# Low Mutation Rate and Atypical Mutation Spectrum in *Prasinoderma coloniale*: Insights From an Early Diverging Green Lineage

Lisa Mettrop [1], Anna Lipzen [2], Celine Vandecasteele [3], Camille Eché [3], Anaïs Labécot [1], Kerrie Barry [2], Igor V. Grigoriev [2,4], Gwenaël Piganeau [1], Marc Krasovec [1,*]

[1]Sorbonne Université, CNRS, Laboratoire de Biodiversité et Biotechnologies Microbiennes, LBBM, F-66650 Banyuls-sur-Mer, France

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

[3]INRAE, US 1426, GeT-PlaGe, Genotoul, France Génomique, Université Fédérale de Toulouse, Castanet-Tolosan, France

[4]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 94598, USA

*Corresponding author: E-mail: marc.krasovec@obs-banyuls.fr.

**Accepted:** February 13, 2025

## Abstract

Mutations are the ultimate source of genetic diversity on which natural selection and genetic drift act, playing a crucial role in evolution and long-term adaptation. At the molecular level, the spontaneous mutation rate ($\mu$), defined as the number of mutations per base per generation, thus determines the adaptive potential of a species. Through a mutation accumulation experiment, we estimate the mutation rate and spectrum in *Prasinoderma coloniale*, a phytoplankton species from an early-branching lineage within the Archaeplastida, characterized by an unusually high genomic guanine-cytosine (GC) content (69.8%). We find that *P. coloniale* has a very low total mutation rate of $\mu = 2.00 \times 10^{-10}$. The insertion–deletion mutation rate is almost 5 times lesser than the single nucleotide mutation rate with $\mu_{ID} = 3.40 \times 10^{-11}$ and $\mu_{SNM} = 1.62 \times 10^{-10}$. *Prasinoderma coloniale* also exhibits an atypical mutational spectrum: While essentially all other eukaryotes show a bias toward GC to AT mutations, no evidence of this AT-bias is observed in *P. coloniale*. Since cytosine methylation is known to be mutagenic, we hypothesized that this may result from an absence of C-methylation. Surprisingly, we found high levels of C-methylation (14% in 5mC, 25% in 5mCG contexts). Methylated cytosines did not show increased mutation rates compared with unmethylated ones, not supporting the prevailing notion that C-methylation universally leads to higher mutation rates. Overall, *P. coloniale* combines a GC-rich genome with a low mutation rate and original mutation spectrum, suggesting the almost universal AT-bias may not have been present in the ancestor of the green lineage.

**Key words:** spontaneous mutation, mutation rate, *Prasinoderma coloniale*, mutation spectrum, green algae, DNA methylation.

## Significance

Mutation is a crucial process in evolution and the mutation rate and spectrum shape the adaptive potential of a species. Cytosine methylation (5mC) has been reported to influence these by increasing C to T mutation rates. However, most studies in eukaryotes have focused on yeast or multicellular species, and little is known in green algae, particularly on species from early-branching lineages.

Here, we present the mutation rate and spectrum of *Prasinoderma coloniale*, an early-branching species in the green lineage. *Prasinoderma coloniale* exhibits a low mutation rate and an atypical mutation spectrum, where T and A nucleotides mutate at similar rates to G and C nucleotides. Interestingly, *P. coloniale* shows no evidence of 5mC hypermutation, potentially explaining its low mutation rate and distinctive mutation spectrum.

These findings suggest that mutation mechanisms may vary more across eukaryotes than previously thought, underscoring the importance of studying nonmodel species.

## Introduction

Mutations are the ultimate source of genetic diversity, restoring standing genetic variation, and allowing adaptation. The spontaneous mutation rate ($\mu$), defined as the number of mutations per base per generation, thus determines the adaptive potential of a species. In addition to this fundamental significance, $\mu$ also serves as a key quantifiable parameter in various fields, including population genetics, phylogenetics, genome evolution studies, genetic disease assessment, and the modeling of evolutionary processes. To date, $\mu$ has been estimated in close to 200 species (Lynch *et al*. 2023; Wang and Obbard 2023), with most studies focusing either on bacteria or, in eukaryotes, animals and the green lineage. The interspecific variation in $\mu$ spans almost 4 orders of magnitude, ranging from about $6.2 \times 10^{-8}$ mutations per site per generation in the tree *Eucalyptus melliodora* (Orr *et al*. 2020) to $7.6 \times 10^{-12}$ in the ciliate *Tetrahymena thermophila* (Long, Winter, *et al*. 2016).

$\mu$ is a quantitative trait subject to selection and genetic drift; the most widely accepted model for evolution of $\mu$ is the drift–barrier hypothesis, formulated by Lynch (2010) and Lynch *et al*. (2016). In this model, selection acts to reduce $\mu$ since most mutations are expected to have deleterious effects on fitness (Barrick and Lenski 2013). However, as $\mu$ decreases, the additional fitness benefits of mutations that lower $\mu$ will also get smaller, with the consequence that $\mu$ will hit a lower limit set by the power of genetic drift. The relative forces of selection and genetic drift are decided by the effective population size, $N_e$, often considerably smaller than absolute population size. Following the drift–barrier hypothesis, species with larger $N_e$ are expected to have lower $\mu$ and this pattern is globally observed (Lynch *et al*. 2023), accounting for part of the variation in $\mu$ across the tree of life. Furthermore, it has become increasingly evident that mutation rates exhibit substantial intraspecific variation and can change in response to environmental conditions (Jiang *et al*. 2014; Shewaramani *et al*. 2017; Liu and Zhang 2019; Belfield et al. 2021; Hasan et al. 2022; Wu et al. 2023). For example, in the midge *Chironomus riparius*, $\mu$ varies almost 5-fold across a temperature range of 9 °C (Waldvogel and Pfenninger 2021).

A second related trait is the mutational spectrum, which can be defined as the distribution of frequencies of different types of mutations. It has been shown to vary between species and environment (Katju and Bergthorsson 2019; Liu and Zhang 2019). In the case of single nucleotide mutations (SNMs), species typically present an AT-biased mutational spectrum, meaning that guanine (G) and cytosine (C) nucleotides exhibit a higher propensity to mutate into adenine (A) and thymine (T) nucleotides than vice versa. This AT-bias is observed in most prokaryotes and nearly all eukaryotes (Katju and Bergthorsson 2019), with the haptophyte *Gephyrocapsa huxleyi* being a notable exception (Krasovec et al. 2020). The higher mutability of G and C nucleotides is commonly attributed to oxidative damage, leading to the spontaneous deamination of methylated cytosines (Duncan and Miller 1980), which causes C to T transitions, and the oxidation of guanine to 8-oxoguanine, which causes G to T transitions (Kasai et al. 1993). By knowing the relative mutation rate of guanine-cytosine (GC) and AT, the equilibrium GC-content can be estimated; any deviation from the equilibrium may elevate $\mu$ (Krasovec et al. 2017).

The green lineage is one of the major groups of photosynthetic eukaryotes, encompassing both green algae (Chlorophyta) and land plants (Streptophyta). To date, mutation rates have been estimated for 20 species in the green lineage, 13 vascular plants, and 7 green algae (Wang and Obbard 2023). Generally, green algae present a $\mu$ one or two orders of magnitude lower than that of multicellular plants. However, 2 species of duckweed have extremely low mutation rates within the Streptophyta, ranging between $8 \times 10^{-11}$ and $9 \times 10^{-11}$ (Xu et al. 2019; Sandler et al. 2020), which may be attributed to their vegetative mode of reproduction.

Despite all the work on mutation rate in the green lineage, there is a lack of data on earlier branching species. In this study, we investigate the mutation rate and

spectrum of *Prasinoderma coloniale*, a marine phytoplankton picoalga first described by Hasegawa et al. (1996). One morphological distinctive trait of the cells from this species is a thick cell wall without any appendages with one or more layers, along with the ability to form loose colonies. *P. coloniale* is part of the Prasinodermophyta, a small group described as a third phylum in the green lineage (Li et al. 2020; Piganeau 2020) although its phylogenetic position is debated: molecular analyses place it either in Chlorophyta (Yang et al. 2023) or suggest that it diverged before the split of Chlorophyta and Streptophyta (Li et al. 2020). The genome of *P. coloniale* is haploid and gene-dense, with an estimated size 25.3 Mb across 20 chromosomes (Li et al. 2020). Remarkably, the nuclear genome consists of 69.8% GC base pairs, making it one of the phytoplanktonic species with the highest GC-content known to date.

## Results

### Mutation Rates and Spectra

The mutation accumulation (MA) experiment in *P. coloniale* lasted a total of 476 d with 34 bottlenecks, ending with 25 MA lines fit for sequencing (Table 1, supplementary fig. S1, Supplementary Material online). The callable nuclear genome $G^*$ is 21,106,141 nt, which corresponds to 82.6% of the total genome size; the per-chromosome coverage ranged from 73% for Chr3 to 92% for Chr1 (supplementary table S1, Supplementary Material online). The GC-content of the callable genome was 69.6% for the total nuclear genome, and 68.1% and 70.7% for coding sequences (CDS) and noncoding regions, respectively (supplementary table S2, Supplementary Material online). The chloroplast and mitochondrial genomes were poorly covered, preventing the estimation of their mutation rates. On average, the mitochondrial genome had a coverage of less than 1× in the $T_0$ genome and the chloroplast genome was very poorly covered in 12 of the MA lines, limiting the $G^*_{chloroplast}$ to 42 sites.

