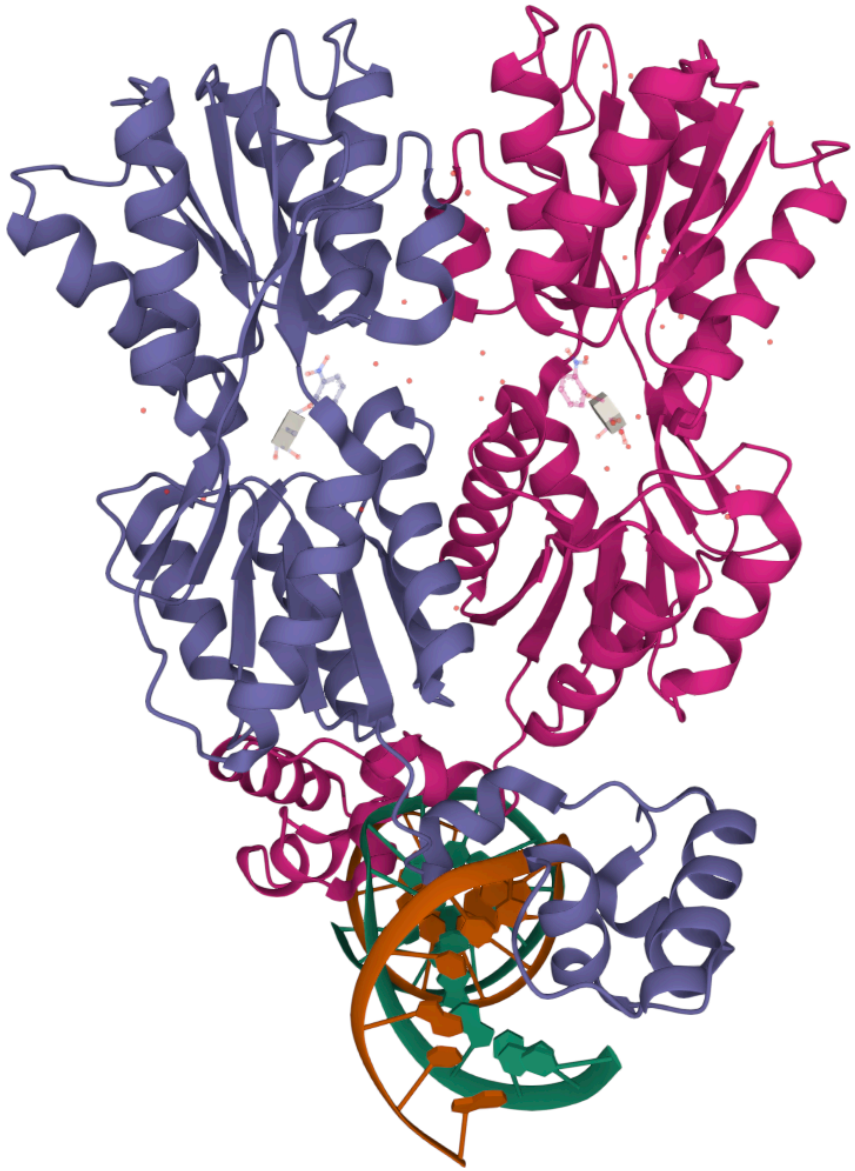
Protein-DNA recognition



https://www.rcsb.org/3d-view/1EFA

Whole LacI https://www.rcsb.org/3d-view/1EFA

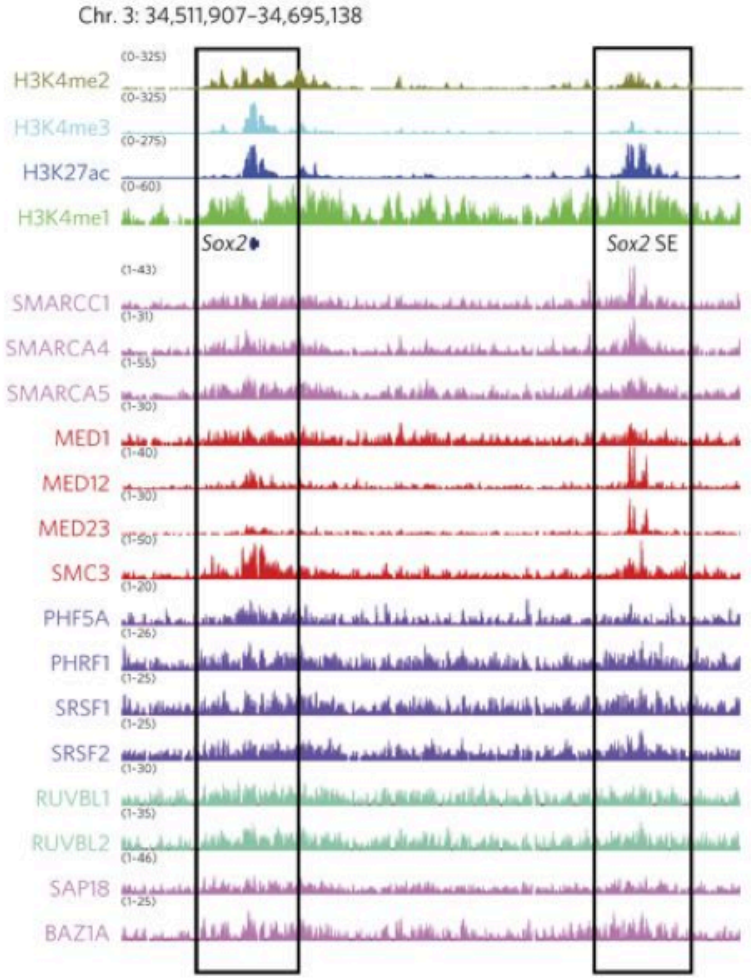
DNA-binding domain https://www.rcsb.org/3d-view/1L1M

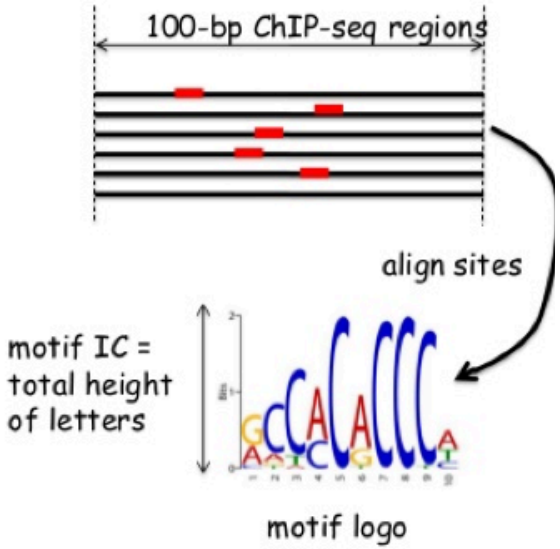
On non-spec DNA https://www.rcsb.org/3d-view/1OSL

Model of transition

https://proteopedia.org/wiki/index.php/Lac_repressor

Methods to characterize protein-DNA specificity





Binding localization, occupancy
Motif identification

Nobel Prize in Physiology or Medicine 1965

Photo from the Nobel Foundation archive.

François Jacob

Prize share: 1/3

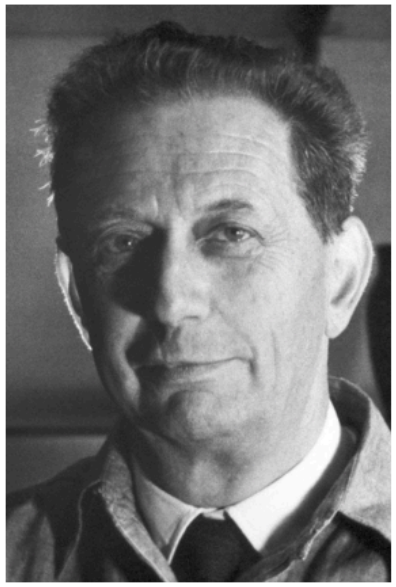


Photo from the Nobel Foundation archive.

André Lwoff

Prize share: 1/3

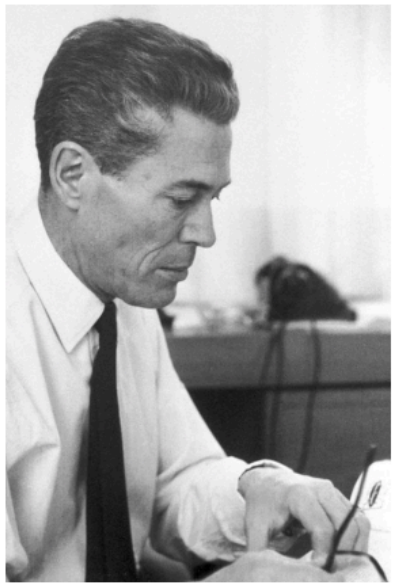


Photo from the Nobel Foundation archive.

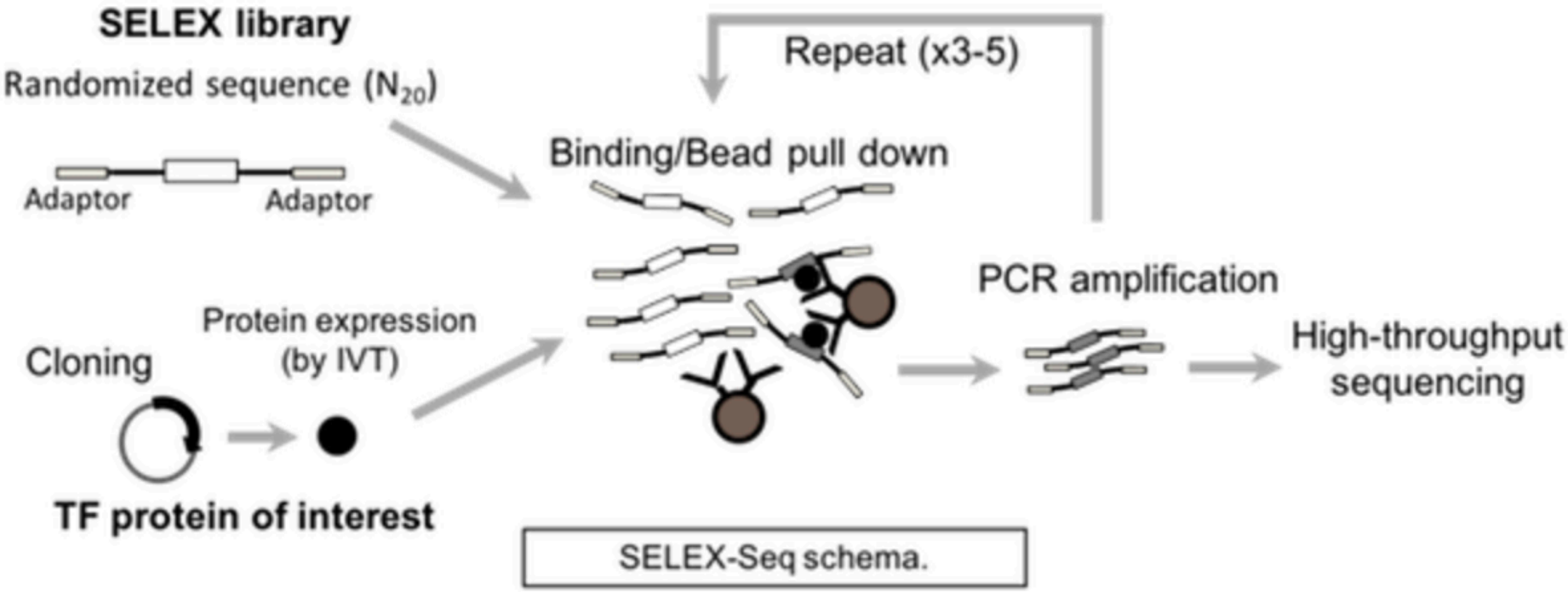
Jacques Monod

Prize share: 1/3

The Nobel Prize in Physiology or Medicine 1965 was awarded jointly to François Jacob, André Lwoff and Jacques Monod "for their discoveries concerning genetic control of enzyme and virus synthesis"

