Using deep learning regulatory models and random DNA for evolutionary inference

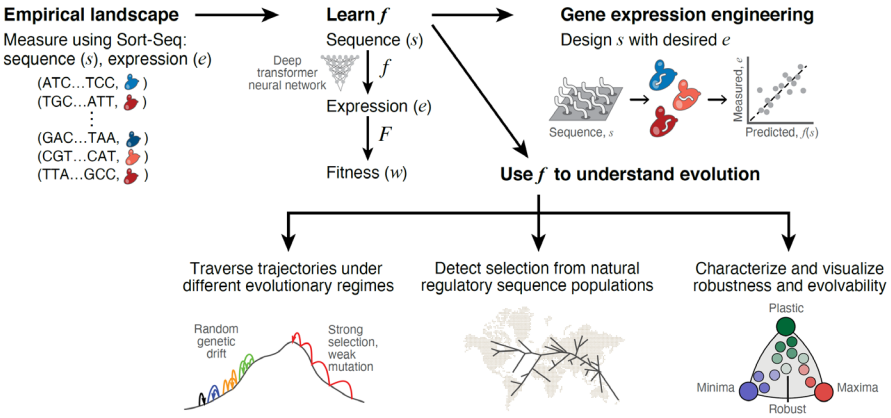
BROAD INSTITUTE

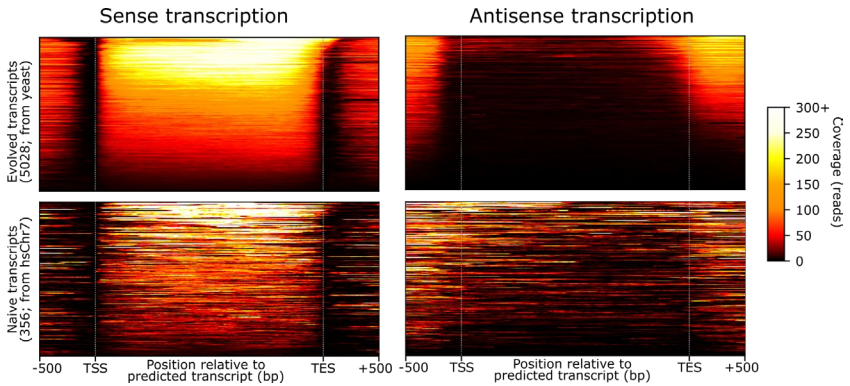
KLARMAN CELL OBSERVATORY
AT BROAD INSTITUTE

Carl de Boer
@Carldeboer@sciencemastodon.com
@CarldeBoerPhD
Carl.deboer@ubc.ca

THE UNIVERSITY OF BRITISH COLUMBIA

SBME UBC
SCHOOL OF BIOMEDICAL ENGINEERING

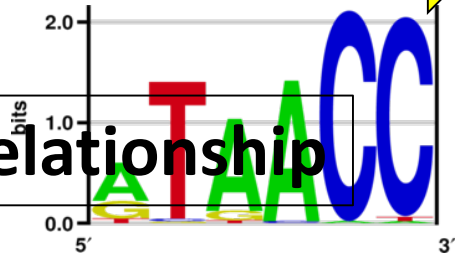




Regulatory variant function

Allele 1

TCGATAACCCCGAGT



Goal: *predict* sequence-expression relationship

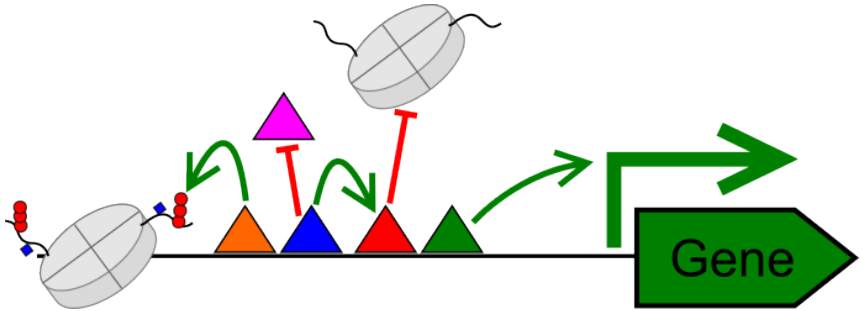
SNP

Allele 2

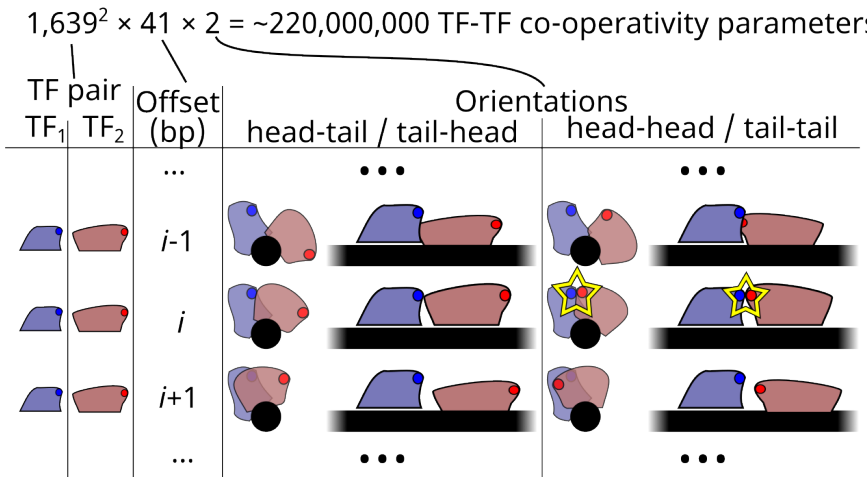
TCGATAATCCCGAGT

The “language” of regulatory DNA

- Alphabet: A, T, G, C
- Words: Transcription Factor binding sites
- Grammar: incompletely known
  - TFs modify expression
  - TFs modify chromatin
  - TFs modify each other



How complex could it be?



	Humans	Yeast
Number of TFs	1639	209
Number of parameters	>220,000,000	>5,000,000
Regulatory regions	~50,000 per cell	~6,000

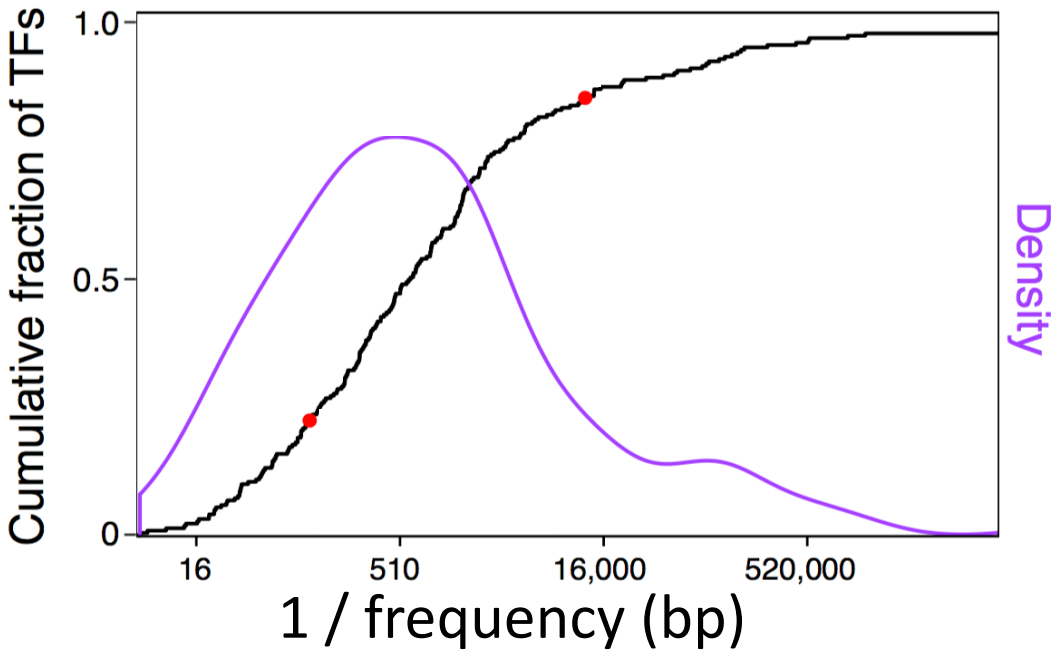
Too few Examples of regulatory DNA!

Data from (Lambert et al. Cell. 2018; de Boer et al. NAR. 2012)

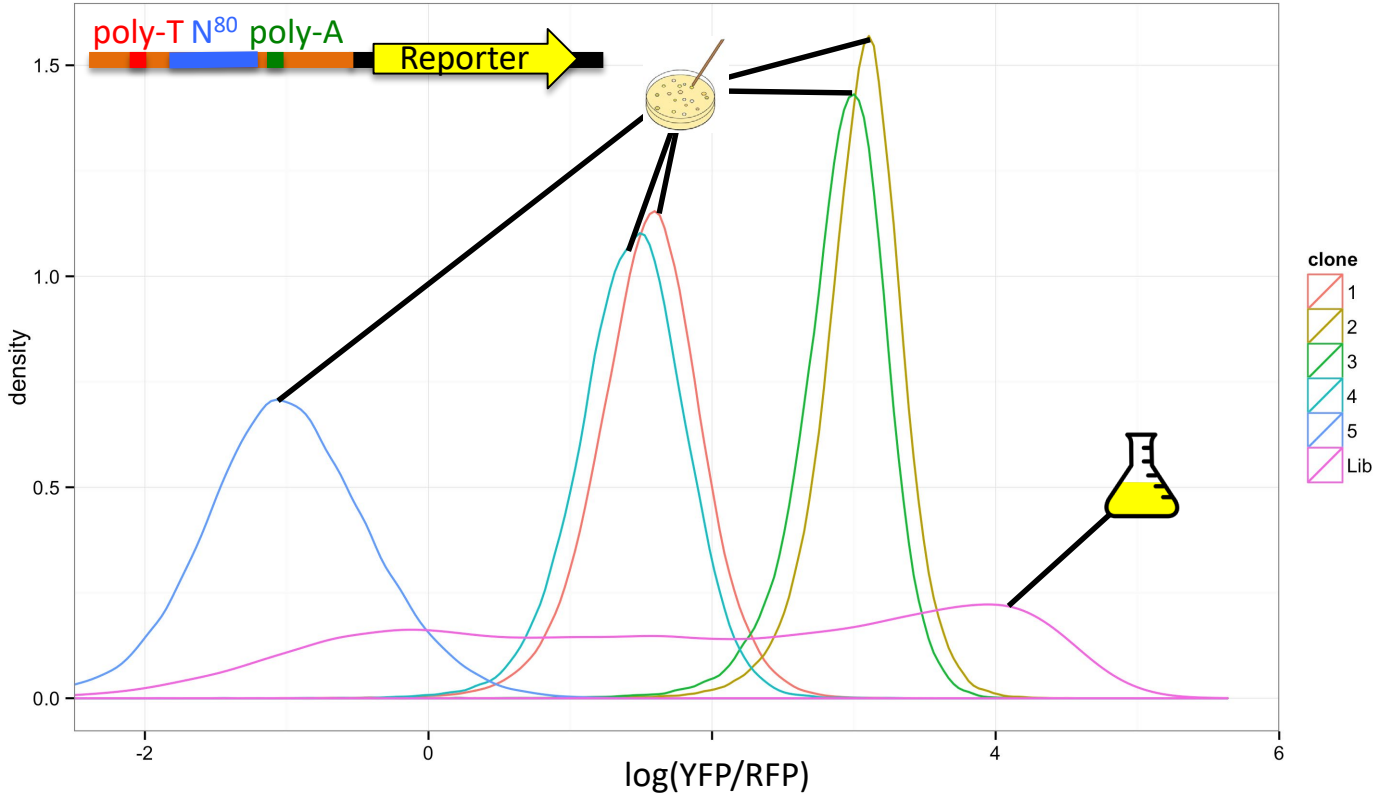
Opportunity: TF motifs predict frequent binding in ***random*** **DNA**

In general:
motif of *x* bits
-> every 2^{x-1} bases

Average random 80-mer: 138 Binding sites for 68 different TFs (!)



