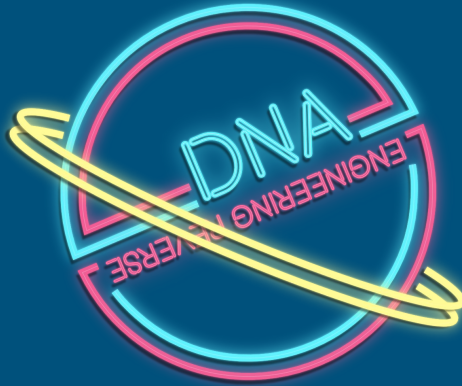
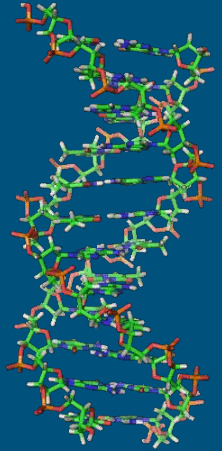
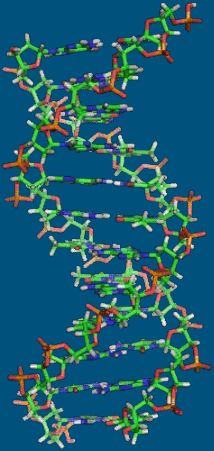
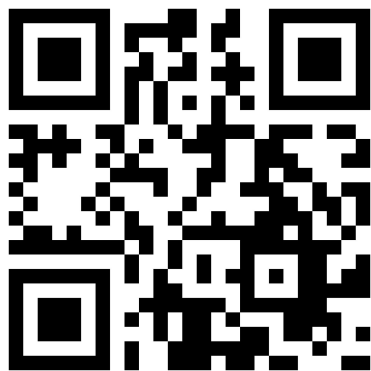


Reverse Engineering Life: What we can learn from the DNA



Credits

- This presentation leans heavily on other people's work and graphics
- All credits are available in the **speaker notes** which you should consult to find out who made all these great movies and images
- **Thank you so much Wikipedia Commons in particular!**



<https://berthub.eu/revdna/>

Online questions!

<https://webchat.oftc.net/?channels=why2025-andromeda>

IRC: oftc.net, channel:
#why2025-andromeda



The ISME Journal
Multidisciplinary Journal of Microbial Ecology

[Login](#) [Cart](#)

Search [go](#) [Advanced search](#)

[Journal home](#) > [Archive](#) > [Original Articles](#) > [Full text](#)

- Journal home**
- Advance online publication**
 - [About AOP](#)
- Current issue**
- Archive**
- Focuses**
- Browse by subject**
- Press releases**
- [Online submission](#)
- [For authors](#)
- [For referees](#)
- [Contact editorial office](#)
- [About the journal](#)
 - [Editors and Editorial Board](#)
- [About the society](#)
- [For librarians](#)
- [Subscribe](#)

Original Article

The ISME Journal (2016) **10**, 30–38; doi:10.1038/ismej.2015.107; published online 3 July 2015

Density-dependent adaptive resistance allows swimming bacteria to colonize an antibiotic gradient

Felix J H Hol¹, Bert Hubert¹, Cees Dekker¹ and Juan E Keymer^{1,2,3}

¹Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, Delft, The Netherlands

²Department of Ecology, Faculty of Biological Sciences, P. Catholic University of Chile, Santiago, Chile

³Institute of Physics, Faculty of Physics, P. Catholic University of Chile, Santiago, Chile

Correspondence: FJH Hol or JE Keymer, Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, Lorentzweg 1, Delft 2628CJ, The Netherlands. E-mail: f.j.h.hol@tudelft.nl or jkeymer@uc.cl

Received 5 January 2015; Revised 9 April 2015; Accepted 19 May 2015

Advance online publication 3 July 2015

Abstract

[Top](#)

During antibiotic treatment, antibiotic concentration gradients develop. Little is known regarding the effects of antibiotic gradients on populations of nonresistant bacteria. Using a microfluidic device, we show that high-density motile *Escherichia coli* populations composed of nonresistant

FULL TEXT

◀ Previous | Next ▶

Table of contents

Download PDF
 Share this article
 View interactive PDF in ReadCube
 Rights and permissions
 CrossRef lists 5 article s citing this article
 Scopus lists 1 article citing this article

▶ Abstract
 ▶ Introduction
 ▶ Materials and methods
 ▶ Results and Discussion
 ▶ Conflict of interest
 ▶ References
 ▶ Acknowledgements
 ▶ Figures and Tables
 ▶ Supplementary info
 ▶ Export citation

[nature](#) > [scientific data](#) > [data descriptors](#) > article

Data Descriptor | [Open access](#) | Published: 22 March 2022

SkewDB, a comprehensive database of GC and 10 other skews for over 30,000 chromosomes and plasmids

[Bert Hubert](#) 

[Scientific Data](#) **9**, Article number: 92 (2022) | [Cite this article](#)

6123 Accesses | **13** Citations | **4** Altmetric | [Metrics](#)

Download PDF

Sections [Figures](#)

[Abstract](#)

[Background & Summary](#)

[Methods](#)

[Data Records](#)



“Imagine a flashy spaceship lands in your backyard. The door opens and you are invited to investigate everything to see what you can learn. The technology is clearly millions of years beyond what we can make.

This is biology.”

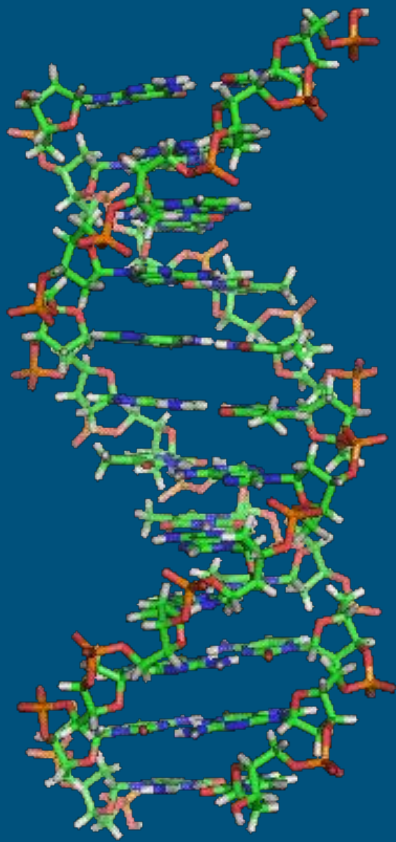




Although...

© NASA (oddly enough)

- Let's study DNA the way we study random binary blobs
- Highlight many cool DNA things
- I want YOU to Join in!



- DNA: Millions, billions of nucleotides or “bases”:
 - A, C, G, T
- Organized in chromosomes & genes
- Absolutely **atom for atom** universal across all life
- >4 billion years old

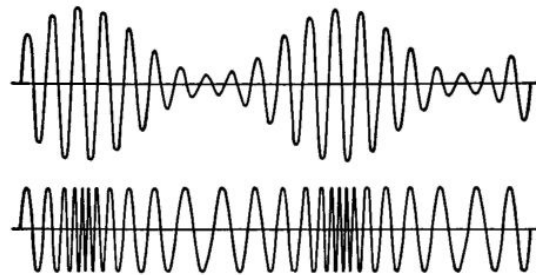
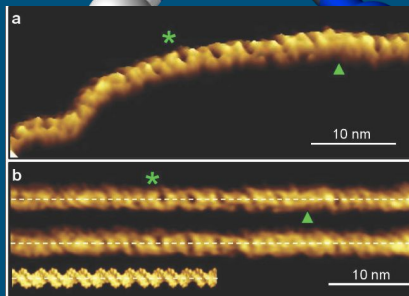
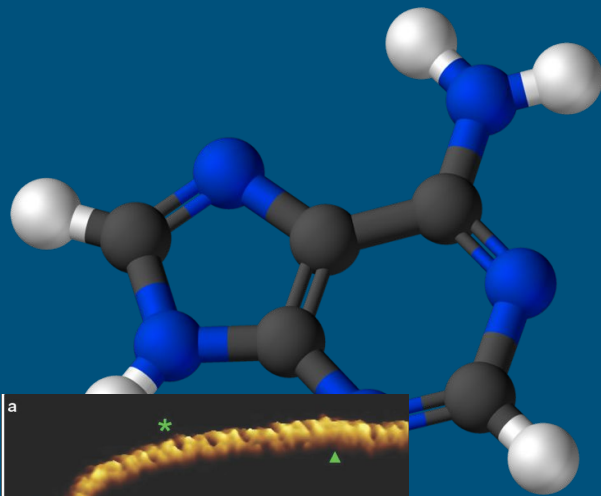
Basics: DNA

A

C

G

T

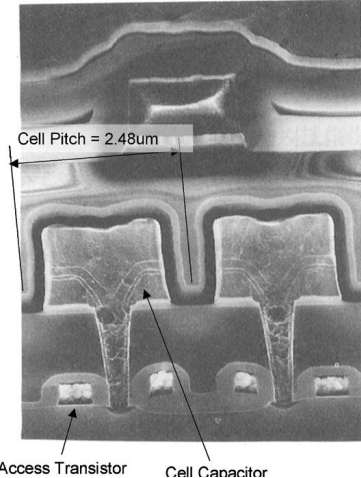


“00”

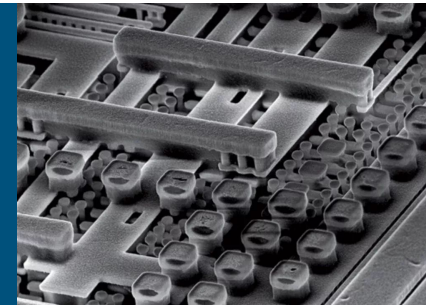
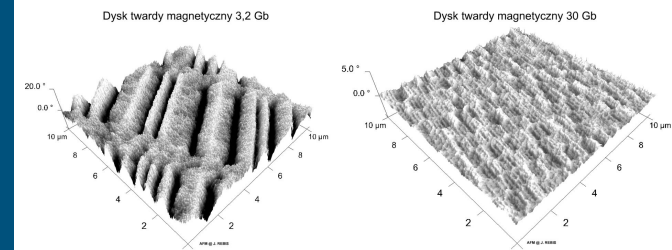
“01”

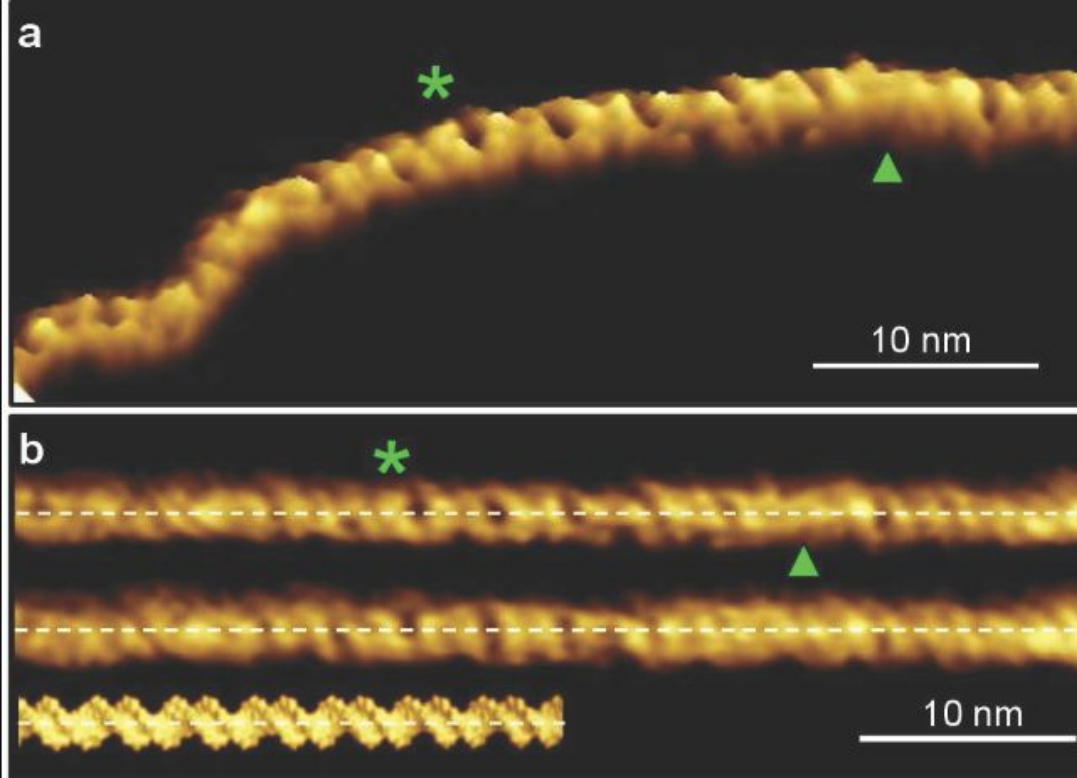
“10”

“11”



MAGNETIC FORCE MICROSCOPY





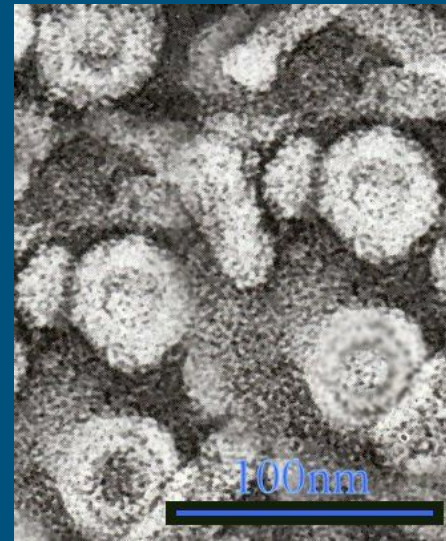
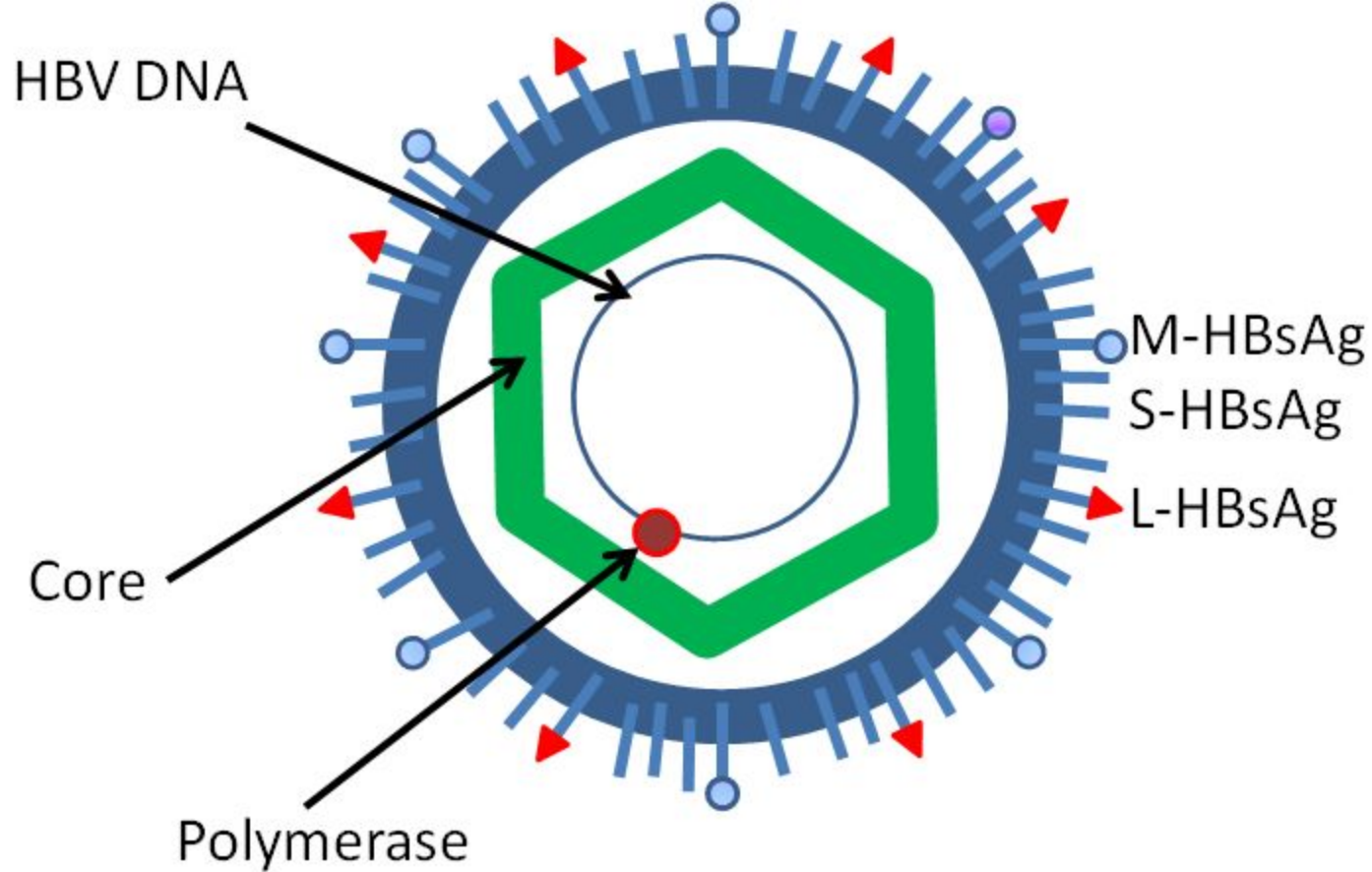
DNA is very much like tape

Sometimes circular tape - no beginning, no end!

No addressing! No alignment!

It is a **nucleotide stream** which can be compared to a **bitstream**

It IS however **content addressable!**



Hepatitis B - 800 bytes (unpleasant)

ctccactgaccttccaccaagctctgtaggatacccaaaagtccaggggctgtatattttcctgctggctggctccaggttcaggaaacacctgctccggaatattgacctctcacatc
tcgtcaatctccgcgaggactggggaccctgtgacgaacatggagaacatcacatcaggatttcttaggacccctgctcgtgttacaggcgggggtttttcttgttgacaagaatcc
tcacaataccgcagagctctagactcgtggaggacttctctcaattttctagggggatcacccgtgtgtcttggccaaaattcgcagtcaccaaacctccaatcactcaccaacctc
ctgtcctccaatttgtcctgggtatcgctggatgtgtctgctggcggttttatcatattcctcttcatcctgctgctatgctcctatcttcttattgggttcttctggattatcaaggt
atgttgcccgtttgtcctctaattccaggatcaacaacaaccagtaggggacctgcaaaacctgcacgactcctgctcaaggcaactctatgtttccctcatgttgcgtgtacaa
aacctacggatggaaattgcacctgtattcccatcccatcgtcctgggctttcgcaaaatacctatgggagtgggcctcagtcctgttctcttggctcagtttacttagtgccatt
tgttcagtgggtcgtagggcgtttccccactgtttggcgttcagctatatggatgatgtggatttggggggccaagtctgtacagcatcgtgagtcctttataaccgctgttacca
attttcttttgtctctgggtatacattttaaaccctaacaaaaacaaaagatgggggttattccctaaacttcaggggttacataattggaagtggggaaactttgccacaggatca
tattgtacaaaagatcaaacactgttttagaaaacttccctgttaacaggcctattgattggaaaagtatgtcaaaagaattgtgggtcgttttgggcgttgcgtcctcatttacaaa
tgtggatatcctgccttaatgcctctgtatgcatgtatacaagctaaacaggcgtttcactttctcgcaacttacaaaggcgtttctaagtaaacagtacatgaacctttaccctcg
ttgctcggcaacggcctggctctgtgccaaagtgtttgctgacgcgaacccccactggctggggcgttggccataggccatcagcgcatgcgtggaacctttgtggctcctctgccgat
ccatactgcggaactccttagccgcttgttttgcctgcagccggctctggagcaaagctcatcggaactgacaattctgtcgtcctctcgcggaataataacatcgtttccatggct
gctaggctgtactgccaaactggatcccttcgcgggacgtcccttgtttacgtccccgtcggcgctgaatccccgcggacgacccccctcgcggggccgcttgggactctctcgtccccct
ctccgtctgccgttccagccgaccacggggcgacacctctctttacgcggctctccccgtctgtgccttctcatctgcgggtccgtgtgcacttcgcttcacctctgcacgttgcat
ggagaccaccgtgaacgcccatacagatcctgcccagggtcttacataagaggactcttggacttccagcaatgtcaacgaccgaccttgaggcctacttcaaagactgtgtgtt
aaggactgggaggagctgggggaggagattaggttaaaggctctttgtattaggaggctgtaggcataaattggctctgcgcaccaacacccatgcaactttttcacctctgccta
catctcttgtacatgtcccactgttcaagcctccaagctgtgccttgggtggcgttggggcatggacattgaccttataaagaatttggagctactgtggagtactctcgttt
ttgccttctgacttctttccctccgtcagagatctcctagacaccgacctcagctctgtatcgagaagccttagagtctcctgagcattgtcacctcaccatactgactcaggc
aagccattctctgctgggggggaattgatgactctagctacctgggtgggtaataatttggaaagatccagcatccagggatctagtagtcaattatgttaataactaacatggggtt
aaagatcaggcaactattgtgggttcatatatcttgccttacttttgggaagagagactgtacttgaatatttgggtctcttccggagtgtggattcgcactcctccagcctataga
ccaccaaatgccctatcttatcaacacttccggaaactactgttgttagacgacgggaccgaggcagggtccccagaagaagaactccccgcctcgcgaacgcagatctcaat
cgccgcgtcgcagaagatctcaatctcgggaatctcaatgttagtattccttggactcataagggtgggaaactttacggggcgtttattcctctacagtagctatctttaatcctg
aatggcaaacctcttcccttccctaagattcatttacaagaggacattattaatagggtgtcaacaatttgtgggcccctcactgtaaaatgaaaagagaagattgaaattaattat
gcctgctagattctatcctacccacactaaatattttcccttagacaaaggaattaaaccttattatccagatcaggtagttaatcattacttccaaaccagacattatttacat
actctttggaaggctgggtattctatataagaggggaaccacacgtagcgcacatatttgcgggtcaccatattcttgggaacaagagctacagcatgggagggttggtcatcaaaa
cctcgcgaaggcatggggacgaatctttctgtttcccaaccctctgggattctttccgatcatcagttggaccctgcattcggagccaaactcaacaatccagattgggacttca
accctcaaggacactggccagcagccaaaggtaggagtgggagcattcgggccagggtcaccctccacacggcggtattttgggggtggagccctcagggtcagggtcag
attgaccacagtgtcaacaattcctcctcctgcctccaccaatcggcagtcaggaaggcagcctactcccatctctccacctctaagagacagtcacctcaggccatgcagtgg
aa

All of Hepatitis-B - 800 bytes

But is life REALLY digital?

Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome

DANIEL G. GIBSON, JOHN I. GLASS, CAROLE LARTIGUE, VLADIMIR N. NOSKOV, RAY-YUAN CHUANG, MIKKEL A. ALGIRE, GWYNEDD A. BENDERS, MICHAEL G. MONTAGUE, LIMA [...], AND J. CRAIG VENTER [+14 authors](#) [Authors Info & Affiliations](#)

SCIENCE • 20 May 2010 • Vol 329, Issue 5987 • pp. 52-56 • DOI:10.1126/science.1190719

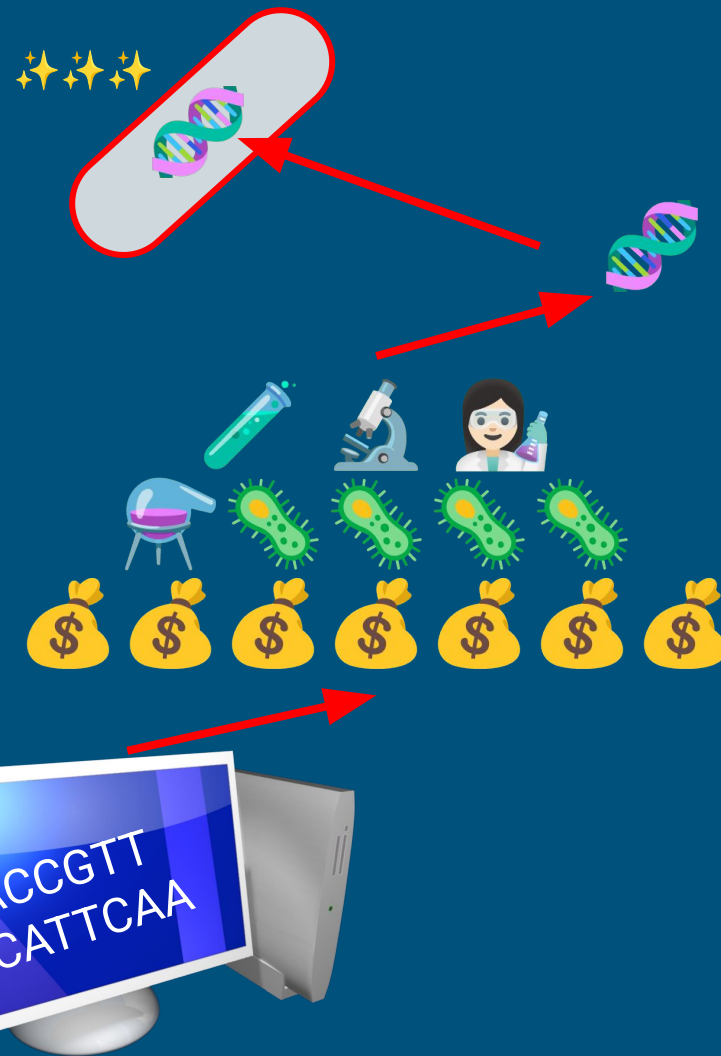
 62,812  1,912



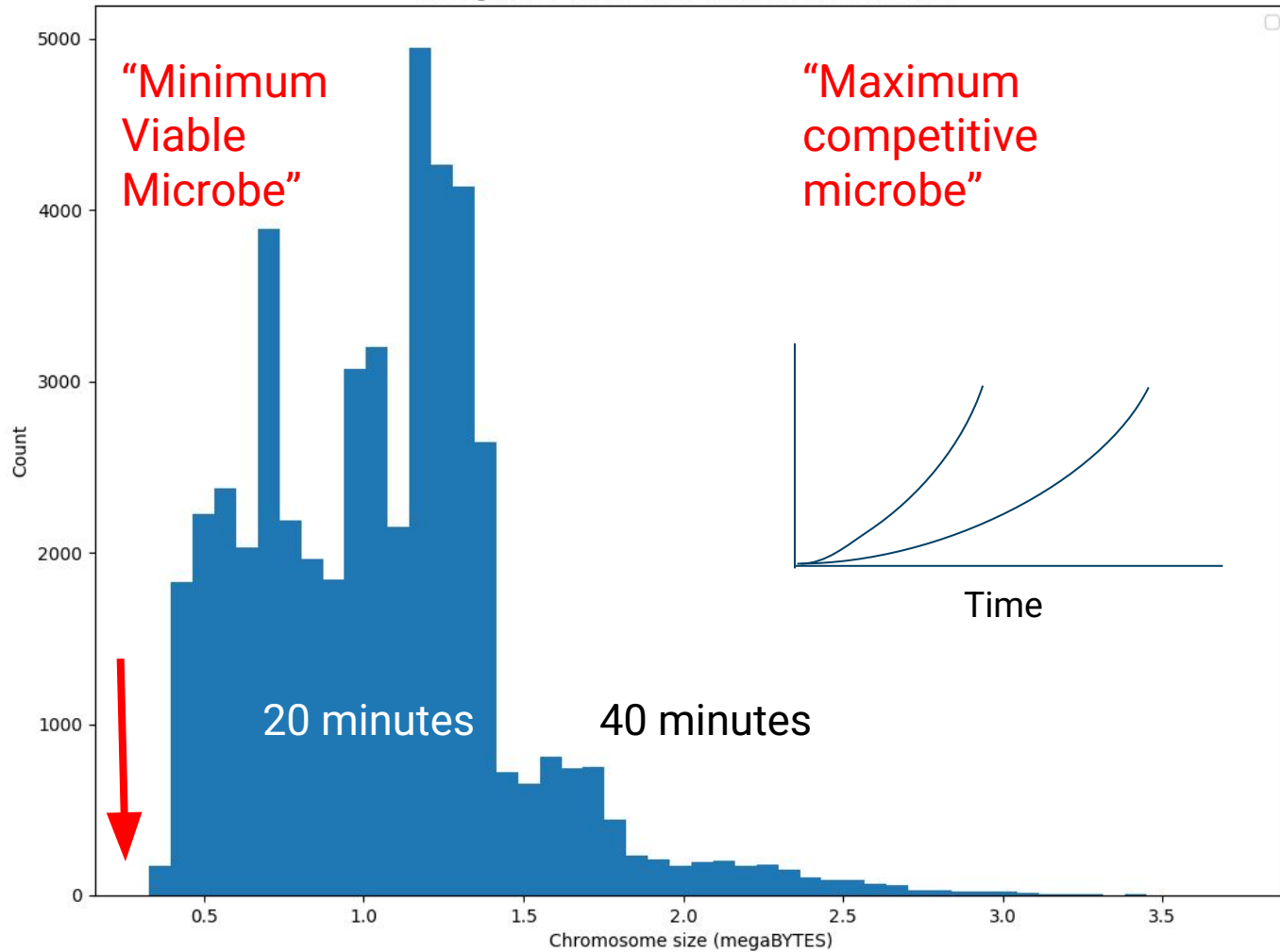
Let There Be Life

The DNA sequence information from thousands of genomes is stored digitally as ones and zeros in computer memory. Now, **Gibson *et al.*** (p. 52, published online 20 May; see the cover; see the Policy Forum by **Cho and Relman**) have brought together technologies from the past 15 years to start from digital information on the genome of *Mycoplasma mycoides* to chemically synthesize the genomic DNA as segments that could then be assembled in yeast and transplanted into the cytoplasm of another organism. A number of methods were also incorporated to facilitate testing and error correction of the synthetic genome segments. The transplanted genome became established in the recipient cell, replacing the recipient genome, which was lost from the cell. The reconstituted cells were able to replicate and form colonies, providing a proof-of-principle for future developments in synthetic biology.

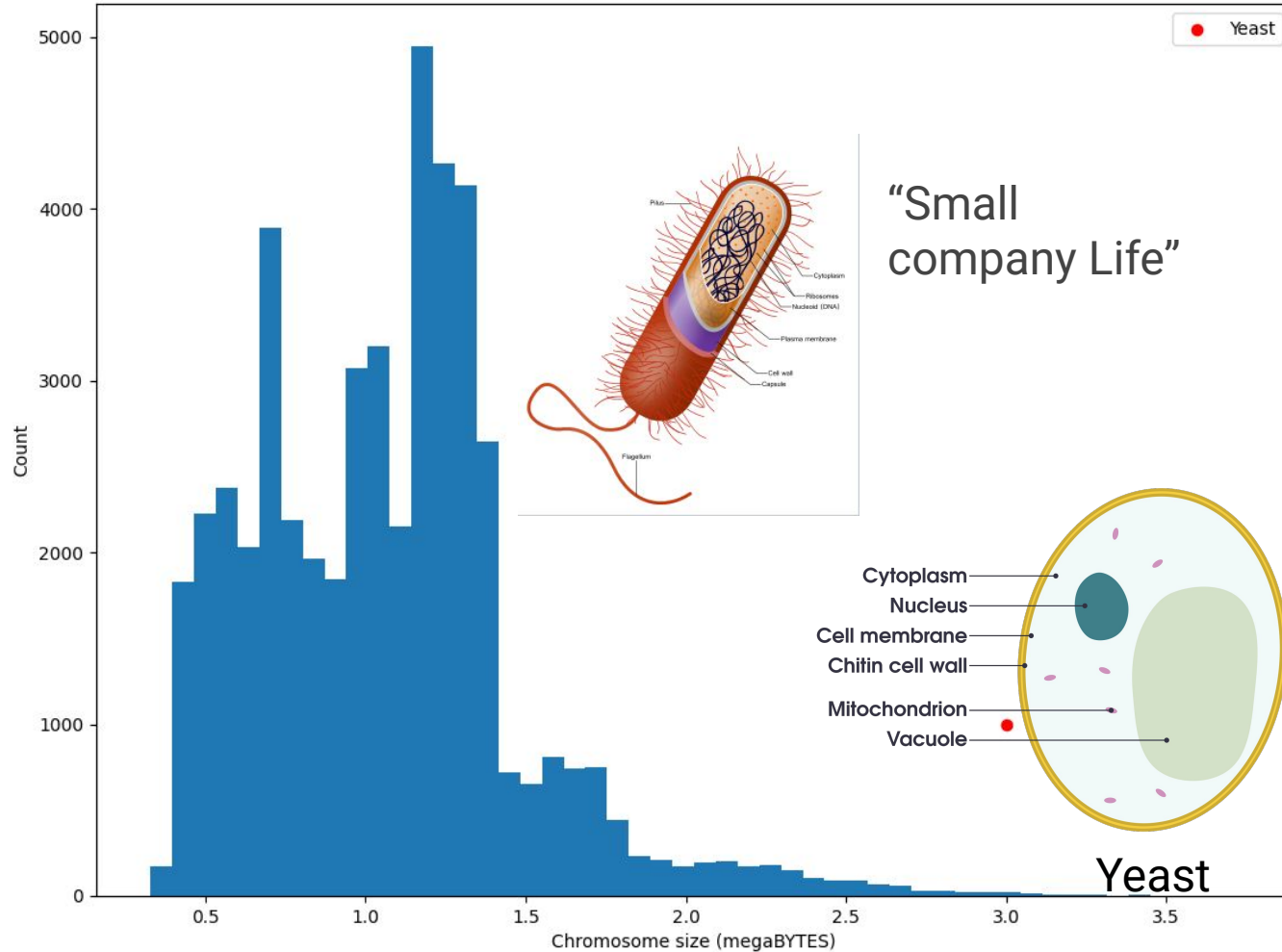
But is life DIGITAL?



Histogram of 49177 BACTERIAL chromosome sizes



Histogram of 49177 BACTERIAL chromosome sizes
Plus select eukaryotes

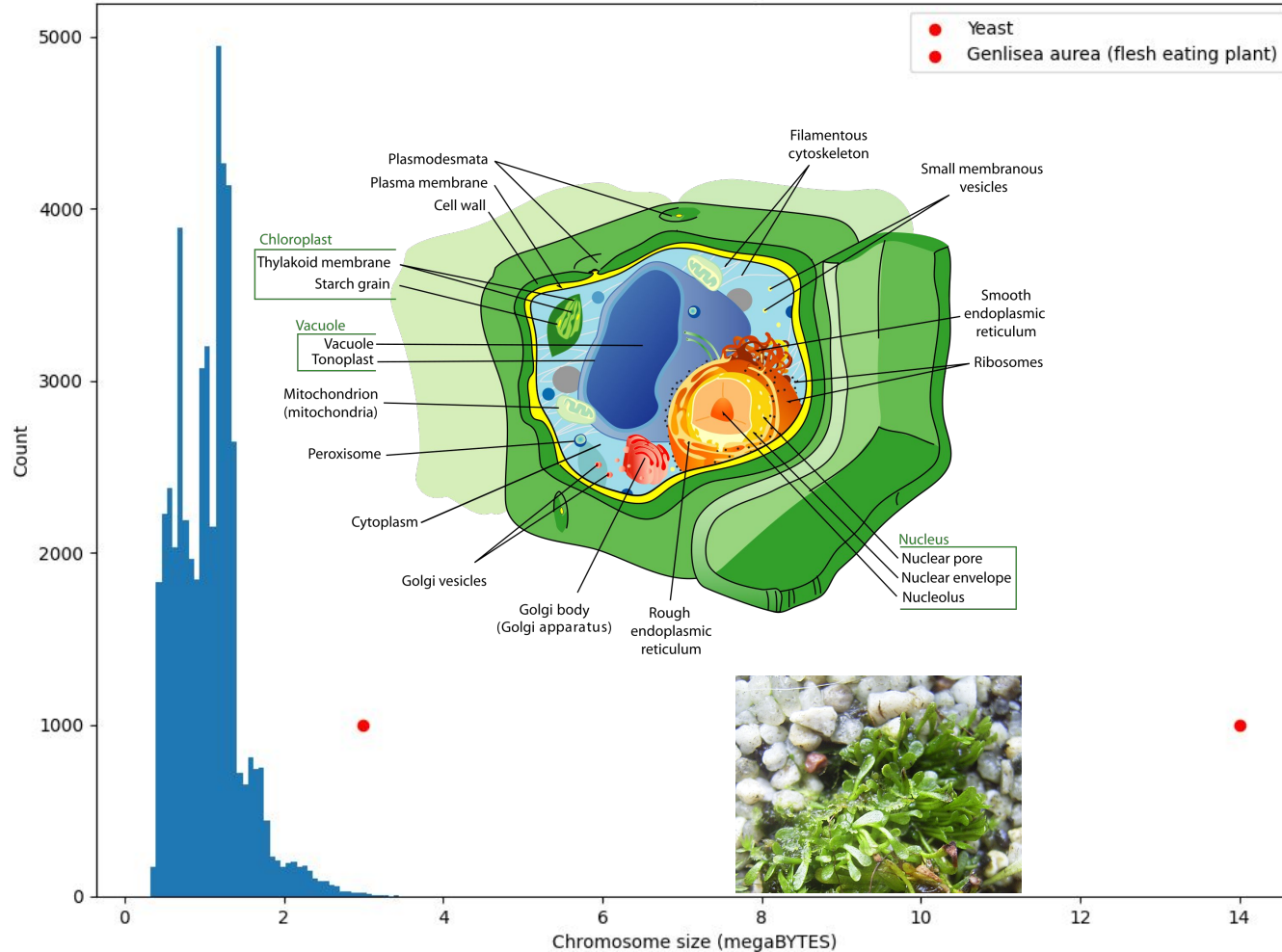


“Small
company Life”

“Enterprise Life”
- various
departments
and procedures

Yeast

Histogram of 49177 BACTERIAL chromosome sizes
Plus select eukaryotes

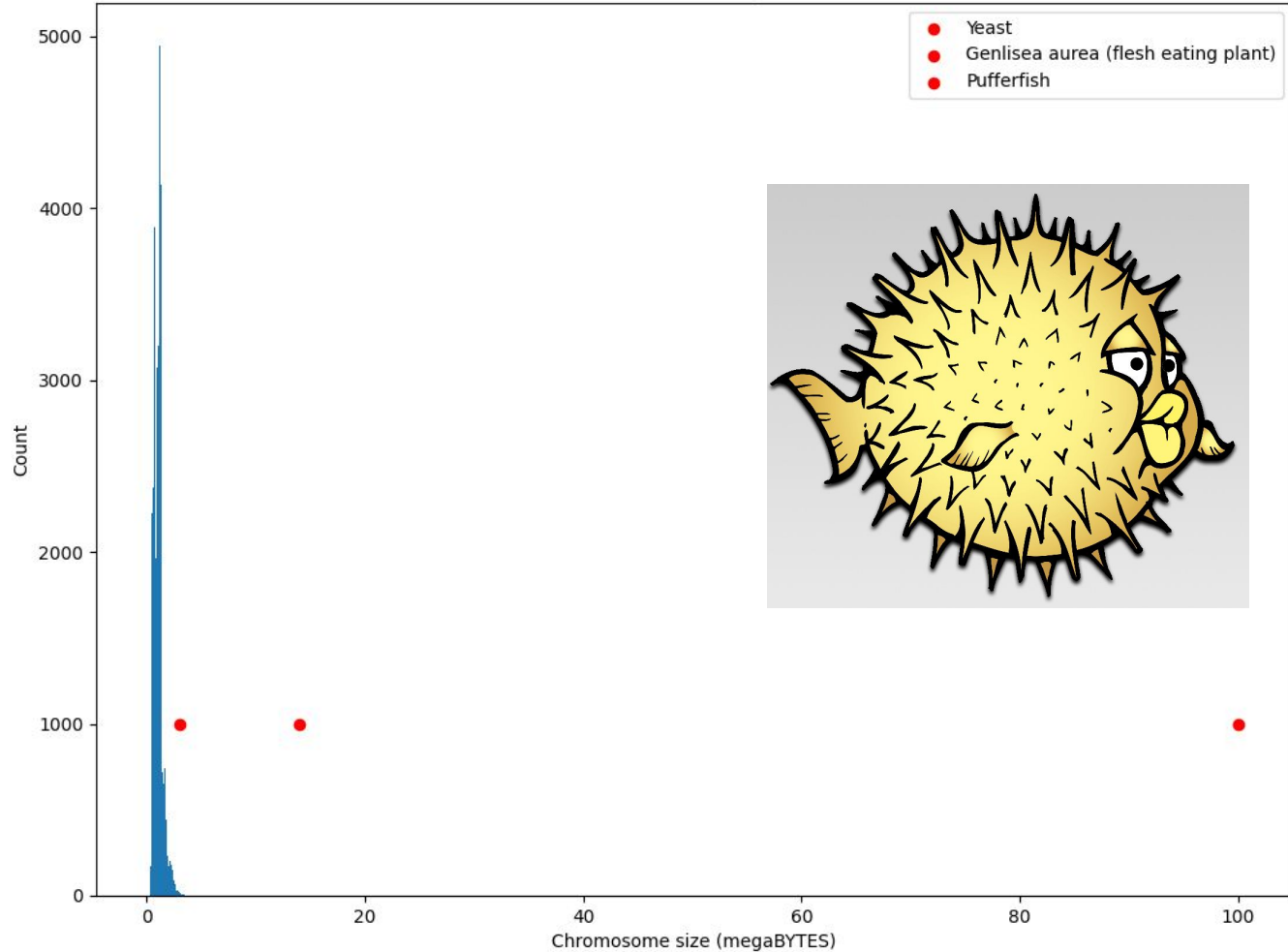


By LadyofHats -
Self-made using Adobe
Illustrator.

