12 recent discoveries that have changed the debate about design in the universe

Part IV

11. Extensive post-transcriptional processing (editing) of genes: the spliceosome and the splicing code.

12. Genes that extensively overlap in the same or opposite directions within a stretch of DNA (overlapping codes)

11. Extensive post-transcriptional processing (editing) of genes: **the spliceosome and the splicing code.**

Some references:

https://bio.libretexts.org/Bookshelves/Cell_and_Molecular_Biology/Book%3A_Cells_-_Molecules_and_Mechanisms_(Wong)/08%3A_Transcription/8.04%3A_Post-Transcriptional_Processing_of_RNA

The highly studied example of tropomyosin:

TRENDS in Cell Biol 2005, 15, 333-341.

Journal of Muscle Res. and Cell Motility (2020) 41:11–22

BioArchitecture 2011, 1:4, 135-164;

Hepatology International (2009) 3:378–383

Alternative splicing in aging and longevity Human Genetics (2020) 139:357–369,
https://doi.org/10.1007/s00439-019-02094-6

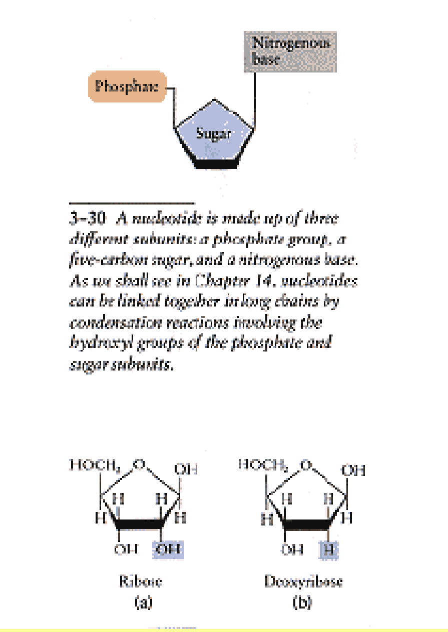
Alternative splicing in the heart Circulation Research 2016;118:454-468.

Some general background:

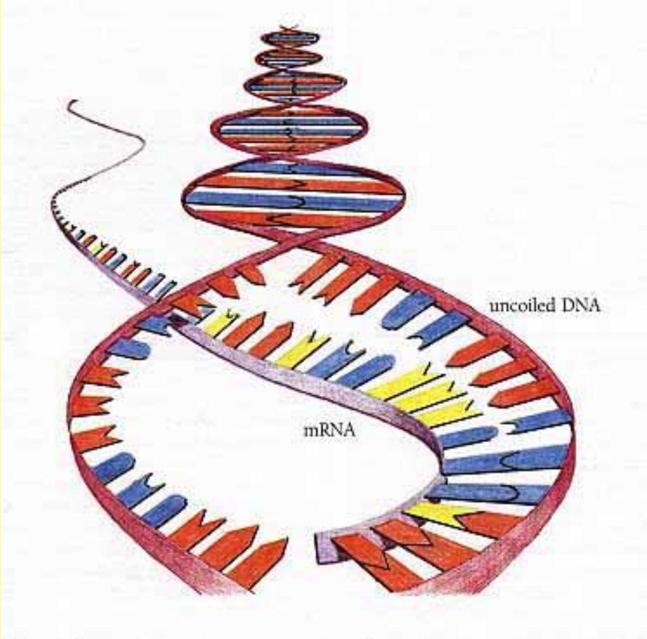
DNA

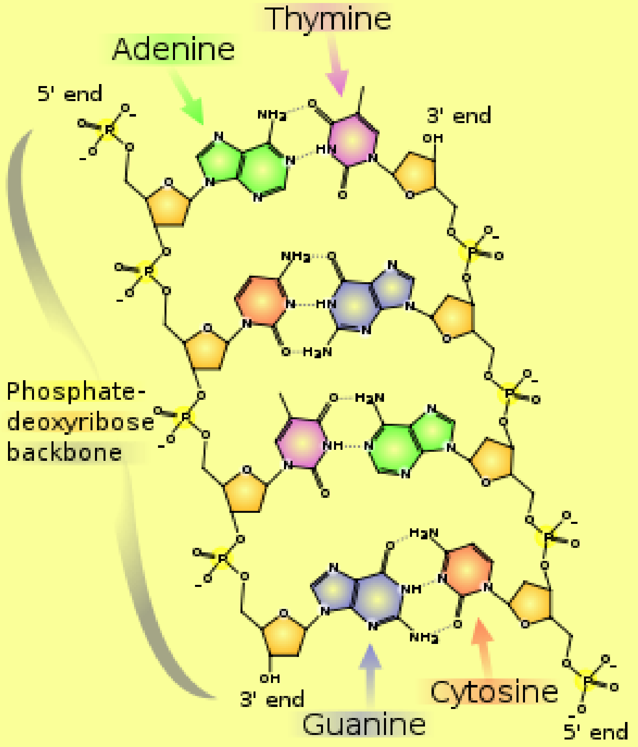
Nucleotides

4 in DNA



3–30 *A nucleotide is made up of three different subunits: a phosphate group, a five-carbon sugar, and a nitrogenous base. As we shall see in Chapter 14, nucleotides can be linked together in long chains by condensation reactions involving the hydroxyl groups of the phosphate and sugar subunits.*

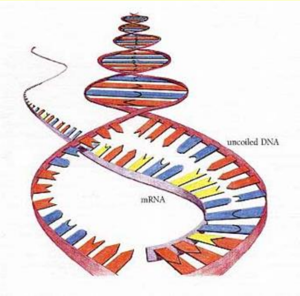


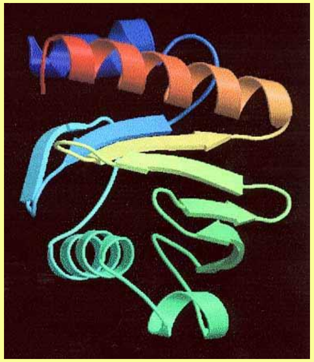


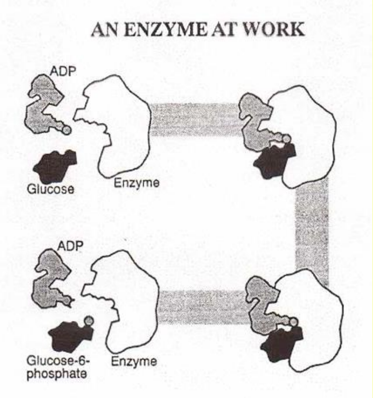
A chain of nucleotides:

ATCGCCTATATAGCGTACAATGGCTACATCGCCTATATAGCGTACAATGGCTACGTAGCTACGATGCTAGCTAGCTAGC
GCTACATCGCCTATATAGCGTACAATGGCTACGTAGCTACGATGCTAGCTAGCTAGCATCGCCTATATAGCGTACAATGGCTAC
ATCGCCTATATAGCGTACAATGGCTACGTAGCTACGATGCTAGCTAGCTAGCATCGCCTATATAGCGTACAATGGCTACATCGC
CTATATAGCGTACAATGGCTACGTAGCTACGATGCTAGCTAGCTAGCATCGCCTATATAGCGTACAATGGCTACATCGCCTATAT
AGCGTACAATGGCTACGTAGCTACGATGCTAGCTAGCTAGCGCTACATCGCCTATATAGCGTACAATGGCTACGTAGCTACGAT
GCTAGCTAGCTAGCATCGCCTATATAGCGTACAATGGCTACATCGCCTATATAGCGTACAATGGCTACGTAGCTACGATGCTAG
CTAGCTAGCATCGCCTATATAGCGTACAATGGCTACATCGCCTATATAGCGTACAATGGCTACGTAGCTACGATGCTAGCTAGC
TAGC

Some general background:







DNA → RNA → protein → structure/function

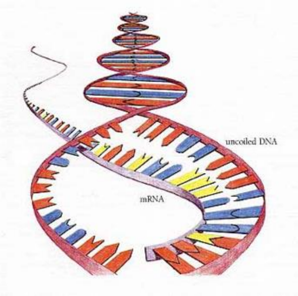
↑ RNA polymerase

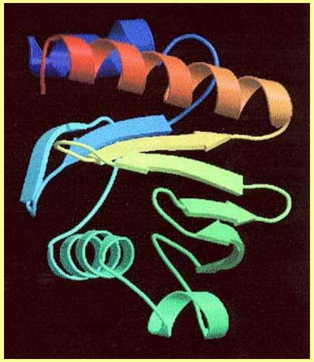
↑ ribosome

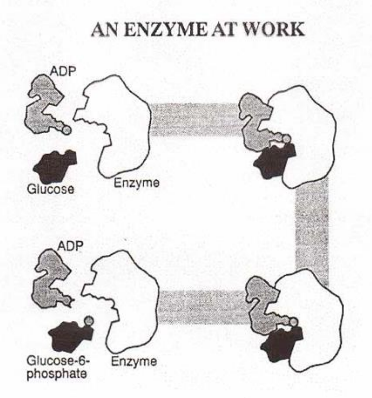
↑ chaperones

RNA also has many other functions

Some general background:



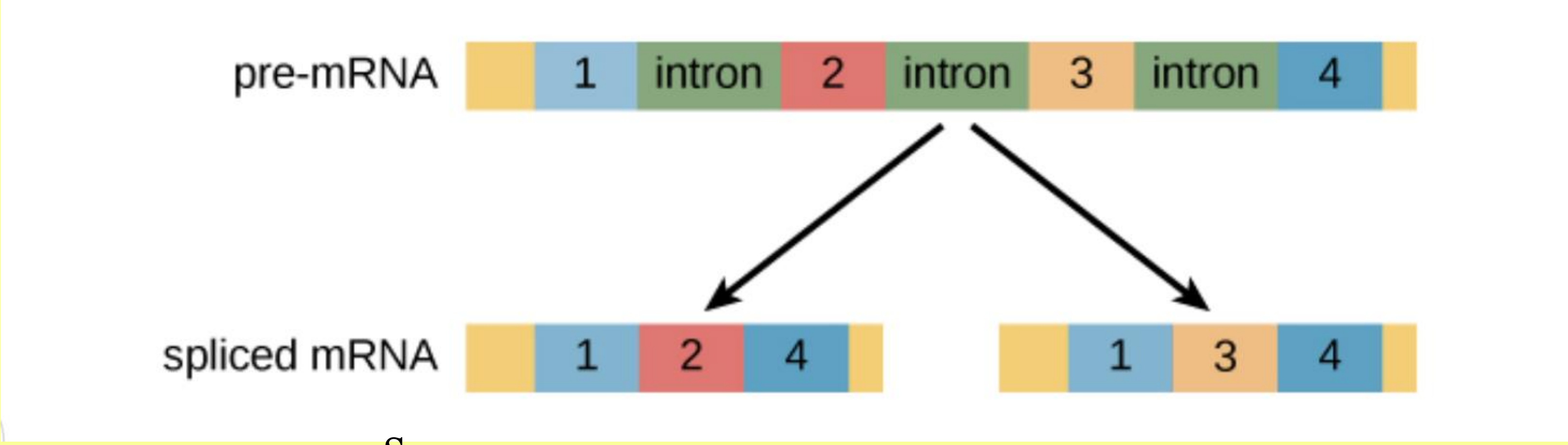




DNA → RNA → protein → structure/function

something else happens here in eukaryotes

as transcribed from DNA sequence



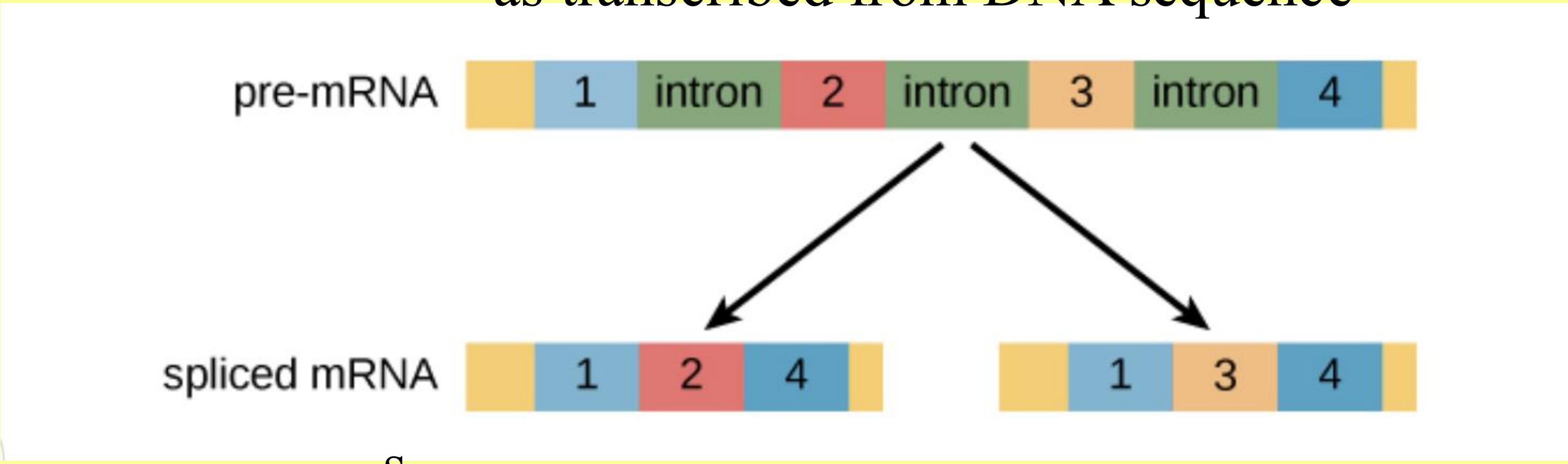
s

after RNA editing

From CHESS 3 database:

