

Cracking the regulation code

Robin Dowell


CGTCTCTTTGCATTTCAGCAAAATATGGTTTTAGCAAAAAAATCTC
CTCATATGTACAAGAATCTTTAAGTGATATAATCTTTCAGCGGA
CGCAGTTTTGCCTTAAATGCCCCTCTGTTGTAGAAGTTAAAT
TTCTTCGCCTCTTTCTCTTGACAGCGTAAGGATCAATTCTTCGCT
CTGCCCTTTTTTCGAAAAATGAAATTAATTAATTCGATTTTG
TATAGTTGTTTAAATTTTGCTCTTATTAATCATCACTTCCATGC
AGTAAGTGACGGTTTCATCATTCGGAAAAAGTAAGATTTGTTTCAA

GGGTTGGCCGAGTGGGCTAAGGCGGCGAGACTTTAAGATATCTGTGGACGGTGTGCGCGGAGTTTGAAGTCTGCATCTCGCATCTTCAGTATTTTTTGTATGATTTCACGTACTATTA
ACTAGAATAATAGGGAATGCAATTTGCGAGTTTGAGAGAAGAAATTTAATAAGCAAAAAAGCGGCTCCTAAGTAAAGTTTTCCTATGGCGCTCTTTGCAATCAGCAATA
ATGGTTTTAGCAAAAAAATCTCTTTCACTACACAAGTTGGAGCATATTTTTTAACAATAACAGATAAACGAAAGGCGATAATAAGTATTAAGTCTCCTATGTACAAGAACTTTTA
AGTGATATAATCTTTGAGCGGACTAGAAGAATAAGGCGTTTCAACTGTTGTCTTGCAGACGTAAACTCCTTCAAGTAGTCATTATTTTGTCTCGCAGTTTGTGCTTAAATGCG
CCTCTGTTGTAGAAGTTAAAAATTTGTTCTATTTTCTATTATTTGCCACCCACTGTCTACCTTCTGTCATCTTTATATTGAGCTTCTCTGTTTCTCGCTCTTTCTCTTGCA
GCGTAAGGATCAATTTCTTGCGTCTATAAGTGGTTCCGCGAGTAGTCTTTTTATTATTACCAGCACTGCTGATCGGTCTCGTTCTTCATCTCTGCTGCTCTTTTTTCGCAAAATG
AAATTTATTTATTTCCGATTTTGGCAGTTTCCGATTAAGTTTCTTGCTTCAAAAATTCCTTACACTTATTTTTTCCATTAATAGAGCCTTCTCTATAGTTTGTGTTCT
TATTATTCATCACTTCCATGCTTCTAGAAACCTCAAAAATG
GGAAAAGTAAGATTTGTTTCAAGTAAGCTTTGGGGTGCCGT
ATATCCAAGTTTTTCAAACCTACCCAAAGGATATTTAACTTG
CAATATGGAATAGTGTAAACCTTCACTTATCAACAACAAAA
AATTTTATAAAAACTATTTATGGACTAAAAACCTACAGCGC
TTAAAGTGGTTTTCTACCGTGTCTAAACGAGTTTTCAATTGCC
ATCTTCCGTACCGGGAGGTAACTTTCTGGCTGCATTACTAC
AGTAGCTCATCAAGTTGATATGGTGTGCAATAAATAGTTCC
TAGGAGATATGTTGTGATATGTGAGCAGTATGATTCTGTTCCG
TTCTAATATAGATGAATTTTCAACGTATTGAGCGCTTAGAG
AGCAAAACGATCATAATAGTTGAAAGCTTTCTTTCTCCTTT
TAGATTTACCTGGTGATACATAAACGAATGTCAATTTGCGCCA
ACTGTAGCCAGTATCTTTGGCAAAAGAAATTTATTTCTAGAA
TTTTGAAAGTCATCAATTTCCCGATATCCGTTCCACATAGC
CCAGCTTTAGAAATGACCCTAATATTTTCAATTGAAGATTAT
TATTCATGATTGCGGACACCGCACCTTTTTTTAGCATGTATT
GAATCATCCAATGAAAATCTGCTAAACTTCAATTGCAGTGC
TGATACCGGCAATTTCCAATGTTTCCGATGCAAAAGGAGATAT
ACTTCACTGTTGTAAGCATCCCATATTCTGACTGAACCAAT
TTAAAAACGATTCACTTTGAGCGATTGTCTATCATCCCCAGC
GAAAATGGAGGCTTGAACATAAAAGAGCCTGCTGGGTATA
GGAGAAGCCTTCGGCAAGGGTAGAAAAATTTATAGTGGAGC
CACAGAATACATTGGTGTGCGCCTAAGTCAATCATGGTTA
TGGCAAGGCCAGCTTACTTTAAAGATAGCGGGTGTCTGT
GTTTACCGCTATTGACATCCCAAAATACCAAGGAATTAAT
TTCTATGTTTTGTGATAGATCAACGCGCCGGAGCATACGGT
GATATAAGTATCGTTCTCTATGTCTTAGGATTCCATTGGAT
ACATTTGATTTTCTATCCACATCATAGATCAATATGGATCCA
AACAGTAGTTAGAATCTGTTTCGAGTGTACTGAAAGAACTA
CGATTTTTTAACGTAATACGACCTCTATTTGCTTTTGCC
GATTCGTCATTTATGCGCGTATGTGCAATATTTTCTTAGTAT
TTCAAATGCTGCTTTTCTTAAACATTTATAAAATTTTTGT
ACTGAGCTGTTTCTTAAATGCTTCTTAAATATGTAACAG
CAAAGATTTATTAATTTATCCAGACTCTGAGGATTCGGT
CGGCAGACTTAAGATCTGTTGGACGGTTGTCTCGCGCGAGTT
ATTGCAGTTTGAGAAGAAGAATTATTAATAAGCCAAAAAGCGGCTCACTAAGTAAAGTTTTCTTCAAGGCTCCTTTGCAATTCAGCAATAATCTCT
TTTCACTACACAAGTTGGAGCATATTTTTTAAACAATAACAGATAAACGAAAGGCGATAATAAGTATACTTCTATGTACAAGAACTTTAAGTGATATAATCTTTGAGCGGA
CTAGAAGAATAAGGCGTTCAACTGTTGTCTTGCAGACGTAACCTCCTCAAGTAGTCAATTATTTGCTCGCAGTTTGTCTTAAATAGCCCTCTGTTGTAGAAGTTAAAT
TTGTTCTATTTTTCTATTATTTGCCACCCACTCTGTACCTTCTGTCATCTTATATTGAGCTTCTGTTTCTCGCTCTTTCTCTTGCAGCGTAAGGATCAATTTCTTGCGT
CTATAAGTGGTTCCGCGAGTAGTCTTTTTATTATTACCAGCATGCTGATCGGTCTGTTCTTCACTCTCTGCTGCTTCTTTTGCAAAATGAAATATATTATTTCCGATTTTG
CCAGTTCCGATATAGTTCTTGCTTCAAAAATCTTCACTACTATTTTTTCCATTAATAGAGCCTTCTCTATAGATTGTTTAAATTTTTGTCTTATTATCATCACTTCCATGCG
CTTAGAAACCTCAAAAATGAATCTATCAAGCTTGGATCAACACATCTTTCTCGCGTAACTGTGATTGTAGTAGAAGTGACGGTTTCATCATTTCGAAAAGTAAGATTTGTTTCAA

Your life depends on the secrets of your DNA. Are you ready?

THE LANGUAGE OF LIFE

"This book sets out hope without hype, and will enrich the mind and uplift the heart."
—JEROME GROOPMAN



DNA AND THE REVOLUTION IN PERSONALIZED MEDICINE

Francis S. Collins

Author of the New York Times Bestseller
The Language of God

Microhabitats save mammals, but not birds, from warming pp. 553 & 633

Gut microbiota modulate immunotherapy pp. 573, 595, & 602

Physically distanced quantum gates pp. 576 & 634

Science

\$15
5 FEBRUARY 2021
sciencemag.org

AAAS

SPECIAL ISSUE

HUMAN GENOME AT



The completion of the first survey of the entire human genome was announced 20 years ago on June 26, 2000.

DNA is a language encoding biology



400,000 words

26 letters

3 Morse Code Symbols
(dash, dot, spaces)



Over 100,000 proteins

20 amino acids

4 nucleotides
(A, T, C, G)

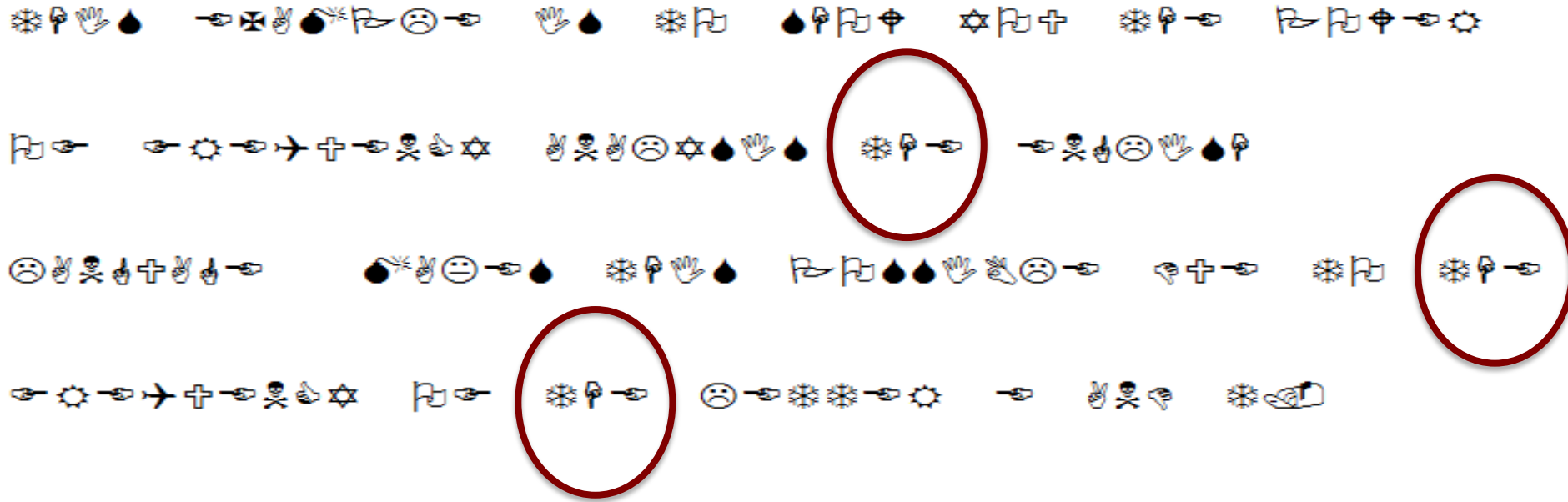
Part I: The written language

So how do we crack a new language?



“Data don’t make any sense,
we will have to resort to statistics.”

Therefore we look for commonly used patterns
(letters or words)

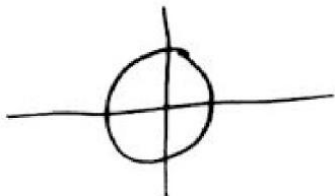


Therefore we look for commonly used patterns
(letters or words)

THIS EXAMPLE IS TO SHOW YOU THE POWER
OF FREQUENCY ANALYSIS THE ENGLISH
LANGUAGE MAKES THIS POSSIBLE DUE TO THE
FREQUENCY OF THE LETTER E AND T.

The message fully decoded.

