



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

De novo mutation rate variation and its determinants in *Chlamydomonas*

Citation for published version:

López-Cortegano, E, Craig, RJ, Chebib, J, Samuels, T, Morgan, AD, Kraemer, SA, Böndel, KB, Ness, RW, Colegrave, N, Keightley, PD & Hendrickson, H (ed.) 2021, 'De novo mutation rate variation and its determinants in *Chlamydomonas*', *Molecular Biology and Evolution*, vol. 38, no. 9, msab140, pp. 3709–3723. <https://doi.org/10.1093/molbev/msab140>

Digital Object Identifier (DOI):

[10.1093/molbev/msab140](https://doi.org/10.1093/molbev/msab140)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

Molecular Biology and Evolution

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



De Novo Mutation Rate Variation and Its Determinants in *Chlamydomonas*

Eugenio López-Cortegano ^{*},¹ Rory J. Craig,¹ Jobran Chebib ¹ Toby Samuels,¹ Andrew D. Morgan,¹ Susanne A. Kraemer,² Katharina B. Böndel ³ Rob W. Ness ⁴ Nick Colegrave,¹ and Peter D. Keightley¹

¹Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

²Department of Biology, Concordia University, Montreal, QC, Canada

³Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany

⁴Department of Biology, University of Toronto Mississauga, Mississauga, ON, Canada

^{*}Corresponding author: E-mail: e.lopez-cortegano@ed.ac.uk.

Associate editor: Heather Hendrickson

Abstract

De novo mutations are central for evolution, since they provide the raw material for natural selection by regenerating genetic variation. However, studying de novo mutations is challenging and is generally restricted to model species, so we have a limited understanding of the evolution of the mutation rate and spectrum between closely related species. Here, we present a mutation accumulation (MA) experiment to study de novo mutation in the unicellular green alga *Chlamydomonas incerta* and perform comparative analyses with its closest known relative, *Chlamydomonas reinhardtii*. Using whole-genome sequencing data, we estimate that the median single nucleotide mutation (SNM) rate in *C. incerta* is $\mu = 7.6 \times 10^{-10}$, and is highly variable between MA lines, ranging from $\mu = 0.35 \times 10^{-10}$ to $\mu = 131.7 \times 10^{-10}$. The SNM rate is strongly positively correlated with the mutation rate for insertions and deletions between lines ($r > 0.97$). We infer that the genomic factors associated with variation in the mutation rate are similar to those in *C. reinhardtii*, allowing for cross-prediction between species. Among these genomic factors, sequence context and complexity are more important than GC content. With the exception of a remarkably high C→T bias, the SNM spectrum differs markedly between the two *Chlamydomonas* species. Our results suggest that similar genomic and biological characteristics may result in a similar mutation rate in the two species, whereas the SNM spectrum has more freedom to diverge.

Key words: *Chlamydomonas incerta*, *Chlamydomonas reinhardtii*, comparative mutability, mutation accumulation, mutation rate, mutation spectrum.

Introduction

Mutation plays a key role in evolution, since it generates genetic variation, providing the raw material for selection and adaptation. New mutational variance and heritability are important for determining the long-term response to selection (Walsh 2004; Mulder et al. 2019), and thus the evolutionary potential of populations. Standing genetic variation is also strongly influenced by variants continuously regenerated by mutation, so that heritability under mutation-drift equilibrium depends directly on the input of mutational variation (Lynch et al. 1999; Walsh and Lynch 2018). Since mutation also underlies genetic differentiation between lineages, it influences evolutionary divergence rates (Keightley 2012). Moreover, when new mutations have direct effects on phenotypes, particularly fitness and health, they also have a major impact in applied fields, such as conservation biology and medicine (Charlesworth 2018; Zhang and Vijg 2018). A better understanding of the rate of mutation and the distribution of mutational effects is one of the key goals in evolutionary biology (Eyre-Walker and Keightley 2007).

From an evolutionary perspective, the mutation rate itself can be regarded as a quantitative trait, which is modulated by natural selection and genetic drift (Lynch and Conery 2003; Lynch 2010). The drift-barrier hypothesis proposes that selection drives mutation rates to low values, minimizing the deleterious load, and that this process is more efficient in populations of higher effective size (N_e) (Lynch 2011; Sung, Ackerman, et al. 2012). However, although N_e is expected to play a central role in mutation rate evolution, there are likely to be many other genetic and biological factors that are involved in the evolution of mutation rates and its differentiation between species, such as the size of the functional genome (Sung, Ackerman, et al. 2012) and an organism's life cycle (Sung, Tucker, et al. 2012; Long et al. 2015). The mutation rate is also affected by environmental conditions (Bjedov et al. 2003), and its evolution could in part be driven by factors including spatial and temporal heterogeneity (de Visser 2002). From a genomic point of view, the mutation rate has been observed to be highly heterogeneous along the genome, and many factors contribute to genomic variation in

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

mutability, including nucleotide context, GC content, and DNA repetitiveness, among others (Mirkin 2008; Stamatoyanopoulos et al. 2009; Muzny et al. 2012; Aggarwala and Voight 2016; Sanjuán and Domingo-Calap 2016; Sassa et al. 2016; Frigola et al. 2017; Leffak et al. 2017; Kessler et al. 2020; McKinney et al. 2020). However, whether or not these correlates of mutation rate are causal, how they compare across species and the extent of their involvement in the evolution of the mutation rate remain open areas of study.

Because individual mutations are very rare, we have only recently been able to study large sets of mutations by combining whole-genome sequencing (WGS) and mutation accumulation (MA) experiments (Katju and Bergthorsson 2019). In an MA experiment, inbred or clonal lines are maintained with minimal N_e so that selection is ineffective and newly arising mutations can drift to fixation. This approach has been used in a variety of organisms (Keightley et al. 2009; Denver et al. 2012; Zhu et al. 2014; Flynn et al. 2017; Hamilton et al. 2017; Long et al. 2018; Krasovec et al. 2019; Weng et al. 2019; Kucukyildirim et al. 2020; Chebib et al. 2021), but has been generally limited to phylogenetically distant model organisms and rarely applied in closely related species to enable comparative investigation of mutation.

For example, in the distantly related yeast species *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, Farlow et al. (2015) observed similar mutation rates, but different spectra of single nucleotide mutations (SNMs). Although the similarity of the mutation rates agreed with the drift-barrier expectation, these yeast species are so distantly related that their genomes share essentially no detectable synteny (Rhind et al. 2011), and a comparative analysis on their mutation properties provide little insight into the phylogenetic scale over which changes in the mutation spectrum evolves. Denver et al. (2012) addressed the importance of estimating mutational properties in more closely related species by studying several genotypes of two *Caenorhabditis* species that diverged approximately 100 Ma (Stein et al. 2003). In their MA experiment, neither the mutation rate nor SNM spectra differed significantly among species or genotypes, indicating that the mutation rate and spectrum may be evolutionarily stable in this clade. More recently, Terekhanova et al. (2017) analyzed sequence data from human and other primates and showed that estimates of the mutation rate in shared genomic windows (i.e., local mutation rates) are similar between closely related species (e.g., human and chimpanzee), but the correlation between mutation rate estimates decays with phylogenetic distance. Nonetheless, the scale over which the mutation rate and spectrum diverge remains unclear, and more information on the evolution of mutation processes in phylogenetically closely related species is needed.

Here, we present an MA experiment in *Chlamydomonas incerta*, the closest known relative of the green alga *Chlamydomonas reinhardtii*, which has emerged as a model for the study of rate and fitness effects of de novo mutation (Ness et al. 2012; Sung, Ackerman, et al. 2012; Morgan et al. 2014; Ness et al. 2015; Kraemer et al. 2017; Böndel et al. 2019).

Although the taxonomic classification of *C. incerta*, which is also referred to as *C. globosa*, is subject to ongoing debate, it has long been recognized as a genetically and biologically distinct species from *C. reinhardtii* (Pröschold et al. 2005; Popescu et al. 2006; Nakada et al. 2010). The two species exhibit ~34% divergence at 4-fold degenerate sites, likely diverged less than 100 Ma and have highly syntenic genomes with similar gene contents (Craig et al. 2021). A highly contiguous and well-annotated *C. incerta* genome assembly has recently been produced using Pacific Biosciences sequencing, enabling comparative genomics analyses between *C. reinhardtii* and its closest relative to be performed for the first time (Craig et al. 2021). Thus, a comparative study on the mutation rate, its spectrum, and the genomic factors related to mutability should lead to a better understanding of the mutation process and its evolution. Based on predictive statistical models, the nearly 6,000 mutations identified in *C. reinhardtii* by Ness et al. (2015) provided several insights into the genomic factors related to mutability. Here, we also study the *C. reinhardtii* data from Ness et al. (2015) together with nearly 2,000 new SNMs identified in *C. incerta* and investigate the evolution of mutational properties in the two species. Specifically, we address the following questions: 1) Is the rate and base spectrum similar between the two species? 2) Is the extent of mutation rate variation between lines and across the genome similar in the two species? 3) Are predictors of mutation rate variation in the genome similar in the two species?

Results

Mutation Rates

A total of 27 MA lines of *C. incerta* were maintained for an average of 788 generations before performing whole-genome resequencing. The analysis of the re-sequencing data aligned to the *C. incerta* reference genome (129.2 Mb) gave an overall proportion of 72% high-quality sites (i.e., callable sites), where candidate mutations could be called. The fraction of callable sites (the callable rate) was consistent across MA lines (Kruskal–Wallis, KW test, $\chi^2_{26} = 30.41$, $P = 0.25$), but varied significantly between contigs (KW test, $\chi^2_{115} = 2.6 \times 10^6$, $P < 2.2 \times 10^{-16}$), and there was generally a higher callability in larger contigs (supplementary fig. S1a–c, Supplementary Material online) and the plastid (82%) and mitochondrial genomes (92%). The positive relationship between contig length and callability is likely to be a consequence of the highly repetitive content of many short contigs (Craig et al. 2021), which foil assembly and lead to low mapping quality (MQ) due to read misalignments (supplementary fig. S1b, Supplementary Material online). Although *C. incerta* has a higher mapped repeat content than *C. reinhardtii*, its callable rate is higher than that previously repeat content (~28% vs. ~22%) than the smaller *C. reinhardtii* genome (~111 Mb), its callable rate is higher than that previously obtained in several strains of *C. reinhardtii* (Ness et al. 2015). This is presumably the result of the *C. incerta* MA lines being derived from the same strain as was used to produce the genome assembly, unlike in *C. reinhardtii* where field isolates exhibiting

substantial genetic variation relative to the reference genome were studied. The callable rate was variable between different classes of genomic sites (supplementary fig. S1d, Supplementary Material online, KW test, $\chi^2_4 = 7.2 \times 10^6$, $P < 2.2 \times 10^{-16}$) and was negatively correlated with the proportion of repetitive sequence (Pearson's product-moment correlation, $t_3 = -3.56$, $r = -0.90$, $P = 3.8 \times 10^{-2}$).