No. of protein-coding genes in humans: 19,839

No. of protein sequences in humans: 73,767 as transcribed from DNA sequence



s

after RNA editing

A machine (spliceosome) does this

A splicing code governs the process

mRNA editing affects:

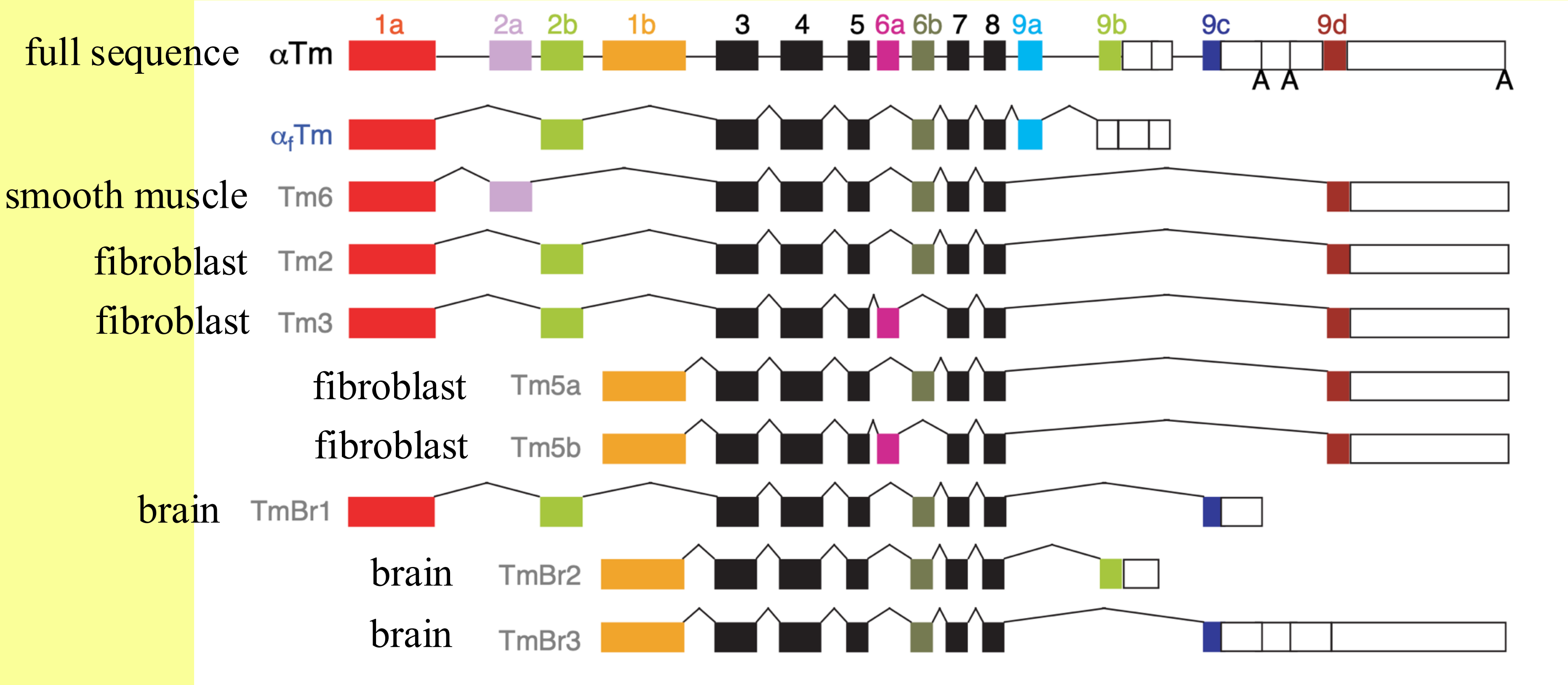
- -sequences of the proteins that are made in different cell types at different times
- -nuclear transport
- -efficiency of translation
- -rate of degradation

Errors in mRNA editing are associated with:

- -cancers
- -neurodegenerative diseases

only occurs in eukaryotes

Example: α-tropomyosin (regulates muscle contraction)



full sequence

smooth muscle

fibroblast

fibroblast

fibroblast

fibroblast

brain

brain

brain

From TRENDS in Cell Biology Vol.15 No.6 June 2005

Some basics of the splicing code

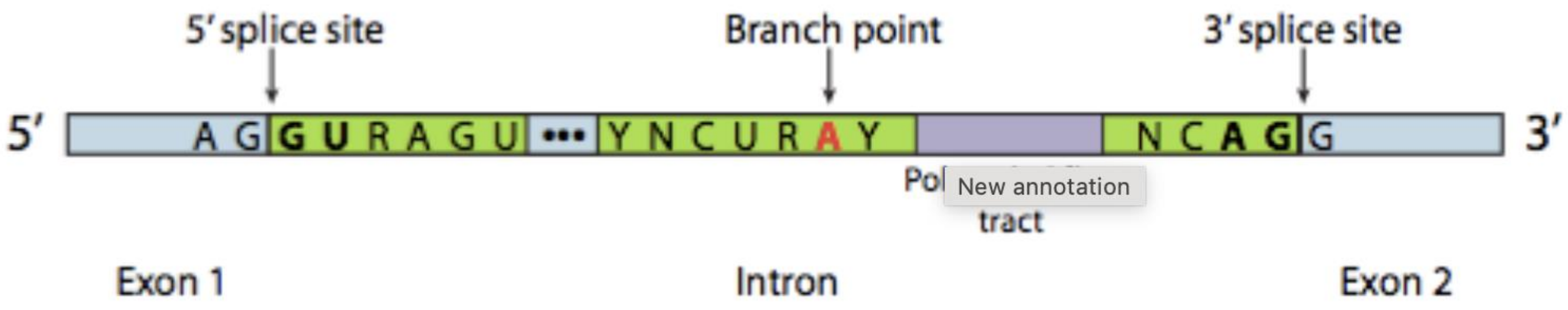


Figure 8.4.8. Consensus sequences for splicing.

U replaces T in RNA

Y is a pyrimidine (C and T)
N is any nucleotide
R is purine (A or G)

from: https://bio.libretexts.org/Bookshelves/Cell_and_Molecular_Biology/Book%3A_Cells_-_Molecules_and_Mechanisms_(Wong)/08%3A_Transcription/8.04%3A_Post-Transcriptional_Processing_of_RNA

Some basics of the splicing code

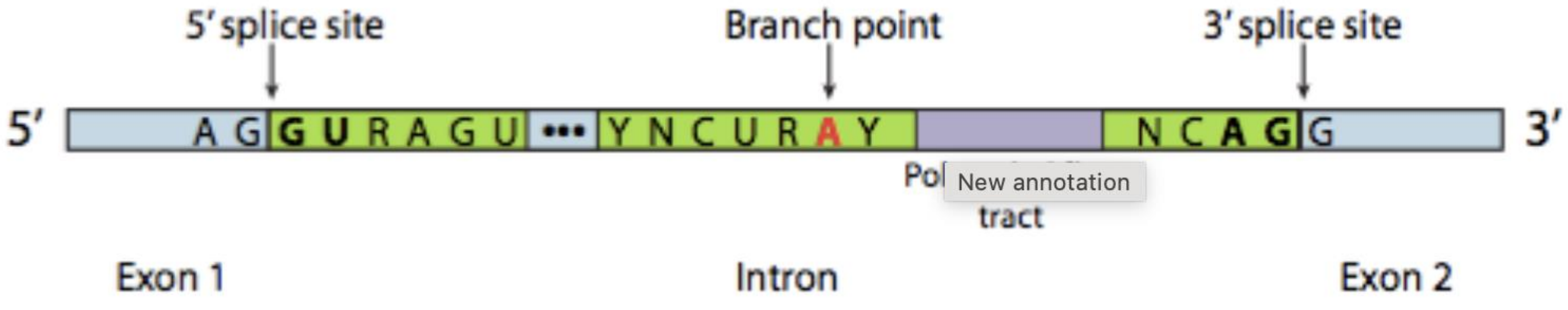


Figure 8.4.8. Consensus sequences for splicing.

Many other factors are involved in the code:

Intron and exon regions contain additional sequence elements that repress or enhance the process by binding to proteins or RNAs

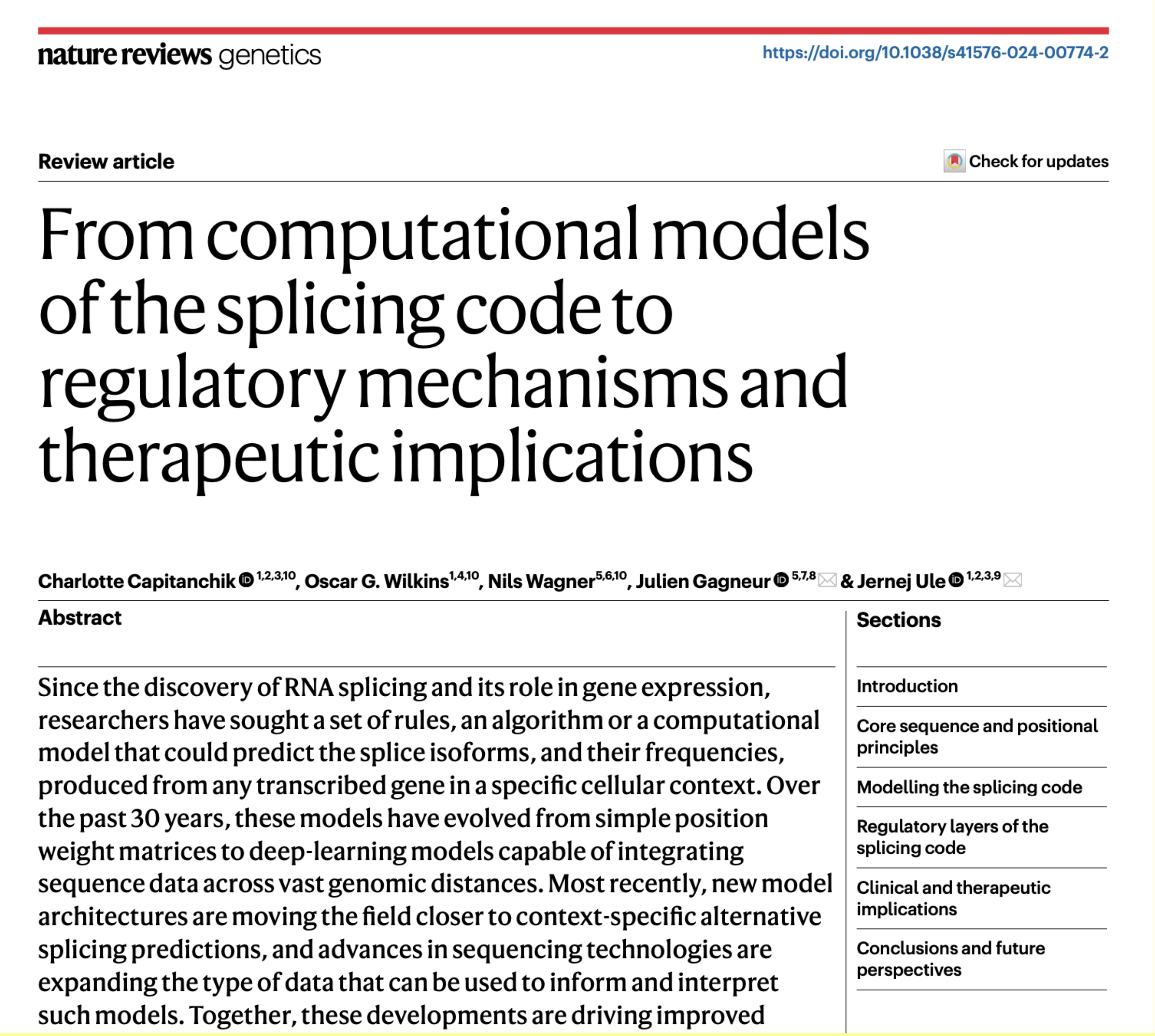
Different cell types express different splicing factors (proteins and RNAs)