H E R > 9 J A V P X I L T G Q
 N 9 + B φ ■ O ■ D W Y · < ■ K 7 ⊖
 B Y Ǝ C M + u z G W φ ⊖ L ■ ⊕ H J
 S 9 9 Δ A J ▲ ▣ V O 9 O + + R K ⊖
 □ Δ M + ⊕ ⊥ τ O I ● F P + P ⊖ X /
 9 ▲ R A F J O - ■ O C 2 F > ⊖ D φ
 ■ ● + K ⊖ ■ Ǝ ⊖ u c X G V · ⊕ L I
 φ G ⊖ J 7 τ ■ O + □ N Y ⊕ + □ L Δ
 O < M + 8 + Z R ⊖ F B c Y A ⊖ ● K
 - ⊕ J u v + A J + O 9 Δ < F B Y -
 u + R / ● ⊥ E I D Y B 9 8 T M K O
 ⊖ < c J R J I ■ ● T ● M · + P B F
 ⊕ ⊖ Δ S Y ■ + N I ● F B c φ Ǝ ▲ R
 J G F N A 7 ● ● ● 8 · c V ● ⊥ + +
 Y B X ● ■ Ǝ ● Δ C E > V u z ● - +
 I c · ⊖ ⊕ B K φ O 9 A · 7 M ⊖ G ●
 R c T + L ● ● c < + F J W B I ⊖ L
 + + ⊖ W c ⊕ W c P O S H T / φ ⊕ 9
 I F X O W < Δ ⊥ B D Y O B ■ - c c
 > M D H N 9 X S ⊕ Z O ▲ A I K Ǝ +



On November 8, 1969, Zodiac mailed the 340 cipher.

“When I first started looking at the
 Zodiac ciphers all those years ago
 (2006), I thought, ‘Oh, I can just write
 a computer program and solve it,’ but
 it’s been kicking my ass all this time.
 Until now.”

--- David Oranchak, December 2020,
 upon solving Zodiac 340 cipher

First letter

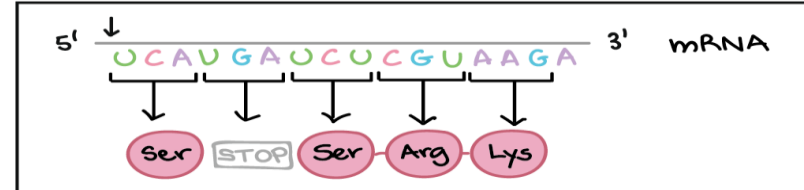
Second letter		A		G	
U	C	U	C	U	C
UUU } Phe	UCU } Ser	UAU } Tyr	UGU } Cys	U	C
UUC }	UCC }	UAC }	UGC }	U	C
UUA } Leu	UCA }	UAA } Stop	UGA } Stop	A	A
UUG }	UCG }	UAG } Stop	UGG } Trp	G	G
C	U	C	U	C	U
CUU } Leu	CCU } Pro	CAU } His	CGU } Arg	U	C
CUC }	CCC }	CAC }	CGC }	C	A
CUA }	CCA }	CAA } Gin	CGA }	A	G
CUG }	CCG }	CAG }	CGG }	G	G
A	C	A	C	A	C
AUU } Ile	ACU } Thr	AAU } Asn	AGU } Ser	U	C
AUC }	ACC }	AAC }	AGC }	C	A
AUA } Met	ACA }	AAA } Lys	AGA } Arg	A	G
AUG }	ACG }	AAG }	AGG }	G	G
G	U	G	U	G	U
GUU } Val	GCU } Ala	GAU } Asp	GGU } Gly	U	C
GUC }	GCC }	GAC }	GGC }	C	A
GUA }	GCA }	GAA } Glu	GGA }	A	G
GUG }	GCG }	GAG }	GGG }	G	G

Third letter

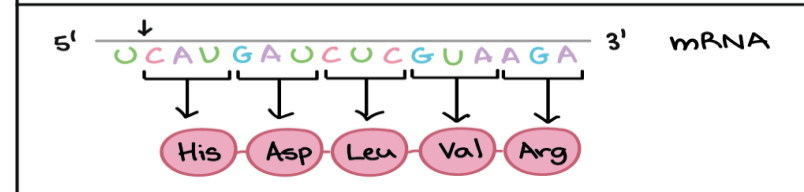
Proteins are encoded in DNA as a simple cipher



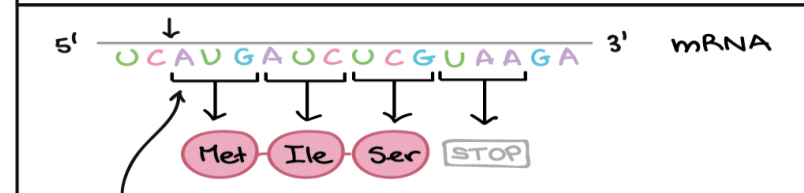
FRAME 1



FRAME 2

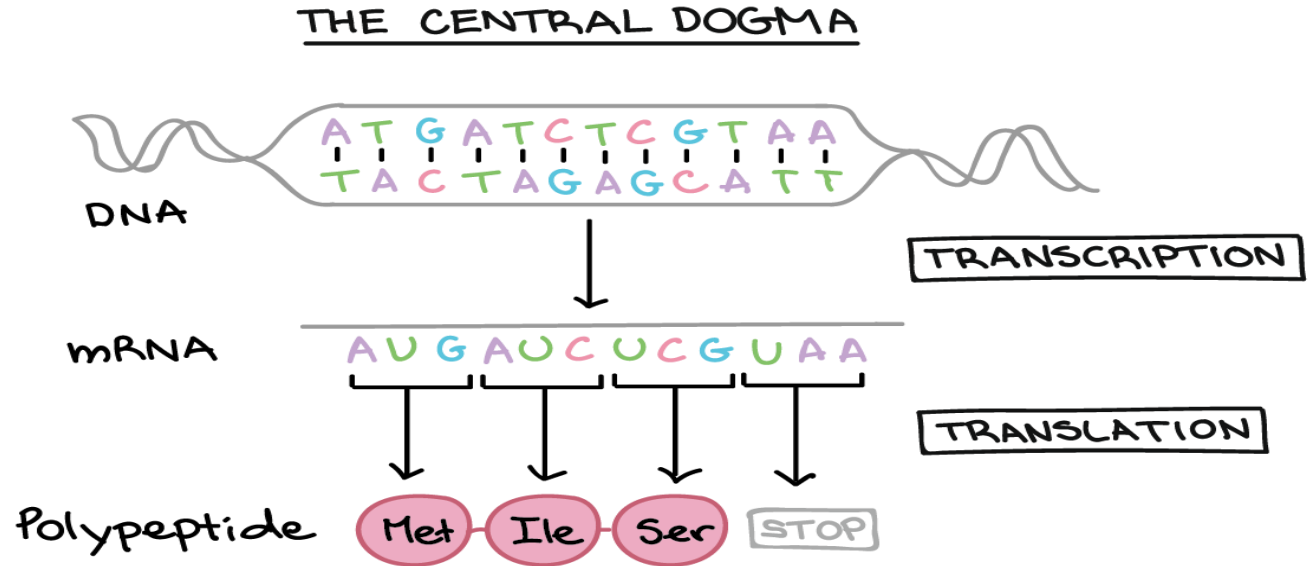


FRAME 3

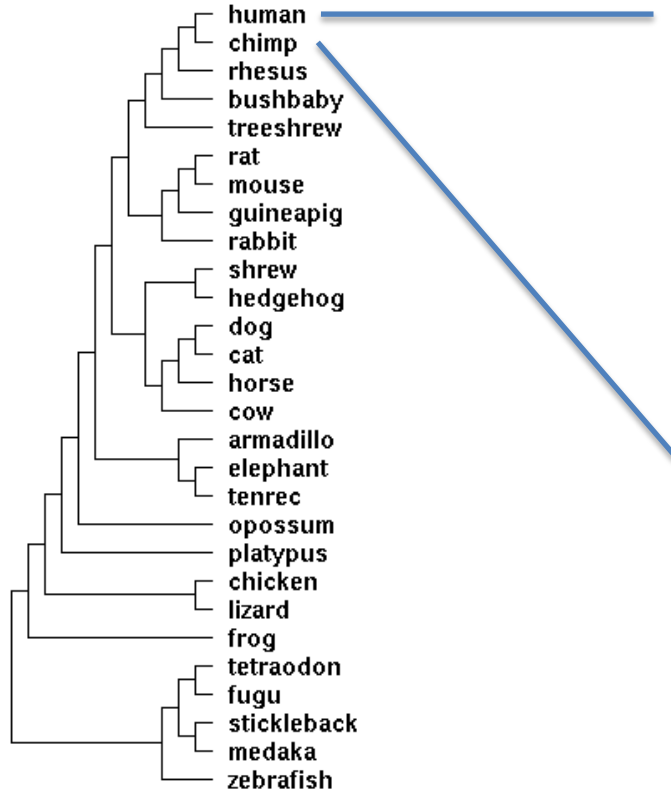


Start codon's position ensures that this frame is chosen

The central dogma describes the biochemistry behind this encoding



Evolution helps filter signal from noise



Can you spot the differences?