Based on the callable portion of the *C. incerta* genome (~84 Mb), a total of 2,609 de novo mutations were found, leading to an average mutation rate estimate per site per generation of $\mu \approx 15.10 \times 10^{-10}$. There were 1,991 SNMs ($\mu_{\text{SNM}} = 11.56 \times 10^{-10}$), 350 deletions ($\mu_{\text{DEL}} = 2.03 \times 10^{-10}$), and 268 insertions ($\mu_{\text{INS}} = 1.56 \times 10^{-10}$). These numbers are similar to those previously obtained in *C. reinhardtii*, that is, Ness et al. (2015) estimated $\mu = 11.5 \times 10^{-10}$ ($\mu_{\text{SNM}} = 9.63 \times 10^{-10}$, $\mu_{\text{INS}} + \mu_{\text{DEL}} = 1.90 \times 10^{-10}$), with an average number of SNMs detected per line and generation similar to the number in the present experiment ($\sim 9.72 \times 10^{-2}$ in *C. incerta* vs. $\sim 7.15 \times 10^{-2}$ in *C. reinhardtii*). The median $\mu_{\text{SNM}} = 7.62 \times 10^{-10}$ in *C. incerta* was also similar to that of *C. reinhardtii* ($\mu_{\text{SNM}} = 5.27 \times 10^{-10}$) and fell within the range observed for *C. reinhardtii* strains (Ness et al. 2015). In contrast, the average number of insertion and deletion variants (INDELs) per line and generation was higher in *C. incerta* (3.02×10^{-2} vs. 1.41×10^{-2}), but this could be caused by differences in the callability of these variants (see below). No mutations were found in the mitochondrial genome, presumably because of its small size (~17.6 kb), and only two SNMs were found in the plastid genome ($\mu_{\text{PLASTID}} = 5.4 \times 10^{-10}$), resulting in an estimate of the mutation rate that is similar to that observed in *C. reinhardtii* ($\mu_{\text{PLASTID}} = 7.7 \times 10^{-10}$). The plastid genome mutation rate is similar to the nuclear genome mutation rate in these species, a finding that contrasts with land plants (Smith and Keeling 2015; Ness et al. 2016). We restrict all further analyses to mutations found in the nuclear genome. Supplementary table S1, Supplementary Material online contains a list of all SNMs and INDELs found.

Although the *C. incerta* MA lines were derived from a single ancestral strain, the mutation rate was highly variable among lines, ranging over more than two orders of magnitude, between $\mu = 0.35 \times 10^{-10}$ in line 3 with only 3 SNMs, to $\mu = 131.7 \times 10^{-10}$ in line 27 with 829 SNMs (fig. 1A). The distribution of the mutation rate among MA lines was highly leptokurtic (supplementary fig. S2a, Supplementary Material online), fitting better a lognormal distribution (meanlog = -3.01, sdlog = 1.15, Kolmogorov-Smirnov test, KS test, $D = 0.14$, $P = 0.60$) than any other distribution tested, including an exponential or gamma distribution. High variability among lines derived from the same genetic background was also observed in *C. reinhardtii* (Ness et al. 2015, see supplementary fig. S2b, Supplementary Material online), and the distribution of mutation rates in MA lines derived from *C. reinhardtii* strains CC-2344 and CC-2931 resembles that of *C. incerta* (KS test, $D < 0.4$, $P > 0.07$, supplementary fig. S2b, Supplementary Material online). After excluding hypermutant line 27, the variance in the number of mutations between lines of *C. incerta*, assuming an equal number of generations between MA lines, was still substantial ($\sigma^2_\mu =$

1070.03). This variation is approximately 22-fold higher than that expected from a Poisson distribution ($\lambda = 48.44$, KS test, $D = 0.38$, $P = 1.45 \times 10^{-3}$), a distribution that is commonly assumed to represent the mutation processes (Charlesworth 2012). We also compared the observed distribution of mutation rates among MA lines with that obtained from computer simulations in which mutations arose independently, using the software SLiM (Haller and Messer 2019). The variance in the number of mutations among simulated lines was much closer to the Poisson expectation (median variance over replicates = 75.05, 95% CI = 39.87–128.04) than to the observed variance among the MA lines. High mutation rate variability between lines derived from the same ancestor genotype has been observed in other species (Dumont 2019; Ho et al. 2020) and supports the idea that substantial variability of the mutation rate may be common. Additional results and discussion on mutation rate variability and the hypermutant line 27 can be found in supplementary file S1, Supplementary Material online.

The mutation rate also varied significantly between contigs (KW test, $\chi^2_{115} = 158.86$, $P = 4.24 \times 10^{-3}$) and was generally higher and more variable in shorter contigs (Pearson's product-moment correlation, $t_{108} = -2.59$, $r = -0.24$, $P = 0.01$, supplementary fig. S3a, Supplementary Material online). Interestingly, the mutation rate also varied between gene-related features (KW test, $\chi^2_4 = 49.54 \times 10^6$, $P = 4.50 \times 10^{-10}$). For example, the mutation rate was 40–50% higher in untranslated regions (UTRs) and intergenic regions than in coding regions, whereas intronic sequences had an intermediate rate (supplementary fig. S3b, Supplementary Material online). In principle, a lower number of mutations in coding sequences could be due to selection during the transfer of colonies from one plate to another. However, the estimate of Ka/Ks in coding sequences was 0.84 (Fisher's exact test, $P = 0.07$), suggesting that little selection occurred, and that other genomic factors were therefore responsible for the variation in the mutation rate between genomic features. One possibility is that the DNA repair machinery is more efficient at preventing mutations in coding regions compared with intergenic sequences, as previously demonstrated in *Escherichia coli* and *Arabidopsis thaliana* MA lines that were mutant for mismatch repair proteins (Lee et al. 2012; Foster et al. 2015; Belfield et al. 2018). Higher mutation rates in UTRs were also observed in *C. reinhardtii*, although the intergenic mutation rate was not higher than that of coding sequence (supplementary fig. S4, Supplementary Material online). The two species have very similar amounts of annotated coding (39.5 Mb *C. incerta*, 37.7 Mb *C. reinhardtii*) and UTR (4.0, 3.9 Mb, respectively) sequence, and the ~18 Mb greater assembly size of *C. incerta* is largely associated with a relative increase in repetitive intergenic sequence (Craig et al. 2021). This may suggest that genome organization plays a role in the observed difference in intergenic mutation rates between the species, although we cannot exclude alternative explanations (e.g., differences in callable rates or annotation quality). Nonetheless, the common pattern observed for coding regions and UTRs supports the existence of a bias toward low mutation rate at coding regions in Chlorophyta, since

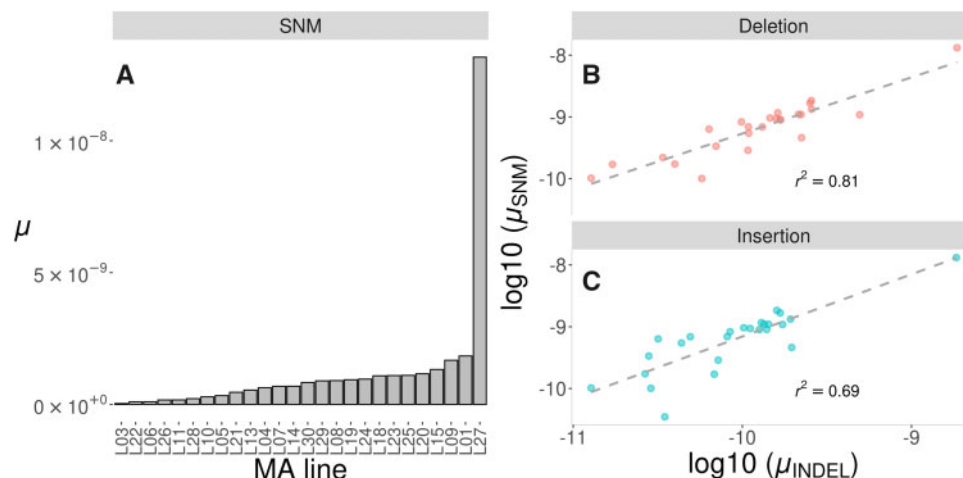


Fig. 1. (A) Mutation rate (μ) estimates for SNMs in the *C. incerta* MA lines. Lines are sorted from lowest (left, $\mu = 0.04 \times 10^{-9}$) to highest μ (right, $\mu = 13.2 \times 10^{-9}$). (B and C) Correlation between μ_{SNM} and μ_{INDEL} (on $\log_{10}$ scale) for deletions (B, in red) and insertions (C, in blue). The dashed lines show linear regression lines for μ_{SNM} as a function of μ_{INDEL} . Note the squared correlation coefficient is lower than that mentioned in the main text, because of the use here of a logarithmic scale.

similar results were observed in other algal species (Krasovec et al. 2017).

The distribution of inter-mutation distance (IMD) differed significantly from the distribution expected if SNMs occurred at random genomic positions (KS test, $D = 0.11$, $P < 2.2 \times 10^{-10}$). This was due to an over-representation of mutation pairs separated by less than 10 bp (supplementary fig. S5a, Supplementary Material online). Among these mutations, 17 SNM pairs out of 27 (63%) occurred at adjacent sites. The most highly represented dinucleotide mutation was CC (35%), which always mutated to AA, TA, or TT. C→A mutations were nearly 2-fold more frequent when compared with other SNM types at dinucleotide mutation sites, after correcting for GC content (supplementary fig. S5b, Supplementary Material online). Differences from random expectation in the number of mutations in genomic segments of different lengths (1, 10, 100, or 500 kb) were nonsignificant (KS test, $D < 0.08$, $P > 0.1$). These results are also in broad agreement with previous findings in *C. reinhardtii* (Ness et al. 2015), indicating not only that mutation rates are similar between *C. incerta* and *C. reinhardtii*, but that mutations are similarly distributed across the genome.