We identified a total of 53 de novo mutations (supplementary table S3, Supplementary Material online) over 12,544 generations in all 25 MA lines combined (211 to 641 generations per line, Table 1), yielding a total mutation rate of $\mu$=$2.00 \times 10^{-10}$ mutations per nucleotide per generation (CI 95% $1.50 \times 10^{-10}$–$2.62 \times 10^{-10}$). Of the 53 mutations, 43 were SNMs ($\mu_{SNM} = 1.62 \times 10^{-10}$, CI 95% $1.18 \times 10^{-10}$–$2.19 \times 10^{-10}$), 9 were small indels (2 insertions and 7 deletions, $\mu_{ID} = 3.40 \times 10^{-11}$, CI 95% $1.55 \times 10^{-11}$–$6.45 \times 10^{-11}$), and 1 was a large duplication (details below). Focusing on the distribution of SNMs, 11 of them occurred in CDS, yielding a $\mu_{SNM}$ of $9.46 \times 10^{-11}$. This rate is significantly lower than the $\mu_{SNM}$ in intergenic sequences, where 25 SNMs resulted in a $\mu_{SNM}$ of $2.07 \times 10^{-10}$ (Fisher's exact test, *P*-value 0.03, see supplementary tables S2 and S4, Supplementary Material online for details on $\mu$ per compartment). Introns, with 6 SNMs, have the highest mutation rate of any compartment with $\mu_{SNM} = 3.89 \times 10^{-10}$; this is significantly higher than in CDS (Fisher's exact test, *P*-value 0.01). Of the 11 SNMs in CDS, 6 were nonsynonymous and 5 were synonymous. To test for evidence of selection against nonsynonymous mutations, we compared the synonymous and nonsynonymous mutation rates, assuming that one-third of the positions are synonymous. There was no evidence for a different mutation rate between nonsynonymous and synonymous sites (Binomial test, $N_{syn\_mutations}/N_{syn\_sites}$:$N_{non\text{-}syn\_mutations}/N_{non\text{-}syn\_sites}$, *P*-value 0.523) and thus no evidence for selection, as expected given the low $N_e$ maintained during the experiment.

Of the SNMs, 31 were transitions and 12 transversions, leading to a Ts/Tv ratio of 2.6. Considering the AT-bias, defined as the ratio of mutation rates $\mu_{GC>AT}/\mu_{AT>GC}$, *P. coloniale* has a bias of 0.87, suggesting that AT bases may have a higher mutation rate than GC bases (Fig. 1). However, the GC > AT mutation rate was not significantly higher than the AT > GC mutation rate (one-sided binomial test, 26 GC > AT mutations on 39 total, *P*-value 0.38, 70% probability). Using $\mu_{GC>AT}$ and $\mu_{AT>GC}$, the equilibrium GC-content is $GC_{eq} = 53.3\%$ (see "Materials and Methods"). It is interesting to note that even if there is no difference in mutation rate between CG and AT nucleotides, the absolute numbers of GC bases that mutate to AT bases and the reverse are not in equilibrium. Here, there are 26 SNMs GC to AT and 13 in the reverse direction, leading to an absolute enrichment in AT of the genome by the mutational process alone (Binomial test, *P*-value 0.03355, 50% probability). *Prasinoderma coloniale* is one of the very rare eukaryotes with an AT-bias that does not exceed 1 and has a GC-content of almost 70%. The correlation between the genomic GC-content and the mutation spectrum bias can be tested using data from 120 species available in Lynch et al. (2023) (Supplementary Material online), showing a significant negative correlation (Pearson correlation, $t = -5.1536$, df = 134, *P*-value = $8.9 \times 10^{-7}$).

There were 2 insertions and 7 deletions (supplementary table S3, Supplementary Material online), leading to a net loss of 27 nt over the course of the experiment and an insertion:deletion ratio of 0.29. One deletion occurred in an exon, causing a 17 nt deletion and an open reading frame (ORF) frameshift. Regarding the mutation spectrum of indels, the seven deletions in this study led to a loss of 14 GC and 16 AT, while the two insertions added 3 AT, resulting in a net loss of 14 GC and 13 AT bases. There were significantly more AT-deletions compared to GC-deletions than expected from the 70% GC composition of the genome (one-sided binomial test, 14 on 30, *P*-value 0.00637, 70% probability), causing genomic GC% to increase in this case. Although it is premature to draw conclusions with our limited data, the AT- or GC-bias of indels may deserve further consideration in MA studies.

**Table 1** Summary of the MA experiment, the genealogy is provided in supplementary fig. S1, Supplementary material online

| MA line names | MA line barcodes | Illumina reads | Generations[a] | Nucleotide mutations | InDel mutations | Total mutation rate | Average sequencing depth |
|---|---|---|---|---|---|---|---|
| $T_0$ | GACTTAGG | 17,537,038 | … | … | … | … | 82 |
| Pc_2 | GTGCCATA | 21,137,382 | 641 | 3 | 1 | $2.96 \times 10^{-10}$ | 116 |
| Pc_3 | CGTACGAA | 15,859,858 | 242 | 0 | 0 | 0 | 67 |
| Pc_4 | ACATTGCG | 20,736,364 | 621 | 1 | 0 | $7.63 \times 10^{-11}$ | 114 |
| Pc_5 | TTGACAGG | 19,283,746 | 265 | 1 | 0 | $1.79 \times 10^{-10}$ | 72 |
| Pc_6 | GGCGTTAT | 21,668,758 | 464 | 0 | 0 | 0 | 112 |
| Pc_9 | ACGACAGA | 21,048,434 | 432 | 2 | 1 | $3.29 \times 10^{-10}$ | 112 |
| Pc_10 | TTAGGTCG | 22,973,022 | 636 | 4 | 0 | $2.98 \times 10^{-10}$ | 118 |
| Pc_11 | CAAGTGCA | 24,994,742 | 436 | 2 | 1 | $3.26 \times 10^{-10}$ | 119 |
| Pc_12 | AATACGCG | 20,893,738 | 615 | 2 | 0 | $1.54 \times 10^{-10}$ | 115 |
| Pc_14 | GTCTGATC | 22,872,580 | 609 | 1 | 2 | $2.33 \times 10^{-10}$ | 126 |
| Pc_15 | GATAGGCT | 23,693,592 | 211 | 1 | 0 | $2.25 \times 10^{-10}$ | 118 |
| Pc_16 | CGACGTTA | 25,929,424 | 631 | 2 | 1 | $2.25 \times 10^{-10}$ | 141 |
| Pc_17 | GATTACCG | 19,961,310 | 523 | 2 | 0 | $1.81 \times 10^{-10}$ | 110 |
| Pc_18 | ATGGAAGG | 20,434,548 | 628 | 4 | 0 | $3.02 \times 10^{-10}$ | 112 |
| Pc_19 | TCAAGGAC | 22,683,670 | 624 | 3 | 0 | $2.28 \times 10^{-10}$ | 117 |
| Pc_20 | CTCTACTC | 20,395,784 | 406 | 1 | 0 | $1.17 \times 10^{-10}$ | 100 |
| Pc_21 | CGTGATCA | 13,594,758 | 410 | 2 | 0 | $2.31 \times 10^{-10}$ | 70 |
| Pc_23 | CCTTGATC | 20,631,662 | 465 | 3 | 1 | $4.07 \times 10^{-10}$ | 107 |
| Pc_24 | TACATCGG | 15,402,068 | 562 | 0 | 0 | 0 | 79 |
| Pc_25 | ACTCTCGA | 17,430,290 | 588 | 4 | 0 | $3.22 \times 10^{-10}$ | 89 |
| Pc_29 | GGACCTAT | 26,969,326 | 621 | 6 | 0 | $4.58 \times 10^{-10}$ | 130 |
| Pc_31 | TTCTCTCG | 18,376,534 | 261 | 0 | 0 | 0 | 77 |
| Pc_32 | CGATGCTT | 31,136,126 | 620 | 4 | 1 | $3.82 \times 10^{-10}$ | 170 |
| Pc_33 | ATACGACC | 20,760,482 | 539 | 1 | 1 | $1.76 \times 10^{-10}$ | 106 |
| Pc_35 | GCTTCTTG | 17,515,838 | 496 | 0 | 1 | $9.55 \times 10^{-11}$ | 92 |

[a]The difference in number of generations is due to the fact that some lines were lost before the end of the experiment (supplementary fig. S1, Supplementary Material online).

