Ancient eukaryotic immunity through genome editing of viral sequences

Lisa Mettrop[1], Anna Lipzen[2], Gilles Mirambeau[1], Kerrie Barry[2], Igor V. Grigoriev[2,3], Gwenaël Piganeau[1], Marc Krasovec[1*]

[1] CNRS, Microbial Biodiversity and Biotechnology Laboratory (UMR 8176), Sorbonne Université, Banyuls sur Mer, France.

[2] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, California, USA.

[3] Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, California, USA

[*]**Corresponding author**: marc.krasovec@obs-banyuls.fr
Keywords: Spontaneous mutation, Mutation rate, *Bigelowiella natans*, Virus, Genome integrity.

ABSTRACT

Mutations provide the raw material for evolution, but mutation rates are not uniform across genomes. Using a mutation accumulation experiment in the marine phytoplankton *Bigelowiella natans*, we discovered extreme local variation in mutation rate: over 1000-fold differences across its nuclear genome. While the baseline single-nucleotide mutation rate is approximately 3.5×10^{-10} per site per generation, a common value for unicellular species, two genomic regions derived from ancient integrated viruses exhibit strikingly elevated rates of about 6×10^{-7}. These regions show a distinctive mutational signature with almost exclusively T/A→C/G transitions, a pattern also found in other virus-derived sequences, contrary to the usual GC to AT mutation bias. Notably, hypermutation occurs only on TA dinucleotides, and only in a subset of experimental lines, suggesting a regulated process rather than random genomic instability. We propose that *B. natans* targets invading DNA through localized hypermutation, reminiscent of ADAR and APOBEC antiviral defense systems in animals. Our results suggest that genome-editing–based immunity may have deep evolutionary roots, predating the divergence of major eukaryotic lineages. This is in line with the idea of ancestral immunity, a concept that argues a common and very ancient origin of immune systems, in particular against viruses, which probably exist since the first cell.

SIGNIFICANCE STATEMENT

De novo mutations provide the raw material for adaptation, but at exceptionally high frequencies they can compromise genome integrity. Here, we describe a hypermutable process targeting an integrated viral genome in a chlorarachniophyte alga, reminiscent hypermutation-based antiviral defenses described in humans against HIV and influenza; whereby host-mediated deamination of the viral genome increases its mutation rate such that the virus loses its infectivity. The idea of genome-editing–based defenses shared in very divergent branches of eukaryotes supports the ancestral immunity hypothesis, which suggests a deeply conserved origin of immune mechanisms in the tree of life.

INTRODUCTION

De novo mutations continually change the genome, sustaining the capacity for evolutionary innovation. The rate of spontaneous mutation (μ), that is the number of mutations per base per generation, and the mutation spectrum, the relative frequency of different types of mutations, define this new diversity. μ is a quantifiable parameter under the influence of both selection and drift, and has been shown to span four orders of magnitude of variation within the tree of life (1). The mutation rate also presents large intracellular variations. In particular, eukaryotic cells contain multiple genomic compartments such as nuclear and mitochondrial genomes, as well as plastid genomes in the case of photosynthetic organisms. For each of these compartments, distinct μ have been estimated. For example, in metazoans and unicellular eukaryotes, μ_{mito} is usually estimated to be higher than $\mu_{nuclear}$ (2–4) while in land plants μ_{mito} is significantly lower than $\mu_{nuclear}$ (4). In contrast, $\mu_{chloroplast}$ is usually lower than $\mu_{nuclear}$ in both multicellular and unicellular photosynthetic organisms (4). Moreover, within the nuclear compartment itself, different types of sequences or regions with specific μ have been observed. For example, transcription can change the spontaneous mutation rate of coding sequences (5) and local mutation rate may be influenced by DNA methylation (6), chromatin structure (7) or histone marks (8). Significant differences in spontaneous mutation rates have also been observed between chromosomes in *Vibrio* bacteria (9). Another interesting sub-compartment is formed by sequences of exogenous origin, such as DNA viruses, which are integrated in many unicellular eukaryotic genomes (10, 11). Normally, viruses have a rapidly evolving genome, with mutation rates of around 10^{-6} -10^{-4} mutations per base per strand copy for RNA viruses and 10^{-8}-10^{-6} for DNA viruses (12), while eukaryotic genome mutation rates mostly range from 10^{-10} to 10^{-8} mutations per base per replication (1). However, the replication of

endogenized viruses being dependent on the host cell, it could be that local mutation rate of integrated sequences approaches the global mutation rate of the host genome.

Here, we conducted a mutation accumulation (MA) experiment in *Bigelowiella natans*. In MA experiments, population sizes of MA lines are kept extremely small with frequent bottlenecks (13). In this way, effective population size (N_e) stays sufficiently small that selection is negligeable; hence, mutations are fixed independently of their effect on fitness. *B. natans* is a chlorarachniophyte, unicellular marine alga in the supergroup Rhizaria. Their chloroplast comes from a secondary endosymbiosis with a green alga. Interestingly, remnants of the symbiont nuclear genome are preserved as a nucleomorph, forming a fourth sub-cellular genomic compartment located between the outer membrane pairs of the *B. natans* chloroplast (14). Moreover, the genome of *B. natans* contains numerous sequences of viral origin, predominantly derived from nucleocytoplasmic large DNA viruses (NCLDV) and virophages (15). The object of the present work is to study intracellular variation in mutation rate and spectrum, both between genomic compartments and within the nuclear genome in *B. natans*.

RESULTS

Mutation accumulation experiment

The MA experiment lasted a total of 378 days with 27 bottlenecks, resulting in 15 successfully sequenced MA lines for a total of 3,132 accumulated generations. The callable nuclear genome (G^*) is 85,048,595 nucleotides (nt), which corresponds to 90% of genome size (14). Callable genomes for organelles and nuclear sub-compartments are detailed in Tables 1 and S1; no *de novo* mutations were found in the organellar genomes. In the nuclear genome, a total of 302 single nucleotide mutations (SNM) were identified, as well as 21 insertions-deletions (ID) (Table S2). The SNM presented an extreme heterogeneity in their distribution in the nuclear genome (Table 1, Fig. S1). Additionally, using raw read coverage as a proxy for sequence copy number, we found three candidates for *de novo* duplication (Fig. S2): in MA line 19 there was a large duplication in the middle of scaffold 53 from approximately position 169,900 to 470,200 (300,300bp), in MA line 15 a part of scaffold 61 was duplicated from approximately position 258,900 to 591,300 (332,400bp), and in MA line 21 the whole scaffold 74 was duplicated (544,257bp).