INDELs and Structural Variants

Three different software packages were used to detect short INDELs and structural variants: Freebayes (Garrison and Marth 2012), GATK (Van der Auwera et al. 2013), and Pindel (Ye et al. 2009). In contrast to SNMs, many of the INDELs and structural variants detected could not be confirmed by visualization in IGV (Thorvaldsdóttir et al. 2013), or the visualized variants did not exactly correspond with the calls in the VCF files. Therefore, to minimize the number of false positives, INDELs, and structural variants were manually curated following visualization by IGV. On this basis, approximately 56% of deletions and 10% of insertions detected by Pindel were rejected, whereas only 6% and 14% of the deletions and insertions called by GATK were excluded. Only 1%

of deletions found by Freebayes were rejected. In general, most INDELs were found using GATK, followed by Pindel and Freebayes (supplementary fig. S6, Supplementary Material online). A false-positive rate in *C. reinhardtii* was estimated from our previous data (Ness et al. 2015) as the percentage of mutations originally called in a line but not verifiable in a visible inspection prior to genotyping them in recombinant lines (Böndel et al. 2019). This rate suggested more false positives for INDELs (7.7%) than SNMs (2.3%), and it is expected to be of the same magnitude in *C. incerta*, given the similar approach used for sequencing, alignment, and calling.

After filtering, deletions were significantly more frequently retained than insertions ($\chi^2_1 = 10.84$, $P < 0.001$), and the estimated mutation rates were $\mu_{\text{DEL}} = 2.03 \times 10^{-10}$ for deletions and $\mu_{\text{INS}} = 1.56 \times 10^{-10}$ for insertions. However, it should be noted that the short-read sequencing used here makes deletions of all sizes easier to detect than insertions. The INDEL mutation rate was higher in *C. incerta* than in *C. reinhardtii* ($\mu_{\text{DEL}} = 1.03 \times 10^{-10}$, $\mu_{\text{INS}} = 0.87 \times 10^{-10}$, Ness et al. 2015), even when variants only called by GATK were considered ($\mu_{\text{DEL}} = 1.58 \times 10^{-10}$, $\mu_{\text{INS}} = 1.40 \times 10^{-10}$). This was the only variant caller used to analyze the *C. reinhardtii* data.

The numbers of deletions and insertions were highly variable among MA lines, but were strongly positively correlated with the number of SNMs (Pearson's product-moment correlation, $t_{24} > 18$, $r_{\text{SNM-DEL}} = 0.97$, $r_{\text{SNM-INS}} = 0.99$, $P < 1 \times 10^{-15}$; fig. 1B and C and supplementary fig. S7a, Supplementary Material online), even after excluding the hypermutant line ($t_{23} > 5$, $r_{\text{SNM-DEL}} = 0.72$, $r_{\text{SNM-INS}} = 0.78$, $P < 1 \times 10^{-4}$). This suggests that the mechanisms responsible for the occurrence of SNMs and INDELs are related. Deletions were generally larger than insertions (Wilcoxon rank-sum test, $W = 58,003$, $P = 1.49 \times 10^{-8}$, supplementary fig. S7b, Supplementary Material online), that is the median length was 1 bp for insertions and 2 bp for deletions, although

larger deletions were also detected. There were 47 deletions longer than 100 bp ($\mu = 2.15 \times 10^{-11}$), and 8 of them were longer than 10 kb ($\mu = 3.66 \times 10^{-12}$). Short INDELs (<50 bp) were slightly shorter than in *C. reinhardtii*, that is mean lengths were 3.6 bp for deletions and 2.3 bp for insertions (vs. 7.9 bp for deletions and for insertions 5.9 bp in *C. reinhardtii*). In addition to INDELs, five large inversions were found by Pindel ($\mu_{INV} = 2.2 \times 10^{-12}$), with sizes ranging from ~700 to ~3,700 bp. No insertions or tandem repeat variants larger than 50 bp were found. As mentioned above, INDEL detection was more challenging than detection of SNM variants, so only SNMs will be considered for further analyses, unless otherwise stated.

Factors Influencing Mutability

To examine the genomic factors associated with mutability, we employed a regularized logistic regression model to predict mutated sites from their genomic properties. The use of a regularized regression algorithm reduces the effect of correlation between the multiple predictor variables fitted in the model. For both *C. incerta* and *C. reinhardtii*, a training set composed of all identified SNMs along with a random sample of 10^5 nonmutated callable sites was used for fitting the model. To test how well the model predicted mutability, a test set containing all SNMs and a larger set of 10^6 randomly selected callable sites was used. A cross-validation scheme between species was implemented by running the model fitted to one species in order to predict mutability in the other. The predictive mutability model always returned statistically significant results, including for the cross-predictions ($P < 1 \times 10^{-3}$). There was also a strong linear relationship between the predicted and the observed mutability, which was higher within species than between species (fig. 2). The correlation between the regression coefficient estimates in the two species was significantly positive (Pearson's product-moment correlation, $t_{141} = 5.77$, $r = 0.44$, $P = 4.86 \times 10^{-8}$), and six out of the ten most important genomic factors associated with mutability were shared by *C. incerta* and *C. reinhardtii* (fig. 3). Supplementary figure S8, Supplementary Material online shows the ten most important predictors for *C. incerta* and *C. reinhardtii*, and a complete list including all raw regression coefficient estimates is given in supplementary table S2, Supplementary Material online. These results indicate that the mechanisms associated with mutation in the two species are highly related.

Remarkably, the trinucleotides CTC and CAC (where the underlined C represents the mutated site) had a similar effect on mutability in the two species. The CTC trinucleotide was the single most important factor in *C. reinhardtii*, and its high mutability has been previously reported in this species (Ness et al. 2015), as well as in other phylogenetic groups, including fungi (Zhu et al. 2014) and animals (Alexandrov et al. 2013). This finding highlights the importance of sequence context variation for variation in mutability (Ling et al. 2020), which is supported by the large amount of variation in the effect of different triplet sequences on mutability (supplementary table S2, Supplementary Material online). For example,

regarding the upstream context of C nucleotides, there were regression coefficients both positively and negatively associated with mutability in *C. incerta* (fig. 4A). In *C. reinhardtii*, most coefficient estimates related to sequence context were close to zero (supplementary fig. S9a, Supplementary Material online), probably due to differences in the penalty term of the regularized regression model in the two species. However, the observed proportion of SNMs in different sequence contexts (fig. 5) was highly variable in the two species, and strongly correlated between them (Pearson's product-moment correlation, $t_{93} = 13.14$, $r = 0.81$, $P < 10^{-15}$).

GC content is the variable most commonly used to reduce local genomic information into a single value, since it is typically correlated with important biological features, including the mutation rate (Krasovec et al. 2017). Here, we explored the role of two additional parameters in relation to mutation, the local Shannon Entropy (E) and sequence linguistic complexity (L), both of which provide insight into the amount of noncompressible information contained in DNA sequences (Schneider 2000; Vinga 2014). In simple terms, E measures nucleotide repetitiveness (it is maximal when all nucleotides have equal frequencies, $p = 0.25$, and minimal when $p = 1$ for one nucleotide), whereas L measures sequence repetitiveness (it is maximal in regions where all possible different sequences of a given length or range of lengths are represented). We also quantified GC content and variation in GC, E and L within genomic windows (named ΔGC , ΔE , and ΔL , respectively). Thus, although GC measures the mean GC content within a genomic window, ΔGC measures the extent of GC variation within that window. Most genomic regions are characterized by having high L and low ΔL , ΔE , and ΔGC . Conversely, mutated sites were usually located in regions with low L , high ΔL , ΔE , and ΔGC , or in regions with unusually high or low values of E or GC (fig. 4B and supplementary fig. S9b, Supplementary Material online). This agrees with previous results suggesting that deviations in GC content from the genomic equilibrium lead to increased mutability (Krasovec et al. 2019). Interestingly, predictors related to nucleotide and sequence complexity showed a stronger association with mutability than GC content in both species (fig. 3 and supplementary table S2, Supplementary Material online). For example, ΔL in small windows (≤ 20 bp) showed the strongest association with mutability in *C. incerta*, and E and L were generally more informative about mutability than measures based on GC in both species (supplementary fig. S8 and table S2, Supplementary Material online). Exceptionally, in *C. reinhardtii* GC content at a mutated site showed a strong association with mutability, probably related to the hypermutability of CTC and CAC trinucleotides, and to the higher probability of mutation at G:C sites than at A:T sites in this species (see below).

In view of previous results on the importance of nucleotide and sequence complexity measures for predicting mutability, we also estimated the correlation between the mutation rate and distance from repetitive sequence regions annotated as low complexity regions and microsatellites. However, correlations were nonsignificant for both low complexity regions

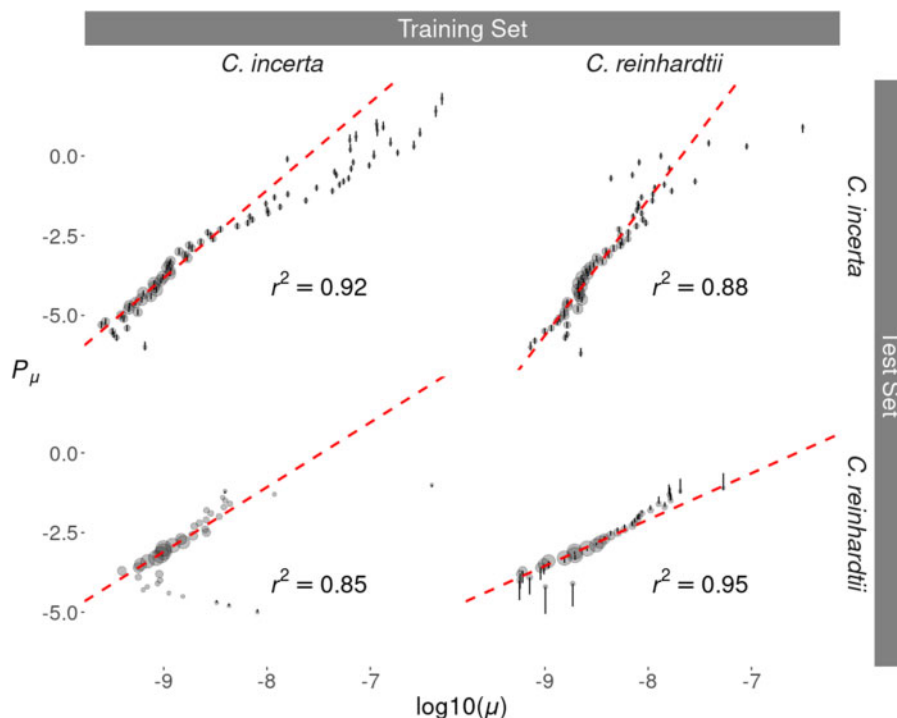


FIG. 2. Correspondence of predicted and observed mutability within and across the *Chlamydomonas incerta* and *Chlamydomonas reinhardtii* genomes. Predicted mutability (P_μ) was obtained for the *C. incerta* and *C. reinhardtii* validation data (top and bottom plots, respectively) using models trained with either *C. incerta* or *C. reinhardtii* data (left and right plots, respectively). Genomic regions were binned in groups with the same P_μ value (rounded to 1 decimal place) and its value was compared with the observed mutation rate (μ) of the same sites. Vertical bars show 95% confidence intervals of P_μ values obtained over ten replicates using different training data sets. Red dashed lines show the linear fit of the predictions, weighted by the number of sites in each bin (larger points indicate more observations). All fits were significant ($P < 10^{-3}$), with coefficients of determination (r^2) shown in the figure.