Deep Learning AI is being used to try to sort this out

Videos of RNA splicing by the spliceosome:

https://www.google.com/search?q=splicosome+video&sca_esv=0a9855b1adeae542&biw=1022&bih=516&tbm=vid&sxsrf=ADLYWIKEizApoXReSL85_tiniP4jMUz6-Q%3A1730234576666&ei=0EghZ5KpKLiA0PEPmJe_iQ8&ved=0ahUKEwiS0ceKurSJAxU4ADQIHZjLL_EQ4dUDCA0&uact=5&oq=splicosome+video&gs_lp=Eg1nd3Mtd2l6LXZpZGVvIhBzcGxpY29zb21lIHZpZGVvMgYQABgWGB4yCxAAGIAEGIYDGIoFMgsQABiABBiGAxiKBTILEAAYgAQYhgMYigUyCBAAGIAEGKIEMggQABiABBiiBDIIEAAYogQYiQUyCBAAGIAEGKIEMggQABiABBiiBEixQ1AAWMc6cAB4AJABAJgBogGgAf8NqgEEMi4xNLgBA8gBAPgBAZgCEKACtg7CAgsQABiABBiRAhiKBcICChAAGIAEGEMYigXCAgsQABiABBixAxiDAcICDhAAGIAEGLEDGIMBGIoFwgIFEAAYgATCAggQABiABBixA8ICDRAAGIAEGLEDGEMYigXCAhAQABiABBixAxhDGIMBGIoFwgIKEAAYgAQYsQMYCsICDRAAGIAEGLEDGIMBGArCAgcQABiABBgKwgIHEAAYgAQYDcICBhAAGA0YHsICCBAAGAgYDRgemAMAkgcEMi4xNKAH82g&sclient=gws-wiz-video#fpstate=ive&vld=cid:e320c8b6,vid:aVgwr0QpYNE,st:0

https://www.google.com/search?q=splicosome+video&sca_esv=0a9855b1adeae542&biw=1022&bih=516&tbm=vid&sxsrf=ADLYWIKEizApoXReSL85_tiniP4jMUz6-Q%3A1730234576666&ei=0EghZ5KpKLiA0PEPmJe_iQ8&ved=0ahUKEwiS0ceKurSJAxU4ADQIHZjLL_EQ4dUDCA0&uact=5&oq=splicosome+video&gs_lp=Eg1nd3Mtd2l6LXZpZGVvIhBzcGxpY29zb21lIHZpZGVvMgYQABgWGB4yCxAAGIAEGIYDGIoFMgsQABiABBiGAxiKBTILEAAYgAQYhgMYigUyCBAAGIAEGKIEMggQABiABBiiBDIIEAAYogQYiQUyCBAAGIAEGKIEMggQABiABBiiBEixQ1AAWMc6cAB4AJABAJgBogGgAf8NqgEEMi4xNLgBA8gBAPgBAZgCEKACtg7CAgsQABiABBiRAhiKBcICChAAGIAEGEMYigXCAgsQABiABBixAxiDAcICDhAAGIAEGLEDGIMBGIoFwgIFEAAYgATCAggQABiABBixA8ICDRAAGIAEGLEDGEMYigXCAhAQABiABBixAxhDGIMBGIoFwgIKEAAYgAQYsQMYCsICDRAAGIAEGLEDGIMBGArCAgcQABiABBgKwgIHEAAYgAQYDcICBhAAGA0YHsICCBAAGAgYDRgemAMAkgcEMi4xNKAH82g&sclient=gws-wiz-video#fpstate=ive&vld=cid:ca71552c,vid:FVuAwBGw_pQ,st:0



nature reviews genetics

https://doi.org/10.1038/s41576-024-00774-2

Review article

Check for updates

From computational models of the splicing code to regulatory mechanisms and therapeutic implications

Charlotte Capitanchik[1,2,3,10], Oscar G. Wilkins[1,4,10], Nils Wagner[5,6,10], Julien Gagneur[5,7,8] & Jernej Ule[1,2,3,9]

Abstract

Since the discovery of RNA splicing and its role in gene expression, researchers have sought a set of rules, an algorithm or a computational model that could predict the splice isoforms, and their frequencies, produced from any transcribed gene in a specific cellular context. Over the past 30 years, these models have evolved from simple position weight matrices to deep-learning models capable of integrating sequence data across vast genomic distances. Most recently, new model architectures are moving the field closer to context-specific alternative splicing predictions, and advances in sequencing technologies are expanding the type of data that can be used to inform and interpret such models. Together, these developments are driving improved

Sections

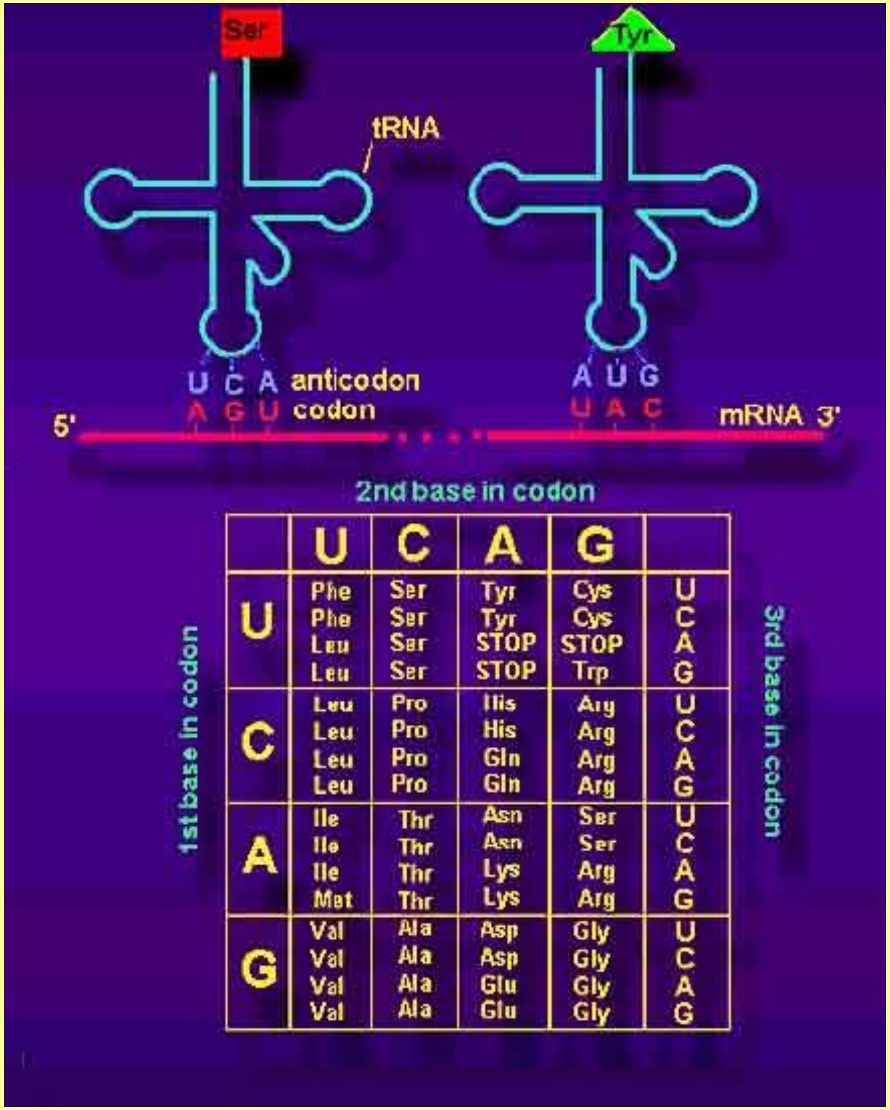
A code exists to govern when and where this happens

Goal of research - a splicing model that can predict from genome sequence alone the complete set of transcripts and their frequencies in any cellular context

Does RNA editing and the spliceosome represent an informational discontinuity?

12. Genes that extensively overlap in the same or opposite directions within a stretch of DNA (overlapping codes)

The Genetic Code



2nd base in codon

1st base in codon

3rd base in codon

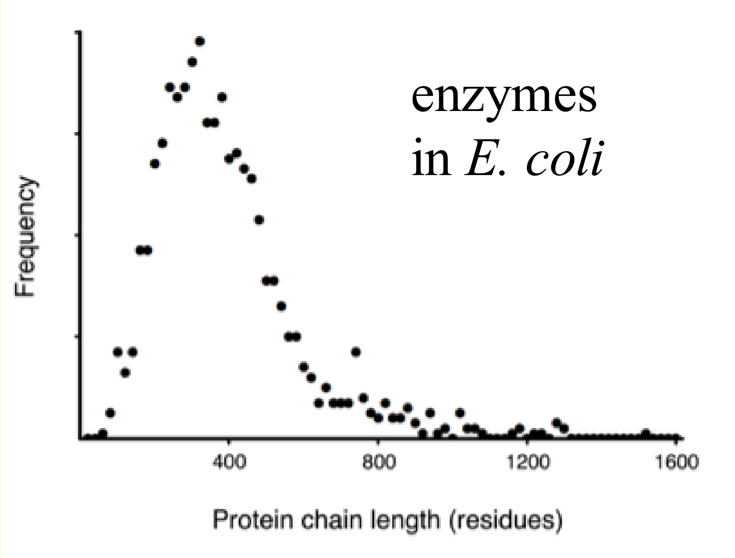
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

DNA: C, A, T, G

RNA: C, A, U, G

an information processing system

A.



20 types of amino acids, chains of amino acids 300 units long

20^{300} possibilities

Only a miniscule fraction of sequence space can be searched!

B. Fraction of sequences that fold (150 amino acids):

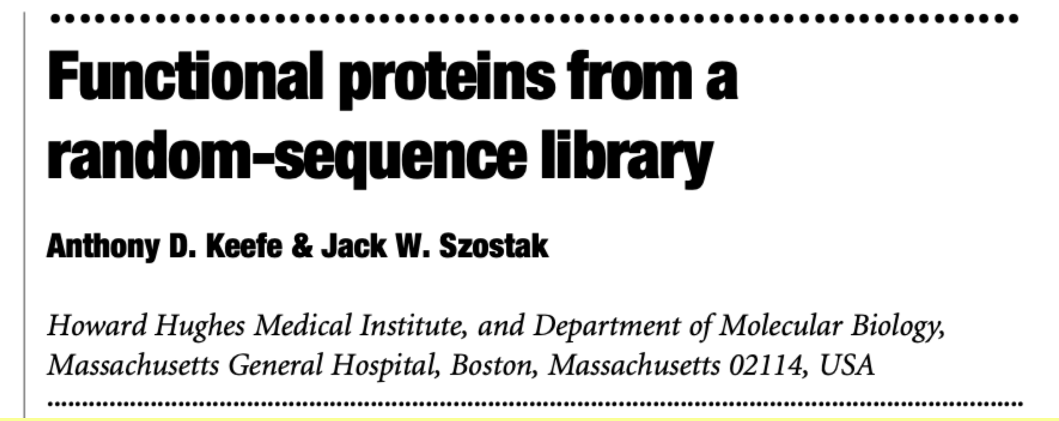
1 in 10^{63} (Sauer et al Proteins: Struct.
Function, and Genetics, 1990)

1 in 10^{74} (Axe J. Molec. Biol. 2004)

Only a miniscule fraction are highly functional (i.e.. enzymes)!

So where do the functional sequences come from?

Caveat – some functions such as binding a small molecule do not require sophisticated folds, and sequences that perform that function are much more prevalent

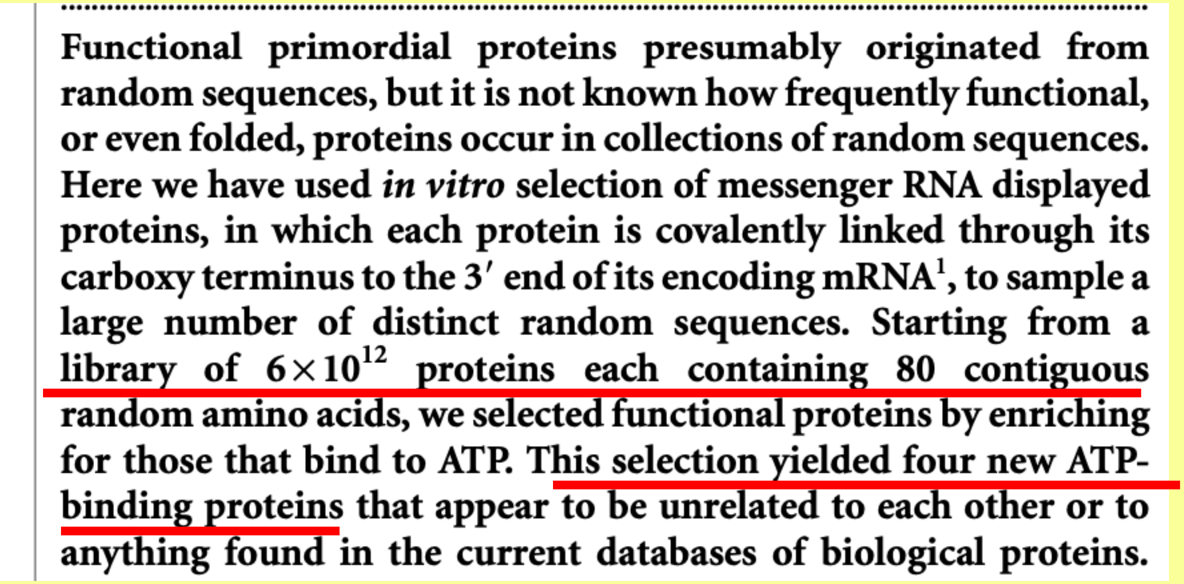


Functional proteins from a random-sequence library

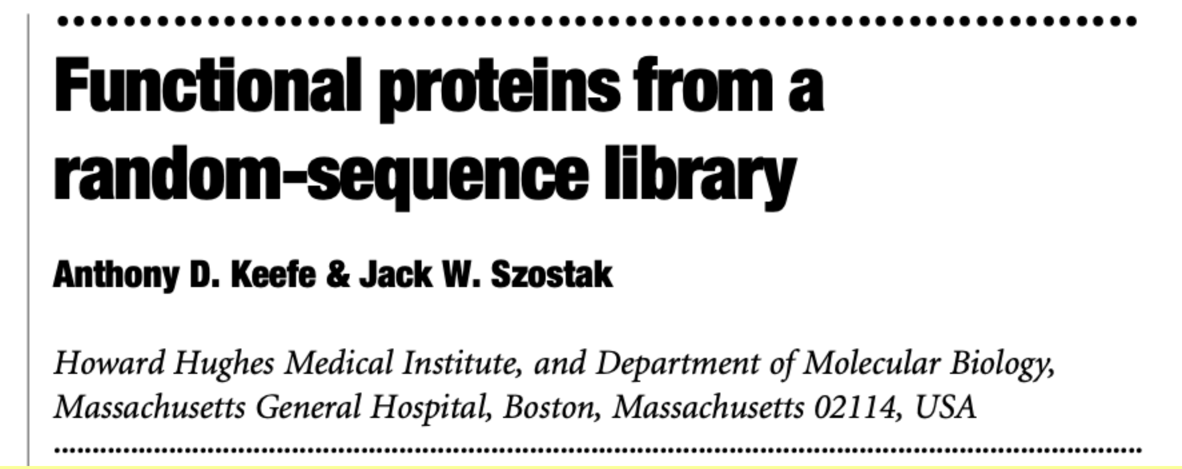
Anthony D. Keefe & Jack W. Szostak

Howard Hughes Medical Institute, and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

Nature
2001, 410, 717



Functional primordial proteins presumably originated from random sequences, but it is not known how frequently functional, or even folded, proteins occur in collections of random sequences. Here we have used *in vitro* selection of messenger RNA displayed proteins, in which each protein is covalently linked through its carboxy terminus to the 3′ end of its encoding mRNA[1], to sample a large number of distinct random sequences. Starting from a library of 6×10^{12} proteins each containing 80 contiguous random amino acids, we selected functional proteins by enriching for those that bind to ATP. This selection yielded four new ATP-binding proteins that appear to be unrelated to each other or to anything found in the current databases of biological proteins.



