992

# A comparative encyclopedia of DNA elements in the mouse genome

A list of authors and their affiliations appears at the end of the paper

- Although much conservation exists, the expression profiles of many mouse genes involved in distinct biological pathways show considerable divergence from their human orthologues.
- A large portion of the *cis*-regulatory landscape has diverged between mouse and human, although the magnitude of regulatory DNA divergence varies widely between different classes of elements active in different tissue contexts.
- Mouse and human transcription factor networks are substantially more conserved than *cis*-regulatory DNA.
- Species-specific candidate regulatory sequences are significantly enriched for particular classes of repetitive DNA elements.
- Chromatin state landscape in a cell lineage is relatively stable in both human and mouse.
- Chromatin domains, interrogated through genome-wide analysis of DNA replication timing, are developmentally stable and evolutionarily conserved.

ARTICLE

OPEN

doi:10.1038/nature13985

# Principles of regulatory information conservation between mouse and human

Yong Cheng[1]*, Zhihai Ma[1]*, Bong-Hyun Kim[2], Weisheng Wu[3,4], Philip Cayting[1], Alan P. Boyle[1], Vasavi Sundaram[5], Xiaoyun Xing[5], Nergiz Dogan[3], Jingjing Li[1], Ghia Euskirchen[1], Shin Lin[1,6], Yiing Lin[1,7], Axel Visel[8,9,10], Trupti Kawli[1], Xinqiong Yang[1], Dorrelyn Patacsil[1], Cheryl A. Keller[3], Belinda Giardine[3], The Mouse ENCODE Consortium†, Anshul Kundaje[1], Ting Wang[5], Len A. Pennacchio[8,9], Zhiping Weng[2], Ross C. Hardison[3]§ & Michael P. Snyder[1]§

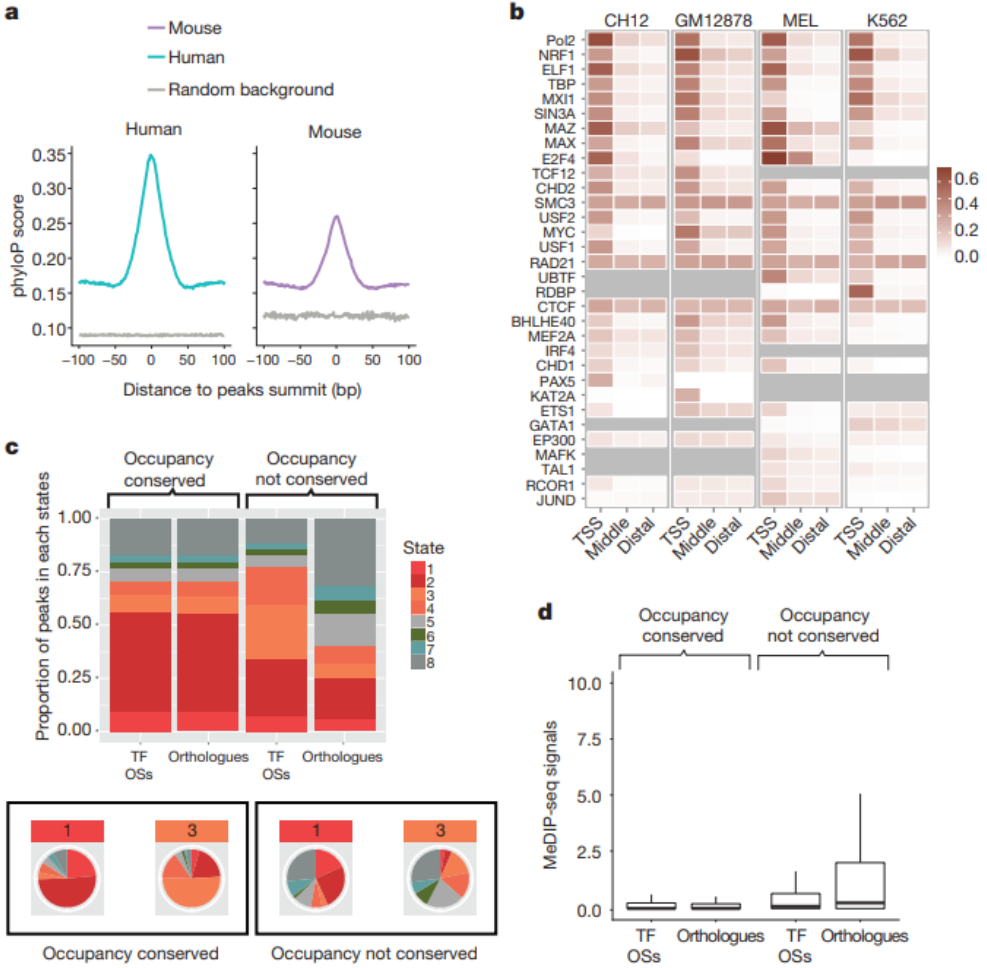



**Figure 2 | Conservation and divergence of TF OSs.** **a**, Blue and purple lines represent the average phyloP score distribution near (±100 bp) the ChIP-seq peak summit in human and mouse. The grey line represents the distribution for randomly selected background sequences. The *x* axis is the distance to the peak summit, and the *y* axis is the average phyloP score. **b**, The heat map represents the occupancy conservation of TF (rows) OSs in the four cell lines. The colour intensity represents the proportion of TF OSs for which occupancy is conserved between mouse and human in different genomic regions (columns). **c**, Comparison of the chromatin state change between TF OSs and orthologous sequences. TF OSs that can be aligned between mouse and human are divided into two groups according to the occupancy conservation status ('occupancy conserved' versus 'occupancy not conserved'). Top, the *y* axis is the proportion of TF OSs and their orthologous sequences in each chromatin state. Bottom, detailed chromatin state change in human orthologues for mouse TF OSs in chromatin states 1 and 3. The pie charts show the distribution of chromatin states in the orthologous sequence in the second species. **d**, Comparison of the DNA methylation change between TF OSs and orthologous sequences. The *y* axis gives the normalized DNA methylation signals (MeDIP-seq). TF OSs are divided into two categories according to the occupancy conservation status as in **c**.

ARTICLE

OPEN

doi:10.1038/nature13972

# Conservation of trans-acting circuitry during mammalian regulatory evolution

Andrew B. Stergachis[1]*, Shane Neph[1]*, Richard Sandstrom[1], Eric Haugen[1], Alex P. Reynolds[1], Miaohua Zhang[2], Rachel Byron[2], Theresa Canfield[1], Sandra Stelhing-Sun[1], Kristen Lee[1], Robert E. Thurman[1], Shinny Vong[1], Daniel Bates[1], Fidencio Neri[1], Morgan Diegel[1], Erika Giste[1], Douglas Dunn[1], Jeff Vierstra[1], R. Scott Hansen[1,3], Audra K. Johnson[1], Peter J. Sabo[1], Matthew S. Wilken[4], Thomas A. Reh[4], Piper M. Treuting[5], Rajinder Kaul[1,3], Mark Groudine[2,6], M. A. Bender[7,8], Elhanan Borenstein[1,9,10] & John A. Stamatoyannopoulos[1,3]

**The basic body plan and major physiological axes have been highly conserved during mammalian evolution, yet only a small fraction of the human genome sequence appears to be subject to evolutionary constraint. To quantify cis- versus trans-acting contributions to mammalian regulatory evolution, we performed genomic DNase I footprinting of the mouse genome across 25 cell and tissue types, collectively defining ~8.6 million transcription factor (TF) occupancy sites at nucleotide resolution. Here we show that mouse TF footprints conjointly encode a regulatory lexicon that is ~95% similar with that derived from human TF footprints. However, only ~20% of mouse TF footprints have human orthologues. Despite substantial turnover of the cis-regulatory landscape, nearly half of all pairwise regulatory interactions connecting mouse TF genes have been maintained in orthologous human cell types through evolutionary innovation of TF recognition sequences. Furthermore, the higher-level organization of mouse TF-to-TF connections into cellular network architectures is nearly identical with human. Our results indicate that evolutionary selection on mammalian gene regulation is targeted chiefly at the level of trans-regulatory circuitry, enabling and potentiating cis-regulatory plasticity.**

