Selective promotion of oligonucleotides in the course of evolution

Bernadette Mathew[1†], Abhishek Halder[2†], Nancy Jaiswal[1#], Smruti Panda[1#], Debjit Pramanik[1#], Sreeram Chandra Murthy Peela[1], Abhishek Garg[4], Sadhana Tripathi[1,5], Prashant Gupta[6], Vandana Malhotra[4], Gaurav Ahuja[1,3] , Debarka Sengupta[1,2,3*]

1. Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India.
2. Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India.
3. Centre for Artificial Intelligence, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India.
4. Sri Venkateswara College, University of Delhi, Delhi - 110021, India.
5. TCS Genomics Lab, Tata Consultancy Services, Lucerna Tower, Sector 125, Noida, UP, 201303, India.
6. Wellcome Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Saffron Walden CB10 1RQ, UK.

[#,†]*Equal contribution*

[*]*Corresponding author: {debarka@iiitd.ac.in}.*

Abstract:

DNA is essentially a collection of short nucleotide sequences with a myriad of functionalities. These oligonucleotides are often studied for their function as molecular binding sites, regulatory motifs or sequence probes. The field of Bioinformatics uses a similar concept under the term "*k*-mers," which is central to various tasks, like comparing the similarity of different DNA sequences in phylogenetic tree reconstruction. The present study explores if *k*-mers experience selection pressure in the course of evolution. Our meta-analysis, spanning multiple organisms, suggests positive selection of genetic changes, where the associated *k-mers* are more redundant in the genome, relative to their wild-type counterparts.

Main Text

Evolution is an ongoing process of adaptation that enables organisms to survive in a plethora of different intracellular and extracellular environments. To unravel the complexities of this process, numerous evolutionary theories have emerged since the 19th century. Among these, Charles Darwin's theory of evolution by natural selection remains the most renowned (*1*). Darwin's seminal work laid the foundation for understanding how heritable variations and differential reproductive success drive the gradual adaptation and speciation of organisms. In the 20th century, the widely accepted Modern Synthesis, or neo-Darwinism (*2*), emerged, emphasizing that genetic variations arising from mutations, recombination, and genetic drift drive evolution. Natural selection acts on these variations, favoring individuals with advantageous traits, leading to gradual adaptation and speciation. Laland and colleagues (*3*) supported this framework by proposing that evolution transcends gene-centric views, asserting that organisms actively co-construct and co-evolve with their environments, thereby altering ecosystems. Recently introduced assembly theory (*4*) proposes that complex biological structures arise from simpler building blocks following efficient assembly pathways, offering a new perspective on evolution alongside natural selection, though supporting evidence is still lacking.

To identify or delineate the key players in evolution, we need to examine molecular level changes such as mutations, which serve as the starting point for evolution. It is known that mutators, which have a high mutational burden, play significant roles in initiating crucial adaptive strategies contributing towards the evolutionary processes. Rapidly accumulating mutations expedite the exploration of mutation space, increasing the likelihood of beneficial mutations emerging and becoming prevalent within a population. The primary disadvantage of the mutator phenotype is the introduction of many detrimental mutations. However, natural selection purges these harmful mutations, resulting in a population with fewer detrimental and more neutral or beneficial mutations (*5*) (*6*). In recent years, Amosova and colleagues discovered a unique ability of short sequences of single strands of DNA in being able to perform depurination, removing nitrogenous bases from the spiral. This process, previously thought

impossible without external catalysts like RNA or proteins, was found to occur consistently and spontaneously at a specific genetic sequence appearing at least 50,000 times in the human genome (*7*). These observations highlight the remarkable flexibility of the DNA molecule and raise questions on the factors that drive these structural changes. It is noteworthy that besides serving the essential role as the repository of genetic information, DNA also contains molecular fossils that can trace our evolutionary legacy(*8*).

To investigate whether DNA directs its own evolution, we analysed oligomeric units known as *k*-mers. These *k*-mers are recognized for their highlight to identify similarities between organisms and exhibit strong correlations with specific genes. They are essential tools in genome alignment and the construction of phylogenetic relationships (*9*). We hypothesize that analyzing *k*-mer frequencies and patterns relative to ancestral sequences can reveal the significance of specific mutations that have undergone positive selection, thereby indicating their roles in adaptation and evolutionary success.

In this work, we elucidate the foundational elements of evolutionary trajectories by analyzing data from the Long-Term Evolution Experiment (LTEE) with *E. coli* and *S. cerevisiae*, along with natural evolutionary data from SARS-CoV-2. These datasets provide a temporal imprint of evolutionary processes. We investigate whether DNA evolves as a large unit or as an ensemble of smaller entities, such as oligomers. Our findings suggest that *k*-mers may serve as fundamental units of measurement for quantifying evolutionary selection based solely on an organism's ancestral information. The phenomenon, termed "**oligo promotion**" or "***k*-mer repurposing**," describes how DNA drives its own evolutionary course by selectively favoring or repurposing *k*-mers across its genome.

By examining the quasispecies theory and the divergent selection pressures on these oligomers, we aim to quantify evolutionary dynamics in diverse microorganisms, including *E. coli*, *S. cerevisiae*, and SARS-CoV-2. This approach offers a comprehensive perspective on the role of *k*-mers and other genetic elements in the evolutionary adaptation and success of organisms. We observe a strong influence of certain mutations driving evolution and conclude that oligonucleotides of length *k* could serve as a building block of evolution.

Results

Oligonucleotide are non-uniformly, and diligently maintained across genomes

DNA is a large polymer of nucleotides arranged in a specific sequence characteristic for a particular organism . Oligonucleotides are short nucleotide sequences that may serve as binding motifs for proteins or RNA and in various technical utilities (such as probes used in targeted sequencing or Polymerase Chain Reaction or PCR). "*k*-mers" are an analogous concept, often used in Bioinformatics analysis for similarity search, sequence alignment (*9*), and homology detection (*10*). In the manuscript both words are used interchangeably. Here, we investigate if *k*-mers undergo selection pressure during evolution. To answer this, we first looked into their non-uniform abundance in individual organismal genomes, and subsequently tried to explore the principle behind such favoritism.

Frequency Chaos Game Representation (FCGR) of a genome is arguably the most insightful visual representation that not only shows *k*-mer frequencies, but also displays coherent groups of *k*-mers with similar frequency levels (*11*) . We used our Python *kaos* package (*12*) to generate FCGR matrices and associated heatmaps to evaluate the potential implications of the abundance and scarcity of these frequencies, we utilized the FCGR method. As shown in Figure 1a, the FCGR captured *k*-mer frequencies, with each *k*-mer's frequency being mapped to its respective position in the FCGR map **(Fig. 1a)**, providing a comprehensive representation of sequence patterns. This fractal pattern is similar across mammalian species such as mouse (*Mus musculus*), gorilla, and human (*H. sapiens*) **(Fig. 1b-d)** while zebrafish (*D. rerio*) and yeast (*S. cerevisiae*) exhibited distinct patterns **(Fig. 1e-f)** .