Functional proteins from a random-sequence library

Anthony D. Keefe & Jack W. Szostak

Howard Hughes Medical Institute, and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

Nature 2001, 410, 717

sequences of 80 amino acids

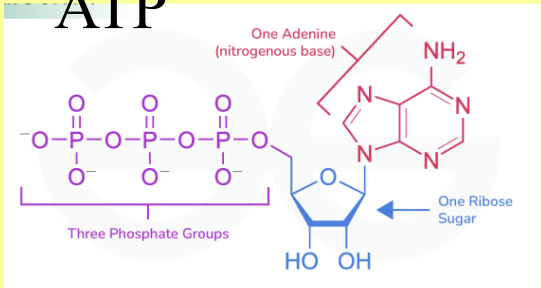
4 out of 6 trillion sequences bound ATP

very low-level functionality

-unclear that sequences must fold to bind ATP

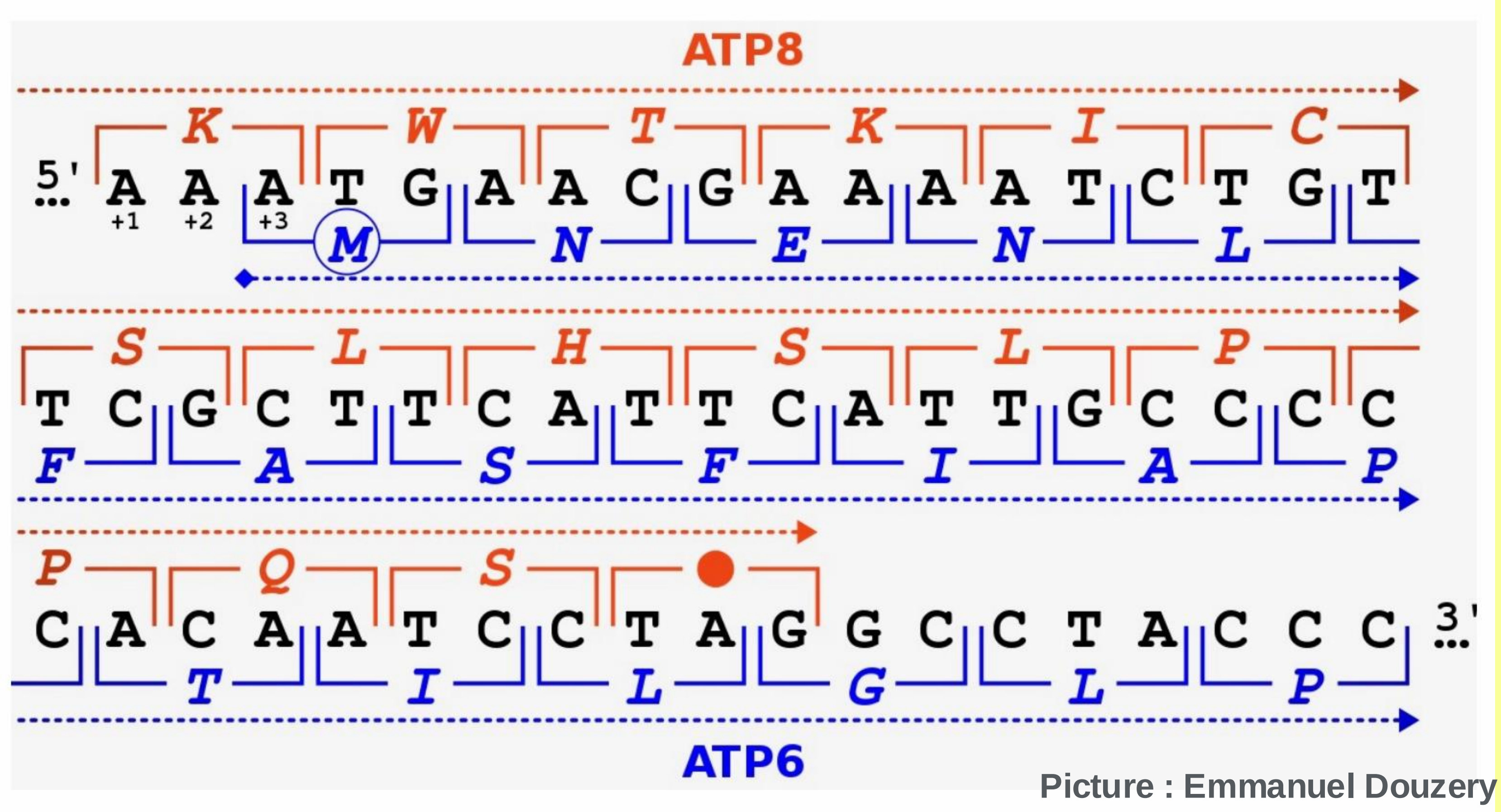
-subsequent work showed that a 12 aa peptide binds to ATP

ATP



Alternative reading frames:

example - human mitochondrial DNA


ATP8
K W T K I C
5' A A A T G A A C G A A A A A T C T G T
+1 +2 +3
M N E N L
S L H S L P
T C G C T T C A T T C A T T G C C C C
F A S F I A P
P Q S ●
C A C A A T C C T A G G C C T A C C C 3'
T I L G L P
ATP6

Picture : Emmanuel Douzery

6 possible reading frames: 3 on each DNA strand

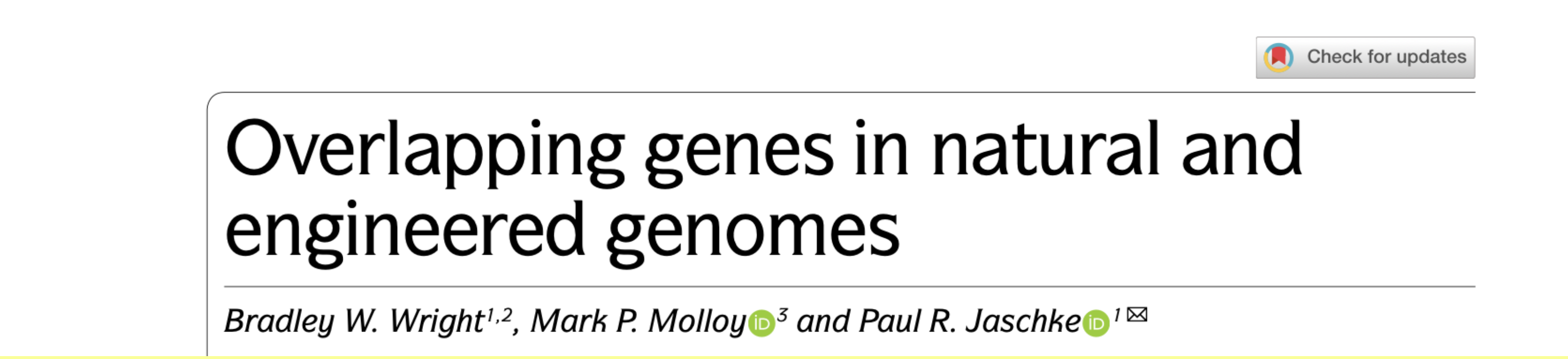
Recent discovery : Genes that extensively overlap in the same or opposite directions within a stretch of DNA (overlapping codes)

Seemingly impossible to explain by random mutation and natural selection: if one message is randomly mutated invariably the other one that is layered on top will be destroyed

Initially discovered in viruses, now known in all domains of life

> 26% of protein-coding genes in humans have some overlapping features (but most overlaps are short)

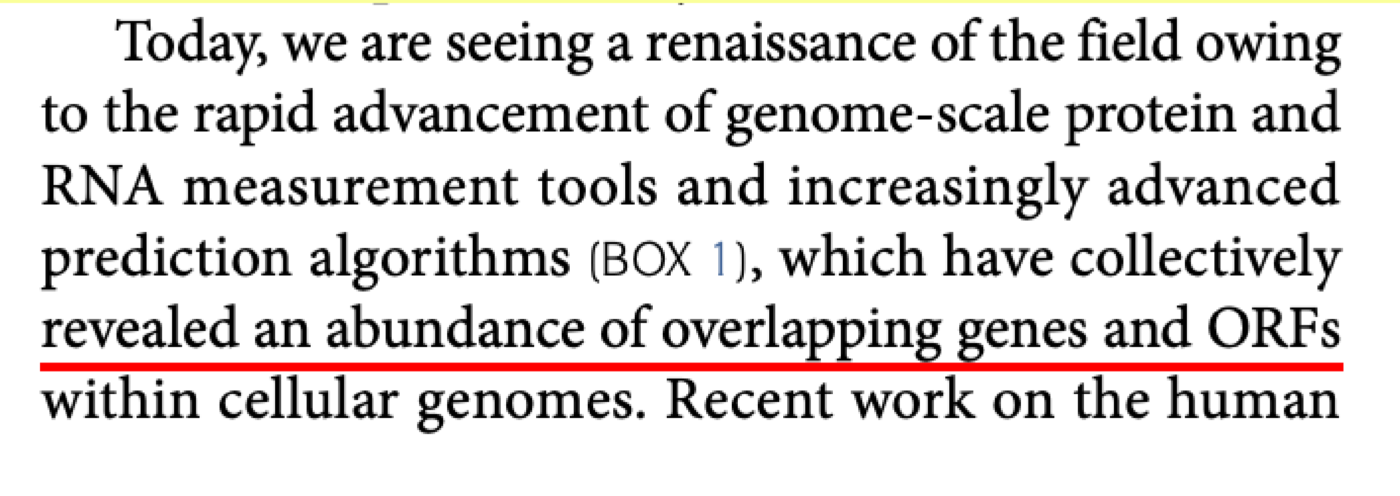
Review article


Check for updates

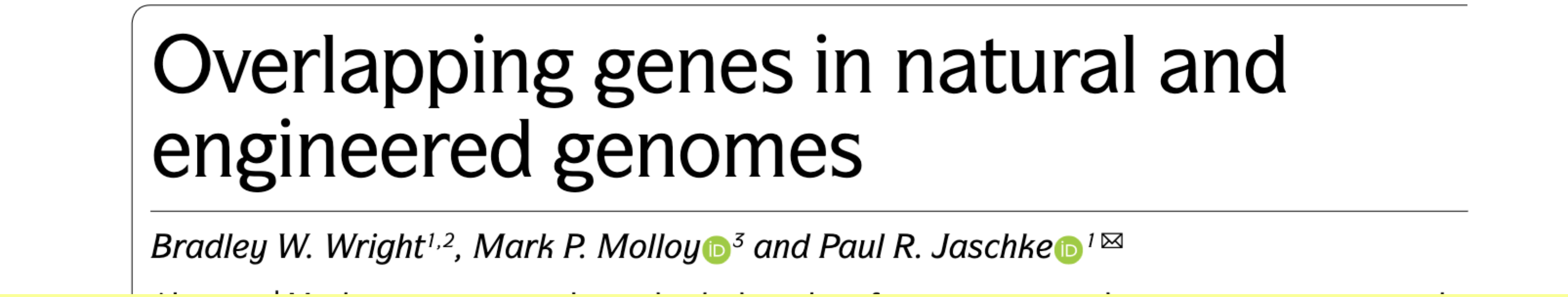
Overlapping genes in natural and engineered genomes