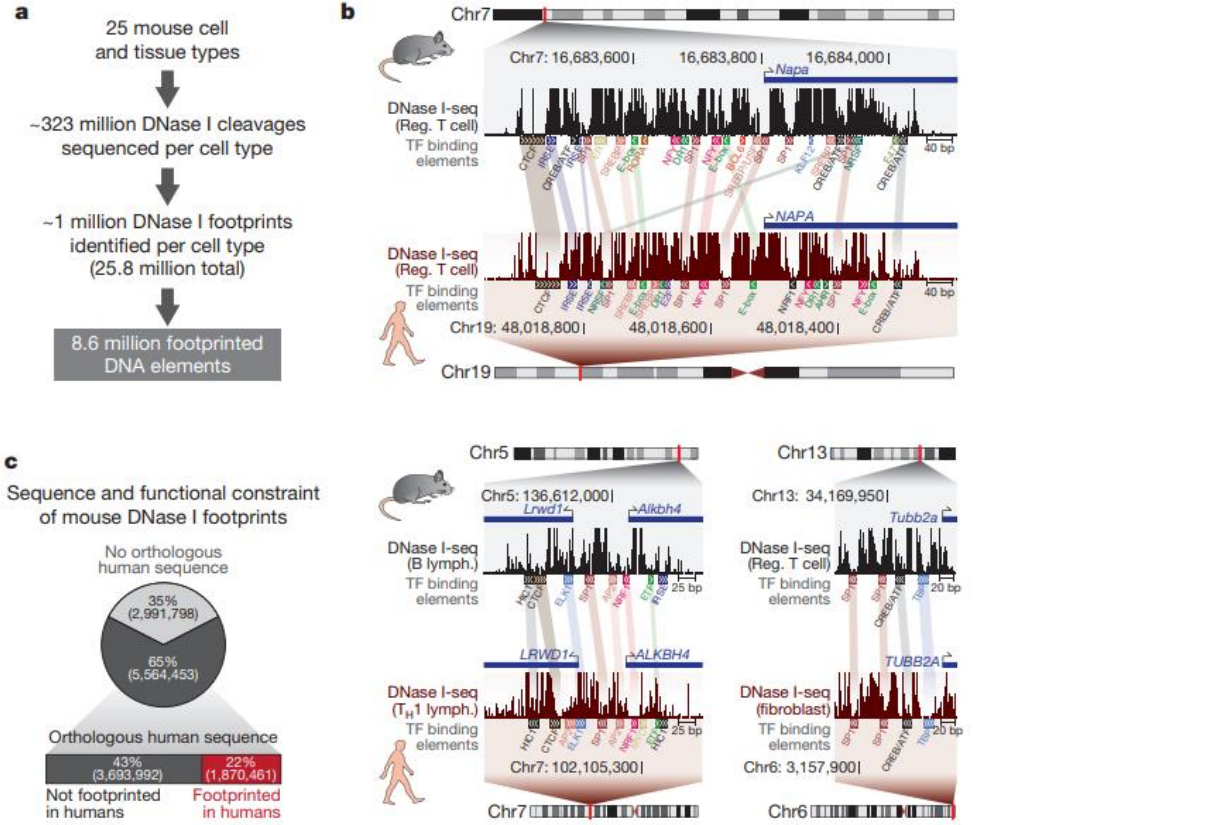



**Figure 1 | Footprinting the mouse genome and comparison with human footprints.** **a**, Derivation of 8.6 million differentially occupied DNase I footprints from 25 mouse cell and tissue types. **b**, Per-nucleotide DNase I cleavage across three gene promoters in both mouse and human cell types; shared TF occupancy sites are indicated by faded boxes. **c**, Percentage of mouse DNase I footprints with sequence aligning to the human genome but not occupied in any human cell type (grey) versus aligning footprints that are occupied in one or more human cell type (red).

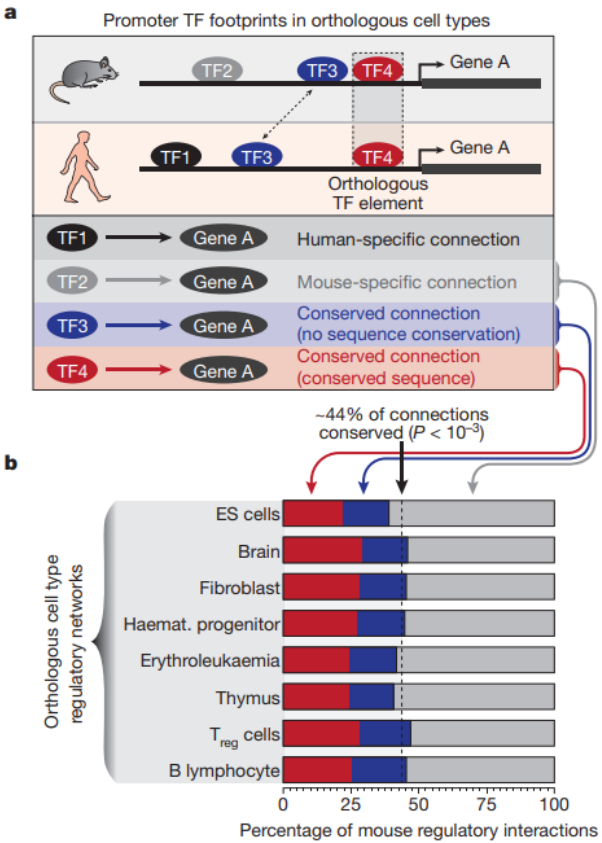



**Figure 4 | Conservation of TF-to-TF regulatory circuitry. a**, Four categories of regulatory interactions identified by comparative analysis of mouse and human TF networks. Functionally conserved connections can be mediated by TF occupancy at orthologous (red) or non-orthologous (blue) binding sites. **b**, Categorization and overall conservation of TF-to-TF connections between orthologous mouse and human cell types. On average 44% of TF-to-TF edges are conserved ($P < 0.001$; empirically calculated using shuffled networks).

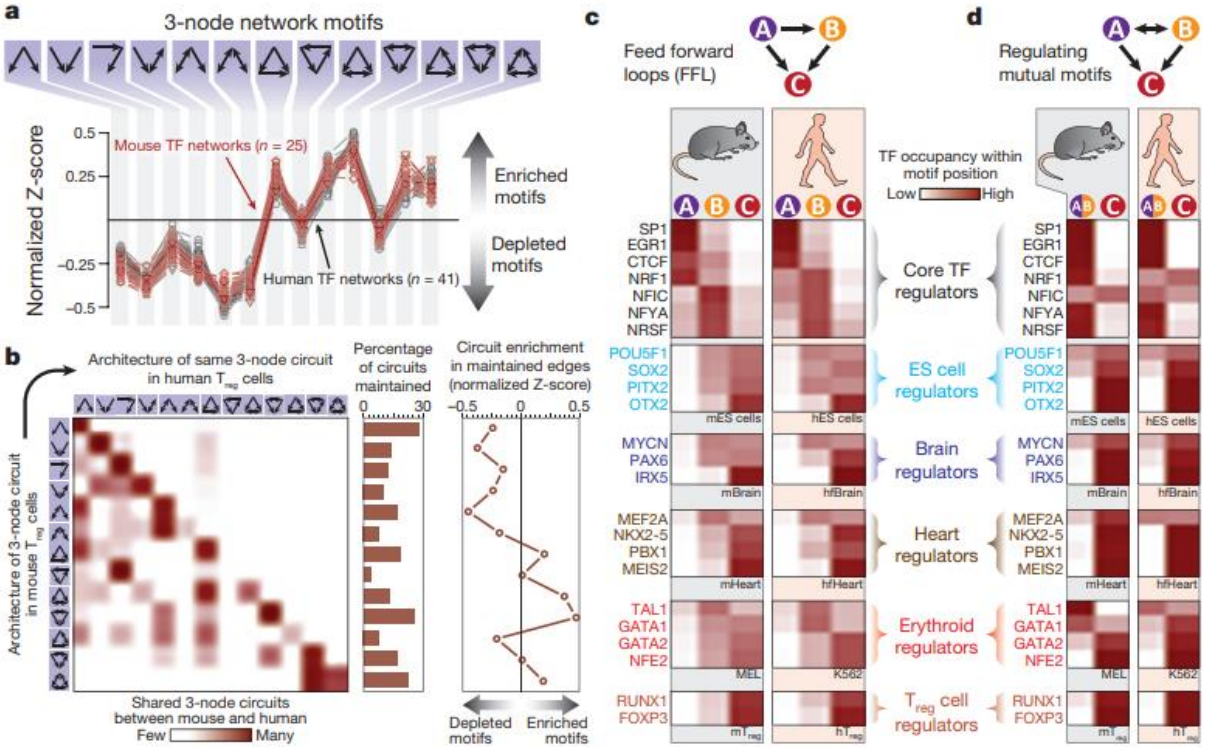



**Figure 5 | Conserved organizing principles of mammalian TF regulatory networks.** **a**, Enrichment of three-node circuits in each mouse (red lines) and human (black lines) TF regulatory network (expanded in Extended Data Fig. 5). **b**, Left: frequency with which individual three-node circuits are identically maintained between the mouse and human $T_{reg}$ network. Middle: percentage of specific three-node circuits identically maintained between the mouse and human $T_{reg}$ network. Right: enrichment of three-node circuits in a network constructed using edges present in both mouse and human $T_{reg}$ networks. **c**, **d**, Frequency with which TFs from six functional classes occupy different positions (driver, first passenger, second passenger) within FFL (**c**) or RM (**d**) circuits in different mouse and human cell-type networks (hfBrain and hfHeart refer to human fetal brain and heart, respectively).

# PREVIOUS NON-ENCODE WORK

# Massive turnover of functional sequence in human and other mammalian genomes

Stephen Meader,[1] Chris P. Ponting,[1,3] and Gerton Lunter[1,2,3]

[1]*MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, United Kingdom;* [2]*The Wellcome Trust Centre for Human Genetics, Oxford OX3 7BN, United Kingdom*

Despite the availability of dozens of animal genome sequences, two key questions remain unanswered: First, what fraction of any species' genome confers biological function, and second, are apparent differences in organismal complexity reflected in an objective measure of genomic complexity? Here, we address both questions by applying, across the mammalian phylogeny, an evolutionary model that estimates the amount of functional DNA that is shared between two species' genomes. Our main findings are, first, that as the divergence between mammalian species increases, the predicted amount of pairwise shared functional sequence drops off dramatically. We show by simulations that this is not an artifact of the method, but rather indicates that functional (and mostly noncoding) sequence is turning over at a very high rate. We estimate that between 200 and 300 Mb (~6.5%–10%) of the human genome is under functional constraint, which includes five to eight times as many constrained noncoding bases than bases that code for protein. In contrast, in *D. melanogaster* we estimate only 56–66 Mb to be constrained, implying a ratio of noncoding to coding constrained bases of about 2. This suggests that, rather than genome size or protein-coding gene complement, it is the number of functional bases that might best mirror our naïve preconceptions of organismal complexity.