Random DNA has diverse expression



Random DNA is ideal for learning sequence -> expression

- Sequences are unrelated
- Diverse TFBSs {binding strength, position, orientation, etc}
- MANY examples

Measuring >10 million promoter expressions/experiment

Grow library

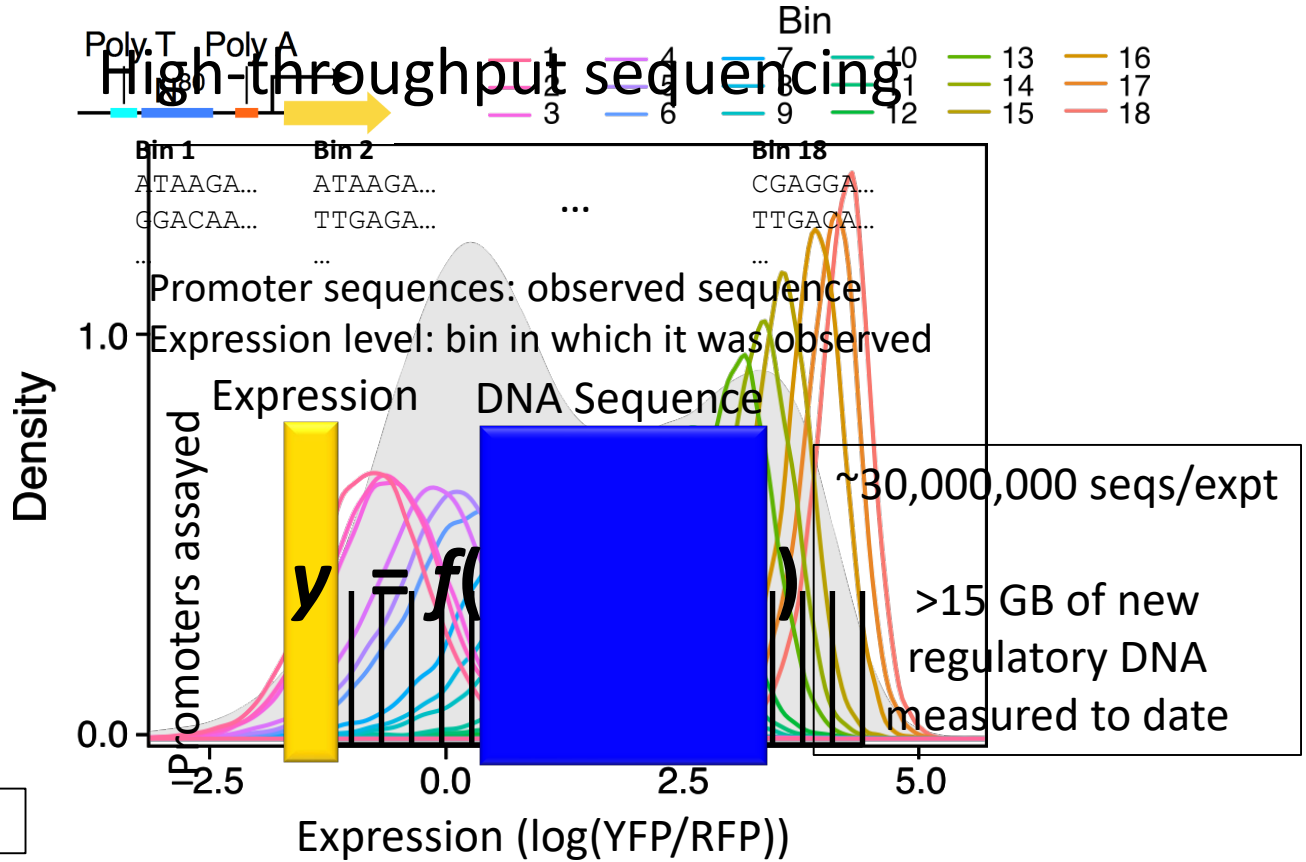
Sort cells by expression

Regrow

Sequence promoters

RFP: constant bckground

High-throughput sequencing



~30,000,000 seqs/expt

>15 GB of new regulatory DNA measured to date

Part 1: Deep learning model illuminates evolution of regulatory sequences

Two sequence-> expression models

Interpretable model

- Explicitly learns TF biochemistry
- Good predictive power
- **Interpretable**

ARTICLES
https://doi.org/10.1038/s41587-019-0315-8
nature biotechnology
There are amendments to this paper

Deciphering eukaryotic gene-regulatory logic with 100 million random promoters

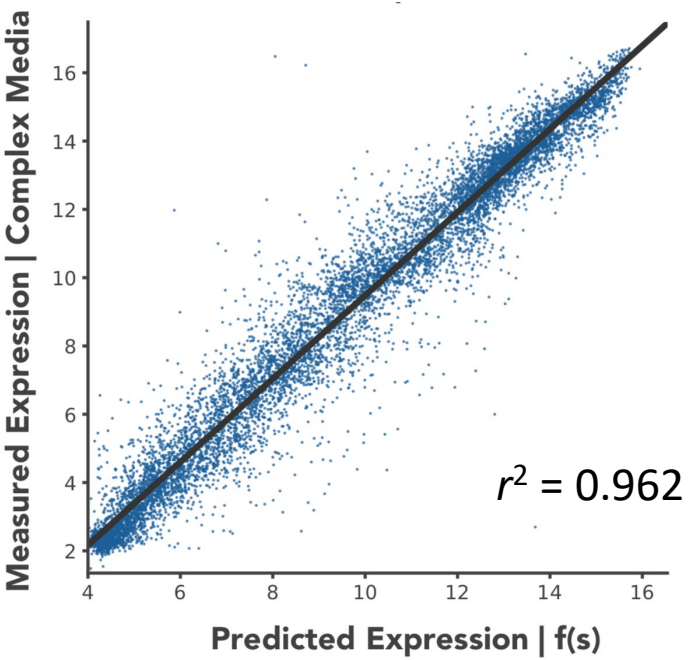
Carl G. de Boer[1*], Eeshit Dhaval Vaishnav[1,2], Ronen Sadeh[3], Esteban Luis Abeyta[4], Nir Friedman[1,3] and Aviv Regev[1,2*]

Neural network

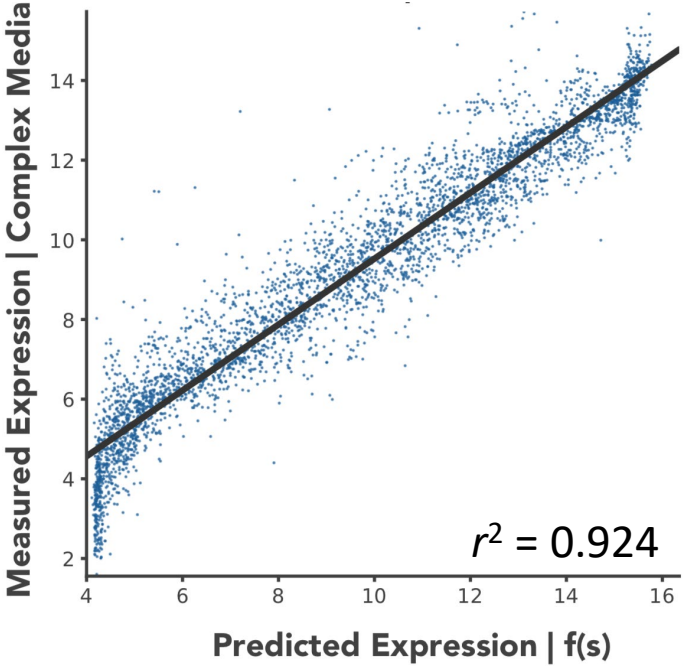
- Can capture unknown mechanisms
- **Excellent predictive power**
- Difficult to interpret

Eeshit Dhaval Vaishnav

Neural network model predicts expression with high accuracy

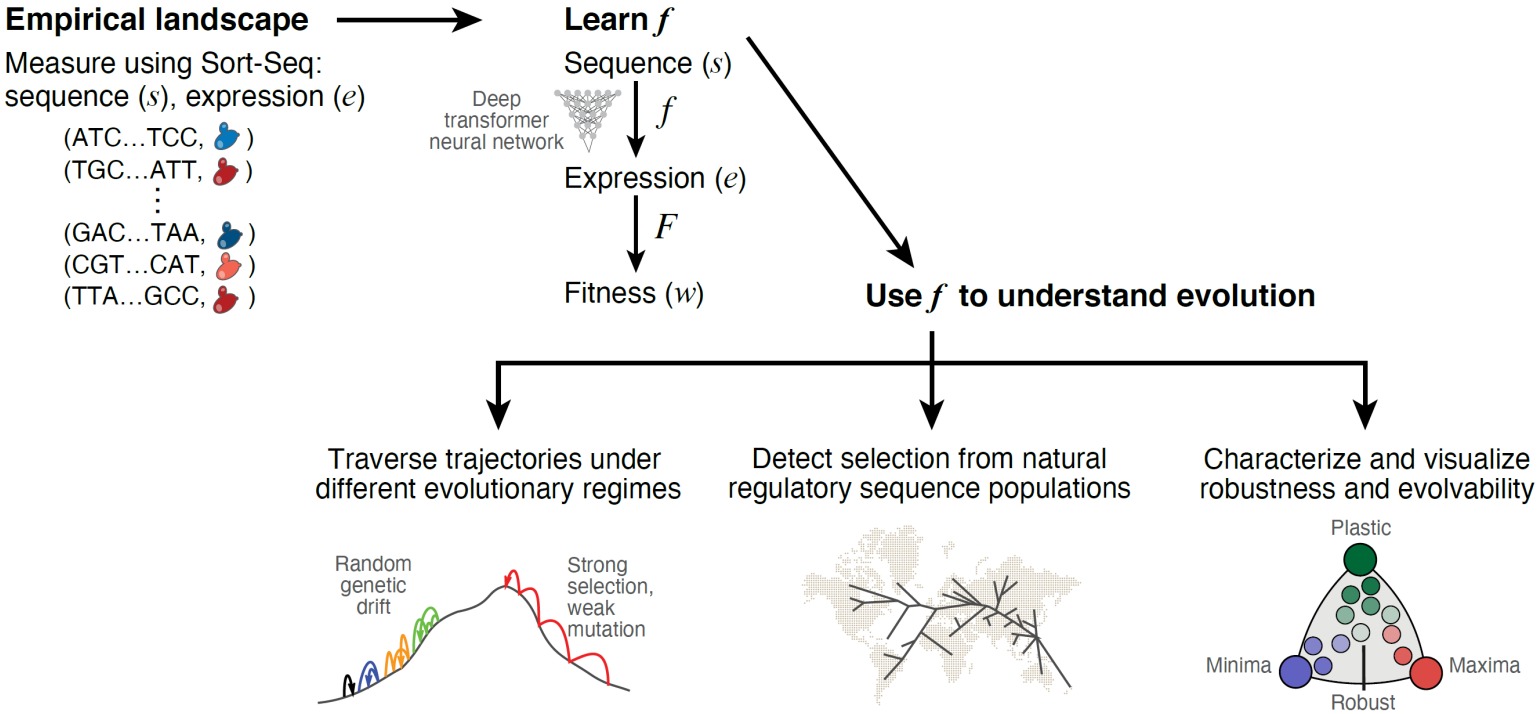


Test data: ~10,000 promoters from different library, separate experiment



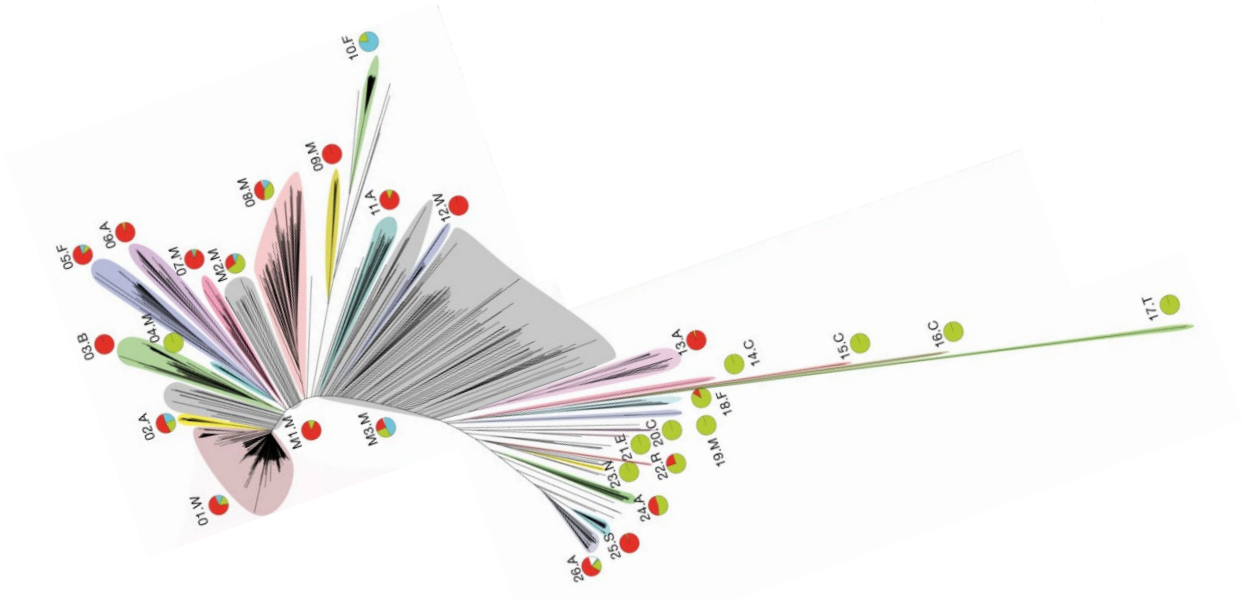
Test data: ~5,000 80 bp fragments of native promoters in pTpA scaffold, separate experiment

Answering evolutionary questions using a gene expression “Oracle”



Question 1: can we quantify selection on promoter sequences using our deep learning oracle?

Learning from regulatory variation

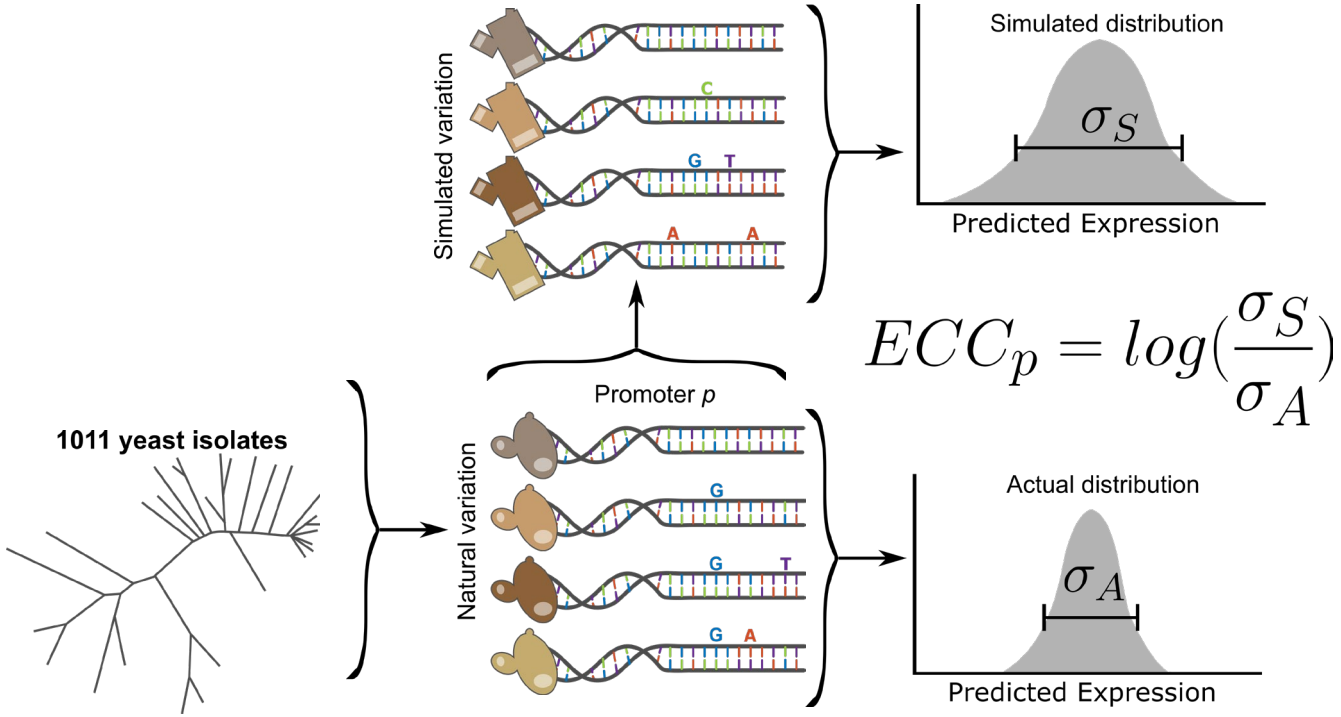


Nature. 2018 Apr;556(7701):339-344. doi: 10.1038/s41586-018-0030-5. Epub 2018 Apr 11.