Bradley W. Wright[1,2], *Mark P. Molloy*[3] *and Paul R. Jaschke*[1] ✉

Nature March 2022, 23, p 154

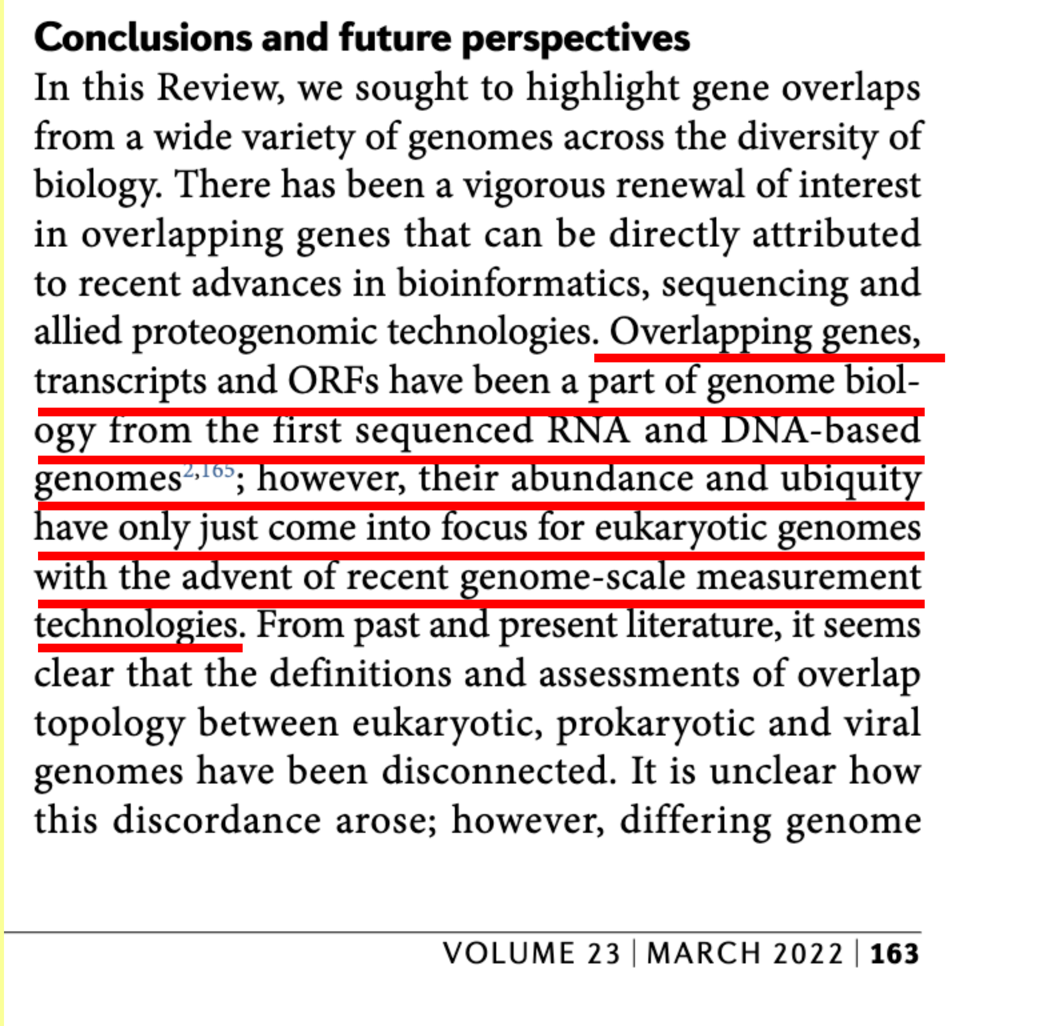


Today, we are seeing a renaissance of the field owing to the rapid advancement of genome-scale protein and RNA measurement tools and increasingly advanced prediction algorithms (BOX 1), which have collectively revealed an abundance of overlapping genes and ORFs within cellular genomes. Recent work on the human



Overlapping genes in natural and engineered genomes

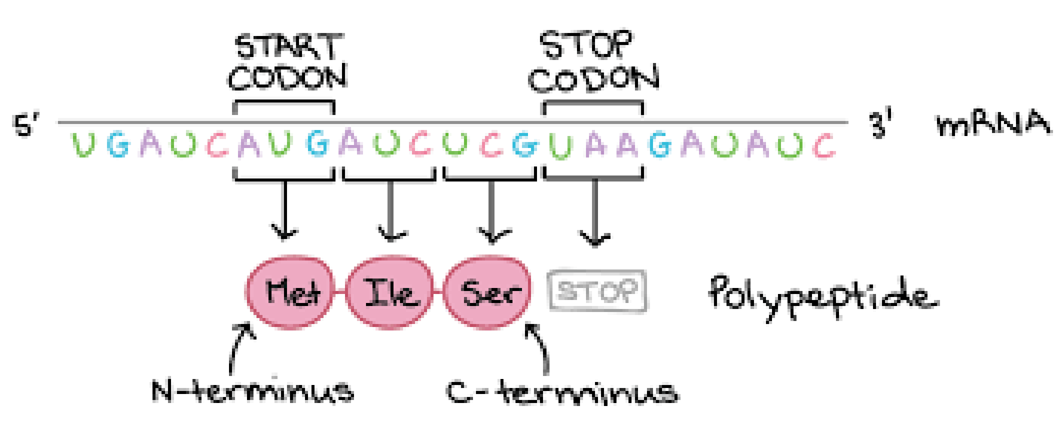
Bradley W. Wright[1,2], Mark P. Molloy[3] and Paul R. Jaschke[1] ✉

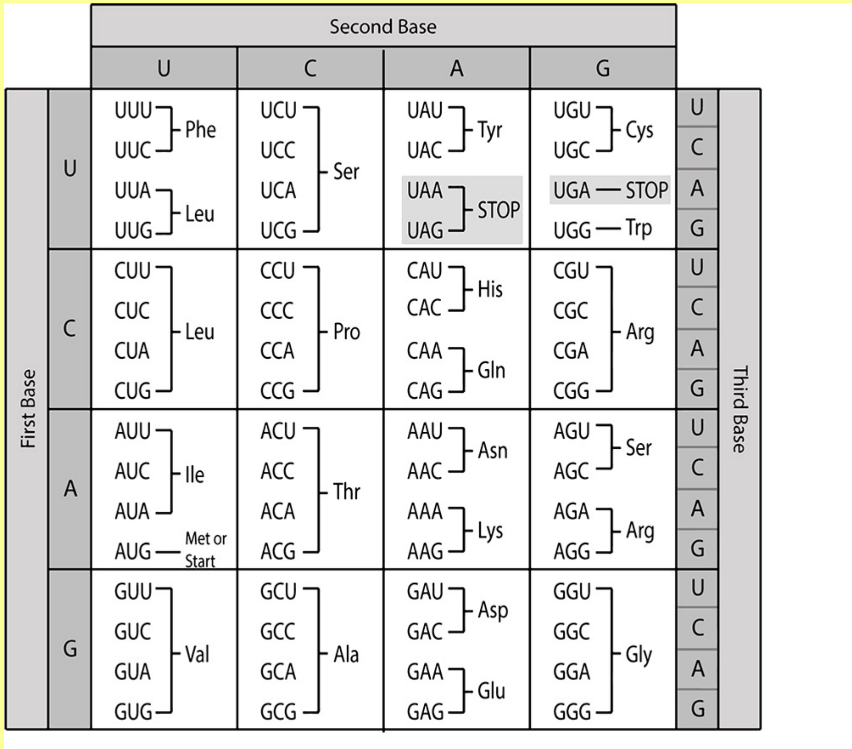


Conclusions and future perspectives

In this Review, we sought to highlight gene overlaps from a wide variety of genomes across the diversity of biology. There has been a vigorous renewal of interest in overlapping genes that can be directly attributed to recent advances in bioinformatics, sequencing and allied proteogenomic technologies. Overlapping genes, transcripts and ORFs have been a part of genome biology from the first sequenced RNA and DNA-based genomes[2,165]; however, their abundance and ubiquity have only just come into focus for eukaryotic genomes with the advent of recent genome-scale measurement technologies. From past and present literature, it seems clear that the definitions and assessments of overlap topology between eukaryotic, prokaryotic and viral genomes have been disconnected. It is unclear how this discordance arose; however, differing genome

Background – essential requirements for a gene

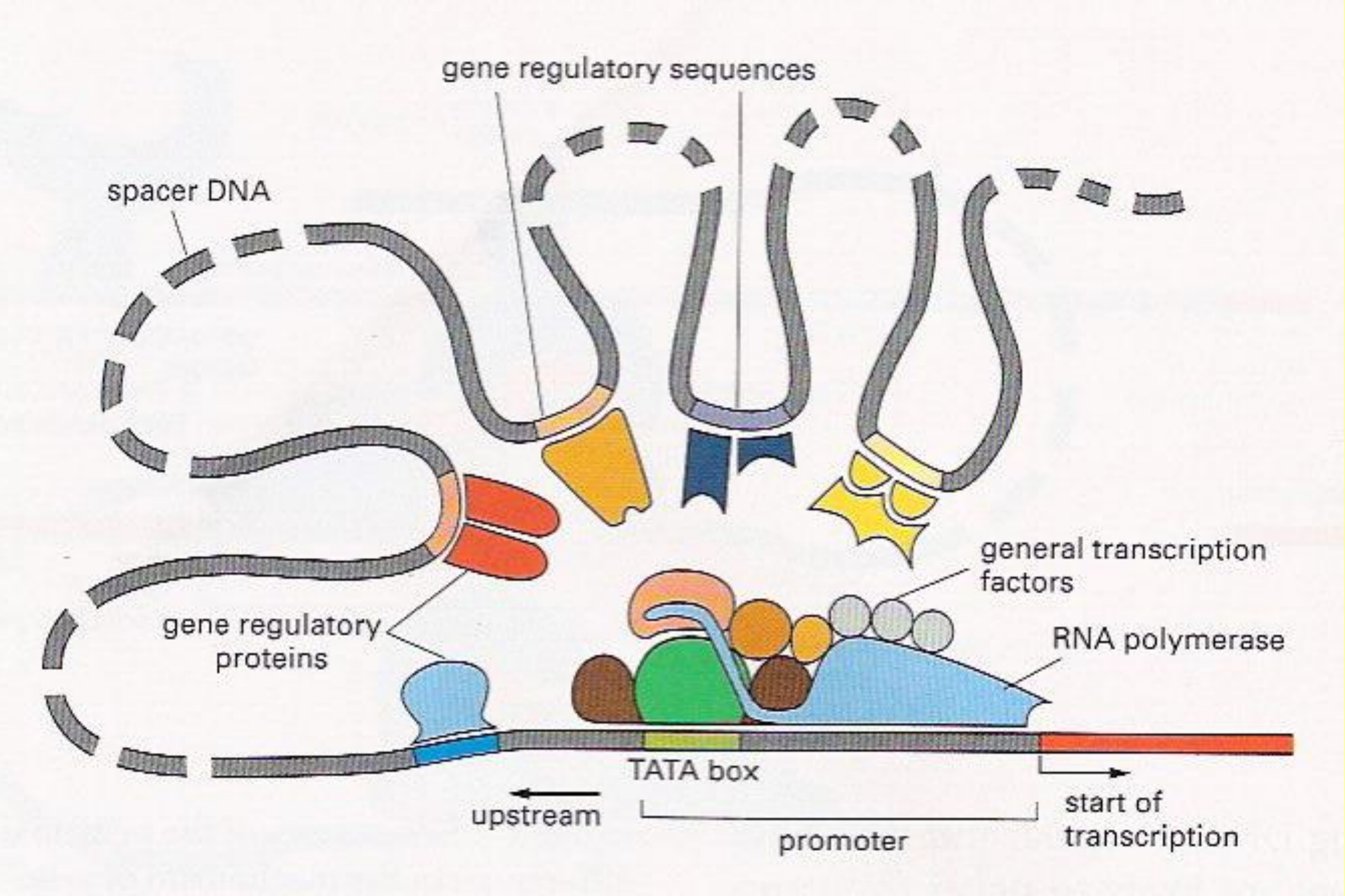




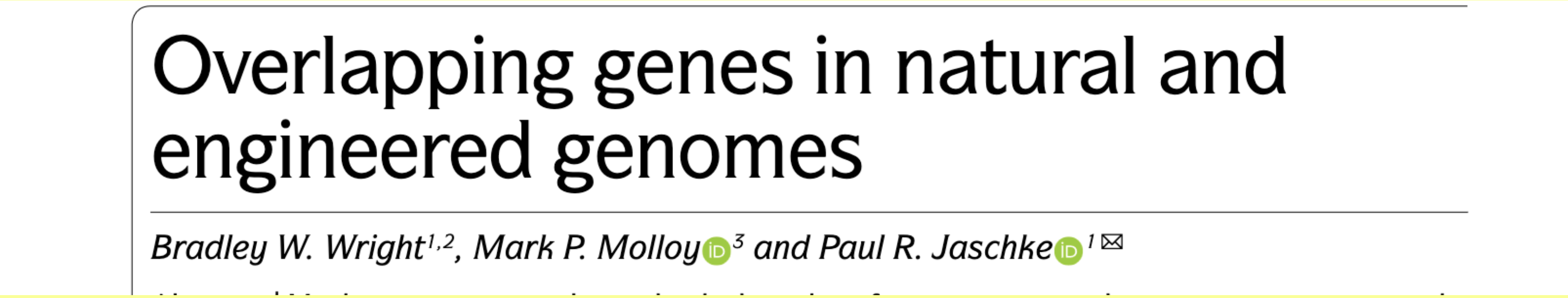
First Base	Second Base								Third Base
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU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