Although it is known that FCGR patterns depict base composition and sequentiality, the biological significance of the specific features of each pattern is not understood (*13*). To unravel the biological significance and to substantiate the non-random nature of DNA patterns, we generated a simulated *E. coli* DNA sequence of *k*-mer length 10 by preserving the nucleotide base count throughout the sequence but rearranging their positions **(Fig. 1g)**. Visualizing the FCGR matrix, we observed that the fractal pattern is not preserved, indicating a lack of fidelity

compared to original DNA sequences **(Fig. 1h)**. This observation corroborates the findings of previous study (*14*) and suggests that DNA sequences maintain specific patterns of *k*-mer abundance and scarcity, pointing towards discernible preservation of certain *k*-mer frequencies across different organisms. Considering that *S. cerevisiae* and *E. coli* genomes seem relatively more stochastic, we hypothesize that maybe these *k*-mer patterns play a significant role in shaping the evolutionary outcome. As a contrast, Figure 1h depicts a randomized FCGR matrix for *E. coli*, which appears significantly more stochastic than the original FCGR matrix. Figure 1i presents a comparison of absolute frequency differences between similar *k*-mers (hamming distance wise) with absolute frequency differences of random *k*-mers . As expected, for highly similar *k*-mers (Hamming distance $\leq$ 2), absolute frequency differences remain significantly lower relative to the randomized scenario.

These data suggest that at a genomic scale, similar *k*-mers arising from different locations of a genome can have similar frequencies, thereby alluding to their greater functional roles and intra-DNA conservation which in turn is relatable to sequence conservation across species. This puts forth the question whether the "survival of fittest" hypothesis applies to oligonucleotides. Subsequent sections shed light on this postulate, by employing systematic *k*-mer meta-analyses across various species.

Beneficial mutations are informed by the gain in associated oligonucleotide frequencies

It could be intuitive to think oligonucleotides/*k*-mers containing an evolutionarily favored variant would be one with more number of copies as compared to the ones containing the reference allele. We developed a simple strategy to substantiate our observation and interpretation of the non-uniform frequency distribution of *k*-mers as discussed in the previous section. Towards this objective, we used various experimental and natural evolution datasets of model organisms such as (*E. coli* and *S. cerevisiae*) as well as SARS-CoV-2.

We introduced a new measure called kGain to capture the fold increase/decrease in copies of *k*-mers associated with substituted allele, relative to the reference. **Figure 2a** illustrates the same. The figure also validates the hypothesis that evolution shapes *k*-mer frequencies across

the genome. In other words, populations with the favored *k*-mers have better odds of survival. To this end, we performed meta-analysis of the genomic sequence data (*15*) from the popular *E. coli* Long-Term Evolution Experiment (LTEE), initiated by Lenski and colleagues (*16*)(*17*, *18*). Past studies determined fitness of all mutator and non mutator (six each) replicate lines and performed genome sequencing for a large number of representative generations covering a substantial part of the evolutionary timeline **(Fig. 2b)**. Mutator genomes harbored far more mutations as compared to non-mutators **(Fig. 2c)**. Both categories showed near-monotonic increase in median fitness across the replicates. Upon quantifying the Area Under the Curve (AUC) associated with the median fitness curves against generations, mutators exhibit slide edges over non-mutators. It's worth noting that the rise in fitness is more pronounced among the early generations **(Fig. 2d)**. Lenski and his colleagues also noted this, observing a 30% increase in fitness by the time they reached the 2000th generation (*16*).

We computed kGain scores for nucleotide substitution across all populations of mutators and non-mutators (Total no. of mutations: 36,923). **Figure 2e** shows kGain score distributions for initial 2000 generations. A clear rise in kGain scores can be observed in both mutators and non-mutators. In mutators, kGain scores up to 2000th generations are assigned lower kGain scores as compared to following generations. This could be due to the persistent genetic drifts among mutators, paving the way for new beneficial mutations. Non-mutators, however, do not exhibit such a tangible increase **(Fig. 2f and fig. S1.a)**. In fact, in non-mutators, median kGain scores for initial 2000 generations is relatively higher. An explanation to this could be that most beneficial mutations, in mutators, are restricted to the initial 2000 generations, some of which get fixed in the population and lead to a plateau. These observations are better understood from the kGain score dynamics along the entire evolutionary time-frame **(Fig. 2g and fig S1.b)**.

As a control, we employed ESM1b Protein Language Model (PLM) on the *E. coli* LTEE dataset. Given that protein language models (PLMs) are specialized deep-learning architectures, trained on several millions of protein sequences from one or more organisms. These models exploit self-supervised sequence representation learning strategies and can be used as foundation models across a myriad of biological problems, including spotting damaging mutations, protein

structure/function prediction and antibody design. These models can also be interrogated to find likelihood of a certain protein sequence (wild type as well as mutant), mirroring sequence probability estimators such as Markov chains. Brandes and colleagues have recently reported ESM1b, a 650 million-parameter protein language model, trained on 250 million sequences from all organisms (*19*, *20*). Authors described log-likelihood ratio (LLR) scores for this purpose. A deleterious variant is typically assigned a relatively low LLR score. We calculated LLR scores only for the missense mutations (18,059 out of 22,857) **(fig. .S1.c)** and observed trends similar to those we noted for the median kGain score **(Fig. 2h and fig. S1d, e, f)**. There is a crucial difference between the kGain and the LLR scores. The former is based on a single reference genome, whereas the latter produces sequence likelihood by observing millions of sequences from across organisms. Further, the LLR score is limited to detecting variants within the coding region and cannot capture information from variants in upstream or downstream genes, non-coding regions, or intergenic spaces.

Essential gene variants are assigned higher kGain scores

Understandably, essential genes are generally resilient to mutations as any mutation can impact the function and therefore organism's survival or fitness. Intuitively, if kGain scores truly indicate evolutionary advantage, in essential genes, we would observe a bias towards variants with higher kGain scores **(Fig. 3a and fig. S2.a)**. This is because essential gene mutations with poor kGain scores would often have deleterious impact on the organism, thereby eliminating populations carrying the variant allele. We sourced essential/non-essential categorization from the study by Gerdes et al. (*21*). Aligned to our hypothesis, we found elevated median kGain scores for single nucleotide substitutions associated with the essential genes, as compared to those associated with the non-essential ones **(Fig. 3b,c)**. The gap in kGain score levels across essential and non-essential genes is evident both in mutators and non-mutators, suggesting its universal nature. Further the difference is prominent throughout the time-course of the LTEE experiment. Median LLR scores show analogous results in case of mutators, whereas, in case of non-mutators, the differences remain insignificant **(fig. S2. b,c)**.