Genome evolution across 1,011 Saccharomyces cerevisiae isolates.

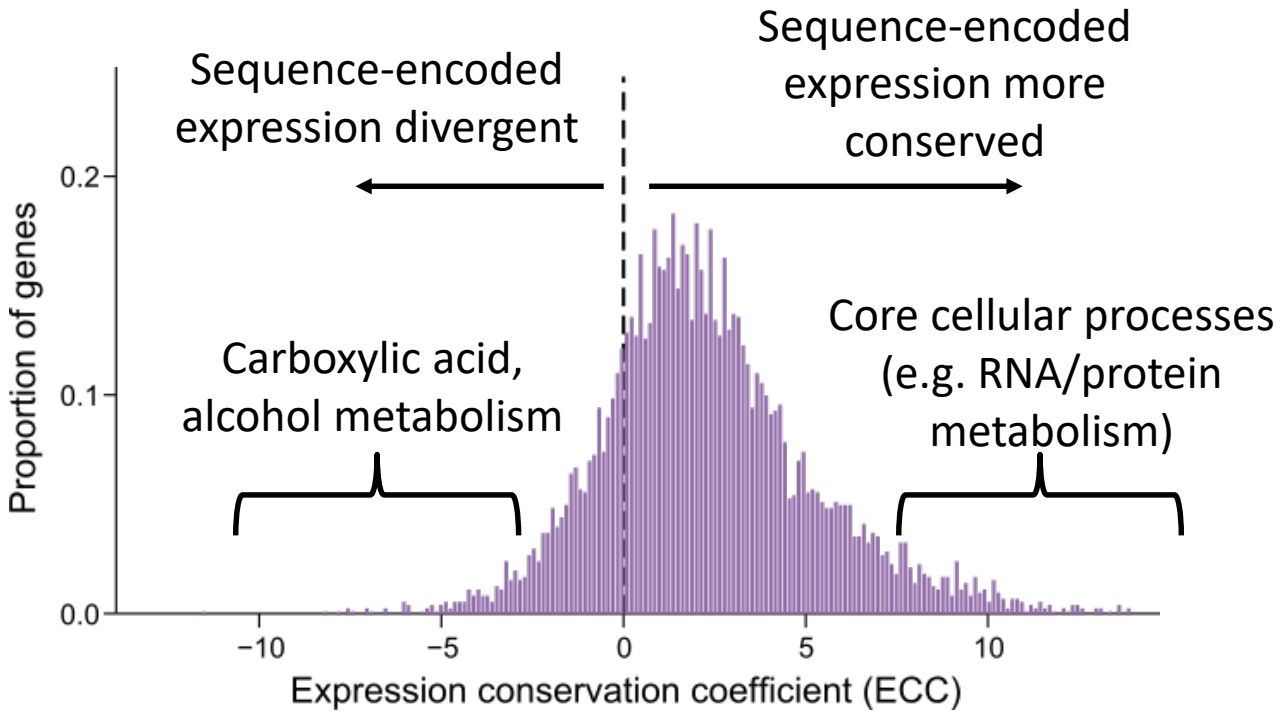
Peter J[1], De Chiara M[2], Friedrich A[1], Yue JX[2], Pflieger D[1], Bergström A[2], Sigwalt A[1], Barre B[2], Freel K[1], Llored A[2], Cruaud C[3], Labadie K[3], Aury JM[3], Istace B[3], Lebrigand K[4], Barbry P[4], Engelen S[3], Lemainque A[3], Wincker P[3,5], Liti G[6], Schacherer J[7].

Quantifying Expression Conservation

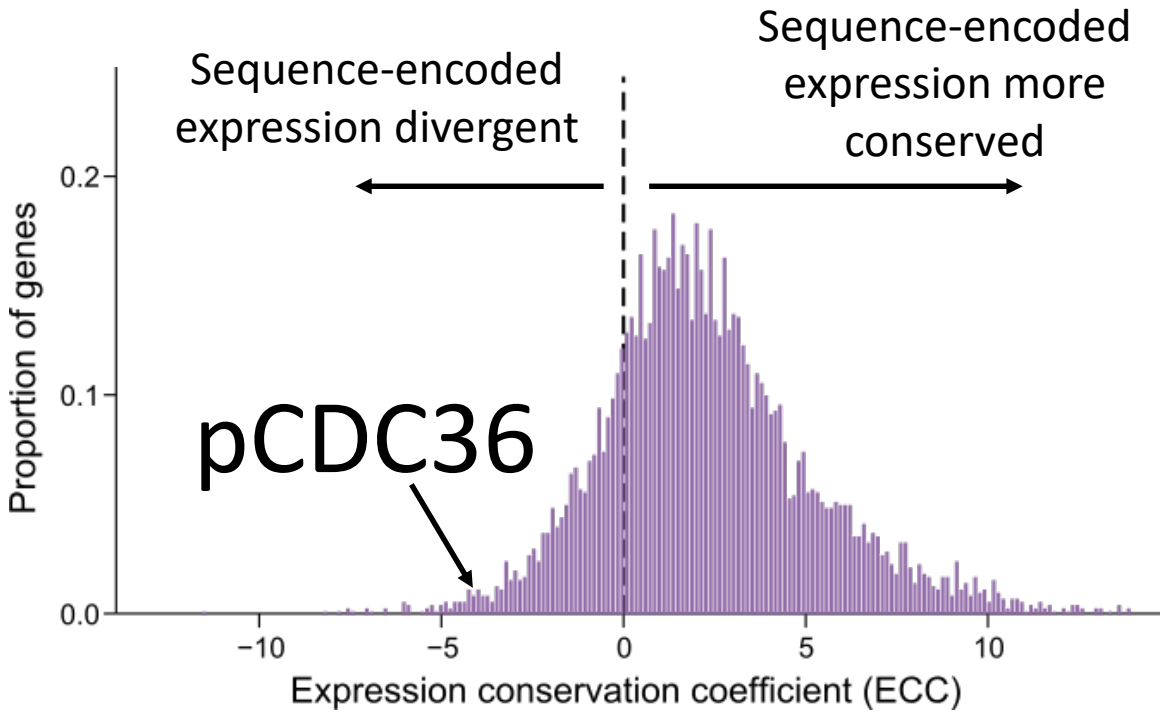


$$ECC_p = log(\frac{\sigma_S}{\sigma_A})$$

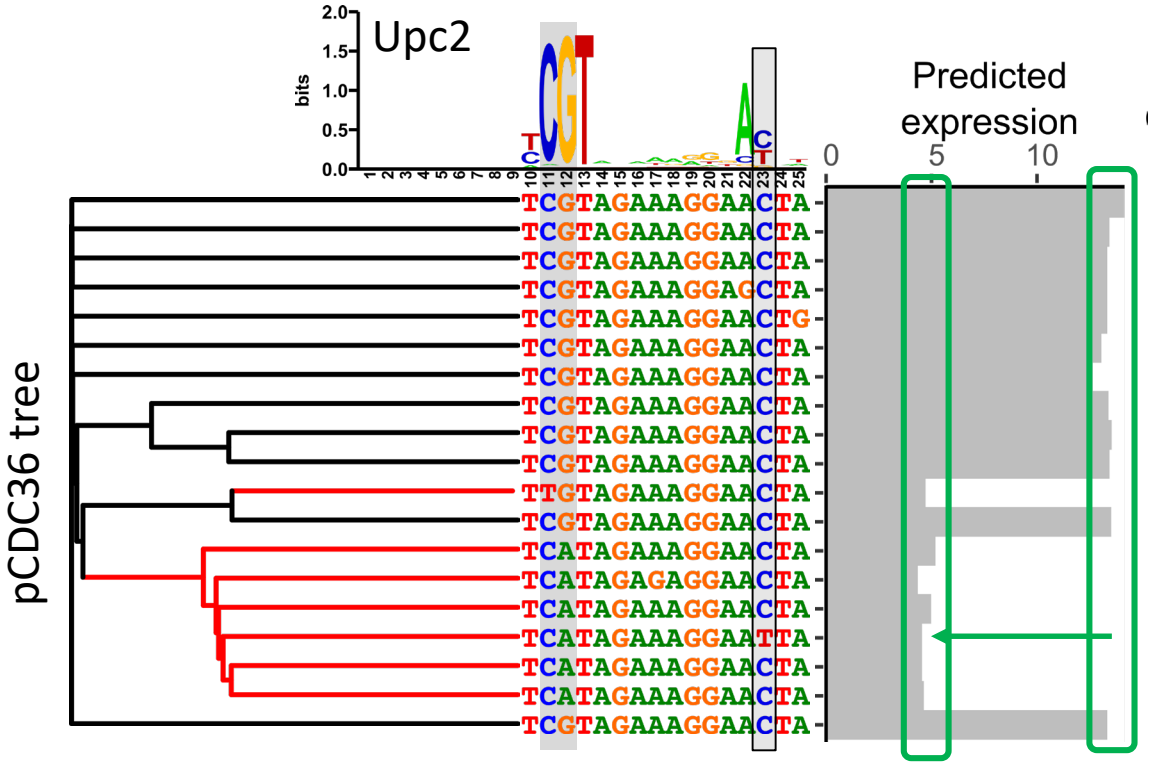
Most genes show expression conservation by ECC