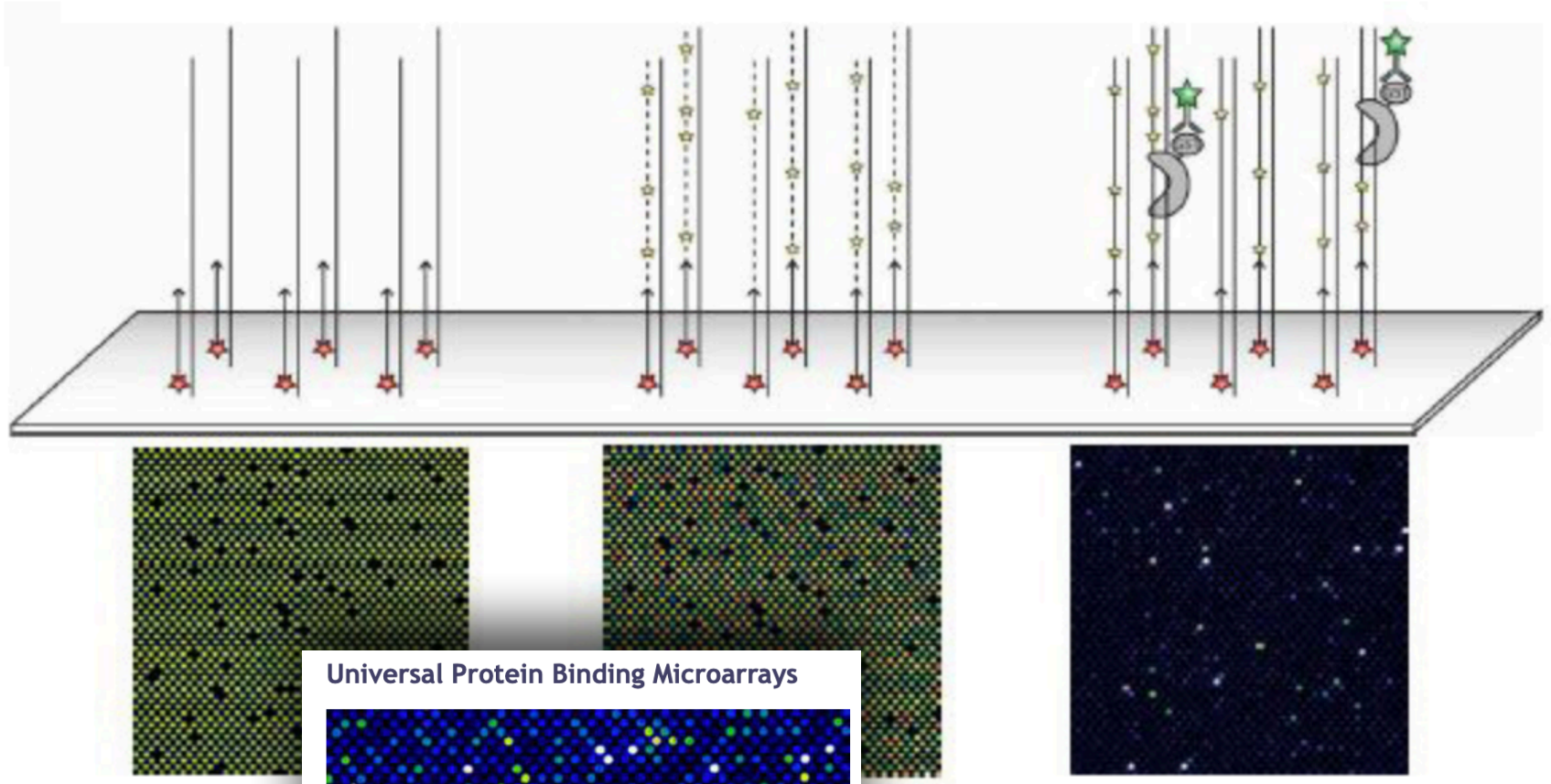
Methods to characterize protein-DNA specificity

SELEX-Seq

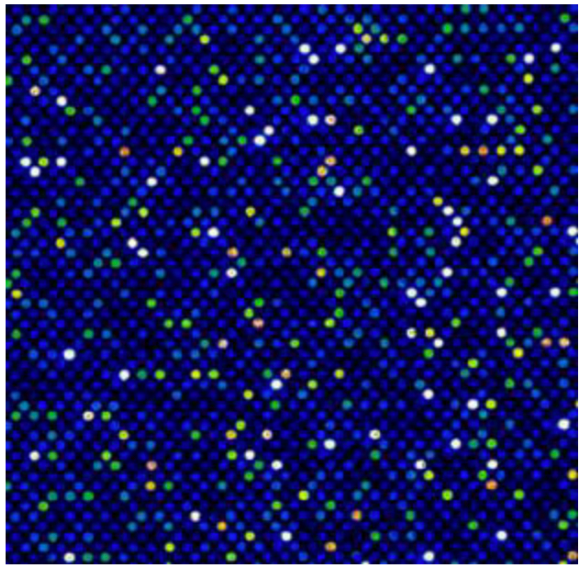


SELEX-Seq schema.

Methods to characterize protein-DNA specificity



Universal Protein Binding Microarrays



BULYK LABORATORY

de Bruijn sequence

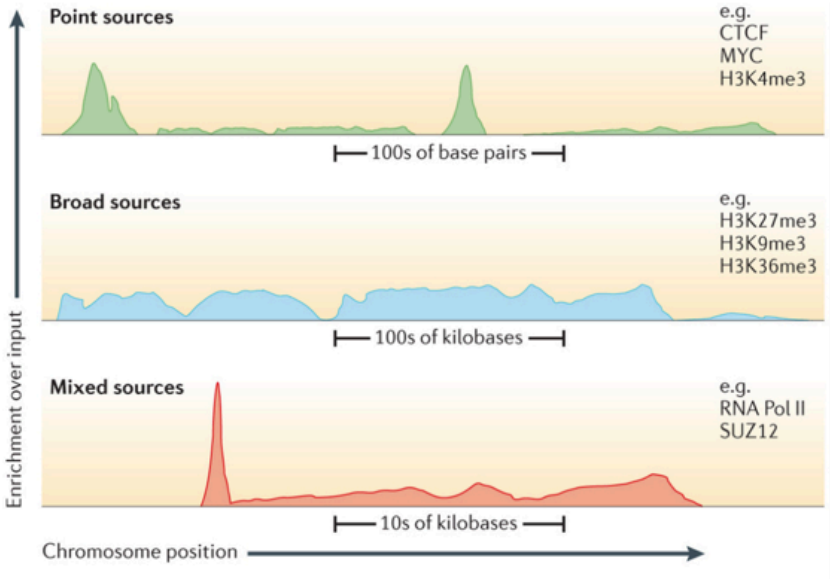
In combinatorial mathematics, a de Bruijn sequence of order n on a size-k alphabet A is a cyclic sequence in which every possible length-n string on A occurs exactly once as a substring. Wikipedia

ChIP-seq

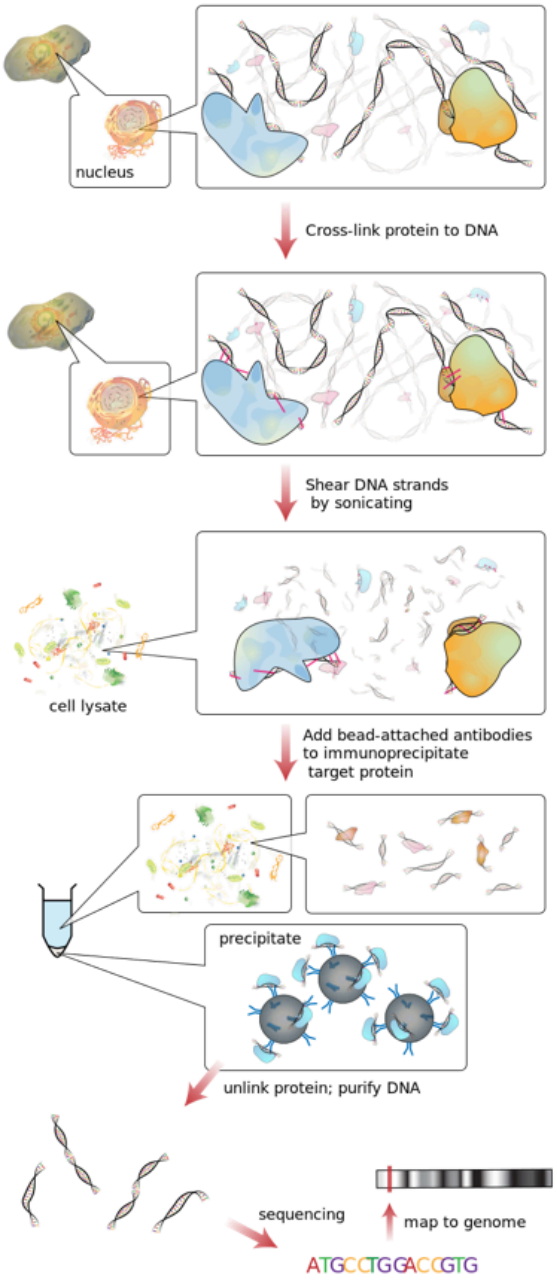
Use antibodies to pull down target of interest!

Related techniques:

- ChIP-exo
- CUT&RUN (no X-linking)



https://www.nature.com/articles/nrg3642



wikipedia

DNase I hypersensitive sites (DHS)

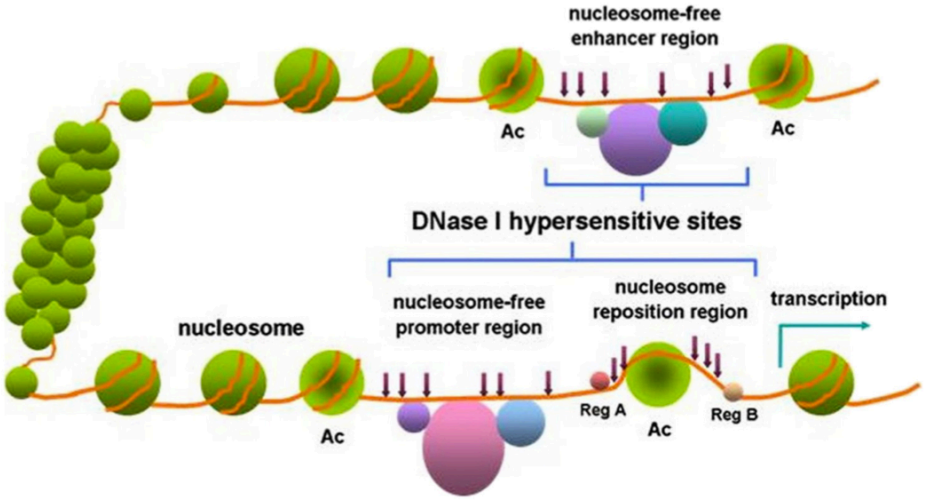
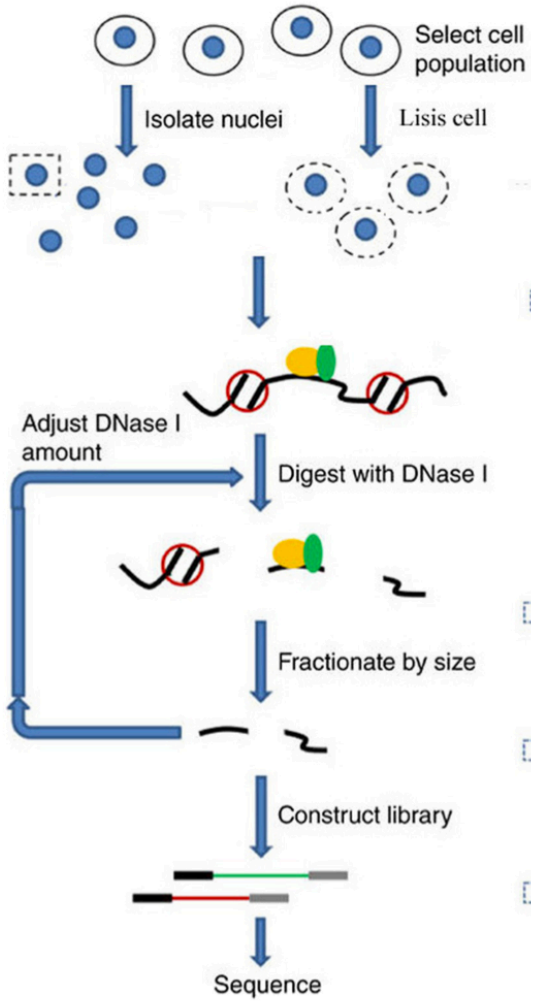


Fig. 1. DHSs within chromatin (Wang et al., 2012).