Next we checked whether variants persisting in populations for a substantial portion of total generations exhibit higher kGain. To explore this, we exclusively examined mutations present in more than 15.6% of generations (The value of 15.6% comes from finding how often each unique mutation appears over 109 generations, then calculating the median of these frequencies across all populations. Genotypes with mutation counts equals/below the median are categorized as non-carry forward (NCF), while those exceeding the median are categorized as carry forward (CF) mutations). We observed a significant elevation in CF kGain scores than NCF across both mutators and non-mutators **(Fig. 3d,e)**. We did not observe such patterns in case of LLR scores **(fig. S2. d,e)**.

Blount and colleagues reported that after ~33K generations, a specific mutator population labeled m3 exhibited a significant increase in turbidity, which continued to rise for several days. The authors attributed this phenomenon to the acquired ability to utilize citrate, which remains unused under oxic conditions. In m3, kGain score levels allude to a noticeable elevation after ~ 33K generations **(Fig. 3f)**.

It is known that GC pairs are generally more stable than AT pairs and thus GC-rich genomes are predictably more adapted to high temperatures than AT-rich genomes. With this in mind, we hypothesized that variants featuring GC to AT switches would be assigned higher kGain scores compared to other possibilities. This hypothesis is consistent with our observations on the *E. coli* LTEE data **(Fig. 3g)**.

Experimental evolution in *S. cerevisiae* reveals consistent kGain trends

Following the assessment of a prokaryotic system, we inquired if eukaryotes would conform with the elevation of the kGain score levels along the evolutionary time-course. To this end, we performed meta-analysis of genomic sequencing dataset from a *S. cerevisiae* LTEE experiment (*22*). The study chose 30 focal populations (12 diploid populations, 12 MATa populations, and 6 MATα populations) across three environments – yeast extract peptone dextrose (YPD) at 30°C, synthetic complete medium (SC) at 30°C, and synthetic complete medium (SC) at 37°C. We computed kGain scores for 133,538 substitutions from across the 90 populations under study.

Notably, authors performed genomic sequencing for samples from only six representative timepoints. We observed a gradual increase in both the median fitness and median kGain scores over successive generations **(Fig. 4a)**. To quantify this trend, we employed least squares regression and observed a strong correlation (Pearson's *rho* = 0.9635, R^2 = 0.928) between median fitness and median kGain **(Fig. 4b)**. Despite the absence of a rise in mutation numbers akin to those observed in the *E. coli* LTEE study, populations did not display hypermutator phenotypes, indicating steady accumulation of beneficial mutations, even in the absence of mutator phenotype.

We closely examined the fixed mutations in the populations. The authors labeled a mutation as "fixed" at a particular time point if it met the criteria of having coverage of at least 5X and being present at a frequency of greater than or equal to 40% (for diploids) or 90% (for haploids). Notably, fixed mutations were assigned higher kGain scores, as compared to ones not fixed **(Fig. 4c)**.

Pathogenic *SARS-CoV-2* variants are assigned elevated kGain scores

The recent outbreak of SARS-CoV-2 arguably offers the most extensive data on natural evolution, supported by comprehensive large-scale genomic profiling initiatives. To comprehensively analyze the mutational landscape of the SARS-CoV-2 virus, we implemented a robust pipeline for mutation analysis, as illustrated in **(fig. S3a)**. We took around 0.5 million high-quality genome sequences from GISAID (last accessed on: 29-08-2022) and made SNP calls, yielding ~54,000 unique SNPs. Within our analysis, we delineated gene length and unique mutation counts **(fig. S3.b,c and table S1)**. We employed two widely recognized approaches to classify SARS-CoV-2 variants: functional categorization of associated genes and their pathogenic characteristics with implications for human health. Functionally, genes are divided into structural, non-structural, and accessory classes. Structural genes, such as Spike (S), Envelope (E), Membrane (M), and Nucleocapsid (N), are essential for the structural integrity and infectivity of the virus. These genes facilitate viral entry into host cells by binding to the ACE2 receptor and contribute to immune evasion and viral persistence. Non-structural genes, exemplified by ORF1ab, are crucial for viral replication and transcription which are vital for its

lifecycle and interaction with the host immune system. Accessory genes, including ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10, enhance the ability of the virus to evade immune responses and modulate host cell functions such as antagonizing the host's interferon response **(Fig. 5a)**. Based on pathogenic roles and public health implications, the World Health Organization (WHO) has classified SARS-CoV-2 variants as Variants of Interest (VOI) or Variants of Concern (VOC) **(table S2)**. VOIs have mutations affecting virus behavior, including transmissibility, severity, detection, or treatment. VOCs meet VOI criteria and cause severe disease, significantly reducing vaccine effectiveness and impact healthcare systems. Examples of lineages associated with VOIs include lineages like Lambda and Mu, while those associated with VOCs include Alpha, Beta, Gamma, Delta, and Omicron **(fig. S3d)**. More details on COVID-19 data processing and filtering can be found in the Materials and Methods section.

We observed a significant contrast between the control variants and the VOC/VOI variants, with the latter exhibiting substantially higher kGain scores **(Fig. 5b)**. Interestingly, gene groups revealed that kGain scores for VOC/VOI variants were higher than those for control variants, particularly within accessory and structural genes **(Fig. 5c)**. We set the kGain cutoff at 0 and divided the variants into two categories: <0 and ≥0. Variants in the latter group were found to be present in considerably higher numbers of COVID-19 genome sequences, suggesting that our score could indicate the type of selection pressure (positive or negative) experienced by a mutation.

Zhang et al. (2022) reported mutations that contribute to the pathogenicity and fitness of the virus. We examined the kGain profile of the reported VOC/VOI mutations and found that ORF3a:S26L, ORF3a:T223I, and ORF3a:S171L had higher kGain scores, whereas ORF3a:Q57H and ORF3a:S253P had lower kGain scores. These variants highlight changes in accessory proteins, which may exhibit a heightened propensity for adaptation through immune evasion.