Table 1. G^* and μ per compartment.

Compartments	G* (% of total)	SNM	ID	μ_{SNM} (CI95%)	μ_{ID} (CI95%)	μ_{total} (CI95%)	GC%
Nuclear NCLDV sc10	176492 (97%)	211	0	$3.75x10^{-7}$ (3.26-$4.29x10^{-7}$)	-	$3.82x10^{-7}$ (3.32-$4.37x10^{-7}$)	49.71
Nuclear virophage sc146	11164 (45%)	20	0	$5.72x10^{-7}$ (3.49-$8.83x10^{-7}$)	-	$5.72x10^{-7}$ (3.49-$8.83x10^{-7}$)	50.09
Remaining nuclear	84860939 (90%)	71[a]	21	$3.46x10^{-10}$ (2.79-$4.26x10^{-10}$)	$7.90x10^{-11}$ ($4.8921x10^{-11}$-$1.21x10^{-10}$)	$3.46x10^{-10}$ (2.79-$4.25x10^{-10}$)	44.87
Nucleomorph	137369 (37%)	-	-	-	-	-	30.01
Chloroplast	29920 (43%)	-	-	-	-	-	33.11
Mitochondrion	36267 (100%)	-	-	-	-	-	42.16

[a]4 SNM are in a single multi-nucleotide mutation

Hypermutable regions of viral origin

There are two hypermutable regions in the nuclear genome that together contain 231 out of the 302 nuclear SNM. The larger region is on scaffold 10, with 211 AT→GC SNM between positions 336,000-521,500 (±186 kb). This region has previously been reported to have NCLDV origins (15), although we extend it by ±21 kb here, based on the continuous high mutation rate. It also stands out from other NCLDV insertions in *B. natans* by its size and its transcriptional silence (15). A second, smaller hypermutable region is present on the end of scaffold 146 (positions 17,000-41,997, ±25 kb), with 20 exclusively AT→GC SNM. This region as well is densely packed with virus-like ORFs, stemming from virophage genomes, and does not differ from other virophage insertions by size (Fig. S3) or gene expression (15). These two regions (hereafter called 'viral regions') have extremely high mutation rates in the order of 10^{-7} mutation per site per generation, three orders of magnitude higher than the rest of the nuclear genome (Table 1, Table S4, Fisher's exact test p-value < $2.2x10^{-16}$); the nuclear μ_{SNM} outside these viral regions is $3.46x10^{-10}$ and μ_{ID} is about 2.6 times lower than μ_{SNM} (Table 1). Interestingly, the hypermutation of viral regions was not observed in all MA lines (Tables 2 and S3): of the 15 lines, 11 showed hyper-mutation of A and T residues on scaffold 10, ranging from 2 to 45 SNM in a single line. On scaffold 146 only 6 lines had mutations in the viral region and all of these also had mutations on scaffold 10. This suggests that the putative mutational mechanism at work is not always active.

Table 2. Number of mutations per MA line.

μ_{SNM} viral considers all SNM in both NCLDV and virophage regions with the combined G^*_{viral} = 187,656 nt.

Line	Generations	SNM nuclear	SNM NCLDV	SNM virophage	ID nuclear	μ_{SNM} viral
1	100	5	35	8	0	2.29×10^{-6}
12	252	6	14	2	2	3.38×10^{-7}
13	101	5	9	0	2	4.75×10^{-7}
14	118	2	2	1	0	1.35×10^{-7}
15	137	7	0	0	0	0
16	133	3	8	3	1	4.41×10^{-7}
18	150	1	3	0	0	1.07×10^{-7}
19	292	11	45	0	2	8.21×10^{-7}
2	278	2	0	0	0	0
20	397	2	0	0	0	0
21	401	6	10	0	3	1.33×10^{-7}
3	133	5	44	5	2	1.96×10^{-6}
17	228	10[a]	29	1	4	7.01×10^{-7}
6	271	5	12	0	5	2.36×10^{-7}
7	141	1	0	0	0	0

[a]4 SNM are in a single multi-nucleotide mutation

Striking change in mutation spectrum

In most known species, the mutation spectrum is AT-biased, meaning that G and C mutate faster than A and T (16). While the spectrum of SNM in the *B. natans* nuclear genome outside of the viral regions is also AT-biased ($\mu_{GC>AT}/\mu_{AT>GC}$ = 2.65, Table S5), the viral regions present extreme biases towards GC (Fig. 1). In the virophage region, the 20 SNM are exclusively A→G and T→C; In the NCLDV region, 210 out of 211 SNM are A→G or T→C (Table S2), giving a mutation bias of $\mu_{GC>AT}/\mu_{AT>GC}$ = 0.006. Interestingly, all of these GC-biased SNM in the viral regions fell on TA dinucleotides (Table S2). In the case of A→G transitions, the mutating A was preceded by a T (T<u>A</u>→T<u>G</u>). The T→C transitions are the reverse complement of this, with each mutated T followed by an A (<u>T</u>A→<u>C</u>A).