SNP

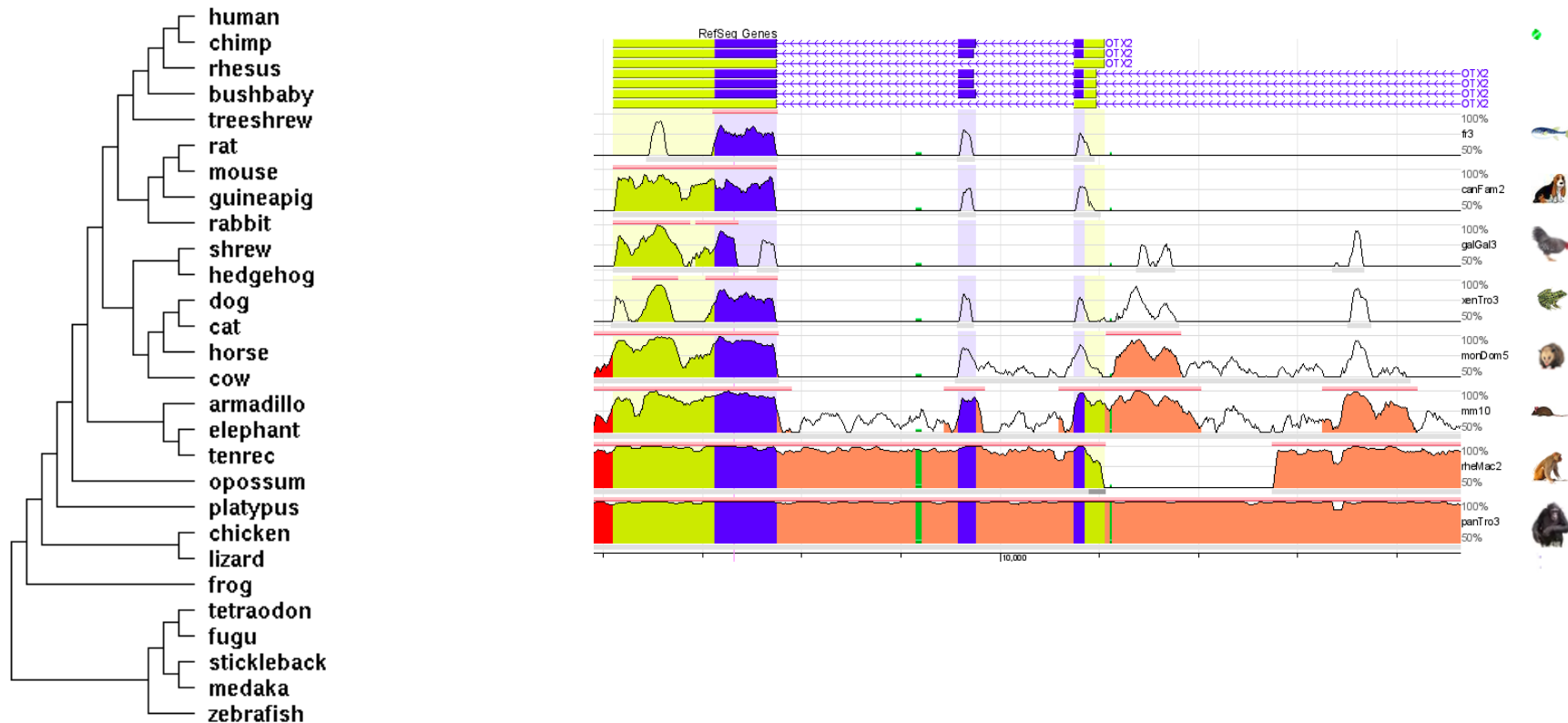


SNP

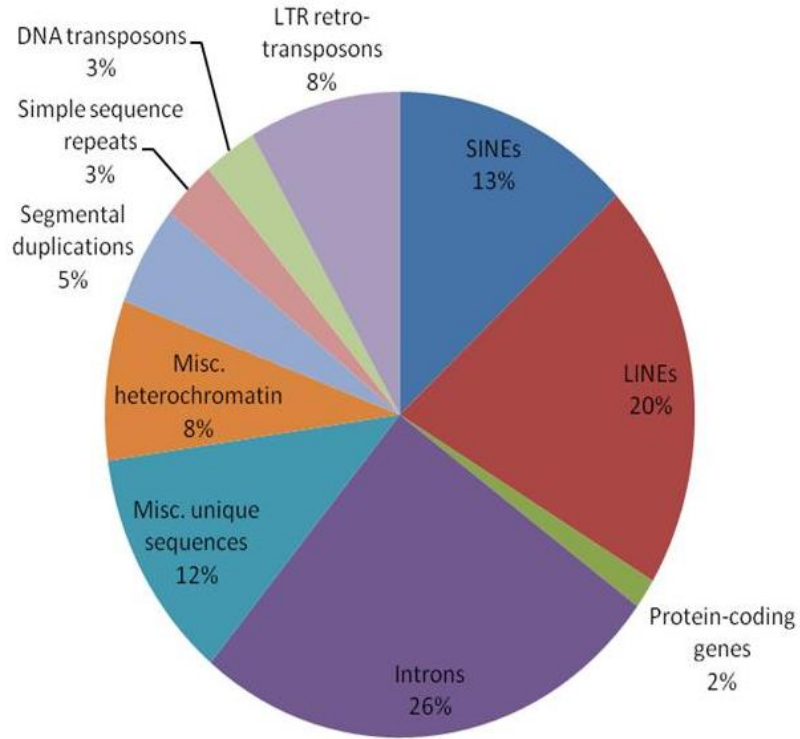


..... TTCG G GGT C.....	AGTC G ACCG.....
..... TTCG A GGT C.....	AGTC A ACCG.....
..... TTCG G GGT C.....	AGTC G ACCG.....
..... TTCG G GGT C.....	AGTC G ACCG.....

Evolution helps filter signal from noise



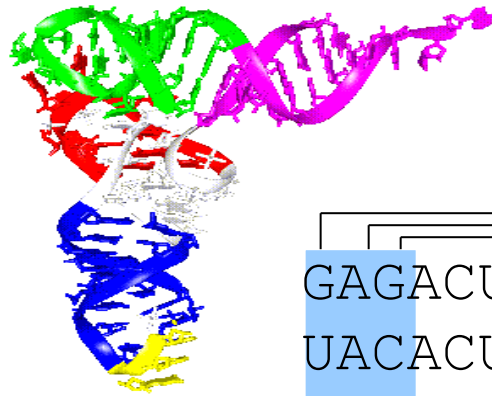
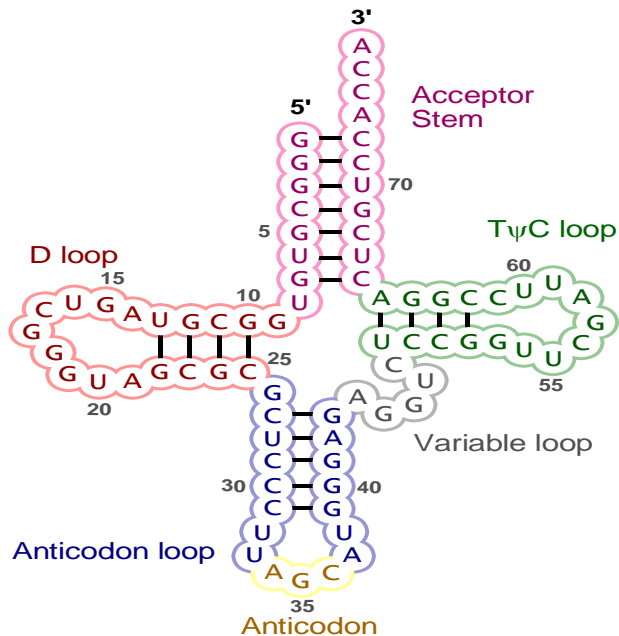
Genes are only a small fraction of the genome.



What else is encoded in the genome?

We can decipher from sequences

Structural non-coding RNAs also have inherent sequence patterns.



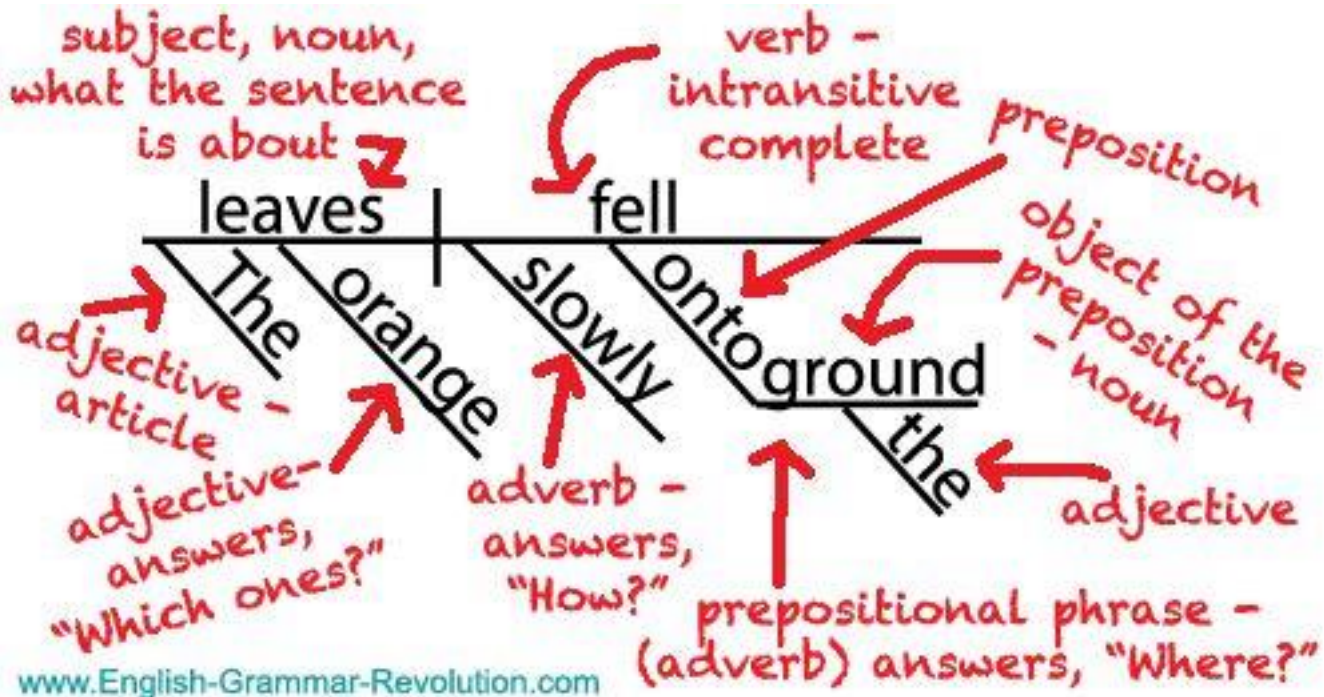
Sequence alignment showing conserved regions (indicated by vertical bars) across five lines of nucleotide sequence:

```

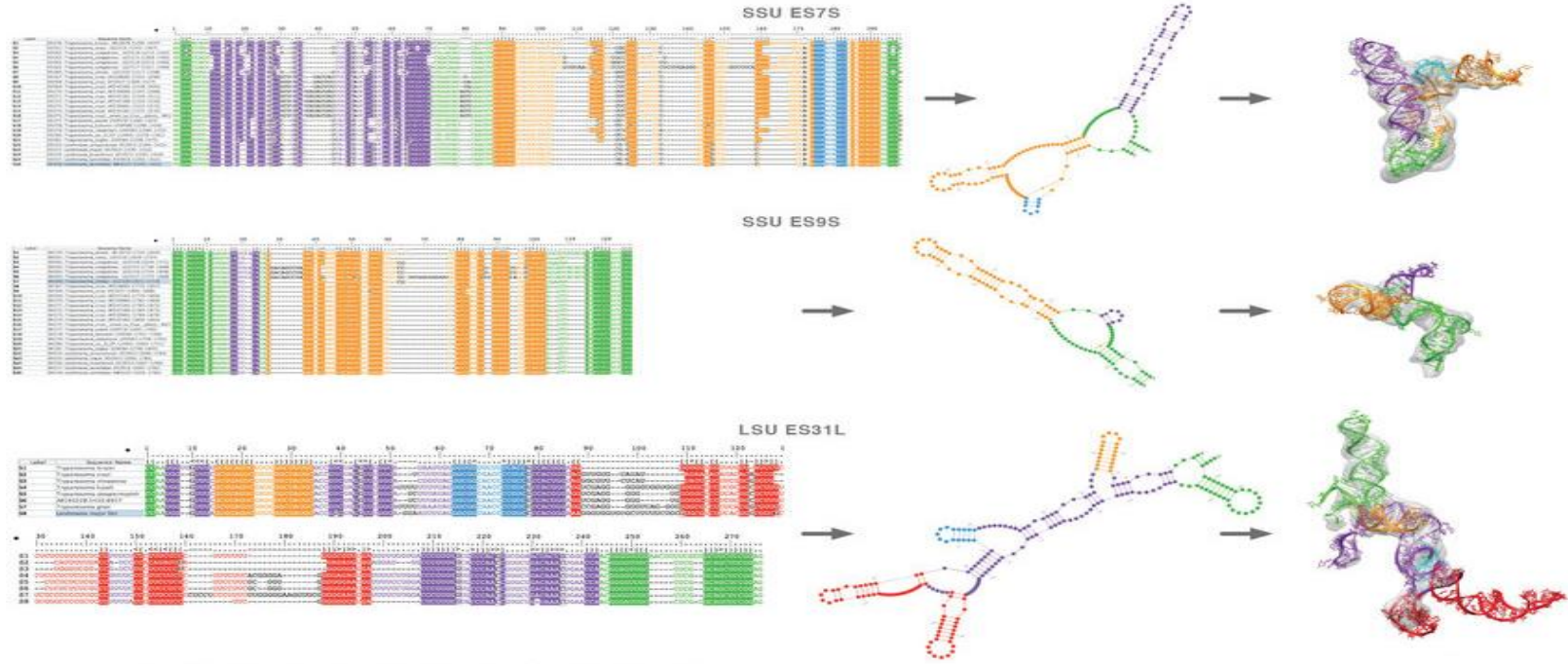
GAGACUUCGGAUCUGGCGACACCC
UACACUUCGGAUGACACCAAAGUG
AGUUCUUCGGCAAGGGCACCAUUC
CCAACUUCGGAUUUUGCUACCAUA
AUGCCUUCGGAGCCAGCGUAACUU
--<<<----->->>-<<-<----->>>
    
```

RNA secondary structure is a context-free grammar.

“The orange leaves fell slowly onto the ground.”