Another study by Obermeyer et al. (2022) established that the SARS-CoV-2 pandemic has been dominated by several genetic changes of functional and epidemiological importance, including the spike (S) D614G mutation. This mutation is associated with higher SARS-CoV-2 loads and has contributed to the increased infectivity and fitness of the virus **(Fig. 5e)**

Discussion

The present study establishes that oligonucleotides, or *k*-mers, of specified length, serve as fundamental genetic units underpinning evolutionary processes. The positive selection of advantageous oligonucleotides suggest that maintaining these patterns may facilitate adaptation of the organism to the dynamic changes in the environment. Leveraging the enhanced FCGR Python package for DNA encoding (*12*), we identified fractal patterns that are associated with wide-spread *k*-mer conservation within genomic sequences. These findings lay the groundwork for further exploration into the structural organization and evolutionary dynamics of *k*-mers. We hypothesize that within a species, oligonucleotides compete and undergo selective pressures, leading to the fixation of those with larger effect sizes and superior competitive fitness. This dynamic explains the observed abundance of certain *k*-mers and the scarcity of others. To quantify evolutionary advantages in the *E. coli* LTEE, we compared mutator and non-mutator phenotypes on various levels. Firstly, we observed the fitness landscape, noting that the AUC of generation-wise median fitness of mutators is slightly higher than that of non-mutators. Secondly, in both mutators and non-mutators, essential genes exhibit higher median kGain scores compared to non-essential genes. Lastly, carry-forward mutations are predominantly associated with higher median kGain scores than non-carry-forward mutations in both mutators and non-mutators. Collectively, these evidences observed across multiple organisms (*E. coli*, *S. cerevisiae*, and SARS-CoV-2), suggest that the kGain score is indicative of whether a *k*-mer contributes favorably to the adaptive strategies driving the evolutionary fitness of an organism.

Historically, evolution of any species is marked by a mutator phenotype wherein the accelerated rate of mutation increases the likelihood of random genetic variations, leading to two possible outcomes. One is the accumulation of deleterious mutations, which can result in being purged from the population. Conversely, beneficial mutations can become fixed, leading to the change becoming dominant within the population (*6*,*23*).

The evolution of SARS-CoV-2 exemplifies how the host-pathogen interactions shape the spontaneous emergence of mutations and new lineages that exhibit higher fitness during replication. Notably, the ORF3a gene is recognized for its diversity and functional adaptation throughout viral evolution. Specific variants of ORF3a, including ORF3a:S26L, ORF3a:T223I, ORF3a:S171L, ORF3a:Q57H, and ORF3a:S253P, have been identified as integral to the overall fitness of the virus in a study by Zhang et al. (2022). By using kGain scores, we can infer which of these variants might be favored during viral evolution.

Our analysis reveals that accessory gene mutations such as ORF3a:S26L, ORF3a:T223I, and ORF3a:S171L are associated with higher kGain scores compared to ORF3a:Q57H and ORF3a:S253P. The former mutations are located in highly conserved regions, such as β-sheets or the S1-S2 domain, where deleterious mutations are typically omitted from the population, and beneficial ones confer survival advantages. In contrast, ORF3a:Q57H and ORF3a:S253P are situated near the C- and N-termini, which are less conserved and thus more tolerant to mutations that do not significantly impact protein function. This differential conservation likely explains the higher kGain scores observed for ORF3a:S26L, ORF3a:T223I, and ORF3a:S171L. Of relevant interest is the D614G mutation within the Spike protein of SARS-CoV-2 that was found to enhance infectivity and transmissibility relative to the original strain largely due to increased affinity for host cell receptors. The higher kGain scores associated with the D614G mutation possibly explain the influence on viral behavior.

A key finding of this study is the robustness of the kGain score that quantifies "*k*-mer repurposing" within an organism's genome, indicating the degree of self-conservation in DNA sequences. This approach incorporates coding and non-coding regions along with ancestral information, to measure the extent to which specific *k*-mers are reused or repurposed. There are other methods such as SIFT, PolyPhen, and PROVEAN, as well as deep learning approaches like ESM1b and ESM1v, which capture conservation across multiple species and are primarily focused on coding regions.

Our findings have subtle resemblance with the recently introduced assembly theory of evolution. According to the assembly theory, sequences that have been favored assemble based

on past information and are thus retained. In our work we propose that the DNA holds this information via *k*-mer such that if the alternative *k*-mer is recognized to have more abundance than the reference *k*-mer then it is highly likely that the mutation containing the alternate *k*-mer would be fixed in the population giving it an evolutionary advantage.

Figures

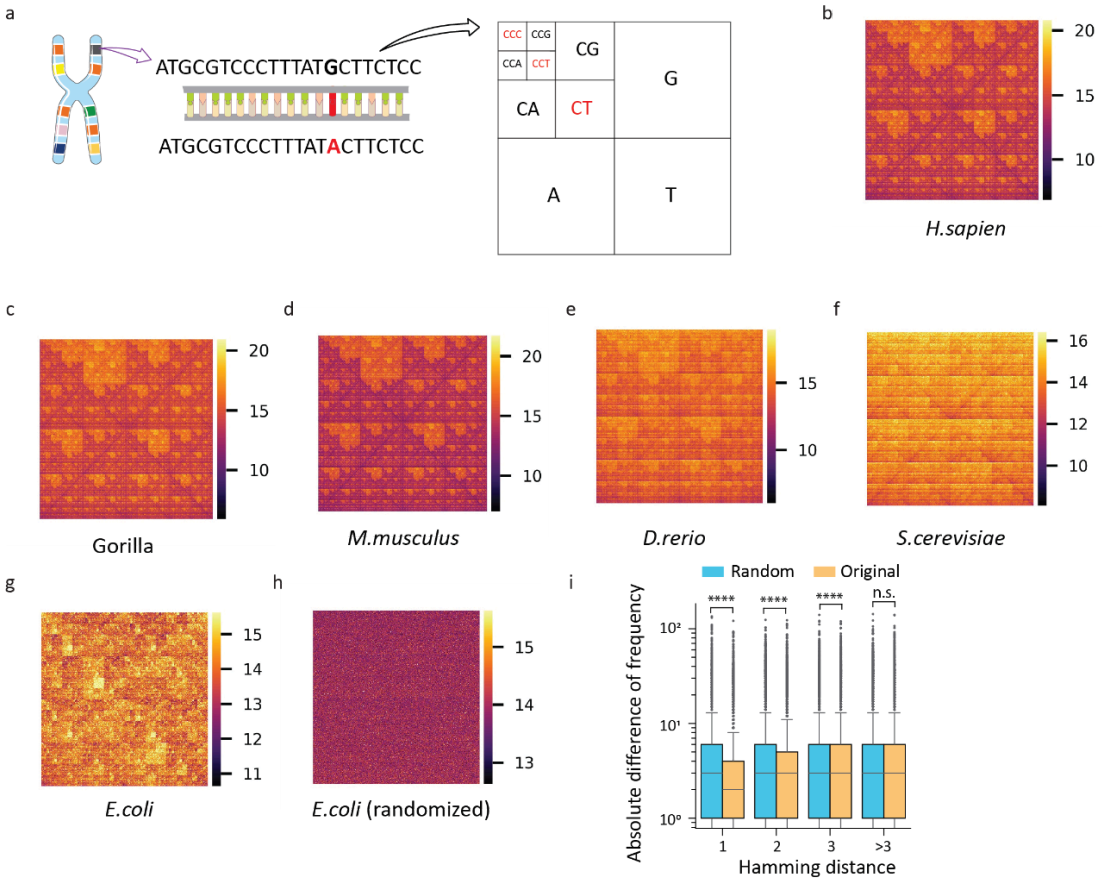
a
ATGCGTCCCTTTATGCTTCTCC
ATGCGTCCCTTTATACTTCTCC
CCC
CCG
CCA
CCT
CG
CA
CT
G
A
T
b
H.sapien
c
Gorilla
d
M.musculus
e
D.rerio
f
S.cerevisiae
g
E.coli
h
E.coli (randomized)
i
Random
Original
****
n.s.
Absolute difference of frequency
Hamming distance
1
2
3
>3

Figure 1

Fig 1. Exploring the fractal pattern of DNA across various organisms and examining its non-random nature.