To test the effect of transcription on the mutation rate, we used the gene expression data from the MMETSP (Keeling et al. 2014) for the same *P. coloniale* strain as used in the MA experiment as a proxy of the gene expression in our lines. Mutated sites are significantly less transcribed than nonmutated site, with an average RNA-seq depth of 36.2 and 99.2 for mutated sites and unmutated sites, respectively (Wilcoxon rank-sum test, *P*-value $8.9 \times 10^{-5}$).

Finally, using raw read coverage as a proxy for sequence copy number (Fig. 2), we found that the *P. coloniale* $T_0$ clonal population has a duplicated Chr18. In addition, there were 2 candidate de novo duplications, one of ±630 kb on Chr3 in MA line 12 and one of ±36 kb on Chr2 in MA line 23 (supplementary fig. S2, Supplementary Material online); the latter duplication was not confirmed by Nanopore sequencing and was therefore discarded. No genes associated with DNA replication or repair, which could have contributed to a reduction in the mutation rate, were identified in the duplicated regions (supplementary table S5, Supplementary Material online).

## Fitness of MA Lines

The correlation between mutation rate and fitness was evaluated using growth rate (GR) as a proxy for fitness. Examining the average GR of the whole experiment, there was no significant variation in fitness (Spearman's correlation, *P*-value 0.07, Rho 0.31, supplementary fig. S3, Supplementary Material online). Some individual lines showed variation over time, with an intriguing slight improvement in GR during the course of the experiment (lines 2, 4, 9, 10, 17, 24, and 33). However, there was no clear relationship between the GR and the number of mutations (see supplementary table S6, Supplementary Material online for *P*-value, Rho, and mutations). None of these lines had indels or nonsynonymous nucleotide mutations in CDS or structural mutations. These results are hard to interpret due to the lack of a control population with larger population size that would have been used to normalize GR.

## DNA Methylation in *Prasinoderma coloniale*

To investigate whether methylated cytosines are hypermutable in this species, DNA methylation in *P. coloniale* was estimated in the $T_0$ and MA lines 12 and 23. Four different contexts were analyzed: 5mC, 5mCG, 4mC, and 6mA. A full overview of average methylation levels per context can be found in supplementary tables S7 and S8, Supplementary Material online. Looking at 5mC and

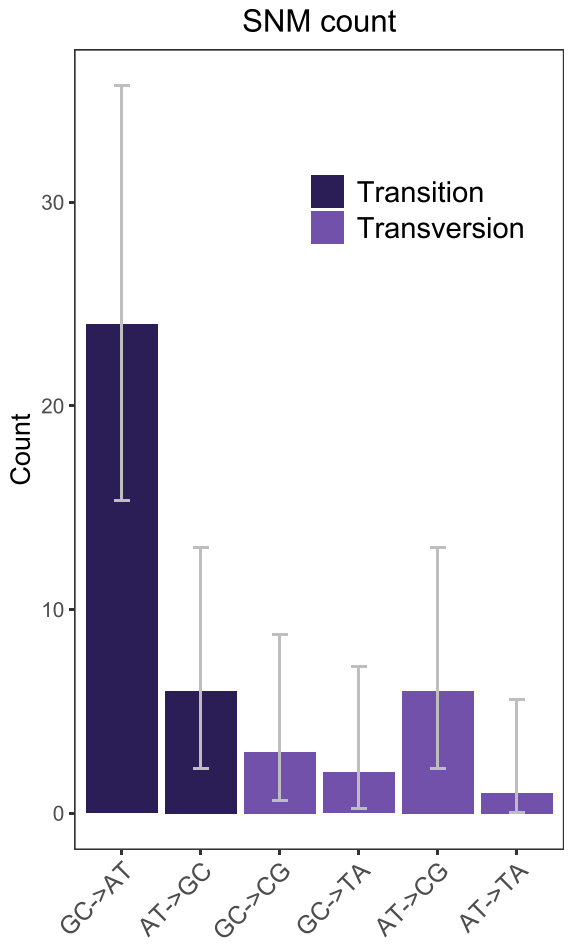


**Fig. 1.** SNM mutation spectrum of *Prasinoderma coloniale* in the nuclear genome. Error bars represent the 95% CI for a Poisson distribution.

5mCG on shared sites that are analyzable in all three samples (10,047,070 and 4,171,117 sites, respectively; Fig. 3), 3 different categories of sites can be defined: Constitutively methylated (9.5% of C, 19.7% of CG, which are methylated in all 3 samples), optionally methylated (4.4% of C, 5.7% of CG, methylated in 1 or 2 samples) and not methylated (86.1% of C, 74.7% of CG, methylated in 0 samples). Looking at 5mC, 13 SNMs fell in the analyzable portion, but none were methylated—this means that mutation rates of C and 5mC are not significantly different (Fisher's exact test, *P*-value = 0.2376). Then, looking at 5mCG, 3 SNMs fell on shared positions and none were methylated—thus, there is no difference in mutability between CG and 5mCG either (Fisher's exact test, *P*-value = 0.5762). Methylation levels in the 3 samples were very similar (supplementary table S8, Supplementary Material online), although statistically different (Pearson's $\chi^2$ test, *P*-value $2.2 \times 10^{-16}$ for all contexts). To assess whether reducing the dataset to shared sites affected the results, the mutability of cytosines was also analyzed in the methylation landscape inferred from each of the three lines individually. There was no evidence of a difference in mutation rates between methylated and nonmethylated sites, either at C or CG positions (supplementary table S8, Supplementary Material online). In short, there is no evidence that C-methylation increases cytosine mutation rate in *P. coloniale*.

Analysis of genome CG architecture revealed 59 clusters (average length: 1,266 bp) enriched in CpG stretches (average number of stretches per cluster: 8, average length of stretches: 35 bp) distributed across 19 chromosomes. Within these clusters, CG sites were significantly more methylated with an average methylation of 61.2%, compared with 25.2% in noncluster CG sites (Wilcoxon rank-sum test, *P*-value $< 2.2 \times 10^{-16}$). No SNMs fell on GC sites within clusters.

## Discussion

### *Prasinoderma coloniale* has a Low Spontaneous Mutation Rate

The mutation rate in *P. coloniale* ($\mu_{total} = 2.00 \times 10^{-10}$, CI 95% $1.50–2.62 \times 10^{-10}$) is one of the lowest mutation rates observed in eukaryotic phytoplankton to date. The next lowest mutation rates are estimated in the green algae *Bathycoccus prasinos* ($\mu_{total} = 4.39 \times 10^{-10}$, CI 95% $3.00–6.20 \times 10^{-10}$) and *Ostreococcus tauri* ($\mu_{total} = 4.79 \times 10^{-10}$, CI 95% $3.91–5.80 \times 10^{-10}$). Mutation rates of similar magnitude or even lower have been reported for *Chlamydomonas reinhardtii* (Ness et al. 2012; Sung et al. 2012), although these were based on very small sample sizes. A subsequent, larger study, however, reported $\mu_{SNM}$ almost 6 times higher than the one in *P. coloniale* (Ness et al. 2015). The $\mu_{SNM}$ of *P. coloniale* ($\mu_{SNM} = 1.62 \times 10^{-10}$, CI 95% $1.18 \times 10^{-10}–2.19 \times 10^{-10}$) is even slightly below that of unicellular model species such as *Saccharomyces cerevisiae* ($\mu_{SNM} = 1.7 \times 10^{-10}–7.2 \times 10^{-10}$) (Lynch et al. 2008; Zhu et al. 2014) and the bacteria *Escherichia coli* ($\mu_{SNM} = 2.0 \times 10^{-10}–2.5 \times 10^{-10}$; Foster et al. 2015; Lee et al. 2012; Long, Miller, et al. 2016), although it remains higher than the extreme values found in some protists. In ciliate species, many asexual reproductive cycles occur during which the germline remains transcriptionally inactive (Long, Winter, et al. 2016), likely contributing to the extremely low mutation rate of $7.6 \times 10^{-12}$ in *T. thermophila*, for example. Low $\mu_{SNM}$ are also observed in the amoeba *Dictyostelium discoideum* (Kucukyildirim et al. 2020), which in contrast has an elevated $\mu_{ID}$, and in the opisthokont *Sphaeroforma arctica* (Long et al. 2018). Here, we explore two factors that could explain the low $\mu_{SNM}$ in *P. coloniale*. The first and most