# Deep conservation of TF motifs

## Conservation of transcription factor binding specificities across 600 million years of bilateria evolution

**Kazuhiro R Nitta[1], Arttu Jolma[1,2], Yimeng Yin[1], Ekaterina Morgunova[1], Teemu Kivioja[2], Junaid Akhtar[3], Korneel Hens[4], Jarkko Toivonen[5], Bart Deplancke[6], Eileen E M Furlong[3], Jussi Taipale[1,2]***

[1]Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden; [2]Genome-Scale Biology Program, University of Helsinki, Helsinki, Finland; [3]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany; [4]Institute of Bioengineering, School of Life Sciences, Swiss Federal Institute of Technology, Lausanne, Switzerland; [5]Department of Computer Science, University of Helsinki, Helsinki, Finland; [6]Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

**Abstract** Divergent morphology of species has largely been ascribed to genetic differences in the tissue-specific expression of proteins, which could be achieved by divergence in *cis*-regulatory elements or by altering the binding specificity of transcription factors (TFs). The relative importance of the latter has been difficult to assess, as previous systematic analyses of TF binding specificity have been performed using different methods in different species. To address this, we determined the binding specificities of 242 *Drosophila* TFs, and compared them to human and mouse data. This analysis revealed that TF binding specificities are highly conserved between *Drosophila* and mammals, and that for orthologous TFs, the similarity extends even to the level of very subtle dinucleotide binding preferences. The few human TFs with divergent specificities function in cell types not found in fruit flies, suggesting that evolution of TF specificities contributes to emergence of novel types of differentiated cells.

DOI: 10.7554/eLife.04837.001

***For correspondence:** jussi.taipale@ki.se

**Competing interests:** The authors declare that no competing interests exist.

**Funding:** See page 17

**Received:** 19 September 2014
**Accepted:** 09 February 2015
**Published:** 17 March 2015

### Introduction

It is estimated that the divergence between vertebrate and invertebrate lineages occurred over 600 million years ago (***Hedges et al., 2006***; ***Peterson et al., 2008***). After the divergence, protein coding sequences have retained a relatively high level of similarity, whereas homology in gene-regulatory

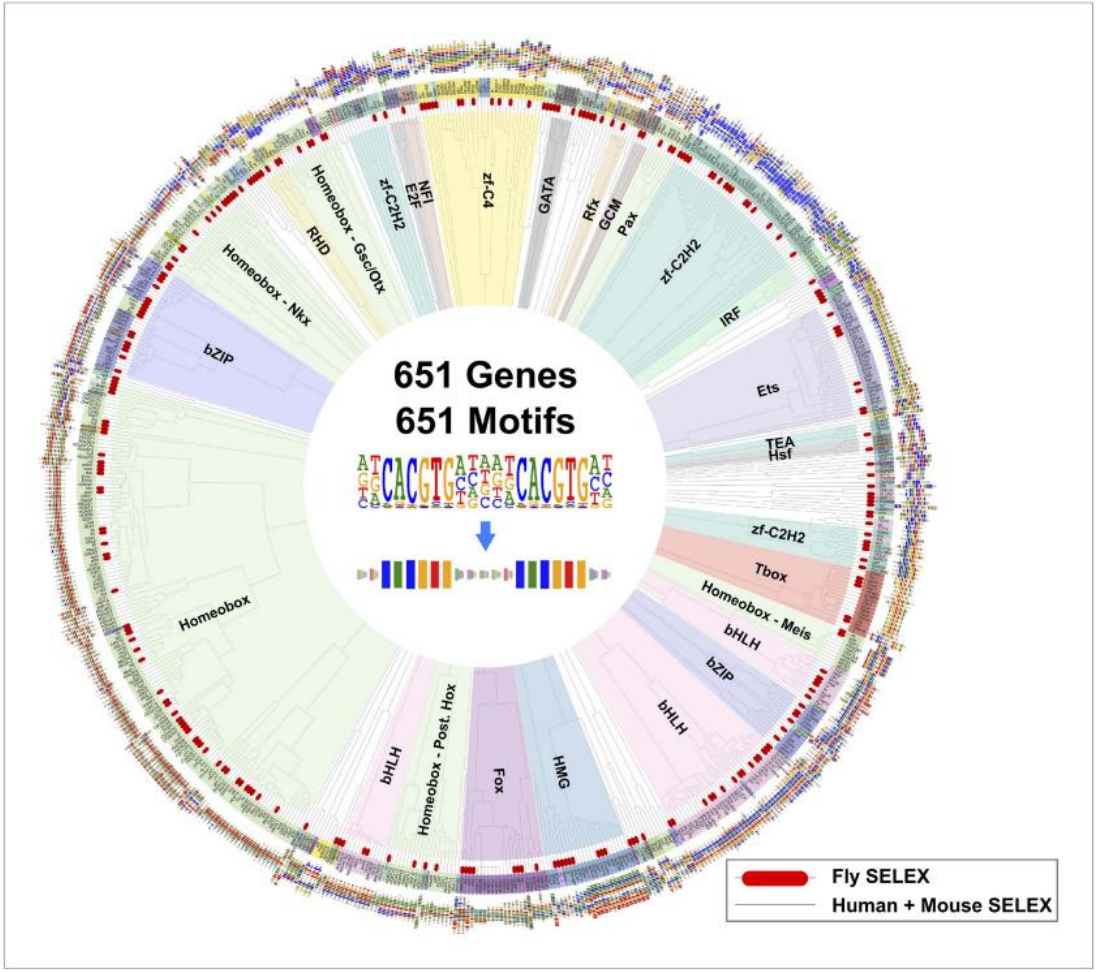



**Figure 4**. Similarity of primary binding profiles between fruit fly and human + mouse. Dendrogram shows motif

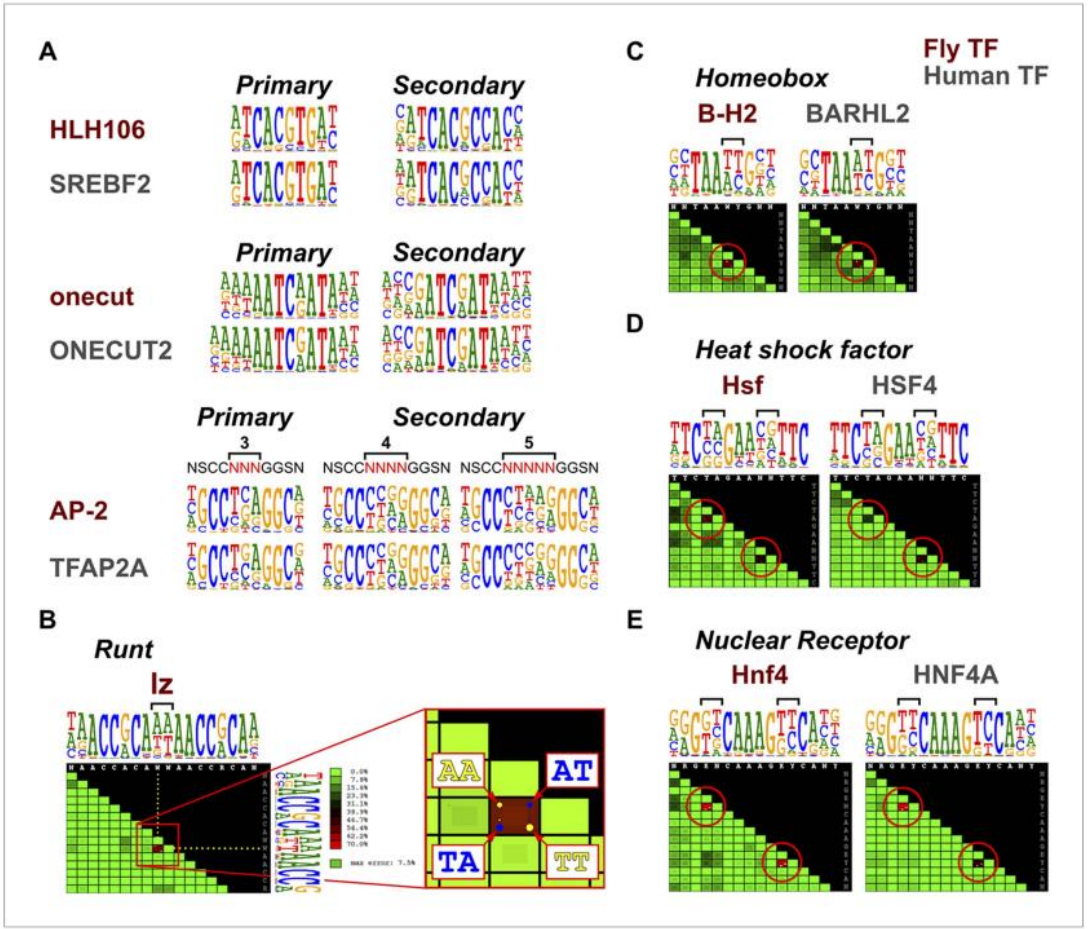



**Figure 5**. Conservation of TF secondary binding modes and dinucleotide preferences. (**A**) Conservation of

2007

BRIEF COMMUNICATIONS

nature genetics

http://www.nature.com/naturegenetics

# Tissue-specific transcriptional regulation has diverged significantly between human and mouse

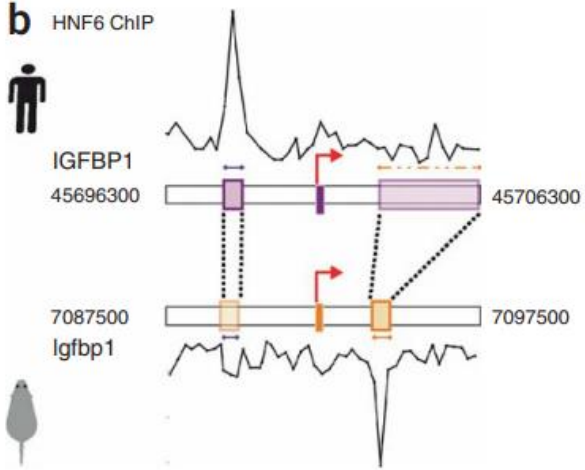
Duncan T Odom[1,5,6], Robin D Dowell[2,6], Elizabeth S Jacobsen[1], William Gordon[3], Timothy W Danford[2], Kenzie D MacIsaac[4], P Alexander Rolfe[2], Caitlin M Conboy[1,5], David K Gifford[1,2] & Ernest Fraenkel[2,3]