<https://commons.wikimedia.org/w/index.php?curid=844682>

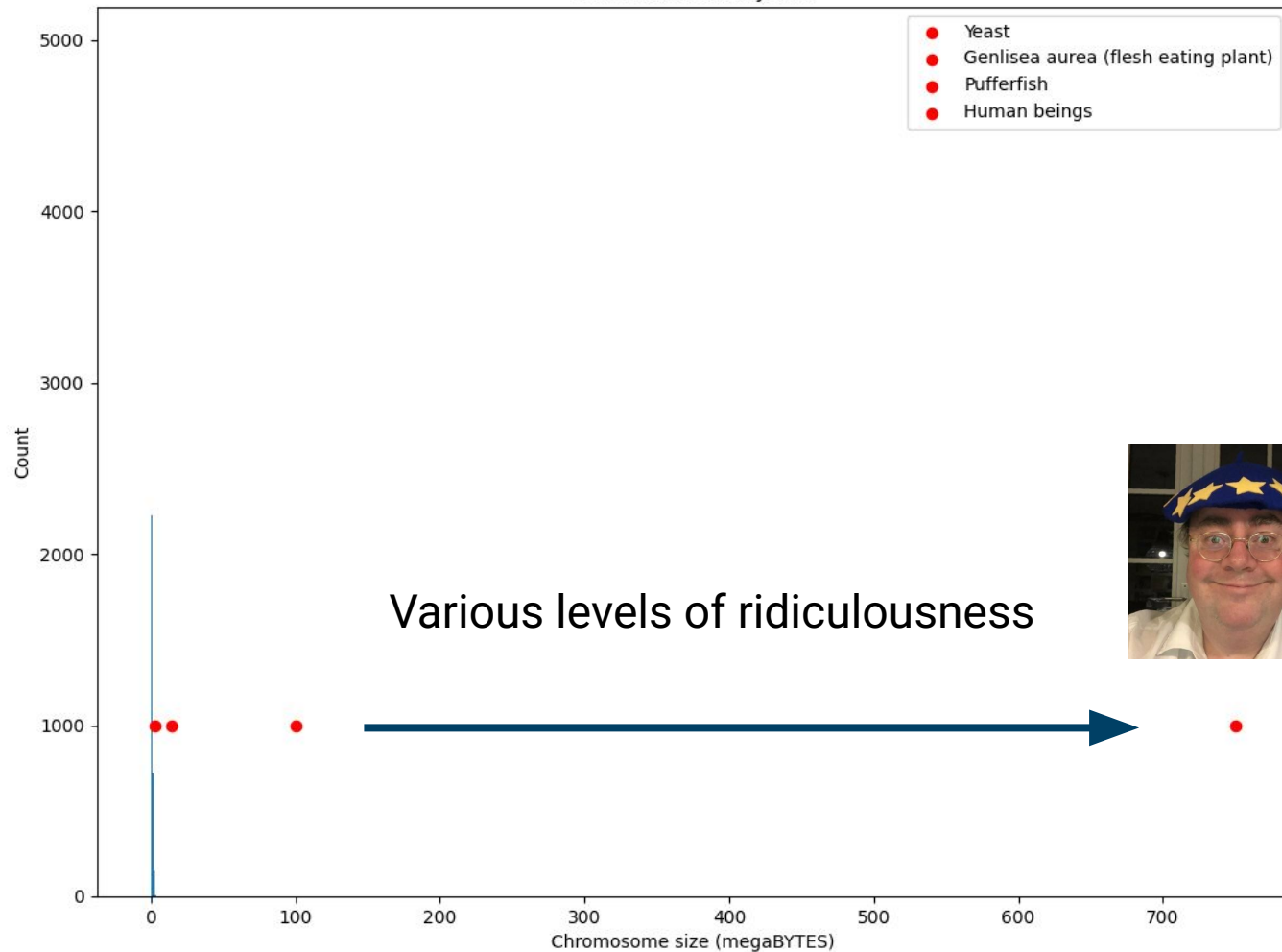
“Mega-corp territory”

Histogram of 49177 BACTERIAL chromosome sizes
Plus select eukaryotes



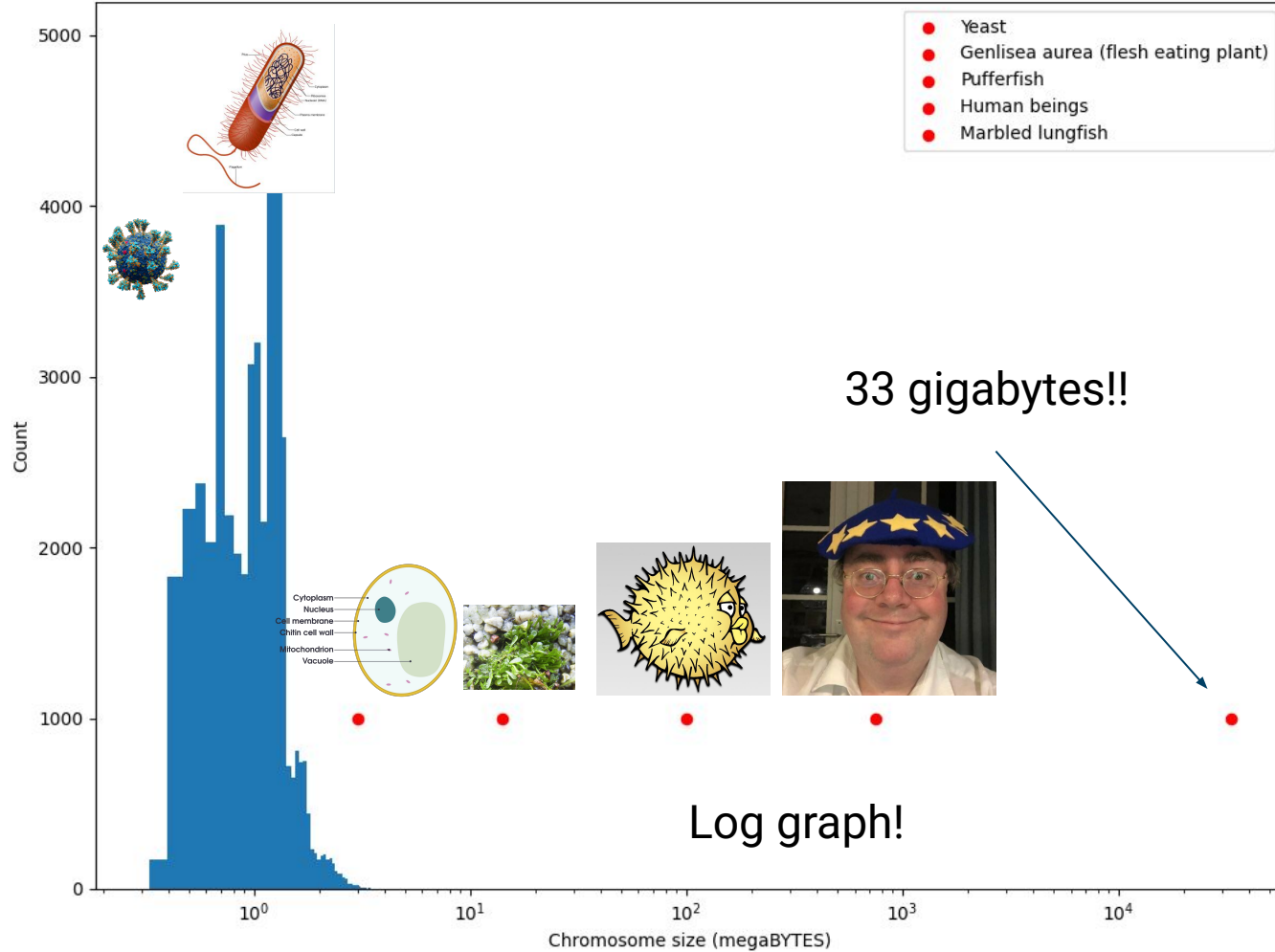
Most compact animal.
Genome almost
completely used for
proteins.

Histogram of 49177 BACTERIAL chromosome sizes
Plus select eukaryotes



“Bloated mega-corp DNA”

Histogram of 49177 BACTERIAL chromosome sizes
Plus select eukaryotes



Unimaginably over the top
bloated megacorp stuff!



Genomes



NCBI's Genome resources include information on large-scale genomics projects, genome sequences and assemblies, and mapped annotations, such as variations, markers and data from epigenomics studies.

How to

[Submit sequence data to NCBI](#)

[Download a complete genome](#)

[Convert feature coordinates between genomic assemblies](#)

[Find an interactive view of a genomic annotation](#)

[more...](#)

Genome Sequences

Genome

information about organisms' genomes

Assembly

genomic assembly statistics

Functional Genomics

GEO DataSets

functional genomics study data

GEO2R

identifies differentially expressed genes

Variation Resources

dbSNP

catalog of short genetic variations

dbVar

genome structural variation studies

Additional Tools

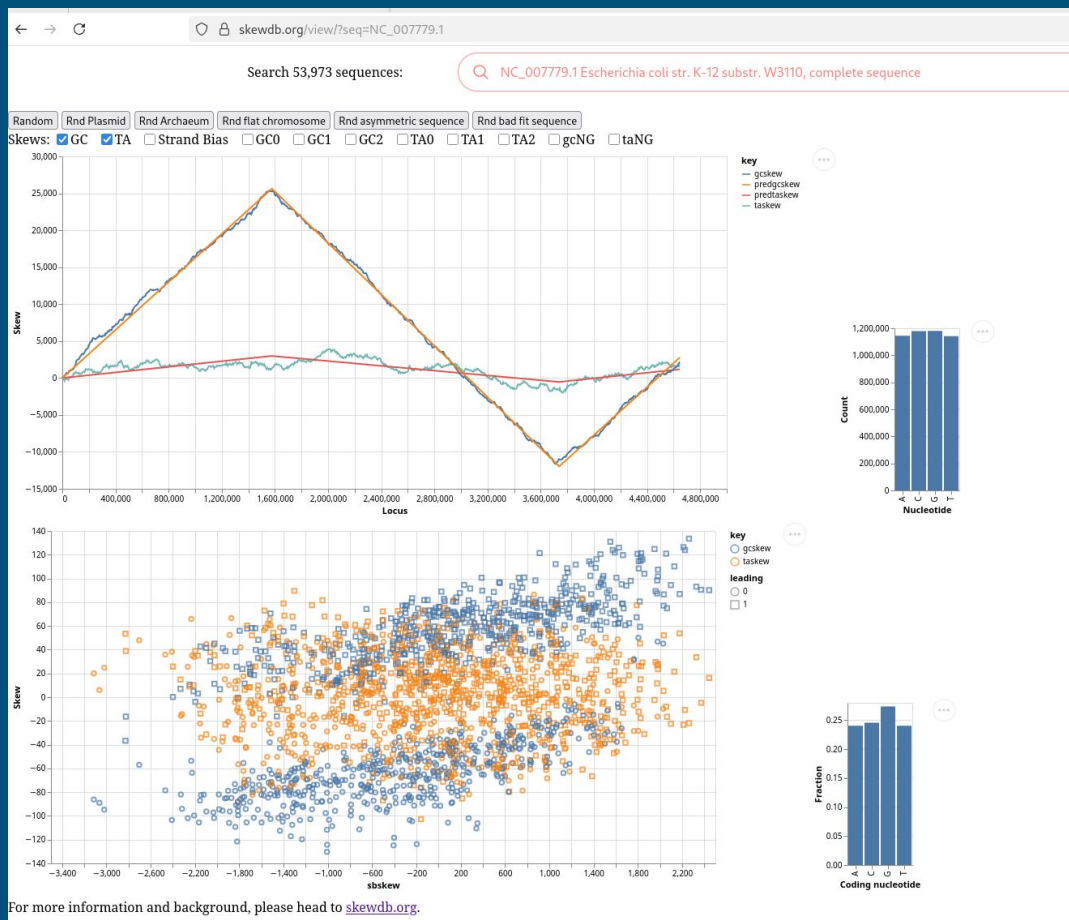
Genome Data Viewer

displays data tracks in an interactive genome browser

Genome Workbench

displays and analyzes sequence data

<https://skewdb.org/> - <https://skewdb.org/view/> - CSV FILES!



scientific data

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [scientific data](#) > [data descriptors](#) > [article](#)

Data Descriptor | [Open access](#) | Published: 22 March 2022

SkewDB, a comprehensive database of GC and 10 other skews for over 30,000 chromosomes and plasmids

[Bert Hubert](#)

[Scientific Data](#) 9, Article number: 92 (2022) | [Cite this article](#)

6590 Accesses | 13 Citations | 5 Altmetric | [Metrics](#)

Abstract

GC skew denotes the relative excess of G nucleotides over C nucleotides on the leading versus the lagging replication strand of eubacteria. While the effect is small, typically around 2.5%, it is robust and pervasive. GC skew and the analogous TA skew are a localized deviation from Chargaff's second parity rule, which states that G and C, and T and A occur with (mostly) equal frequency even within a strand. Different bacterial phyla show different kinds of skew, and differing relations between TA and GC skew. This article introduces an open access database (<https://skewdb.org>) of GC and 10 other skews for over 30,000 chromosomes and plasmids. Further details like codon bias, strand bias, strand lengths and taxonomic data are also included. The SkewDB can be used to generate or verify hypotheses. Since the origins of both the second parity rule and GC skew itself are not yet satisfactorily explained, such a database may enhance our understanding of prokaryotic DNA.

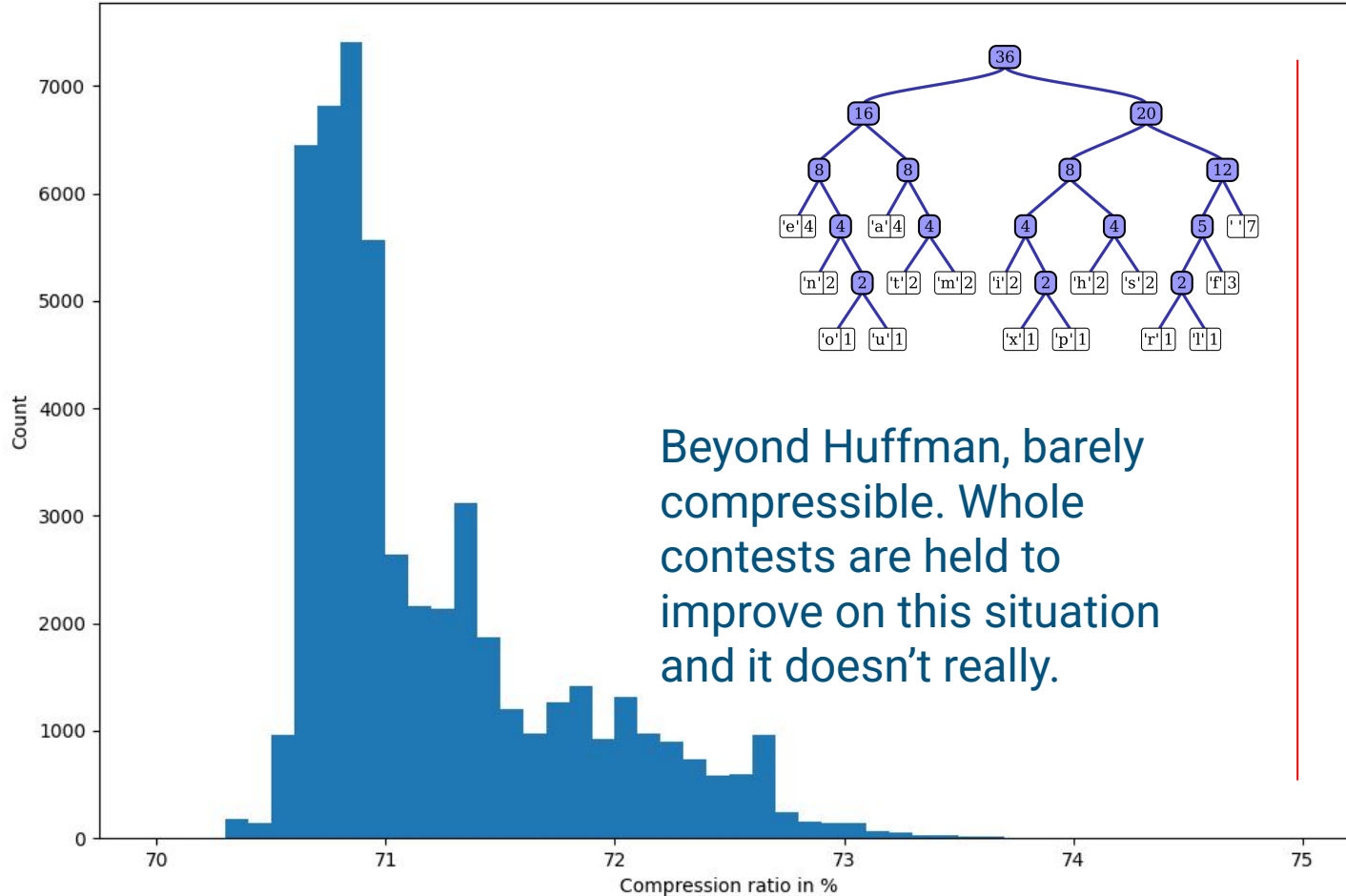
Let's dive into the >50,000 binaries!
(in ASCII)

>NZ_CP150338.1 Sorangium sp. So ce388 chromosome, complete genome
ATGACGATTCCGAAACACGAGCCCCGCGAAGTCTTCGATCGGGCGATCGAGCATACGCGGGCCCTTTCTCCCGCAACTTT
TGATCAGTGGTTTGGGGGAGTTCAGTTCGATGACCTGACCGACGGCGTGCTCACGCTGCGAGTCCAGAACGAGTTCGTCC
TCGAGTGGGTCAGGGACAATTTCTGCCCCGCGCTGACCGACAAGATCCGCGAGATCACGGGCTGGTCGGTCCAGGTGGCG
TGGACGGTGGATCAGCACCTTGAGTCGCCGATCGCGCAGCGCGTCGAGCTCACCCCGGTTTCGCCCCGCGGGCGCTCGTGGT
GCGTCCGACGAGCACGGCGCCGACGCCTGCGCCCCGCCCCGAGCAGCCGCTGCGGGCGCGTCTCGCCGATGCCCGACGACC
TCAACCCGAAGCACACCTTCGCGAGCTTCGTCTGTGGGCCCGTCGAACCAGCTGGCGCACGCGGGCCGCGATCGCGGCCGCG
GGCGGGCGGGGGTCGCCGGTACAACCCGCTCTTCATCTGCGGCGGAACGGGGCTCGGCAAGACCCACCTGATGCACGCGAT
CGCGCATCGCGTCTTCGAGGGCAGGCCGGACGCGCGGATCATCTACGTCTCGGCGGAGAAGTTCACGAACGACTTCATCA
CGGCCATCCAGCACCAACCGGATGGACGACTTCCGCACGAGGTACAGGTCGAGCTGCGACGTGCTGCTCGTCGACGACATC
CAGTTCCTGGCCGGGCGCGAGCAGACGCGAGGAGGAGTTCTTCCACACCTTCAACGCGCTCCACACGCTCGACCGGCAGAT
CGTGGTGACGAGCGACAAGTACCCCCAGAACCTCGAGCGCATGGAGGAGCGCCTCGTCTCGCGTTTCTCGTGGGGGCTCG
TCGCCGACATCCAGGTGCCGGAGCTGGAGACGCGCGTCGCGATCGTCCGGAACAAGGCGGGCGCTCGAGGGCATCGCGCTC
ACGGACGACGTGGCGCTCTACCTCGCGCAGATGGTCCGCTCGAACGTCCGCGAGCTCGAGGGGACGCTGATCCGGCTTGC
GGCCAAGAGCTCGCTCACGGGCCGCCCCGTGGATCTCCCGTTTCGCGCGCGCCGAGATCACGGCCACGTCGCCGCCGCGGG

<https://ftp.ncbi.nlm.nih.gov/genomes/refseq/>

GCF_051474125.1_ASM5147412v1_genomic.fna.gz

gzip compression ratio of 52171 microbial ASCII chromosome files



Beyond Huffman, barely compressible. Whole contests are held to improve on this situation and it doesn't really.

```
>NC_004337.2 Shigella flexneri 2a str. 301 chromosome, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTG
GTTACCTGCCGTGAGTAAATTTAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGAC
...
GACGAAATGCTGAACCAGGGCTGAAGCGTTGGTTTCTTTCCACNNNNNNCAGCTTCAGCCGTTATTGGTGCGATTGTCGT
GCTATTTATCTACAGGAAGATTAAAAGTTAACGCTTAAATTGCACAAAGGCTGCACACAGGCAGCCTTTGCTATTTTTTA
GAGGTGACGTTACCACCATCCCAGCTCAGTGCCGAGTACGACAATAATCGCCACACCCATCATAATAATCGTTGTTTTTT
TCATCTTTATGCTCCCGGGGCAGCATAGCAGCCAATAAAAAACCATGCTAAAAATACGACCGCTGTAATGAAGATTACAA
CCGGAAAAATAATACCGATTTCGCATGTTATCACCAATCAAAGGATGTGAATACGCCTCAGGAATGTAGCGCTGGATGCG
```

“2bit” formats would achieve 75% compression, and allow better further compression, but the field doesn’t care.