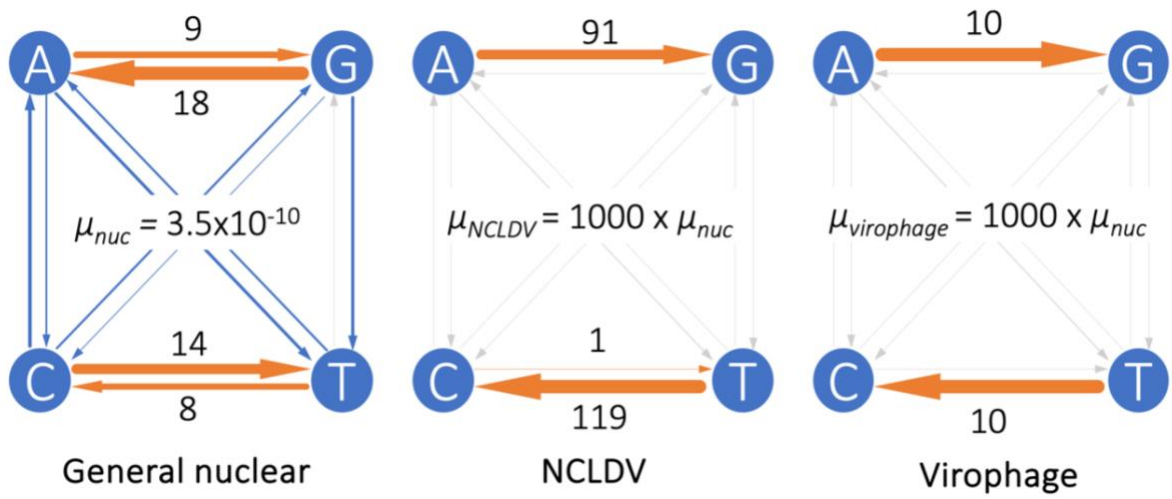

Fig. 1. SNM spectrum in the different nuclear regions. While the general nuclear genome presents all 4 types of transitions plus 7 transversions ($\mu_{AT\rightarrow GC}$ = 1.30x10^{-10} and $\mu_{GC\rightarrow AT}$ = 3.44x10^{-10}, thus AT-bias = 2.65, Ts/Tv = 7), SNM in the viral regions are almost exclusively A→G and T→C transitions, with $\mu_{AT\rightarrow GC}$ = 6.85x10^{-7} and $\mu_{GC\rightarrow AT}$ = 3.90x10^{-9} in the NCLDV region and $\mu_{AT\rightarrow GC}$ = 9.02x10^{-7} and $\mu_{GC\rightarrow AT}$ = 0 in the virophage region.

Dinucleotide frequencies in B. natans genes

If the mutation spectrum with TA motif we observe here indeed reflects an evolutionary active targeted mutational mechanism, traces of its signature on previous insertion events should be noticeable in the *B. natans* genome, present as TA depletion and CA+TG enrichment.

First, we looked at genes previously identified as having transpoviron, virophage or NCLDV origins in *B. natans* (15), 197 genes in total. The mean relative frequency of CA+TG dinucleotides in these 197 genes was 0.153, significantly higher than the mean frequency of 0.137 for all nuclear genes (Wilcoxon rank sum test, p-value 4.17x10^{-8}). Surprisingly, the mean relative TA frequency in the 197 genes was 0.052, also significantly higher than the mean frequency of 0.041 in all genes (Wilcoxon rank sum test, p-value 5.39x10^{-5}).

Second, we looked at the distribution of TA dinucleotides in all predicted *B. natans* transcripts using the relative frequency of each dinucleotide and found a subgroup depleted in TA (Fig. 2A, Fig. S4, average TA-frequency 0.0018, max TA-frequency 0.005). There was no clear subgroup enriched in CA+TG. 324 transcripts fall in the low TA group (Table S6) and a total of 245 different Pfam domains was found for 115 of them. Double hits for the same sequence were removed, and the frequency of each Pfam in the TA depleted transcripts was compared to the frequency of the Pfam in all 21,708 transcripts (Fisher's exact test with Benjamini-Hochberg correction for multiple testing). Four Pfam domains were significantly overrepresented in TA depleted sequences (adjusted p-value <0.05, Table S7). Of these, three

Pfam domains (DDE-superfamily endonuclease, Endonuclease-reverse transcriptase, and Reverse transcriptase) are consistent with a viral or transposable element ancestry and the fourth is a Zinc-finger motif.

Third, we did a similar analysis with four adjustments (Methods). Using 5%-extreme cut-off values, we extracted the genes that were both TA-depleted and CA+TG-enriched (Table S8). This time, 56 sequences had Pfam hits, for a total of 119 different Pfams. 49 Pfams had significant adjusted p-values, with enrichments ranging from 0 to 53.74 (Table S9). Of the Pfams with an enrichment higher than 0, 11 suggested foreign origins, 24 were of likely host origin and 11 others were ambiguous. Three of the four previously found Pfams were recovered. The Pfams with a high fold enrichment were mainly of foreign origin: of the 17 significant Pfams with >25-fold enrichment in the studied subset, 8 were almost certainly of viral, transposon, bacterial or fungal origins.

Dinucleotide frequencies at whole genome scale

We then extended our search for sequences with low TA and high CA+TG to the whole genome using 1kb windows. Plotting Rho_{TA} versus the Rho of other dinucleotides shows that for a supgroup of windows, Rho_{CA+TG} increases strongly as Rho_{TA} decreases in a linear fashion (Fig. S5). To increase specificity, we re-calculated Pearson regression coefficients for all windows with Rho_{TA} <0.5. Rho_{TA} variations below 0.5 are significantly correlated to all possible combinations of dinucleotides (Fig. 2B). However, the negative relationship between Rho_{TA} and Rho_{CA+TG} is the clearly the strongest (Rho=-0.5, p=0), suggesting that for genomic regions that are depleted in TA, CA+TG increases in a linear fashion.