We performed ChIPs independently in primary hepatocytes directly isolated from mouse and human liver using antibodies against four tissue-specific transcription factors (FOXA2, HNF1A, HNF4A and HNF6) involved in liver development and regulation (**Fig. 1b** and **Supplementary Table 1** online)[7]. Hepatocytes were chosen as a representative tissue for these experiments because (i) they are functionally and structurally conserved among mammals[8], (ii) their gene expression programs are similar across species (**Supplementary Table 1**), (iii) their gene expression patterns are largely unperturbed by isolation procedures[9] and (iv) the transcription factors responsible for hepatocyte development and function are highly conserved[8]. We

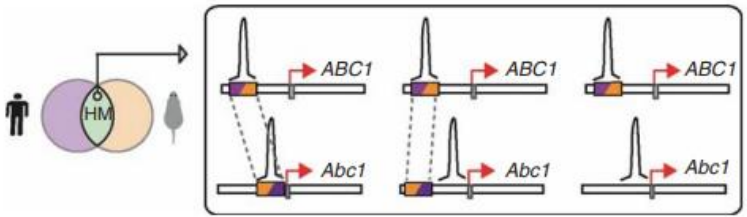
**a**

| Regulator | PFAM category | HS bound | MM bound | Intersection | *P* value | HS binding sequence | MM binding sequence |
|---|---|---|---|---|---|---|---|
| FOXA2 | Forkhead | 151 | 574 | 68 | 1.0E–45 |  |  |
| HNF1A | POU-homeodomain | 251 | 224 | 45 | 1.0E–29 |  |  |
| HNF4A | Nuclear receptor | 1,251 | 654 | 387 | 1.0E–136 |  |  |
| HNF6 | CUT-homeodomain | 157 | 324 | 41 | 1.0E–27 |  |  |

**b**




**c**




| | | | |
|---|---|---|---|
| FOXA2 | 34% | 39% | 28% |
| HNF1A | 41% | 31% | 28% |
| HNF4A | 23% | 38% | 38% |
| HNF6 | 30% | 41% | 30% |

2010



REPORT

# Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding

DOMINIC SCHMIDT, MICHAEL D. WILSON, BENOIT BALLESTER, PETRA C. SCHWALIE, GORDON D. BROWN, AILEEN MARSHALL, CLAUDIA KUTTER, STEPHEN WATT, CELIA P. MARTINEZ-JIMENEZ, [...] DUNCAN T. ODOM +4 authors Authors Info & Affiliations

*SCIENCE* · 8 Apr 2010 · Vol 328, Issue 5981 · pp. 1036-1040 · DOI: 10.1126/science.1186176

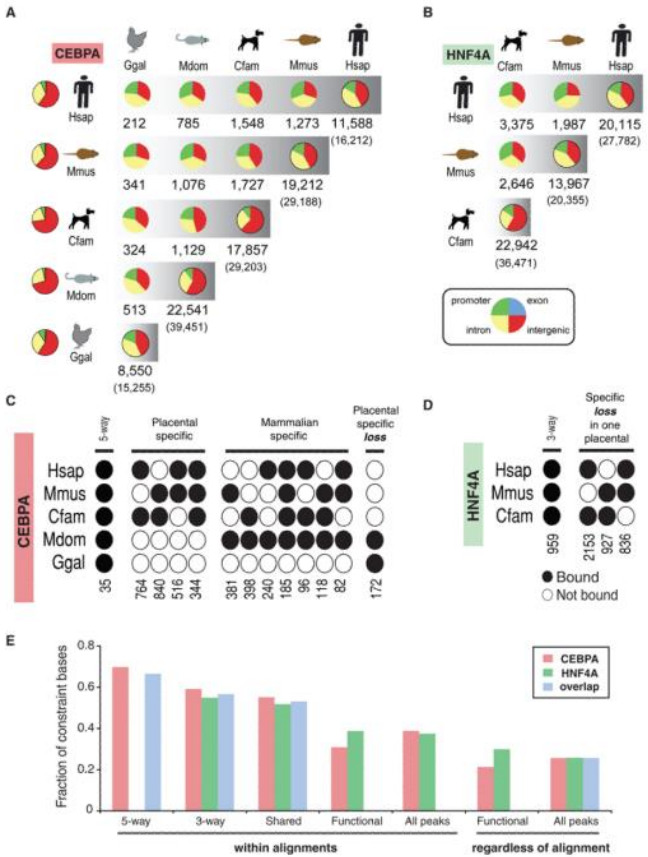



**Figure 2.**
Conservation and divergence of transcription factor binding. (**A**) For CEBPA and (**B**)

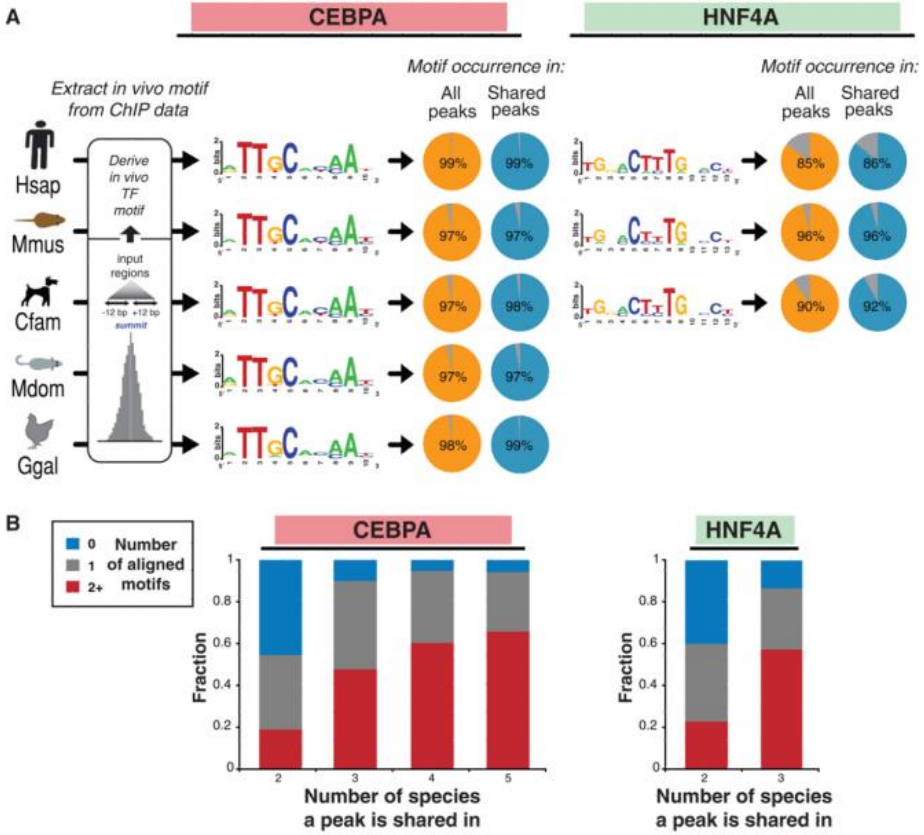



**Figure 3.**
DNA binding specificities of CEBPA and HNF4A are highly conserved during vertebrate

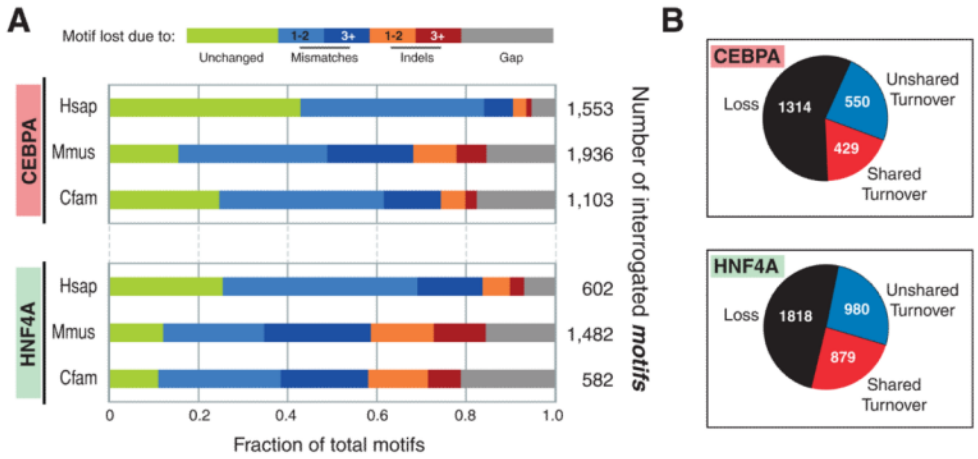



**Figure 4.**
Lineage-specific loss and turnover of transcription factor binding events. **(A)** The unbound regions in each placental mammal that align to regions showing TF binding in the other two placental mammals were collected, and the mechanisms by which the underlying motifs were disrupted were summarized. **(B)** Turnovers occurred near lineage-specific lost binding events approximately half the time; shared turnovers represent cases where a cluster of binding events likely occurred in a common ancestor (see text, Figure S16).