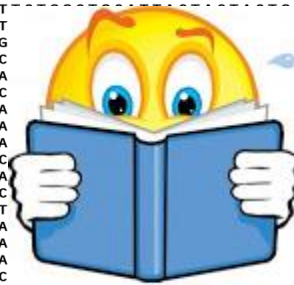
Uncover RNA structure through evolutionary alignments (no simple cipher)



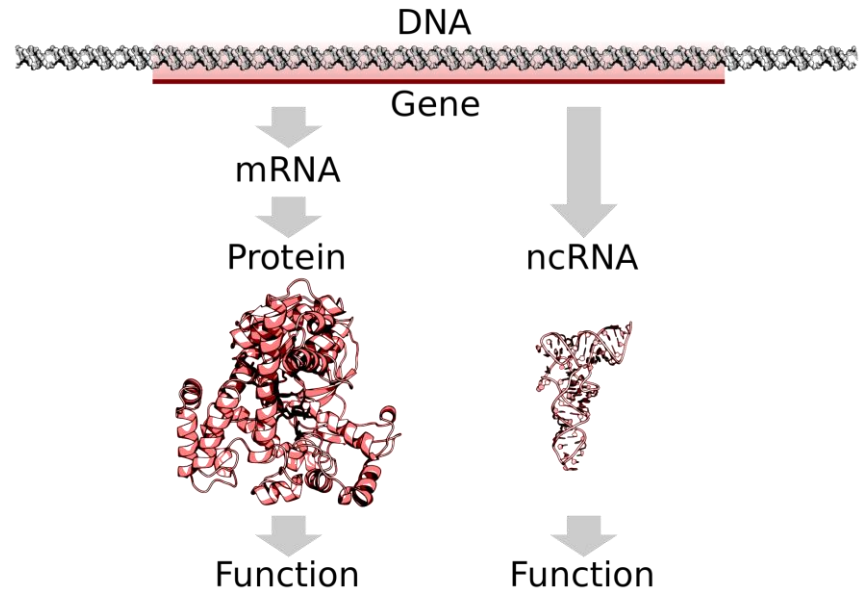
The written language of life ...

ACCCCTTACCAACAATGGAATCTCAAAGATTATTAATTTATTCACAGACTCTGAGGATTTCGGGTAAATAGGGTATTTAACTGGTTACCGGAAAGGTTTAGAAAAATTCGTGGA
GGGTTGGCCGAGTGGTCTAAGGCCGCGAGACTTAAGATCTGTTGGACGGTTGTCGCGCGGAGTTTCGAACCTCGCATCCTTCAGTATTTTTTTGATGATTTAACGTACTATTA
ACTAGAATAATAGGGAATGCAATTGCAGTTTGAGAAGAAGAATTATTAATAAGCCAAAAGCGGCTCACTAAGTAAGTTTTCTTCATGGCGTCTTTGCAATTCAGCAATA
ATGGTTTTAGCAAAAAATCTCTTTCACTACACAAGTTGGAGCATATTTTTAAACAAATACAGATAAACGAAAGGCGATAATAAGTATACTTCCTATGTACAAGAATCTTTA
AGTGATATAATCTTTTCAGCGGACTAGAAGAATAGGGCGTTCAACTGTTGTCCTTGCAGACGTAACCTCTCAAGTACTCATTATTTTCTCGGAGTTTTGCTCTAAAAATGC
CCTCTGTTGTAGAAGTTAAAAATTTGTTCTATTTTTCTATTATTTGCCACCCACTCTGTCACTTCGTCC,
GCGTAAC.....TGA
AAATTA.....TTT
TATTAT.....TCT
GGAAAA.....CCG
ATATCC.....AAA
CAATAT.....TCT
AATTTTATAAAACTATTTATGGACTAAAAACCTACAGCGCTCTTCATCATATCCCTTCCCGTTTCTTC
TTAAAGTGGTTTCTACCGTGTCTAAACGAGTTTCAATTGCCCTAGAAACACTTTTAAGGACACCATATC
ATCTTCCGTACCGGGAGGTAACCT
AGTAGCTCATCAAGTTGATATGGT
TAGGAGTATGTTGTGATATGTCA
TTCTAATATAGATGAATTTTCAAC
AGCAAAACGATCATAATAGTTGAA
TAGATTTACCTGGTGATACAAAC
CTGTAGCCAGTATCTTTGGCAAA
TTTTGAAAGTCATCAATTCGCCGA
CCAGCTTTAGAAATGACCTTAATA
TATTCATGATTGCGGACAGCGCAC
GAATCATCCAATGAAATCTGCTA
TGATACCGCAATTTCCAATGTTTC
ACTTCACGTGTGGAAGCATCCCAT
TTAAAAACGATTCACTTTGAGCGA
GAAAAATGGAGGCCCTTGGAACTAA
GGAGAAGCCTTCGGCAAGGGTAGA
CACAGAATACATTGGTGTGCCGCC
TGGCAAGGCCAGCTTACTTTAAAG
GTTTACCGCTATTGACATCCCAAA
TTCTATGTTTGTGGATAGATCAC
GATATAAGTATCGTTCCCTATGCT
ACATTTGATTTCTATCCACATCAT
AACAGTAGTTAGAATCTGTTTGA
CGATTTTTTAACGTAAATACGACC
GATTTCGTCATTTATGCCGTATGT
TTCAAATGCCGTCTTTCTTAAACATTTATAAAATTTTTGTATCTGTTCAATTGACAATTTTGTAACTT
ACTGAGCTGTTTCTTAAATGCTTCTTAAATAATGTAACAGAAATGCGCATTGTTGAACATACGGCTTCG
CAAAGATTATTAATTTATTCACAGACTCTGAGGATTCGGGTAAATAGGGTATTTAACTGGTTACCGGAAAGGTTTAGAAAAATTCGTGGA
CGGCAGACTTAAGATCTGTTGGACGGTTGTCGCGCGGAGTTTCGAACCTCGCATCCTTCAGTATTTTTTAACTGGTTACCGGAAAGGTTTAGAAAAATTCGTGGA
ATTTGCAGTTTGAAGAAGAATTTAATAAGCCAAAAGCGGCTCACTAAGTAAGTTTTCTTCATGGCGTCTTTGCAATTCAGCAATAATAGGTTTTAGCAAAAAATCTTC
TTTTCACTACACAAGTTGAGCATATTTTTTAAACAAATACAGATAAACGAAAGGCGATAATAAGTATACTTCCTATGTACAAGAATCTTTAAGTATATACTTCCTCAGCGGA
CTAGAAGAATAGGGCGTTCAACTGTTGTCCTTGCAGACGTAACCTCTTCAAGTAGTCAATATTTTGGCTCGCAGTTTTGCCCTTAAATGCCCCTCTGTTGTAGAAGTTAAAA
TTGTTCTATTTTTCTATTTTTCACCCCACTCTGTCACTCTTGGTCCATCTTATATAGACTTCTGTTTTCTCGCGTCTTTCTTCGACGTAAGGATCAATTCCTGGCT
CTATAAGTGGTTTCGCGAGTAGTCTTTTTATTTATTCAGCCCACTGTTCATCGGCTCTGTTCTTCACTCTGCTGCTCTTTTTCCGCAAAATGAAATATATTTCCGATTTTG
CCAGTTCCGATATAGTCTTGGCTTTCGTTCAAAATCTTCACTATTTTTCATTAATAGACGCTTTCCCTATAGTTGTTTAAATTTTGTCTTATTTATCATCACTTCCATGC
CTTAGAAACCTCAAAAATGAATCTATCAAGCTTGGATCAACACATCTTTCTCGCGTAAGTGTGATTGTAGTAAGTGACGGTTTCATCATTCCGAAAAAGTAAGATTTGTTTCA

Genome Sequence Patterns

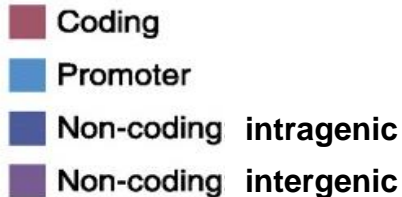
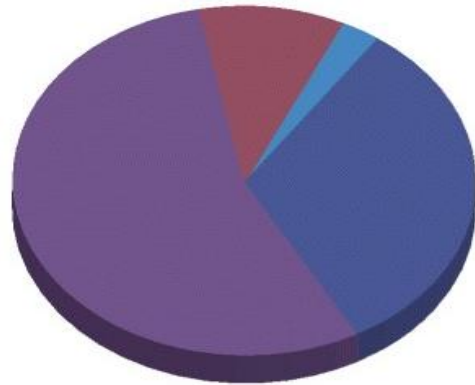


READ

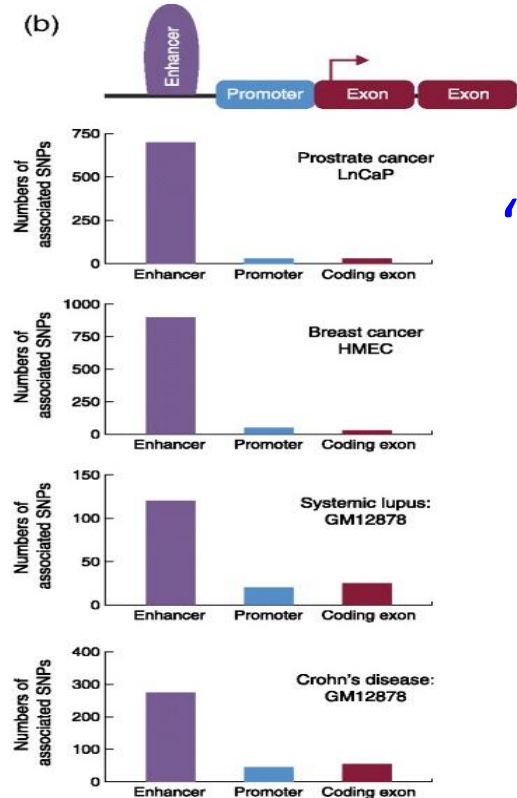


Most disease associated variations reside in *regulatory* regions

(a) Distribution of GWAS variants



(b)



‘Oh, I can just write a computer program and solve it.’

--- Robin Dowell
(2005)

We don't communicate solely through writing

“I don't think he should get that job.”

Somebody else thinks he should get the job.

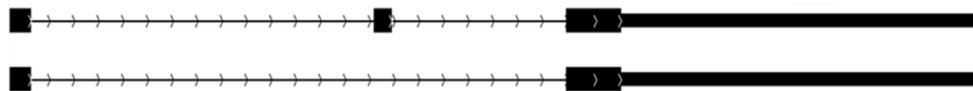
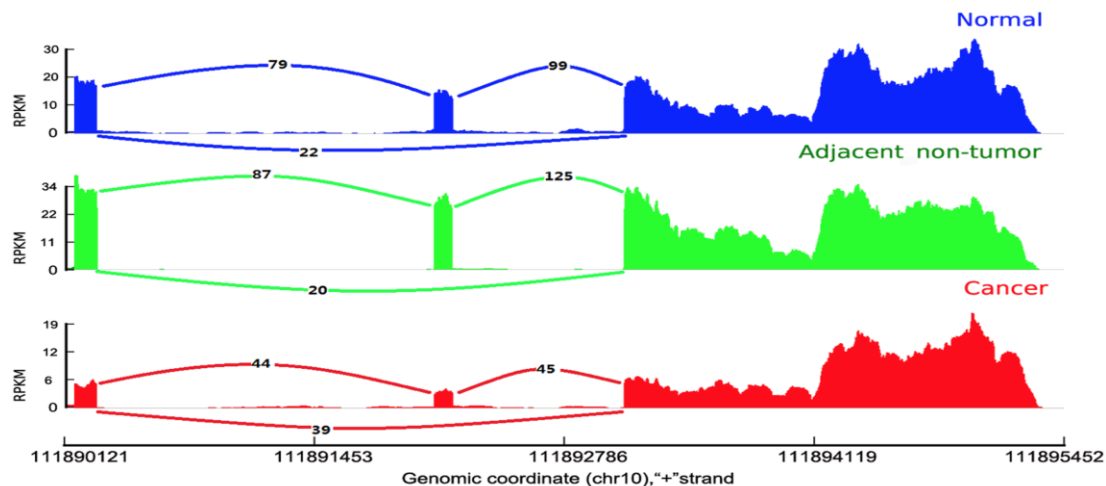
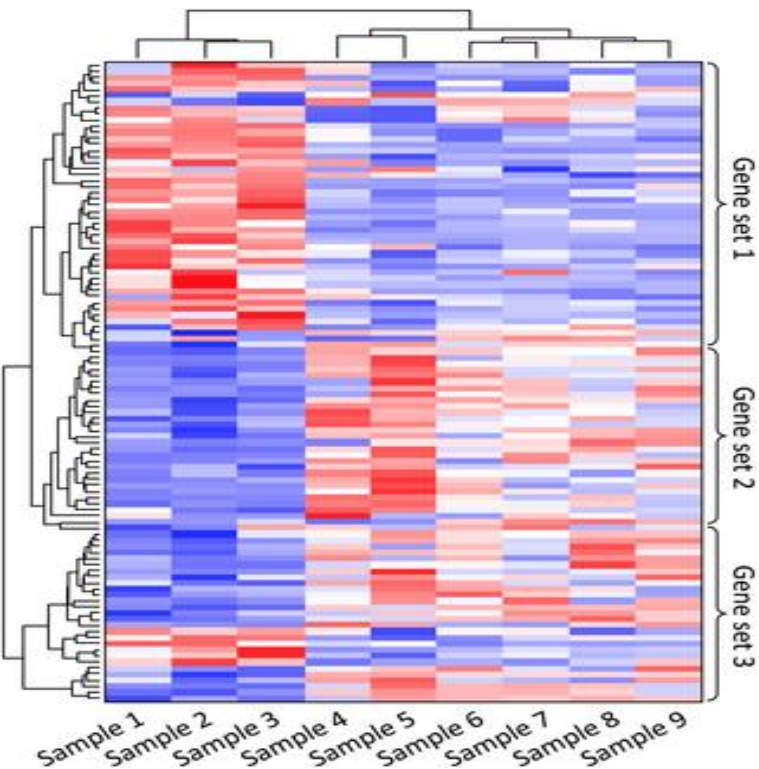
He should get another job.