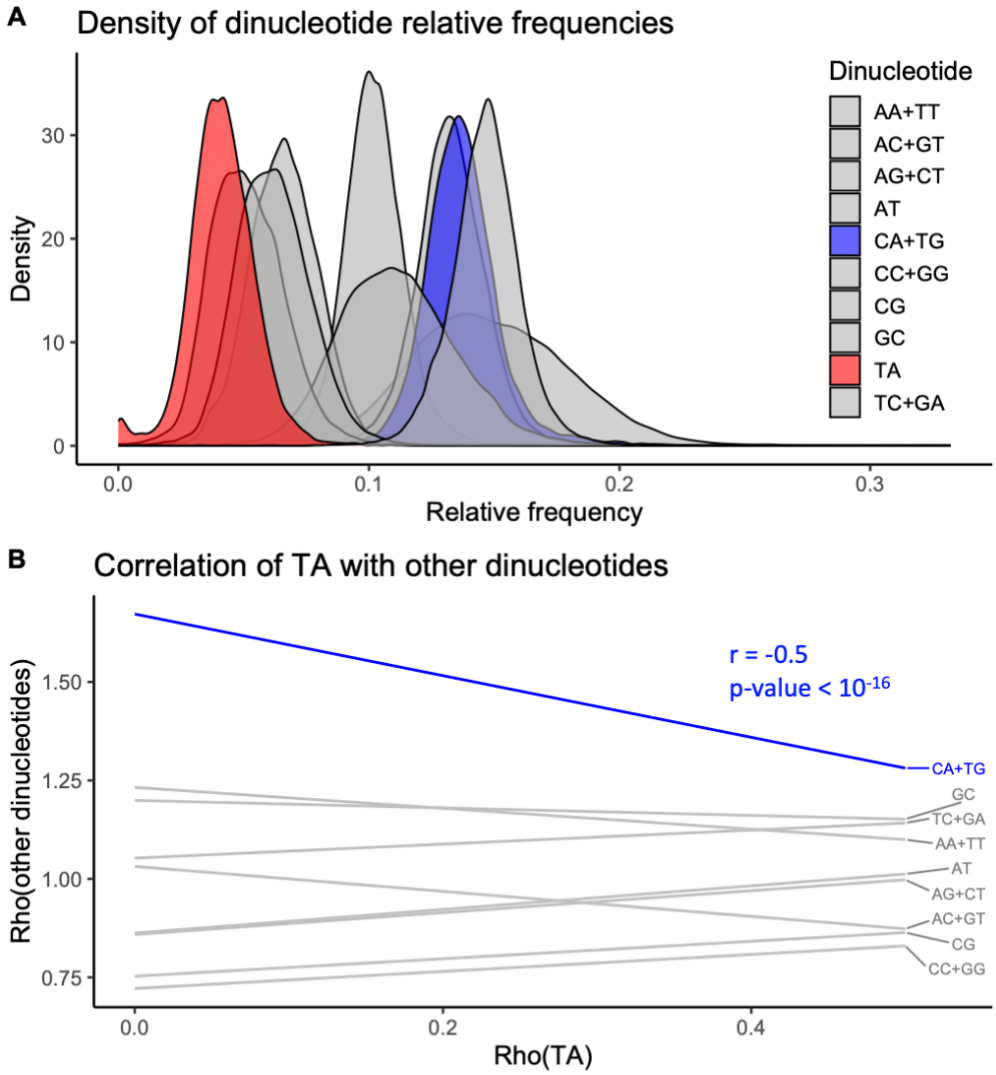

Fig. 2 Dinucleotide analyses. A) Distribution of the 10 different dinucleotide groups in the whole transcriptome; TA is the only group to have a visible depleted sub-population. B) Pairwise relationships between Rho_{TA} and other dinucleotide Rho in the whole genome per 1kb genomic intervals, only considering windows with low TA frequency (Rho_{TA} <0.5). Red and grey lines represent linear regressions. The regression in blue highlights the correlation between Rho_{TA} and Rho_{CA+TG}, the dinucleotide group with the strongest negative correlation to TA; for this correlation, the Pearson correlation coefficient (r) with associated p-value is displayed in blue.

Then, we tested whether the TA-depleted/CA+TG-enriched windows form clusters. Again, we used 5% cut-off values to define these windows ($Rho_{TA} \leq 0.43$ and $Rho_{CA+TG} \geq 1.52$, Fig. S6). 'Clusters' were defined as at least three subsequent TA-depleted/CA+TG-enriched windows, based on cluster length distribution (Figs. S7 and S8, Table S10). To evaluate whether the number of clusters was greater than expected by chance, we performed 10,000 permutations, shuffling Rho values within scaffolds. No permuted dataset produced as many clusters as observed (max 3 clusters versus 124 observed clusters, permutation p-value <0.0001). Interestingly, the whole scaffold 226 is a cluster (5 windows, 4360 bp). The 124 observed clusters contain 96 genes of which 18 have IPR annotations: 4 are viral/transposon related, 2

others possibly have bacterial origins (Table S11). The remaining 78 proteins had no annotation, including the one on scaffold 226. Clusters contain a significantly higher proportion of non-annotated proteins than the general genome (12,577/21,707, Fisher's exact test p-value = 2.7×10^{-5}).

Signatures of long-term hypermutation in the* T_0 *genome

To see if our observations reflect long-term hypermutability, we analyzed the genetic diversity in the NCLDV region between the reference *B. natans* genome (14) and our T_0. Only SNPs in intergenic regions were considered. 161 SNPs were located in the NCLDV region on scaffold 10, resulting in a high diversity of 4.31×10^{-3} SNP per nucleotide. For the remaining nuclear genome this was 1.58×10^{-4} with 3,866 SNPs, which is 27 times lower than on the viral region (Fisher's exact test, p-value $< 2.2 \times 10^{-16}$). The SNP spectrum was highly biased as well, with an AT-bias 10 times lower in the NCLDV region than in the remaining nuclear genome (0.14 and 1.33, respectively, Fisher's exact test, p-value $< 2.2 \times 10^{-16}$) (Table S12).

Then, for both viral regions, the GC-content is ~5% higher than in the remaining nuclear genome (Table 1, Chi-square test, p-values < 0.003). The NCLDV region being of Mimivirus origins (15), its original GC-content should be low: generally, Mimiviridae have GC-contents ranging from 21.5 to 40.5%, with most genomes around 30% (retrieved from ncbi.nlm.nih.gov/datasets/genome on 13 Aug. 2025). Thus, the current high GC-content of the viral regions suggests that the integrated viruses have been subjected to hypermutation. However, TA mutational targets persist, since 3.3% (N = 6169) of dinucleotides in the NCLDV region are TA in the T_0 strain.