2011

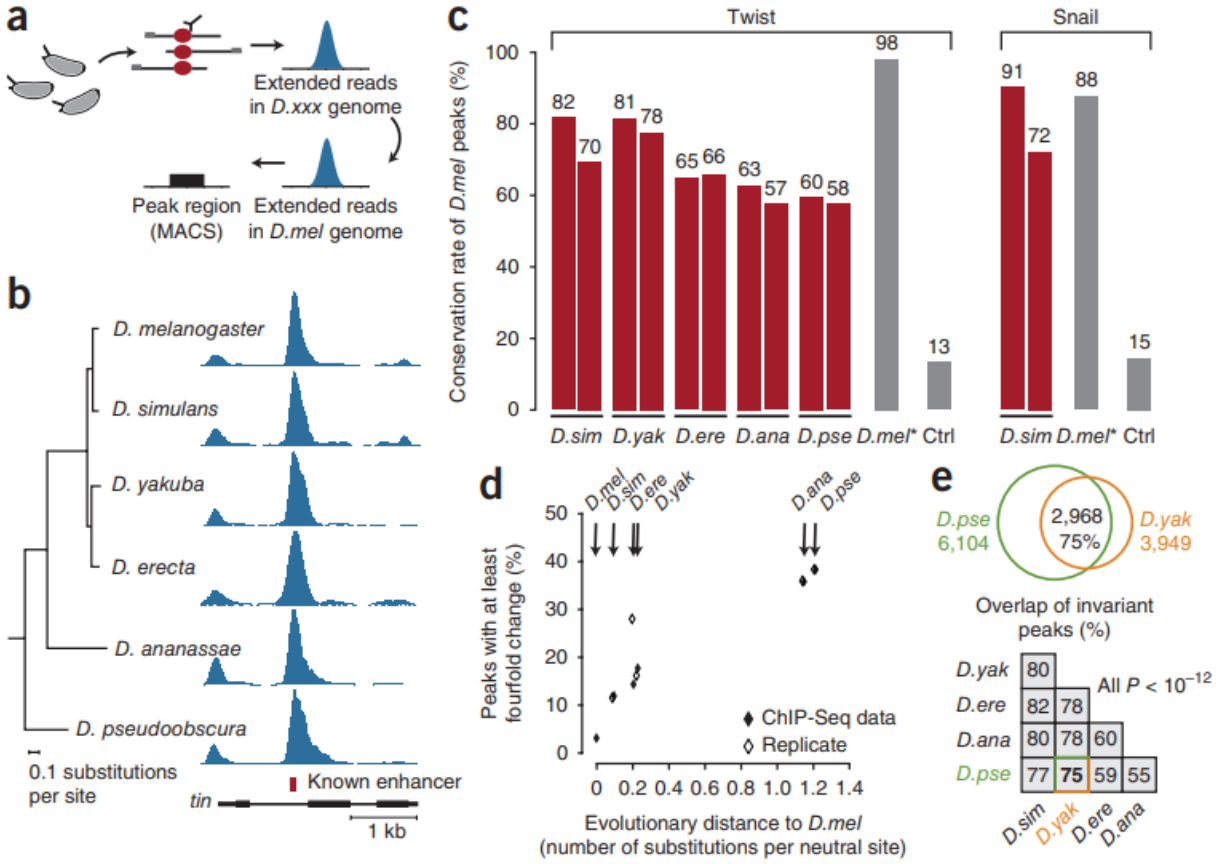
ARTICLES

nature genetics

# High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species

Qiye He[1,4], Anaïs F Bardet[2,4], Brianne Patton[1], Jennifer Purvis[1], Jeff Johnston[1], Ariel Paulson[1], Madelaine Gogol[1], Alexander Stark[2] & Julia Zeitlinger[1,3]

**The binding of some transcription factors has been shown to diverge substantially between closely related species. Here we show that the binding of the developmental transcription factor Twist is highly conserved across six *Drosophila* species, revealing strong functional constraints at its enhancers. Conserved binding correlates with sequence motifs for Twist and its partners, permitting the *de novo* discovery of their combinatorial binding. It also includes over 10,000 low-occupancy sites near the detection limit, which tend to mark enhancers of later developmental stages. These results suggest that developmental enhancers can be highly evolutionarily constrained, presumably because of their complex combinatorial nature.**


a
Extended reads in D.xxx genome
Peak region (MACS)
Extended reads in D.mel genome
b
D. melanogaster
D. simulans
D. yakuba
D. erecta
D. ananassae
D. pseudoobscura
0.1 substitutions per site
Known enhancer
tin
1 kb
c
Twist
Snail
Conservation rate of D.mel peaks (%)
82 70 81 78 65 66 63 57 60 58 98 13 91 72 88 15
D.sim D.yak D.ere D.ana D.pse D.mel* Ctrl
D.sim D.mel* Ctrl
d
Peaks with at least fourfold change (%)
D.mel D.sim D.ere D.yak D.ana D.pse
ChIP-Seq data
Replicate
Evolutionary distance to D.mel (number of substitutions per neutral site)
e
D.pse 6,104
2,968 75%
D.yak 3,949
Overlap of invariant peaks (%)
All P < 10−12
D.yak 80
D.ere 82 78
D.ana 80 78 60
D.pse 77 75 59 55
D.sim D.yak D.ere D.ana

2010

Cell

# Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals

Klara Stefflova,[1,8] David Thybert,[2,8] Michael D. Wilson,[3] Ian Streeter,[2] Jelena Aleksic,[4,5] Panagiota Karagianni,[6] Alvis Brazma,[2] David J. Adams,[7] Iannis Talianidis,[6] John C. Marioni,[2] Paul Flicek,[2,7,*] and Duncan T. Odom[1,7,*]

[1]Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge CB2 0RE, UK

[2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[3]Genetics & Genome Biology Program, Hospital for Sick Children (SickKids) and Department of Molecular Genetics, University of Toronto, 101 College Street, East Tower, Toronto, ON M5G 1L7, Canada

[4]Department of Genetics, University of Cambridge, Cambridge CB1 3QA, UK

[5]Cambridge Systems Biology Centre, University of Cambridge, Cambridge CB2 1QR, UK

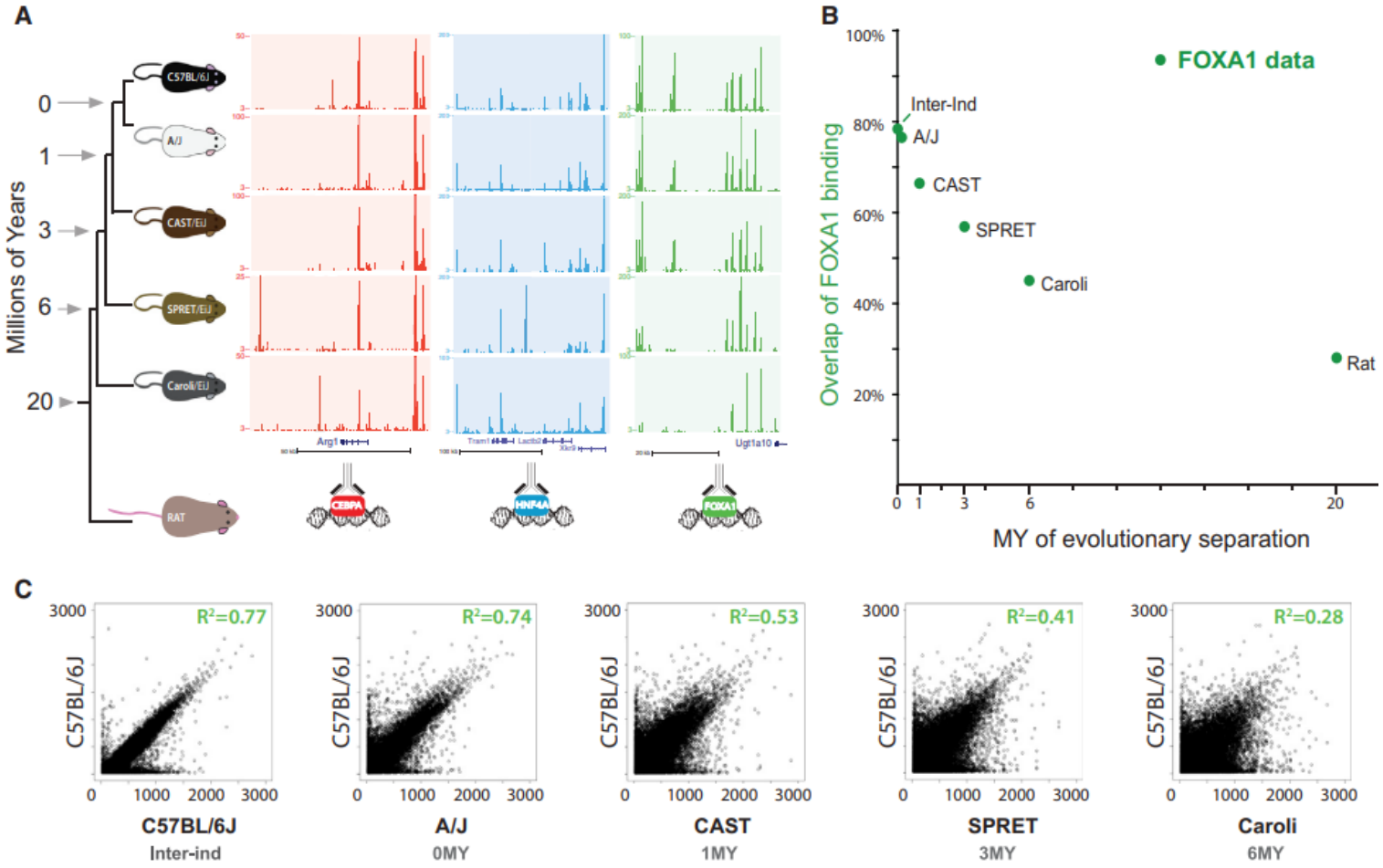
[6]Biomedical Sciences Research Center Alexander Fleming, 16672 Vari, Greece