I'm not sure he'll get that job.

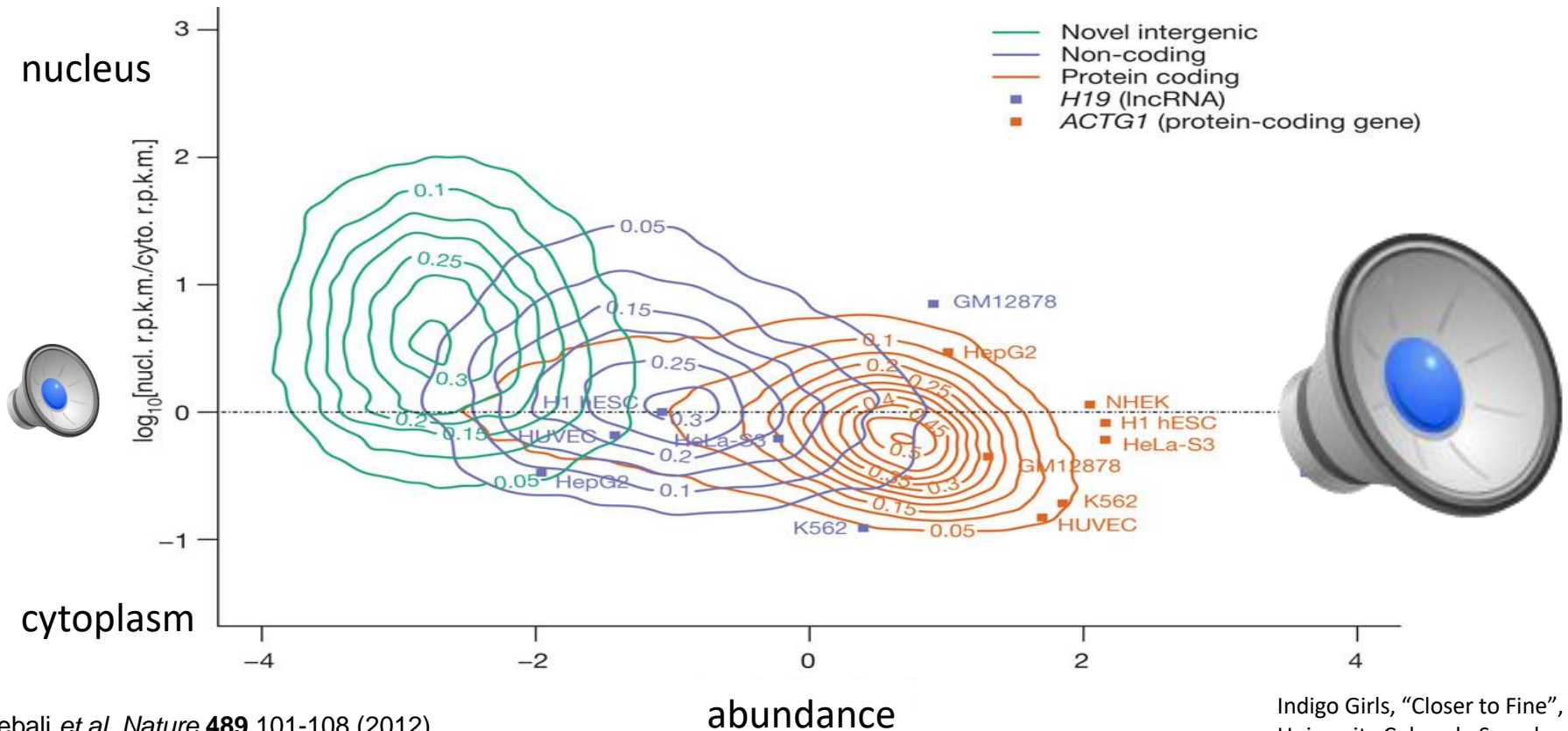
In my opinion it's wrong that he's going to get that job.



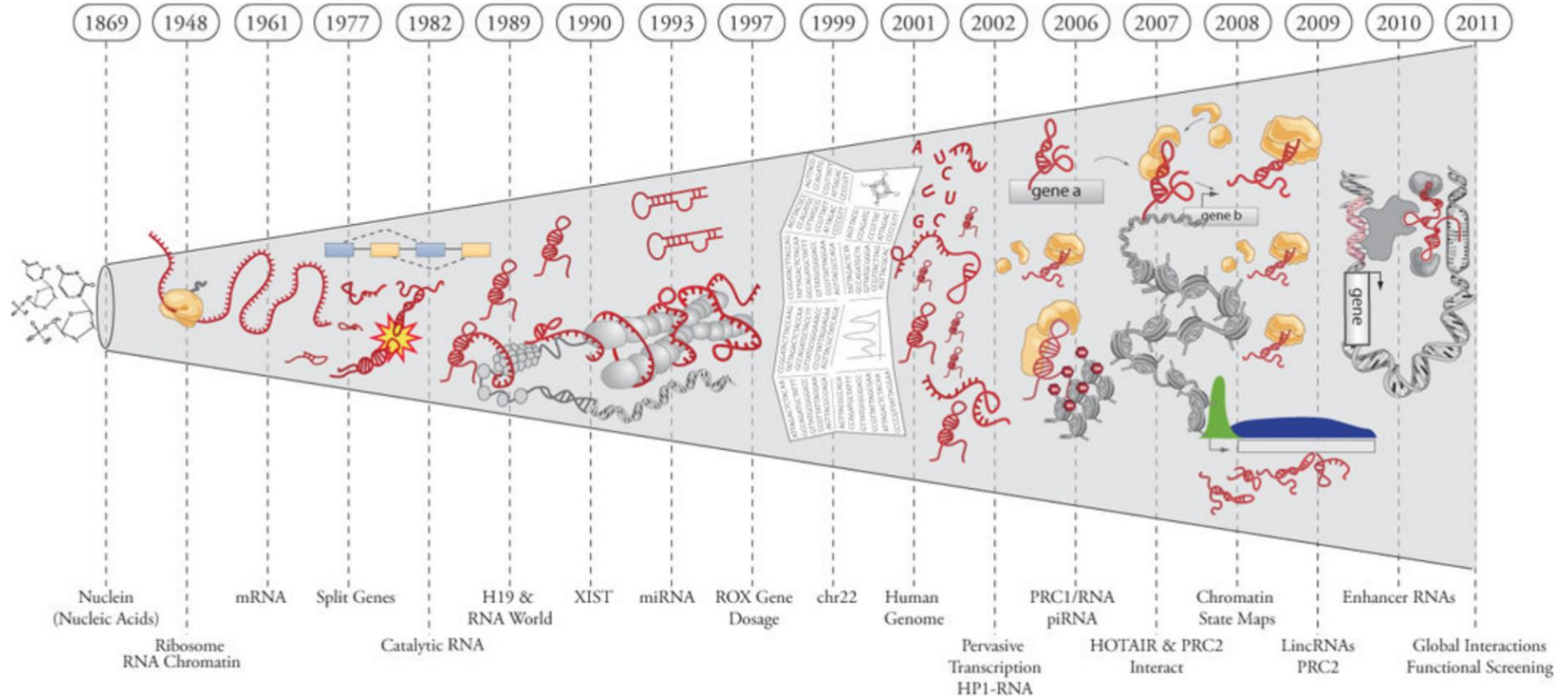
We listen to the genome via transcriptomics



As our ability to listen gets better, we uncover more classes of RNA

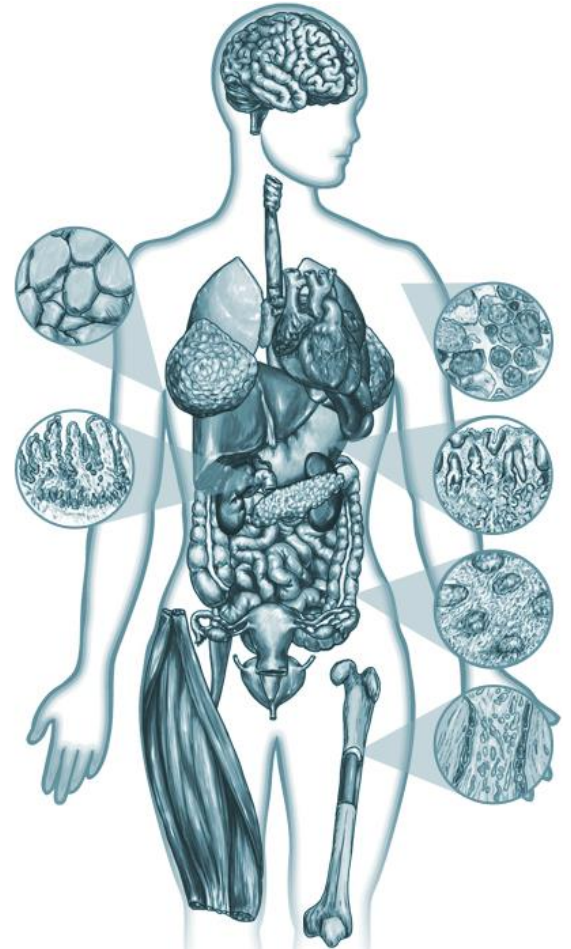
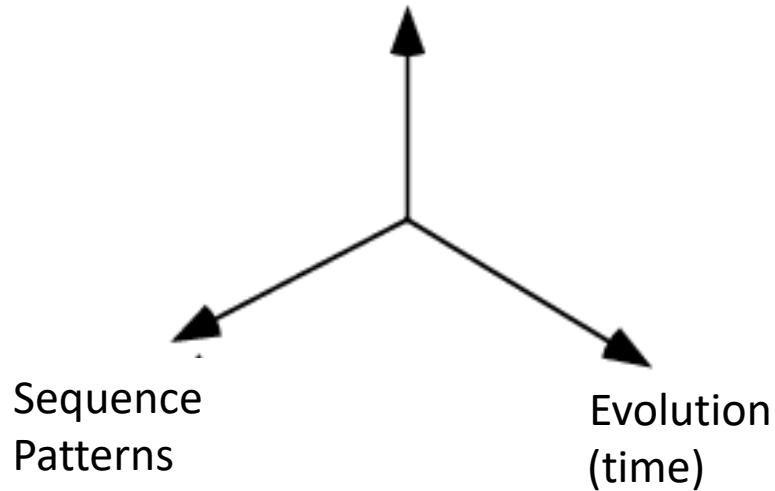


Many of newly discovered RNAs are regulatory



What do we need to crack the regulatory code?