Background – essential requirements for a gene

"promotor"
or
"enhancer"
sequences

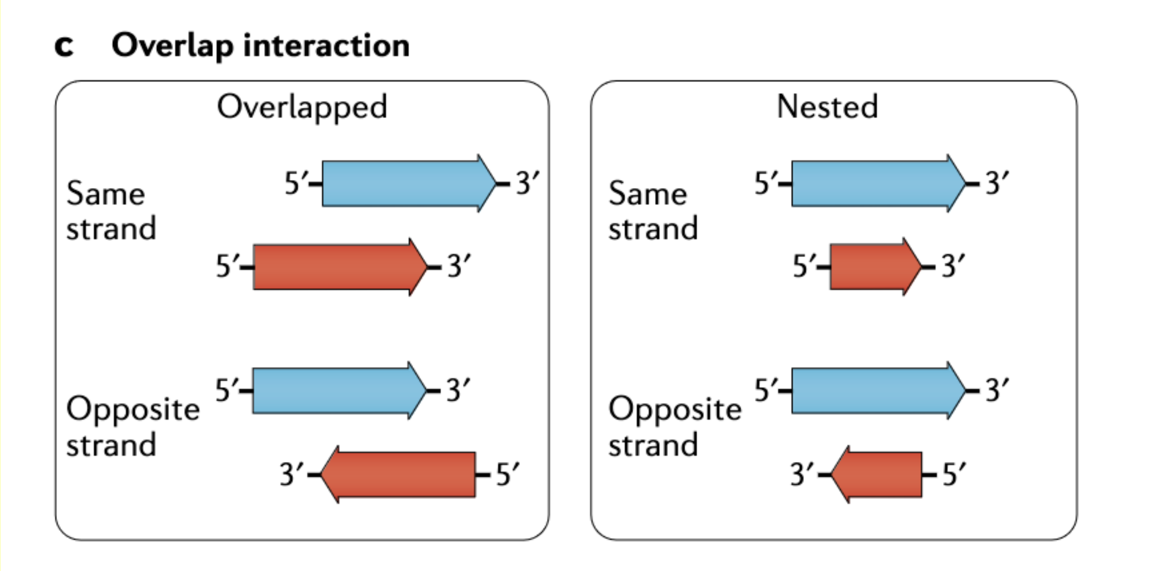


from "Essential Cell Biology"
Alberts et al



Overlapping genes in natural and engineered genomes

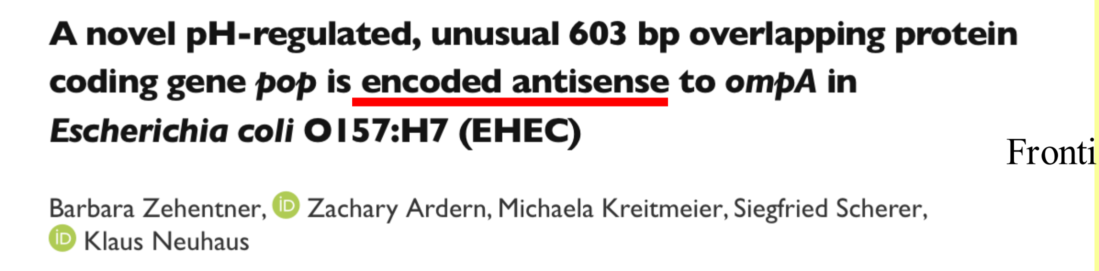
Bradley W. Wright[1,2], Mark P. Molloy[3] and Paul R. Jaschke[1] ✉



c Overlap interaction

NATURE REVIEWS | GENETICS VOLUME 23 | MARCH 2022 | 155

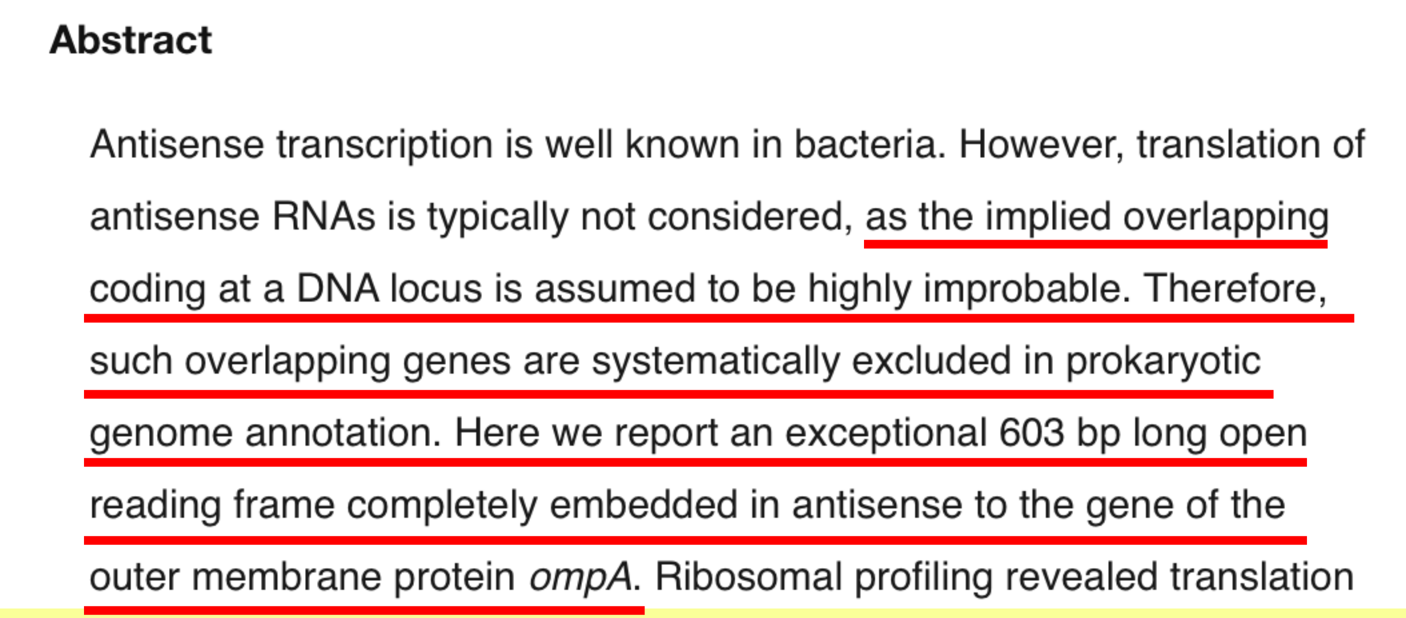
Example:



A novel pH-regulated, unusual 603 bp overlapping protein coding gene *pop* is encoded antisense to *ompA* in *Escherichia coli* O157:H7 (EHEC)

Barbara Zehentner, Zachary Ardern, Michaela Kreitmeier, Siegfried Scherer, Klaus Neuhaus

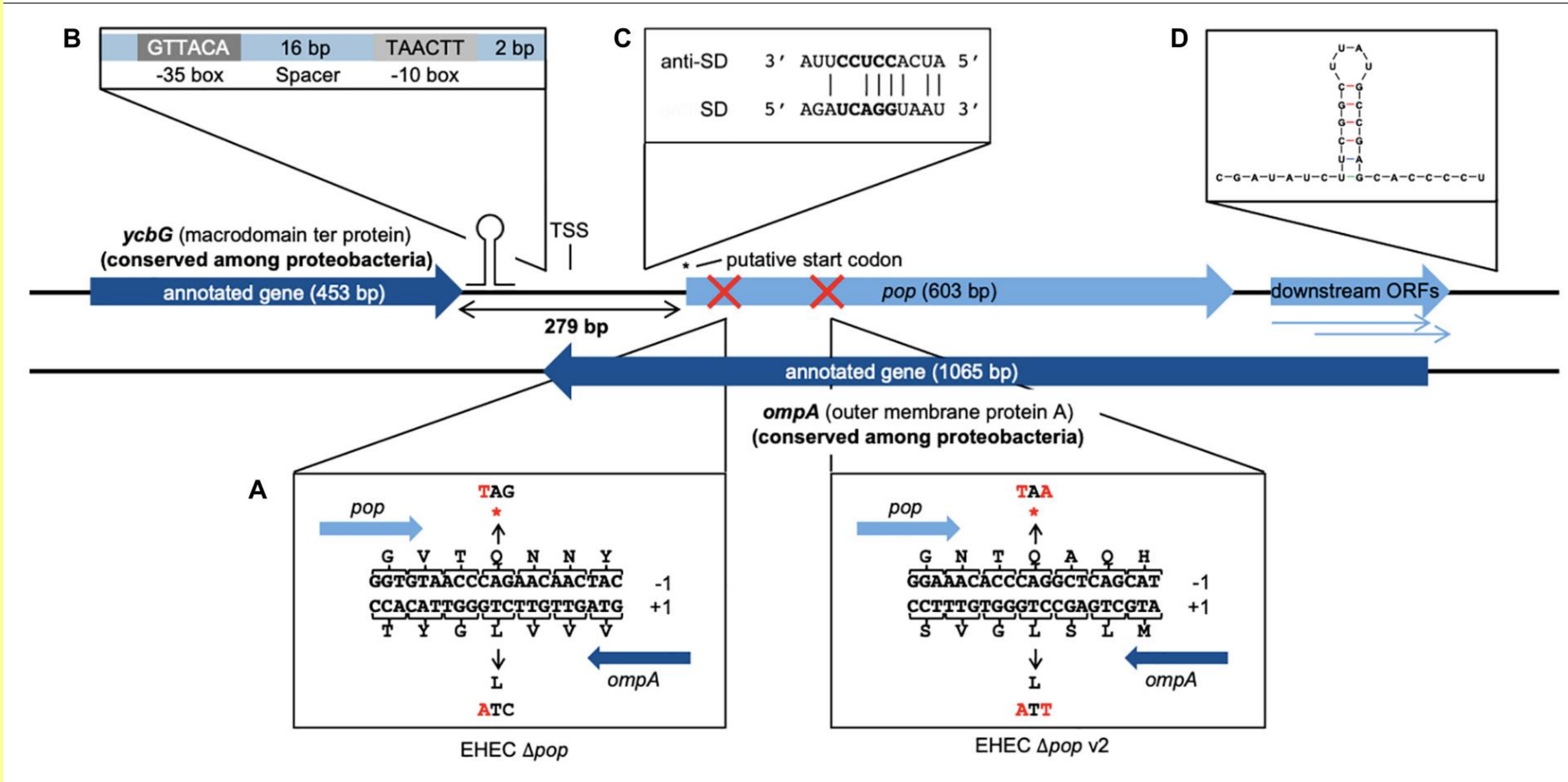
Frontiers in Microbiology
2020, 11, 377



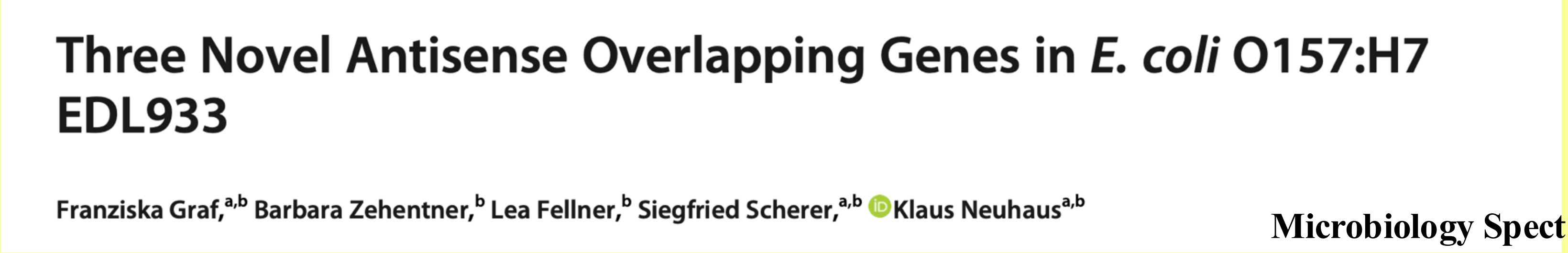
Abstract

Antisense transcription is well known in bacteria. However, translation of antisense RNAs is typically not considered, as the implied overlapping coding at a DNA locus is assumed to be highly improbable. Therefore, such overlapping genes are systematically excluded in prokaryotic genome annotation. Here we report an exceptional 603 bp long open reading frame completely embedded in antisense to the gene of the outer membrane protein *ompA*. Ribosomal profiling revealed translation

Example:


B
GTTACA
16 bp
TAACTT
2 bp
-35 box
Spacer
-10 box
C
anti-SD 3′ AUUCCUCCACUA 5′
SD 5′ AGAUCAGGUAAU 3′
D
U-A
U U
C-G
G-C
G-C
C-G
U-A
C-G-A-U-A-U-C-U-G-C-A-C-C-C-C-U
TSS
ycbG (macrodomain ter protein)
(conserved among proteobacteria)
annotated gene (453 bp)
279 bp
* putative start codon
pop (603 bp)
downstream ORFs
annotated gene (1065 bp)
ompA (outer membrane protein A)
(conserved among proteobacteria)
A
pop
TAG
*
G V T Q N N Y
GGTGTAACCCAGAACAACTAC -1
CCACATTGGGTCTTGTTGATG +1
T Y G L V V V
L
ATC
ompA
EHEC Δpop
pop
TAA
*
G N T Q A Q H
GGAAACACCCAGGCTCAGCAT -1
CCTTTGTGGGTCCGAGTCGTA +1
S V G L S L M
L
ATT
ompA
EHEC Δpop v2

Example:

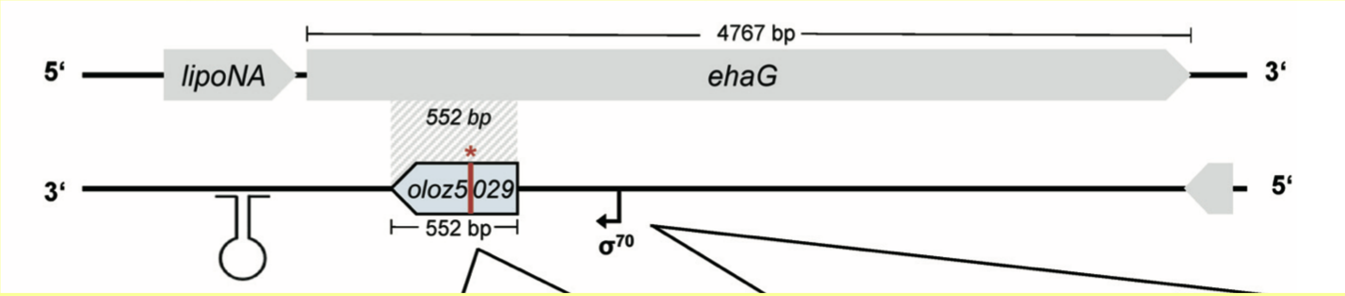


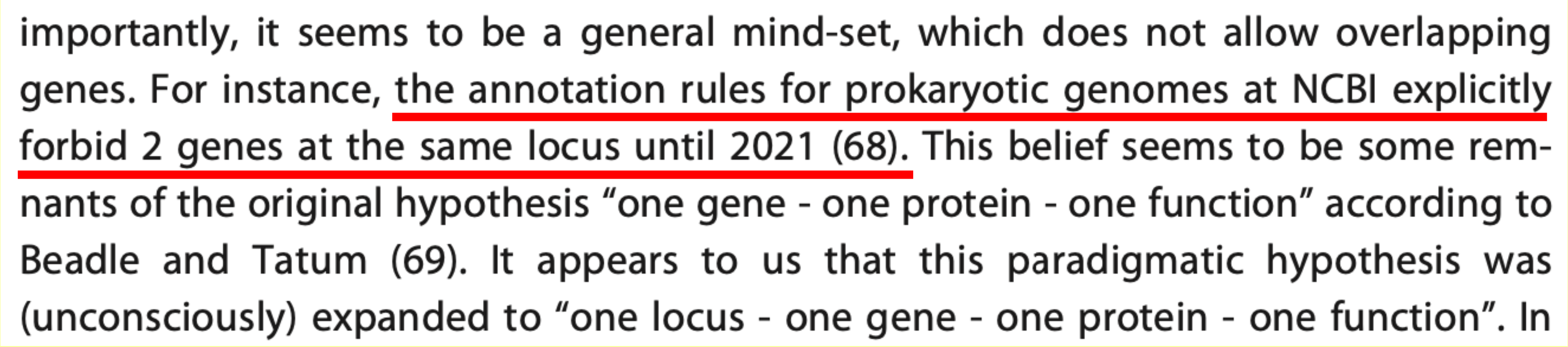
Three Novel Antisense Overlapping Genes in *E. coli* O157:H7 EDL933

Franziska Graf,[a,b] Barbara Zehentner,[b] Lea Fellner,[b] Siegfried Scherer,[a,b] Klaus Neuhaus[a,b]

Microbiology Spectrum
2023, 11, e0235122

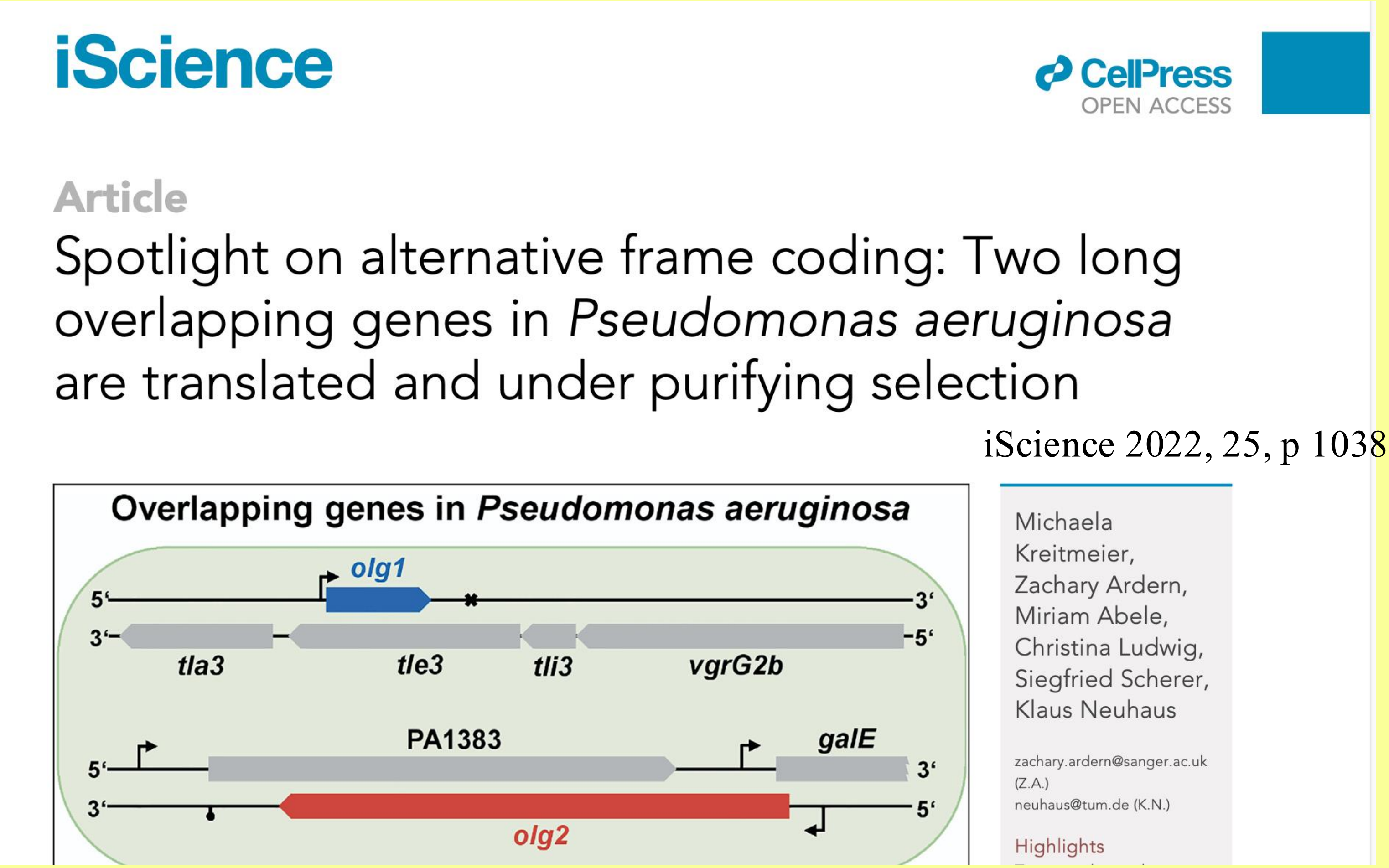
Length: (177 nt, 303 nt, and 552 nt) genes, antisense





importantly, it seems to be a general mind-set, which does not allow overlapping genes. For instance, the annotation rules for prokaryotic genomes at NCBI explicitly forbid 2 genes at the same locus until 2021 (68). This belief seems to be some remnants of the original hypothesis "one gene - one protein - one function" according to Beadle and Tatum (69). It appears to us that this paradigmatic hypothesis was (unconsciously) expanded to "one locus - one gene - one protein - one function". In

Example:



iScience

CellPress OPEN ACCESS

Article

Spotlight on alternative frame coding: Two long overlapping genes in *Pseudomonas aeruginosa* are translated and under purifying selection

Overlapping genes in *Pseudomonas aeruginosa*

Michaela Kreitmeier, Zachary Ardern, Miriam Abele, Christina Ludwig, Siegfried Scherer, Klaus Neuhaus

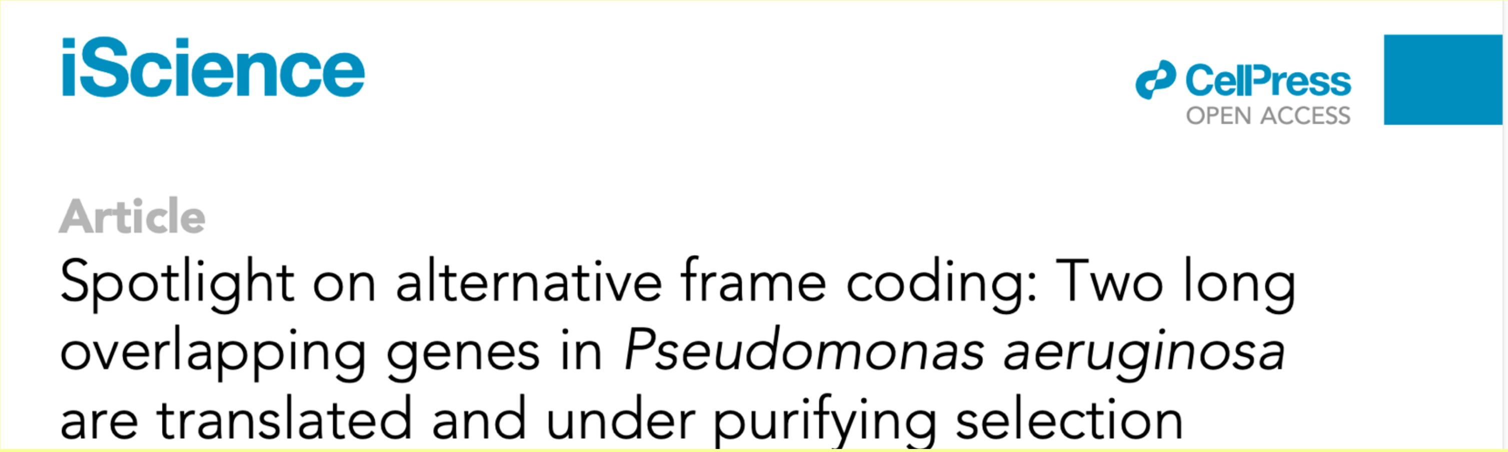
zachary.ardern@sanger.ac.uk (Z.A.)
neuhaus@tum.de (K.N.)

Highlights

iScience 2022, 25, p 103844

Exceptionally long (957nt and 1536 nt) genes, antisense

Example:



iScience

CellPress
OPEN ACCESS

Article

Spotlight on alternative frame coding: Two long overlapping genes in *Pseudomonas aeruginosa* are translated and under purifying selection

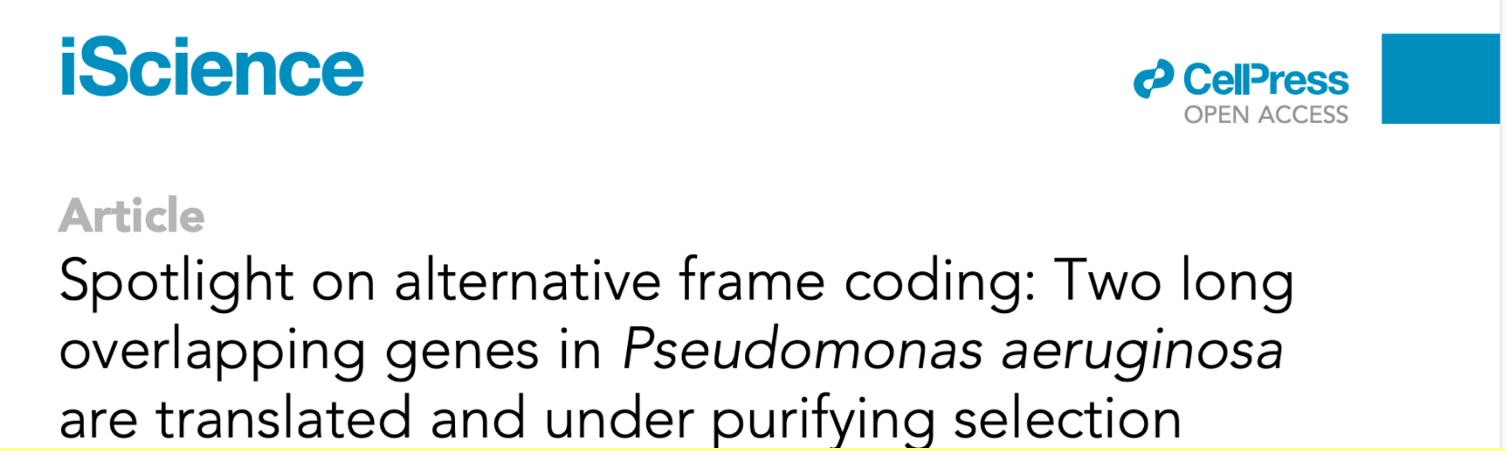
iScience 2022, 25, p 103844

Each has a promoter sequence, start and stop codons

Transcription and translation were verified

Expression is regulated during growth

Example:



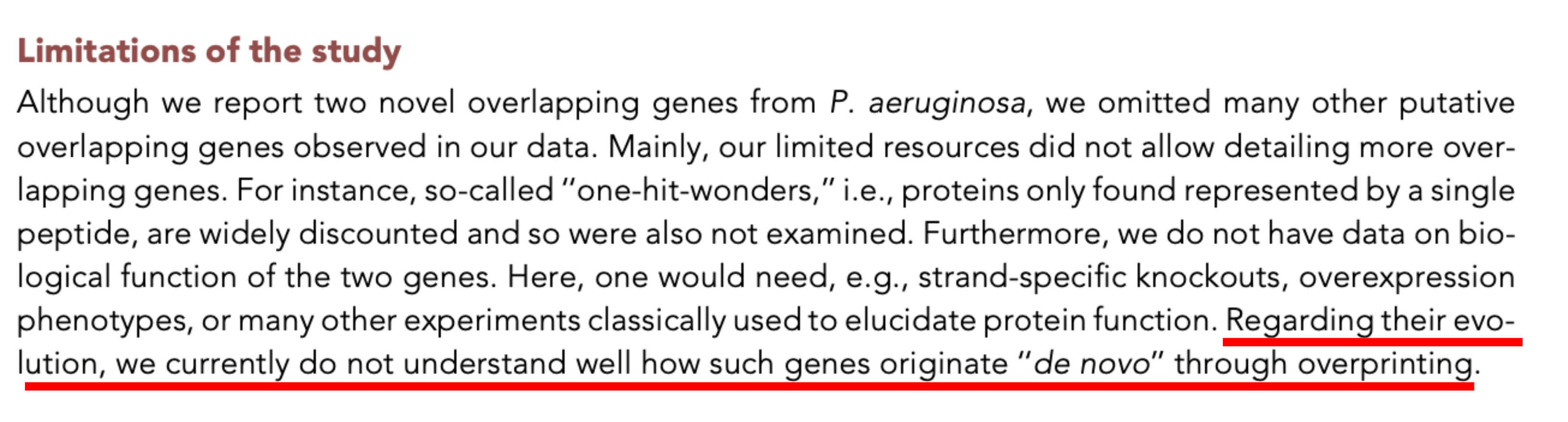
iScience

CellPress OPEN ACCESS

Article

Spotlight on alternative frame coding: Two long overlapping genes in *Pseudomonas aeruginosa* are translated and under purifying selection