Adenosine (A)-deaminase candidates in B. natans

One explanation for the observed high TA→CG SNM rate in viral regions could be the activity of an A-deaminase. Adenosine can be deaminated into inosine, which is recognized as guanine during the next replication round (17). Thus, we searched for a DNA A-deaminase candidate in *B. natans*. In metazoans, certain A-deaminases, called ADAR, can hyperdeaminate viral RNA as a defense mechanism (18) but in amphioxus, BjADAT2 has recently been shown to edit DNA as well (19). Only one of the predicted *B. natans* proteins (Bn78255) possessed the conserved domains described for BjADAT2 (19) (File S1), most importantly the HxE-PCxxC motif and an N at position 119 (N113 in BjADAT2). This N seems pivotal for DNA editing capacity, as shown experimentally in *E. coli* EcTadA A-deaminase (20). We aligned the structural AlphaFold predictions of Bn78255 and BjADAT2 in Pymol and found an average

RMSD of 0.644 Å, meaning the protein structures are highly similar (Fig. S9). Since BjADAT2 works together with BjADAT3 for DNA editing in the nucleus (19), we looked for a homologue of BjADAT3 and found one candidate (Bn77986, 37.4% protein sequence identity over 155 amino acids).

DISCUSSION

A defense against foreign sequences?

The extreme mutation rate and GC-mutation-bias in the viral regions, very different from the rest of the nuclear genome, indicate that specific mutational processes are at work. Since the hypermutable sequences are of viral origin, this hypermutation may be as a host defense mechanism against invading sequences. Indeed, Pfams of TA-depleted and CA+TG-enriched transcripts more frequently suggest foreign origin than other transcripts, although many enriched Pfams are common eukaryotic domains. Three elements must be considered when interpreting these results. First, many TA-depleted/CA+TG-enriched genes lack annotation, likely because they are already highly degraded. Second, transposons and integrated viruses are notoriously hard to annotate and may therefore be overrepresented in unannotated sequences (21, 22). Third, some domains, like Zinc-fingers, are intrinsically rich in TA, which is why we targeted sequences that are both TA-depleted and CA+TG enriched. Altogether, our results suggest that hypermutation activity extends to other foreign sequences beyond viruses. It is interesting that both virophage and NCLDV regions mutate at roughly the same rate, since these viruses play different roles during infection. NCLDV are known to prey on protists (23) and play an important role in biogeochemical cycles by lysis of eukaryotic phytoplankton (24). Virophages, on the other hand, are miniscule viruses that hijack NCLDV machinery (25) and have been proposed to act as a host defense system against NCLDV (15). Here, the putative hypermutation mechanism does not make any distinction between NCLDV and virophage.

Inconsistent hypermutation of viral sequences

Hypermutation of viral sequences was not observed in all MA lines, which suggests that the mechanism is not constitutive. Then, why did it activate in some lines? There were no signs of viral infection during the MA experiment. However, a threat could come from the genome itself through the activation of transposable elements or lysogenic viruses, that integrate into the host genome and persist latently. Lysogenic viruses in unicellular eukaryotes have sparked interest recently; especially virophages are suspected of having lysogenic lifecycles (15, 25) but lysogenic NCLDV have been described (26, 27). It is also possible that the hypermutation

system was triggered by epigenetic mutations (28); however, such changes were not tracked in this study.

Genome editing, an ancestral immunity system in eukaryotes?

The targeted mutation of TA dinucleotides could correspond to an adenosine-deaminase activity, which causes A→G mutations. Enzymatic deamination could also explain the specificity for the TA-motif in viral regions, and is reminiscent of immune systems in animals: ADAR proteins been shown to act on RNA sequences of influenza and measles viruses as a defence in humans (18). The similarities between the ADAR in metazoans and the genome editing in *B. natans* can suggest a common and ancient origin of this editing system. In addition, ADAR is not the only deaminase implicated in viral defense. In metazoans, the conserved AID/APOBEC proteins are also involved in immunity. They form a diverse group of enzymes that deaminate cytosines in both RNA and DNA substrates, leading to C→T mutation (29). The discovery of APOBEC was a key step in understanding cellular immunity to viruses, identifying hypermutation by DNA deamination as a mechanism inhibiting HIV replication (30). APOBEC-like genes have been found out of metazoans, notably in dinoflagellates, green algae, slime molds and haptophytes (31). The origin of APOBEC in eukatyotes has been proposed to come from bacterial toxin systems, operating as mutators of hyper-variable genes, viruses and selfish elements (32). The similarities between these deamination systems and the observations in *B. natans* are consistent with the recently proposed concept of ancestral immunity, which posits that many eukaryotic immune modules originates from prokaryotic defenses (33). Thus, although immune systems are often considered clade-specific, this concept argues for a common and very ancient origin, in particular against viruses, which probably exist since the first cell (34, 35). The ancestral immunity theory may explain the recurrence of similar defense systems across the tree of life and provides a compelling rationale for the resemblance between genome editing defenses in metazoans and *B. natans.*

Conclusion

Here, we propose an often-overlooked role of a hypermutation mechanism in the chlorarachniophyte *B. natans*: safeguarding genome integrity against foreign sequences. These exciting prospects need to be studied further but suggest the existence of a specialized TA hyper-deamination pathway as a *B. natans* defense against integrating viruses and potentially transposable elements. The similarity with systems discovered more than 20 years ago in

humans by studying HIV and influenza genome hypermutability is consistent with the concept of ancestral immunity, which explains the conservation of immune modules across deeply divergent branches of eukaryotes. Together, these insights underscore the deep evolutionary roots of antiviral defense.

MATERIALS AND METHODS

Mutation accumulation (MA) experiment

Bigelowiella natans strain RCC623 (synonym to CCMP621) came from the Roscoff culture collection (36); its identity was confirmed by 18S rDNA PCR coupled with Sanger sequencing prior to the experiment. Cells were cultivated in L1 medium (Bigelow L1 Culture Medium Kit) generated from sterile natural sea water with salinity of 33 g/L in a light:dark cycle of 12:12 hours at 20°C. The mutation accumulation (MA) was done with a flow cytometry protocol adapted to phytoplankton, where single cells can be isolated by dilution after estimation of cell density (37). To generate a clonal ancestral population (T_0) for the MA experiment, a single cell was isolated and left to propagate for 2 weeks. Then, cell density was measured again, which allowed for the sampling of a volume corresponding to 10 cells, that was added to 10 mL of medium and used to inoculate 6 wells for each MA line on a 48-wells plate with 1 mL per well. The 6 replicates limit the loss of MA lines due to technical errors or lethal mutations. Single-cell bottlenecks were performed every 14 days, by isolating 10 cells by dilution as described above from the first well with living cells encountered for each MA line; each bottleneck, 6 replicates from the same well were inoculated.