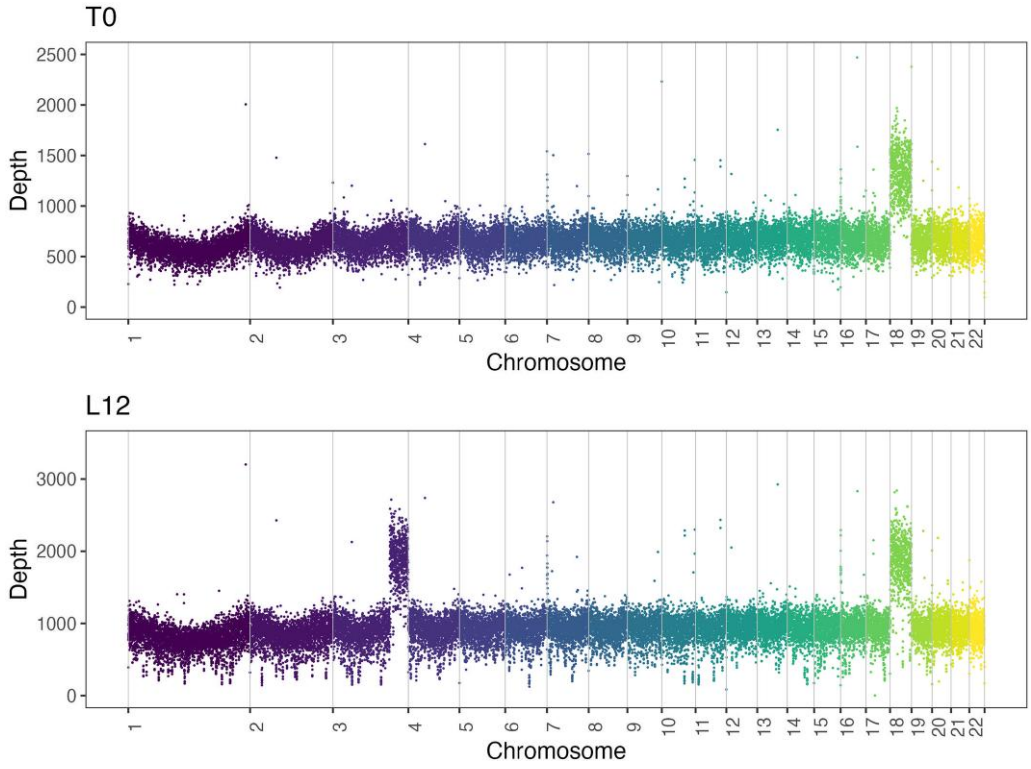


**Fig. 2.** Illumina sequencing depth per 1,000 bp window of the nuclear genome in the ancestral line ($T_0$) and MA line 12, which presents a duplicated region of ~600,000 bp on chromosome 3, starting at around position 1,625,000 and extending to the end of the chromosome. This duplication was confirmed by both Nanopore and Illumina sequencing. Chromosome 18 is duplicated in the ancestral line as well as in all MA lines. See supplementary table S5, Supplementary Material online for the complete list of duplicated genes.

widely accepted is a large effective population size ($N_e$), which is generally a characteristic of unicellular species. A large $N_e$ can significantly lower the mutation rate, as postulated by the drift–barrier hypothesis. Other species of marine phytoplankton are known to have large $N_e$ values, such as $2.7 \times 10^6$ in *G. huxleyi* (Krasovec et al. 2020), $8.9 \times 10^6$ in *Phaeodactylum tricornatum* (Krasovec et al. 2019) and $1.2 \times 10^7$ in *O. tauri* (Blanc-Mathieu et al. 2017). Similarly, the protist *T. thermophila* has an even larger $N_e$ of $1.1 \times 10^8$ (Katz et al. 2006). While an accurate estimate of $N_e$ for *P. coloniale* is currently unavailable, a very large $N_e$ likely explains the low mutation rate.

Second, transcription has been shown to have an effect on mutation rate; however, whether it increases or decreases seems to depend on the species (Park et al. 2012; D'Alessandro and d'Adda di Fagagna 2017; Sebastian and Oberdoerffer 2017; Foster et al. 2021). In *P. coloniale*, mutated sites have a lower transcription rate than nonmutated sites, at least in the available MMETSP dataset, which could change from the lines in our experiment. A similar negative correlation between transcription and mutation rate has been observed in *B. prasinos* and *O. tauri* (Krasovec et al. 2017) and is a mark of transcription coupled repair (TCR; Selby et al. 2023). This protective effect of transcription combined with the gene-dense genome of *P. coloniale* might be another factor contributing to lower $\mu_{SNM}$. Interestingly, introns exhibit a higher mutation rate than other compartments in the genome, suggesting that there are additional forces that may lower mutation rates in exons beyond TCR. One likely explanation is differential mismatch repair between exons and introns, which has been shown to reduce the mutational load in exons compared to introns in somatic cells (Frigola et al. 2017), although this has not yet been confirmed in germline cells (Rodriguez-Galindo et al. 2020).

A third explanation may be the lack of methylated C hypermutability, which is detailed below.

## *Prasinoderma coloniale* has an Atypical Mutation Spectrum and High GC-Content

As summarized in Long et al. (2018), genome composition can mainly be explained by 3 nonmutually exclusive mechanisms, which we will discuss briefly for *P. coloniale*: the mutation spectrum, direct selection on GC-content and GC-biased gene conversion (gBGC).

First, the ratio of $\mu_{GC>AT}/\mu_{AT>GC}$ is 0.87 in *P. coloniale*. In most studied species (Hershberg and Petrov 2010;

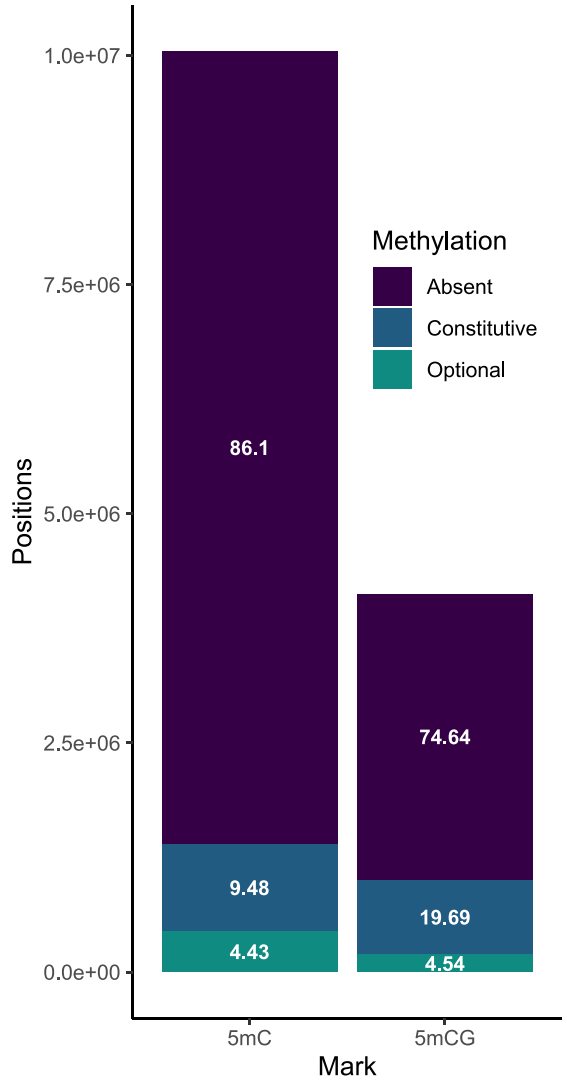


**Fig. 3.** Cytosine methylation in *Prasinoderma coloniale* $T_0$ and MA lines 12 and 23 were sequenced with an ONT PromethION device to analyze base modifications. In *P. coloniale*, global 5mC and specifically 5mCG are the predominant methylation marks. For both marks, a distinction can be made between constitutively methylated sites (methylated at this position in all the 3 samples) and optionally methylated sites (present in only 1 or 2 samples). The *y* axis shows the total count of genomic positions for each group of methylated sites, while relative percentages are displayed in white on the bar plots. Neither of these 2 marks showed a significant relationship with the mutation rate.

Hildebrand et al. 2010; Katju and Bergthorsson 2019), the $\mu_{GC>AT}/\mu_{AT>GC}$ ratio is higher than 1, which is called the AT-mutation bias. Within eukaryotes, there are very few exceptions to this trend (one species in Long, Behringer, et al. [2016]; one in Liu and Zhang [2019]; 3 in Bergeron et al. (2023); see Lynch et al. 2023 for an overview) and the only species to present a strong inverse bias is *G. huxleyi* (Krasovec et al. 2020) with a notable ratio of 0.38. Like *P. coloniale*, *G. huxleyi* has a GC-rich genome (65.5%) that may be explained in part by its atypical mutation bias. In prokaryotes, a few other examples exist and all are in species with a high GC-content (Lynch et al. 2023, Supplementary tables). Indeed, the GC-biased mutation spectrum in *P. coloniale* could very well explain, in part, the GC-rich genome.