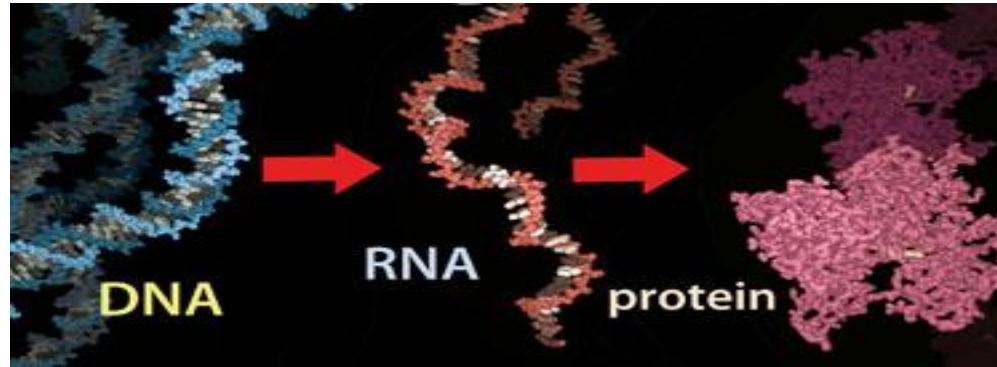
Measure the right thing!





"So things are good, stuff is OK, and I reiterate my request for more specific data."

Ironically, despite years of expression studies, we haven't been looking at transcription!



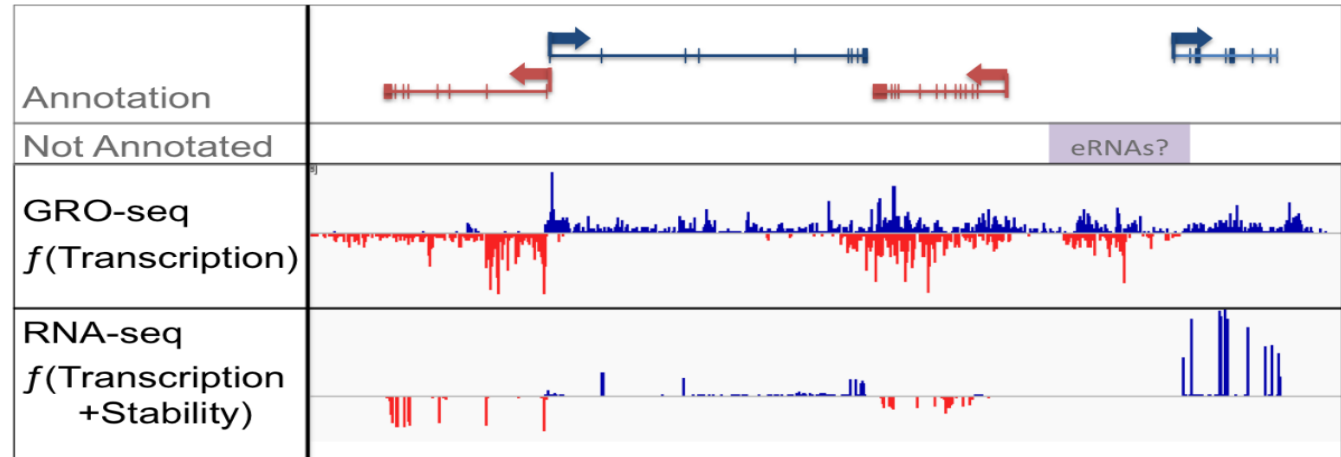
Nascent Transcription



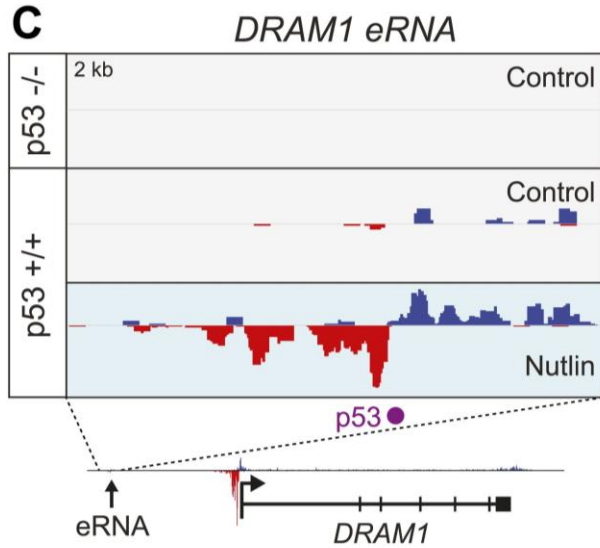
Steady state RNA = Expression studies

Direct measures of transcription indicate a large fraction of the genome is transcribed but
unstable

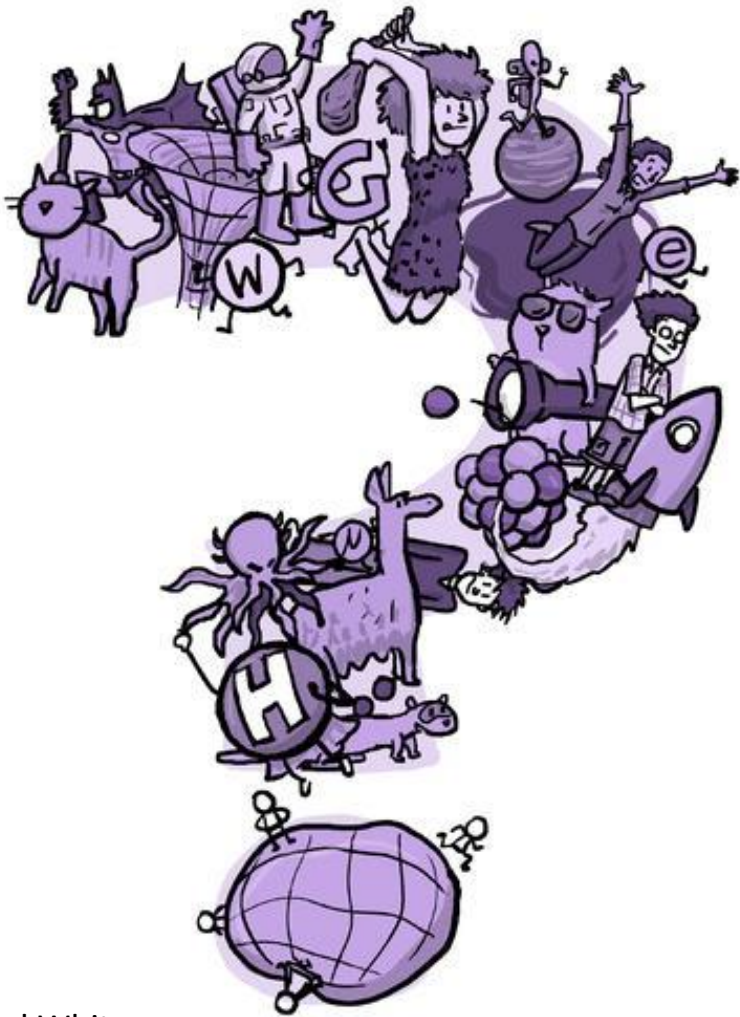
DNA
↓
Nascent RNA
↓
Mature RNA



■ forward strand (>0)
■ reverse strand (<0)



WHAT DOES ALL
THIS UNSTABLE
TRANSCRIPTION
DO?

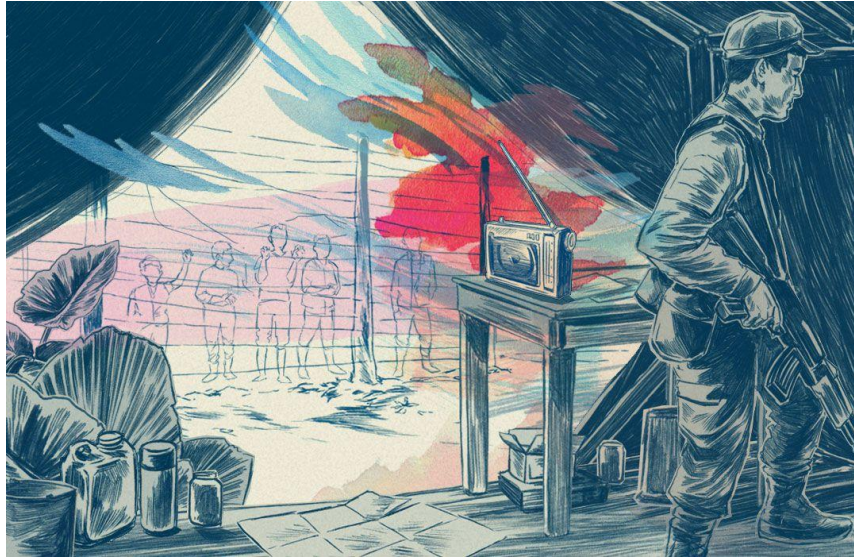


Fact of the
matter: We have
no idea!



Signal vs. Noise

There are hidden gems ...



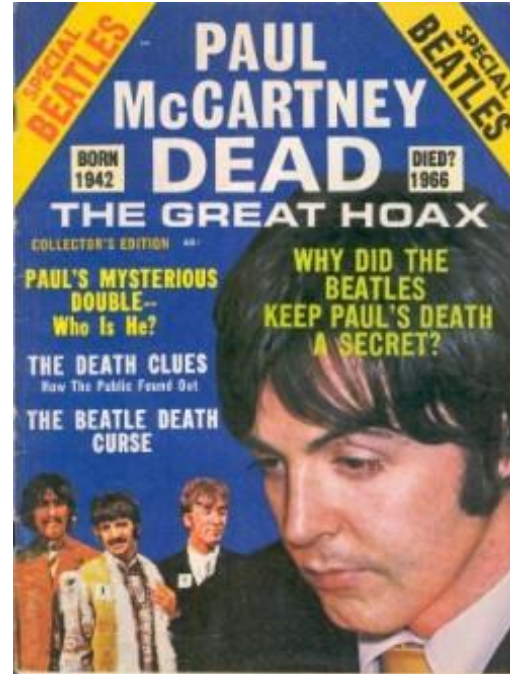
Inevitably, some of this unstable transcription is just noise.