[7]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

[8]These authors contributed equally to this work and are listed alphabetically

*Correspondence: flicek@ebi.ac.uk (P.F.), duncan.odom@cruk.cam.ac.uk (D.T.O.)

http://dx.doi.org/10.1016/j.cell.2013.07.007


A
Millions of Years
0
1
3
6
20
C57BL/6J
A/J
CAST/EiJ
SPRET/EiJ
CaroliEiJ
RAT
Arg1
Tram1
Lactb2
Xkr8
Ugt1a10
CEBPA
HNF4A
FOXA1
B
FOXA1 data
Inter-Ind
A/J
CAST
SPRET
Caroli
Rat
Overlap of FOXA1 binding
100%
80%
60%
40%
20%
0
1
3
6
20
MY of evolutionary separation
C
R²=0.77
R²=0.74
R²=0.53
R²=0.41
R²=0.28
C57BL/6J
3000
1000
2000
C57BL/6J
Inter-ind
A/J
0MY
CAST
1MY
SPRET
3MY
Caroli
6MY

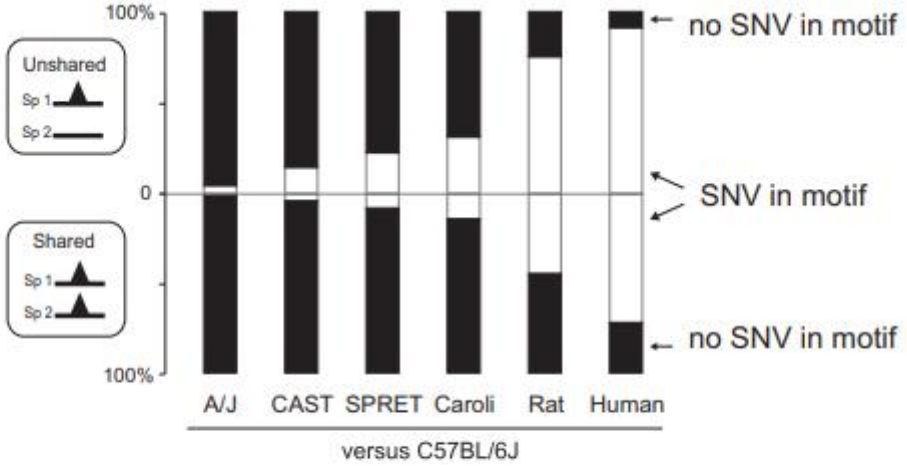



**Figure 2. Evolutionary Differences in TF Binding Cannot Be Explained Purely by Genetic Variation in Directly Bound Sequence Motifs**

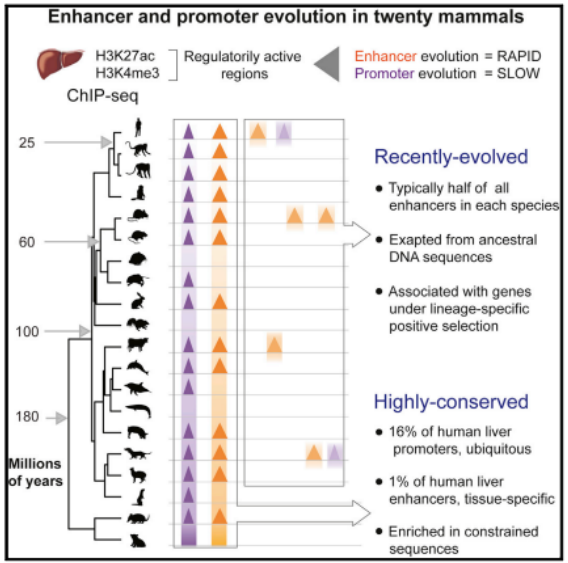
2015

Cell



# Enhancer Evolution across 20 Mammalian Species

## Graphical Abstract




## Highlights

- Rapid enhancer and slow promoter evolution across genomes of 20 mammalian species
- Enhancers are rarely conserved across these mammals
- Recently evolved enhancers dominate mammalian regulatory landscapes
- Unbiased mapping links candidate enhancers with lineage-specific positive selection


## Authors

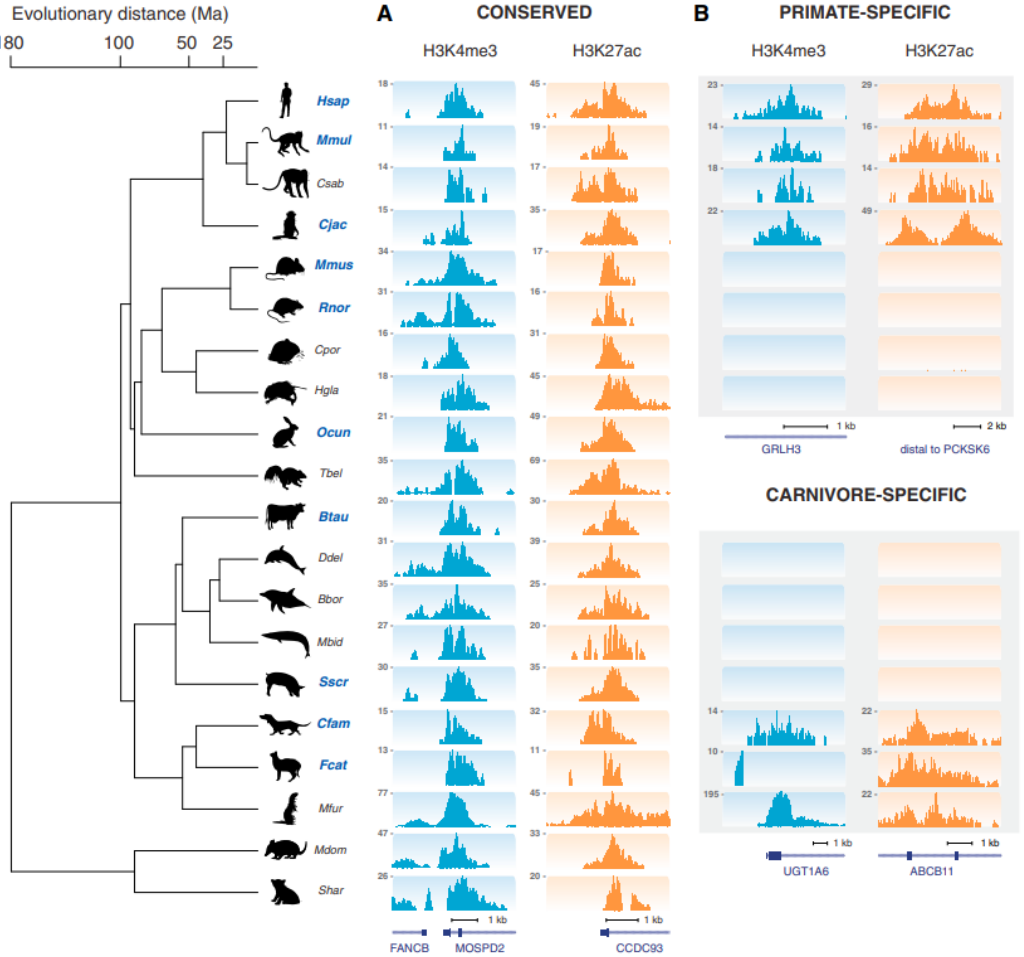
Diego Villar, Camille Berthelot, ...,
Paul Flicek, Duncan T. Odom