An explanation for this absence of AT mutation bias in *P. coloniale* could be the lack of or limited methylation of cytosines, because methylated Cs have been shown to have a higher mutation rate than unmethylated Cs as a consequence of spontaneous deamination (Coulondre et al. 1978; Duncan and Miller 1980; Shen et al. 1994); in *Arabidopsis thaliana*, mutation rate of methylated Cs is estimated to be 1.3 to 1.5 times higher than in nonmethylated Cs (Kusmartsev et al. 2020; Lu et al. 2021). We thus hypothesized that C methylation is absent in *P. coloniale*. However, the genome of *P. coloniale* contains 5 possible DNA methyltransferases (PFAM domain PF00145), at least 3 of which are expressed at modest levels (>800 RNA reads mapped, within 55% to 35% of lowest transcribed genes, supplementary table S9, Supplementary Material online), providing evidence that DNA methylation is active. We found that *P. coloniale* exhibits notable levels of C methylation, particularly in the 5mCG context. In green algae, 5mCG can range from 5.4% in *C. reinhardtii* (Feng et al. 2010) to as much as 76.9% in *Scenedesmus acutus* (Ferrari et al. 2020) but typically ranges around 20% (Huff and Zilberman 2014), which is consistent with the results observed in *P. coloniale*. The pioneering studies on 5mC mutation were performed on bacteria or animals and are not necessarily representative for algal species, where C methylation may serve different causes and is often maintained by other enzymes (Huff and Zilberman 2014; Hoguin et al. 2023). In many algae, CG methylation is mediated by DNMT5 rather than DNMT1 and CG methylation does not necessarily increase cytosine mutability—rather the inverse, it seems that contrary to the GC depletion expected from hypermutable Cs, there is significant enrichment of GC dinucleotides in *O. tauri* (Huff and Zilberman 2014). We suspect similar mechanisms take place in *P. coloniale*. The genome contains a good candidate for an active DNMT5, the PRCOL_00007237-RA gene, presenting SNF2-related domains as described previously (Huff and Zilberman 2014), and we observe no correlation between C methylation and C mutability in *P. coloniale*. The absence of methylated cytosine hypermutation is a likely explanation for the distinct mutation spectrum we observe in *P. coloniale*, where no difference in mutation rate is observed between AT and GC bases. Furthermore, this absence may contribute to a lower mutation rate in this species, especially when compared with species where methylated cytosines are hypermutable.

The interplay between a long-term absence of an AT-bias and direct selection on GC in *P. coloniale* may induce

an exceptionally high codon usage bias (CUB). The strength of CUB can be estimated by a low effective number of codons (ENC), which corresponds to the number of used codons (between 20 and 61, the lower the ENC, the stronger the CUB) and high GC-content of the third silent position of codons (GC3; Wright 1990). Selection generally increases GC-content at silent sites, which stabilizes RNA structures, increases translation efficacy, and creates more RNA:RNA interaction sites (Kudla et al. 2006; Shabalina et al. 2013). The almost universal AT-biased mutation spectrum and the almost universal GC-selection on CUB are thus opposing forces in most species when it comes to GC-content of synonymous sites. However, in species with both an AT > GC mutation bias and a selection toward GC-rich codons, the resulting CUB may be very strong, as is the case of *G. huxleyi* with an ENC of 38 (Krasovec et al. 2020). This is the lowest ever measured ENC in phytoplankton; most phytoplankton species have values between 45 and 58 (Krasovec and Filatov 2022). Here, *P. coloniale* exhibits an even lower CUB with an ENC of 36. Another hallmark of CUB is the GC-content at synonymous sites compared to other sites; especially a difference between GC3 and intron GC-content is often interpreted as a signature of selection (Eyre-Walker 1999; Hershberg and Petrov 2009; Shabalina et al. 2013). Indeed, for *P. coloniale*, the GC-content of the third, silent position of codons (GC3) is 85.7%, significantly higher than GC1 + GC2 (59.6%), intron GC-content (73.5%), and total genome GC-content (69.6%) (Fisher's exact test, *P*-value $< 2.2 \times 10^{-16}$ for all). This supports a strong CUB in *P. coloniale*, shaped by both selection and a non-AT-biased mutation spectrum.

Finally, the exceptionally high GC-content of *P. coloniale* may be explained, to a lesser extent, by gBGC that is typically biased toward GC in most eukaryotes (Duret and Galtier 2009). While this process is mostly active during meiosis, it can occur during mitosis (Chen et al. 2007) and is also suspected to be present in eukaryotes with predominant asexual cycles (Jancek et al. 2008; Marsolier-Kergoat 2013) and even in bacteria (Lassalle et al. 2015). A negative correlation between chromosome length and GC-content is consistent with gBGC and meiotic recombination (Jancek et al. 2008). In *P. coloniale*, this correlation was indeed observed (Spearman's correlation, *P*-value 0.033, Rho −0.46), further supporting gBGC and meiotic recombination as mechanisms contributing to the high GC-content in this genome.

## Materials and Methods

### MA Experiment

The MA experiment has been performed on *P. coloniale* strain *CCMP1413* using a flow cytometry protocol described previously for liquid medium (Krasovec et al. 2017). The identity of the strain was confirmed by 18S rDNA PCR coupled with Sanger sequencing at the start of the experiment. Cells were cultivated in L1 medium generated from sterile natural sea water enriched with $NaNO_3$, $NaH_2PO_4$, trace minerals, and vitamins (Bigleow L1 Culture Medium Kit, according to manufacturer's instructions) with salinity of 33 g/L in a light:dark cycle of 12:12 h at 20 °C. The ancestral population used for the experiment was started with an estimated single cell isolated by dilution after measuring cell density with a Beckman Coulter Cytoflex cytometer (Krasovec et al. 2017). After 2 weeks, the cell density of the starting culture was measured again by flow cytometry to inoculate MA lines. The measured cell density allowed to sample the volume corresponding to ten cells, which were diluted in 10 mL culture medium and then used to inoculate 6 wells on a 48-well plate, each containing 1 mL, with statistically one cell per well. The 6 replicates limit the loss of MA lines due to technical errors or lethal mutations (Krasovec et al. 2016). To reduce selection between mutated and nonmutated cells, single cell bottlenecks were performed every 14 d for each MA line. At each bottleneck time, cell density of MA line populations was measured by flow cytometry and the first well with living cells encountered was used for reinoculation with statistically one cell per well, isolated by dilution as described above. All cell and media manipulations were performed under a sterilized laminar flow hood to avoid contamination and all containers were either autoclaved or brand-new. The bottlenecks were repeated 34 times within the 476 d experiment. The cell density was also used to estimate the average daily GR between each bottleneck using the following formula:

$$\mathrm{GR} = e^{\frac{\ln\left(n \text{ cells} \times \frac{\text{dilution}}{\text{flow rate in mL}}\right)}{\text{days}}}.$$

Thus, the number of generations between each bottleneck equals Gen = (GR−1)×14. The average $N_e$ per MA line between each bottleneck was calculated using the harmonic mean of cell number, starting from one cell to the final cell number, with the number of cells per day calculated with GR. The $N_e$ in each well never exceeded eight cells. The GR was also used as a proxy of MA line fitness during the experiment.

At the end of the MA experiment, MA line populations were transferred to flasks containing 30 mL of L1 medium and allowed to grow for 2 to 3 weeks in order to obtain sufficient cell numbers for DNA extraction for multiple uses (i.e. quality control, whole genome sequencing and PCR followed by Sanger sequencing). High molecular weight DNA was extracted from 25 MA lines plus the ancestral line $T_0$ using a cetyltrimethylammonium bromide (CTAB) extraction protocol modified from Winnepenninckx (1993)

and Sahu et al. (2012): Cell lysis was performed in two steps, first with Triton X-100 2% (Sigma-Aldrich, 9036 to 9019-5) and in a second step with CTAB 4% (Sigma-Aldrich, H6269) and proteinase K 0.1 mg/mL (from *Tritirachium album*, Sigma-Aldrich, SAE0151); see Winnepenninckx (1993) for other buffer ingredients. Total DNA libraries were prepared and sequenced by the Joint Genome Institute (JGI) at Berkeley Lab, CA, USA. Sequencing was done with Illumina NovaSeq 4000 in 150 paired-end reads, generating a raw read count of between 14 and 32 million reads per sample (Table 1).

## De Novo Mutation Identification

Raw Illumina reads were trimmed at the JGI, and the read quality was checked using FastQC v0.11.7 (Andrews 2025). Reads were mapped against the reference genome with bwa-mem2 v2-2.2 (Vasimuddin et al. 2019), and the resulting binary alignment mapping (BAM) files were treated using samtools v1.18 (Danecek et al. 2021). The callable genome ($G^*$) was defined through three steps: (i) samtools depth was used to keep the positions with at least 20 coverage considering both base and mapping quality thresholds of 20 (options -q 20 -q 20); (ii) positions with a coverage of >221, which corresponds to 2× the average depth, were removed to exclude erroneous mappings in highly repetitive regions; (iii) only the positions that fulfilled these criteria in all 26 samples were kept as callable sites. An exception to step (ii) was made for Chr18, which was shown to be duplicated in the $T_0$ *P. coloniale* culture (Fig. 2) and was thus analyzed separately, using exactly the same criteria but with a maximum depth threshold of 443 reads instead of 221. Variant calling was done with HaplotypeCaller from GATK v4.2.6.1 (McKenna et al. 2010) with the *-ploidy 1* option. The resulting VCF files were manually treated to extract de novo mutation candidates from all variants at callable sites, in the first place retaining variants found in only one MA line (or 2 in the case of sister lines, supplementary fig. S1, Supplementary Material online).