(a) The genomic DNA is fragmented into oligonucleotides of a defined length k, which are subsequently mapped to the FCGR matrix of the reference genome. **(b-f)** Heatmaps illustrating normalized frequency values (in negative logarithmic scale) to reduce the range which demonstrates that various organisms exhibit distinct patterns of *k*-mer abundance within their DNA sequences. **(g-h)** To evaluate the significance of these fractal patterns, we utilized a heatmap of genomic DNA from *E. coli*, comparing it to a randomized counterpart, thereby elucidating discernible differences in pattern preservation. **(i)** Boxplot showing the absolute difference (log-scale) in frequency distribution between two randomly selected *k*-mers and one randomly selected *k*-mer with its counterpart altered by n Hamming distances (original altered)

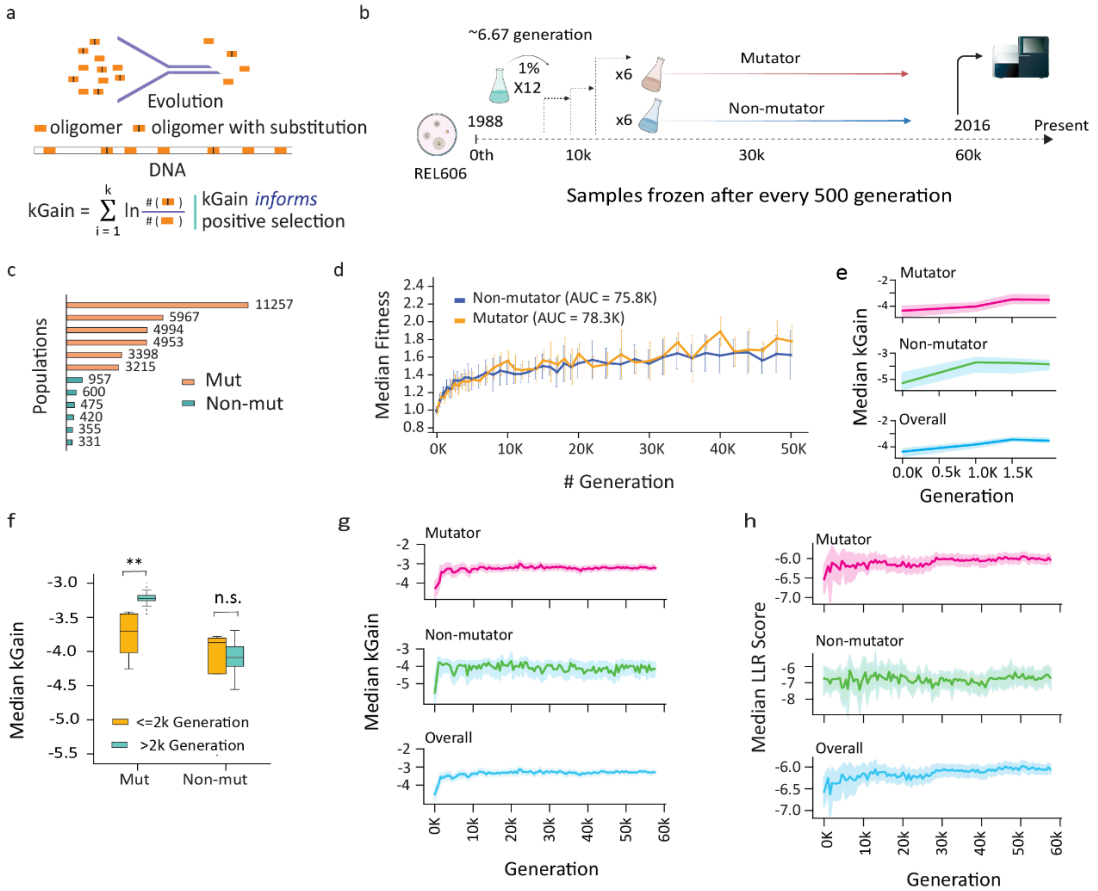
a
Evolution
oligomer
oligomer with substitution
DNA
kGain = $\sum_{i=1}^{k} \ln \frac{\#(\blacksquare)}{\#(\blacksquare)}$
kGain informs positive selection
b
~6.67 generation
1%
X12
x6
x6
Mutator
Non-mutator
1988
2016
Present
0th
10k
30k
60k
REL606
Samples frozen after every 500 generation
c
Populations
11257
5967
4994
4953
3398
3215
957
600
475
420
355
331
Mut
Non-mut
d
Median Fitness
Non-mutator (AUC = 75.8K)
Mutator (AUC = 78.3K)
Generation
e
Median kGain
Mutator
Non-mutator
Overall
Generation
f
Median kGain
**
n.s.
<=2k Generation
>2k Generation
Mut
Non-mut
g
Median kGain
Mutator
Non-mutator
Overall
Generation
h
Median LLR Score
Mutator
Non-mutator
Overall
Generation

Figure 2

Fig 2. Evolutionary Dynamics

(a) A visual depiction illustrating that, under bottleneck conditions, survival is limited to populations with specific *k*-mer abundances that are associated with higher kGain scores, which undergo high positive selection pressure. **(b)** A visual depiction of the Long-Term Evolution Experiment (LTEE) involving *Escherichia coli* B strain REL606. Initiated in 1988 by Lenski et al., this experiment has spanned over 30 years and includes 12 replicate populations labeled ara-1 to ara-6 (m1 to m6) and ara+1 to ara+6 (p1 to p6). These populations were propagated for 60,000 generations (1998-2016) and subjected to sequencing at intervals of 500 generations. **(c)** After several generations, some of the population developed a hypermutator phenotype, known as mutators, while the rest were known as non-mutators,captured in the barplot. **(d)** Median fitness remains relatively constant across all generations in non-mutators while the same is highly dynamic (with higher AUC) among the mutators**. (e)** The line plot represents the median kGain scores across generations up to 2,000 for mutator and non-mutator phenotypes, as well as the entire population. **(f)** The box plot between mutator and non-mutator phenotypes is categorized based on whether they occurred before or after 2,000 generations. **(g)** Median kGain scores across all generations for mutator, non-mutator, and the overall population. **(h)** Median LLR scores across all generations for mutator, non-mutator, and the overall population.