The cytometric count of phytoplankton populations gives access to the number of generations g between single-cell bottlenecks with $g=e(\log N_t/\log N_0)/t)-1$, with $N_0=1$ and $t=14$. At the final time of the experiment, each MA line was transferred to 30 mL L1 medium and left to grow for 2 weeks in order to obtain enough cells for DNA extraction, yielding material for quality control, Illumina sequencing and PCR. DNA was extracted from 15 MA lines plus the ancestral line T_0 using a modified protocol from (38) and (39): cell lysis was performed in two steps, first with Triton X-100 2% and in a second step with CTAB 4% and proteinase K 0.1 mg/mL.

De novo mutation identification

Sequencing and library preparation were done at the DOE Joint Genome Institute (California, USA) using Illumina NovaSeq S4 to generate 150 pb paired-end reads that were mapped on the reference genome (14) with BWA mem2 v2.2.1 (40), after quality control with FastQC v0.12.1 (41). BAM files were sorted with samtools v1.20 (42). The callable genome (G^*) was

determined through 3 steps: i) only sites with a minimal depth of 10 for all MA lines and 20 for the T_0 (samtools depth with -Q 20 -q 20 options); ii) only sites with less than 2x the average depth were kept to exclude erroneous mapping in highly repetitive regions; iii) only the positions that met these criteria in all MA lines were kept in $G*$. $G*$ analysis was done separately for the nucleomorph and the organelles, taking into account their relative ploidies (43), as well as scaffold 43 which appears to be duplicated in this strain (Fig. S10) and the hypermutable regions. SNM and short ID calling was done with GATK Haplotypecaller v4.2.6.1 (44). VCF files containing all variants were then manually sorted to extract de novo mutations considering several criteria: i) candidate mutations are called in only 1 MA line, with up to 2 similar reads in other lines as sequencing errors; ii) at most 2 reads that do not support the candidate mutation; iii) the allele corresponding to the mutation is not present in the T_0 genome; iv) the site is in $G*$. All insertions and deletions (ID) were manually checked in IGV (45). Then, primers were designed to verify all ID and a subset of SNM by PCR and Sanger sequencing (Table S13). All successfully sequenced amplicons (7 ID and 19 SNM) showed the expected mutation. The total mutation rate as well as mutation rates per type of mutation were calculated using the following formula:

$$\mu = \frac{Nmut}{G * \times MAgen}$$

Where N_{mut} is the number of mutations, $G*$ is the number of callable sites, and MA_{gen} is the sum of all independent generations across the 15 MA lines.

Additionally, we looked at the raw genomic coverage with bedtools v2-2.18.0 (46) using a window size of 1kb. This analysis provided evidence of large duplications in three MA lines (Fig. S2), as well as the duplication of scaffold 43 mentioned above.

Dinucleotide analysis

To identify sequences that might already have undergone TA→CA and TA→TG hypermutation we applied four different methods. First, we looked at the relative frequencies of these dinucleotides in genes previously identified as having transpoviron, virophage or NCLDV origins in (15). We extracted transcript sequences from the 'Bigna1_all_transcripts.fasta' on Phycocosm (47). We calculated the relative frequency of each dinucleotide, with reverse complements combined using 1bp steps.

Second, we searched for sequences that are depleted in TA or depleted in TA and enriched in CA+TG in the whole transcriptome. A supgroup depleted in TA (Fig. 2A) was then extracted with Mclust in R v.4.4.3 (48) (Fig. S4). The 324 transcripts identified were then scanned for

known Pfam domains (49) using HMMER hmmscan v3.2.1 (50) to see which domains were enriched in the TA-depleted supgroup; only hmmscan hits with E-value $<10^{-2}$ were kept and each Pfam was counted once per sequence. Protein family and taxonomy were retrieved from InterPro (51).

Third, genes of ≤300 bp in length were omitted to increase the representation of protein-coding genes, resulting in 21,037 transcripts. Relative dinucleotide frequencies were calculated as before to extract the 5% sequences lowest in TA and the 5% highest in CA+TG; 470 sequences were shared. Hmmscan was run on these sequences with a stricter E-value $<10^{-5}$. This time, each Pfam could figure multiple times per sequence for statistical analysis.

Last, to extend the analysis to the whole genome, the genome was cut in 1kb windows and for each window, we estimated relative enrichment of both TA and CA+TG dinucleotides using the Rho-value (52). This method uses nucleotide composition to predict expected dinucleotide composition within a sequence:

$$Rho_{XY} = \frac{N_{XY}}{N_X * N_Y}$$

where N_{XY} is the raw count of any dinucleotide. A Rho >1 indicates enrichment and a Rho <1 depletion of the dinucleotide. Clusters depleted in TA and enriched in CA+TG were defined as at least 3 subsequent windows with $Rho_{TA} \leq 0.43$ and $Rho_{CA+TG} \geq 1.52$. Gene annotations were retrieved from Phycocosm (47).

Single nucleotide polymorphism (SNP) analysis in the *T_0* strain

To determine if the long-term NCLDV sequence evolution reflects both exceptional spontaneous mutation rate and bias we observe, we analysed the neutral SNPs in the NCLDV region between the reference *B. natans* genome and the *T_0* culture from our experiment (14). Only intergenic positions with ≥20 depth (samtools depth, -Q 20 -q 20 options) in the *T_0* genome were considered. SNPs were called by GATK HaplotypeCaller and diversity was calculated as N_{SNP}/N_{total}. All SNPs that showed a change from the reference genome from A/T to G/C in *T_0* were divided by the total number of A or T sites analysed, giving SNP GC-rate. SNP AT-rate was calculated similarly and SNP AT-bias was calculated as AT-rate/GC-rate.