>NZ_CP043307.1 *Acinetobacter johnsonii* strain Acsw19
ATGCTTTGGACGGACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGGGAATGTCTTTACAATGT

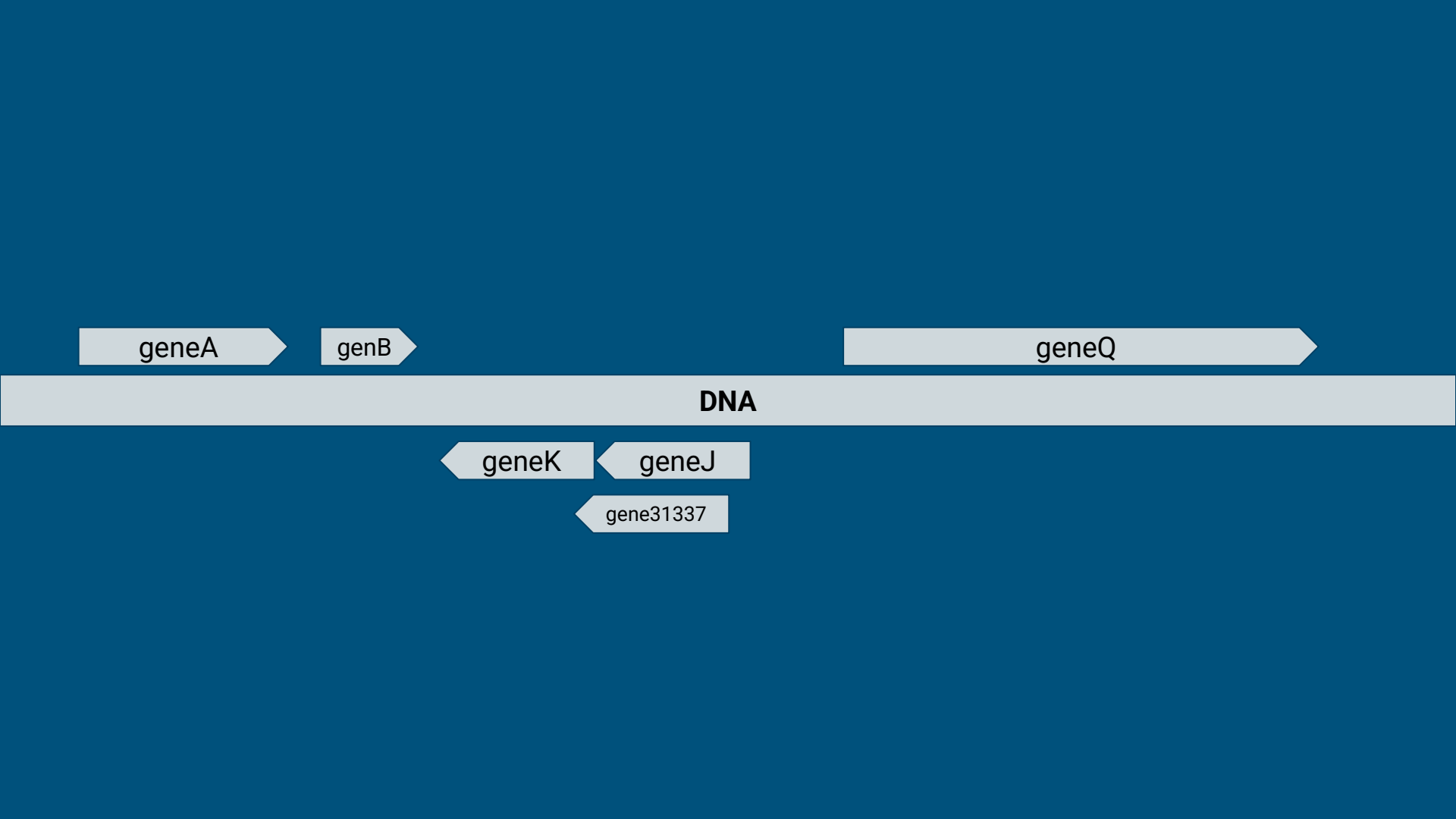
A	C
T	G



ATGCTTTGGACGGACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGGGAATGTCTTTACAATGT
|||||
TACGAAACCTGCCTGACGAATTGAGCGAACGCTGTTCTCGAGAGACCCTTACAGAAATGTTACA



>NZ_CP043307.1.**rev** *Acinetobacter johnsonii* strain Acsw19
ACATTGTAAAGACATTCCCAGAGAGCTCTTGTCGCAAGCGAGTTAAGCAGTCCGTCCAAAGCAT



geneA

genB

geneQ

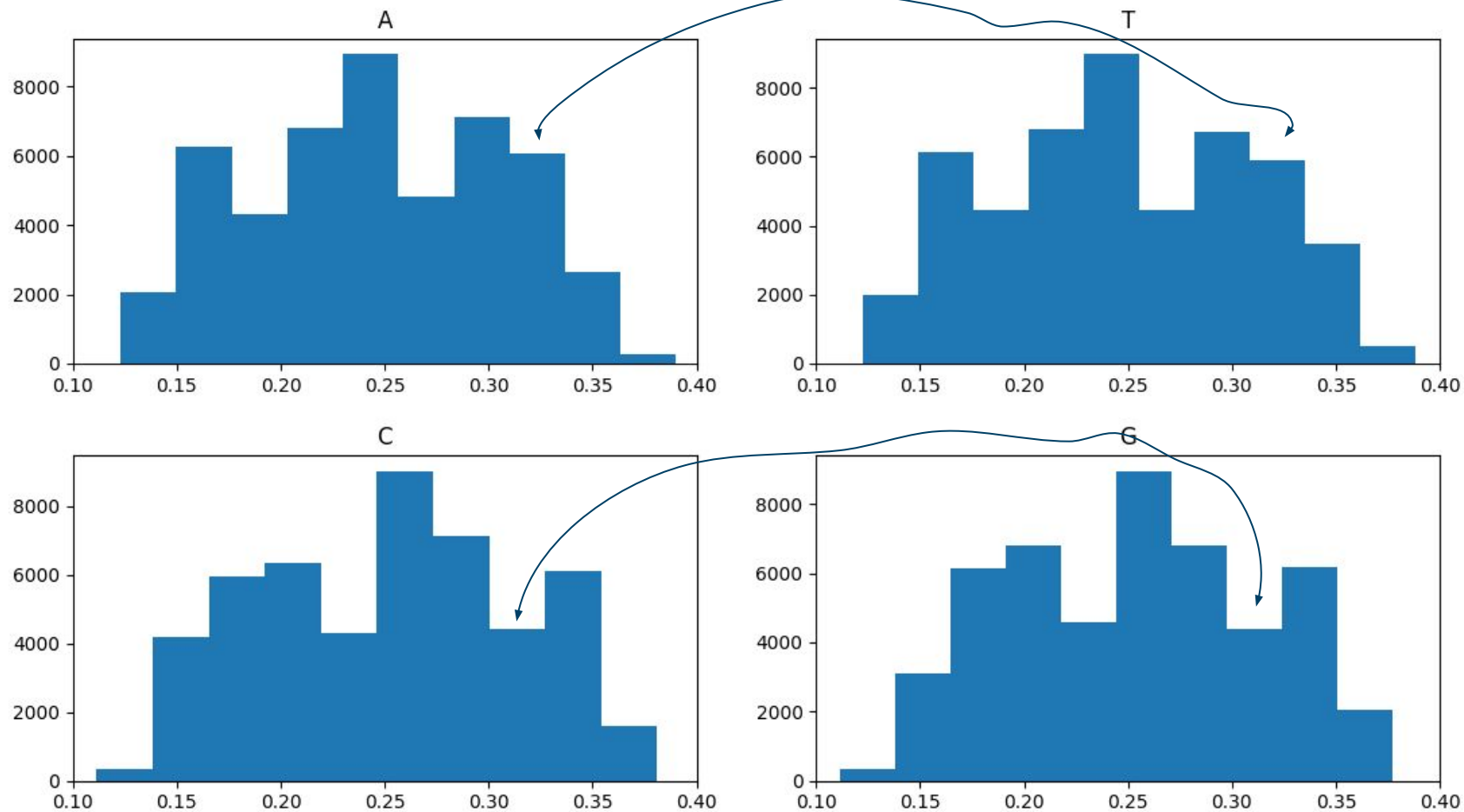
DNA

geneK

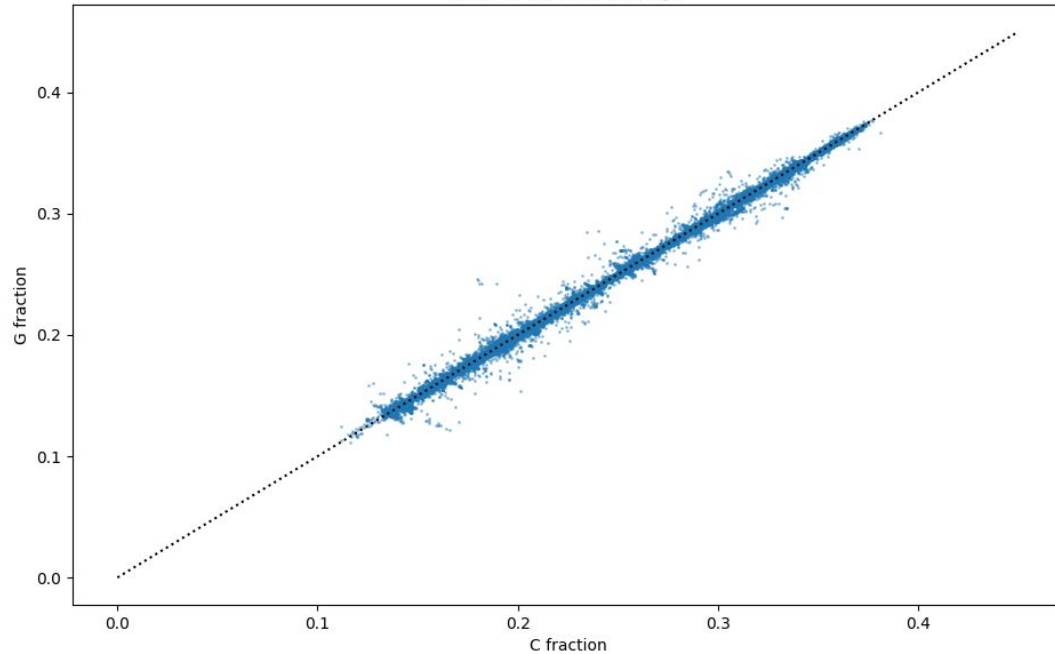
geneJ

gene31337

Histogram of nucleotide fraction for 49366 chromosomes



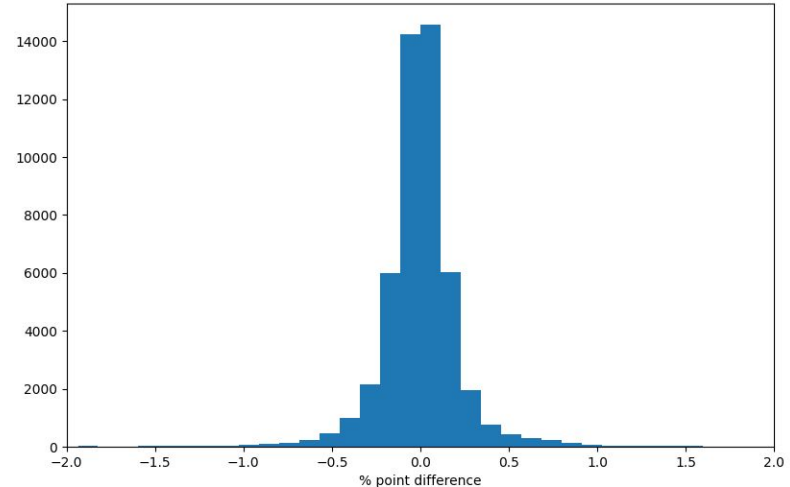
For 49366 bacterial genomes
And no one knows why!



“The second Chargaff rule holds that both $A\% \approx T\%$ and $G\% \approx C\%$ are valid for **each** of the two DNA strands”

“The basis for this rule is still under investigation” (!!!)

Histogram of C fraction minus G fraction for 49366 genomes



But wait, it gets weirder

Complementary DNA:

C -> G

A -> T

- Do the complement thing
- Reverse the string

Reverse & complement:

“Reverse Complement” (RC)

CTA -> GAT -> TAG

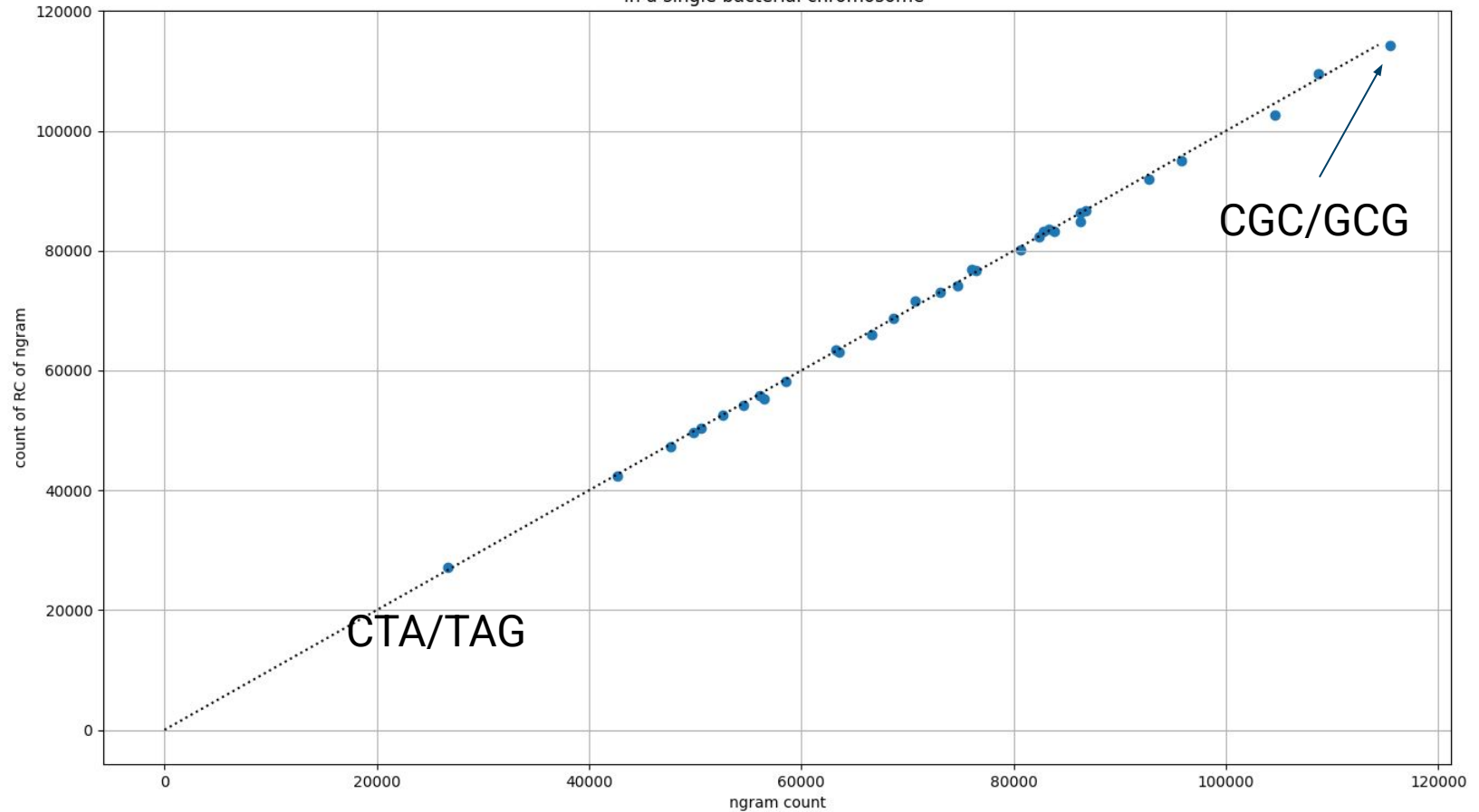
ATG -> TAC -> CAT



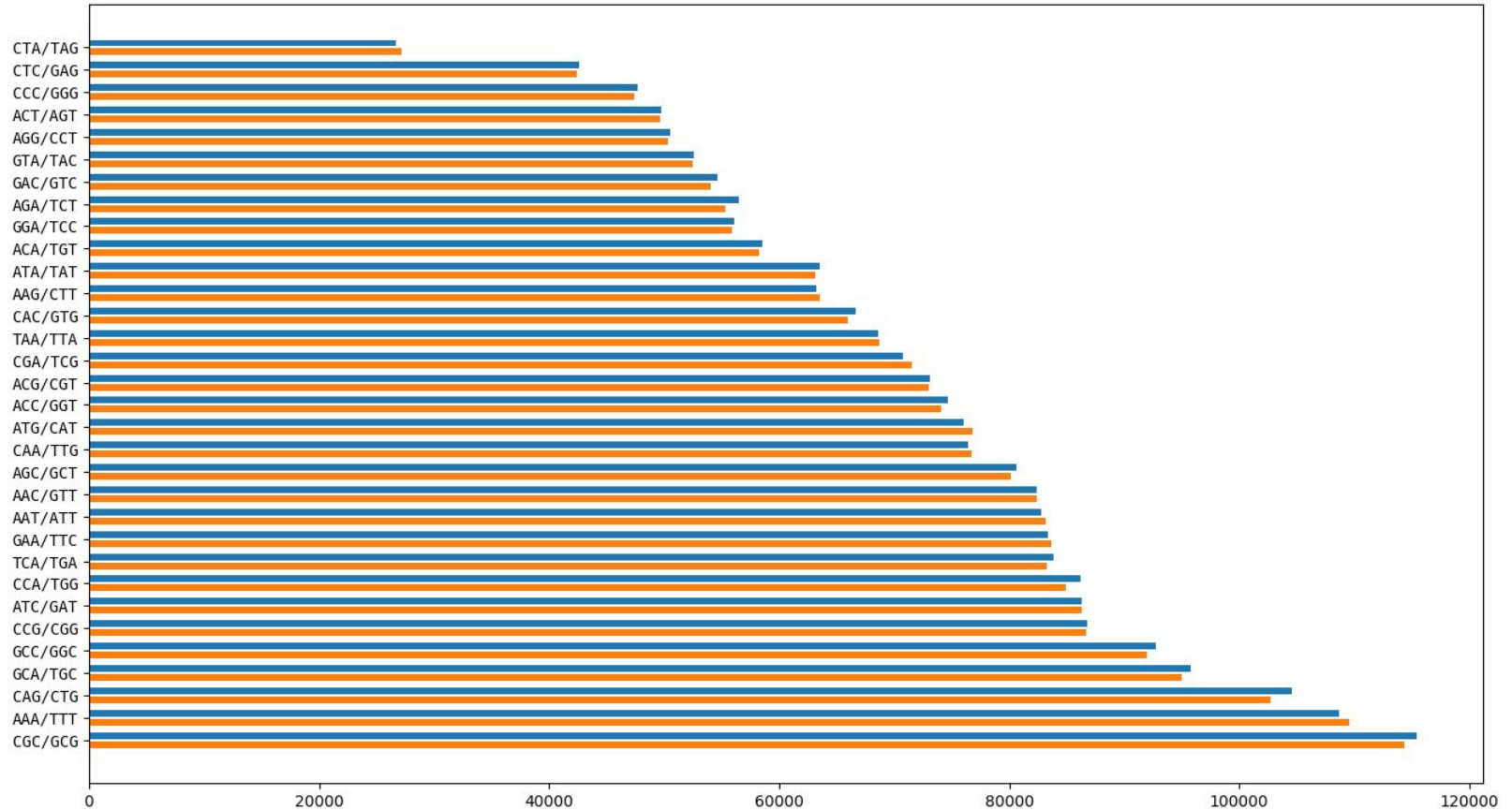
The diagram shows two DNA sequences aligned. The top sequence is **ATGCTTTGGACGGACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGGGAATGTCTTTACAATGT** and the bottom sequence is **TACGAAACCTGCCTGACGAATTGAGCGAACGCTGTTCTCGAGAGACCCTTACAGAAATGTTACA**. Vertical lines connect corresponding bases between the two sequences, illustrating base pairing (A with T, C with G). A red arrow at the top points from left to right above the top sequence. A red arrow at the bottom points from right to left below the bottom sequence. A small red star is located between the two sequences, positioned under the 15th base of the top sequence (the first 'A' in 'AACT').