Inference of e(i,a) using expectation maximization

Precise physical models of protein–DNA interaction from high-throughput data

Justin B. Kinney, Gašper Tkačik, and Curtis G. Callan, Jr.*

Physics Department and Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544

Contributed by Curtis G. Callan, Jr., November 8, 2006 (sent for review September 30, 2006)

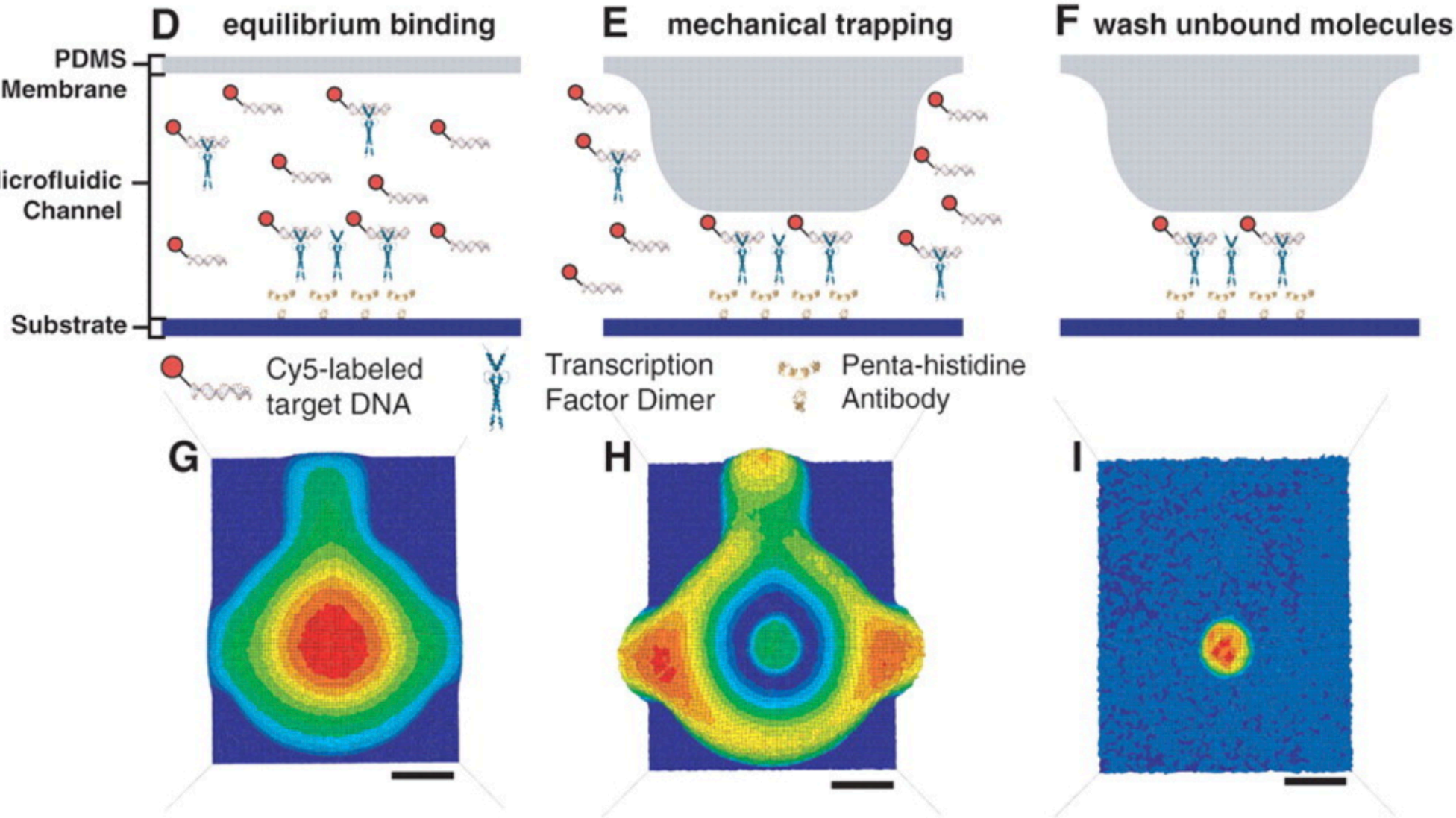
BIOINFORMATICS

Vol. 22 no. 14 2006, pages e141–e149
doi:10.1093/bioinformatics/btl223

Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE

Barrett C. Foat[1], Alexandre V. Morozov[2] and Harmen J. Bussemaker[1,3,*]

[1]Department of Biological Sciences, Columbia University, New York, NY 10027, USA, [2]Center for Studies in Physics and Biology, The Rockefeller University, New York, NY 10021, USA and [3]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA



A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors

Sebastian J. Maerkl[1,2] and Stephen R. Quake[2]*

SCIENCE VOL 315 12 JANUARY 2007

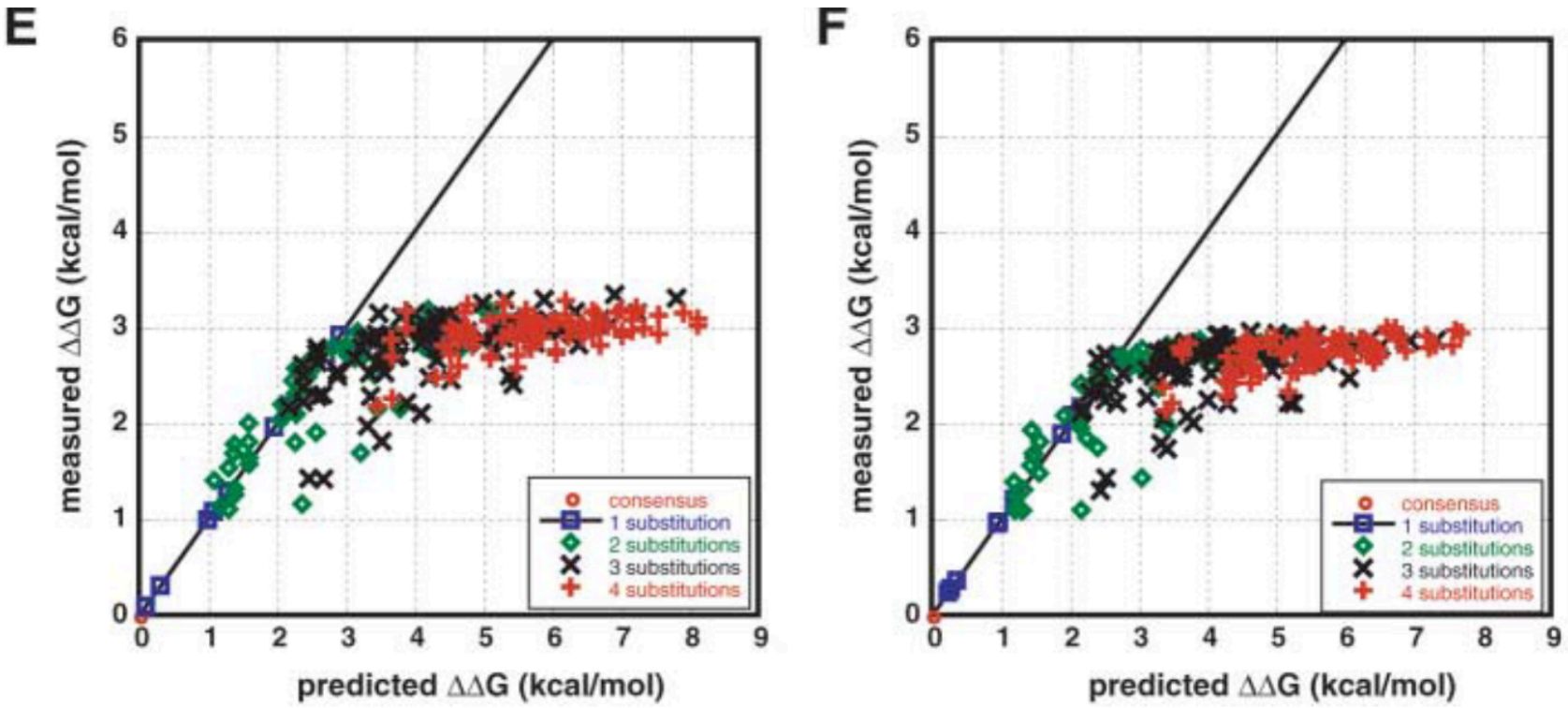


Fig. 2. Binding affinities of C-terminally tagged TFs MAX iso A (**A**), MAX iso B (**B**), Pho4p (**C**), and Cbf1p (**D**) to all sequence permutations of N_{-4} to N_{-1}. Sequences N_{-3} to N_{-1} are plotted on the category axis, with the fourth base, N_{-4}, displayed as clusters of four columns per category. (**E** and **F**) Comparisons of predicted changes in the Gibbs free energy ($\Delta\Delta G$) against measured values for MAX isoforms A and B are shown, respectively. All predicted values were calculated from PWMs assuming base independence.