For better oversight, VCF data of candidate positions were compiled using samtools mpileup v.1.18. Next, 3 lists of mutation candidates were generated in order of diminishing stringency. List 1 contains 24 candidate SNMs that met the following criteria: (i) 0 alternative reads in the ancestral line $T_0$; (ii) at most one alternative read in another MA line with the alternative read differing from the candidate SNM. List 2 contains 19 candidate SNMs that satisfied the following criteria: (i) alternative reads in the ancestral line $T_0$ are tolerated but cannot be the same as the candidate SNM; (ii) more alternative reads than for list 1 are tolerated in nonmutated MA lines with up to three sequencing errors (which can be the same as the candidate SNM) in at most 5 MA lines (excluding the mutated MA line). List 3 contains 9 candidate InDels that were manually reviewed using the IGV genome browser v.2.15.1 (Thorvaldsdóttir et al. 2013).

After compiling these lists, primers were designed to check a subset of candidate mutations by Sanger resequencing. Primers were designed for all InDels and all SNMs of list 2 as well as for 8 randomly chosen candidate positions from list 1. For the SNMs, 13 of 26 chosen positions were successfully amplified, including 3 of 8 from list 1, and 10 of 18 from list 2 (see supplementary table S10, Supplementary Material online for primers). For the InDels, 8/9 positions were amplified (see supplementary table S10, Supplementary Material online for primers). Sanger sequencing showed a true positive rate of 100%. Thus, our criteria for candidate positions were strict enough, and all 43 SNMs and 9 InDels were considered true. Then, SnpEff v.5.1 (Cingolani) was used to identify mutation effects.

The total mutation rate as well as mutation rates per type of mutation were calculated using the following formula:

$$\mu = \frac{N_{\text{mut}}}{G^* \times \text{MA}_{\text{gen}}},$$

where $N_{\text{mut}}$ is the number of mutations, $G^*$ is the number of callable sites, and $\text{MA}_{\text{gen}}$ is the sum of all generations of the 25 MA lines.

For the analysis of mutation biases in the mutation spectrum, the different mutation rates between nucleotides were estimated: the $\mu_{\text{GC}>\text{AT}}$ mutation rate is the number of GC to AT mutations per generation per number of GC bases in the callable genome; and the $\mu_{\text{AT}>\text{GC}}$ mutation rate is the number of AT to GC mutations per generation per number of AT bases in the callable genome. The ratio between these specific mutation rates gives the AT-bias of the mutation spectrum, which has a net effect on genomic GC-content: $\mu_{\text{GC}>\text{AT}}/\mu_{\text{AT}>\text{GC}}$. With these rates, we can also calculate the equilibrium GC-content at which nucleotide composition is stable if we consider mutational process alone, or, at which the number of mutations from GC to AT and AT to GC are equal: $\text{GC}_{\text{eq}} = \mu_{\text{AT}>\text{GC}}/(\mu_{\text{GC}>\text{AT}} + \mu_{\text{AT}>\text{GC}})$.

Finally, we examined structural mutations using the Nanopore long-read sequencing data from 2 MA lines and the $T_0$ with 2 different strategies. First, we mapped the reads trimmed with Chopper v0.7.0 (Coster 2025) onto the reference genome with Minimap2 v2.28 (Li 2018) and called variants using Sniffles2 v2.2 (Smolka et al. 2024). Second, we did de novo assemblies of the genomes with Flye v2.9.2 (Kolmogorov et al. 2019) on the raw fastq files and used MUM&Co v3.8 (O'Donnell and Fischer 2020) to detect structural mutations. Additionally, for both long and short read sequences, we looked at the raw genomic coverage with bedtools v2-2.18.0 (Quinlan and Hall 2010), using a window size of 1 kb. This analysis provides

evidence of a duplication in one of four MA lines (Fig. 2), as well as the duplication of Chr18 mentioned earlier.

### Codon Bias Analysis

Codon bias analyze was done using the fasta file with CDS (https://genome.jgi.doe.gov/portal/Praco1/download/Praco1_GeneCatalog_CDS_20200803.fasta.gz), after verifying that all sequences were in ORF +1 (8 out of 7,139 sequences removed). The ENC on the remaining 7,131 genes was calculated with the coRdon package (v1.2.0) (Elek 2024) in R (v4.3.2) (R Core Team 2021), with a length threshold of 120 nt to remove short noncoding and fragmented RNA. GC-content on third position of codons (GC3) was calculated with the seqinr package (v4.2-36; Lobry et al. 2024).

### Methylation Analysis

To see if DNA methylation affects the mutation spectrum of *P. coloniale*, the genomes of the $T_0$ as well as MA lines L12 and L23 were long-read sequenced with Oxford Nanopore Technology. Lines L12 and L23 were chosen to confirm eventual de novo large duplications found with short read coverage (supplementary fig. S2, Supplementary Material online). Oxford nanopore technologies (ONT) libraries were prepared according to the manufacturer's instructions "genomic-dna-by-ligation-sqk-lsk114" and loaded onto R10.4.1 flow cell and sequenced on PromethION instrument. Modified base-calling and alignment of long Nanopore reads were done with Dorado v.0.7.2 (Malton et al.), using the dna_r10.4.1_e8.2_400bps_sup@v5.0.0 model with four different DNA methylation models (5mCG_5hmCG, 5mC_5hmC, 4mC_5mC, and 6mA, all v1). BAM files were sorted and indexed with samtools v1.19 and converted to bedMethyl files using Modkit v0.3.1 (Rand 2024). The default thresholds set by Modkit were quite low (supplementary table S8, Supplementary Material online) but maintained, since increasing the filter-threshold can introduce a bias, artificially increasing methylation levels as modified base probability is generally higher for 5mC than canonical bases. We analyzed the positions covered at least 5× (Ziller et al. 2015), first for each sample separately (supplementary table S8, Supplementary Material online) and then on the joined data (supplementary table S7, Supplementary Material online) to get a more confident methylation call, using only the positions covered ≥5× times in all samples. Positions were considered methylated if ≥80% of reads supported the modification. Last, CpG rich regions were identified using gCluster (Gómez-Martín et al. 2018).

### Transcription Analysis

To detect any effect of transcription rate on the mutation rate, which can be either mutagenic through transcription associated mutagenesis (Jinks-Robertson and Bhagwat 2014) or protective through TCR (Hanawalt and Spivak 2008), we used the available transcriptome of *P. coloniale* CCMP1413 (MMETSP0806/SRR1296862; Keeling et al. 2014) to make a correlation between mutation rate and expression level. Raw reads were mapped against the reference genome using Hisat2 v.2.2.1 with the –sra-acc command and default parameters (Kim et al. 2019).

## Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

## Acknowledgments

We acknowledge the GenoToul Bioinformatics platform (Toulouse, France) for bioinformatics analysis support and cluster availability, the BIOPIC platform for support with the flow cytometry and all members of the GENOPHY team.

## Author Contributions

M.K. performed the MA experiment and managed the project. A.La. assisted with experimental work. L.M. analyzed the data and drafted the manuscript. A.Li., C.V., and C.E. analyzed the data. I.V.G., K.B., and G.P. managed sequencing and data analysis. L.M., M.K., and G.P. wrote the manuscript, and all authors contributed to the final version.

## Funding

The work (proposal: to M.K.) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, was 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 Agence Nationale de la Recherche through the grant ANR-23-CE20-0013 to M.K.

## Data Availability

The raw data underlying this article are available in the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/), see supplementary table S11, Supplementary Material online for an overview of the different SRA accessions.

## Literature Cited

Andrews S. 2025. s-andrews/FastQC. https://github.com/s-andrews/FastQC [Accessed 2025 January 6].

Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nat Rev Genet. 2013:14(12):827–839. https://doi.org/10.1038/nrg3564.

Belfield EJ, Brown C, Ding ZJ, Chapman L, Luo M, Hinde E, van Es SW, Johnson S, Ning Y, Zheng SJ, et al. Thermal stress accelerates *Arabidopsis thaliana* mutation rate. Genome Res. 2021:31(1): 40–50. https://doi.org/10.1101/gr.259853.119.

Bergeron LA, Besenbacher S, Zheng J, Li P, Bertelsen MF, Quintard B, Hoffman JI, Li Z, St Leger J, Shao C, et al. Evolution of the germline mutation rate across vertebrates. Nature. 2023:615(7951): 285–291. https://doi.org/10.1038/s41586-023-05752-y.

Blanc-Mathieu R, Krasovec M, Hebrard M, Yau S, Desgranges E, Martin J, Schackwitz W, Kuo A, Salin G, Donnadieu C, et al. Population genomics of picophytoplankton unveils novel chromosome hypervariability. Sci Adv. 2017:3(7):e1700239. https://doi.org/10.1126/sciadv.1700239.

Chen J-M, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet. 2007:8(10):762–775. https://doi.org/10.1038/nrg2193.

Cingolani P. Home - SnpEff & SnpSift. https://pcingola.github.io/SnpEff/ (Accessed 2025 January 6).