Search for candidate deaminases involved in immunity

To investigate if *B. natans* possesses candidates for A-deaminases working on DNA, we aligned all 5 adenosine deaminases predicted in the *B. natans* genome (ProteinIDs: 142701, 73984, 135286, 57641, 78255) against a previously published alignment (19) (File S1). The

following sequences were retrieved from GenBank: ADAT2 from *Mus musculus* (MmADAT2, NP_080024.3), *Danio rerio* (DrADAT2, XP_005160675.1), *Branchiostoma japonicum* (BjADAT2, UDV79371.1), *Mizuhopecten yessoensis* (MyADAT2, XP_021373128.1), *Trypanosoma brucei* (TbADAT2, RHW71099.1), as well as their prokaryotic ortholog TadA from *Escherichia coli* (EcTadA, VWQ04824.1) and AID from *M. musculus* (MmAID, AAD41793.1). Protein sequences were aligned with the ClustalW algorithm (53) in MEGA11 (54), default parameters. Alignments were visualized with ESPript 3.0 (55). AlphaFold predictions of Bn78255 and BjADAT2 were done on AlphaFoldServer.com (56), using default parameters. Pymol v3.1.0 (57) was used for structural alignment of 'model 0' of each protein, default parameters. To find a BjADAT3 homologue, its protein sequence was retrieved from GenBank (UDV79370.1) and was blasted against the *B. natans* genome with tblastn (58), minimum E-value $<10^{-2}$, default parameters.

ACKNOWLEDGEMENTS

We acknowledge the GenoToul Bioinformatics platform (Toulouse, France) for bioinformatics analysis support and cluster availability and all members and former members of the GENOPHY team, in particular Frédéric Sanchez and Anaïs Labécot for experimental procedures. We also thank the BIOPIC, BSBII and Bio2Mar platforms from the Observatory of Banyuls sur mer. We thank Julien Henri for his help with the predicitions and alignment of three-dimensional protein structures. The work (proposal: 10.46936/10.25585/60008631) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. This work was funded by the ANR ANR-23-CE20-0013 and ANR-21-CE02-0026.

AUTHOR CONTRIBUTIONS

LM performed the bioinformatic analyses and drafted the first version of the manuscript. AL performed the initial bioinformatics analyses on raw reads. KB and IG managed sequencing. GP and MK managed the project, MK performed the mutation accumulation experiment. All authors contributed to the final version of the manuscript.

COMPETING INTERESTS

No conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

All raw reads are available in the SRA database of the NCBI (Table S14).

SUPPLEMENTARY MATERIAL

Table S1. Details of the callable genomes.

Table S2. List of mutations.

Table S3. Details of the MA lines.

Table S4. Significance of the difference of μ_{SNM} between compartments.

Table S5. Mutation biases.

Table S6. Gene IDs of 324 genes with low relative TA dinucleotide frequency.

Table S7. Significant Pfam hits of the genes with low relative TA dinucleotide frequency.

Table S8. Transcript IDs of 470 genes within both the 5% with lowest relative TA and the 5% with highest relative CA+TG.

Table S9. Pfam hits of the 470 genes within both the 5% with lowest relative TA and the 5% with highest relative CA+TG.

Table S10. Coordinates of genomic clusters with RhoTA ≤ 0.43 and RhoCA+TG ≥ 1.52.

Table S11. Annotation of the genes in low TA / high CA+TG clusters.

Table S12. Genetic diversity of the different nuclear regions in the ancestral strain T_0.

Table S13. Primers used to check candidate mutations.

Table S14. SRA accessions

File S1. Adenosine deaminase sequence alignment

Fig. S1. SNM per nuclear compartment

Fig. S2. Large duplications in *B. natans* MA lines

Fig. S3. Length of viral elements in *B. natans*

Fig. S4. Transcript population depleted in TA dinucleotides

Fig. S5. Scatterplots of the pairwise relationships between RhoTA and other dinucleotide Rho

Fig. S6. Distribution of RhoTA and RhoCA+TG in 1kb genomic windows

Fig. S7. Distribution of the number of subsequent 1kb windows

Fig. S8. Distribution of RhoTA and RhoCA+TG values in 1kb windows across the *B. natans* 50 largest scaffolds

Fig. S9. Adenosine deaminase structural alignment

Fig. S10. Duplication of scaffold 43 in the *B. natans* T_0 line

REFERENCES

1. M. Lynch, *et al.*, The divergence of mutation rates and spectra across the Tree of Life. *EMBO Rep.* **24**, e57561 (2023).

2. W. M. Brown, M. George, A. C. Wilson, Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci.* **76**, 1967–1971 (1979).

3. G. Saxer, *et al.*, Whole Genome Sequencing of Mutation Accumulation Lines Reveals a Low Mutation Rate in the Social Amoeba Dictyostelium discoideum. *PLOS ONE* **7**, e46759 (2012).

4. D. R. Smith, Mutation rates in plastid genomes: they are lower than you might think. *Genome Biol. Evol.* **7**, 1227–1234 (2015).

5. G. D'Alessandro, F. d'Adda di Fagagna, Transcription and DNA damage: holding hands or crossing swords? *J. Mol. Biol.* **429**, 3215–3229 (2017).

6. V. Kusmartsev, M. Drożdż, B. Schuster-Böckler, T. Warnecke, Cytosine methylation affects the mutability of neighboring nucleotides in germline and soma. *Genetics* **214**, 809–823 (2020).

7. K. D. Makova, R. C. Hardison, The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **16**, 213–223 (2015).

8. M. Habig, C. Lorrain, A. Feurtey, J. Komluski, E. H. Stukenbrock, Epigenetic modifications affect the rate of spontaneous mutations in a pathogenic fungus. *Nat. Commun.* **12**, 5869 (2021).

9. M. M. Dillon, W. Sung, R. Sebra, M. Lynch, V. S. Cooper, Genome-Wide Biases in the Rate and Molecular Spectrum of Spontaneous Mutations in Vibrio cholerae and Vibrio fischeri. *Mol. Biol. Evol.* **34**, 93–109 (2017).

10. M. Moniruzzaman, A. R. Weinheimer, C. A. Martinez-Gutierrez, F. O. Aylward, Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 141–145 (2020).

11. C. Bellas, *et al.*, Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. *Proc. Natl. Acad. Sci.* **120**, e2300465120 (2023).

12. R. Sanjuán, M. R. Nebot, N. Chirico, L. M. Mansky, R. Belshaw, Viral mutation rates. *J. Virol.* **84**, 9733–9748 (2010).

13. D. L. Halligan, P. D. Keightley, Spontaneous Mutation Accumulation Studies in Evolutionary Genetics. *Annu. Rev. Ecol. Evol. Syst.* **40**, 151–172 (2009).