A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors

Sebastian J. Maerkl[1,2] and Stephen R. Quake[2*]

SCIENCE VOL 315 12 JANUARY 2007

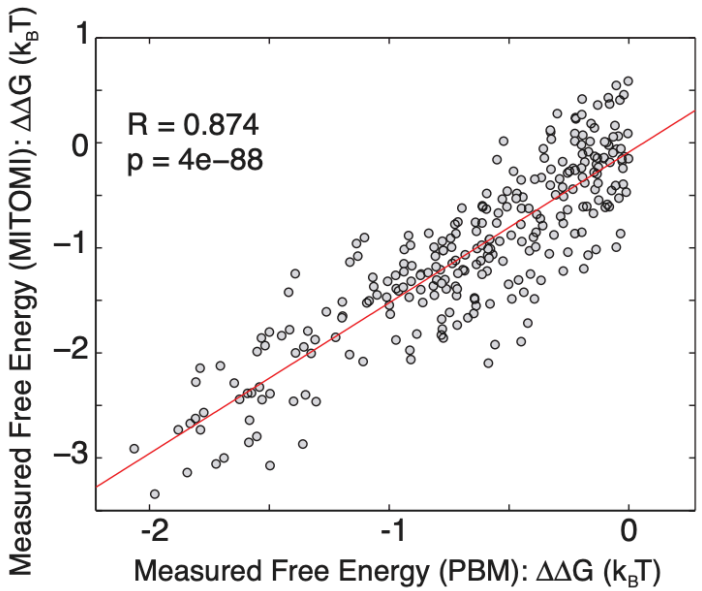
Non-specific binding

Protein—DNA binding in the absence of specific base-pair recognition

Ariel Afek[a], Joshua L. Schipper[b], John Horton[b], Raluca Gordân[b,1], and David B. Lukatsky[a,1]

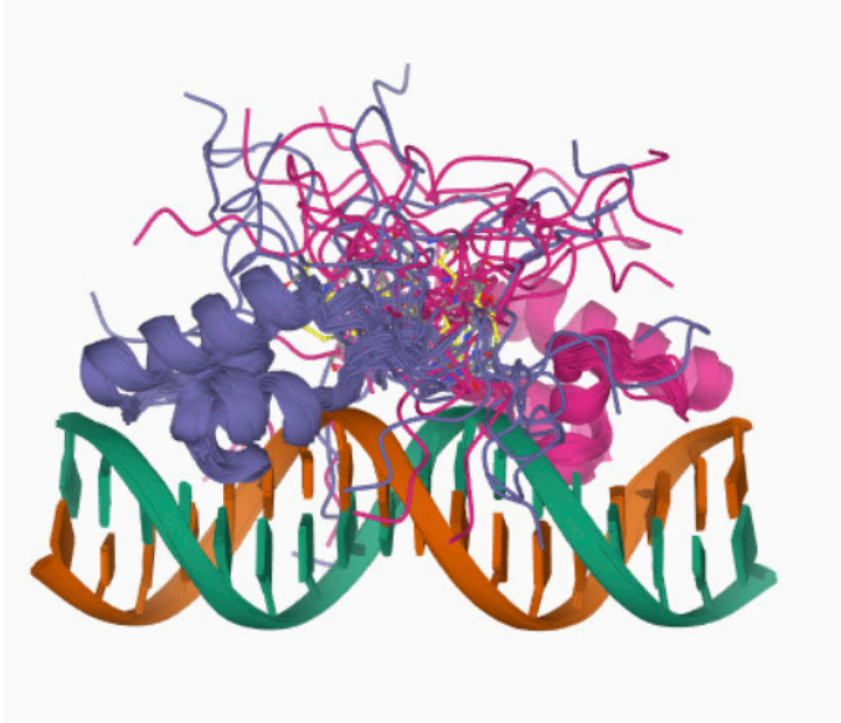
[a]Department of Chemistry, Ben-Gurion University of the Negev, Be'er Sheva 8410501 Israel; and [b]Center for Genomic and Computational Biology, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27708

analysis. First, the presence of the specific motif leads statistically, on average, to at most $\sim -2\ k_BT$ free-energy difference compared with the negative control. Second, the magnitude of the identified



~1 kT per bp

Non-specific binding



https://www.rcsb.org/3d-view/1OSL

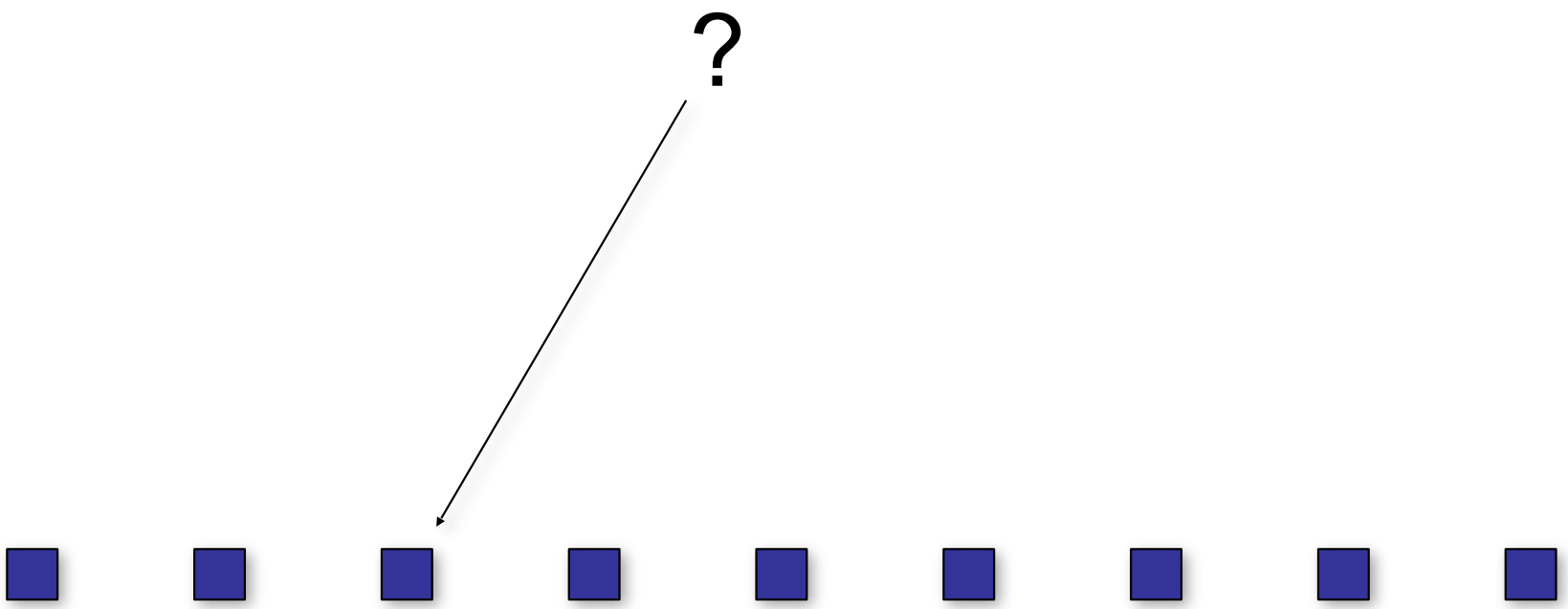
https://www.rcsb.org/3d-view/1L1M

Information theory

How much information do you need?

To find an object among N decoys

Need at least: $\log_2 N$ bits



How much information do you need?

To find an object among N decoys

Need at least: $I_{min}=\log_2 N$ bits

?

Microbes N=$10^{6\text{-}7}$ I_{min}= *20-23 bits*

Multicell euk N=$10^{8\text{-}10}$ I_{min}=*27-33 bits*

How much information is contained in a motifs

- Conserved position = 2 bits
- T or C are equally likely = 1bit
- more generally:

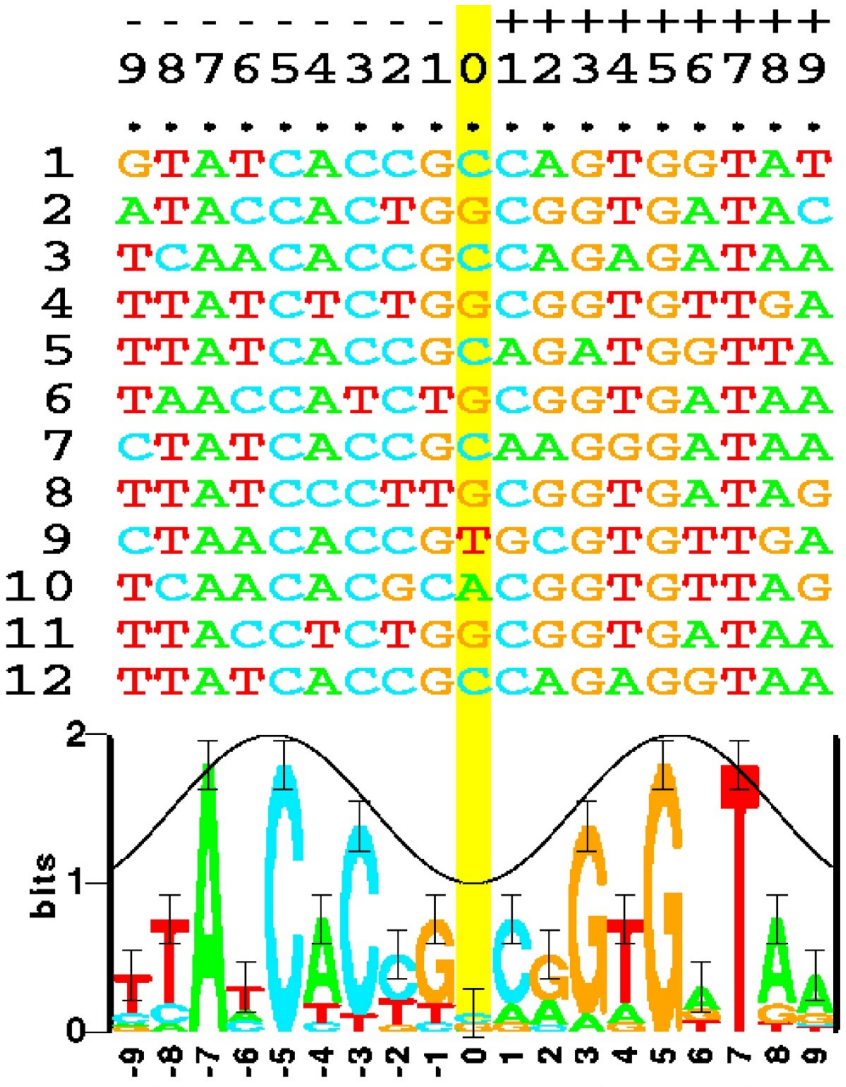
$$I_k = \sum_{x=A,T,G,C} p_k(x)\log_2\left[p_k(x)/g(x)\right]$$