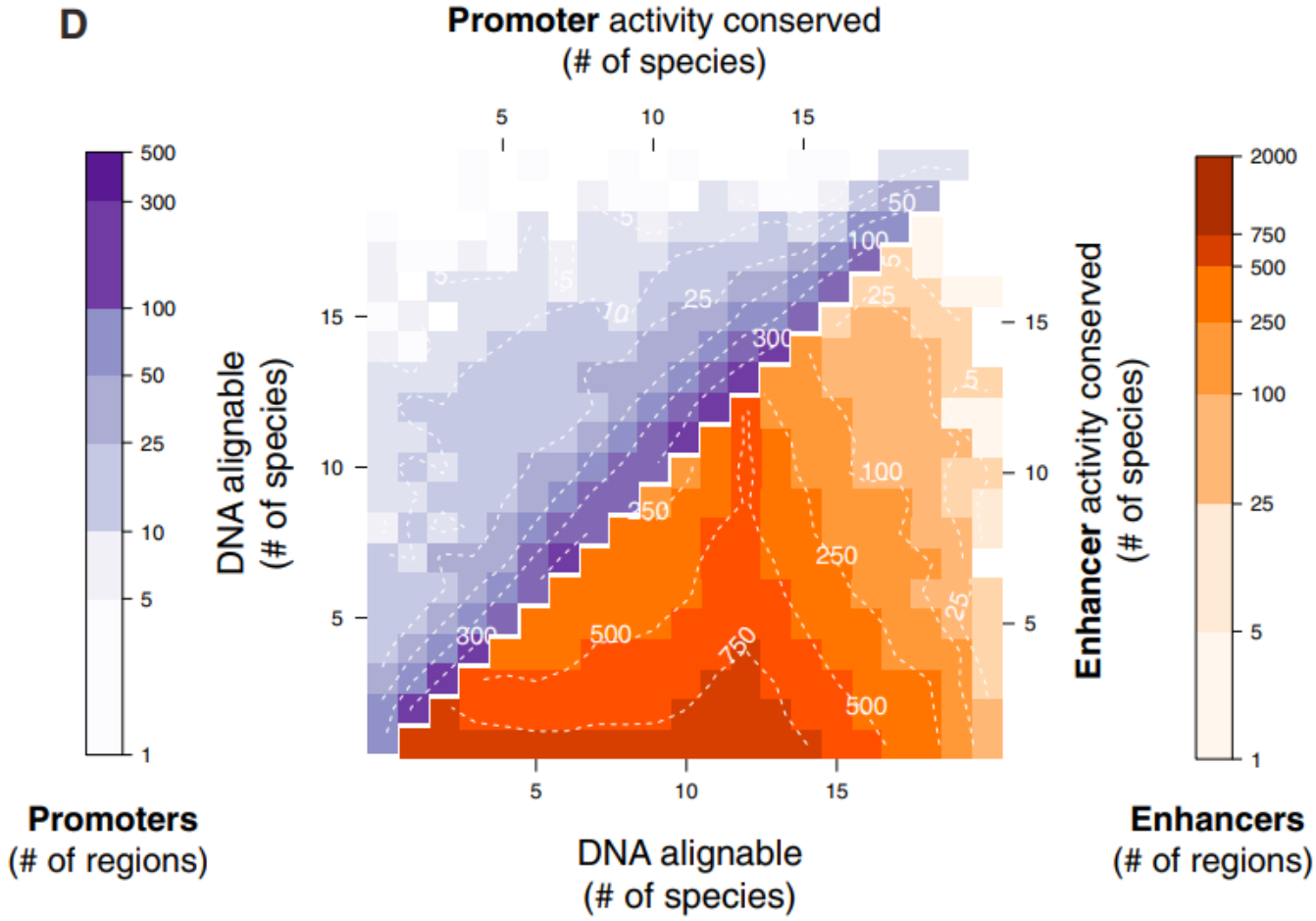
## Correspondence

flicek@ebi.ac.uk (P.F.),
duncan.odom@cruk.cam.ac.uk (D.T.O.)


## In Brief

Comparative functional genomic analysis in 20 mammalian species reveals distinct features for the evolution of enhancers, in comparison to those of promoters, across 180 million years.


Evolutionary distance (Ma)
180
100
50
25
Hsap
Mmul
Csab
Cjac
Mmus
Rnor
Cpor
Hgla
Ocun
Tbel
Btau
Ddel
Bbor
Mbid
Sscr
Cfam
Fcat
Mfur
Mdom
Shar
A
CONSERVED
H3K4me3
H3K27ac
1 kb
FANCB
MOSPD2
CCDC93
B
PRIMATE-SPECIFIC
H3K4me3
H3K27ac
1 kb
2 kb
GRLH3
distal to PCKSK6
CARNIVORE-SPECIFIC
1 kb
1 kb
UGT1A6
ABCB11


D
Promoter activity conserved
(# of species)
DNA alignable
(# of species)
Enhancer activity conserved
(# of species)
Promoters
(# of regions)
DNA alignable
(# of species)
Enhancers
(# of regions)
500
300
100
50
25
10
5
1
2000
750
500
250
100
25
5
1
5
10
15
50
100
25
300
250
500
750

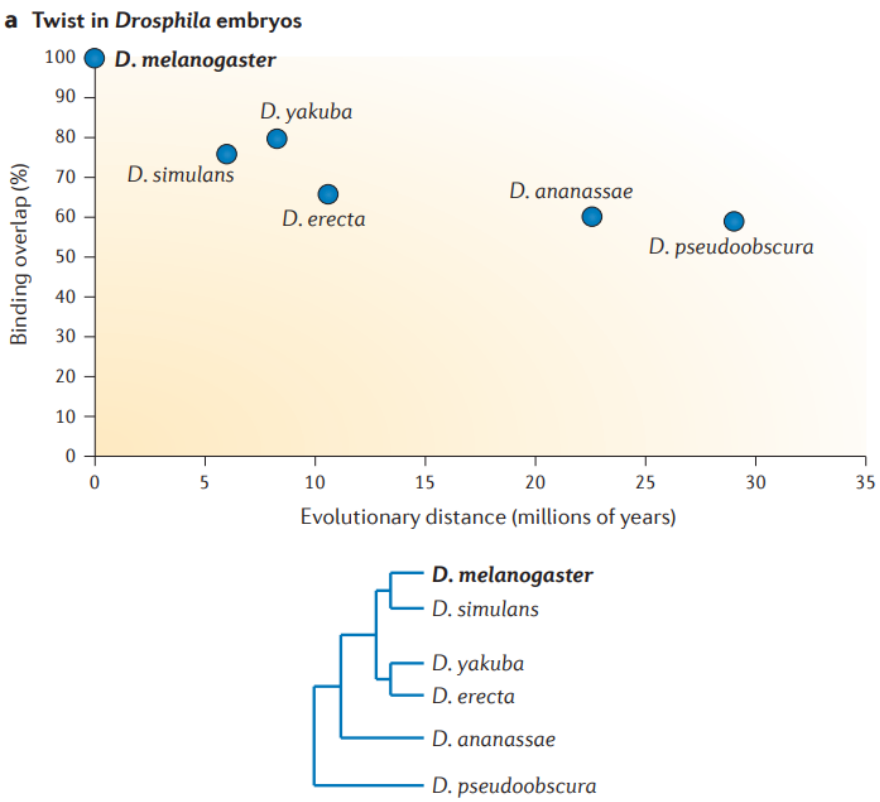
REVIEWS

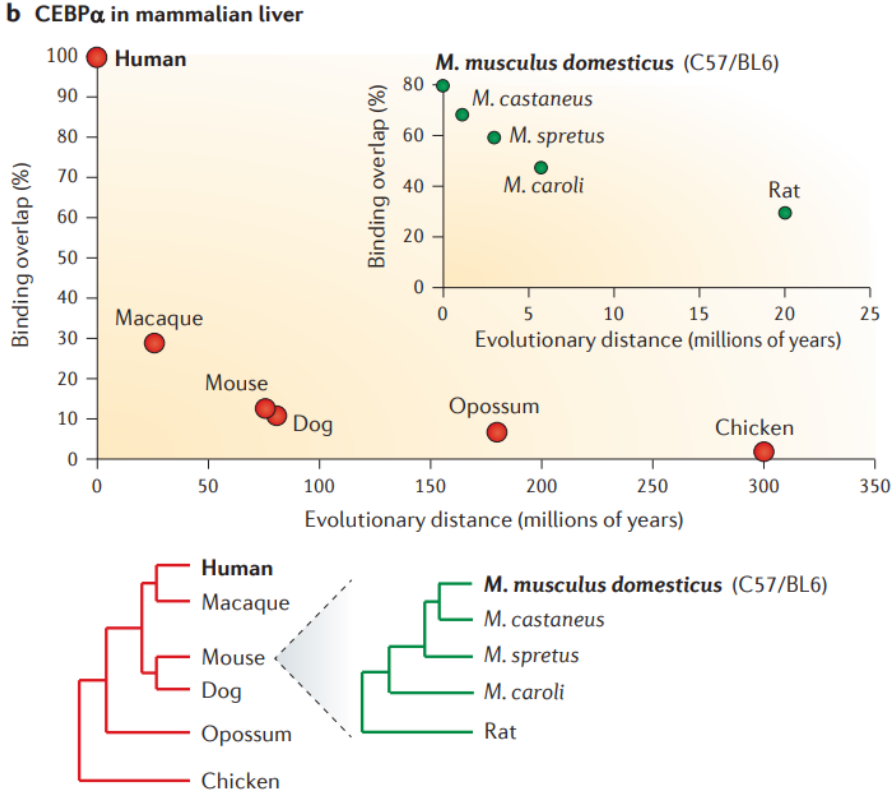
MODES OF TRANSCRIPTIONAL REGULATION

# Evolution of transcription factor binding in metazoans — mechanisms and functional implications

*Diego Villar[1], Paul Flicek[2] and Duncan T. Odom[1]*

Abstract | Differences in transcription factor binding can contribute to organismal evolution by altering downstream gene expression programmes. Genome-wide studies in *Drosophila melanogaster* and mammals have revealed common quantitative and combinatorial properties of *in vivo* DNA binding, as well as marked differences in the rate and mechanisms of evolution of transcription factor binding in metazoans. Here, we review the recently discovered rapid 're-wiring' of *in vivo* transcription factor binding between related metazoan species and summarize general principles underlying the observed patterns of evolution. We then consider what might explain the differences in genome evolution between metazoan phyla and outline the conceptual and technological challenges facing this research field.