ATGCTTTGGACGGACTGCTTAACTCGCTTGCGACAAGAGCTCTCTGGGAATGTCTTTACAATGT
|||||
TACGAAACCTGCCTGACGAATTGAGCGAACGCTGTTCTCGAGAGACCCTTACAGAAATGTTACA

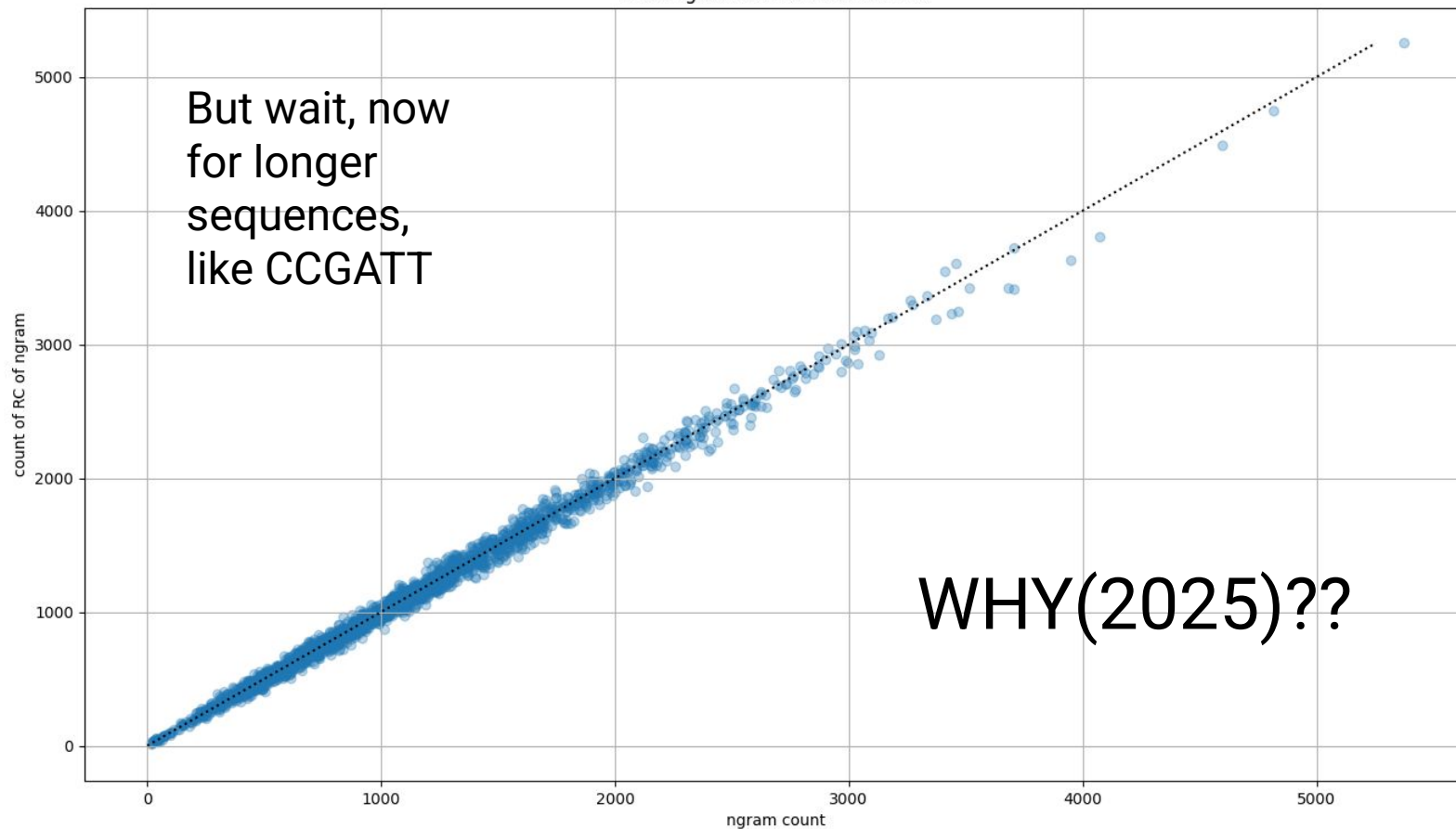
Count of trigram/codon versus count of its reverse complement
In a single bacterial chromosome



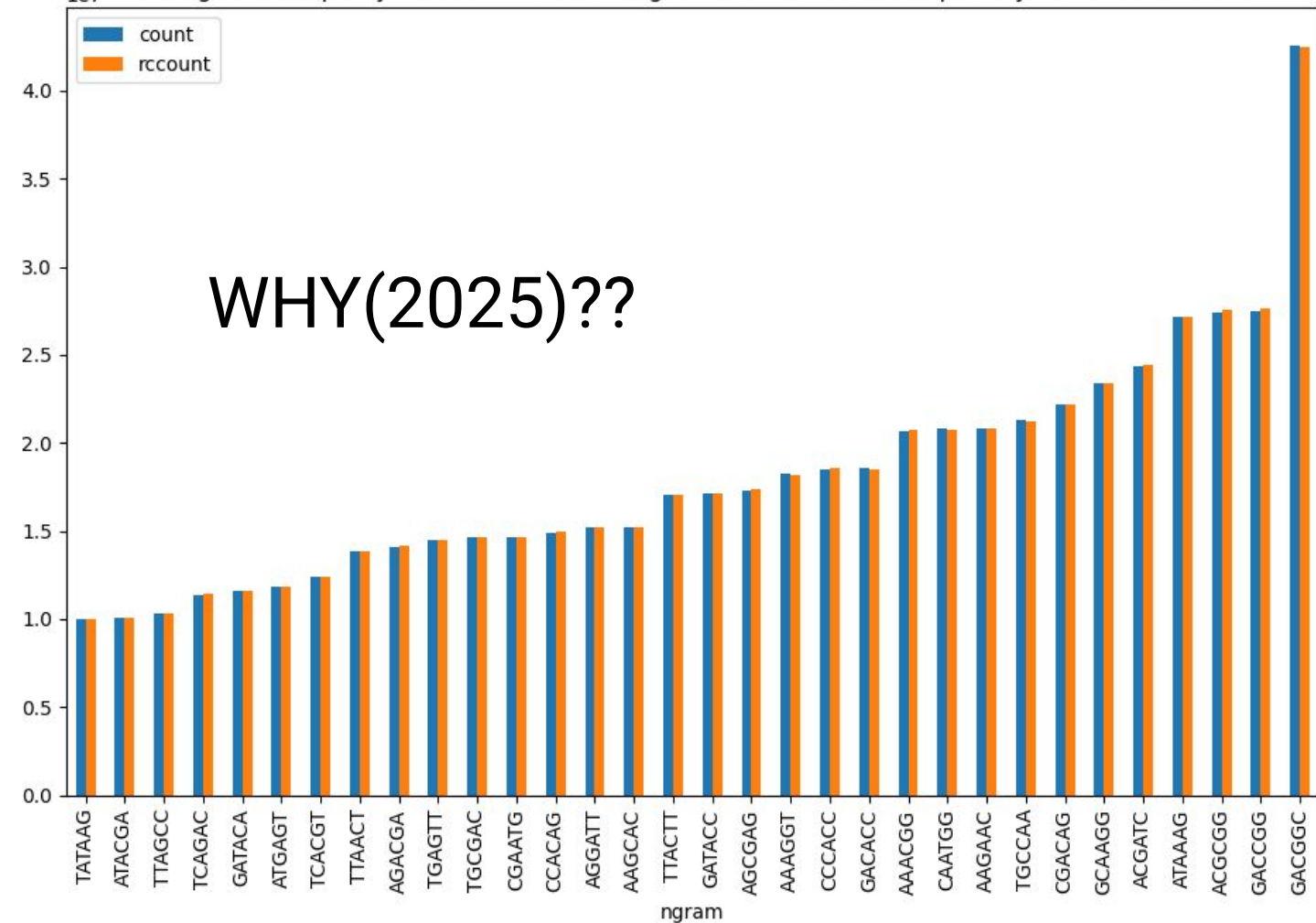
Occurrence of ngrams in DNA compared to their reverse complement
In a single bacterial chromosome



Count of 6-mer versus count of its reverse complement
In a single bacterial chromosome



WHY(2025)??

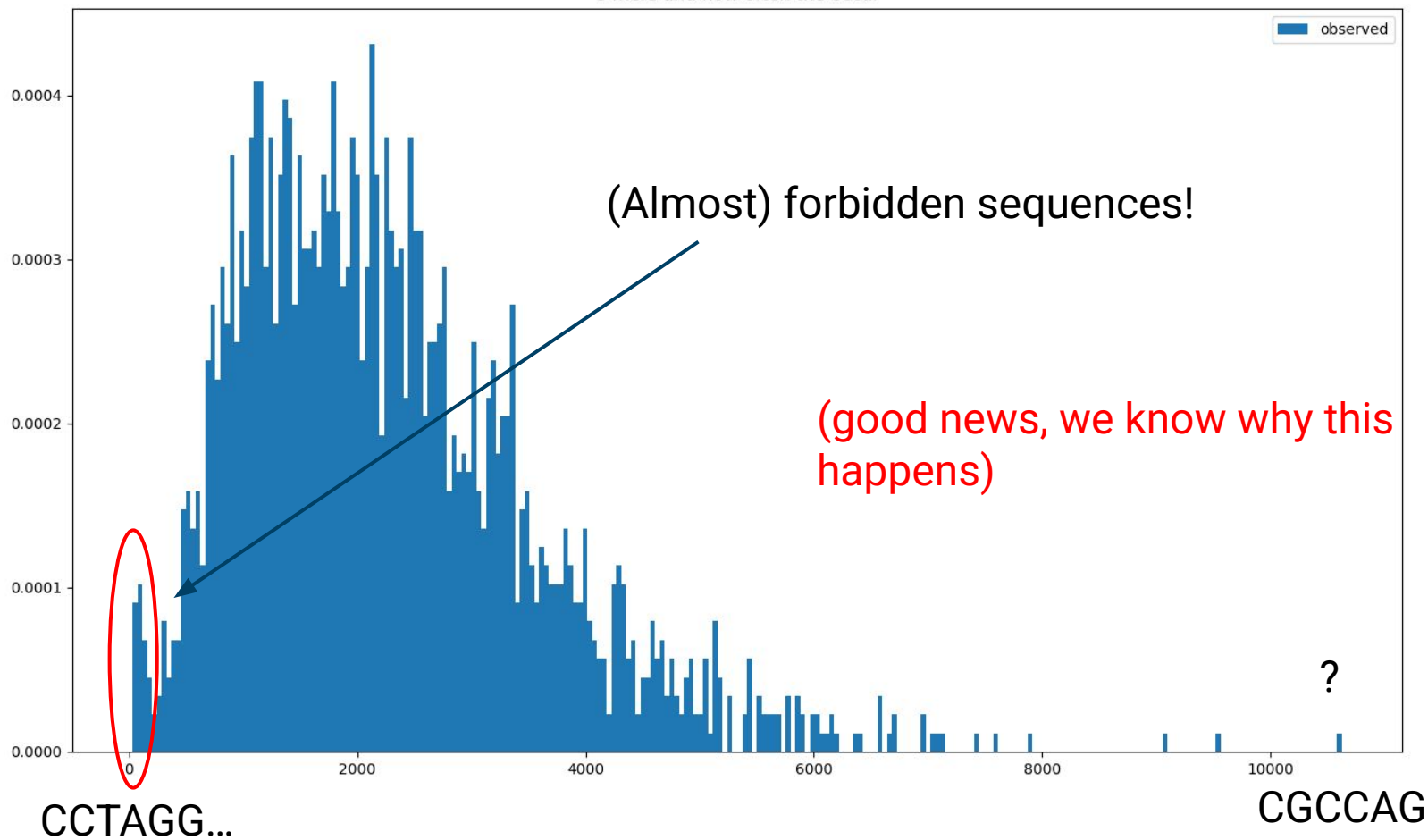


We need to definitively find out why this is so, and if it means something!

It works up to *10* characters at least.

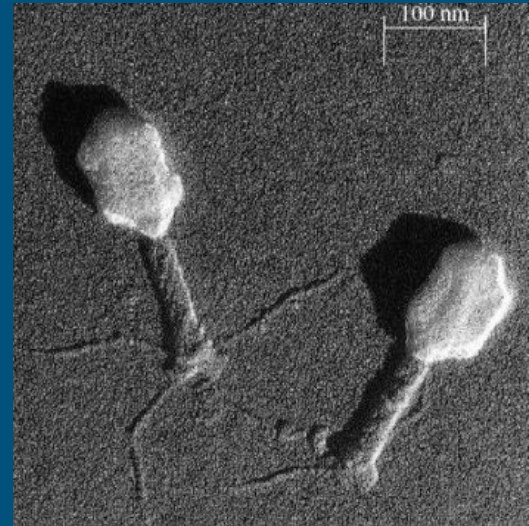
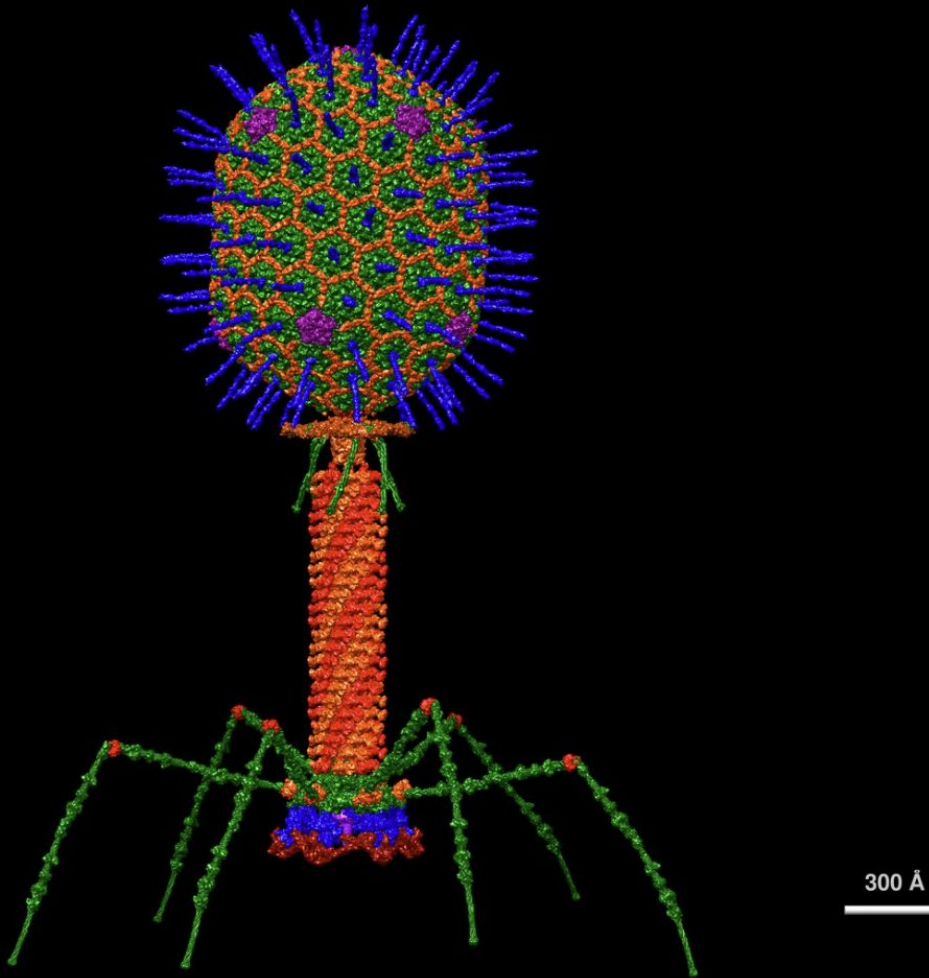
	ngram	rcngram	count	rccount	procdiff	totcount
839	CCTAGG	CCTAGG	16	16	0.00000	32
1004	CTAGGA	TCCTAG	17	21	23.52940	38
837	CCTAGA	TCTAGG	23	19	-17.39130	42
997	CTAGAC	GTCTAG	19	27	42.10530	46
1006	CTAGGG	CCCTAG	28	37	32.14290	65
...	...	...	...	...	...	...
679	CAGCGC	GCGCTG	3947	3629	-8.05675	7576
736	CCAGCA	TGCTGG	4072	3804	-6.58153	7876
738	CCAGCG	CGCTGG	4598	4489	-2.37060	9087
1291	GCCAGC	GCTGGC	4815	4750	-1.34995	9565
913	CGCCAG	CTGGCG	5372	5253	-2.21519	10625

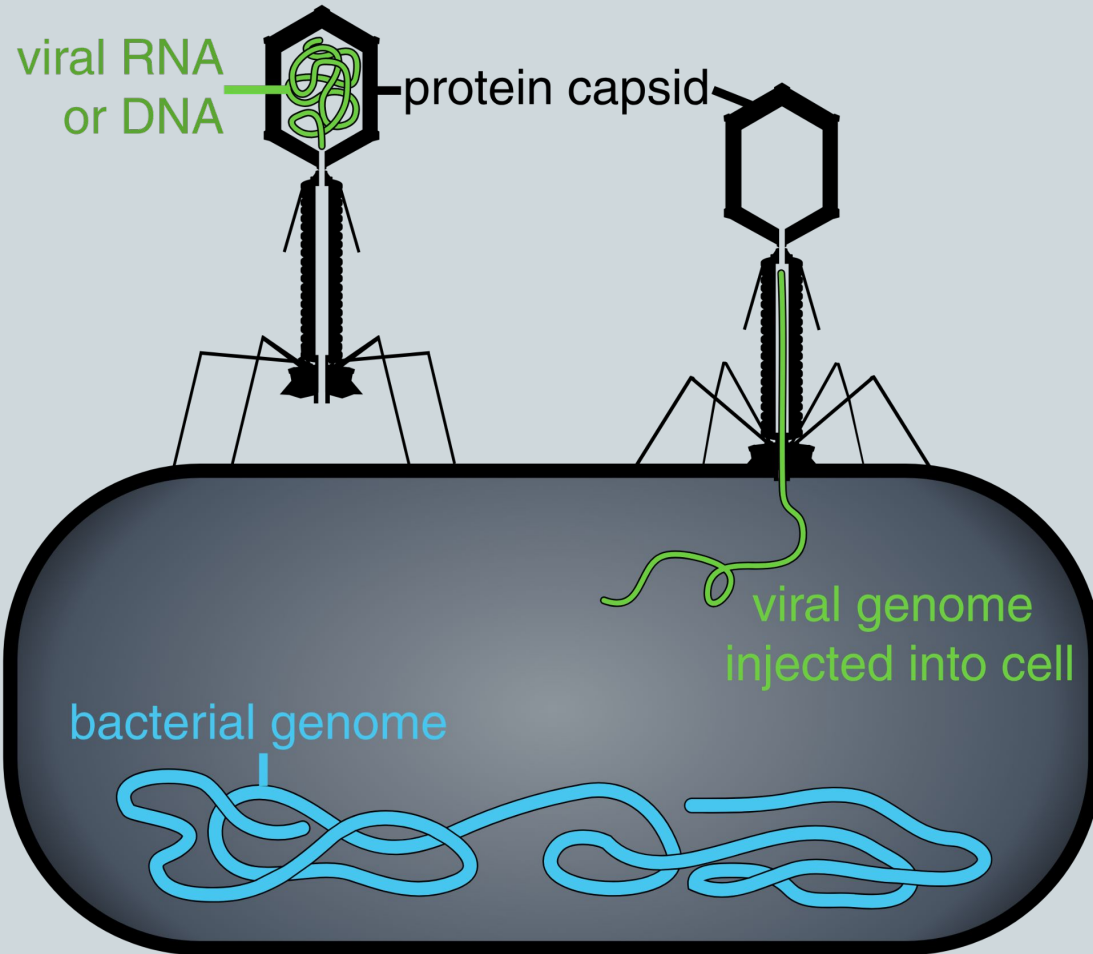
6-mers and how often the occur



Bacteria have viruses too..