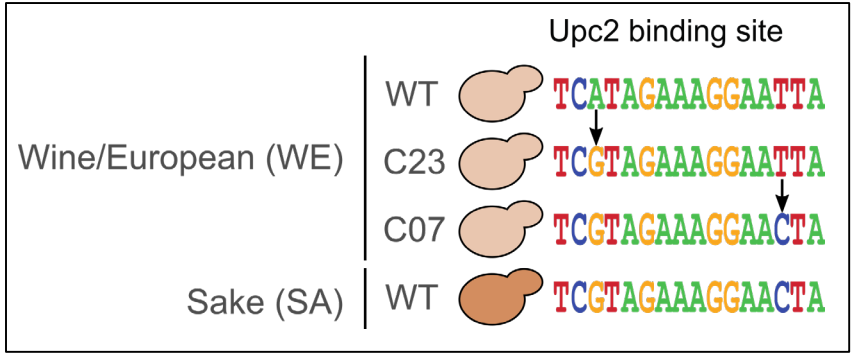
CDC36: an interesting example



Natural selection targeted same UPC2 binding site twice

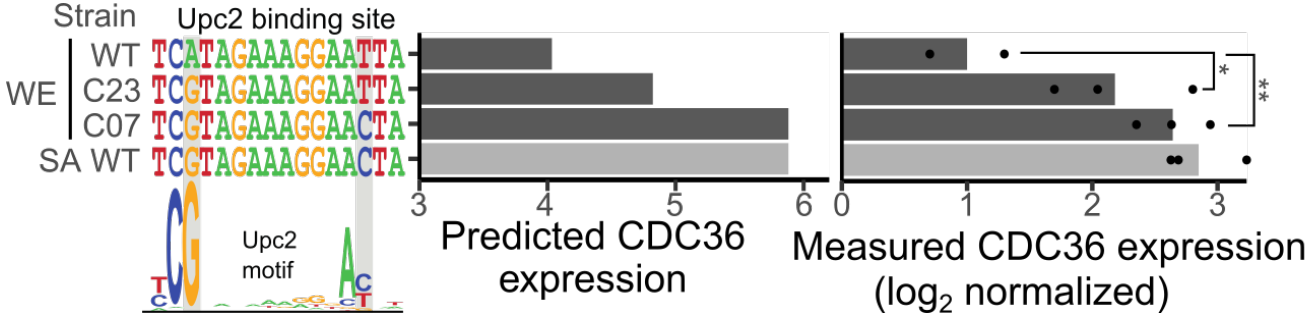


CDC36 promoter variants affect expression and fitness



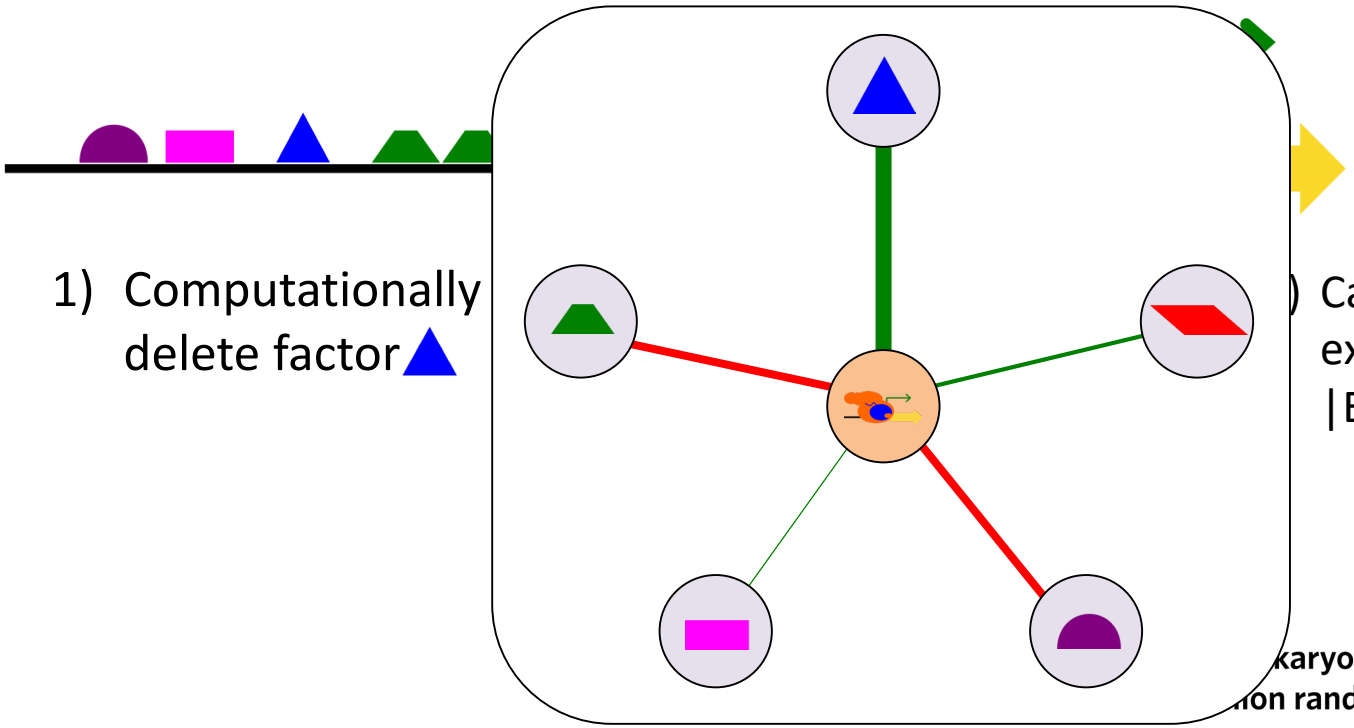
Jennifer Molinet

Francisco Cubillos



Question 2: does selection shape the complexity of regulation, or just expression level?

Predicting regulatory interactions with an interpretable “biochemical” model



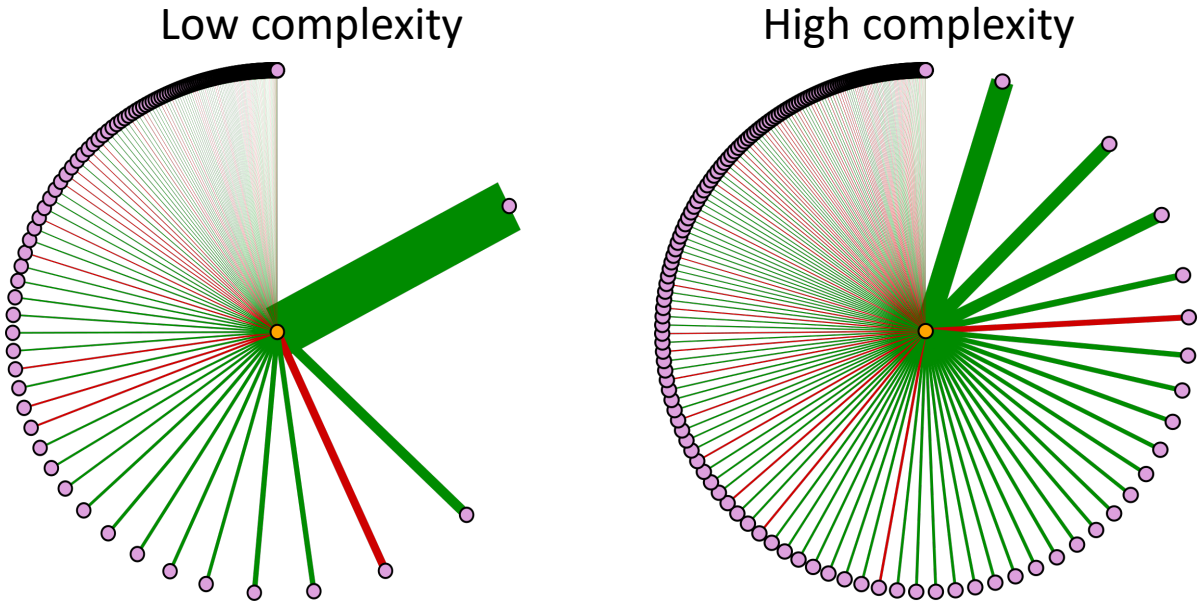
1) Computationally delete factor ▲

) Calculate interaction expression “effect” $|EL_0 - EL(\Delta▲)|$

karyotic gene-regulatory logic
ion random promoters

Carl G. de Boer[1*], Eeshit Dhaval Vaishnav[1,2], Ronen Sadeh[3], Esteban Luis Abeyta[4], Nir Friedman[1,3] and Aviv Regev[1,2*]

Identical expression levels encoded by different regulatory logic

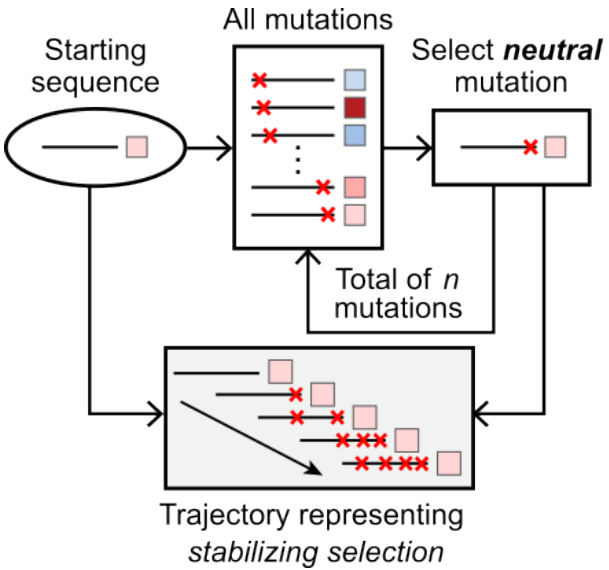


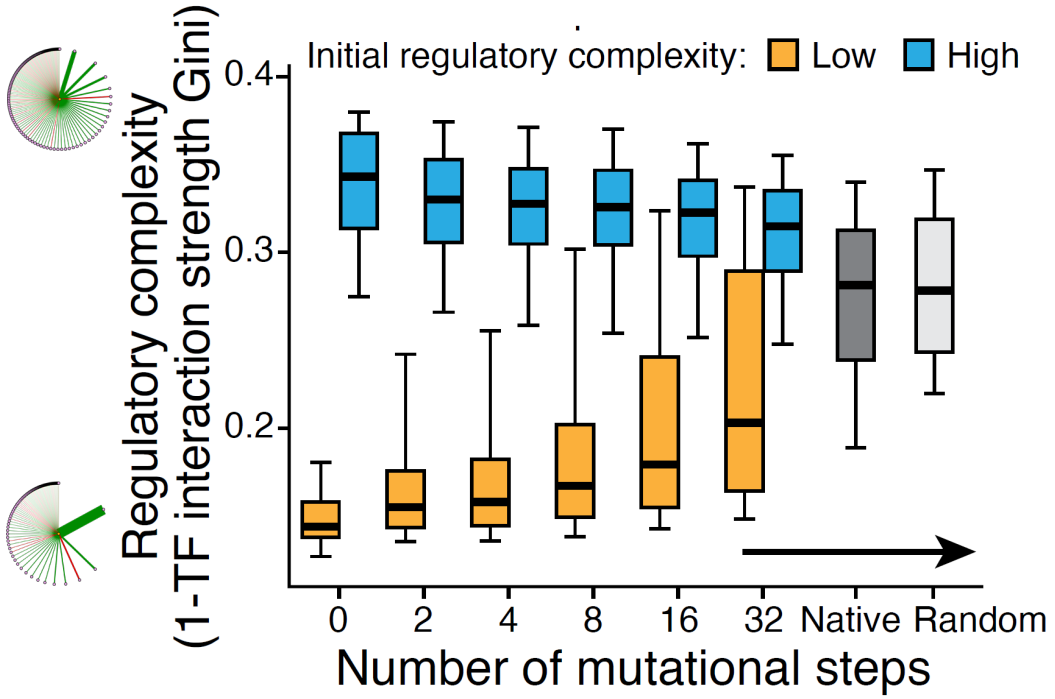
Complexity = 1-Gini

Transcription factor

Regulated sequence

Little evidence for selection on regulatory complexity





Conclusions Part 1: Deep learning Oracle

- Gene expression oracle can identify expression evolution using population genetics data
  - CDC36: Example of convergent expression adaptation in yeast
- Little evidence for selection on complexity of regulation:
  - Expression level is more important than how you get it
- See paper for details and ~80% more evolutionary insight

The evolution, evolvability and engineering of gene regulatory DNA

Eeshit Dhaval Vaishnav, Carl G. de Boer, Jennifer Molinet, Moran Yassour, Lin Fan, Xian Adiconis, Dawn A. Thompson, Joshua Z. Levin, Francisco A. Cubillos & Aviv Regev

Nature **603**, 455–463 (2022) | Cite this article

Predicting gene expression using millions of random promoter sequences
DREAM Challenge 2022

SBME UBC School of Biomedical Engineering · IBM Research · deep genomics · Google Research TPU Research Cloud · SageBionetworks

Abdul Muntakim Rafi

With:
- Pablo Meyer (IBM Research)
- Jake Albrecht (Sage Biosystems)
- Paul Boutros (UCLA)
- Julie Bletz (Sage Biosystems)
- Payman Yadollahpour (Broad/Genentech)

Final results

Top performers:
- Autosome.org
- BHI – dream challenge
- Unlock_DNA

Importantly: the teams *kicked our butts*

Detailed results presented at RSG/DREAM (Nov. 2022 / Las Vegas)

Preprint coming soon…

https://www.synapse.org/#!Synapse:syn28469146/wiki/617075

Part 2: Biochemical activity is the default DNA state

Transcripts

Random sequences have biochemical activity in reporter assays

But this is based on reporter systems

-tested in promoter/enhancer-like contexts, with a protein coding (reporter) gene

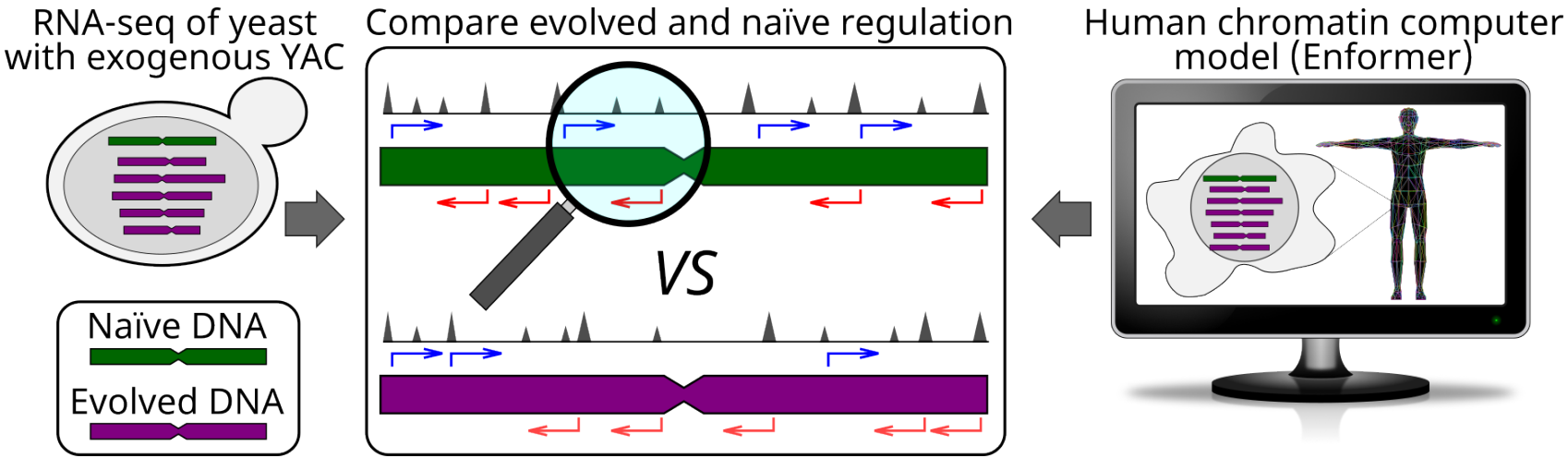
How much biochemical activity would we see in ***entire chromosomes*** of naïve DNA?

Biswajyoti Sahu, Tuomo Hartonen, Päivi Pihlajamaa, Bei Wei, Kashyap Dave, Fangjie Zhu, Eevi Kaasinen, Katja Lidschreiber, Michael Lidschreiber, Carsten O. Daub, Patrick Cramer, Teemu Kivioja & Jussi Taipale ✉

Controversy over the degree of adaptation/function in the genome

- Disagreement over fraction of the genome that is functional: 80% (ENCODE, 2012) vs <<25% (Graur, 2017)
- lncRNAs: 10,000-100,000; few with important function, others uncharacterized or biochemical noise?
- “Suppose we put a few million bases of entirely random synthetic DNA into a human cell, and do an ENCODE project on it. Will it be reproducibly transcribed into mRNA-like transcripts, reproducibly bound by DNA-binding proteins, and reproducibly wrapped around histones marked by specific chromatin modifications? I think yes.” (Eddy, 2013)

Making a “Draft Random ENCODE” for yeast and humans



bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

Biochemical activity is the default DNA state in eukaryotes

Ishika Luthra, Xinyi E. Chen, Cassandra Jensen, Abdul Muntakim Rafi, Asfar Lathif Salaudeen, Carl G. de Boer

doi: https://doi.org/10.1101/2022.12.16.520785

Yeast Artificial Chromosomes as a source of evolutionarily naïve DNA

- Used for sequencing the human genome
  - ~1MB of human DNA already in yeast
- Human and yeast:
  - Both eukaryotes (and opisthokonts)
  - ~1 billion years of evolution separate us
  - TFs function similarly, but few TFs conserved
  - Distinct gene regulatory structure

The quest for YACs

Hi Dr. Scherer. Got any YACs?