a Twist in Drosphila embryos
D. melanogaster
D. yakuba
D. simulans
D. erecta
D. ananassae
D. pseudoobscura
Binding overlap (%)
Evolutionary distance (millions of years)
D. melanogaster
D. simulans
D. yakuba
D. erecta
D. ananassae
D. pseudoobscura


b CEBPα in mammalian liver
Human
Macaque
Mouse
Dog
Opossum
Chicken
M. musculus domesticus (C57/BL6)
M. castaneus
M. spretus
M. caroli
Rat
Binding overlap (%)
Evolutionary distance (millions of years)
Human
Macaque
Mouse
Dog
Opossum
Chicken
M. musculus domesticus (C57/BL6)
M. castaneus
M. spretus
M. caroli
Rat

eLIFE elifesciences.org

SHORT REPORT

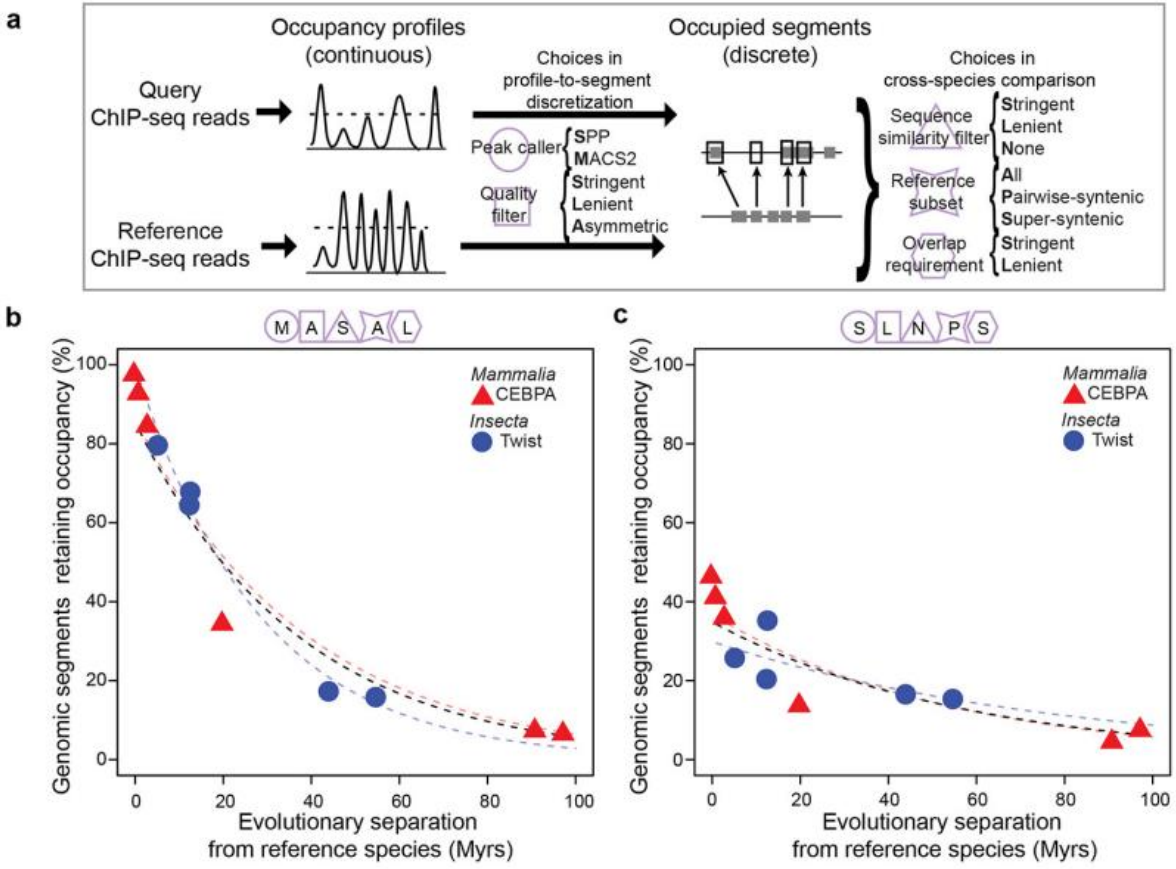
# Evidence for a common evolutionary rate in metazoan transcriptional networks

**Anne-Ruxandra Carvunis†, Tina Wang†, Dylan Skola†, Alice Yu, Jonathan Chen, Jason F Kreisberg, Trey Ideker***

Department of Medicine, University of California, San Diego, La Jolla, United States

**Abstract** Genome sequences diverge more rapidly in mammals than in other animal lineages, such as birds or insects. However, the effect of this rapid divergence on transcriptional evolution remains unclear. Recent reports have indicated a faster divergence of transcription factor binding in mammals than in insects, but others found the reverse for mRNA expression. Here, we show that these conflicting interpretations resulted from differing methodologies. We performed an integrated analysis of transcriptional network evolution by examining mRNA expression, transcription factor binding and *cis*-regulatory motifs across >25 animal species, including mammals, birds and insects. Strikingly, we found that transcriptional networks evolve at a common rate across the three animal lineages. Furthermore, differences in rates of genome divergence were greatly reduced when restricting comparisons to chromatin-accessible sequences. The evolution of transcription is thus decoupled from the global rate of genome sequence evolution, suggesting that a small fraction of the genome regulates transcription.

DOI: 10.7554/eLife.11615.001


a
Occupancy profiles
(continuous)
Occupied segments
(discrete)
Query
ChIP-seq reads
Reference
ChIP-seq reads
Choices in
profile-to-segment
discretization
Peak caller
SPP
MACS2
Quality
filter
Stringent
Lenient
Asymmetric
Choices in
cross-species comparison
Sequence
similarity filter
Stringent
Lenient
None
Reference
subset
All
Pairwise-syntenic
Super-syntenic
Overlap
requirement
Stringent
Lenient
b
M A S A L
Genomic segments retaining occupancy (%)
Evolutionary separation
from reference species (Myrs)
Mammalia
CEBPA
Insecta
Twist
c
S L N P S
Genomic segments retaining occupancy (%)
Evolutionary separation
from reference species (Myrs)
Mammalia
CEBPA
Insecta
Twist

# CURRENT DNAZOO DATA SUMMARY

# ATAC-seq data summary:

- **120** species (**118** + human and mouse)
- Libraries generated for **118**
- Sequenced for **77**
- Aligned against official release: **16**
- Aligned against draft/pre-release: **23**
- Any alignment: **39**

| Subclass / Infraclass | order | # species |
|---|---|---|
| **Monotremata** | Monotremata | |
| **Metatheria** | Didelphimorphia | 1 |
| | Paucituberculata | |
| | Microbiotheria | |
| | Dasyuromorphia | |
| | Peramelemorphia | |
| | Notoryctemorphia | |
| | Diprotodontia | 3 |
| **Eutheria** | Macroscelidea | |
| | Afrosoricida | |
| | Tubulidentata | |
| | Proboscidea | 0 |
| | Hyracoidea | |
| | Sirenia | 0 |
| | Cingulata | |
| | Pilosa | 0 |
| | Scandentia | 1 |
| | Dermoptera | |
| | Primates | 21 |
| | Rodentia | 39 |
| | Lagomorpha | 6 |
| | Erinaceomorpha | |
| | Soricomorpha | |
| | Chiroptera | 13 |
| | Pholidota | |
| | Carnivora | 26 |
| | Artiodactyla | 6 |
| | Perissodactyla | 2 |
| | Total | 118 |

# WHAT TO DO WITH THE DATA

# Current proposal:

- Individual, methodologically or otherwise focused projects are completed using a portion of the data as it is already or becomes available in the future

    - ATAC vs. Hi-C
    - Aman's modelling efforts
    - Etc.

- Eventually we also have a big all-encompassing study that uses all the data and makes grand comparisons across the mammalian tree-of-life

# CRE-level conservation

-- identify the conserved and divergent sets of fibroblast CREs of across all species

-- degree of enhancer and promoter conservation

-- how much conservation of number of CREs associated with a gene vs. their location

-- how many changes in location of homologous CREs (distance, position, orientation vs. cognate TSSs, etc.)

-- rate of CRE turnover as a function of evolutionary distance

-- (split by Enh, Prom, Ins, etc.)

# Reference-guided analysis of TF footprints

The references are human, where we have/will have the ENCODE TFs atlases and the ChromBPNet atlas, and to a lesser extent mouse, which will provide a reference set of occupied and footprinted motifs

-- One thing to focus on separately is the analysis of CTCF site evolution
    -- Gain/Loss/Conservation of individual sites
    -- Changes in the motifs, the spacing between ZFs, etc.
    -- What changes and conservation of motif number, type and orientation do we see at conserved TAD boundaries?
    -- And at non-conserved TAD boundaries

-- Second thing to focus on separately is REST/NRSF.
    -- Gain/Loss/Conservation of individual sites
    -- Around which genes is it most conserved?
    -- internal spacing changes?

-- Third thing is ZNFs, which should have changed a lot, and will become progressively more difficult to analyze the further one gets from human, but this is the interesting part:

-- how far back does the conservation of motifs and footprints go for each?

-- how plastic are individual motifs?

-- rate of evolution vs. evolutionary distance for each.

The ZNF paper is the starting resources for that.

-- Fourth, we try to assess occupied and footprinted motif conservation at the sequence and footprint levels for all the ENCODE TFs.

-- rate of evolution vs. evolutionary distance for each TF

# Repeats

-- Which repeats are accessible

-- What TF motifs can be seen footprinted there?

This will require a separate processing, and we should be seeing a lot of difference as different mammals have different repeat classes.

# Grammar conservation

This is ChromBPNet-based and focuses on evolution within homologous CREs.

-- overall sequence level conservation as a function of evolutionary time

-- motif content conservation

-- motif conservation at the sequence level

-- importance score motif conservation

-- footprinted motif conservation, i.e. is the footprint as opposed to the presence of the motif conserved

-- core analysis, taking the set of motifs highlighted by importance scores and footprints:

- -- how much conservation of combinations of motifs with a CRE do we see as a function of time?
- -- how much are different properties conserved -- orientation, positioning, number motifs, etc.
- -- which combinations of such motifs are most constrained?
- -- do we see them preferentially in certain types of CREs or around particular genes?

# What is needed next:

-- Sequence all ATAC

-- Assemblies for all species

-- Finalized whole genome alignments against at least human.

-- Further suggestions for analysis, etc.?