(Pearson's product-moment correlation, $t_{361177} = -0.34$, $r = -5.68 \times 10^{-4}$, $P = 0.73$) and microsatellites ($t_{20111} = 0.30$, $r = 2.14 \times 10^{-3}$, $P = 0.76$). It should be noted that our annotation of repetitive sequences did not include large satellite DNA, because these regions are not expected to be callable.

To further explore repetitive patterns that could help to explain the higher mutability of genomic regions of low linguistic complexity, we expanded our analysis to include alternate DNA conformations, which are usually characterized by repetitive sequences. This analysis was done using NeSSie (Berselli et al. 2018) and QPARSE (Berselli et al. 2020) to detect the absence/presence of potential motifs for DNA triplexes, G-quadruplexes, mirrors, and palindromes. When considering small genomic windows of 10 bp extending either side of a mutated or a randomly sampled site, the proportion of genomic regions containing potential DNA triplex motifs was approximately 4-fold higher in regions containing a mutated site than other genomic regions (KW test, $\chi^2_1 = 30.76$, $P = 2.92 \times 10^{-8}$, fig. 6). Sequence mirrors, potentially associated with DNA hairpins (Gajarský et al. 2017), were also more frequently found near mutated sites (KW test, $\chi^2_1 = 7.95$, $P = 4.8 \times 10^{-3}$), whereas palindromes were slightly under-represented (KW test, $\chi^2_1 = 4.51$, $P = 0.03$). When analyzing genomic windows longer than 20 bp (0.2 and 2 Kb), no significant enrichment was found for any of the DNA sequence motifs considered (supplementary fig. S10, Supplementary

Material online). Thus, it is possible that mutability is influenced by close proximity to DNA motifs that lead to alternate conformations such as DNA triplexes or hairpins. The bioinformatic analysis presented here, however, assumed that these motifs are truly indicative of the presence of DNA secondary structure in *C. incerta*, but there is no experimental confirmation, and therefore results should be interpreted with caution. More studies addressing sequence complexity and DNA secondary structure are needed, particularly in the context of de novo mutation.

SNM Spectrum

The SNM spectrum departed from the expectation of equally frequent SNM types, after correction for the genomic mean GC content of 66% ($\chi^2_5 = 1,124.9$, $P < 2.2 \times 10^{-16}$). Specifically, transition point mutations were much more frequent than transversions, and C→T transitions were 2.35 times more frequent than the random expectation, and twice as frequent as A→G transitions (fig. 7A). Overall, C→T mutations represented nearly 52% of all SNMs. A similar pattern was observed by Ness et al. (2015) in *C. reinhardtii*. However, whereas transitions were the most frequent mutation type in *C. incerta*, C sites had a higher mutability in *C. reinhardtii* (i.e., C→A transversions were more frequent than A→G transitions), mainly due to the hypermutability of the trinucleotide

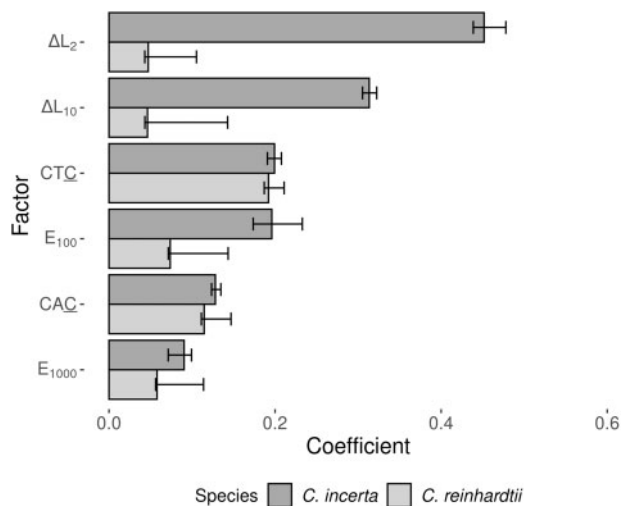


Fig. 3. Regression coefficient estimates obtained from the predictive model of genomic mutability. Median estimates and 95% confidence intervals are shown based on ten training set replicates of *C. incerta* (dark bars) and *C. reinhardtii* (light bars). Only factors that were common between the ten most important ones estimated in *C. incerta* and *C. reinhardtii* are shown. The factors shown in the figure are: Variation in sequence repetitiveness (ΔL_2 and ΔL_{10} , measured in windows of 2 and 10 bp, respectively, extending downstream and upstream), nucleotide repetitiveness (E_{100} and E_{1000} , measured in windows of 100 and 1,000 bp, respectively, extending downstream and upstream), and trinucleotides CTC and CAC (where the underlined C is the reference site containing the mutation). Factors are sorted by their median effect sizes estimated in *C. incerta*. All predictors were standardized prior to regression, so their effect sizes are comparable.

CTC. Consequently, the mutation spectrum of the two species is qualitatively different (fig. 7B).

As previously done for mutability, a regularized predictive model was run to determine the genomic factors that explain the occurrence of different types of SNMs. Thus, the data set only included observations at SNM sites, and the response variable was first defined as the SNM type, a multinomial variable with six levels, one per type of SNM (A→C, A→G, A→T, C→A, C→G, and C→T). Models tested included regularized multinomial regression and classification machine learning algorithms, such as neural networks and random forests (details on the tuning of the different models used are given in Materials and Methods). A random sample of 75% SNMs was used for training and the remaining 25% for testing, so that accuracy could be measured as the proportion of SNM types correctly classified in the test set, relative to the total number of observations. A cross-validation scheme between species was also included. For both species, however, all classification models considered produced low (<33%) and nonsignificant accuracy values. Alternatively, when the response variable was set as a binomial variable indicating whether or not an SNM was of type C→T, accuracy increased to values close to 65% both in *C. incerta* (exact binomial test, EB test, $P < 1.6 \times 10^{-3}$) and in *C. reinhardtii* (EB test, $P = 0.02$). Accuracy was always nonsignificant for predictions across species, and the most important predictors

determining the occurrence of C→T mutations were different for each species (supplementary fig. S11 and table S3, Supplementary Material online). This indicates that the C→T bias is either associated with different factors in each species, the relative importance of the factors analyzed differ between the two species, or the most important explanatory variables regarding C→T prediction were not included in our model.

Discussion

We have conducted a comparative study of mutability in *Chlamydomonas* green algae using WGS data from an MA experiment in *C. incerta* and from an experiment previously carried out in *C. reinhardtii* (Ness et al. 2015). Given their relatively large genomes (111–129 Mb) and their short generation intervals, these unicellular species represent excellent models for investigating the nature of de novo mutations, since large numbers of mutations can be accumulated in a short time (~ 0.097 SNMs per line per generation), allowing the factors associated with the properties of new mutations to be investigated. Our analyses revealed that the mutation rate, the spatial distribution of mutations across the genome, and the genomic factors associated with mutability were similar between the two species, whereas the SNM spectra differed and were nonpredictable across species. The mutation rate was also found to be highly variable between lines derived from the same ancestral strain in both species.

It is perhaps unsurprising that the median SNM rate estimated in *C. incerta* (7.6×10^{-10}) is similar to that observed across strains of *C. reinhardtii* (5.3×10^{-10} , Ness et al. 2015), since they are closely related and have similar genomic architectures and total lengths of coding sequence. Earlier estimates of the mutation rate in *C. reinhardtii* were substantially smaller, that is 2.1×10^{-10} (Ness et al. 2012) and 0.7×10^{-10} (Sung, Ackerman, et al. 2012), but this difference could either be due to methodological differences in mutation detection or to between-strain variation in the mutation rate. Krasovec et al. (2017, 2018) estimated the mutation rate in five more distantly related green algal species that have much smaller genomes (in the range 12–21 Mb) and more variable GC contents (46–64%) than *Chlamydomonas*. In these species, estimates were nonetheless of the same order of magnitude as *C. incerta* and *C. reinhardtii* ($3.02 \times 10^{-10} \leq \mu\text{SNM} \leq 9.19 \times 10^{-10}$). Unfortunately, there are only two known isolates of *C. incerta* and no genetic diversity data are available for the species, so it is not currently possible to estimate N_e for the species, which can be used to test the “drift barrier” hypothesis (Sung, Ackerman, et al. 2012; Lynch et al. 2016).