LOL! 30 year old YACs?!?! Unlikely!

Do us a biggie and check, pls?

OMG! Found the YACs!

OMG! 4 Real?!?!? I can has?

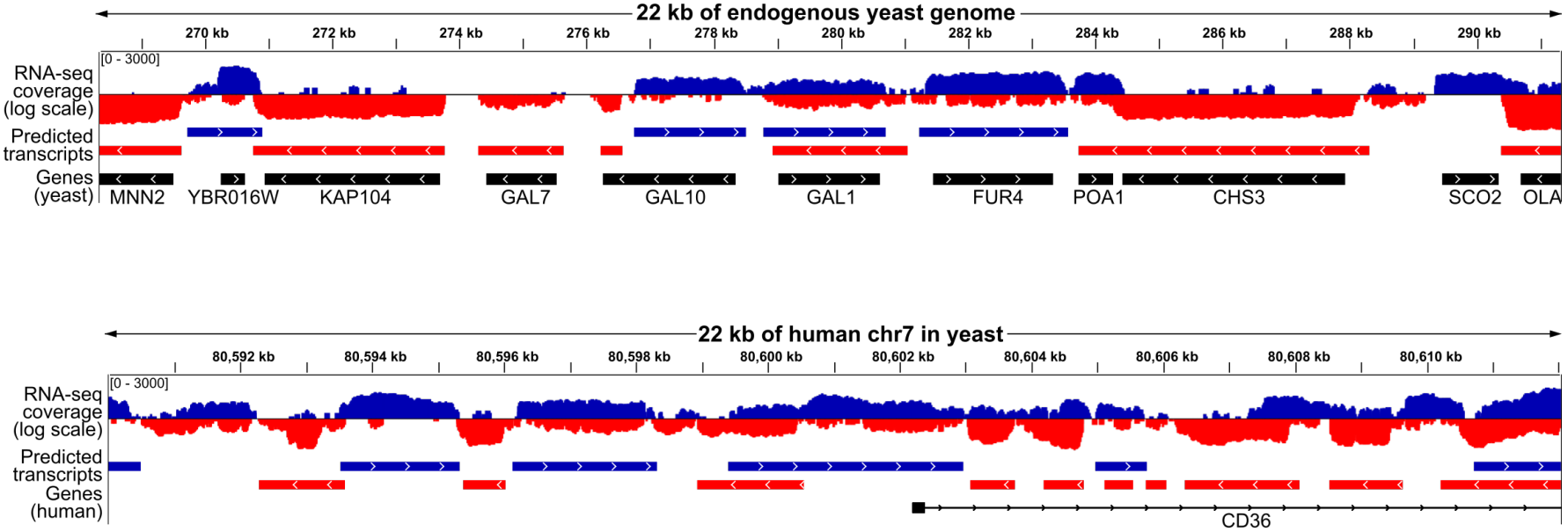
Cassandra Jensen
(de Boer lab)

Steve Scherer
(SickKids Toronto)

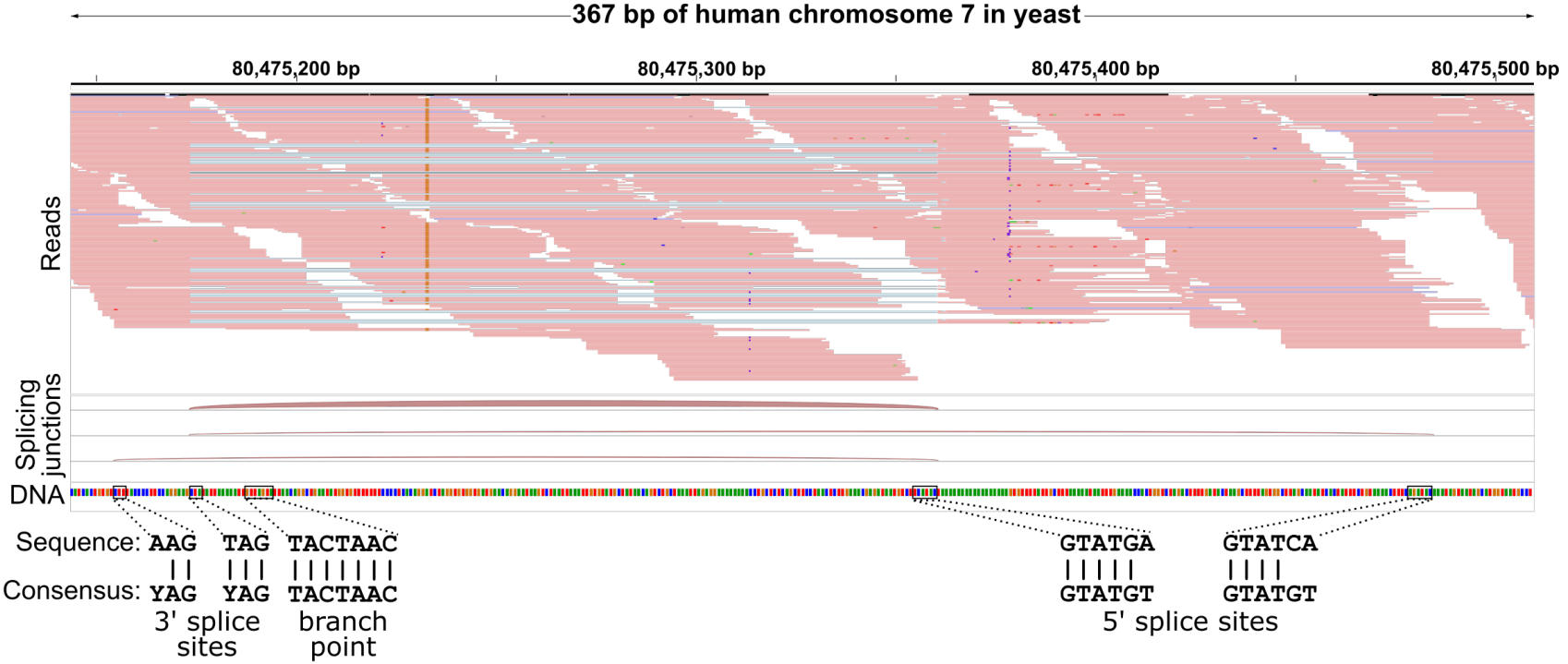
Transcriptional profiling of human DNA expressed in yeast

- Got 10 YACs, but only needed one:
  - ~760 kb of CD36 locus (chromosome 7)
- Performed strand-specific polyA+ RNA-seq
- Align to hybrid yeast genome+human chr7
- Analysis approach: compare expression from **Evolved** (endogenous yeast genome) to **Naïve** (human YAC DNA)

Naïve DNA is transcribed *extensively*

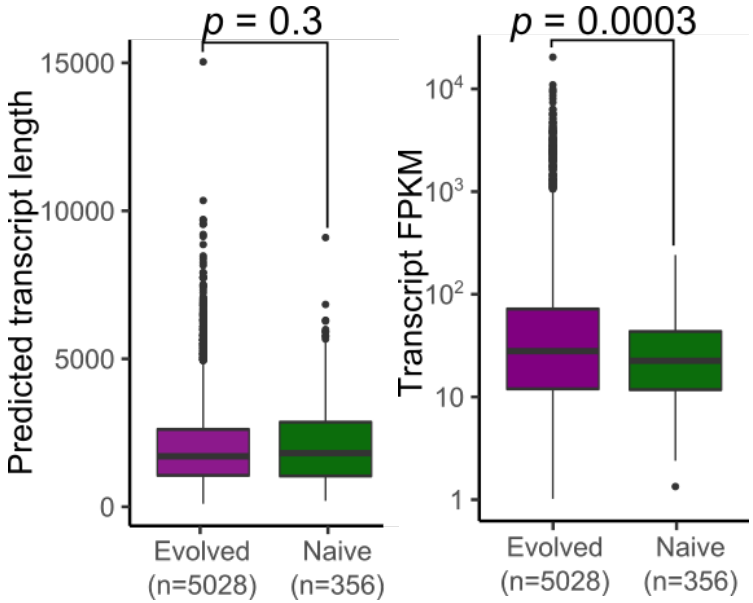


Some naïve transcripts are spliced



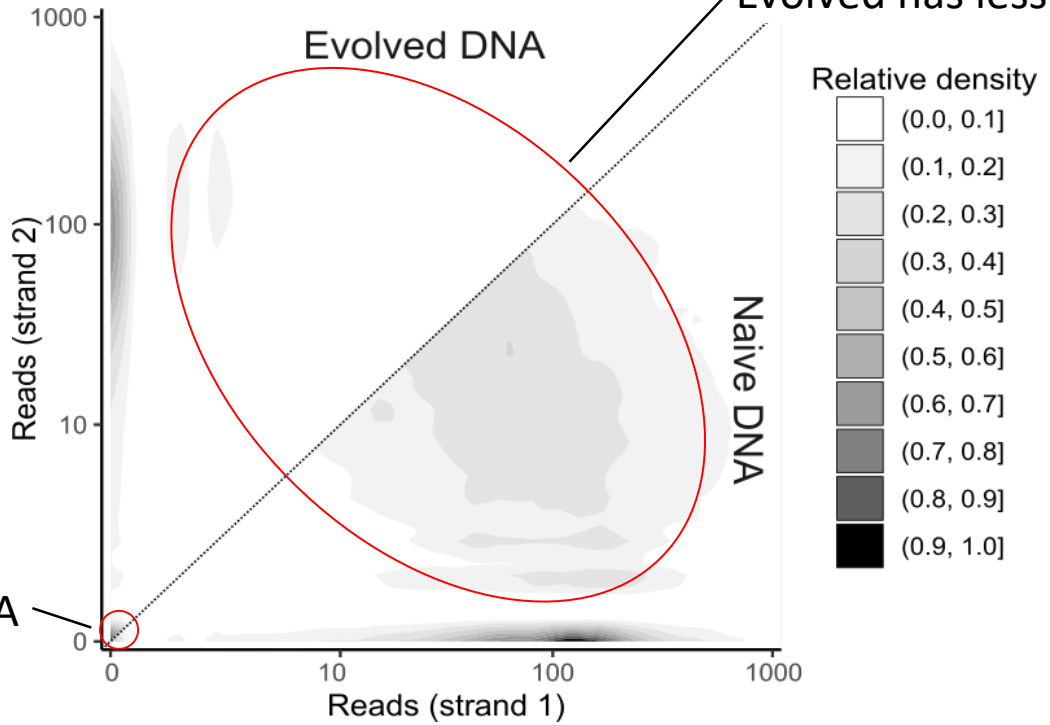
Naïve DNA produces gene-like transcripts

	Evolved	Naïve
DNA amount	13Mb	760 kb
Transcripts	5028	356
Transcripts/kb	0.42	0.47
Transcripts spliced	6%	2.8%



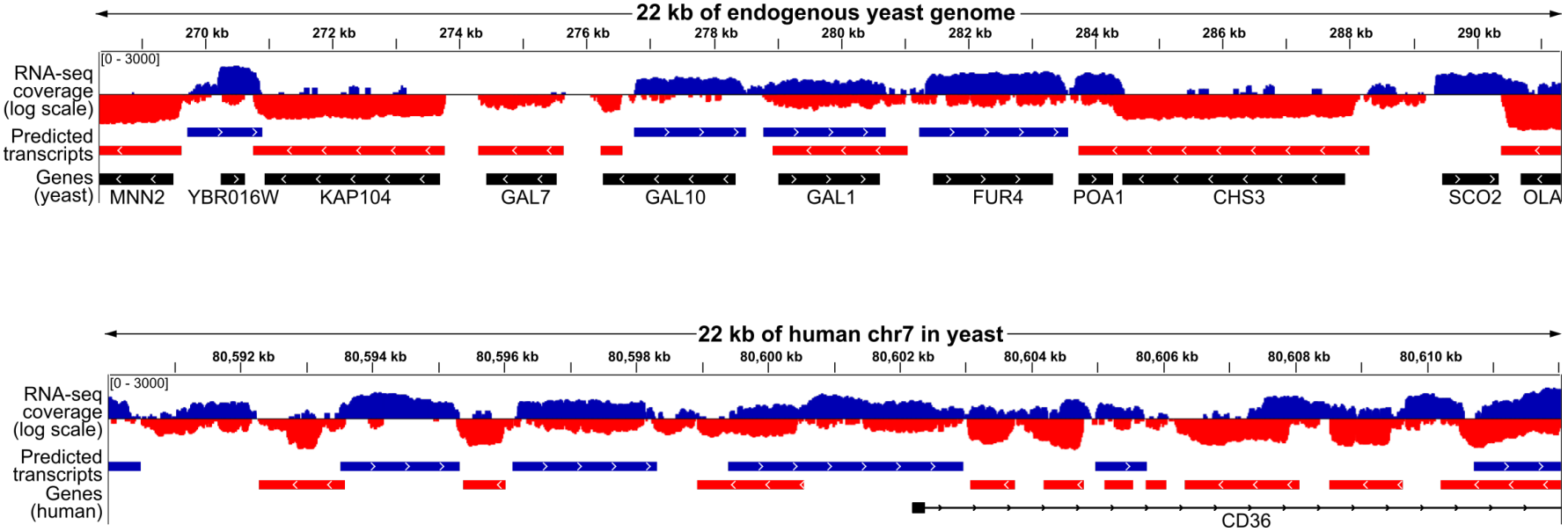
Endogenous DNA has evolved a coherent gene structure