Note: The P-value cutoff for all the plots is 0.05. *, **, ***, and **** refer to P-values <0.05, <0.01, <0.001, and <0.0001, respectively.

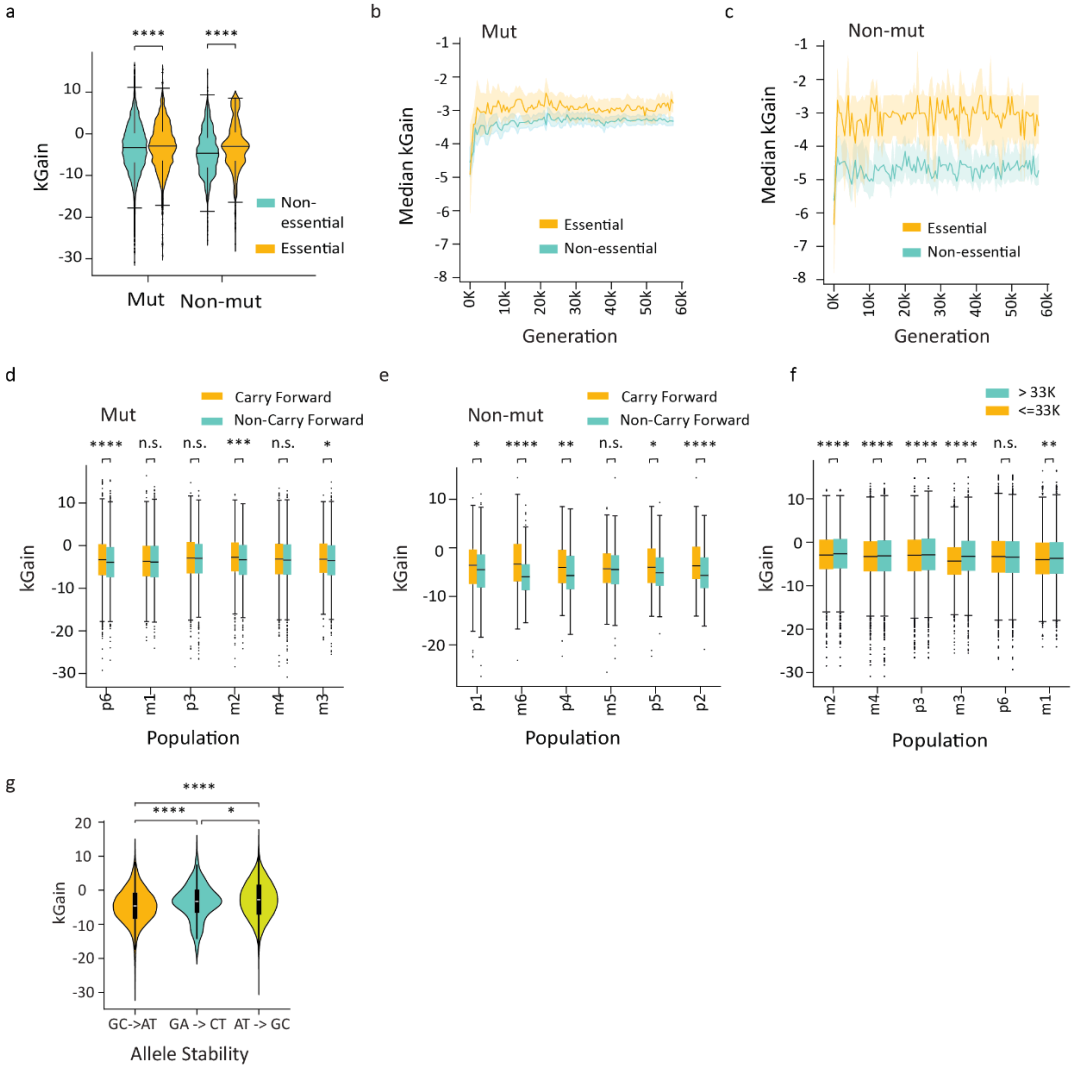
a
****
****
kGain
Non-essential
Essential
Mut
Non-mut
b
Mut
Median kGain
Essential
Non-essential
Generation
c
Non-mut
Median kGain
Essential
Non-essential
Generation
d
Mut
Carry Forward
Non-Carry Forward
****
n.s.
n.s.
***
n.s.
*
kGain
p6
m1
p3
m2
m4
m3
Population
e
Non-mut
Carry Forward
Non-Carry Forward
*
****
**
n.s.
*
****
kGain
p1
m6
p4
m5
p5
p2
Population
f
> 33K
<=33K
****
****
****
****
n.s.
**
kGain
m2
m4
p3
m3
p6
m1
Population
g
****
****
*
kGain
GC->AT
GA -> CT
AT -> GC
Allele Stability

Figure 3

Fig 3. Beneficial mutational spectrum shift

(a) Violin plot depicting overall differences in kGain scores of essential/non-essential genes among mutator and non-mutator populations. **(b-c)** Generation-wise differences in median kGain scores of essential and non-essential genes among mutators and non-mutators populations. **(d-e)** The box plot presented here captures the kGain score of variants CF and NCF during the course of its evolution between mutator and non-mutator populations. **(f)** The box plot shows the kGain scores in the mutator population, divided at the 33k generation mark, which is when citrate utilization was first observed in the population.**(g)** A violin plot of kGain score vs. allele stability illustrates the kGain distribution across three categories: Unstable (G/C switching to A/T), Controls (DNA complementary), and Stable (A/T switching to G/C). One-sided Mann-Whitney U test p-values were applied between stable vs. control, stable vs. unstable, and control vs. unstable.