where $p(x)$ is the frequency of x at positions k

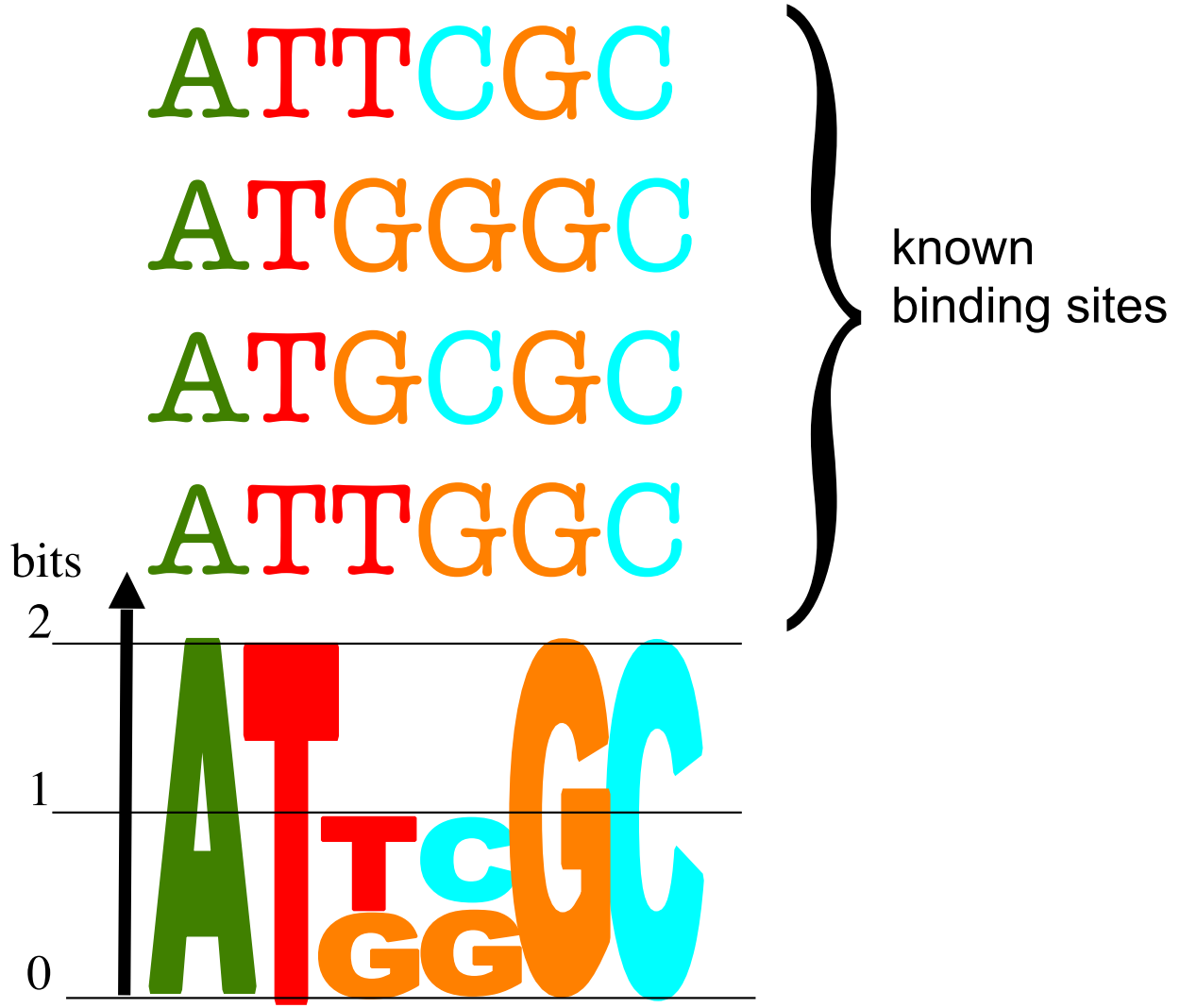
The Information Content

of the whole motif

$$I = \sum_{k=1}^{L} I_k$$



How much information is contained in a motifs



$$I = \sum_{k} \sum_{x=A,T,G,C} p_k(x) \log_2 \left[p_k(x) / g(x) \right]$$

Results

We analyzed 969 TF motifs

- calculate their Information Contents I
- compare to I_{min}

Bacteria N=10^{6-7} I_{min}= *20-23 bits*

Yeast N=10^{7} I_{min}= *24 bits*

Multicell euk N=10^{8-10} I_{min}= *27-33 bits*

Bacteria: 20-25 bits

Yeast: 14 bits

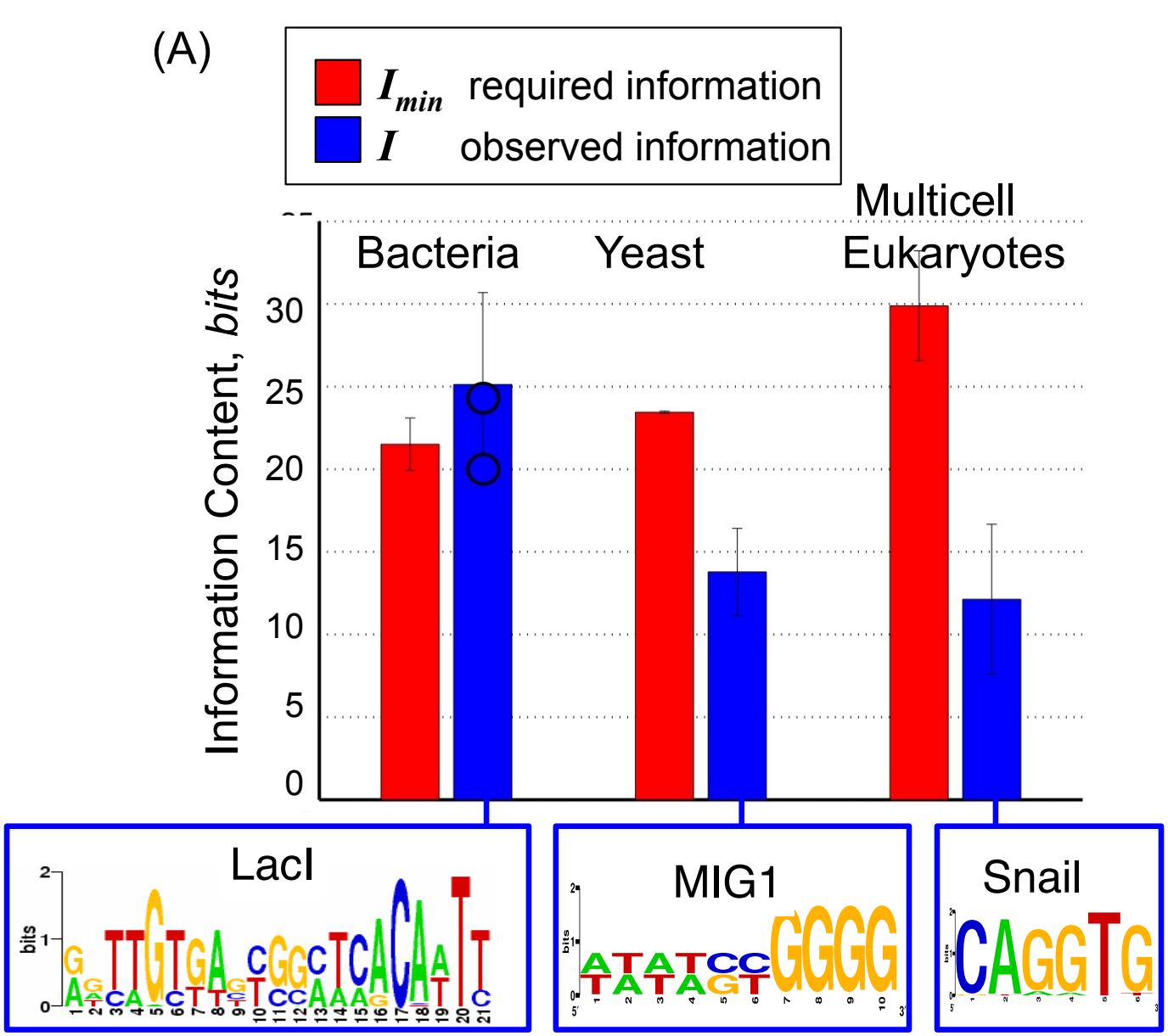
Multicell eukaryotes: 12 bits

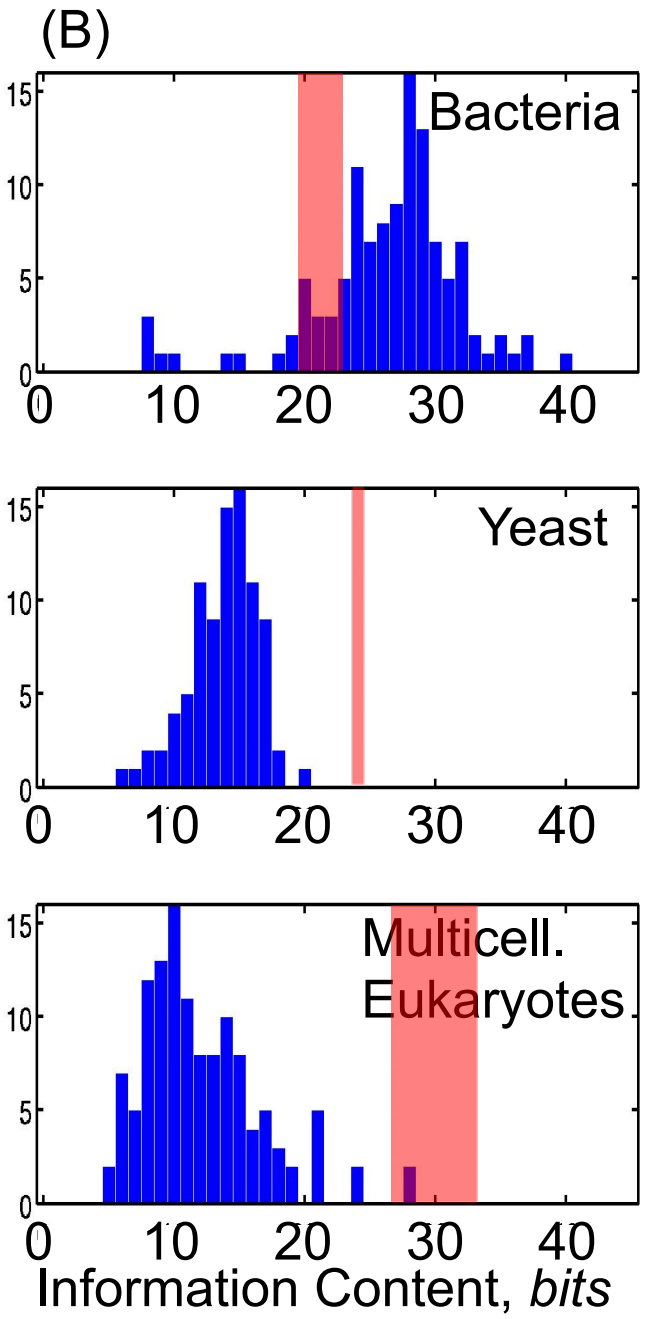
Trends Genet. 2009 Oct;25(10):434-40. doi: 10.1016/j.tig.2009.08.003. Epub 2009 Oct 6.

Different gene regulation strategies revealed by analysis of binding motifs.

Wunderlich Z[1], Mirny LA.

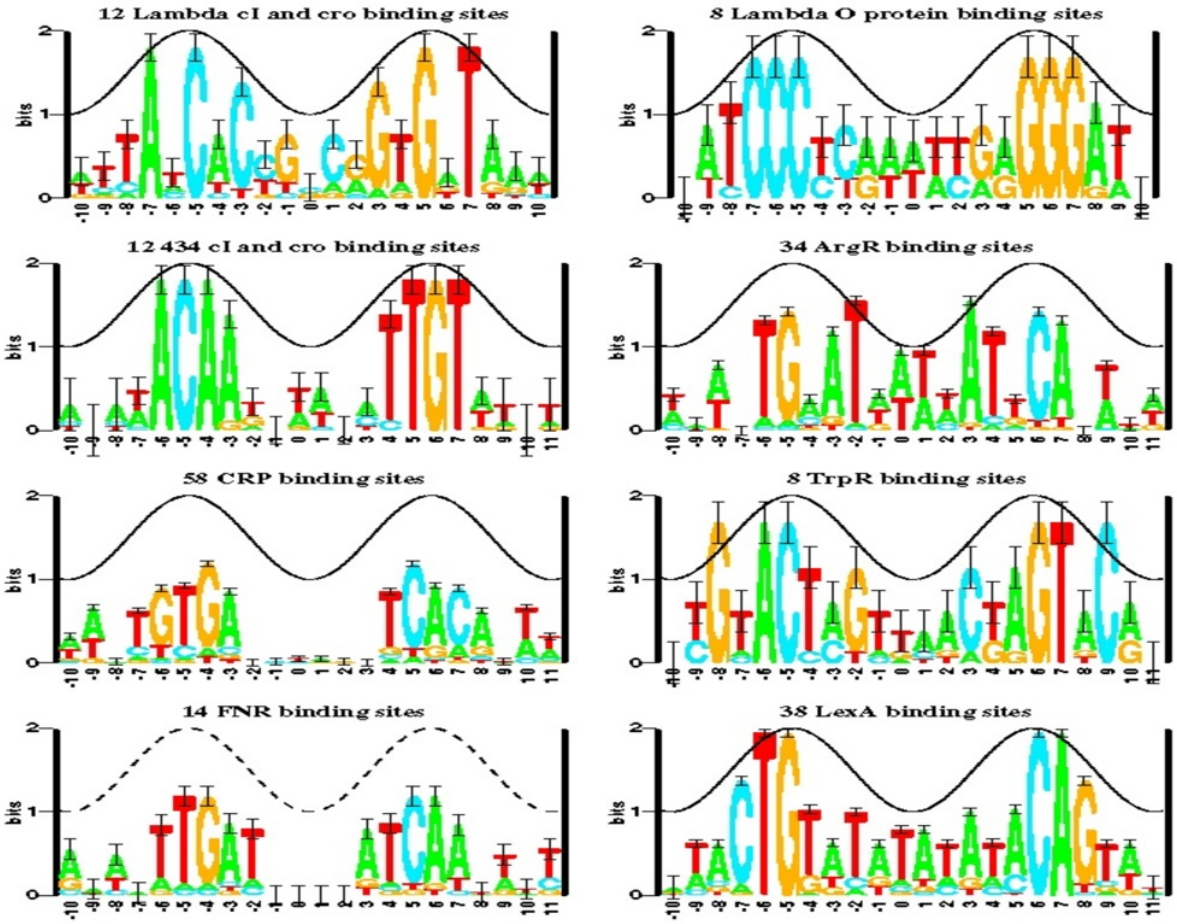
Figure1


(A)
I_{min} required information
I observed information
Information Content, *bits*
Bacteria
Yeast
Multicell Eukaryotes
LacI
MIG1
Snail


(B)
Bacteria
Yeast
Multicell. Eukaryotes
Information Content, *bits*

Examples

Bacteria: 26 bits



Eukaryotes: 12-14 bits

Cbf1 (*S. cerevisiae* bHLH)

Zif268 (*M. musculus* C_2H_2 zinc finger)

Literature consensus: GCGKGGGCG

Ceh-22 (*C. elegans* NK homeodomain)

Literature consensus: CACTNNA

Oct-1 (*H. sapiens* POU homeodomain)

Literature consensus: ATGCAAAT

Information deficiency => binding decoys

Prob of a **hits** (mistake) ≥ 2^{-I}

Number of **hits** per genome ≥ $N2^{-I} = 2^{I_{\min} - I}$

	I_{min}-I	***Number of hits***
Bacteria:	~0 bits	~1 per genome
Multicell eukaryotes:	18 bits	~10^4-10^6 per genome

assuming 90% chromatinized DNA

Trends Genet. 2009 Oct;25(10):434-40. doi: 10.1016/j.tig.2009.08.003. Epub 2009 Oct 6.