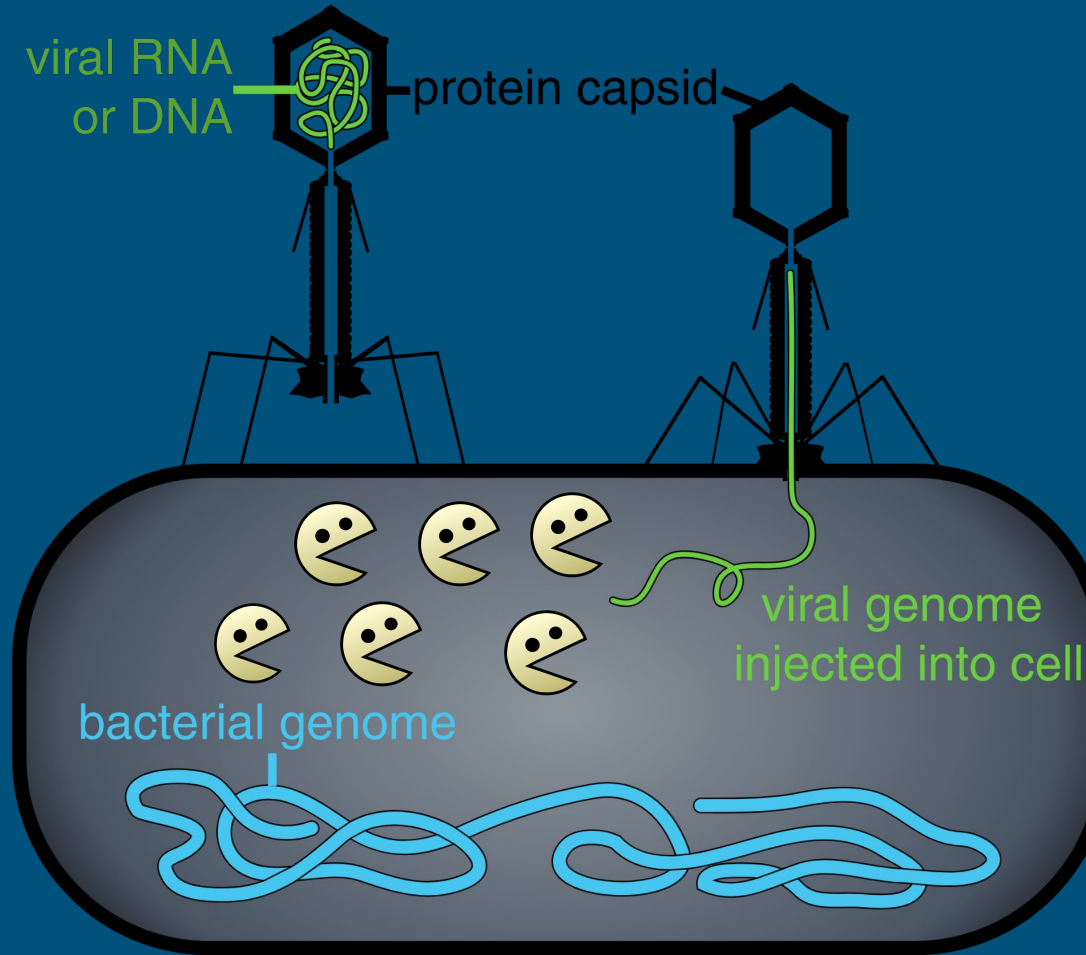
T4





“Does not end well...”





$2^{-12} = 1$ in every 4096 DNA characters at random... SNIP

This **restricts** the attacking DNA

Bacteria apply **modifications** to their own CCTAGG instances for protection

Not perfect though... so they avoid the sequence themselves

[Home](#) > [Restriction Endonucleases](#) > [AvrII](#)

AvrII       

AvrII has been reformulated with Recombinant Albumin (rAlbumin) beginning with Lot #10128047. [Learn more.](#)

We are excited to announce that all reaction buffers are now BSA-free. NEB began switching our BSA-containing reaction buffers in April 2021 to buffers containing **Recombinant Albumin** (rAlbumin) for restriction enzymes and some DNA modifying enzymes. Find more details at www.neb.com/BSA-free.

5'... C[▼] T A G G ... 3'
3'... G G A T C C ... 5'
 ▲

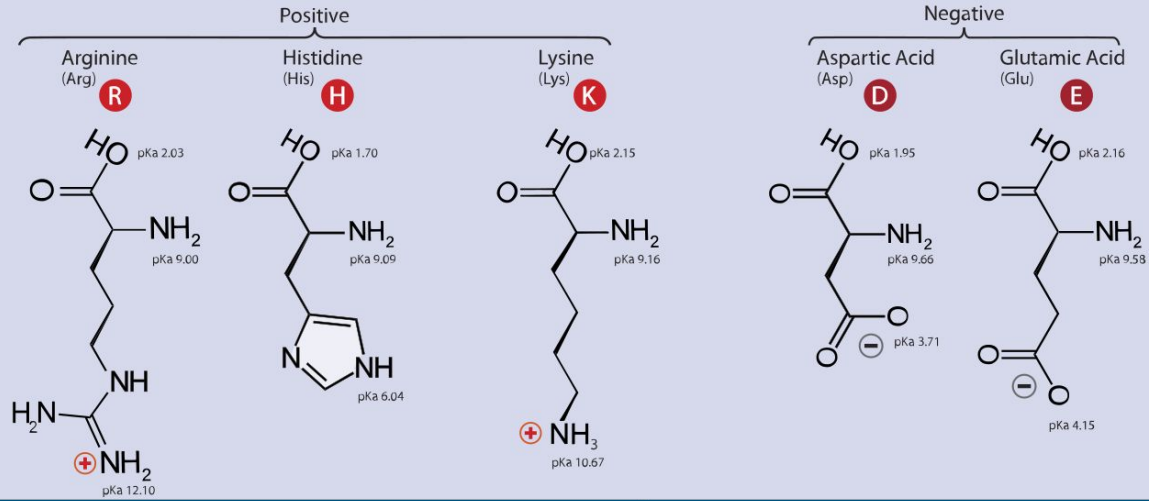
[Isoschizomers](#) | [Single Letter Code](#) | Pronunciation: 

- **Time-Saver™** qualified for digestion in 5-15 minutes
- 100% activity in **rCutSmart™ Buffer** (over 210 enzymes are available in the same buffer) simplifying double digests
- Supplied with 1 vial of **Gel Loading Dye, Purple (6X)**
- Restriction Enzyme Cut Site: C/CTAGG

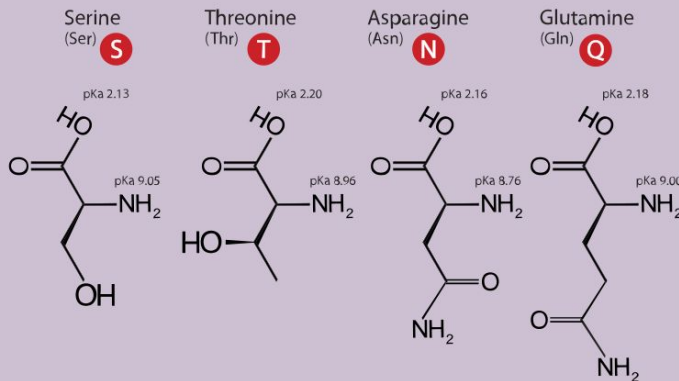
Twenty-One Amino Acids

⊕ Positive
⊖ Negative
• Side chain charge at physiological pH 7.4

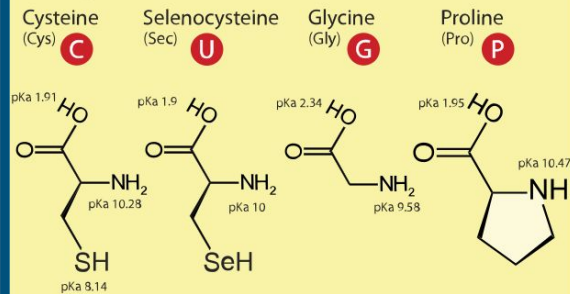
A. Amino Acids with Electrically Charged Side Chains



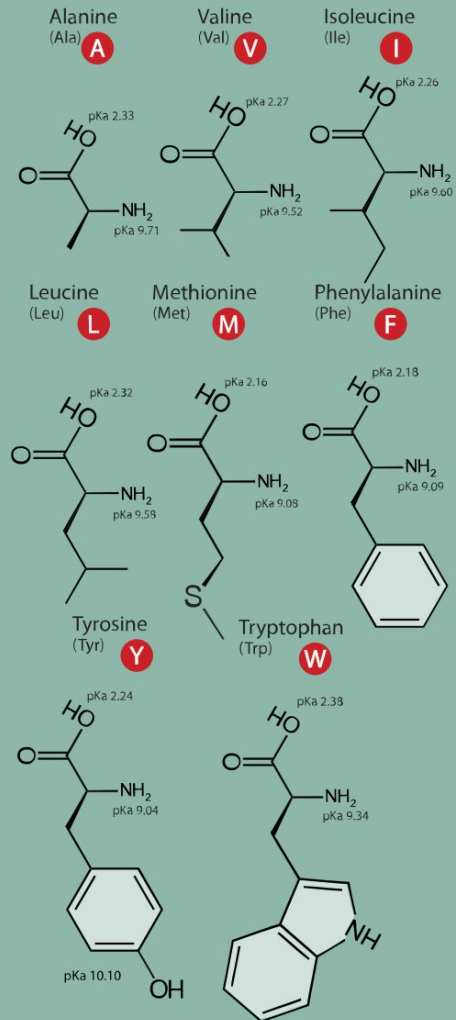
B. Amino Acids with Polar Uncharged Side Chains



C. Special Cases



D. Amino Acids with Hydrophobic Side Chain



1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine (p)	TAT	(Tyr/Y) Tyrosine (p)	TGT	(Cys/C) Cysteine (p)	T
	TTC	(np)	TCC		TAC		TGC		C
	TTA		TCA		TAA	Stop (Ochre) *[note 2]	TGA	Stop (Opal) *[note 2]	A
	TTG ⇒		TCG		TAG	Stop (Amber) *[note 2]	TGG	(Trp/W) Tryptophan (np)	G
C	CTT	(Leu/L) Leucine (np)	CCT	(Pro/P) Proline (np)	CAT	(His/H) Histidine (b)	CGT	(Arg/R) Arginine (b)	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine (p)	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine (np)	ACT	(Thr/T) Threonine (p)	AAT	(Asn/N) Asparagine (p)	AGT	(Ser/S) Serine (p)	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine (b)	AGA	(Arg/R) Arginine (b)	A
	ATG ⇒		ACG		AAG		AGG		G
G	GTT	(Val/V) Valine (np)	GCT	(Ala/A) Alanine (np)	GAT	(Asp/D) Aspartic acid (a)	GGT	(Gly/G) Glycine (np)	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid (a)	GGA		A
	GTG ⇒		GCG		GAG		GGG		G

Multi-billion year old table!

Multiple codons for same amino acids

This allows for **dialects** and shaping DNA

Where do genes
begin and end?

```
>NZ_CP150338.1 Sorangium sp. So ce388 chromosome, complete genome
ATGACGATTCCGAAACACGAGCCCCGCGAAGTCTTCGATCGGGCGATCGAGCATACGCGGGCCCTTTCTCCCGCAACTTT
TGAACAGTGGTTTGGGGGAGTTTCAGTTCGATGACCTGACCGACGGCGTGCTCACGCTGCGAGTCCAGAACGAGTTCGTCC
TCGAGTGGGTACAGGGACAATTTCTGCCCGCGCTGACCGACAAGATCCGCGAGATCACGGGCTGGTTCGGTCCAGGTGGCG
TGGACGGTGCATCAGCACCTTTCAGTTCGGCGATCGCGACGCGCTCGAGCTACCCCGGTTTCGCGCGCGGGCGCTCGTGGT
GCGTCCGACGAGACGCGGCG
TCAACCCGAAGCACACCTT
GGCGGCGGGGGTCGCCGG
CGCGCATCGCGTCTTCGAG
CGGCCATCCAGCACCAACC
CAGTTCCTGGCCGGGCGCG
CGTGGTGACGAGCGACAAC
TCGCCGACATCCAGGTGCC
ACGGACGACGTGGCGCTC
GGCCAAGAGCTCGCTCACC
```

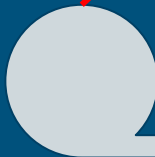
##gff-version 3
#lgff-spec-version 1.21
#lprocessor NCBI annotwriter
#lgenome-build ASM5147412v1
#lgenome-build-accession NCBI_Assembly:GCF_051474125.1
#lannotation-date 07/23/2025 04:03:25
#lannotation-source NCBI RefSeq GCF_051474125.1-RS_2025_07_23
##sequence-region NZ_CP150338.1 1 14889354
##species <https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=3133309>
NZ_CP150338.1 RefSeq region 1 14889354 + ID=NZ_CP150338.1:1..14889354;Dbxref=tax
on:3133309;Is_circular=true;Name=ANONYMOUS;collection-date=1988;country=Japan;gbkey=Src;genome=chromosome;isolation-sou
rce=mud;mol_type=genomic DNA;strain=So ce388
NZ_CP150338.1 RefSeq gene 1 1416 + ID=gene-WMF16_RS00005;Name=dnaA;gbkey=Gene;gene
=dnaA;gene_biotype=protein_coding;locus_tag=WMF16_RS00005;old_locus_tag=WMF16_00005
NZ_CP150338.1 Protein Homology CDS 1 1416 + 0 ID=cds-WP_437306908.1;Parent=ge
ne-WMF16_RS00005;Dbxref=GenBank:WP_437306908.1;Name=WP_437306908.1;Ontology_term=GO:0006270,GO:0006275,GO:0003677,GO:00
03688,GO:0005524;gbkey=CDS;gene=dnaA;go_function=DNA binding|0003677||IEA,DNA replication origin binding|0003688||IEA,ATP binding|0005524||IEA;go_process=DNA replication initiation|0006270||IEA,regulation of DNA replication|0006275||IEA;i
nference=COORDINATES: similar to AA sequence:RefSeq:WP_020464715.1;locus_tag=WMF16_RS00005;product=chromosomal replicat
ion initiator protein DnaA;protein_id=WP_437306908.1;transl_table=11
NZ_CP150338.1 RefSeq gene 1516 5439 - ID=gene-WMF16_RS00010;Name=WMF16_RS00010;gbkey=
Gene;gene_biotype=protein_coding;locus_tag=WMF16_RS00010;old_locus_tag=WMF16_00010
NZ_CP150338.1 Protein Homology CDS 1516 5439 - 0 ID=cds-WP_437306909.1;Parent=ge
ne-WMF16_RS00010;Dbxref=GenBank:WP_437306909.1;Name=WP_437306909.1;Ontology_term=GO:0006468,GO:0004674,GO:0005524,GO:00
16887;gbkey=CDS;go_function=protein serine/threonine kinase activity|0004674||IEA,ATP binding|0005524||IEA,ATP hydrolysis activity|0016887||IEA;go_process=protein phosphorylation|0006468||IEA;inference=COORDINATES: protein motif:HMM:NF012
298.6;locus_tag=WMF16_RS00010;product=serine/threonine-protein kinase;protein_id=WP_437306909.1;transl_table=11
NZ_CP150338.1 RefSeq gene 5569 7431 - ID=gene-WMF16_RS00015;Name=dnaK;gbkey=Gene;gene
=dnaK;gene_biotype=protein_coding;locus_tag=WMF16_RS00015;old_locus_tag=WMF16_00015
NZ_CP150338.1 Protein Homology CDS 5569 7431 - 0 ID=cds-WP_437306910.1;Parent=ge

HNGGGGGG!!!

FASTA

GFF

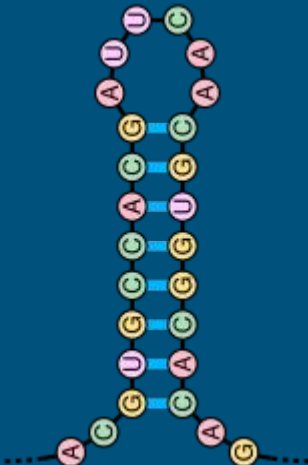
DNA layer
RNA layer
Amino acid layer
..
RNA layer
DNA layer



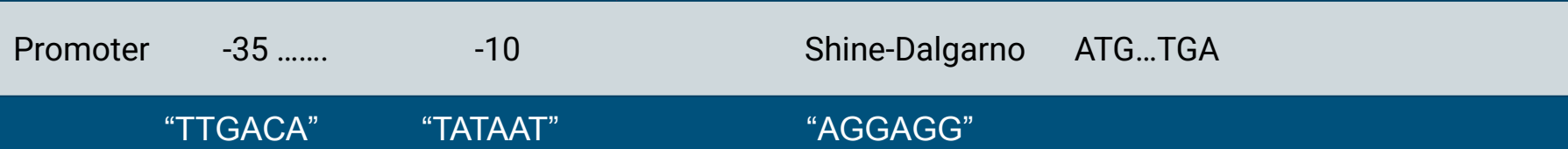
RNA

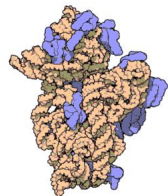


Amino Acids

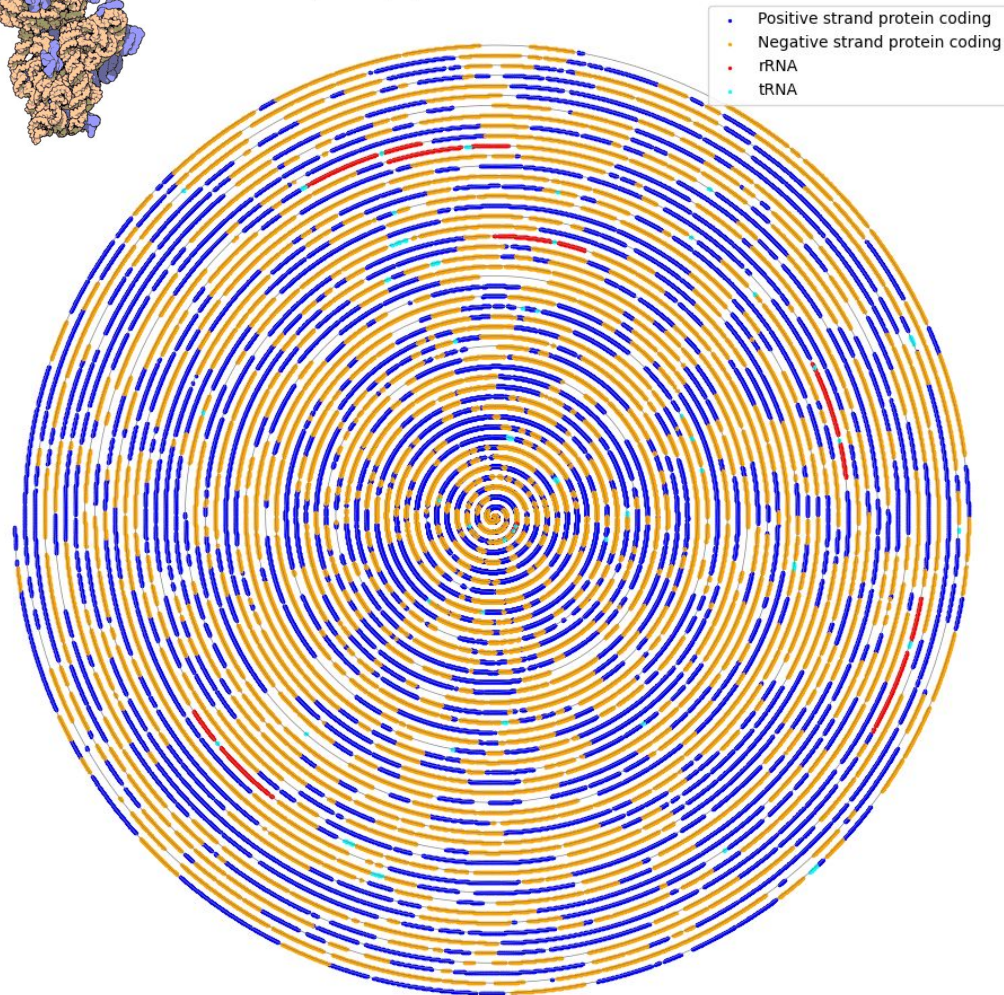


Amino acids
code





Spiral display of a bacterial chromosome



Escherichia coli str. K-12 substr.
W3110

Topologically this chromosome is
actually a circle.