[**Note**: The p-value cutoff for all the plots is 0.05. *, **, ***, and **** refer to p-values <0.05, <0.01, <0.001, and <0.0001, respectively]

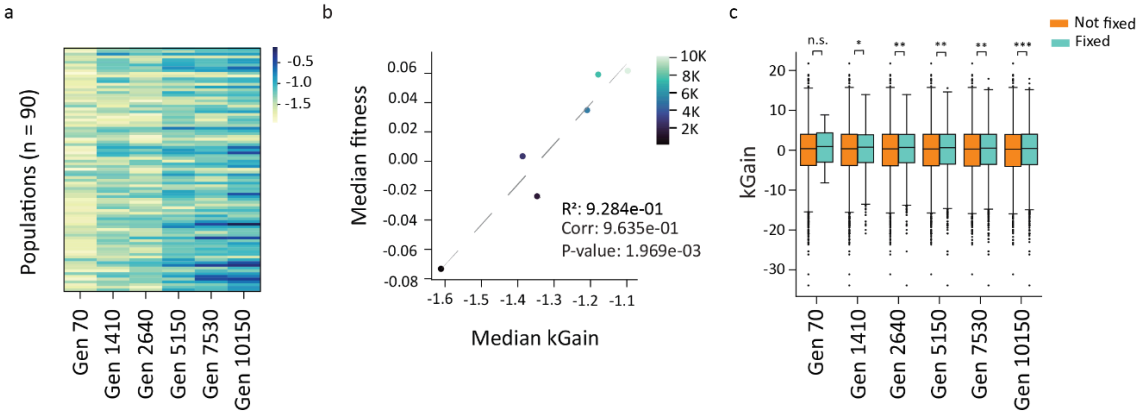
a
Populations (n = 90)
-0.5
-1.0
-1.5
Gen 70
Gen 1410
Gen 2640
Gen 5150
Gen 7530
Gen 10150
b
Median fitness
Median kGain
10K
8K
6K
4K
2K
R²: 9.284e-01
Corr: 9.635e-01
P-value: 1.969e-03
c
Not fixed
Fixed
kGain
n.s.
*
**
**
**
***
Gen 70
Gen 1410
Gen 2640
Gen 5150
Gen 7530
Gen 10150

Figure 4

Fig 4. Mutational spectrum of Yeast LTEE experiment.

(a) Heatmap depicting an increasing trend in the median kGain score across the six time points where color intensity indicates the magnitude of the median kGain score. **(b)** For every available generation, we computed the median kGain and median fitness across the entire population. These values were then visualized in the form of a scatter plot, where the color gradient represents generation. **(c)** Boxplot was constructed to depict mutations that were fixed in the population across all six time points. Each box represents the distribution of fixed mutations, with "Fixed" indicating those that were fixed and "Unfixed" indicating those that were not fixed in the population.

[**Note**: The P-value cutoff for all the plots is 0.05. *, **, ***, and **** refer to P-values <0.05, <0.01, <0.001, and <0.0001, respectively]

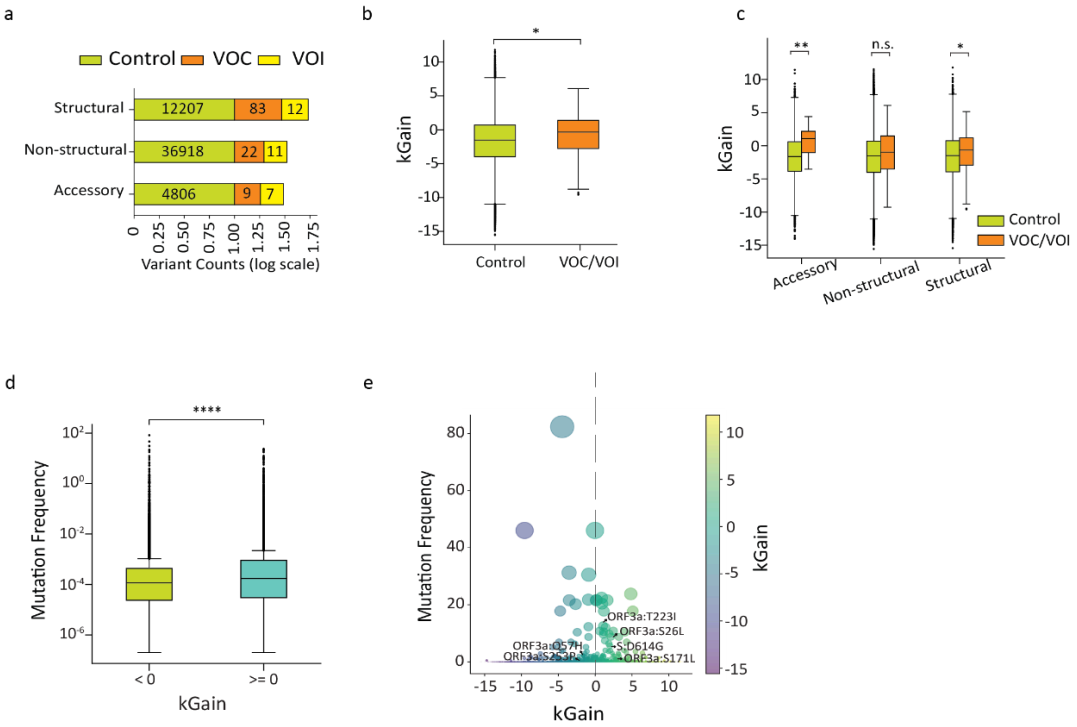
a
Control
VOC
VOI
Structural
Non-structural
Accessory
12207
83
12
36918
22
11
4806
9
7
0
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Variant Counts (log scale)
b
*
kGain
Control
VOC/VOI
c
**
n.s.
*
kGain
Accessory
Non-structural
Structural
Control
VOC/VOI
d
****
Mutation Frequency
< 0
>= 0
kGain
e
Mutation Frequency
ORF3a:T223I
ORF3a:S26L
S:D614G
ORF3a:Q57H
ORF3a:S171L
kGain
kGain

Figure 5

Fig 5. kGain scores to understand dynamics of SARS-CoV2 genomes

(a) The variant counts are depicted using a logarithmic scale for control, VOC and VOI categories within the structural, non-structural and accessory gene groups. **(b)** We calculated the kGain values for each variant and categorized them into VOC (Variants of Concern) and VOI (Variants of Interest), collectively forming the "VOC/VOI" group. The remaining variants were labeled as "control". These values were then represented using box plots, with the color gradient indicating the influence of kGain on both control and VOC/VOI. **(c)** All variants are classified into accessory, structural and non-structural based on gene and variant classification (control and VOC/VOI) and then visualized with kGain using box plots. **(d)** kGain of 0 is taken as cutoff and based on that, variants <0 and => 0 are visualized with log of mutation frequency using box plot.**(e)** Scatter plot for Mutation frequency vs kGain scores showing some of the important genes which are moving towards adaptation and are pathogenic.

[**Note**: The p-value cutoff for all the plots is 0.05. *, **, ***, and **** refer to p-values <0.05, <0.01, <0.001, and <0.0001, respectively]

Acknowledgements

We would like to express our gratitude to Dr. Michael Desi from the Department of Organismic and Evolutionary Biology at Harvard University, Cambridge, and Dr. Milo Johnson, Postdoctoral Fellow at UC Berkeley, for their invaluable assistance in understanding and acquiring the data.