Different gene regulation strategies revealed by analysis of binding motifs.

Wunderlich Z[1], Mirny LA.

Cell Biology

- Widespread non-functional binding
- Binding ≠ gene expression
- Carefully interpret experimental data

Expected spurious binding ~20% of promoter fro any TF.

ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes

Adam A. Margolin[a,b,c,1], Teresa Palomero[d,e], Pavel Sumazin[b], Andrea Califano[a,b,d,2,3], Adolfo A. Ferrando[d,e,f,2,3], and Gustavo Stolovitzky[b,c,2,3]

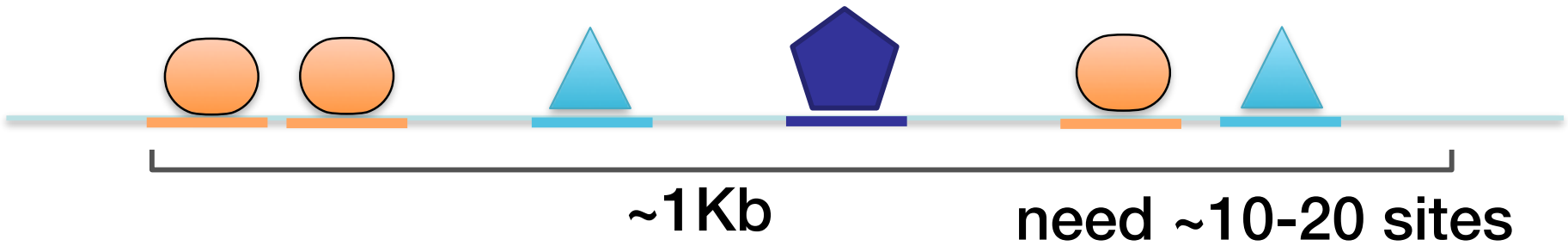
MYC binds 48% of all promoters, (!!!)
NOTCH1 19%,
HES1 18%.

Cell Biology

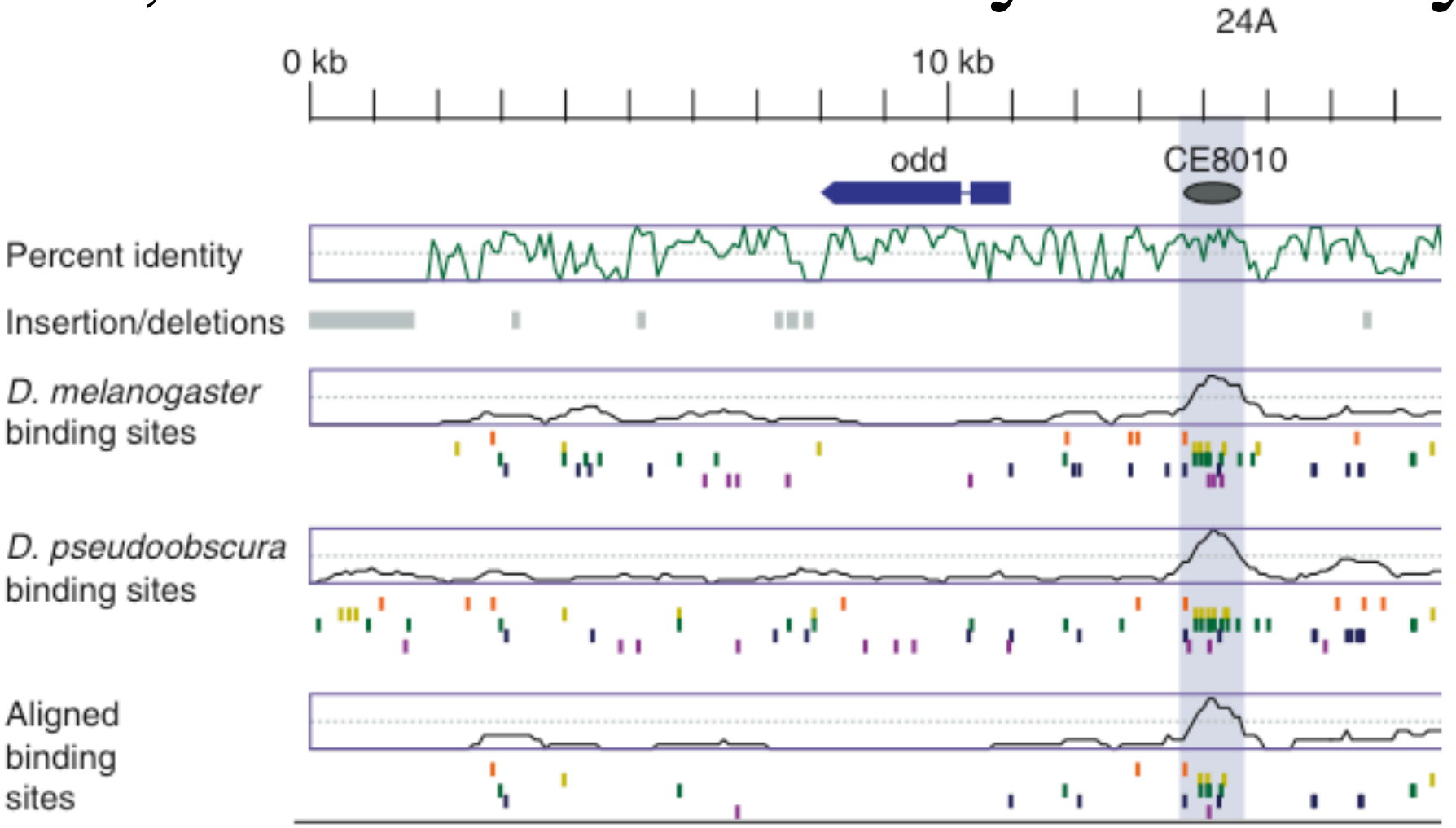
- Widespread non-functional binding
- Binding ≠ gene expression
- Carefully interpret experimental data

Need cluster of site to specify a genomic region

In a region of 1 kb composed of the sites of 3–10 different TFs, we calculate $n_{cluster}$ = 10–20 sites. This lower limit on the number of required binding sites is remarkably consistent with the mean of ~20 sites per 1 kb observed in fly developmental enhancers [28].



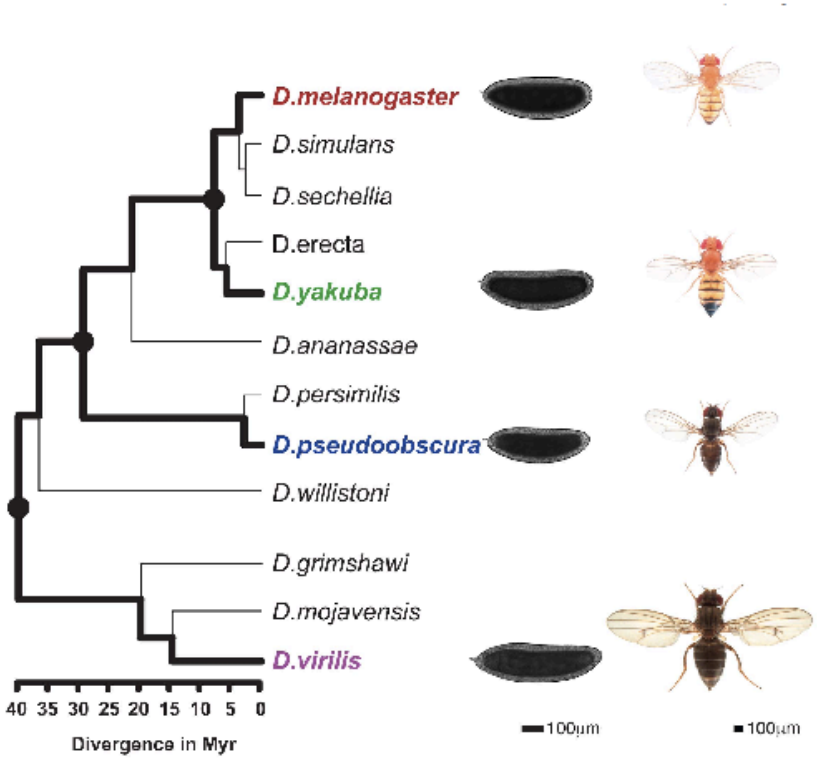
Significant evolutionary flexibility



Genome Biology 2004, 5:R61

Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*

Benjamin P Berman[¤*], Barret D Pfeiffer[¤†], Todd R Laverty[‡], Steven L Salzberg[§], Gerald M Rubin[*†‡], Michael B Eisen[¤*¶¥] and Susan E Celniker[¤†]



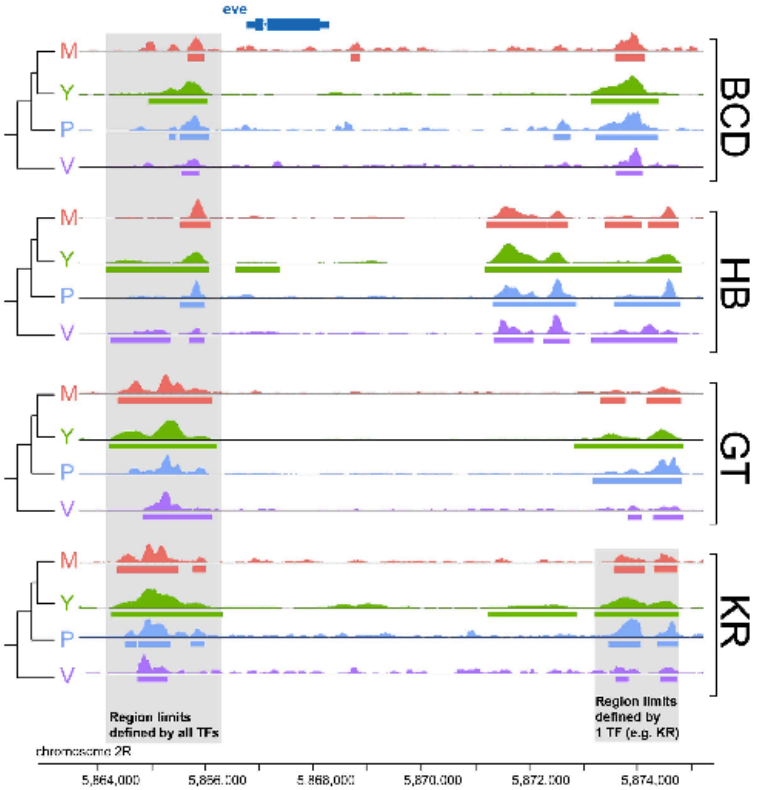
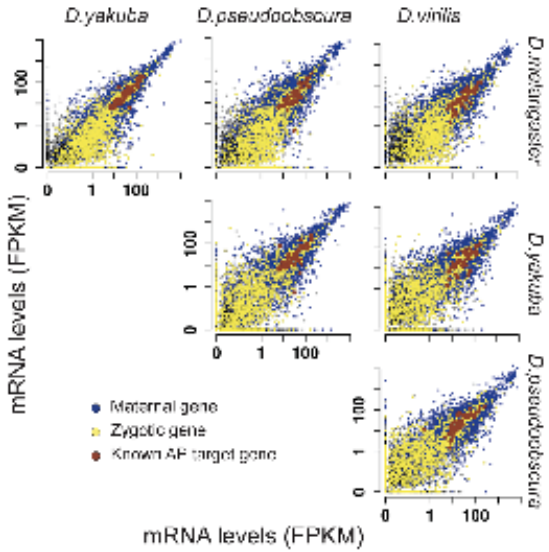