Not only was the mutation rate similar between the two *Chlamydomonas* species, but also the genomic factors associated with mutability. There was a high correlation between the genomic predictors of mutability in the two species and an accurate cross-prediction of mutability between the species, suggesting that the modeled genomic factors were associated with the mutation rate in the species’ common ancestor. We found that sequence context and complexity

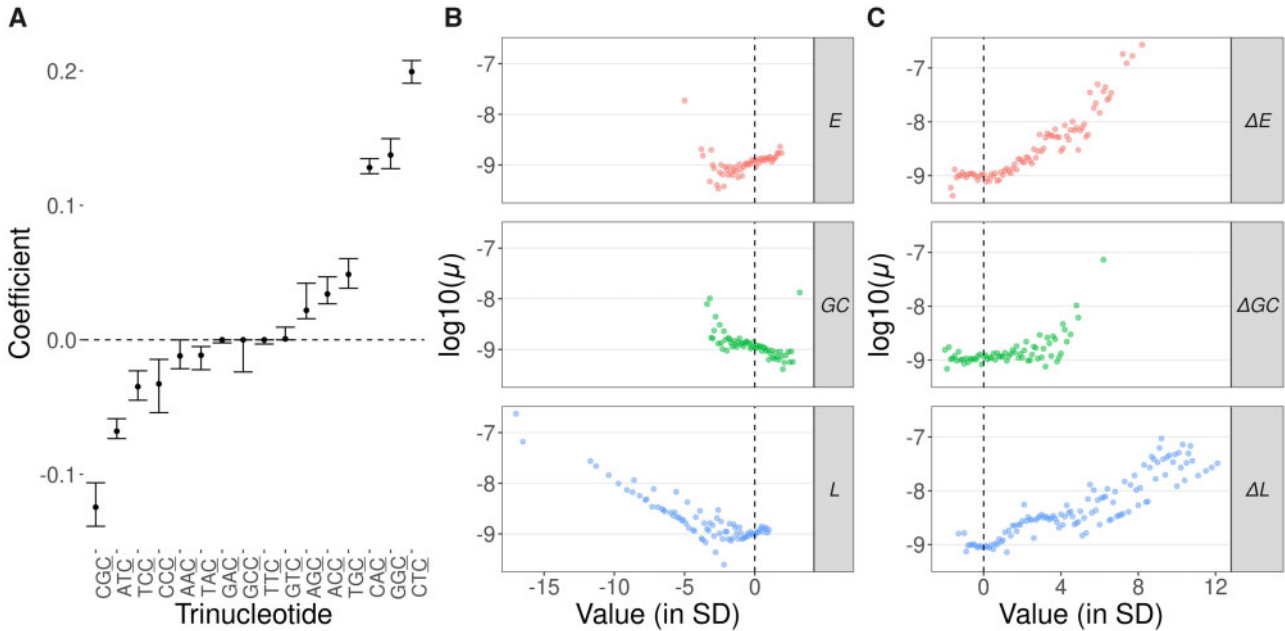


FIG. 4. Relationships between sequence context, base composition, sequence complexity, and mutability in *C. incerta*. (A) Regression coefficient estimates for the 16 possible dinucleotides upstream of reference C sites. (B) Relationship between scaled mean nucleotide repetitiveness (E), GC content (GC), and sequence repetitiveness (L), measured in genomic windows of 2 Kb. (C) Relationship between the scaled variation in nucleotide repetitiveness (ΔE), variation in GC content (ΔGC), and variation in sequence repetitiveness (ΔL), measured in a genomic windows of 20 bp. Note that standard deviations are used as the unit of measurement for the genomic parameters in plots (B) and (C). Only genomic sections where $\mu > 0$ are shown in (B) and (C).

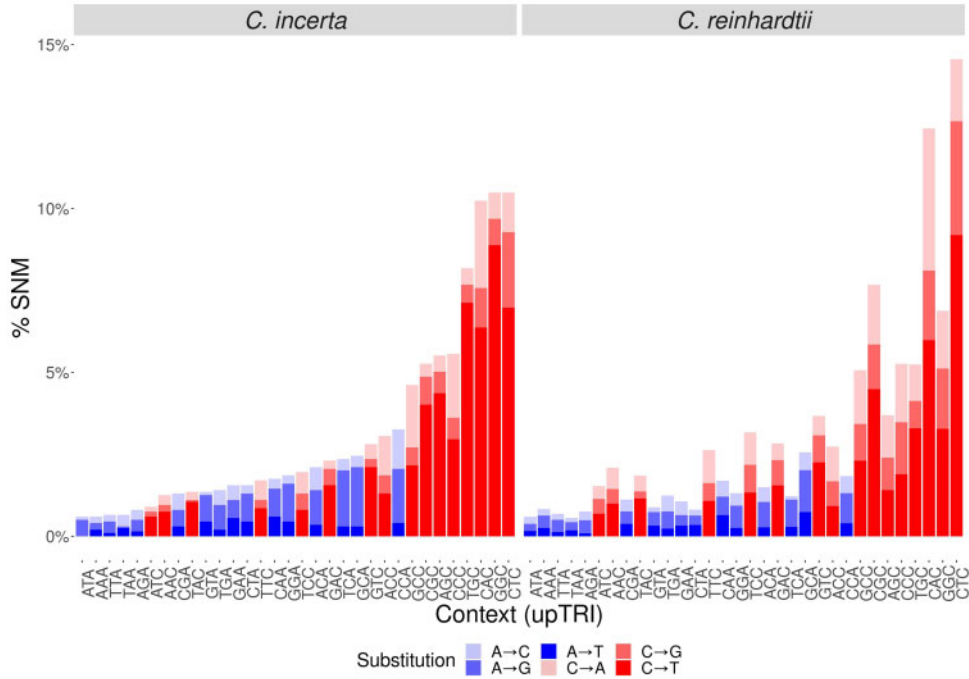


FIG. 5. Percentage of different types of SNM by upstream genomic context, defined as 2 bp upstream from a reference A or C site. Bars are colored by nucleotide composition at the reference site (A/T: blue, C/G: red). For each bar the different types of SNM are further colored in light (A→C and C→A), intermediate (A→G and C→G) or dark (A→T and C→T) colors. Sequence context is sorted on the x-axis by mutation frequency in *C. incerta*.

were the most important factors associated with mutability in *Chlamydomonas*. The mutation rate varied substantially among different trinucleotide sequence contexts (fig. 5) and

was especially high in the CTC and CAC contexts. Variation in sequence editing proteins with context-dependent activity, such as the APOBEC family of cytidine deaminases (Sanjuán

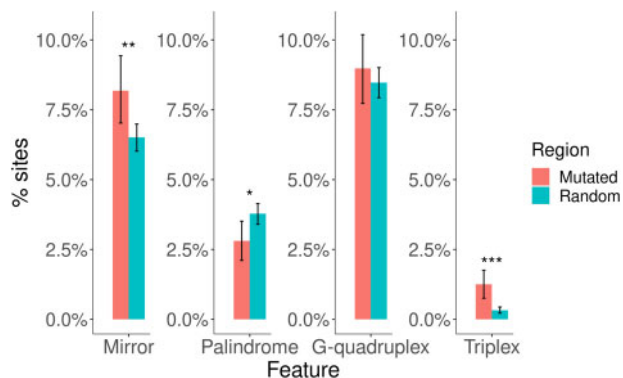


FIG. 6. Percentage of genomic sites containing at least one of the following features: mirror, palindrome, G-quadruplex, and triplex in windows of 21 bp. Regions are grouped into those containing known mutated sites, containing a total of 1,991 SNMs (in red), and 10^4 randomly sampled genomic locations (in blue). Significance is calculated using the Kruskal–Wallis test (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$). Confidence intervals (95%) are based on 1,000 bootstrap samples.

and Domingo-Calap 2016), could explain the strong association of particular sequence contexts with mutability. For example, *APOBEC3* is known to introduce C→T mutations and shows affinity for TC and CC substrates (Refsland and Harris 2013; Roberts et al. 2013). However, the true nature of the hypermutability of CTC and CAC contexts in *Chlamydomonas* remains unknown. Regarding measures of nucleotide and sequence complexity, these factors contributed more to the prediction of mutability than measures based on GC content alone (fig. 4 and supplementary fig. S9 and table S2, Supplementary Material online). The strong association between complexity-related measures and mutability might arise from the higher level of sequence information they provide compared with GC content. For example, nucleotide complexity (i.e., sequence entropy) is more sensitive to differences in the arrangement of nucleotides than GC content, and variation in complexity within genome sequences can be used to predict the presence of repetitive

sequences (Thanos et al. 2018). Linguistic sequence complexity might be associated with nonstandard DNA conformations, which usually involve some degree of repetitiveness (Wells 1988; Bacolla and Wells 2004). High mutagenicity for secondary structure-forming sequences might then explain the association between low complexity sequences and mutability, for example via double-strand breaks (DSBs) or interference with the DNA repair machinery. This association has been described for Z-DNA, which is associated with DSBs and mutation in humans and yeast (McKinney et al. 2020), and whose presence seems to be associated with alternate purine-pyrimidine repeats. Although it is clear that higher-order patterns within sequences strongly influence stability and mutation rates, the causal mechanisms remain to be determined.

In spite of the close phylogenetic relationship between *C. reinhardtii* and *C. incerta*, and the aforementioned similarities of their mutation rates and associated genomic factors, the SNM spectra of the two species showed substantial differences. Only a high C→T bias was in common between the two SNM spectra, but this bias is nearly universal, since it has been found both in prokaryote and eukaryote organisms (Hershberg and Petrov 2010; Ossowski et al. 2010; Farlow et al. 2015; Krasovec et al. 2019). More generally, transitions are the most common SNM type in *C. incerta*, whereas in *C. reinhardtii* SNMs at C sites represent the most common type of SNMs. Although differences in the SNM spectra may evolve as a consequence of environmental change (Liu and Zhang 2019), we do not expect this to be the case here, since *C. incerta* and *C. reinhardtii* experiments were performed at the same time, under the same environmental conditions, and we have no evidence on these conditions being more stressful for one species than the other. In an evolutionary context, similar mutation rates, but different SNM spectra, have been observed in other taxa, such as in the yeast species *S. cerevisiae* and *S. pombe* (Farlow et al. 2015), and in diverse green algal species (Krasovec et al. 2017). However, previous studies have involved species that are phylogenetically more distantly related than *C. incerta* is to *C. reinhardtii*. Thus, our results highlight that the SNM

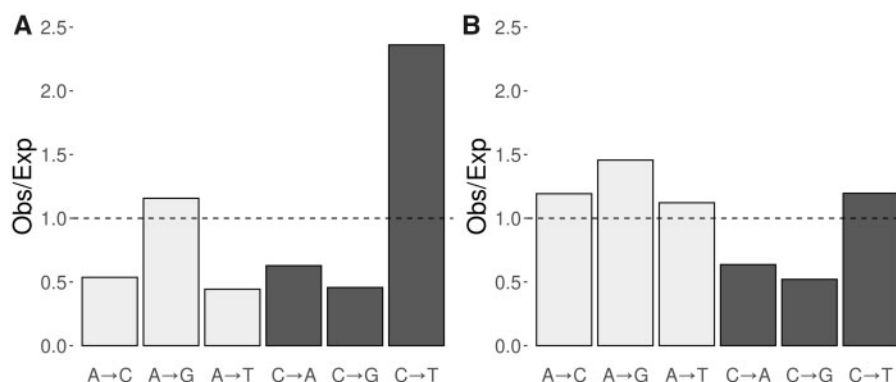


FIG. 7. Spectrum of SNMs. (A) SNM spectrum for *C. incerta*. The height of the bars represents the deviation of the observed value from the expectation for all SNMs being equally frequent, after correcting for GC content (mean GC = 66%). (B) SNM spectrum for *C. incerta* relative to *C. reinhardtii*. The height of the bars is calculated using the observed/expected deviation calculated in (A) for *C. incerta*, divided by the same measurement estimated for *C. reinhardtii*, using data from Ness et al. (2015). Light bars refer to mutations at A:T sites, whereas dark bars refer to C:G sites.