Revolution #9

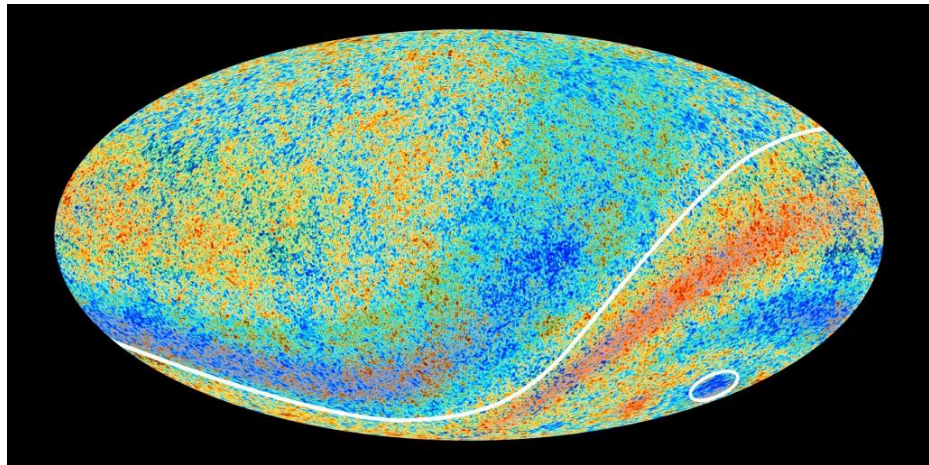


"turn me on, dead man"



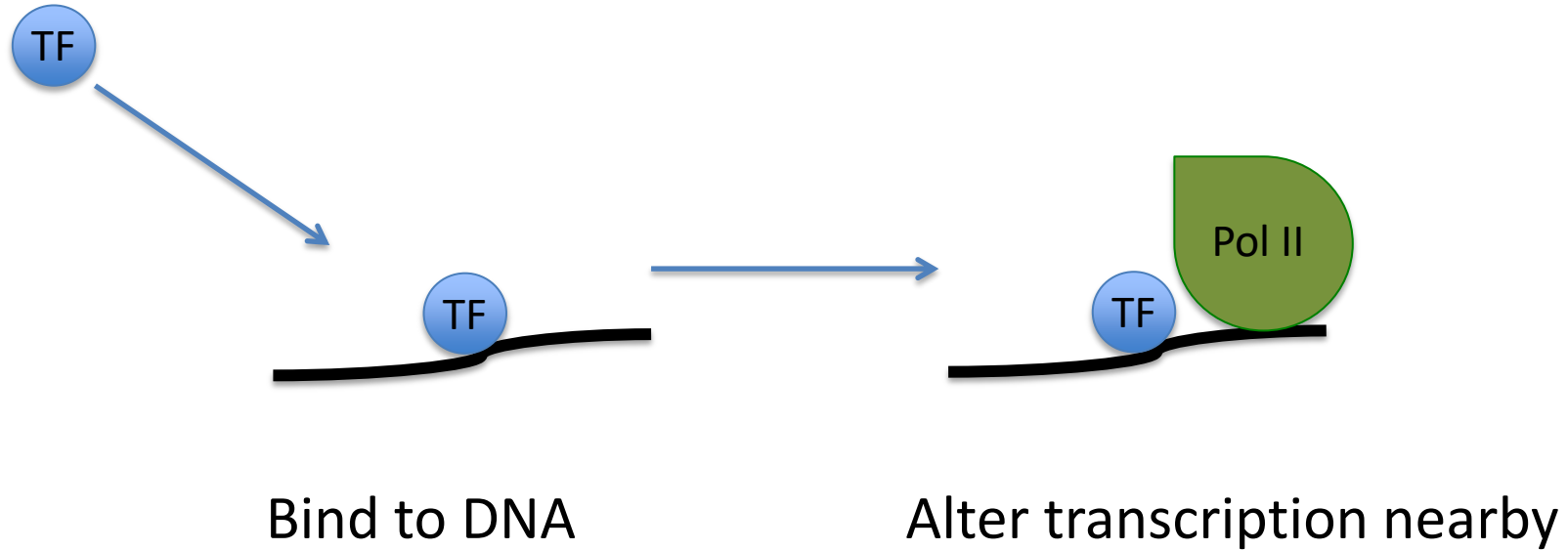
Sometimes “noise” is beautiful

In 1960's, Bell Labs' Holmdale Horn Antenna in New Jersey picked up an odd buzzing sound that came from all parts of the sky at all times.



Penzias and Wilson worked endlessly to remove the noise but eventually realized this cosmic microwave background radiation was the thermal echo of the universe's explosive birth.

Transcription factors are the regulatory machines of the genome, they bind DNA and alter transcription nearby.

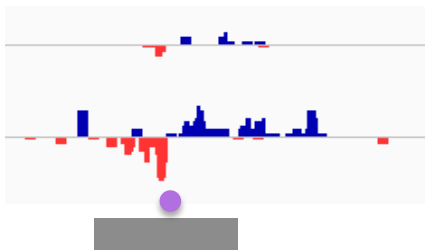


Alterations in transcription observed at many active TFs (enhancer RNAs)

p53 (Allen 2014)

Control

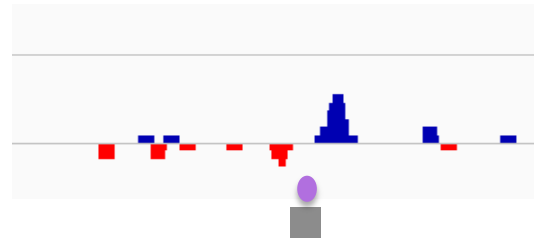
Nutlin



Androgen (Puc 2015)

Control

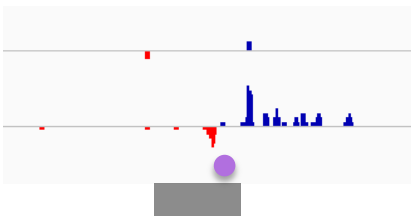
DHT



Estrogen (Hah 2013)

Control

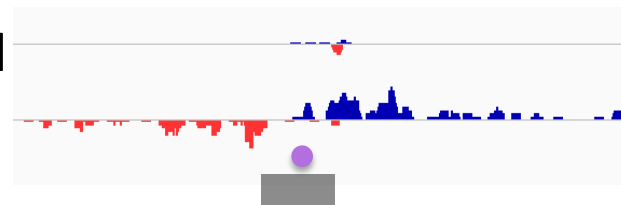
Estradiol





NF- κ B (Luo 2014)

Control

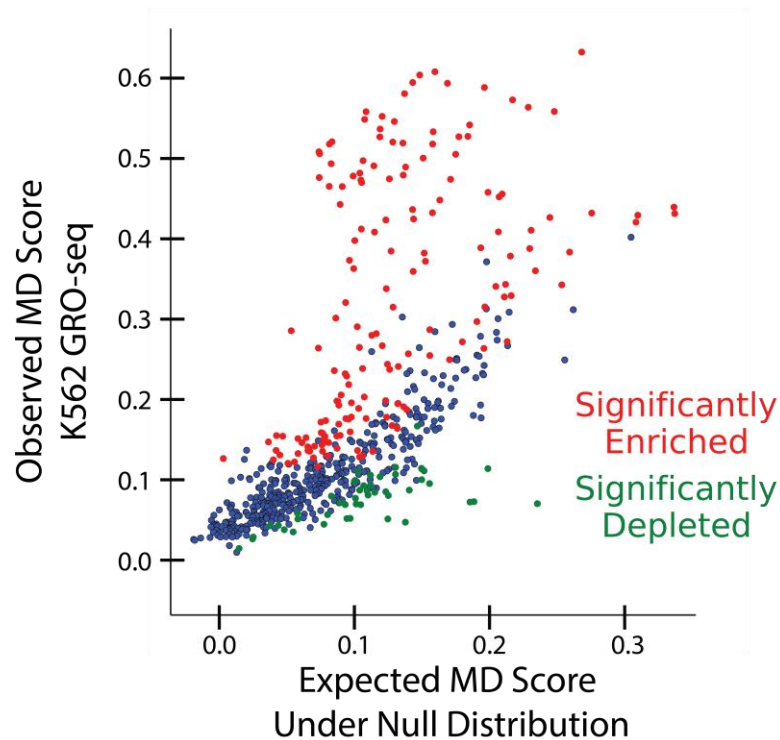
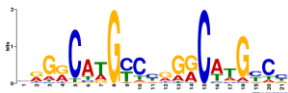
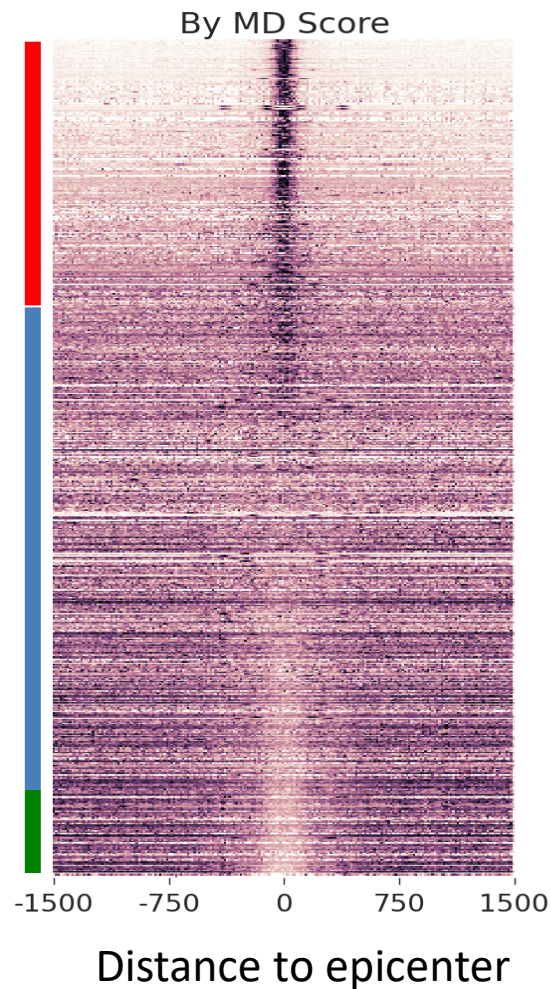
TNF α

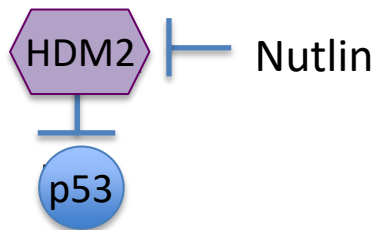


 Transcription factor motif
 ChIP-seq peak for indicated TF

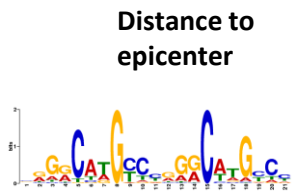
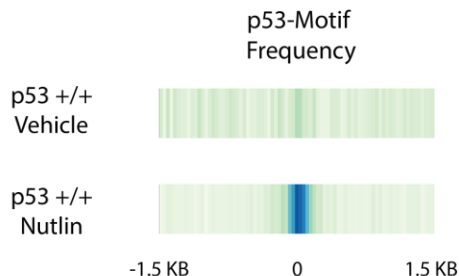
 forward strand (>0) Nascent transcription
 reverse strand (<0)