Spiral shape however is better for an
overview

The blank areas are **not genes**.
Partially we know what this is.
Partially not!

Biologists study genes.. “Looking for
your keys where the light is”

**But us nerds could take a look at the
data!**

```
Usage: prodigal [-a trans_file] [-c] [-d nuc_file] [-f output_type]
           [-g tr_table] [-h] [-i input_file] [-m] [-n] [-o output_file]
           [-p mode] [-q] [-s start_file] [-t training_file] [-v]
```

So standard that Debian ships it! (Also, congrats on Trixie!)

- a: Write protein translations to the selected file.
- c: Closed ends. Do not allow genes to run off edges.
- d: Write nucleotide sequences of genes to the selected file.
- f: Select output format (gbk, gff, or sco). Default is gbk.
- g: Specify a translation table to use (default 11).
- h: Print help menu and exit.
- i: Specify FASTA/Genbank input file (default reads from stdin).
- m: Treat runs of N as masked sequence; don't build genes across them.
- n: Bypass Shine-Dalgarno trainer and force a full motif scan.
- o: Specify output file (default writes to stdout).
- p: Select procedure (single or meta). Default is single.
- q: Run quietly (suppress normal stderr output).
- s: Write all potential genes (with scores) to the selected file.
- t: Write a training file (if none exists); otherwise, read and use the specified training file.
- v: Print version number and exit.

```
ahu@xeon:~/skewdb/skewdb-articles/antonie2$ prodigal -i in -o genes
```

```
-----  
PRODIGAL v2.6.3 [February, 2016]
```

```
Univ of Tenn / Oak Ridge National Lab
```

```
Doug Hyatt, Loren Hauser, et al.  
-----
```

```
Request:  Single Genome, Phase:  Training
```

```
Reading in the sequence(s) to train...4646332 bp seq created, 50.80 pct GC
```

```
Locating all potential starts and stops...241263 nodes
```

```
Looking for GC bias in different frames...frame bias scores: 1.54 0.18 1.27
```

```
Building initial set of genes to train from...done!
```

```
Creating coding model and scoring nodes...done!
```

```
Examining upstream regions and training starts...done!  
-----
```

```
Request:  Single Genome, Phase:  Gene Finding
```

```
Finding genes in sequence #1 (4646332 bp)...done!
```

```
ahu@xeon:~/skewdb/skewdb-articles/antonie2$ valgrind prodigal -i in -o genes
==4112615== Memcheck, a memory error detector
==4112615== Copyright (C) 2002-2022, and GNU GPL'd, by Julian Seward et al.
==4112615== Using Valgrind-3.19.0 and LibVEX; rerun with -h for copyright info
==4112615== Command: prodigal -i in -o genes
==4112615==
-----
PRODIGAL v2.6.3 [February, 2016]
Univ of Tenn / Oak Ridge National Lab
Doug Hyatt, Loren Hauser, et al.
-----
Request: Single Genome, Phase: Training
Reading in the sequence(s) to train...4646332 bp seq created, 50.80 pct GC
Locating all potential starts and stops...241263 nodes
Looking for GC bias in different frames...frame bias scores: 1.54 0.18 1.27
Building initial set of genes to train from...==4112615== Conditional jump or move depends on uninitialised value(s)
==4112615==    at 0x11708D: ??? (in /usr/bin/prodigal)
==4112615==    by 0x109E60: ??? (in /usr/bin/prodigal)
==4112615==    by 0x4976249: (below main) (libc_start_call_main.h:58)
==4112615==
==4112615== Conditional jump or move depends on uninitialised value(s)
==4112615==    at 0x116FC8: ??? (in /usr/bin/prodigal)
==4112615==    by 0x10BF48: ??? (in /usr/bin/prodigal)
```

OOPS!!!!

Realloc zero, fixes undefined behaviour #119



Open

berthubert wants to merge 2 commits into [hyattpd:GoogleImport](#) from [berthubert:realloc-zero](#)



Conversation 0



Commits 2



Checks 0



Files changed 1



berthubert commented [4 days ago](#) ...

Valgrind saw Prodigal access uninitialized memory. This was because resizing the nodes array using ~~realloc~~ did not zero the newly allocated memory. This could have had consequences for generated gene predictions if you were unlucky.

This PR makes sure that the new memory is zeroed as well. In addition, an error message if /dev/stdin was not available was fixed to not crash.



berthubert added 2 commits [4 days ago](#)



[fix error report if /dev/stdin could not be opened](#)

[d4e3d76](#)



[when resizing the nodes array, the newly allocated space was not zero...](#) ...

[4e3b87b](#)

```

485 491
486 492     /* Reallocate memory if this is the biggest sequence we've seen */
487 493     if(slen > max_slen && slen > STT_NOD*8) {
488 -         nodes = (struct _node *)realloc(nodes, (int)(slen/8)*sizeof(struct _node));
494 +         size_t newnodesize = (int)(slen/8)*sizeof(struct _node);
495 +         nodes = (struct _node *)realloc(nodes, newnodesize);
489 496         if(nodes == NULL) {
490 497             fprintf(stderr, "Realloc failed on nodes\n\n");
491 498             exit(11);
492 499         }
500 +         memset( ((char*) &nodes[0]) + nodesize, 0, newnodesize-nodesize);
501 +         nodesize = newnodesize;
493 502         max_slen = slen;
494 503     }
495 504

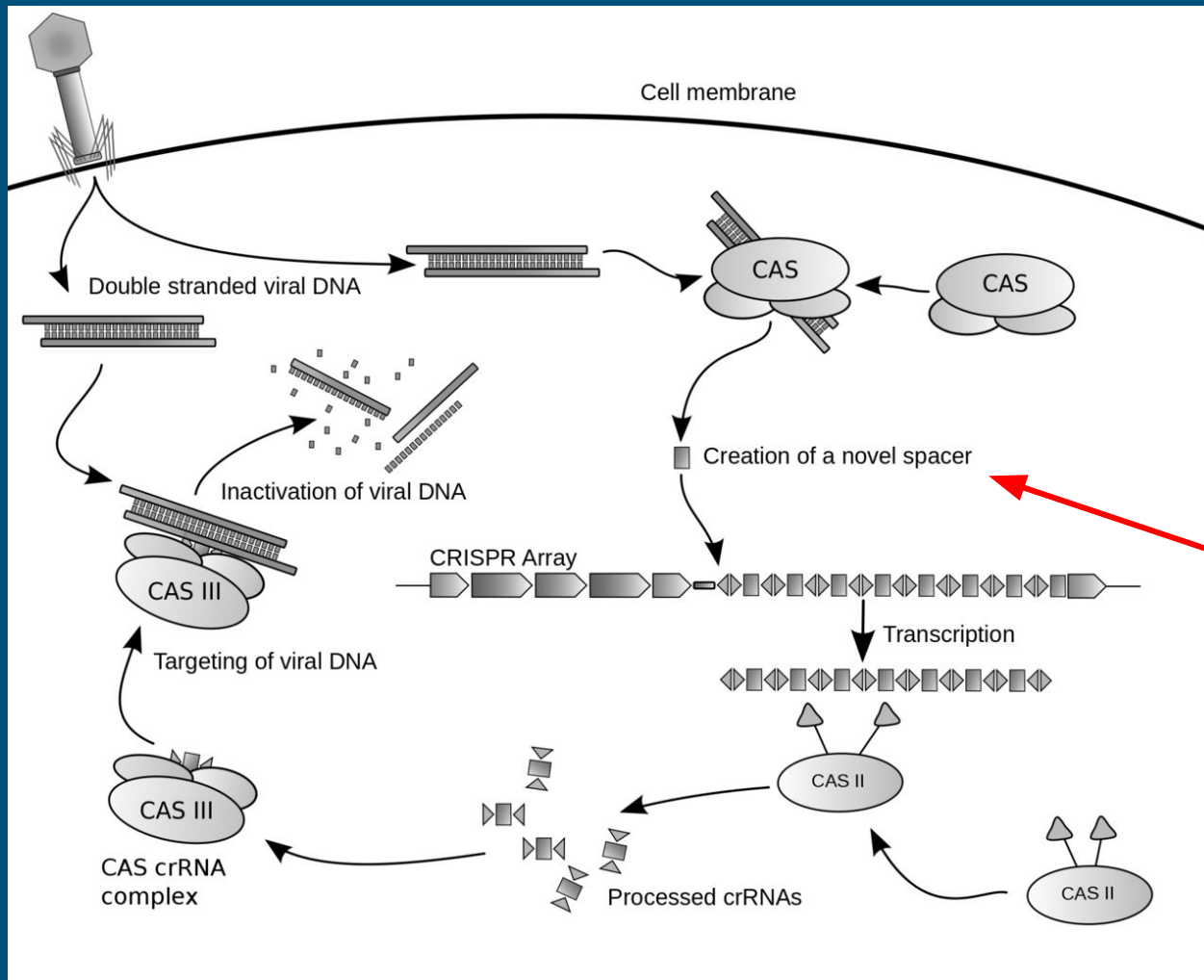
```

The state of bioinformatics software is... not great (second time this happened to me)

More bacterial anti-viral
defenses.

This war has been raging for 2
billion years at least!

Maybe we could learn..



A multi-generational immune system.

A forensic record of previously survived viruses!

Can we see it?

People looked at this for 20 years w/o knowing what it was

By James atmos - Own work, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=7821536>

CRISPR: “The origin of the spacer sequences remains unknown” - 2002

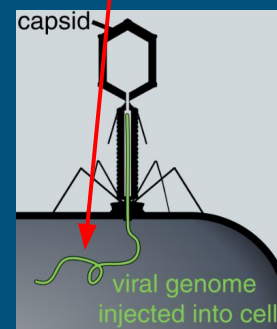
```
CTGGATAATCACCAACATATTGATAAGGCTCTTCCTGGAAGAATAGAACGTCGCTGCCGTGACGTTATGCGGATAATGCT
ACCTCTGGTGAAGGAGTTGGCGAAGGCGTCTTGATGGGTTTGAAAATGGGAGCTGGGAGTTCTACCGCAGAGGCGGGGA
ACTCCAAGTGATATCCATCATCGCATCCAGTGCGCCCGGTTTATCCCCGCTGATGCGGGGAACACCAGCGTCAGGCGTGA
AATCTCACCGTCGTTGCCGTTTATCCCTGCTGGCGCGGGGAACCTCTCGGTTTACAGGCGTTGCAAACCTGGCTACCGGGCG
GTTTATCCCCGCTAACGCGGGGAACCTCGTAGTCCATCATTCCACCTATGTCTGAACTCCCGGTTTATCCCCGCTGGCGCG
GGGAACCTCCCGGGGGATAATGTTTACGGTCATGCGCCCCCGGTTTATCCCCGCTGGCGCGGGGAACCTCTGGGCGGCTTG
CCTTGACAGCCAGCTCCAGCAGCGGTTTATCCCCGCTGGCGCGGGGAACCTCAAGCTGGCTGGCAATCTCTTTCGGGGTGAG
TCCCGGTTTATCCCCGCTGGCGCGGGGAACCTCTAGTTTCCGTATCTCCGGATTATATAAAGCTGACCGGTTTATCCCCGCTGG
CGCGGGGAACCTCGCAGGCGGCGACGCGCAGGGTATGCGCGATTCCCGGTTTATCCCCGCTGGCGCGGGGAACCTCGCGACC
GCTCAGAAATTCCAGACCCGATCCAAACCGGTTTATCCCCGCTGGCGCGGGGAACCTCTCAACATTATCAATTACAACCGAC
AGGGAGCCCGGTTTATCCCCGCTGGCGCGGGGAACCTCAGCGTGTTTCGGCATCACCTTTGGCTTCGGCTGCGGTTTATCCC
CGCTGGCGCGGGGAACCTCTGCGTGAGCGTATCGCCGCGCGTCTGCGAAAGCGGTTTATCCCCGCTGGCGCGGGGAACCTCT
CTAAAAGTATACATTTGTTCTTAAAGCATTTTTTCCCATAAAAACAACCCACCAACCTTAATGTAAACATTTCTTATTAT
TAAAGATCAGCTAATTCTTTGTTTTCAAACAGGTAAAAAAGACACCAACCTTAACCATCCAATCTACCGGGGTACGCC
```

cas genes

Leader

Repeat-spacer
array

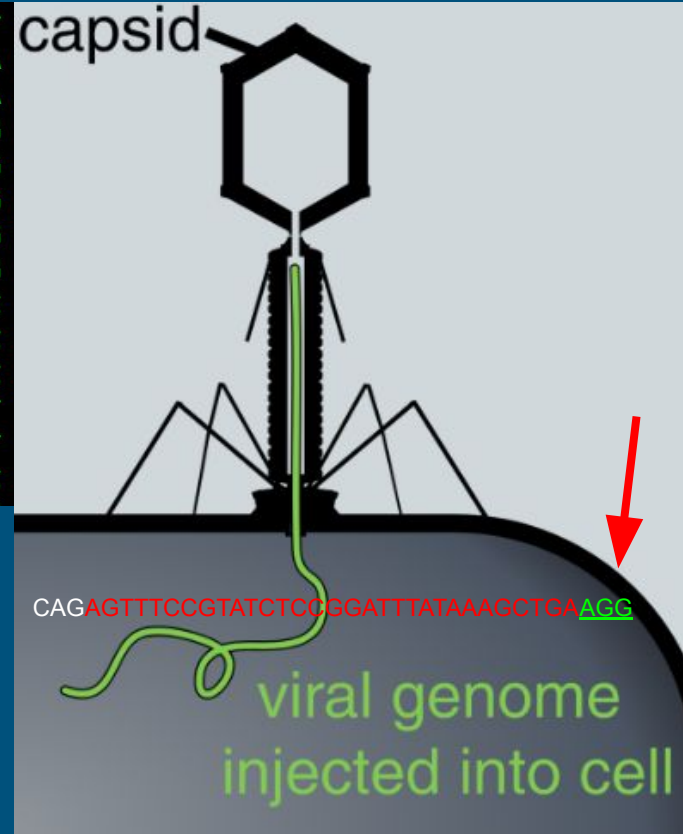
```
CTGGATAATCACCAACATATTGATAAGGCTCTTCTGGAAGAAATAGAACGTCCTGCGGTGACGTTATGCGGATAATGCT
ACCTCTGGTGAAGGAGTTGGCGAAGGCGTCTTGATGGGTTTGAAAATGGAGCTGGGAGTTCTACCGCAGAGGCGGGGGA
ACTCCAAGTGATATCCATCATCGCATCCAGTGCGCCGGGTTTATCCCCGCTGATGCGGGGAACACGAGCGTCAGGCGTGA
AATCTCACCGTCGTTGCCGTTTATCCCTGCTGGCGCGGGGAACCTCTCGGTTCAAGCGTTGCAAACTGCTACCGGGCG
GTTTATCCCCGCTAACGCGGGGAACCTGATGCCATCATTCACCTATGTCTGAACCTCCGGTTTATCCCCGCTGGCGCG
GGGAACCTCCGCGGGGATAATGTTTACGGTCATGCGCCCCCGGTTTATCCCCGCTGGCGCGGGGAACCTCTGGCGGCTT6
CCTTGACGCCAGCTCCAGCAGCGGTTTATCCCCGCTGGCGCGGGGAACCTCAAGCTGGCTGGCAATCTCTTTCGGGTGAG
TCGGGTTTATCCCCGCTGGCGCGGGGAACCTAGTTTCCGTATCTCCGGATTATAAAGCTGACGGTTTATCCCCGCTGG
CGCGGGGAACCTCGCAGGCGGCGACGCGCAGGGTATGCGCGATTCCGGTTTATCCCCGCTGGCGCGGGGAACCTCGCGACC
GCTCAGAAATTCAGACCCGATCAAAAAGGTTTATCCCCGCTGGCGCGGGGAACCTCTCAACATTATCAATTACAACCGAC
AGGGAGCCCGGTTTATCCCCGCTGGCGCGGGGAACCTCAGCGTGTTCGGCATCACCTTTGGCTTCGGCTCGGTTTATCCC
CGCTGGCGCGGGGAACCTCTGCGTGAGCGTATCGCCGCGCTCTGCGAAAGCGGTTTATCCCCGCTGGCGCGGGGAACCTCT
CTAAAAGTATACATTTGTTCTTAAAGCATTTTTCCCATAAAAACAACCCACCAACCTTAATGTAACATTTCTTATTAT
TAAAGATCAGCTAATTCTTTGTTTTCAACAGGTAAGGACACCAACCTTAAACATCAAAATCTACCGGGGTACGCC
```

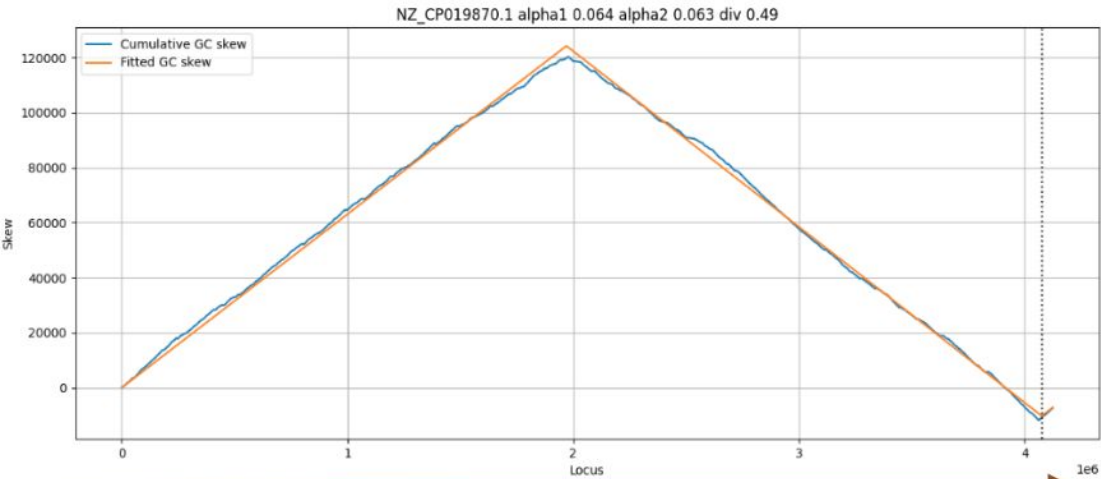


CTGGATAATCACCAACATATTGATAAGGCTCTTCTCTGGAAGAATAGAACGTCGCTGCCGTGACGTTATGCGGATAATGCT
 ACCTCTGGTGAAGGAGTTGGCGAAGGCGTCTTGATGGGTTTAAAAATGGGAGCTGGGAGTTCTACCGCAGAGGCGGGGA
 ACTCCAAGTGATATCCATCATCGCATCCAGTGCGCCCGGTTTATCCCCGCTGATGCGGGGAACACCAGCGTCAGGCGTGA
 AATCTCACCGTCGTTGCCGTTTATCCCTGCTGGCGCGGGGAACCTCTCGTTTCAGGCGTTGCAAACCTGGCTACCGGGCG
 GTTTATCCCCGCTAACGCGGGGAACCTCGTAGTCCATCATTCCACCTATGTCTGAACTCCCGGTTTATCCCCGCTGGCGCG
 GGGAACTCCCGGGGATAATGTTTACGGTCATGCGCCCCCGGTTTATCCCCGCTGGCGCGGGGAACCTCTGGGCGGCTTG
 CCTTGCGAGCCAGCTCCAGCAGCGGTTTATCCCCGCTGGCGCGGGGAACCTCAAGCTGGCTGGCAATCTCTTTCGGGGTGAG
 TCCGGTTTATCCCCGCTGGCGCGGGGAACCTCTAGTTTCCGTATCTCCGATTATATAAAGCTGACGGTTTATCCCCGCTGG
 CGCGGGGAACCTCGCAGGCGGCGACGCGCAGGGTATGCGCGATTCCCGGTTTATCCCCGCTGGCGCGGGGAACCTCGCGACC
 GCTCAGAAATTCAGACCCGATCCAAAACGGTTTATCCCCGCTGGCGCGGGGAACCTCTCAACATTATCAATTACAACCGAC
 AGGGAGCCCGGTTTATCCCCGCTGGCGCGGGGAACCTCAGCGTGTTTCGGCATCACCTTTGGCTTCGGCTGCGGTTTATCCC
 CGCTGGCGCGGGGAACCTCTGCGTGAGCGTATCGCCGCGCTCTGCGAAAGCGGTTTATCCCCGCTGGCGCGGGGAACCTCT
 CTAAGATCAGCTAATTCTTTGTTTCAAACAGGTAAAAAGACACCAACCTTAATGTAACATTTCCTTATTAT
 TAAAGATCAGCTAATTCTTTGTTTCAAACAGGTAAAAAGACACCAACCTTAACCATCCAATCTACCGGGGTACGCC

Signature:
 “AGTTTCCGTATCTCCGGATTATATAAAGCTGA”

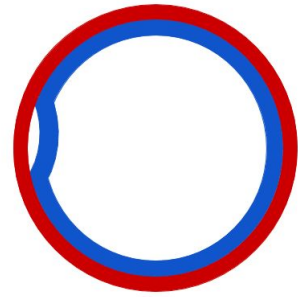
Why does the CRISPR system not destroy itself?