Evolved has less antisense expression

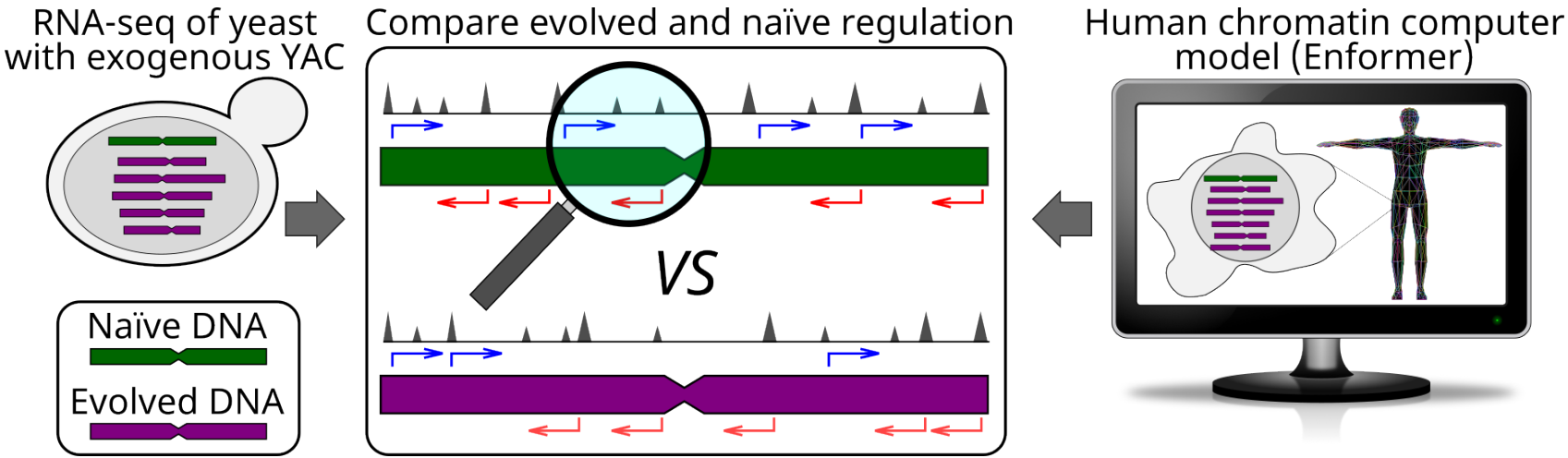


Evolved has more non-expressed DNA (3.4% vs 0.6%)

Endogenous DNA has evolved a coherent gene structure



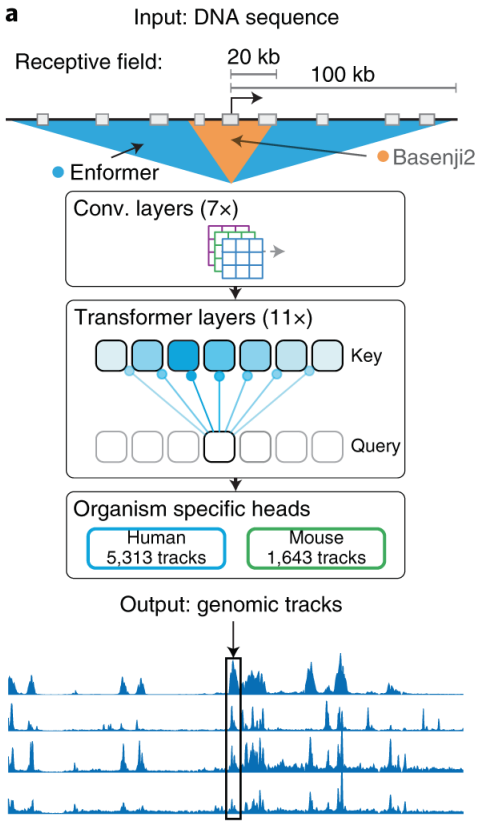
Making a “Draft Random ENCODE” for yeast and humans

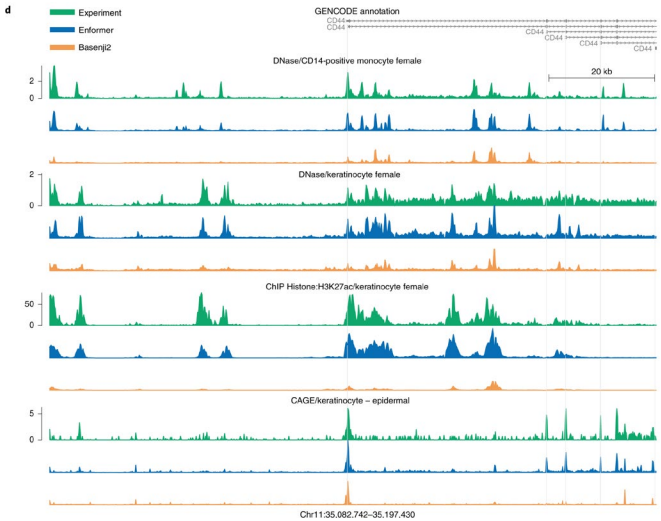


Experiments are hard…

- Experimentally challenging to experimentally test even 100 kb of naïve DNA
- Average 100 kb of human genomic DNA has ~0.6 genes, not all expressed, so need much more

Predicting regulatory activity with Enformer





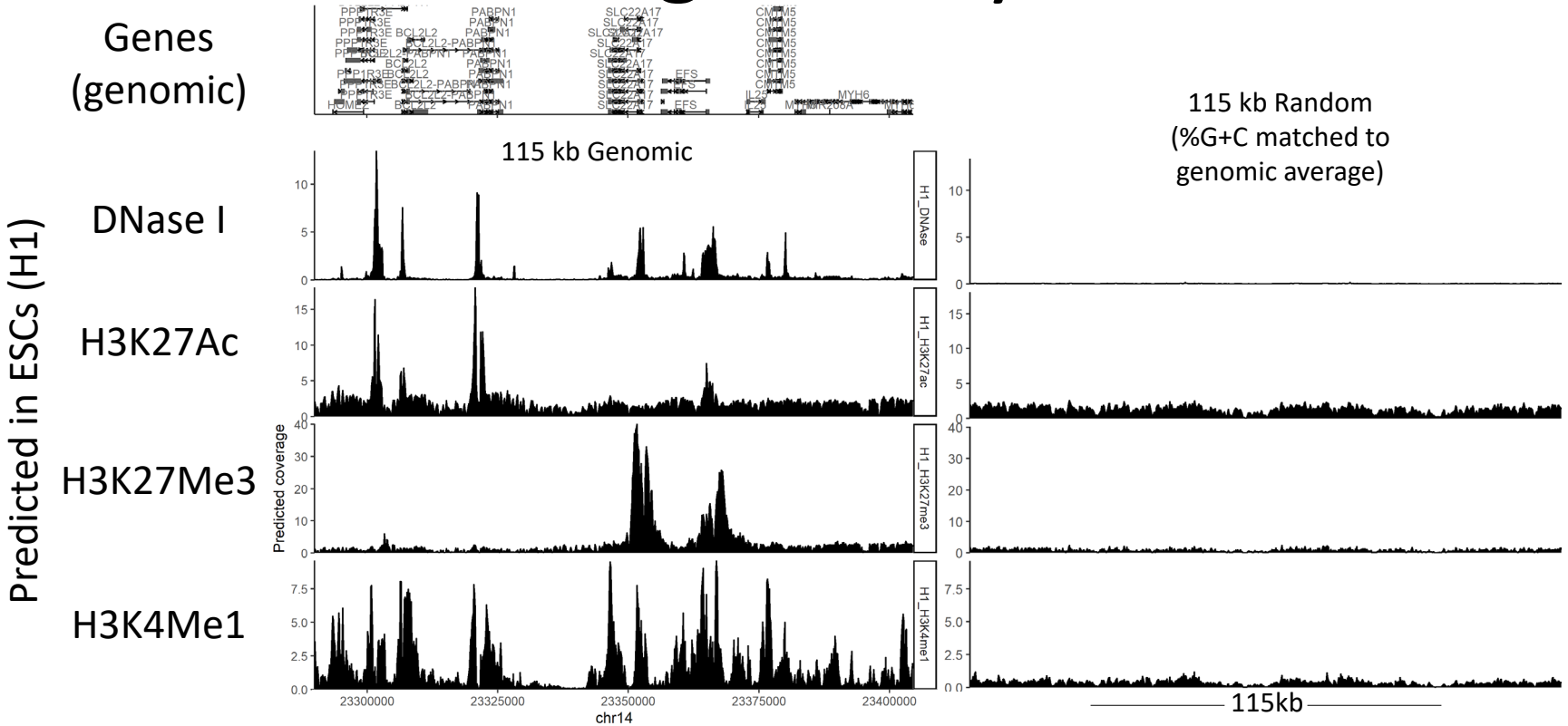
Our analysis: only considering genomic sequences in Enformer test set

Effective gene expression prediction from sequence by integrating long-range interactions

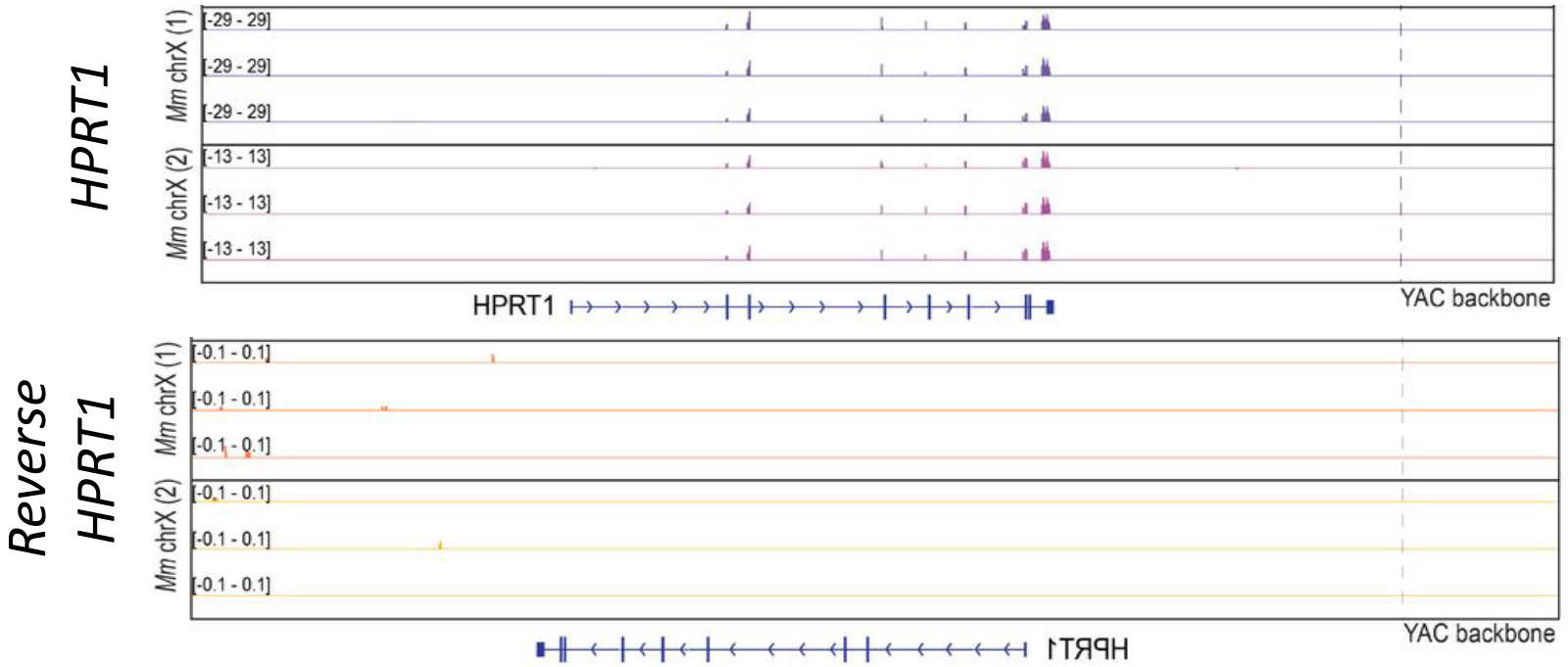
Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli & David R. Kelley