Figure 2. Comparison of binding profiles of BCD, GT, HB and KR at the *even-skipped* locus in the four species *D.melanogaster*, *D.yakuba*, *D.pseudoobscura* and *D.virilis*. An illustration of the two types of comparisons made in this study are highlighted in grey: trans-species comparison for each single TF (right) or trans-TFs comparisons (left). For simplicity, the species names were shortened using their initial: *D.melanogaster* (M), *D.yakuba* (Y), *D.pseudoobscura* (P) and *D.virilis* (V).
doi:10.1371/journal.pgen.1003748.g002

A



OPEN ACCESS Freely available online

PLOS | GENETICS

Extensive Divergence of Transcription Factor Binding in *Drosophila* Embryos with Highly Conserved Gene Expression

Mathilde Paris[1]*, Tommy Kaplan[1,2], Xiao Yong Li[1,3], Jacqueline E. Villalta[2], Susan E. Lott[1,4], Michael B. Eisen[1,2,3]*

September 2013 | Volume 9 | Issue 9 | e1003748

D

Significant evolutionary flexibility

Abstract

To better characterize how variation in regulatory sequences drives divergence in gene expression, we undertook a systematic study of transcription factor binding and gene expression in blastoderm embryos of four species, which sample much of the diversity in the 40 million-year old genus *Drosophila*: *D. melanogaster*, *D. yakuba*, *D. pseudoobscura* and *D. virilis*. We compared gene expression, measured by mRNA-seq, to the genome-wide binding, measured by ChIP-seq, of four transcription factors involved in early anterior-posterior patterning. We found that mRNA levels are much better conserved than individual transcription factor binding events, and that changes in a gene's expression were poorly explained by changes in adjacent transcription factor binding. However, highly bound sites, sites in regions bound by multiple factors and sites near genes are conserved more frequently than other binding, suggesting that a considerable amount of transcription factor binding is weakly or non-functional and not subject to purifying selection.

OPEN ACCESS Freely available online

PLOS | GENETICS

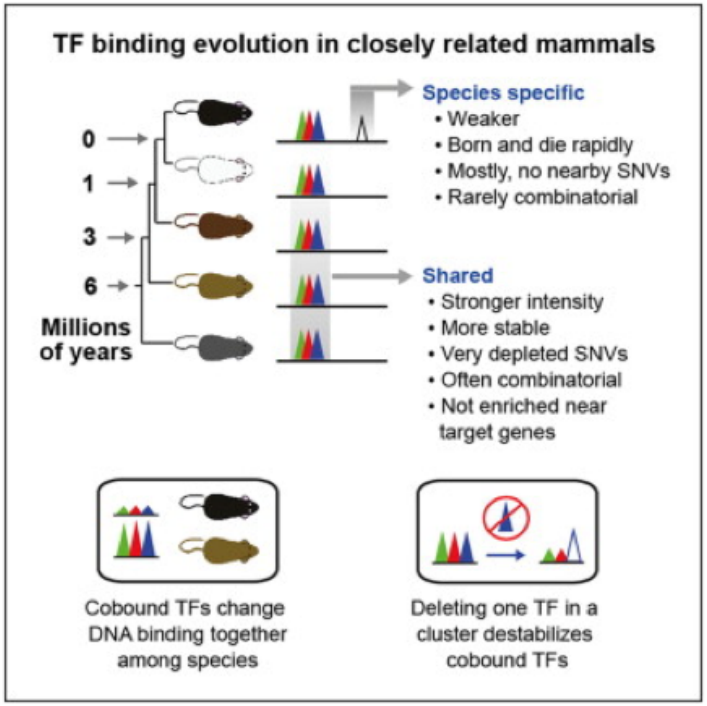
Extensive Divergence of Transcription Factor Binding in *Drosophila* Embryos with Highly Conserved Gene Expression

Mathilde Paris[1]*, Tommy Kaplan[1,2], Xiao Yong Li[1,3], Jacqueline E. Villalta[2], Susan E. Lott[1,4], Michael B. Eisen[1,2,3]*

September 2013 | Volume 9 | Issue 9 | e1003748

Significant evolutionary flexibility

Graphical Abstract



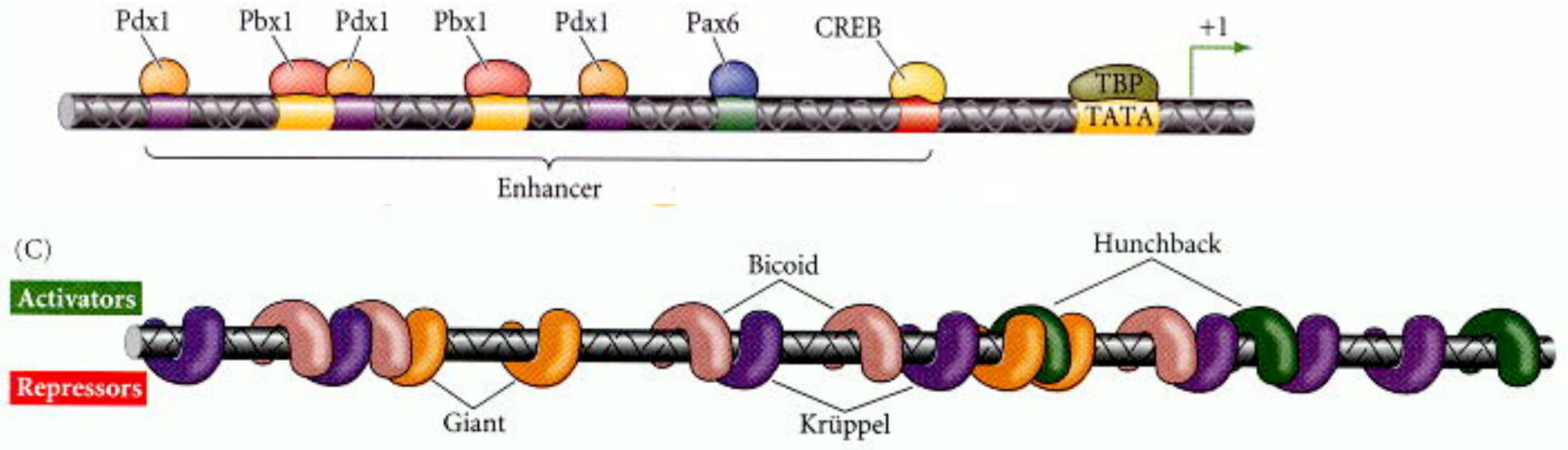
Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals

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,*]

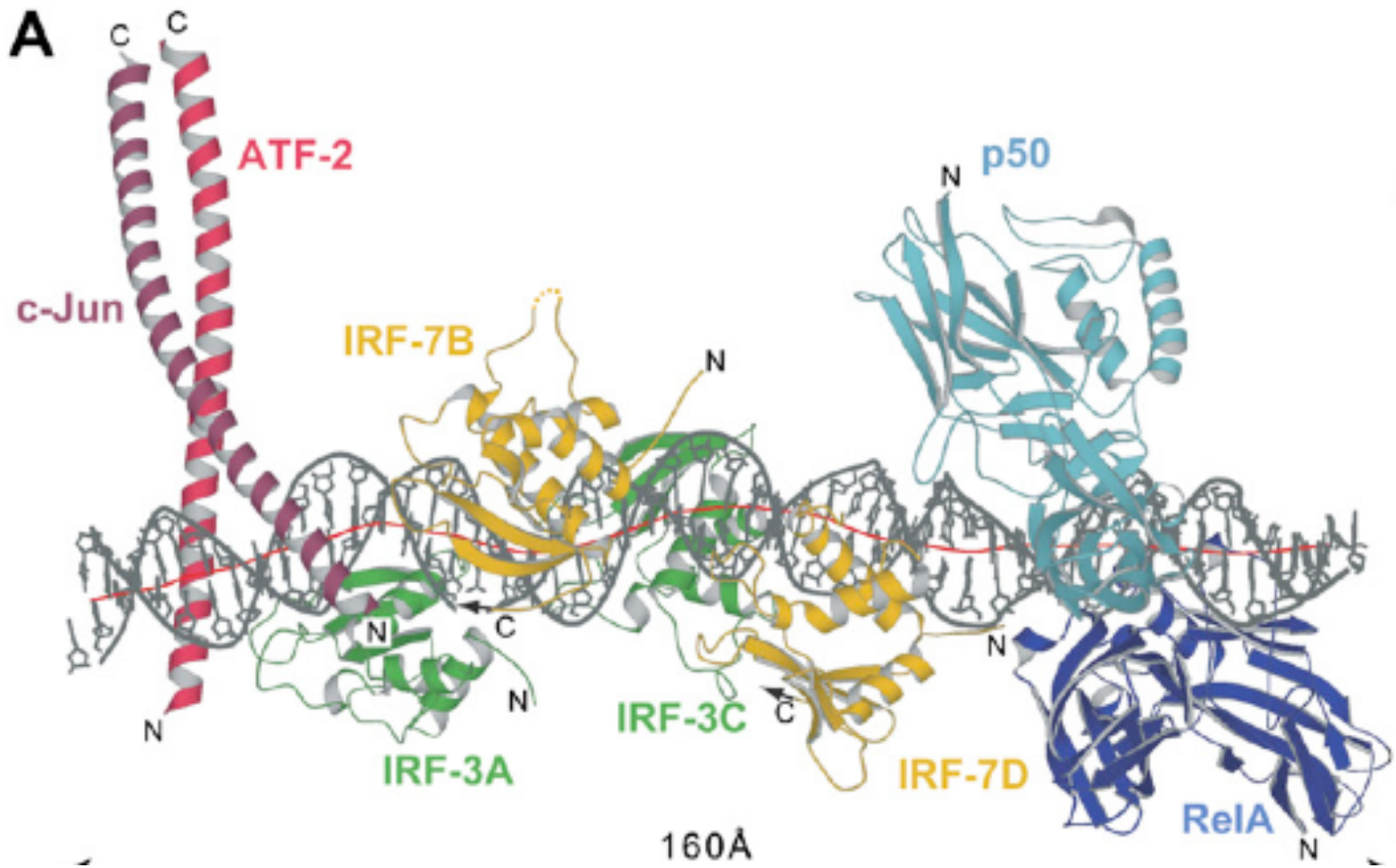
Cell *154*, 530–540, August 1, 2013 ©2013