14. B. A. Curtis, *et al.*, Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492**, 59–65 (2012).

15. G. Blanc, L. Gallot-Lavallée, F. Maumus, Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. *Proc. Natl. Acad. Sci.* **112**, E5318–E5326 (2015).

16. V. Katju, U. Bergthorsson, Old trade, new tricks: insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biol. Evol.* **11**, 136–165 (2019).

17. I. Alseth, B. Dalhus, M. Bjørås, Inosine in DNA and RNA. *Curr. Opin. Genet. Dev.* **26**, 116–123 (2014).

18. R. Suspène, *et al.*, Double-Stranded RNA Adenosine Deaminase ADAR-1-Induced Hypermutated Genomes among Inactivated Seasonal Influenza and Live Attenuated Measles Virus Vaccines. *J. Virol.* **85**, 2458–2462 (2011).

19. Z. Gao, *et al.*, Amphioxus adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U and A-to-I deamination of DNA. *Commun. Biol.* **6**, 1–12 (2023).

20. N. M. Gaudelli, *et al.*, Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).

21. S. A. Widen, *et al.*, Virus-like transposons cross the species barrier and drive the evolution of genetic incompatibilities. *Science* **380**, eade0705 (2023).

22. A. P. Camargo, *et al.*, Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024).

23. H. A. M. Mönttinen, C. Bicep, T. A. Williams, R. P. Hirt, The genomes of nucleocytoplasmic large DNA viruses: viral evolution writ large. *Microb. Genomics* **7**, 000649 (2021).

24. K. M. J. Mayers, *et al.*, Grazing on Marine Viruses and Its Biogeochemical Implications. *mBio* **14**, e01921-21 (2023).

25. M. Berjón-Otero, A. Koslová, M. G. Fischer, The dual lifestyle of genome-integrating virophages in protists. *Ann. N. Y. Acad. Sci.* **1447**, 97–109 (2019).

26. E. C. Henry, R. H. Meints, A persistent virus infection in *Feldmannia* (Phaeophyceae). *J. Phycol.* 28:517–26 (1992).

27. M. P. Erazo-Garcia, *et al.*, Cryptic infection of a giant virus in a unicellular green alga. *Science* (2025).

28. A. van der Graaf, *et al.*, Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proc. Natl. Acad. Sci.* **112**, 6676–6681 (2015).

29. R. Pecori, S. Di Giorgio, J. Paulo Lorenzo, F. Nina Papavasiliou, Functions and consequences of AID/APOBEC-mediated DNA and RNA deamination. *Nat. Rev. Genet.* **23**, 505–518 (2022).

30. R. S. Harris, *et al.*, DNA Deamination Mediates Innate Immunity to Retroviral Infection. *Cell* **113**, 803–809 (2003).

31. A. Krishnan, L. M. Iyer, S. J. Holland, T. Boehm, L. Aravind, Diversification of AID/APOBEC-like deaminases in metazoa: multiplicity of clades and widespread roles in immunity. *Proc. Natl. Acad. Sci.* **115**, E3201–E3210 (2018).

32. L. M. Iyer, D. Zhang, I. B. Rogozin, L. Aravind, Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems. *Nucleic Acids Res.* **39**, 9473–9497 (2011).

33. A. Bernheim, J. Cury, E. Z. Poirier, The immune modules conserved across the tree of life: Towards a definition of ancestral immunity. *PLOS Biol.* **22**, e3002717 (2024).

34. P. Forterre, The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res.* **117**, 5–16 (2006).

35. A. Nasir, G. Caetano-Anollés, A phylogenomic data-driven exploration of viral origins and evolution. *Sci. Adv.* **1**, e1500527 (2015).

36. D. Vaulot, F. Gall, D. Le Marie, L. Guillou, F. Partensky, The Roscoff Culture Collection (RCC): a collection dedicated to marine picoplankton. *Nova Hedwig.* 49–70 (2004). https://doi.org/10.1127/0029-5035/2004/0079-0049.

37. M. Krasovec, *et al.*, Fitness Effects of Spontaneous Mutations in Picoeukaryotic Marine Green Algae. *G3 GenesGenomesGenetics* **6**, 2063–2071 (2016).

38. D. W. Winnepenninckx, Extraction of high molecular weight DNA from molluscs. *Trends Genet.* **9**, 407 (1993).

39. S. K. Sahu, M. Thangaraj, K. Kathiresan, DNA extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. *Int. Sch. Res. Not.* **2012**, e205049 (2012).

40. Md. Vasimuddin, S. Misra, H. Li, S. Aluru, Efficient architecture-aware acceleration of BWA-MEM for multicore systems in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, (2019), pp. 314–324.

41. S. Andrews, s-andrews/FastQC. (2025). Deposited 6 January 2025.

42. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

43. Y. Hirakawa, K.-I. Ishida, Polyploidy of Endosymbiotically Derived Genomes in Complex Algae. *Genome Biol. Evol.* **6**, 974–980 (2014).

44. A. McKenna, *et al.*, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

45. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

46. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

47. I. V. Grigoriev, *et al.*, PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Res.* **49**, D1004–D1011 (2021).

48. L. Scrucca, C. Fraley, T. B. Murphy, A. E. Raftery, *Model-Based Clustering, Classification, and Density Estimation Using mclust in R* (Chapman and Hall/CRC, 2023).

49. J. Mistry, *et al.*, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

50. S. R. Eddy, A NEW GENERATION OF HOMOLOGY SEARCH TOOLS BASED ON PROBABILISTIC INFERENCE | Genome Informatics 2009. (2009). https://doi.org/10.1142/9781848165632_0019.

51. M. Blum, *et al.*, InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res.* **53**, D444–D456 (2025).

52. S. Karlin, L. R. Cardon, Computational DNA sequence analysis. *Annu. Rev. Microbiol.* **48**, 619–654 (1994).

53. J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).

54. K. Tamura, G. Stecher, S. Kumar, MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* **38**, 3022–3027 (2021).

55. X. Robert, P. Gouet, Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).

56. J. Abramson, *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

57. Schrödinger, LLC, PyMOL The PyMOL Molecular Graphics System, Version 3.1, Schrödinger, LLC. (2015).

58. C. Camacho, *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).