spectrum may be subject to evolutionary divergence between closely related species, possibly contributing to the early differentiation of genomic and biological characteristics. Evolution of DNA repair machinery is likely to be related to the evolution of the SNM spectrum, as suggested by evidence from experiments in bacteria. For example, Dillon et al. (2017) showed that the SNM spectra in wild-type *Vibrio cholerae* and *Vibrio fischeri* strains are substantially different, but they converge in strains mutant for the DNA mismatch repair gene *mutS*. In *E. coli*, C→T mutations are the most common SNM, but in *mutL* mutants A→G mutations are the most abundant (Lee et al. 2012). Thus, it is possible that the differences between the *C. incerta* and *C. reinhardtii* SNM spectra have evolved as a consequence of a small number of substitutions in DNA repair genes, such as *mutS* homologs. Orthologs of known DNA repair genes have experienced non-synonymous substitutions between the species (up to a rate of 2.4%, supplementary table S4, Supplementary Material online), and it is possible that some of these amino acid changes may contribute to SNM divergence in *Chlamydomonas*. However, confirming the involvement of DNA repair machinery divergence in shaping the SNM spectra would require experimental validation, making use of strains mutant for the DNA repair machinery.

The mutation rate and its spectrum are variable in nature and probably respond to the same evolutionary forces, including genetic drift and selection, as regular quantitative traits (Lynch 2010). In particular, the evolution of mutational properties is likely to be driven by the evolution of the DNA repair machinery. Understanding the genetic architecture of the mutation rate and the causes of its variation both at the population and the genomic level is a fundamental problem in evolutionary biology. Here, we have studied the determinants of the mutation rate in *Chlamydomonas*, and characterize differences in the SNM spectra of *C. incerta* and *C. reinhardtii*. The two species' genomes are highly syntenic, but show an average synonymous divergence of ~34% (Craig et al. 2021), which means they are on a similar scale of divergence as humans and rodents (Lindblad-Toh et al. 2011). We show that the mutation rate and its associated genomic factors have been maintained in *Chlamydomonas*, whereas the SNM spectra have substantially diverged, likely contributing to the appearance of genomic differences between species. Thus, our results contribute to the understanding of the evolution of mutational properties in closely related species. More work is needed, however, focusing on the evolution of mutational properties in the context of the evolution of the DNA repair machinery. Future research shall also benefit from using additional sources of de novo mutations, such as structural mutations, using recent long-read technology.

Materials and Methods

MA Experiment and WGS

MA lines were initiated from the *C. incerta* strain SAG 7.73, which was obtained from the SAG culture center (Germany). During the MA experiment, cell suspensions were spread out

on Bold's agar plates, and lines were bottlenecked regularly by picking single colonies at random and transferring them from one plate to another at intervals of 3–5 days for an average of 74 transfers. The effective population size was therefore expected to be low, reducing the effectiveness of natural selection. In order to calculate the number of generations that occurred during the MA experiment, the generation times of colonies growing over 3-, 4-, and 5-day periods were determined for two replicates of 14 of the MA line endpoints. Since the generation time is expected to increase as mutations accumulate, this procedure is likely to underestimate the generation time over the whole course of MA. However, the mean generation time of the MA lines was close to that measured for five replicates of the ancestor (Wilcoxon rank-sum test, $W = 1,119$, $P = 0.71$, supplementary fig. S12, Supplementary Material online), and therefore we expect this bias to be small. Liquid cultures were plated and colonies were allowed to grow for 3, 4, or 5 days. The total number of colonies (N_0) on each plate was counted, and then the plates were flooded with medium in order to suspend the cells. Cell suspensions were diluted and replated, incubated for 5 days, and newly growing colonies counted. Then, the total colony forming units (N_t) in the undiluted cell suspensions were calculated. The number of generations (t) was computed as $t = (\log N_t - \log N_0) / \log 2$. Growth rate per day for each transfer period was then used to compute the total number of generations over the course of the entire MA experiment. For MA lines that did not have direct measurement of growth rate, the average rate of the other MA lines was used. Additional details on the maintenance of the MA lines can be found in Morgun et al. (2014).

A total of 27 lines were generated from the MA experiment. DNA was extracted from frozen cells using the phenol-chloroform protocol as described in Ness et al. (2012). Sequencing followed Ness et al. (2015). Briefly, whole-genome resequencing was performed at ~25X coverage of 100 bp paired-end reads on the Illumina HiSeq 2500 platform by BGI (China), using modified PCR conditions (Aird et al. 2011) to accommodate the high GC content (66%) of the *C. incerta* genome.

Alignment and Variant Calling

We used the ~129 Mb *C. incerta* reference genome of the ancestral strain (SAG 7.73), which is a highly contiguous (contig-level N50 ~ 1.6 Mb) assembly based on Pacific Biosciences sequencing (Craig et al. 2021). Plastid and mitochondrial genomes were included in the alignment. The short reads were aligned to the reference genome using BWA-MEM v.0.7.17 (Li and Durbin 2009). The resultant BAM files were sorted with SAMtools v.1.9 (Li et al. 2009; Li 2011) and further processed using Picard tools v.2.21.1 (Broad Institute 2019). The tool MarkDuplicates was used to tag duplicate reads, and AddOrReplaceReadGroups was used to update the files' metadata. After processing, the average coverage measured with SAMtools was ~21X.

To directly compare the mutations inferred in *C. incerta* with those from *C. reinhardtii*, contigs with no evidence of synteny between the two species were excluded. These

represented approximately 9% of the *C. incerta* reference genome (Craig et al. 2021). Variant calling was first done using GATK v4.1.4.0 (McKenna et al. 2010; Van der Auwera et al. 2013), but variants found by Freebayes 1.3.2 (Garrison and Marth 2012) were also included. In GATK, the HaplotypeCaller command was used to generate a genome variant call format (GVCF) file for each line separately using the optional parameter for a haploid genome (-ploidy 1), and every genomic site was called, including nonvariant ones (-ERC BP_RESOLUTION). GVCF files were merged using GATK CombineGVCFs. The combined variant call format (VCF) file was indexed with IndexFeatureFile, and variants were called with GenotypeGVCFs (-ploidy, -all-sites). Freebayes was run with ploidy set to 1 (-p 1), and invariant sites (-report-monomorphic) were included in the output. In addition to GATK and Freebayes, Pindel v0.2.5b9 (Ye et al. 2009) was used to call INDELS, including large insertions, and inversions and tandem repeats. Pindel was configured using insert sizes of 250, 500, and 2,500 bp.

Mutations and Callable Sites

Candidate mutations were detected following a similar procedure as described by Ness et al. (2015), making use of a custom Cython script and the cyvcf2 0.11.5 package (Pedersen and Quinlan 2017) (see Data Availability). We summarize the main steps for detecting SNMs and structural variants below. Callable sites were required to have an MQ of at least 50, a threshold chosen on the basis of the distribution of its observed values. Similarly, a minimum value of the Phred quality score (QUAL) of 100 was set for all nuclear contig sites, but a lower threshold of 70 was set for the organelle genomes, based on their distribution of QUAL values. The distribution of contig read depths (DP) did not show clearly distinct peaks, and thus a minimum combined DP for all lines was set to 167 for the nuclear contigs, which is one standard deviation below their mean DP value. Similarly, this threshold was set to a combined DP of 26,000 for the plastid and 20,000 for the mitochondrial genome. In addition, only callable sites marked as haploid and with Phred-scaled genotype quality (GQ) equal to its maximum value of 99 in at least three lines were called. In order to compare the frequency of callable sites between MA lines, a line-specific callable rate was also estimated from a separate set of alignment files (one per line) containing synthetic mutations at known sites (Farlow et al. 2015; Keightley et al. 2015). These mutations were distributed every 27 kb, and called as regular mutations (see below), so a callable rate was estimated from the number of mutations retrieved relative to the total number introduced.

Since mutations are assumed to be extremely rare events, variant sites were only considered as mutation candidates when the alternate allele was present in only one line. Candidate mutation sites were further required to have a GQ of 99, be biallelic, have a minimum depth of six reads, and with no more than one in six reads containing the reference allele. The alignment context was also taken into account, by removing candidate mutations that occurred no farther than 10 base pairs from a site where at least 14 lines

contained more than 1 out of 6 reads with alternate alleles. Mutation candidates were further validated by visualization of snapshots generated using the batch mode of the Integrative Genomics Viewer, IGV (Robinson et al. 2011; Thorvaldsdóttir et al. 2013).

For comparative analyses, a data set of callable sites from *C. reinhardtii* was also used. This corresponded to the one used by Ness et al. (2015), including genome-wide information for reference and mutated alleles and their position, and included more than 5,000 SNMs from approximately 95 Mb of callable positions (combined over ancestral strains). Callable and mutated sites were defined as above. Other parameters related to sequence context and annotation of genomic features were recalculated using the same tools and methods as used for *C. incerta* (see below).

The Mutation Rate and Its Distribution

The mutation rate (μ) was estimated as $\mu = N\mu / (Nc \times Nlines \times t)$, where $N\mu$ is the number of mutations found, Nc the length of callable genome (in base pairs), $Nlines$ is the number of MA lines, and t the number of generations. When μ was estimated for individual lines, contigs or genomic features, the numerator and denominator were adjusted accordingly. The distribution of SNMs among lines was fitted to parametric distributions using the R package fitdistrplus 1.1-1 (Delignette-Muller and Dutang 2015). Normal, log-normal, gamma, Poisson, and exponential distributions were tested, and the one with the lowest value of the Akaike information criterion was selected. The Poisson distribution was chosen as a null hypothesis for the expected variance of the number of mutations among lines (Charlesworth 2012). This variation was also evaluated by simulating the MA experiment using SLiM 3 (Haller and Messer 2019). The callable genome length and mutation rate of *C. incerta* were calculated assuming a haploid, neutral model with no recombination. A total of 1,000 replicates were simulated for the estimated average number of generations of the experiment (788 generations).