Coster WD. 2025. wdecoster/chopper. https://github.com/wdecoster/chopper (Accessed 2025 January 6).

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature. 1978:274(5673):775–780. https://doi.org/10.1038/274775a0.

D'Alessandro G, d'Adda di Fagagna F. Transcription and DNA damage: holding hands or crossing swords? J Mol Biol. 2017:429(21): 3215–3229. https://doi.org/10.1016/j.jmb.2016.11.002.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021:10(2): giab008. https://doi.org/10.1093/gigascience/giab008.

Duncan BK, Miller JH. Mutagenic deamination of cytosine residues in DNA. Nature. 1980:287(5782):560–561. https://doi.org/10.1038/287560a0.

Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009:10(1):285–311. https://doi.org/10.1146/annurev-genom-082908-150001.

Elek A. 2024. BioinfoHR/coRdon. https://github.com/BioinfoHR/coRdon (Accessed 2025 January 6).

Eyre-Walker A. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics. 1999:152(2):675–683. https://doi.org/10.1093/genetics/152.2.675.

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A 2010:107(19):8689–8694. https://doi.org/10.1073/pnas.1002720107.

Ferrari M, Torelli A, Marieschi M, Cozza R. Role of DNA methylation in the chromium tolerance of *Scenedesmus acutus* (Chlorophyceae) and its impact on the sulfate pathway regulation. Plant Sci. 2020:301:110680. https://doi.org/10.1016/j.plantsci.2020.110680.

Foster PL, Lee H, Popodi E, Townes JP, Tang H. Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. Proc Natl Acad Sci U S A 2015:112(44): E5990–E5999. https://doi.org/10.1073/pnas.1512136112.

Foster PL, Niccum BA, Lee H. DNA replication-transcription conflicts do not significantly contribute to spontaneous mutations due to replication errors in *Escherichia coli*. mBio. 2021:12(5):e0250321. https://doi.org/10.1128/mBio.02503-21.

Frigola J, Sabarinathan R, Mularoni L, Muiños F, Gonzalez-Perez A, López-Bigas N. Reduced mutation rate in exons due to differential mismatch repair. Nat Genet. 2017:49(12):1684–1692. https://doi.org/10.1038/ng.3991.

Gómez-Martín C, Lebrón R, Oliver JL, Hackenberg M. Prediction of CpG islands as an intrinsic clustering property found in many eukaryotic DNA sequences and its relation to DNA methylation. In: Vavouri T, Peinado MA, editors. CpG islands: methods and protocols. New York (NY): Springer; 2018. p. 31–47.

Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. Nat Rev Mol Cell Biol. 2008:9(12): 958–970. https://doi.org/10.1038/nrm2549.

Hasan AR, Lachapelle J, El-Shawa SA, Potjewyd R, Ford SA, Ness RW. Salt stress alters the spectrum of *de novo* mutation available to selection during experimental adaptation of *Chlamydomonas reinhardtii*. Evol Int J Org Evol. 2022:76(10):2450–2463. https://doi.org/10.1111/evo.14604.

Hasegawa T, Miyashita H, Kawachi M, Ikemoto H, Kurano N, Miyachi S, Chihara M. *Prasinoderma coloniale* gen. et sp. nov., a new pelagic coccoid prasinophyte from the western Pacific Ocean. Phycologia. 1996:35(2):170–176. https://doi.org/10.2216/i0031-8884-35-2-170.1.

Hershberg R, Petrov DA. General rules for optimal codon choice. PLoS Genet. 2009:5(7):e1000556. https://doi.org/10.1371/journal.pgen.1000556.

Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 2010:6(9):e1001115. https://doi.org/10.1371/journal.pgen.1001115.

Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 2010:6(9):e1001107. https://doi.org/10.1371/journal.pgen.1001107.

Hoguin A, Yang F, Groisillier A, Bowler C, Genovesio A, Ait-Mohamed O, Vieira FRJ, Tirichine L. The model diatom *Phaeodactylum tricornutum* provides insights into the diversity and function of microeukaryotic DNA methyltransferases. Commun Biol. 2023:6(1):253. https://doi.org/10.1038/s42003-023-04629-0.

Huff JT, Zilberman D. Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. Cell. 2014:156(6): 1286–1297. https://doi.org/10.1016/j.cell.2014.01.029.

Jancek S, Gourbière S, Moreau H, Piganeau G. Clues about the genetic basis of adaptation emerge from comparing the proteomes of two *Ostreococcus* ecotypes (Chlorophyta, Prasinophyceae). Mol Biol Evol. 2008:25(11):2293–2300. https://doi.org/10.1093/molbev/msn168.

Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP. Environmentally responsive genome-wide accumulation of *de novo Arabidopsis thaliana* mutations and epimutations. Genome Res. 2014:24(11):1821–1829. https://doi.org/10.1101/gr.177659.114.

Jinks-Robertson S, Bhagwat AS. Transcription-associated mutagenesis. Annu Rev Genet. 2014:48(1):341–359. https://doi.org/10.1146/annurev-genet-120213-092015.

Kasai H, Yamamoto F, Chung MH, Ohtsuka E, Laval J, Grollman AP, Nishimura S. Formation of 8-hydroxyguanine in DNA by oxygen radicals and its repair enzyme. In: Nigam S, Honn KV, Marnett LJ, Walden TL, editors. Eicosanoids and Other Bioactive Lipids in Cancer, Inflammation and Radiation Injury: Proceedings of the 2nd International Conference; 1991 September 17–21, Berlin, FRG. Boston (MA): Springer US; 1993. p. 447–451.

Katju V, Bergthorsson U. 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. 2019:11(1):136–165. https://doi.org/10.1093/gbe/evy252.

Katz LA, Snoeyenbos-West O, Doerder FP. Patterns of protein evolution in *Tetrahymena thermophila*: implications for estimates of effective population size. Mol Biol Evol. 2006:23(3):608–614. https://doi.org/10.1093/molbev/msj067.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, et al. The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 2014:12(6): e1001889. https://doi.org/10.1371/journal.pbio.1001889.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019:37(8):907–915. https://doi.org/10.1038/s41587-019-0201-4.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019:37(5): 540–546. https://doi.org/10.1038/s41587-019-0072-8.

Krasovec M, Eyre-Walker A, Grimsley N, Salmeron C, Pecqueur D, Piganeau G, Sanchez-Ferandin S. Fitness effects of spontaneous mutations in picoeukaryotic marine green algae. G3 (Bethesda). 2016:6(7):2063–2071. https://doi.org/10.1534/g3.116.029769.

Krasovec M, Eyre-Walker A, Sanchez-Ferandin S, Piganeau G. Spontaneous mutation rate in the smallest photosynthetic eukaryotes. Mol Biol Evol. 2017:34(7):1770–1779. https://doi.org/10.1093/molbev/msx119.

Krasovec M, Filatov DA. Codon usage bias in phytoplankton. J Mar Sci Eng. 2022:10(2):168. https://doi.org/10.3390/jmse10020168.

Krasovec M, Rickaby REM, Filatov DA. Evolution of mutation rate in astronomically large phytoplankton populations. Genome Biol Evol. 2020:12(7):1051–1059. https://doi.org/10.1093/gbe/evaa131.

Krasovec M, Sanchez-Brosseau S, Piganeau G. First estimation of the spontaneous mutation rate in diatoms. Genome Biol Evol. 2019:11(7):1829–1837. https://doi.org/10.1093/gbe/evz130.

Kucukyildirim S, Behringer M, Sung W, Brock DA, Doak TG, Mergen H, Queller DC, Strassmann JE, Lynch M. Low base-substitution mutation rate but high rate of slippage mutations in the sequence repeat-rich genome of *Dictyostelium discoideum*. G3 (Bethesda). 2020:10(9):3445–3452. https://doi.org/10.1534/g3.120.401578.

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol. 2006:4(6):e180. https://doi.org/10.1371/journal.pbio.0040180.

Kusmartsev V, Drożdż M, Schuster-Böckler B, Warnecke T. Cytosine methylation affects the mutability of neighboring nucleotides in germline and soma. Genetics. 2020:214(4):809–823. https://doi.org/10.1534/genetics.120.303028.

Lassalle F, Périan S, Bataillon T, Nesme X, Duret L, Daubin V. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet. 2015:11(2):e1004941. https://doi.org/10.1371/journal.pgen.1004941.

Lee H, Popodi E, Tang H, Foster PL. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. Proc Natl Acad Sci U S A 2012:109(41): E2774–E2783. https://doi.org/10.1073/pnas.1210309109.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018:34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li L, Wang S, Wang H, Sahu SK, Marin B, Li H, Xu Y, Liang H, Li Z, Cheng S, et al. The genome of *Prasinoderma coloniale* unveils the existence of a third phylum within green plants. Nat Ecol Evol. 2020:4(9):1220–1231. https://doi.org/10.1038/s41559-020-1221-7.

Liu H, Zhang J. Yeast spontaneous mutation rate and spectrum vary with environment. Curr Biol. 2019:29(10):1584–1591.e3. https://doi.org/10.1016/j.cub.2019.03.054.