Transcription factors and promoters

Eukaryotes



Few protein-protein interactions

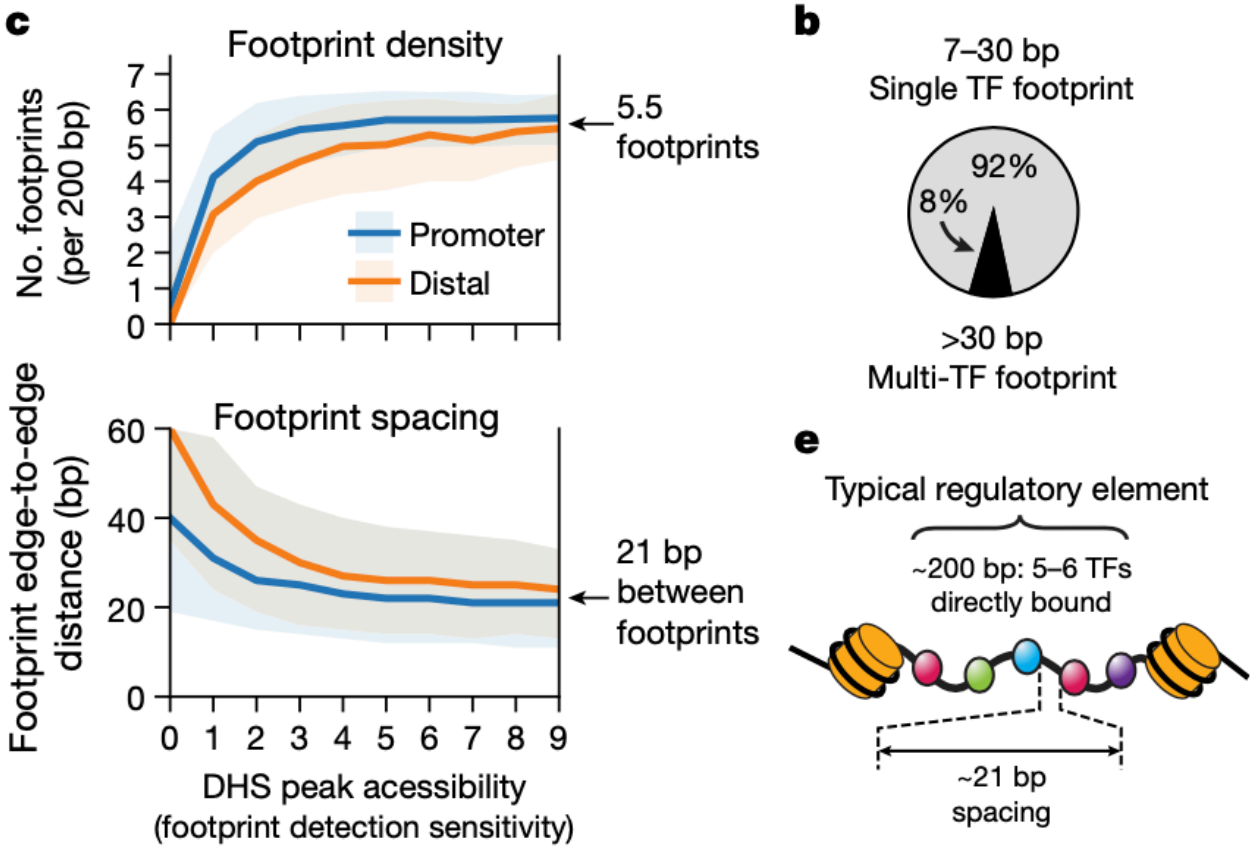


An Atomic Model of the Interferon-β Enhanceosome

Daniel Panne,[1] Tom Maniatis,[2] and Stephen C. Harrison[1,*]

Cell

Clusters of regulatory motifs seen by DHS footprinting



Global reference mapping of human transcription factor footprints

https://doi.org/10.1038/s41586-020-2528-x
Received: 30 January 2020
Accepted: 25 June 2020
Published online: 29 July 2020

Jeff Vierstra[1]✉, John Lazar[1,2], Richard Sandstrom[1], Jessica Halow[1], Kristen Lee[1], Daniel Bates[1], Morgan Diegel[1], Douglas Dunn[1], Fidencio Neri[1], Eric Haugen[1], Eric Rynes[1], Alex Reynolds[1], Jemma Nelson[1], Audra Johnson[1], Mark Frerker[1], Michael Buckley[1], Rajinder Kaul[1], Wouter Meuleman[1] & John A. Stamatoyannopoulos[1,2,3]✉

Article

Low overlap of transcription factor DNA binding and regulatory targets

https://doi.org/10.1038/s41586-025-08916-0

Received: 12 July 2023

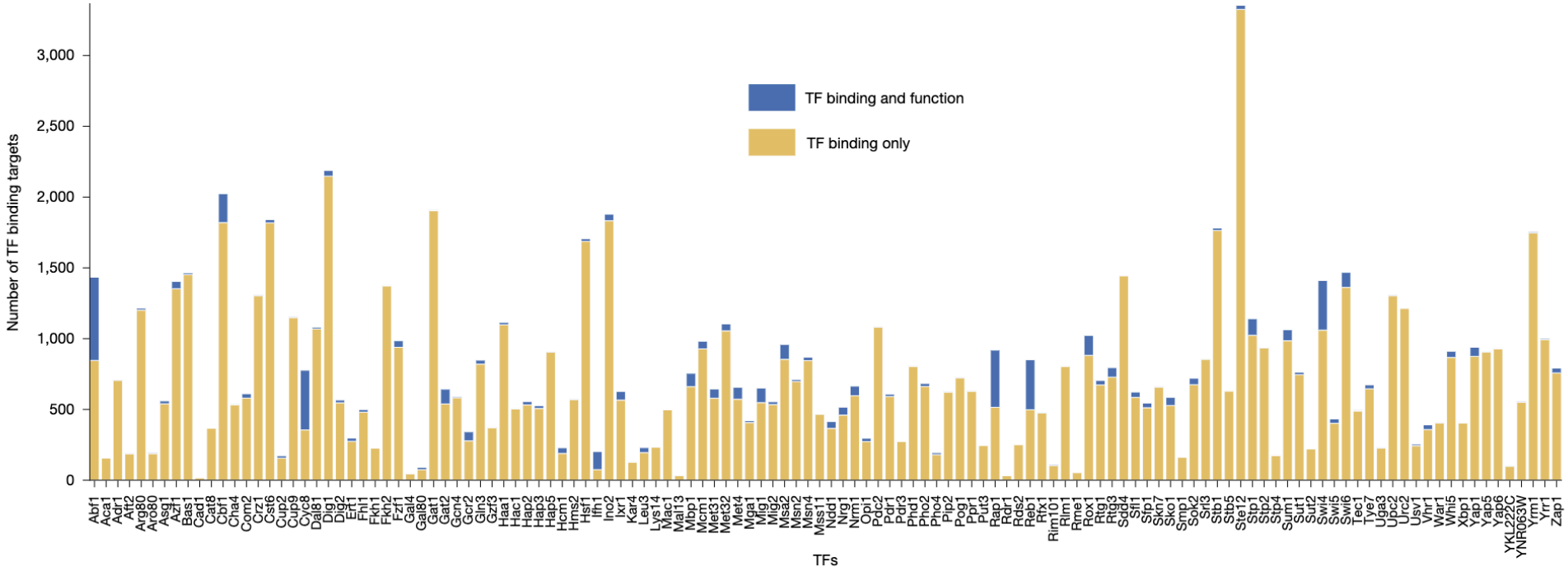
Accepted: 19 March 2025

Published online: 16 April 2025

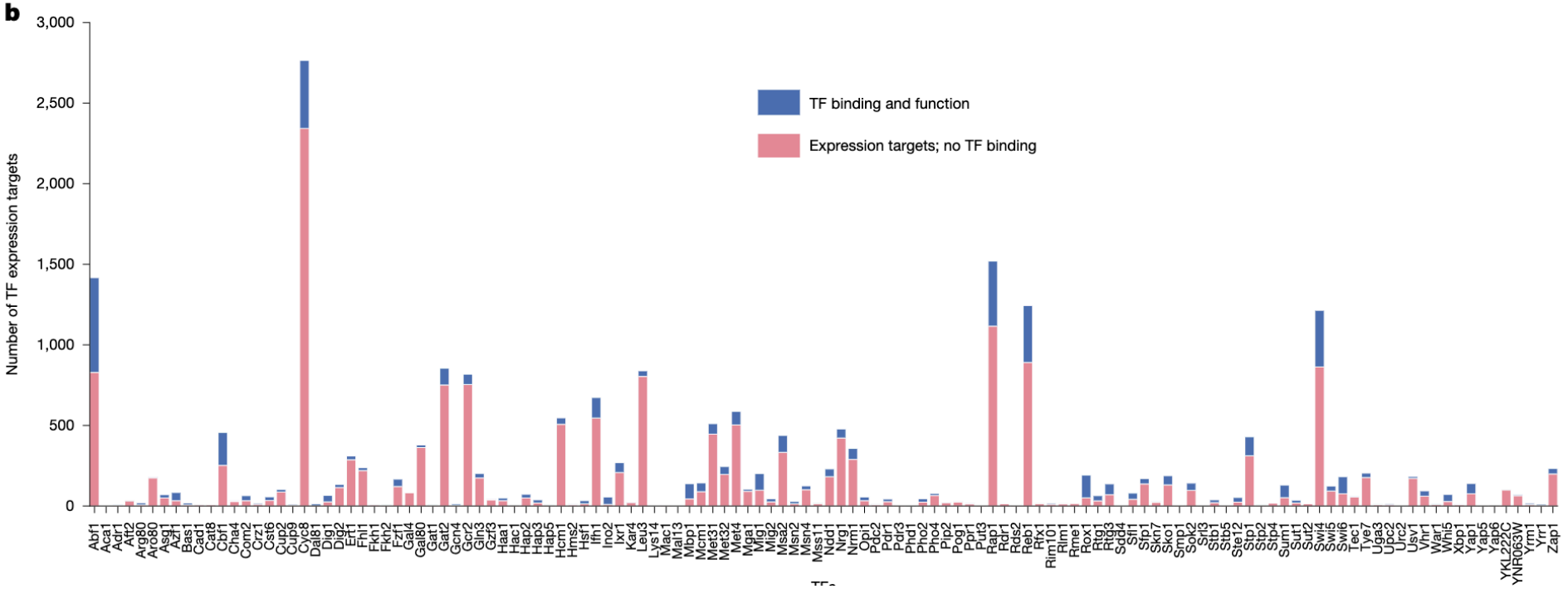
Check for updates

Lakshmi Mahendrawada[1], Linda Warfield[1], Rafal Donczew[1,2] & Steven Hahn[1]✉

DNA sequence-specific transcription factors (TFs) modulate transcription and chromatin architecture, acting from regulatory sites in enhancers and promoters of eukaryotic genes[1,2]. How multiple TFs cooperate to regulate individual genes is still unclear. In yeast, most TFs are thought to regulate transcription via binding to upstream activating sequences, which are situated within a few hundred base pairs upstream of the regulated gene[3]. Although this model has been validated for individual TFs and specific genes, it has not been tested in a systematic way. Here we integrated information on the binding and expression targets for the near-complete set of yeast TFs and show that, contrary to expectations, there are few TFs with dedicated activator or repressor roles, and that most TFs have a dual function. Although nearly all protein-coding genes are regulated by one or more TFs, our analysis revealed limited overlap between TF binding and gene regulation. Rapid depletion of many TFs also revealed many regulatory targets that were distant from detectable TF binding sites, suggesting unexpected regulatory mechanisms. Our study provides a comprehensive survey of TF functions and offers insights into interactions between the set of TFs expressed in a single cell type and how they contribute to the complex programme of gene regulation.


Number of TF binding targets
3,000
2,500
2,000
1,500
1,000
500
0
TF binding and function
TF binding only
TFs

b


Number of TF expression targets
3,000
2,500
2,000
1,500
1,000
500
0
TF binding and function
Expression targets; no TF binding