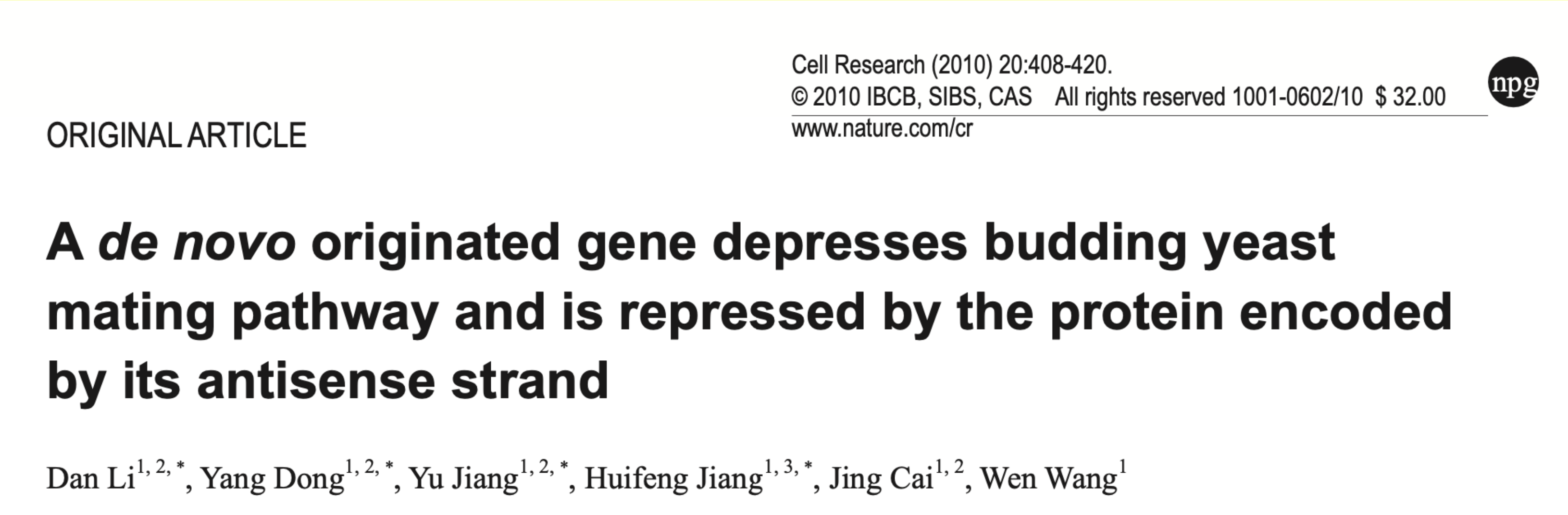
iScience 2022, 25, p 103844



Limitations of the study

Although we report two novel overlapping genes from *P. aeruginosa*, we omitted many other putative overlapping genes observed in our data. Mainly, our limited resources did not allow detailing more overlapping genes. For instance, so-called "one-hit-wonders," i.e., proteins only found represented by a single peptide, are widely discounted and so were also not examined. Furthermore, we do not have data on biological function of the two genes. Here, one would need, e.g., strand-specific knockouts, overexpression phenotypes, or many other experiments classically used to elucidate protein function. Regarding their evolution, we currently do not understand well how such genes originate "*de novo*" through overprinting.

Example:



Cell Research (2010) 20:408-420.
© 2010 IBCB, SIBS, CAS All rights reserved 1001-0602/10 $ 32.00
www.nature.com/cr

ORIGINAL ARTICLE

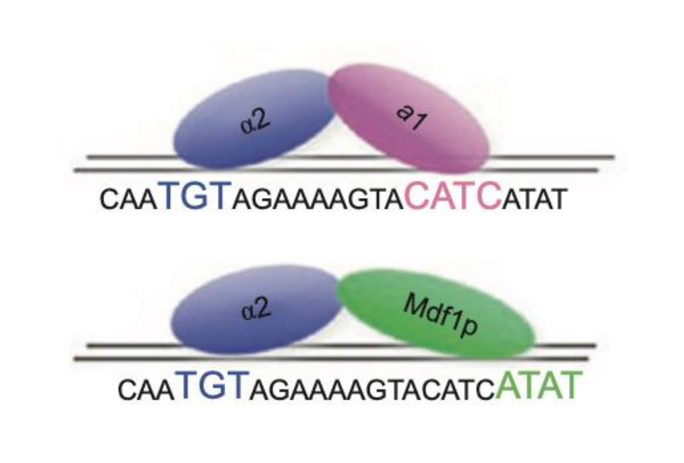
A *de novo* originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand

Dan Li[1, 2, *], Yang Dong[1, 2, *], Yu Jiang[1, 2, *], Huifeng Jiang[1, 3, *], Jing Cai[1, 2], Wen Wang[1]

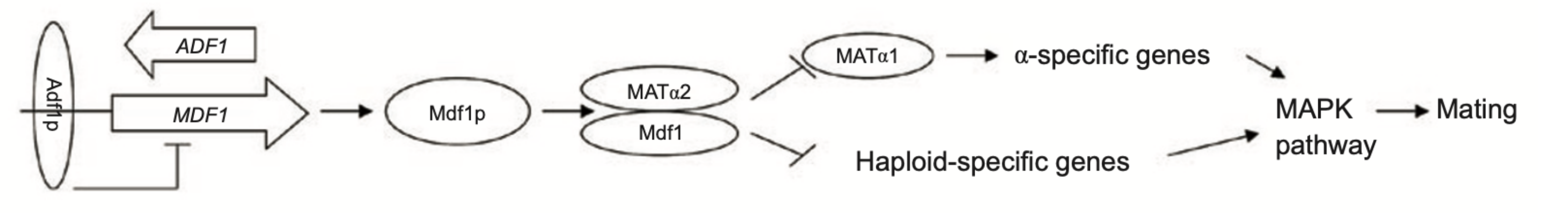
Example:

Mdf1p binds to Matα2p and the complex binds to DNA to suppress genes in the mating pathway

model for the function of Mdf1p



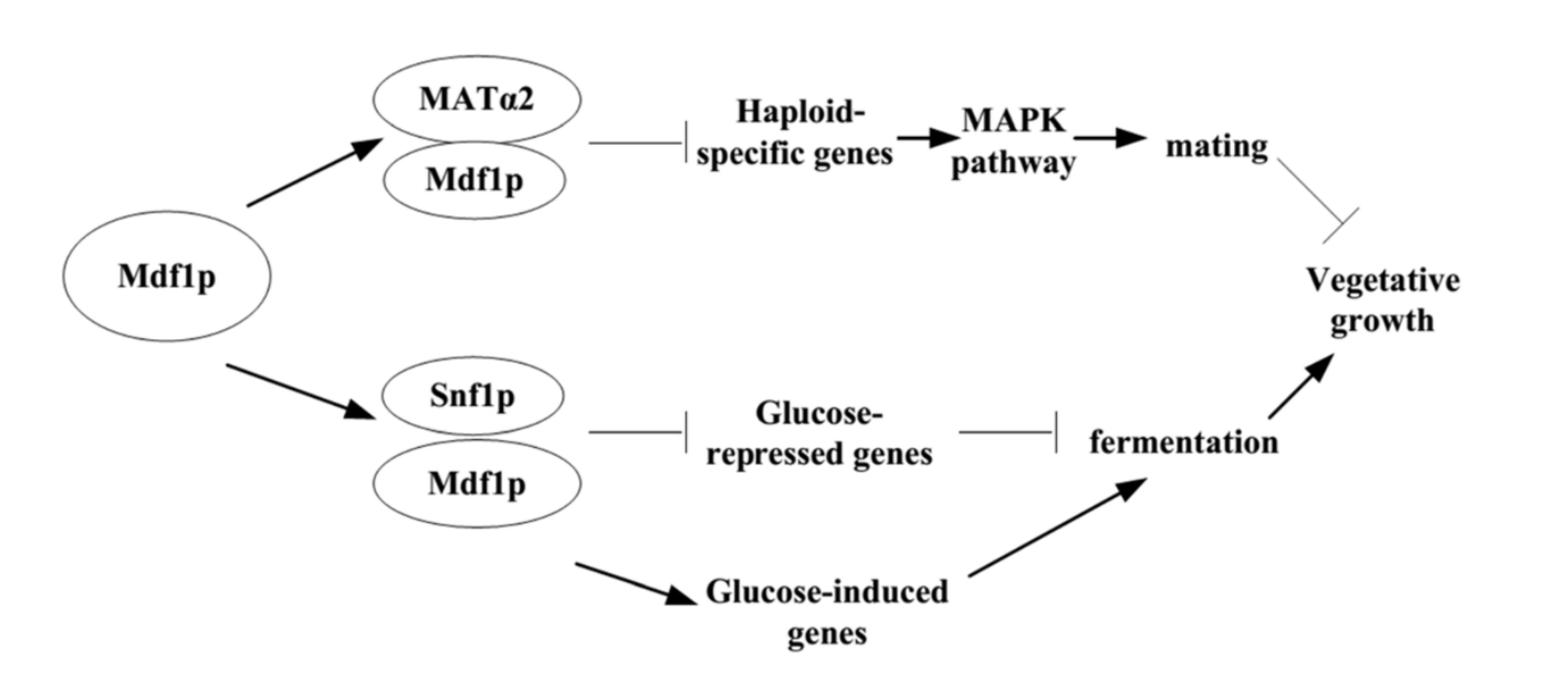
from Li et al
Cell Research 2010



The MDF1 gene is regulated by the ADF1 gene that **overlaps antisense to MDF1**

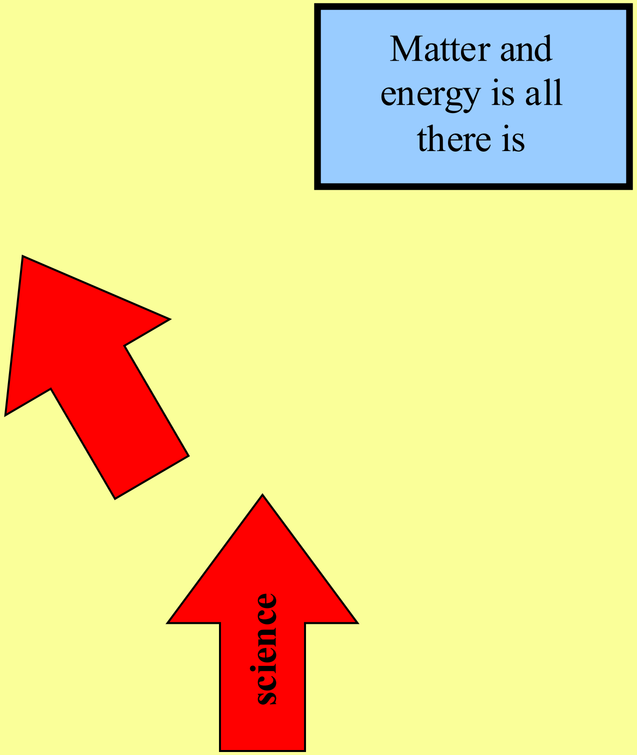
Example:

Mdf1p also promotes vegetative growth by binding to Snf1p blocking de-repression of glucose-repressed genes



from Li et al., Scientific Reports 2014

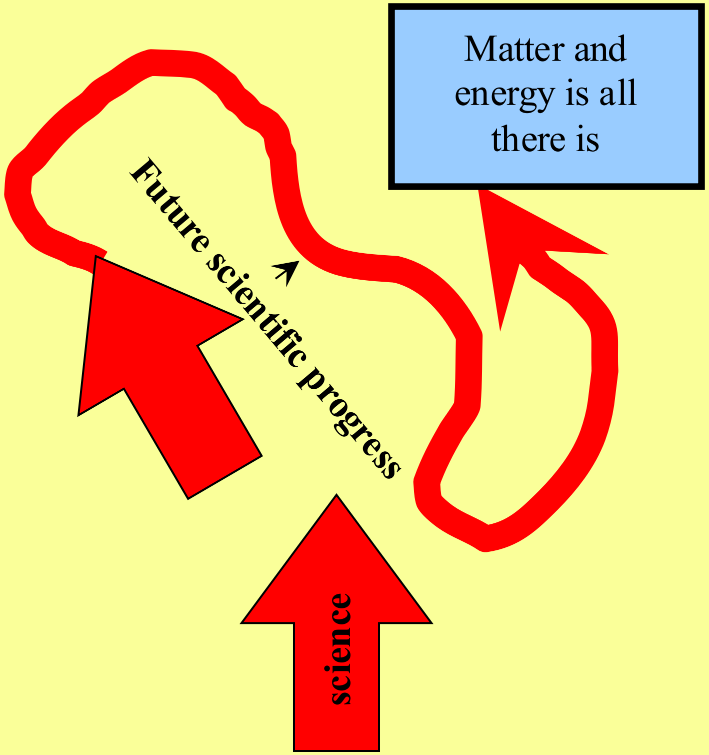
Do extensively overlapping genes represent an informational discontinuity?



Matter and energy is all there is

science

OR



Matter and energy is all there is

Future scientific progress

science

An important question:

Do the two overlapping genes code for proteins with functions that require sophisticated folds and/or rare sequences?