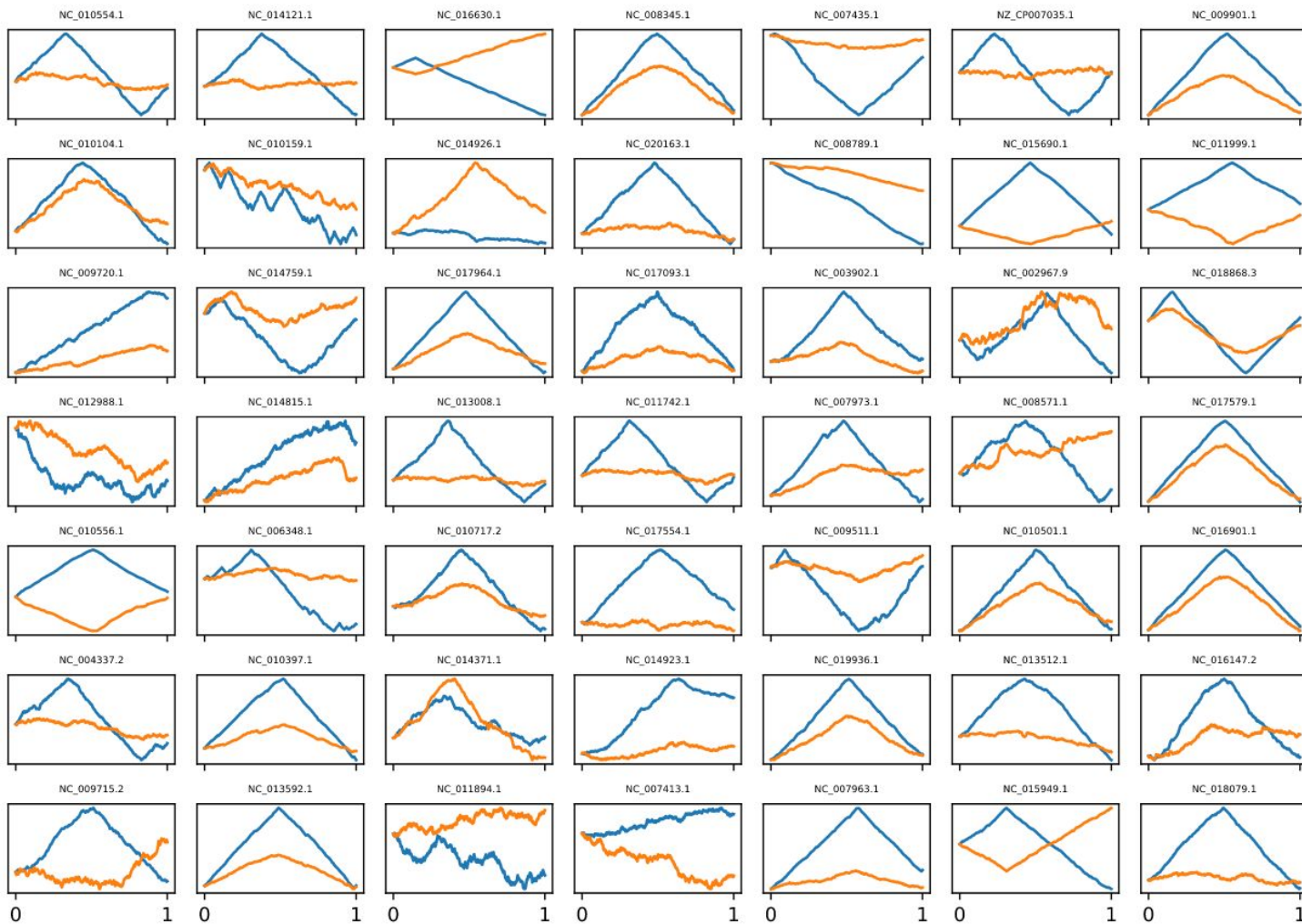


GC SKEW!

<https://skewdb.org/view>

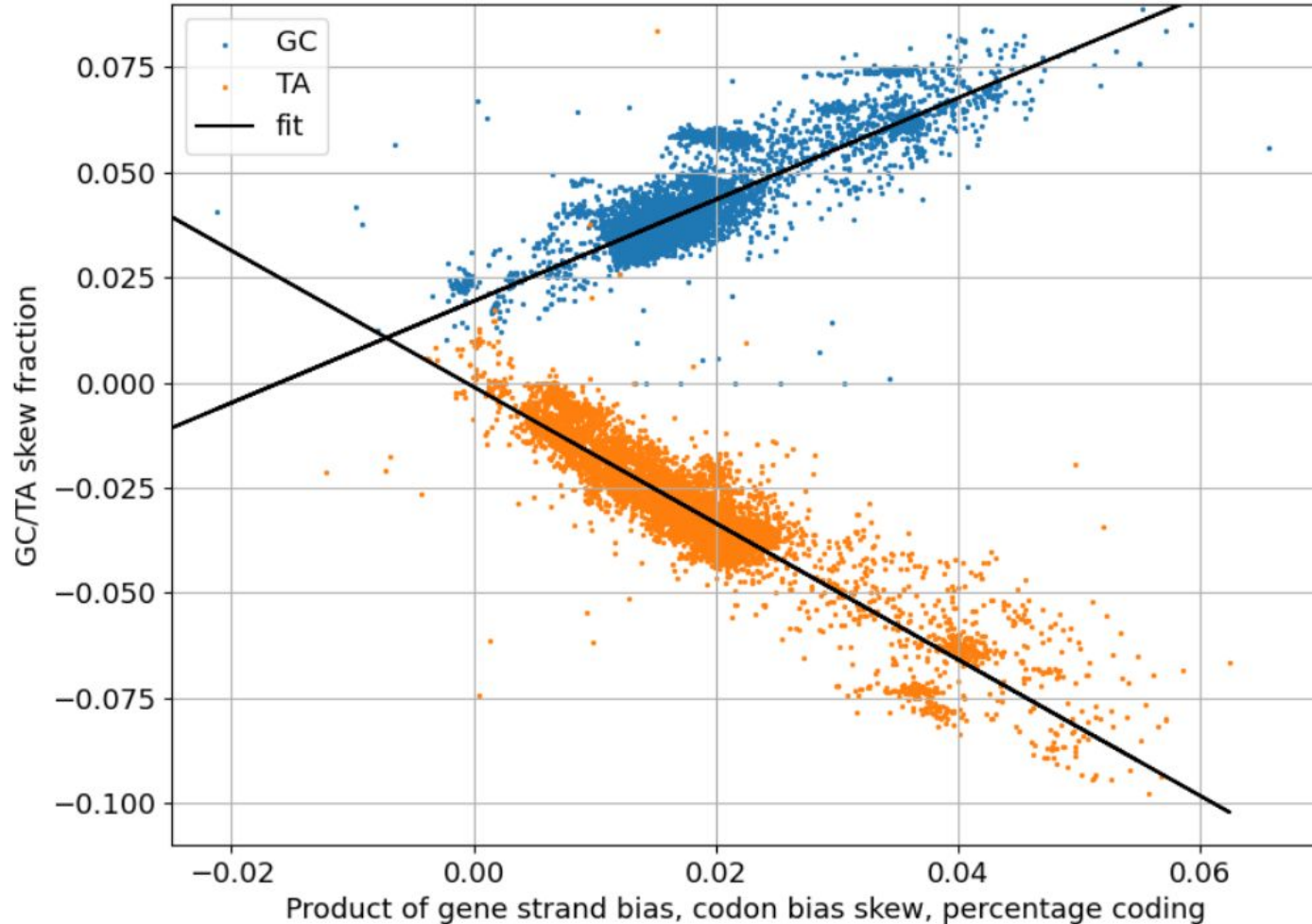


GC and TA skew in random bacterial chromosomes



**And no one
really knows
why!**

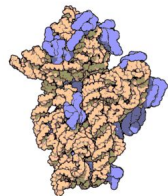
Data for 7970 Firmicute chromosomes



Lots of data
available to test
hypotheses

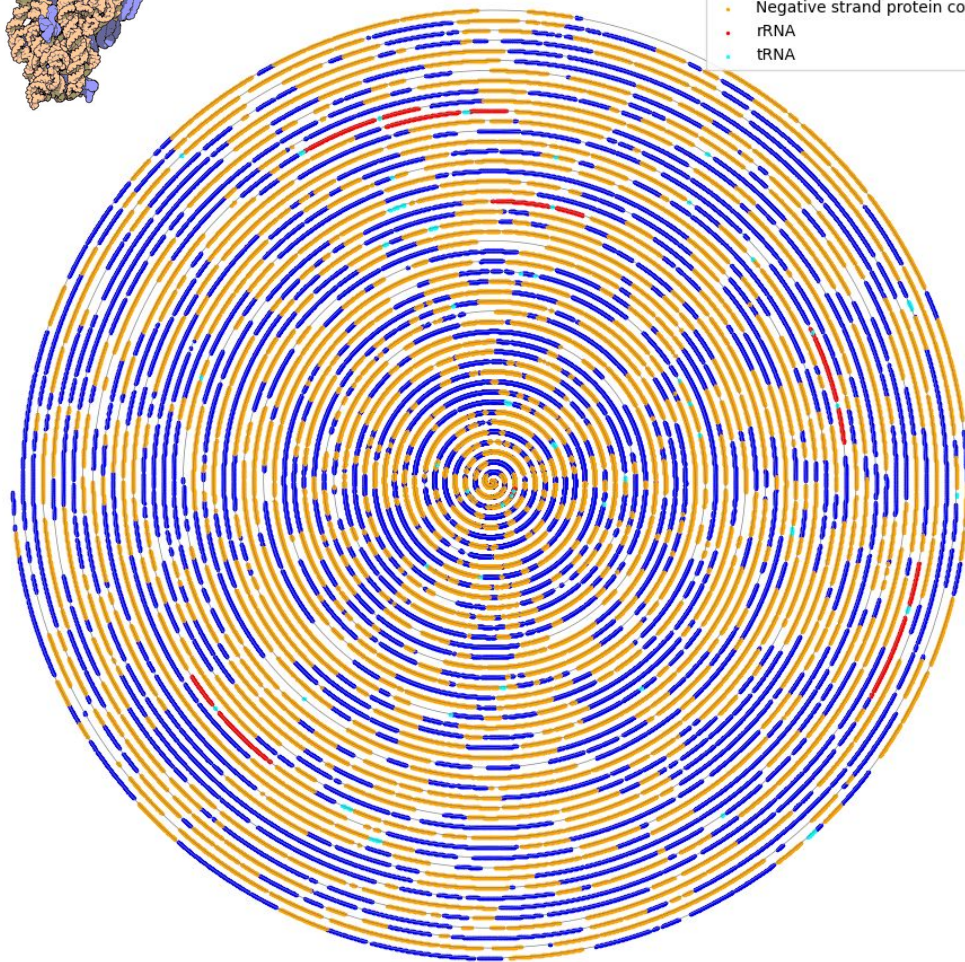
I also have a few

Have a go with the
data from
skewdb.org!



Spiral display of a bacterial chromosome

- Positive strand protein coding
- Negative strand protein coding
- rRNA
- tRNA



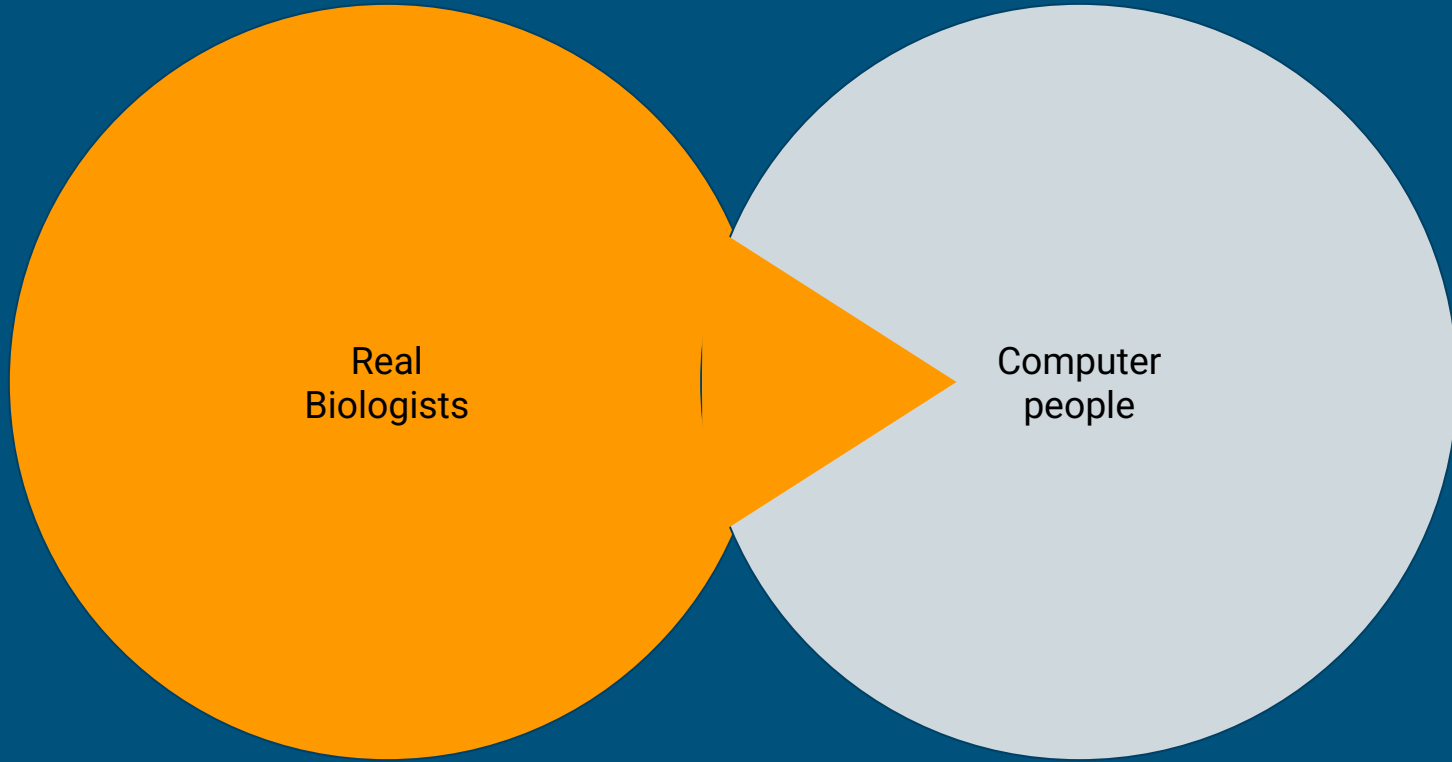
Escherichia coli str. K-12 substr.
W3110

Look at the white spaces

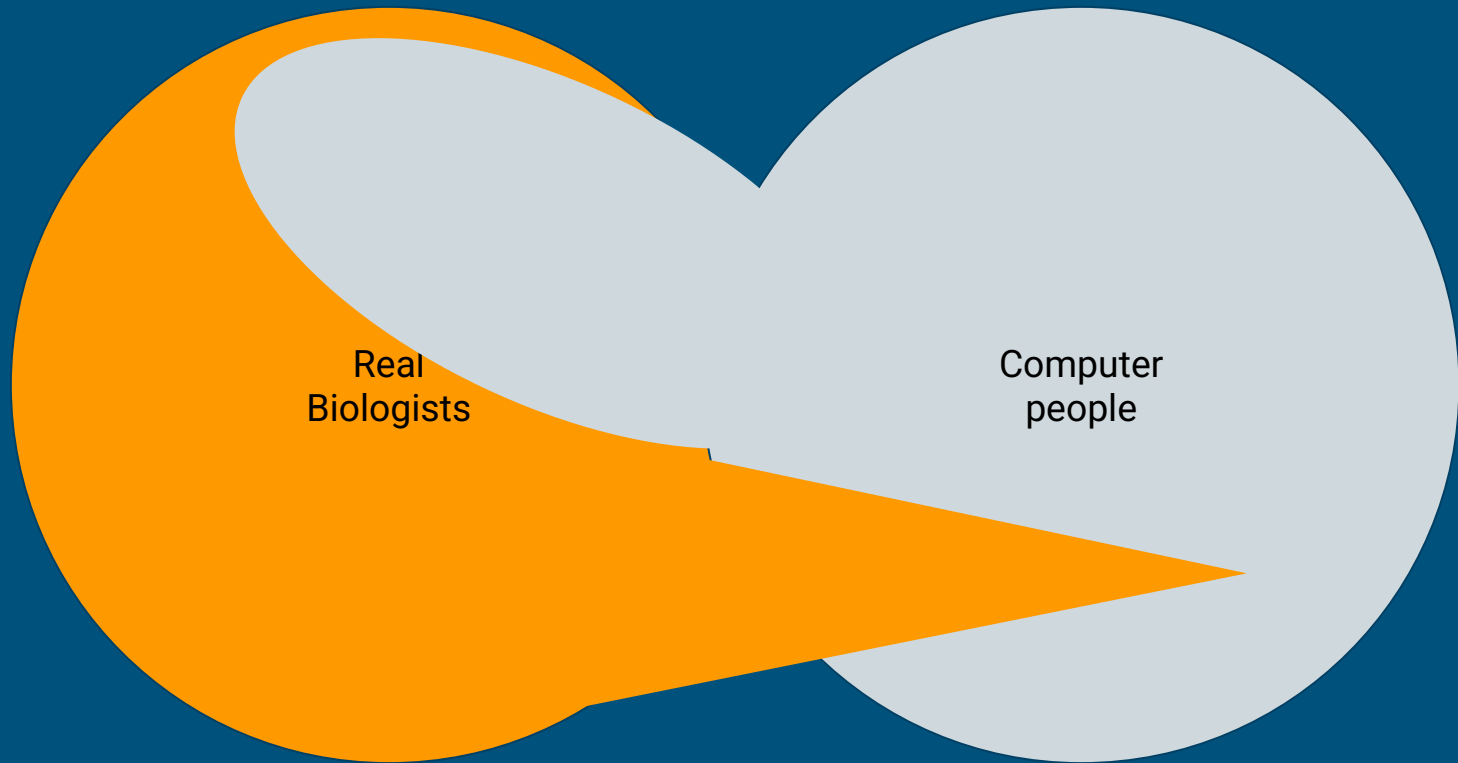
Very interesting things could be
hiding there

**And you could help find out what it
is!**

Knowledge



Knowledge



Why does GC-skew exist?

Why even Chargaff's 2nd rule?

What is in the bacterial
whitespace?

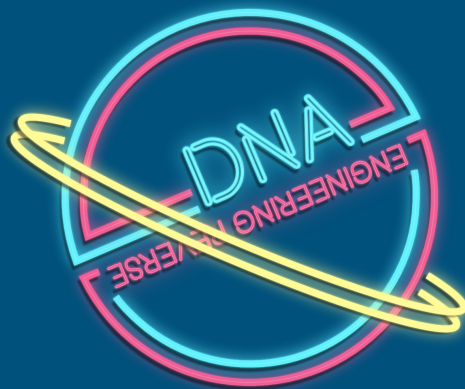
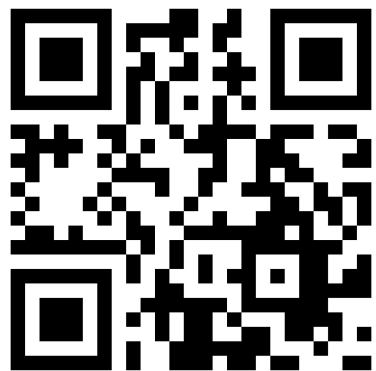
Why is so much bioinformatics
tooling so terrible?

GOOD LUCK!

Come to the afterparty!

Tomorrow, Monday, 2025-08-11 15:00–15:50, Cassiopeia

Interactive session, featuring all the skipped slides and lots of room for questions and answers!



<https://berthub.eu/revdna/>