Nature Methods **18**, 1196–1203 (2021) | Cite this article

Completely random DNA is predicted to be regulatorily silent

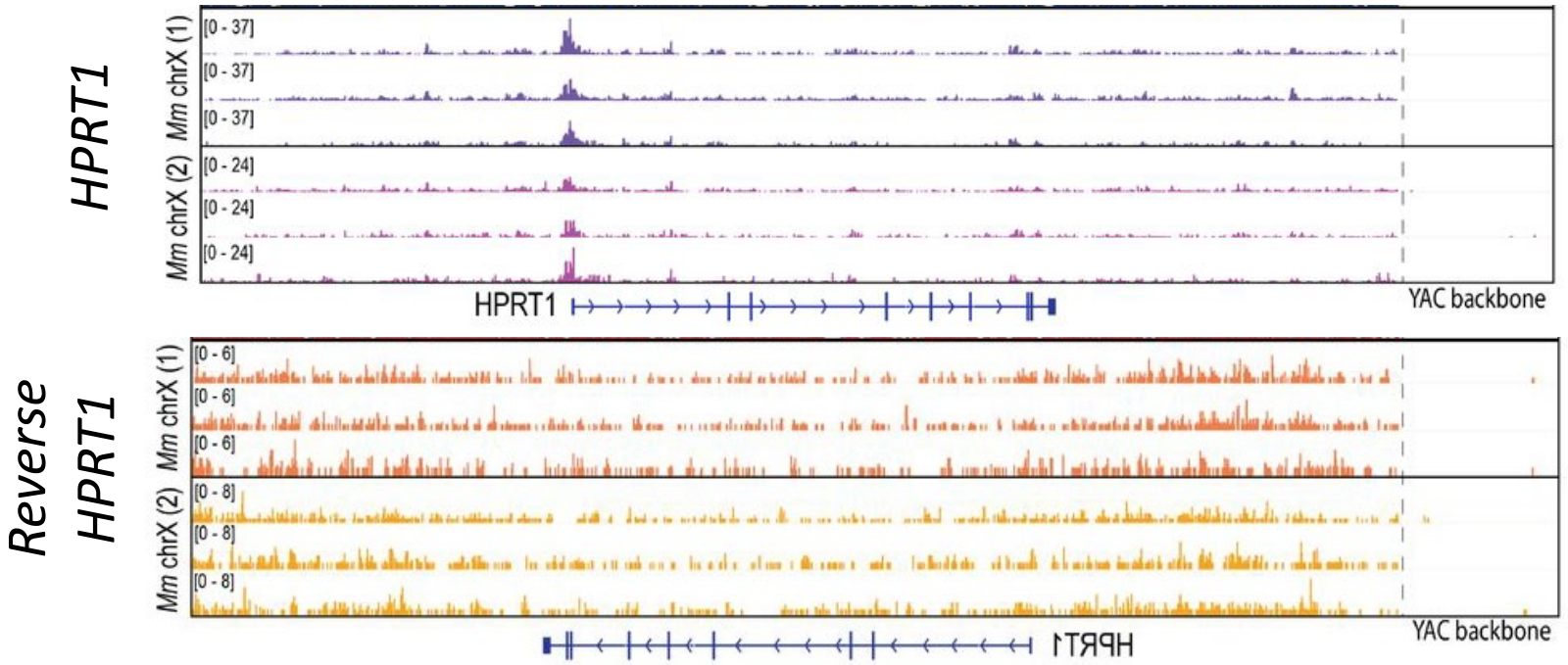


RNA-seq signal in sense but not reversed *HPRT1* locus



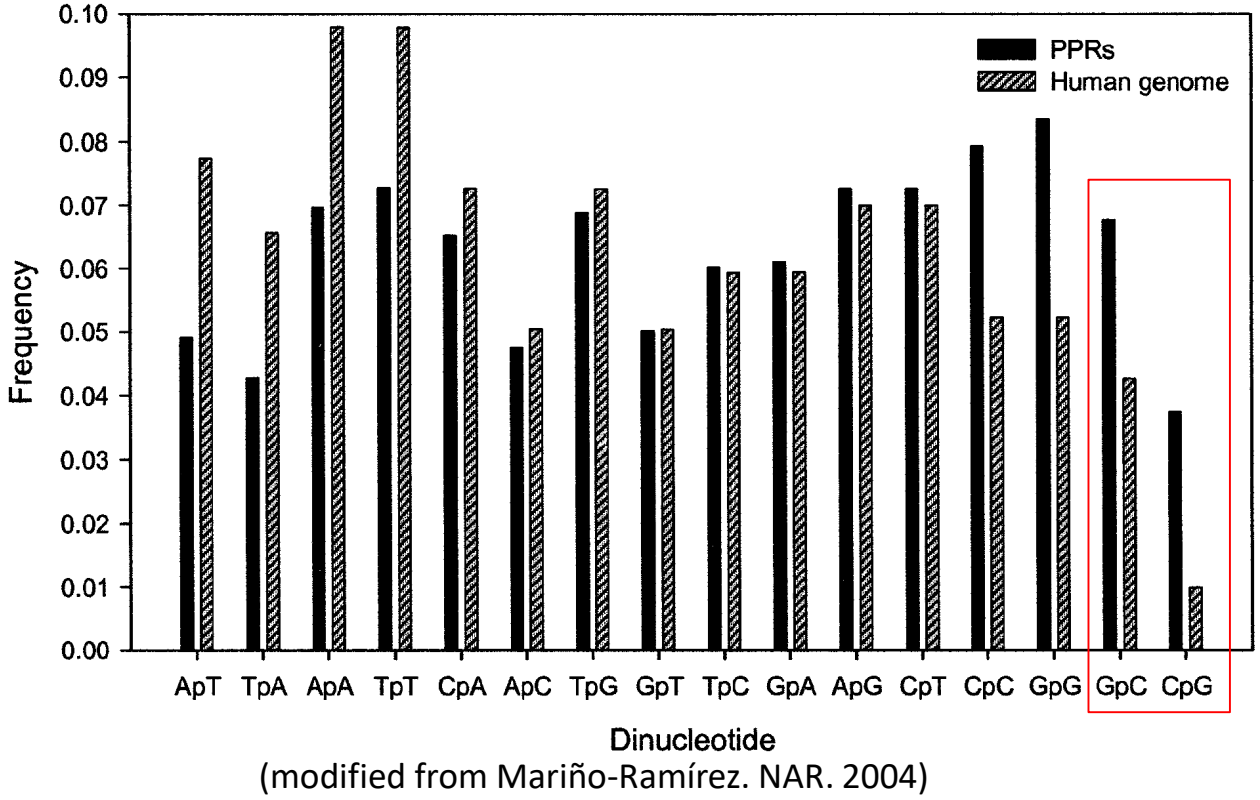
(Camellato, Brosh, Maurano, and Boeke. BioRxiv. 2022)

ATAC-seq signal in sense but not reversed *HPRT1* locus



(Camellato, Brosh, Maurano, and Boeke. BioRxiv. 2022)

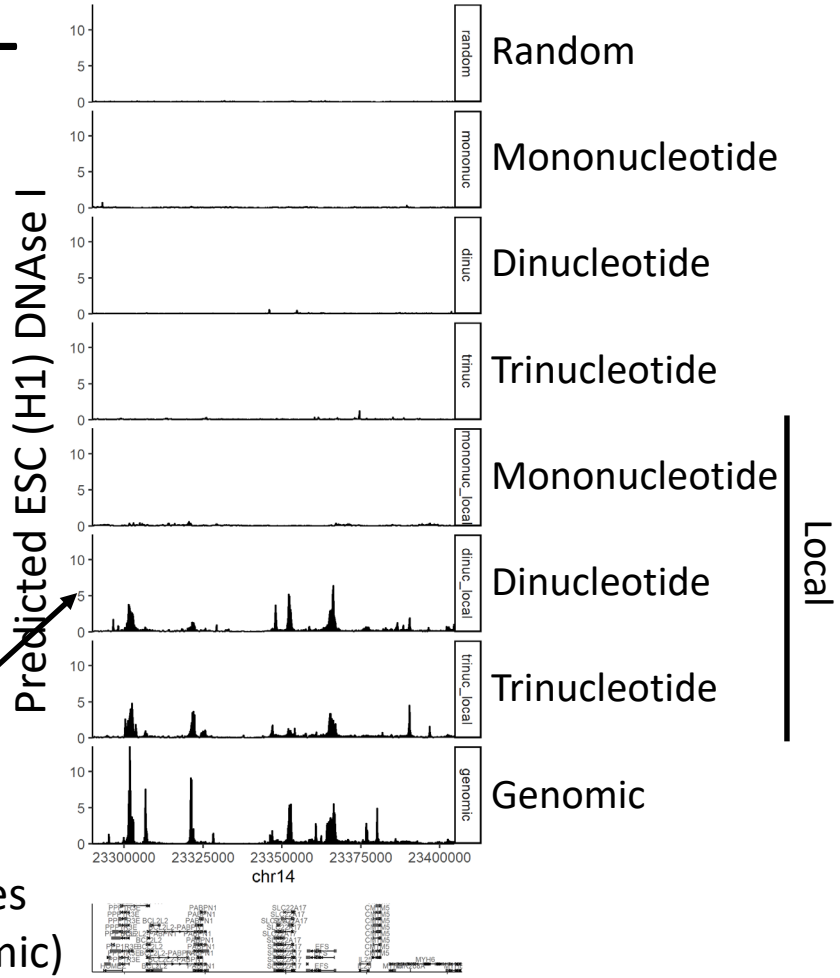
The human genome is non-random even at the dinucleotide level



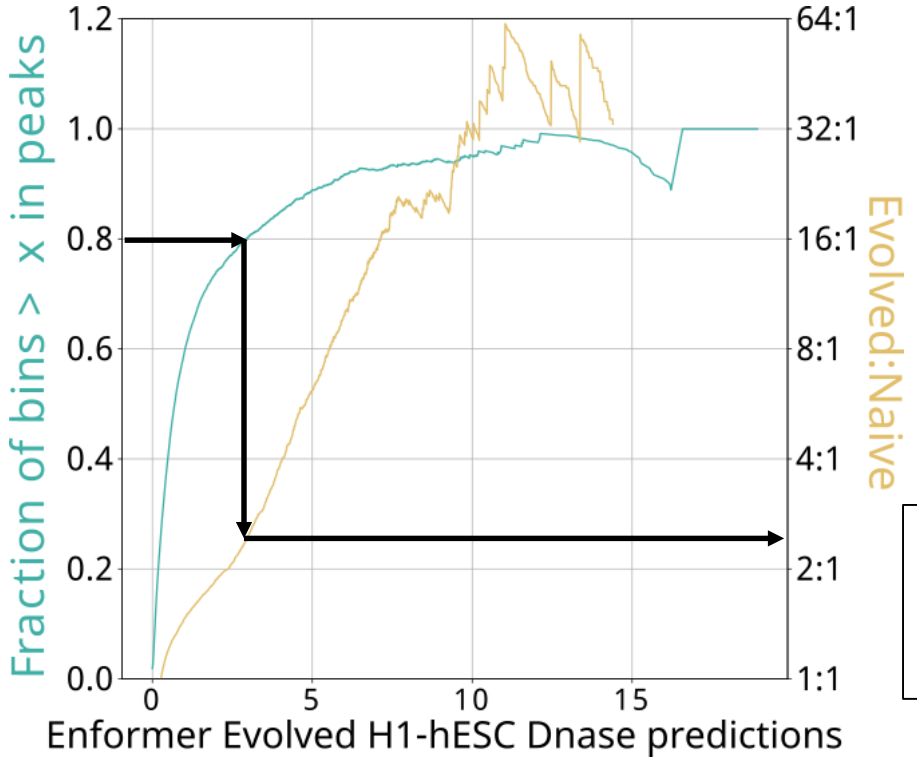
(modified from Mariño-Ramírez. NAR. 2004)

Matching local di/tri-nucleotide content “rescues” activity distribution

Use local dinucleotide shuffling as “naïve”:
- Scrambles TF binding sites
- Captures mutational bias of genome

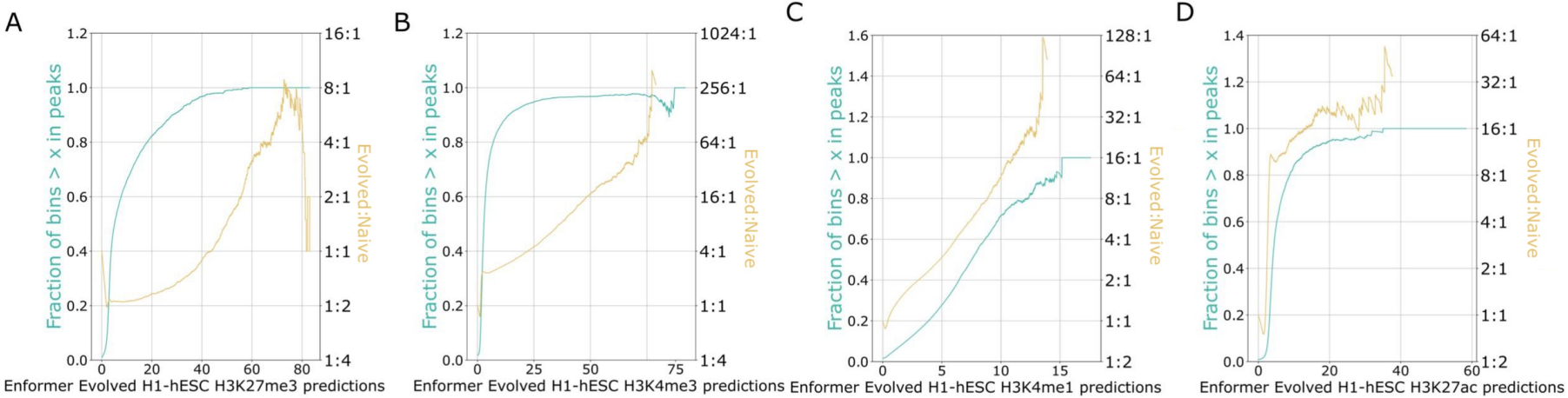


Selection required for extremes of DNase I

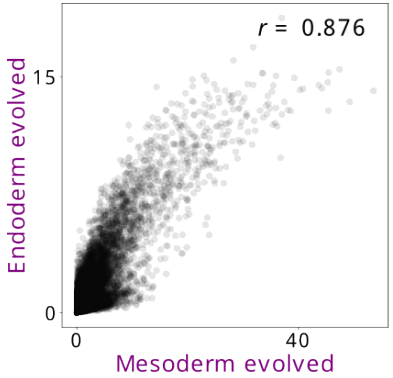


~2.5:1
Evolved:Naïve @
80% in peaks

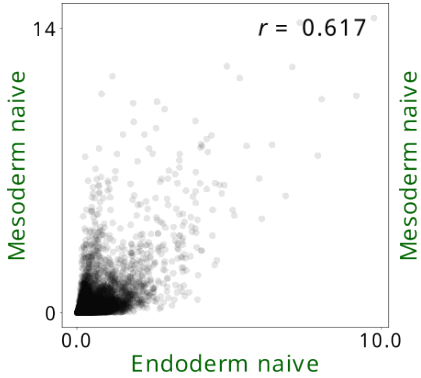
Selection required for extremes of chromatin marks