Fundings

DS acknowledges the support of the iHub-Anubhuti-IIITD Foundation set up under the NM-ICPS scheme of the DST. DS. also acknowledges support from the SERB (CRG/2022/007706) and DBT (IC-12044(12)/4/2022-ICD-DBT)

Authors contributions

Conceptualization: DS.Methodology: DS, BM, AH.Investigation: BM, AH, SP, NJ, DP, ST, AG.Visualization: BM, AH, SP. Supervision: DS, GA, VM. Writing – original draft: DS ,BM ,AH. Writing – review & editing: DS, BM, AH, SMP, SP, AG, GA, VM.

Competing interests

DS is a co-founder and shareholder at GeneSilico, Inc. and Dkraft Pte. Ltd.

Data and materials availability

The kGain analysis code and data for *E. coli* and *S. cerevisiae* are available on GitHub https://github.com/cellsemantics/Oligo_Promotion. LLR analysis code and data for *E. coli* is available on GitHub https://github.com/cellsemantics/esm1b. The kGain analysis code and data for *SARS-CoV-2* are available on Github https://github.com/cellsemantics/Oligo_Promotion. GISAID sequence data for *SARS-CoV-2* is publicly available at GISAID Initiative (epicov.org). Source code for variant calling, ORF prediction and genome annotation are available at rintukutum/strainflow_manuscript (github.com) and boxiangliu/covseq: CoV-Seq: COVID-19 Genomic Sequence Database and Visualization (github.com). Reference genomes for *SARS-CoV-2* are sourced from Severe acute respiratory syndrome coronavirus 2 isolate

Wuhan-Hu-1, co - Nucleotide - NCBI (nih.gov). The mutation count and lineage data are available at Welcome to the Python Outbreak.info package docs! — Python Outbreak.info API 0.1 documentation (outbreak-info.github.io). WHO guidelines for *SARS-CoV-2* variant classification are available at COVID-19 variants | WHO COVID-19 dashboard and SARS-CoV-2 Variant Classifications and Definitions (cdc.gov).

Supplementary Materials

This PDF includes material and methods along with Figs. S1 to S3 and Table S1 - S2

REFERENCE AND NOTES

1. D. Partridge, Darwin's two theories, 1844 and 1859. *J. Hist. Biol.* **51**, 563–592 (2018).

2. Fisher, R. Aylmer, Sir, 1890-, 009: The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).

3. K. Laland, T. Uller, M. Feldman, K. Sterelny, G. B. Müller, A. Moczek, E. Jablonka, J. Odling-Smee, G. A. Wray, H. E. Hoekstra, D. J. Futuyma, R. E. Lenski, T. F. C. Mackay, D. Schluter, J. E. Strassmann, Does evolutionary theory need a rethink? *Nature* **514**, 161–164 (2014).

4. A. Sharma, D. Czégel, M. Lachmann, C. P. Kempes, S. I. Walker, L. Cronin, Assembly theory explains and quantifies selection and evolution. *Nature* **622**, 321–328 (2023).

5. G. Lobinska, Y. Pilpel, Y. Ram, Phenotype switching of the mutation rate facilitates adaptive evolution. *Genetics* **225** (2023).

6. F. Taddei, M. Radman, J. Maynard-Smith, B. Toupance, P. H. Gouyon, B. Godelle, Role of mutator alleles in adaptive evolution. *Nature* **387**, 700–702 (1997).

7. O. Amosova, R. Coulter, J. R. Fresco, Self-catalyzed site-specific depurination of guanine residues within gene sequences. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 4392–4397 (2006).

8. S. Olson, *Evolution and the Biosphere* (National Academies Press (US), 1989; https://www.ncbi.nlm.nih.gov/books/NBK218603/).

9. A. Sievers, K. Bosiek, M. Bisch, C. Dreessen, J. Riedel, P. Froß, M. Hausmann, G. Hildenbrand, K-mer Content, Correlation, and Position Analysis of Genome DNA Sequences for the Identification of Function and Evolutionary Features. *Genes* **8** (2017).

10. M. L. MacDonald, S. W. Polson, K. H. Lee, k-mer-Based Metagenomics Tools Provide a Fast and Sensitive Approach for the Detection of Viral Contaminants in Biopharmaceutical and Vaccine Manufacturing Applications Using Next-Generation Sequencing. *mSphere* **6** (2021).

11. H. F. Löchel, D. Heider, Chaos game representation and its applications in bioinformatics. *Comput. Struct. Biotechnol. J.* **19**, 6263–6271 (2021).

12. A. Halder, Piyush, B. Mathew, D. Sengupta, Improved Python Package for DNA Sequence Encoding using Frequency Chaos Game Representation, *bioRxiv* (2024)p.

2024.04.14.589394.

13. K. A. Hill, N. J. Schisler, S. M. Singh, Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. *J. Mol. Evol.* **35**, 261–269 (1992).

14. D. Lichtblau, Alignment-free genomic sequence comparison using FCGR and signal processing. *BMC Bioinformatics* **20**, 742 (2019).

15. B. H. Good, M. J. McDonald, J. E. Barrick, R. E. Lenski, M. M. Desai, The dynamics of molecular evolution over 60,000 generations. *Nature* **551**, 45–50 (2017).

16. R. E. Lenski, Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *ISME J.* **11**, 2181–2194 (2017).

17. E. Crozat, N. Philippe, R. E. Lenski, J. Geiselmann, D. Schneider, Long-term experimental evolution in Escherichia coli. XII. DNA topology as a key target of selection. *Genetics* **169**, 523–532 (2005).

18. H. Jeong, V. Barbe, C. H. Lee, D. Vallenet, D. S. Yu, S.-H. Choi, A. Couloux, S.-W. Lee, S. H. Yoon, L. Cattolico, C.-G. Hur, H.-S. Park, B. Ségurens, S. C. Kim, T. K. Oh, R. E. Lenski, F. W. Studier, P. Daegelen, J. F. Kim, Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). *J. Mol. Biol.* **394**, 644–652 (2009).

19. A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021).

20. N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, V. Ntranos, Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).

21. Supplementary Online Material: Sveta Gerdes et al. https://www.genome.wisc.edu/Gerdes2003/genetic.html.

22. M. S. Johnson, S. Gopalakrishnan, J. Goyal, M. E. Dillingham, C. W. Bakerlee, P. T. Humphrey, T. Jagdish, E. R. Jerison, K. Kosheleva, K. R. Lawrence, J. Min, A. Moulana, A. M. Phillips, J. C. Piper, R. Purkanti, A. Rego-Costa, M. J. McDonald, A. N. Nguyen Ba, M. M. Desai, Phenotypic and molecular evolution across 10,000 generations in laboratory budding yeast populations. *Elife* **10** (2021).

23. H. A. Orr, Fitness and its role in evolutionary genetics. *Nat. Rev. Genet.* **10**, 531–539 (2009).