MA line 2 was excluded when computing mutation rates, as it was found to share an unusually high number of alternate alleles with MA line 3 (75%, compared with the mean 1% shared with the remaining lines), suggesting possible contamination during the maintenance or sequencing of the lines. A neighbor-joining dendrogram generated using vc-kit phylo (Cook and Andersen 2017) illustrates this genetic relationship (supplementary fig. S13, Supplementary Material online). Two SNMs found uniquely in MA line 2 were included for all remaining analyses.

Counts of SNMs causing synonymous and nonsynonymous changes within coding sequences were used to estimate the Ka/Ks ratio (Hurst 2002) and to test for the possible existence of selection during the MA experiment. Since only one or a few mutations are expected within each coding sequence, all coding sequences with at least one SNM were concatenated. The processing of the files was done using the bedtools v2.29.2 intersect and getfasta utilities (Quinlan and Hall 2010), and the Ka/Ks ratio was estimated with KaKs_Calculator 2.0 (Wang et al. 2010). Amino acid mutation effect types were also classified as low (synonymous),

moderate (missense), or high (stop lost/gain) using SnpEff 4.3t (Cingolani et al. 2012), and their functional annotation was obtained both with biomaRt 2.44.0 (Durinck et al. 2005, 2009) using the Phytozome database (Goodstein et al. 2012), and BLAST (Boratyn et al. 2013) using the standard nucleotide database.

The IMD was used as an indication of the level of clustering of mutations. This parameter was estimated by measuring the distance (in base pairs) between SNMs belonging to the same contig, irrespective of the MA line carrying it. To determine whether the observed IMD was distributed differently from a random expectation, the IMD was also computed for the same number of mutations, located at random callable sites. This process was repeated for 1,000 iterations. Mutation clustering was also measured by counting the number of SNMs in genomic sections of 1, 10, 100, and 500 kb, and numbers were compared with the expected number of mutations in these sections, following the same procedure as described above.

Genomic Context

The genomic context of the mutations was explored in two ways, first by means of parameters that collapse genomic information into simple parameters and second by identifying sequence motifs and features.

Local genomic information was first summarized by measuring GC content, but other parameters such as Shannon Entropy (E) and linguistic complexity (L) were also calculated (Schneider 2000; Vinga 2014). These parameters were computed in windows of 2, 10, 100, and 1,000 bp extending upstream and downstream from a mutation site, including the mutation site. Genomic intervals and GC content were measured using the bedtools makewindows and nuc utilities. E and L were measured using NeSsie (Berselli et al. 2018). Variability in GC, E , and L , measured as the standard deviations of these parameters within windows overlapping a given position were also computed and then averaged for windows of 5, 21, 201, and 2,001 bp (obtained as above).

Regarding sequence motifs, we considered the trinucleotide context of every site. For a given site, three different contexts were obtained: upstream, downstream, and surrounding contexts, all based on trinucleotide sequences, including the site whose context was extracted. For example, taking the sequence AGATA, where the middle and underlined A is the reference site whose context is of interest, three different contexts are obtained: AGA (upstream context), GAT (surrounding context), and ATA (downstream context). In order to reduce the total number of possible context sequences, trinucleotide context sequences were edited so that the reference site was always either A or C. For example, given the sequence TATCT, the surrounding context of the underlined T, ATC, would be converted to GAT using the reverse complement. The upstream and downstream contexts were also deduced using the reverse complementary sequence, but including a swap between these contexts, for example the upstream context TAT would be considered as downstream context using the reverse complement: ATA.

In addition, annotations for different genomic features were used to provide genomic context. Gene annotations

for *C. incerta* were obtained from Craig et al. (2021) and the *C. reinhardtii* v5.6 annotation (Merchant et al. 2007) was obtained from Phytozome. We limited the analyses to coding sequences, introns, UTRs, and intergenic regions. When features overlapped, coding sequences took precedence over other site classes, followed by 5' and 3'UTRs, and finally introns. Repetitive sequences were identified by providing RepeatMasker v4.0.3 (Smit et al. 2013–2015) with a custom repeat library containing manually curated transposable elements (TEs) from *C. reinhardtii*, *C. incerta*, *Chlamydomonas schloesseri*, *Edaphochlamys debaryana*, and *Volvox carteri* (see Craig et al. [2021]). Repeats were classified as TEs, low complexity regions (defined as single nucleotide repeats), and microsatellites. TEs were further divided by class (i.e., class I and II) and order/superfamily following Wicker et al. (2007). Distance (in base pairs) was also calculated from each genomic feature to the closest mutation site.

Modeling Mutability

Generalized linear models allow the estimation of the effects of genomic factors on mutability based on data sets containing sites classified either as mutated or nonmutated. Here, we used a regularized regression approach that introduced a penalization factor to constrain the magnitude of the coefficient estimates, which reduces the correlation between the variables used as predictors in the model. A database was first built for all callable genomic positions. Only presence/absence of nuclear SNMs was evaluated as the response variable in the model. As predictor variables, we included variables related to genomic context, including GC, E , and L content in different genomic window sizes, trinucleotide motifs, genomic features, and distances, as detailed above. Positions where any of these variables were missing were removed from the data set, along with near-zero variance predictors. All factors were converted into dummy numerical variables, centered, and scaled so that a unit change corresponded to a change of one standard deviation. Given the large number of parameters, and the likely correlation among them, a logistic regression model was run via penalized maximum likelihood making use of elastic net regularization using the glmnet R package 4.0-2 (Friedman et al. 2010). This allows first for optimizing penalization factors using a 10-fold cross-validation method, where 75% of data are used for training and 25% for validation. The generalized model was then run in a similar fashion as described in Ness et al. (2015), using a training set composed of all the detected mutations and 10^5 random nonmutated sites. The performance of the model was tested against a larger data set containing all mutated sites and 10^6 nonmutated random callable positions. In addition, fits and predictions were run ten times, using independent training and test sets, including model prediction across species (see below). Confidence intervals were built for the predicted response and regression coefficient estimates based on their variation over replicates.

A second set of models was fitted in order to model the occurrence of the six different types of SNM ($A \rightarrow C$, $A \rightarrow G$, $A \rightarrow T$, $C \rightarrow A$, $C \rightarrow G$, and $C \rightarrow T$). The data set used here was obtained from the previous one with the following changes.

First, only mutation sites were considered in the data set. Variables related to GC content were excluded, and those related to trinucleotide context were edited so that the mutated site was excluded from them. The data set was split in two, one for training containing 75% of the observations, and an independent data set with the remaining 25% for testing. Given the multinomial nature of the response variable, we trained different classification and regression models using the caret package v6.0-86 (Kuhn 2020), including penalized regression as described above (glmnet), but also including machine learning methods based on regularized random forests and neural networks (nnet). Since the different SNM types were unevenly represented in our data set (e.g., C→T mutations were largely over-represented), different strategies to correct data imbalance were considered, including random down-sampling and up-sampling and the synthetic minority over-sampling technique, SMOTE (López et al. 2013). In addition, a model was run where the response variable only had two levels, C→T mutations (the most abundant SNM type), and other SNM types grouped together. The same training set was used in each case, by setting the same seed for sampling random numbers. An R environment (version 3.6.1, R Core Team 2020) was used for all statistical analyses, including prediction of mutation properties. Predictive models were also run on a genomic data set with *C. reinhardtii* data from Ness et al. (2015), and cross predictions were performed using one species' data set for training, and the other's for testing (e.g., by fitting the model with *C. incerta* data and making predictions on *C. reinhardtii* data).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgements

We thank Rodrigo Bacigalupe for useful comments and suggestions. Analyses performed here made use of the high-performance computing (HPC) resources at the Edinburgh Compute and Data Facility (ECDF). This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 694212).

Data Availability

Raw FASTQ files for the *C. incerta* MA lines are publicly available through the NCBI Sequence Read Archive (SRA), with BioProject ID PRJNA687646. Mutation candidates were called from VCF files using the Cython script vcf2mut.pyx, available at GitLab (<https://gitlab.com/elcortegano/vcf2mut>, last accessed May 2021).

References

Aggarwala V, Voight BF. 2016. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat Genet*. 48(4):349–355.

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*. 12(2):R18.

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*. 3(1):246–259.

Bacolla A, Wells RD. 2004. Non-B DNA conformations, genomic rearrangements, and human disease. *J Biol Chem*. 279(46):47411–47414.

Belfield EJ, Ding ZJ, Jamieson FJC, Visscher AM, Zheng SJ, Mithani A, Harberd NP. 2018. DNA mismatch repair preferentially protects genes from mutation. *Genome Res*. 28(1):66–74.

Berselli M, Lavezzo E, Toppo S. 2018. NeSSie: a tool for the identification of approximate DNA sequence symmetries. *Bioinformatics*. 34(14):2503–2505.

Berselli M, Lavezzo E, Toppo S. 2020. QPARSE: searching for long-looped or multimeric G-quadruplexes potentially distinctive and druggable. *Bioinformatics*. 36(2):393–399.

Bjedov I, Tenaillon O, Gérard B, Souza V, Denamur E, Radman M, Taddei F, Matic I. 2003. Stress-induced mutagenesis in bacteria. *Science*. 300(5624):1404–1409.

Böndel KB, Kraemer SA, Samuels T, McClean D, Lachapelle J, Ness RW, Colegrave N, Keightley PD. 2019. Inferring the distribution of fitness effects of spontaneous mutations in *Chlamydomonas reinhardtii*. *PLoS Biol*. 17(6):e3000192.

Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezuk Y, et al. 2013. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res*. 41(W1):W29–W33.

Broad Institute. 2019. Picard Toolkit. Available from: <https://broadinstitute.github.io/picard/>. Accessed May 2021.

Charlesworth B. 2012. The effects of deleterious mutations on evolution at linked sites. *Genetics*. 190(1):5–22.

Charlesworth B. 2018. Mutational load, inbreeding depression and heterosis in subdivided populations. *Mol Ecol*. 27(24):4991–5003.

Chebib J, Jackson BC, López-Cortegano E, Tautz D, Keightley PD. 2021. Inbred lab mice are not isogenic: genetic variation within inbred strains used to infer the mutation rate per nucleotide site. *Heredity*. 126(1):107–116.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w¹¹¹⁸; iso-2; iso-3. *Fly*. 6(2):80–92.

Cook DE, Andersen EC. 2017. VCF-kit: assorted utilities for the variant call format. *Bioinformatics*. 33(10):1581–1582.

Craig RJ, Hasan AR, Ness RW, Keightley PD. 2021. Comparative genomics of *Chlamydomonas*. *Plant Cell*. doi: 10.1093/plcell/koab026

de Visser JAGM. 2002. The fate of microbial mutators. *Microbiology*. 148(5):1247–1252.