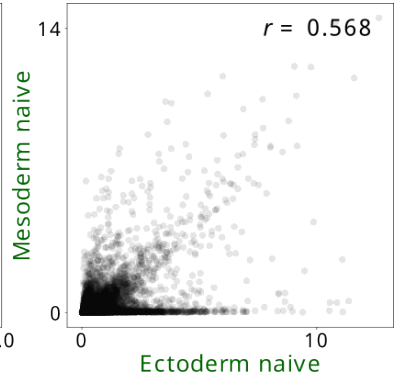


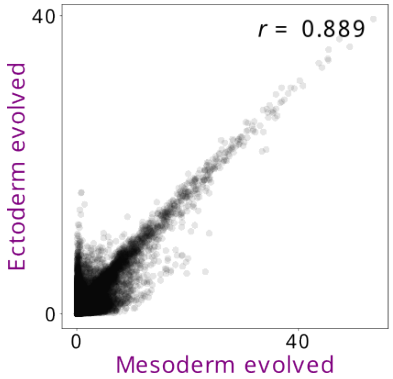
Naïve DNA is *more* cell type specific than evolved (DNase I)

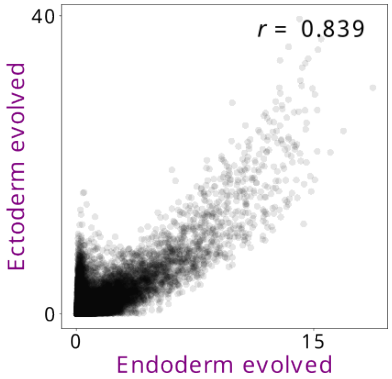


Evolved



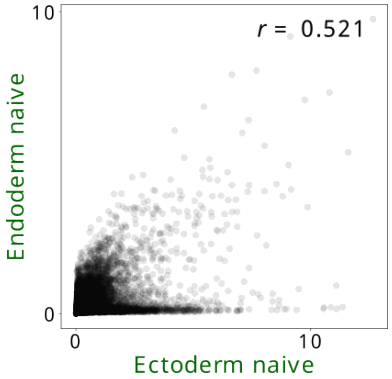




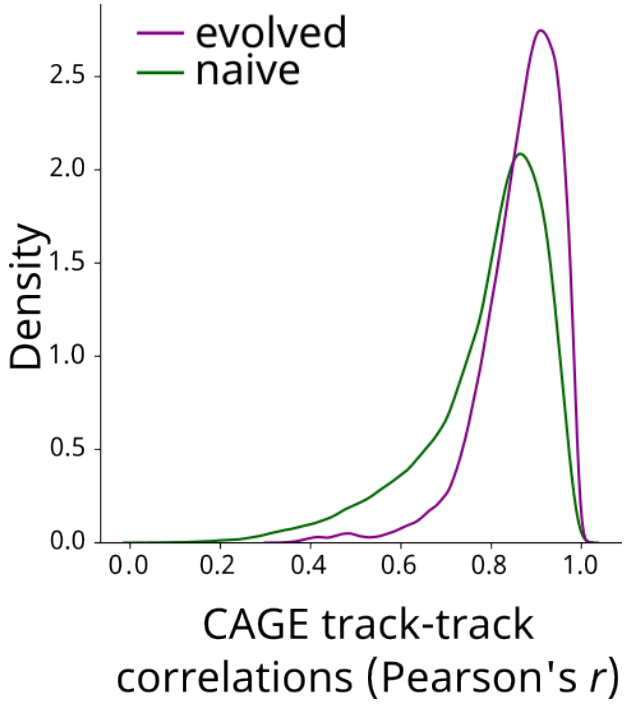


All axes: predicted DNaseI

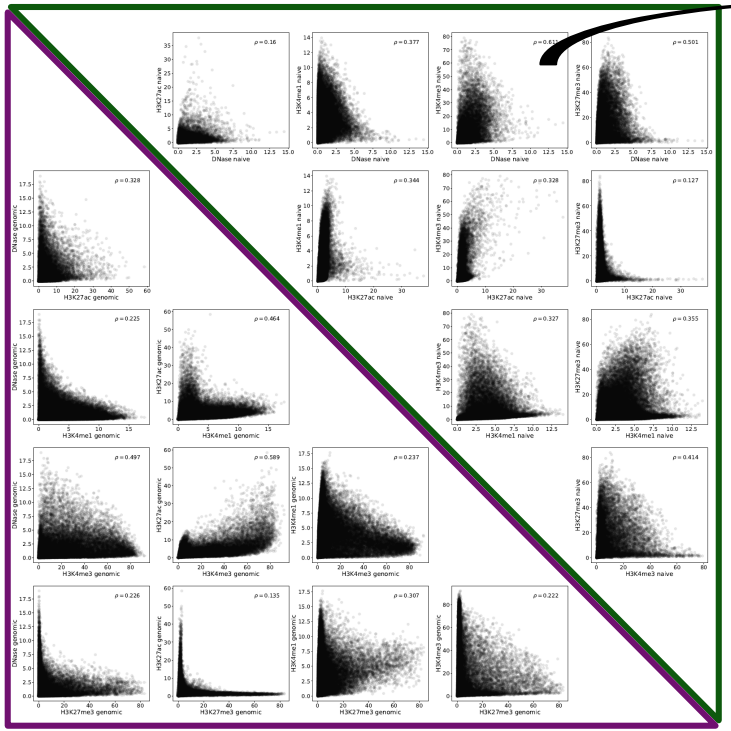
Naïve

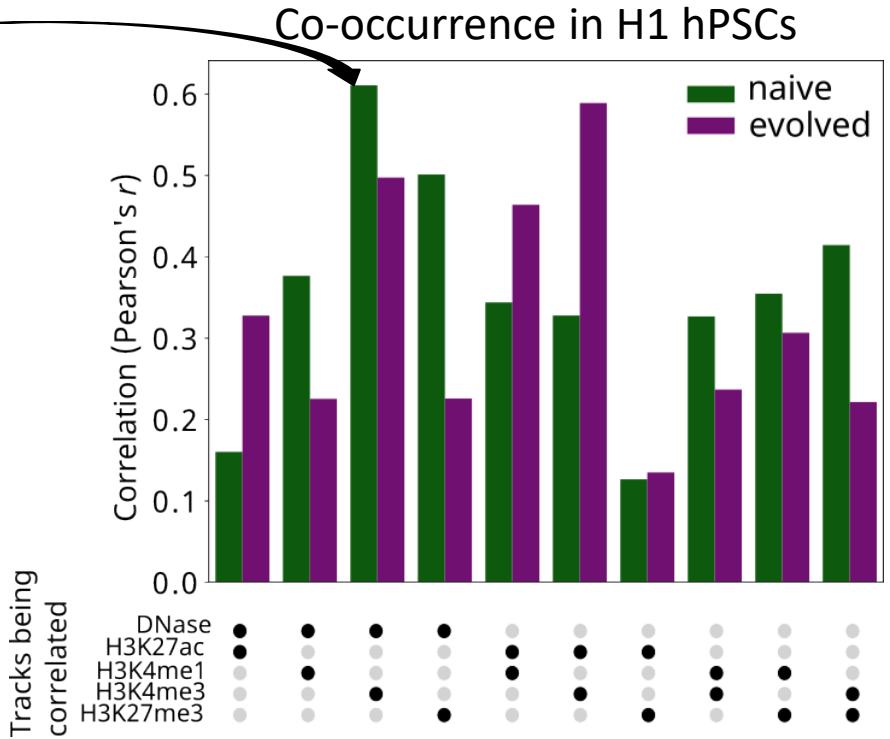


Naïve DNA is *more* cell type specific than evolved (CAGE)

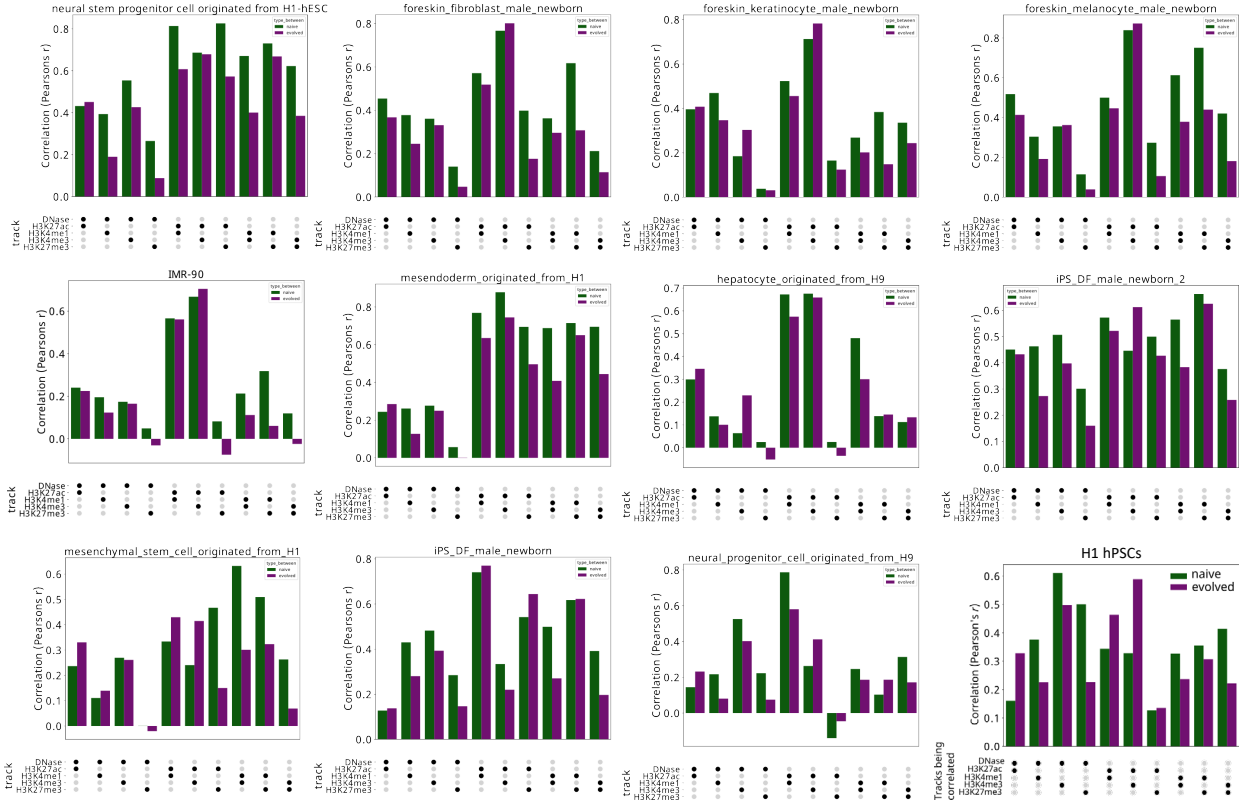


Active chromatin marks co-occur similarly in naïve and evolved DNA



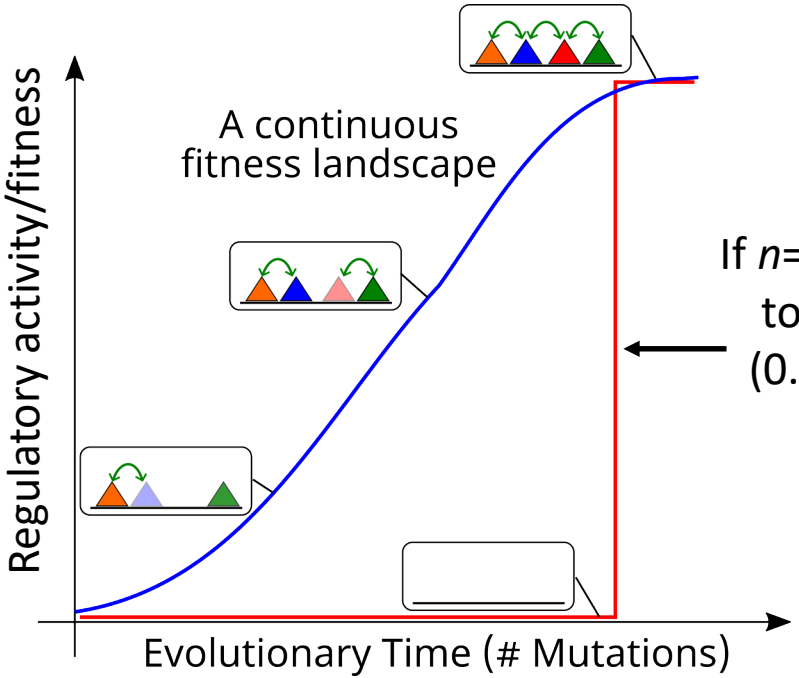


Active chromatin marks co-occur similarly in naïve DNA, across cell types



Regulatory activity *must* occur frequently by chance

But, regulatory regions evolve rapidly!



If *n*=2, need **3E15 people** to find this genotype (0.5M years at today's population size)

Conclusions: A functional genomics null hypothesis

- Naïve DNA is biochemically active in yeast and humans, and looks a lot like evolved DNA
- Biochemical activity, cell type specificity, and co-occurrence are unreliable markers of function
- Extreme activity is a marker of function
- Regulatory activity must occur often by chance in humans, or evolving new regulatory regions would be highly improbable

Acknowledgements

- **de Boer Lab (UBC)**
  - **Cassandra Jensen**
  - **Ishika Luthra**
  - **Emilia Chen**
  - Najmeh Nikpour
  - Abdul Muntakim Rafi
  - Asfar Lathif Salaudeen
- **Aviv Regev (Broad/Genentech)**
  - **Eeshit Dhaval Vaishnav**
  - Esteban Luis Abeyta
  - Moran Yassour (Hebrew U of J)
  - Jenna Pfiffner
  - Dawn Thompson
- Joshua Levin (Broad)
  - Lin Fan
  - Xian Adiconis
- Broad Sorting facility
  - Patricia Rogers
- Eran Segal (Weizmann)
- **Francisco Cubillos (U Santiago de Chile)**
  - **Jennifer Molinet**
- DREAM Challenge
  - Abdul Muntakim Rafi (de Boer)
  - Pablo Meyer (IBM Research)
  - Jake Albrecht (Sage Biosystems)
  - Paul Boutros (UCLA)
  - Julie Bletz (Sage Biosystems)
  - Payman Yadollahpour (Broad/Genentech)

DREAM Challenge

Funding

Primary locations