Co-localization predicts which TFs are active in a given cell type



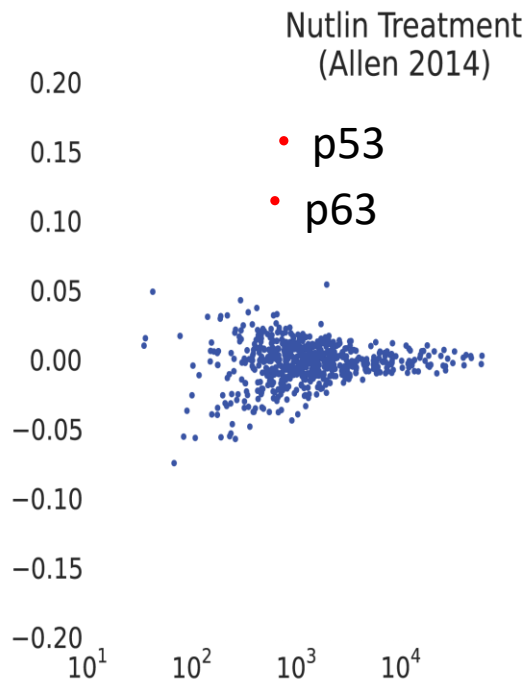


Changes in co-localization identify TFs altered in response to perturbation



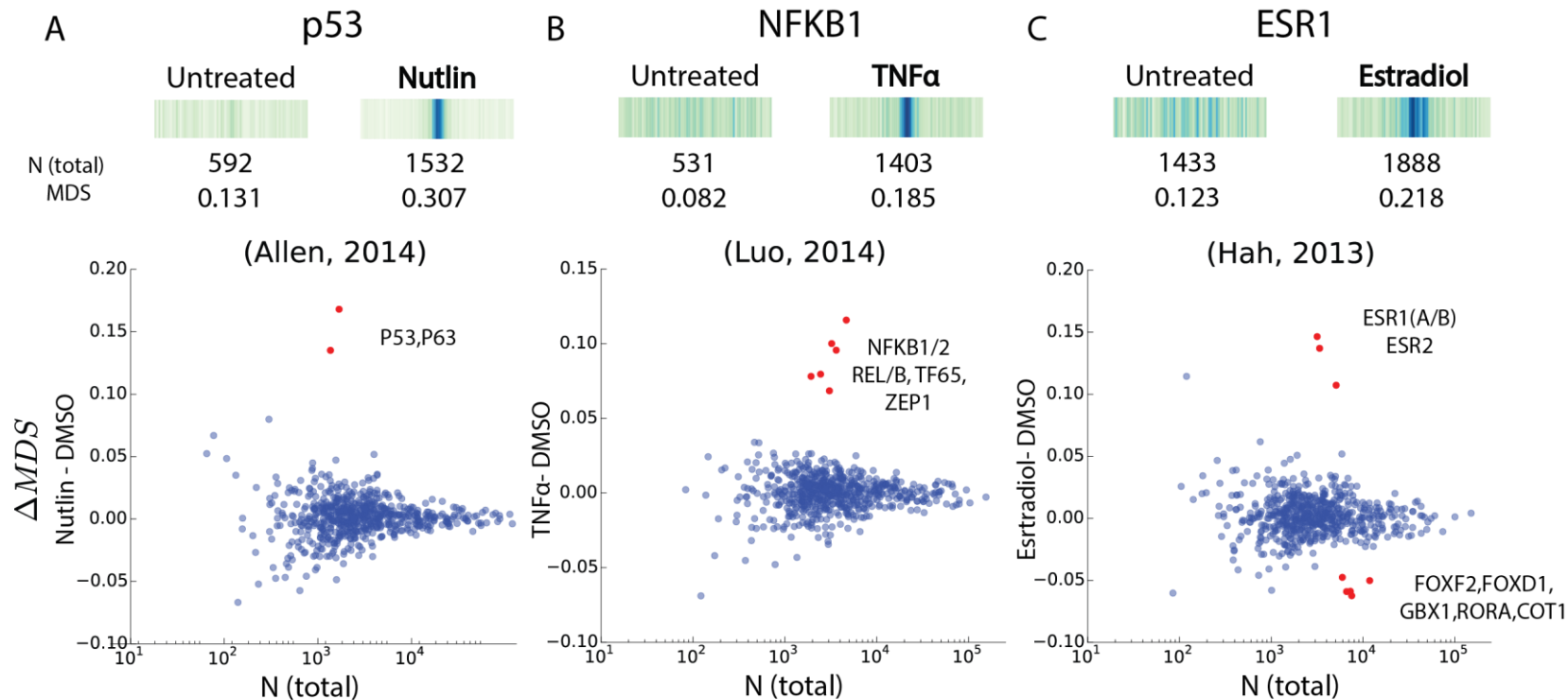
- TF motif (HOCOMOCO database)
- Statistically significant MD score shift

Difference in MD Score
(Treatment-Control)

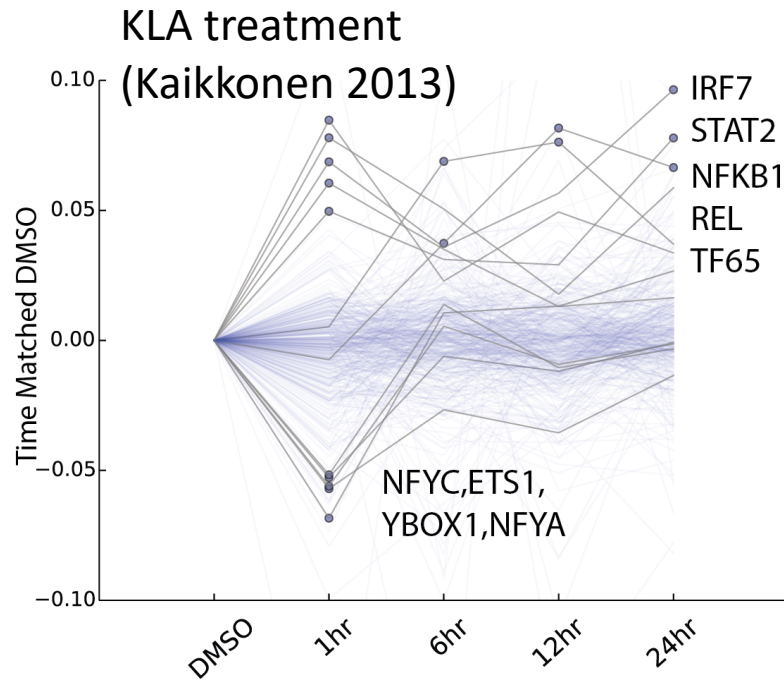
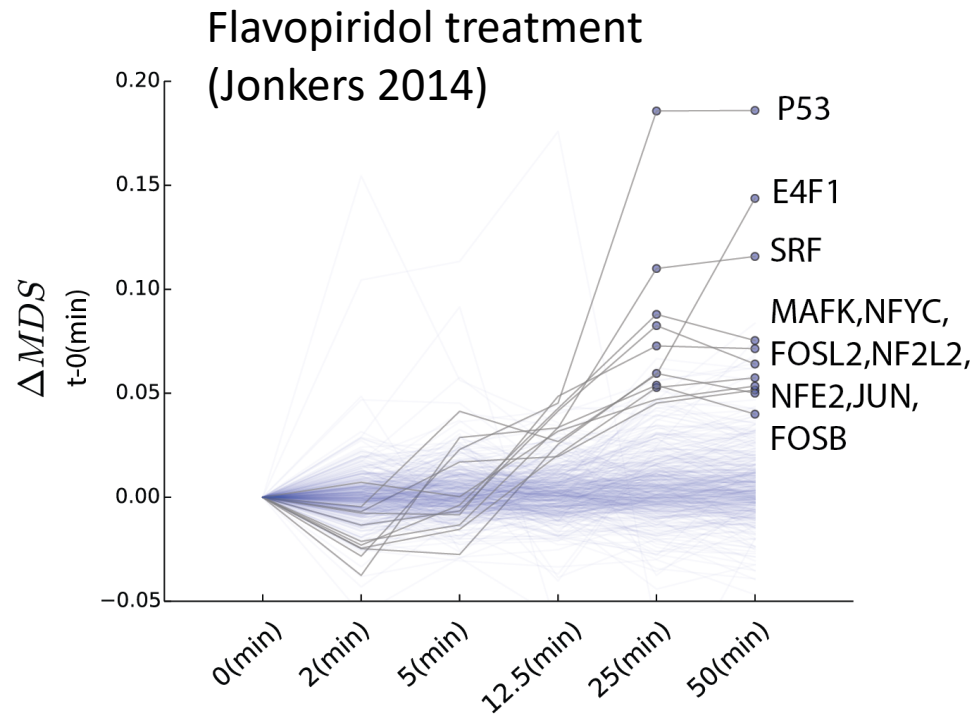


$(n_k + n_j)$: Number of epicenters with motif

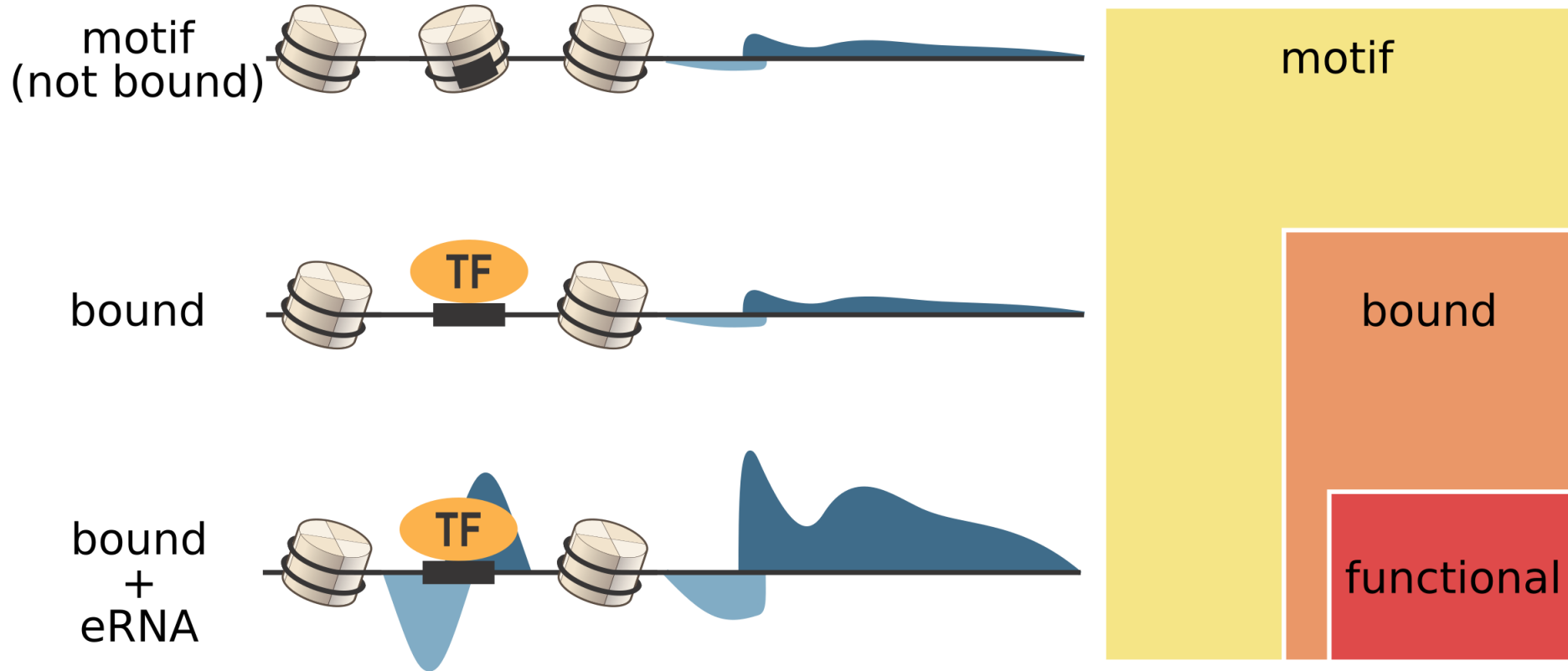
Identifying alterations in TF activity across distinct perturbations



Identifying alterations in TF activity across distinct time scales



Context matters in regulation ...



In language, meaning is influenced by context

Tear



E.g. I'll **tear** the package open.

Tear



E.g. She left the room in **tears**.

Bass



E.g. I play **bass** in a jazz band.

Bass



E.g. Fresh sea **bass** is a great delicacy.

Live



E.g. We used to **live** in London.

Live



E.g. The club has **live** music most nights.

Context **matters** to our comprehension.

Summary: 20 years of deciphering the genome



Genome Sequence Patterns

Transcriptional Regulation

“When I first started looking at **transcription regulation** all those years ago (2005), I thought, ‘Oh, I can just write a computer program and solve it,’ but it’s been kicking my ass all this time.”

--- Robin Dowell (today)

Moving towards fluency....

Genome Sequence Patterns

Transcriptional Regulation



Reading

Writing

Listening

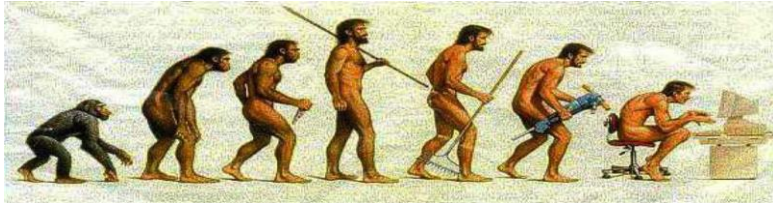
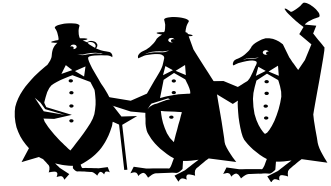
Speaking

Genome Modification

Synthetic Biology

(2020 Nobel Prize)

Acknowledgements



[DnA Laboratory](#),

Drs. Robin Dowell and Mary Allen
University of Colorado Boulder

Jacob Stanley, Gilson Sanchez, Lynn Sanford, Daniel Ramirez, Chi Zhang, Samuel Hunter, Jessica Westfall, Rutendo Sigauke, Qing Yang, Marko Melnick, Kendra Meer, Taylor Jones, Jesse Kurland, Zach Maas



Disclosure: Founder of