Lobry et al. 2024. lbbe-software/seqinr. https://github.com/lbbe-software/seqinr (Accessed 2025 January 6).

Long H, Behringer MG, Williams E, Te R, Lynch M. Similar mutation rates but highly diverse mutation spectra in Ascomycete and Basidiomycete yeasts. Genome Biol Evol. 2016:8(12):3815–3821. https://doi.org/10.1093/gbe/evw286.

Long H, Miller SF, Strauss C, Zhao C, Cheng L, Ye Z, Griffin K, Te R, Lee H, Chen CC, et al. Antibiotic treatment enhances the genome-wide mutation rate of target cells. Proc Natl Acad Sci U S A 2016:113(18): E2498–E2505. https://doi.org/10.1073/pnas.1601208113.

Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C, Strauss C, Stone C, et al. Evolutionary determinants of genome-wide nucleotide composition. Nat Ecol Evol. 2018:2(2):237–240. https://doi.org/10.1038/s41559-017-0425-y.

Long H, Winter DJ, Chang AY, Sung W, Wu SH, Balboa M, Azevedo RBR, Cartwright RA, Lynch M, Zufall RA. Low base-substitution mutation rate in the germline genome of the ciliate *Tetrahymena thermophila*. Genome Biol Evol. 2016:8(12):3629–3639. https://doi.org/10.1093/gbe/evw223.

Lu Z, Cui J, Wang L, Teng N, Zhang S, Lam HM, Zhu Y, Xiao S, Ke W, Lin J, et al. Genome-wide DNA mutations in *Arabidopsis* plants after multigenerational exposure to high temperatures. Genome Biol. 2021:22(1):160. https://doi.org/10.1186/s13059-021-02381-4.

Lynch M. Evolution of the mutation rate. Trends Genet. 2010:26(8): 345–352. https://doi.org/10.1016/j.tig.2010.05.003.

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet. 2016:17(11):704–714. https://doi.org/10.1038/nrg.2016.104.

Lynch M, Ali F, Lin T, Wang Y, Ni J, Long H. The divergence of mutation rates and spectra across the Tree of Life. EMBO Rep. 2023:24(10): e57561. https://doi.org/10.15252/embr.202357561.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A 2008:105(27):9272–9277. https://doi.org/10.1073/pnas.0803466105.

Malton, Daw J, Lawrence B, Seymour C. nanoporetech/dorado: Oxford Nanopore's Basecaller. https://github.com/nanoporetech/dorado/tree/release-v0.7 (Accessed September 5, 2024).

Marsolier-Kergoat M-C. Models for the evolution of GC content in asexual fungi *Candida albicans* and *C. dubliniensis*. Genome Biol Evol. 2013:5(11):2205–2216. https://doi.org/10.1093/gbe/evt170.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010:20(9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Ness RW, Morgan AD, Colegrave N, Keightley PD. Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii*. Genetics. 2012:192(4):1447–1454. https://doi.org/10.1534/genetics.112.145078.

Ness RW, Morgan AD, Vasanthakrishnan RB, Colegrave N, Keightley PD. Extensive *de novo* mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. Genome Res. 2015:25(11):1739–1749. https://doi.org/10.1101/gr.191494.115.

O'Donnell S, Fischer G. MUM&Co: accurate detection of all SV types through whole-genome alignment. Bioinformatics. 2020:36: 3242–3243. https://doi.org/10.1093/bioinformatics/btaa115.

Orr AJ, Padovan A, Kainer D, Külheim C, Bromham L, Bustos-Segura C, Foley W, Haff T, Hsieh JF, Morales-Suarez A, et al. A phylogenomic approach reveals a low somatic mutation rate in a long-lived plant. Proc R Soc B Biol Sci. 2020:287(1922):20192364. https://doi.org/10.1098/rspb.2019.2364.

Park C, Qian W, Zhang J. Genomic evidence for elevated mutation rates in highly expressed genes. EMBO Rep. 2012:13(12): 1123–1129. https://doi.org/10.1038/embor.2012.165.

Piganeau G. A planktonic picoeukaryote makes big changes to the green lineage. Nat Ecol Evol. 2020:4(9):1160–1161. https://doi.org/10.1038/s41559-020-1244-0.
Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010:26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033.
Rand A. 2024. nanoporetech/modkit. https://github.com/nanoporetech/modkit (Accessed 2024 August 22).
R Core Team. 2021. R: The R Project for Statistical Computing. R Lang Environ Stat Comput. https://www.r-project.org/ (Accessed 2025 January 6).
Rodriguez-Galindo M, Casillas S, Weghorn D, Barbadilla A. Germline *de novo* mutation rates on exons versus introns in humans. Nat Commun. 2020:11(1):3304. https://doi.org/10.1038/s41467-020-17162-z.
Sahu SK, Thangaraj M, Kathiresan K. DNA extraction protocol for plants with high levels of secondary metabolites and polysaccharides without using liquid nitrogen and phenol. Int Sch Res Not. 2012:2012:e205049. https://doi.org/10.5402/2012/205049.
Sandler G, Bartkowska M, Agrawal AF, Wright SI. Estimation of the SNP mutation rate in two vegetatively propagating species of duckweed. G3 (Bethesda). 2020:10(11):4191–4200. https://doi.org/10.1534/g3.120.401704.
Sebastian R, Oberdoerffer P. Transcription-associated events affecting genomic integrity. Philos Trans R Soc B Biol Sci. 2017:372(1731):20160288. https://doi.org/10.1098/rstb.2016.0288.
Selby CP, Lindsey-Boltz LA, Li W, Sancar A. Molecular mechanisms of transcription-coupled repair. Annu Rev Biochem. 2023:92(1):115–144. https://doi.org/10.1146/annurev-biochem-041522-034232.
Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res. 2013:41(4):2073–2094. https://doi.org/10.1093/nar/gks1205.
Shen J-C, Rideout WM III, Jones PA. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. 1994:22(6):972–976. https://doi.org/10.1093/nar/22.6.972.
Shewaramani S, Finn TJ, Leahy SC, Kassen R, Rainey PB, Moon CD. Anaerobically grown *Escherichia coli* has an enhanced mutation rate and distinct mutational spectra. PLoS Genet. 2017:13(1):e1006570. https://doi.org/10.1371/journal.pgen.1006570.
Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024:42(10):1571–1580. https://doi.org/10.1038/s41587-023-02024-y.
Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci U S A 2012:109(45):18488–18492. https://doi.org/10.1073/pnas.1216223109.
Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013:14(2):178–192. https://doi.org/10.1093/bib/bbs017.
Vasimuddin Md, Misra S, Li H, Aluru S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: 2019 IEEE international parallel and distributed processing symposium (IPDPS); Rio de Janeiro, Brazil. IEEE; 2019. p. 314–324. https://doi.org/10.1109/IPDPS.2019.00041.
Waldvogel A-M, Pfenninger M. Temperature dependence of spontaneous mutation rates. Genome Res. 2021:31(9):1582–1589. https://doi.org/10.1101/gr.275168.120.
Wang Y, Obbard DJ. Experimental estimates of germline mutation rate in eukaryotes: a phylogenetic meta-analysis. Evol Lett. 2023:7(4):216–226. https://doi.org/10.1093/evlett/qrad027.
Winnepenninckx DW. Extraction of high molecular weight DNA from molluscs. Trends Genet. 1993:9(12):407. https://doi.org/10.1016/0168-9525(93)90102-N.
Wright F. The 'effective number of codons' used in a gene. Gene. 1990:87(1):23–29. https://doi.org/10.1016/0378-1119(90)90491-9.
Wu K, Li H, Wang Y, Liu D, Li H, Zhang Y, Lynch M, Long H. Silver nanoparticles elevate mutagenesis of eukaryotic genomes. G3 (Bethesda). 2023:13:jkad008. https://doi.org/10.1093/g3journal/jkad008.
Xu S, Stapley J, Gablenz S, Boyer J, Appenroth KJ, Sree KS, Gershenzon J, Widmer A, Huber M. Low genetic variation is associated with low mutation rate in the giant duckweed. Nat Commun. 2019:10(1):1243. https://doi.org/10.1038/s41467-019-09235-5.
Yang Z, Ma X, Wang Q, Tian X, Sun J, Zhang Z, Xiao S, De Clerck O, Leliaert F, Zhong B. Phylotranscriptomics unveil a paleoproterozoic-mesoproterozoic origin and deep relationships of the *Viridiplantae*. Nat Commun. 2023:14(1):5542. https://doi.org/10.1038/s41467-023-41137-5.
Zhu YO, Siegal ML, Hall DW, Petrov DA. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci U S A 2014:111(22):E2310–E2318. https://doi.org/10.1073/pnas.1323011111.
Ziller MJ, Hansen KD, Meissner A, Aryee MJ. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. Nat Methods 2015:12(3):230–232. https://doi.org/10.1038/nmeth.3152.

**Associate editor:** Rebecca Zufall