Delignette-Muller ML, Dutang C. 2015. fitdistrplus: an R package for fitting distributions. *J Stat Softw*. 64(4):1–34.

Denver DR, Wilhelm LJ, Howe DK, Gafner K, Dolan PC, Baer CF. 2012. Variation in base-substitution mutation in experimental and natural lineages of *Caenorhabditis* nematodes. *Genome Biol Evol*. 4(4):513–522.

Dillon MM, Sung W, Sebra R, Lynch M, Cooper VS. 2017. Genome-wide biases in the rate and molecular spectrum of spontaneous mutations in *Vibrio cholerae* and *Vibrio fischeri*. *Mol Biol Evol*. 34(1):93–109.

Dumont BL. 2019. Significant strain variation in the mutation spectra of inbred laboratory mice. *Mol Biol Evol*. 36(5):865–874.

Durinck S, Moreau Y, Kasprzyk A, Davis S, de Moor B, Brazma A, Huber W. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*. 21(16):3439–3440.

Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*. 4(8):1184–1191.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet*. 8(8):610–618.

Farlow A, Long H, Arnoux S, Sung W, Doak TG, Nordborg M, Lynch M. 2015. The spontaneous mutation rate in the fission yeast *Schizosaccharomyces pombe*. *Genetics*. 201(2):737–744.

- Flynn JM, Chain FJJ, Schoen DJ, Cristescu ME. 2017. Spontaneous mutation accumulation in *Daphnia pulex* in selection-free vs. competitive environments. *Mol Biol Evol.* 34(1):160–173.
- Foster PL, Lee H, Popodi E, Townes JP, Tang H. 2015. Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. *Proc Natl Acad Sci U S A.* 112(44):E5990–E5999.
- Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J Stat Soft.* 33(1):1–22.
- Frigola J, Sabarinathan R, Mularoni L, Muiños F, Gonzalez-Perez A, López-Bigas N. 2017. Reduced mutation rate in exons due to differential mismatch repair. *Nat Genet.* 49(12):1684–1692.
- Gajarský M, Živković ML, Stadlbauer P, Pagano B, Fiala R, Amato J, Tomáška L, Šponer J, Plavec J, Trantírek L. 2017. Structure of a stable G-hairpin. *J Am Chem Soc.* 10(15):3591–3594.
- Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv: 1207.3907 [q-bio.GN].
- Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40(D1):D1178–D1186.
- Haller BC, Messer PW. 2019. Slim 3: forward genetic simulations beyond the Wright-Fisher model. *Mol Biol Evol.* 36(3):632–637.
- Hamilton WL, Claessens A, Otto TD, Kekre M, Fairhurst RM, Rayner JC, Kwiatkowski D. 2017. Extreme mutation bias and high AT content in *Plasmodium falciparum*. *Nucleic Acids Res.* 45(4):1889–1901.
- Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet.* 6(9):e1001115.
- Ho EKH, Macrae F, Latta LL, McIlroy P, Ebert D, Fields PD, Benner MJ, Schaack S. 2020. High and highly variable spontaneous mutation rates in *Daphnia*. *Mol Biol Evol.* 37(11):3258–3266.
- Hurst LD. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18(9):486–487.
- Katju V, Bergthorsson U. 2019. 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(1):136–165.
- Keightley PD. 2012. Rates and fitness consequences of new mutations in humans. *Genetics.* 190(2):295–304.
- Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, Davey JW, Jiggins CD. 2015. Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Mol Biol Evol.* 32(1):239–243.
- Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res.* 19(7):1195–1201.
- Kessler MD, Loesch DP, Perry JA, Heard-Costa NL, Taliun D, Cade BE, Wang H, Daya M, Ziniti J, Datta S, et al. 2020. National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Consortium. De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population. *Proc Natl Acad Sci U S A.* 117(5):2560–2569.
- Kraemer SA, Böndel KB, Ness RW, Keightley PD, Colegrave N. 2017. Fitness change in relation to mutation number in spontaneous mutation accumulation lines of *Chlamydomonas reinhardtii*. *Evolution* 71(12):2918–2929.
- Krasovec M, Eyre-Walker A, Sanchez-Ferandin S, Piganeau G. 2017. Spontaneous mutation rate in the smallest photosynthetic eukaryotes. *Mol Biol Evol.* 34(7):1770–1779.
- Krasovec M, Sanchez-Brosseau S, Grimsley N, Piganeau G. 2018. Spontaneous mutation rate as a source of diversity for improving desirable traits in cultured microalgae. *Algal Res.* 35:85–90.
- Krasovec M, Sanchez-Brosseau S, Piganeau G. 2019. First estimation of the spontaneous mutation rate in diatoms. *Genome Biol Evol.* 11(7):1829–1837.
- Kucukyildirim S, Behringer M, Williams EM, Doak TG, Lynch M. 2020. Estimation of the genome-wide mutation rate and spectrum in the archaeal species *Haloferax volcanii*. *Genetics* 215(4):1107–1116.
- Kuhn M. 2020. caret: classification and regression training. R package version 6.0-85. Available from: <https://CRAN.R-project.org/package=caret>. Accessed May 2021.
- Lee H, Popodi E, Tang H, Foster PL. 2012. 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.* 109(41):E2774–E2783.
- Leffak M, Gadgil R, Barthelemy J, Lewis T. 2017. Replication stalling and DNA microsatellite instability. *Biophys Chem.* 225:38–48.
- Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21):2987–2993.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map (SAM) format and SAMtools. *Bioinformatics* 25(16):2078–2079.
- Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al; Broad Institute Sequencing Platform and Whole Genome Assembly Team. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370):476–482.
- Ling G, Miller D, Nielsen R, Stern A. 2020. A Bayesian framework for inferring the influence of sequence context on point mutations. *Mol Biol Evol.* 37(3):893–903.
- Liu H, Zhang J. 2019. Yeast spontaneous mutation rate and spectrum vary with environment. *Curr Biol.* 29(10):1584–1591.
- Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C, Strauss C, Stone C, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. *Nat Ecol Evol.* 2(2):237–240.
- Long H, Winter DJ, Chang AYC, Sung W, Wu SH, Balboa M, Azevedo RBR, Cartwright RA, Lynch M, Zufall RA. 2015. Low base-substitution mutation rate in the germline genome of the ciliate *Tetrahymena thermophila*. *Genome Biol Evol.* 8(12):3629–3639.
- López V, Fernández A, García S, Palade V, Herrera F. 2013. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Inf Sci.* 250:113–141.
- Lynch M. 2010. Evolution of the mutation rate. *Trends Genet.* 26(8):345–352.
- Lynch M. 2011. The lower bound to the evolution of mutation rates. *Genome Biol Evol.* 3:1107–1118.
- Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet.* 17(11):704–714.
- Lynch M, Blanchard J, Houle D, Kibota T, Schultz S, Vassilieva L, Willis J. 1999. Perspective: spontaneous deleterious mutation. *Evolution* 53(3):645–663.
- Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302(5649):1401–1404.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9):1297–1303.
- McKinney JA, Wang G, Mukherjee A, Christensen L, Subramanian SHS, Zhao J, Vasquez KM. 2020. Distinct DNA repair pathways cause genomic instability at alternative DNA structures. *Nat Commun.* 11:236.
- Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, et al. 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 318 (5848):245–250.
- Mirkin SM. 2008. Discovery of alternative DNA structures: a heroic decade (1979–1989). *Front Biosci.* 13(13):1064–1071.
- Morgan AD, Ness RW, Keightley PD, Colegrave N. 2014. Spontaneous mutation accumulation in multiple strains of the green alga, *Chlamydomonas reinhardtii*. *Evolution* 68(9):2589–2602.

- Mulder HA, Lee SH, Clark S, Hayes BJ, van der Werf JHJ. 2019. The impact of genomic and traditional selection on the contribution of mutational variance to long-term selection response and genetic variance. *Genetics* 213(2):361–378.
- Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, Kovar CL, Lewis LR, Morgan MB, Newsham IF, et al. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487(7407):330–337.
- Nakada T, Shinkawa H, Ito T, Tomita M. 2010. Recharacterization of *Chlamydomonas reinhardtii* and its relatives with new isolates from Japan. *J Plant Res*. 123(1):67–78.
- Ness RW, Kraemer SA, Colegrave N, Keightley PD. 2016. Direct estimate of the spontaneous mutation rate uncovers effects of drift and recombination on the *Chlamydomonas reinhardtii* plastid genome. *Mol Biol Evol*. 33(3):800–808.
- Ness RW, Morgan AD, Colegrave N, Keightley PD. 2012. Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii*. *Genetics* 192(4):1447–1454.
- Ness RW, Morgan AD, Vasanthakrishnan RB, Colegrave N, Keightley PD. 2015. Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. *Genome Res*. 25(11):1739–1749.
- Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327(5961):92–94.
- Pedersen BS, Quinlan AR. 2017. cyvcf2: fast, flexible variant analysis with Python. *Bioinformatics* 33(12):1867–1869.
- Popescu CE, Borza T, Bielawski JP, Lee RW. 2006. Evolutionary rates and expression level in *Chlamydomonas*. *Genetics* 172(3):1567–1576.
- Pröschold T, Harris EH, Coleman AW. 2005. Portrait of a species: *Chlamydomonas reinhardtii*. *Genetics* 170(4):1601–1610.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
- R Core Team. 2020. R: a language and environment for statistical computing. R foundation for statistical computing. Available from: <https://www.R-project.org>. Accessed May 2021.
- Refsland EW, Harris RS. 2013. The APOBEC3 family of retroelement restriction factors. *Curr Top Microbiol Immunol*. 371:1–27.
- Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, Wapinski I, Roy S, Lin MF, Heiman DI, et al. 2011. Comparative functional genomics of the fission yeast. *Science* 332(6032):930–936.
- Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, et al. 2013. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet*. 45(9):970–976.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol*. 29(1):24–26.
- Sanjuán R, Domingo-Calap P. 2016. Mechanisms of viral mutation. *Cell Mol Life Sci*. 73(23):4433–4448.
- Sassa A, Kanamaru Y, Kamoshita N, Honma M, Yasui M. 2016. Mutagenic consequences of cytosine alterations site-specifically embedded in the human genome. *Genes Environ*. 38(1):17.
- Schneider TD. 2000. Evolution of biological information. *Nucleic Acids Res*. 28(14):2794–2799.
- Smit AFA, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. Available from: <https://www.repeatmasker.org>. Accessed May 2021.
- Smith DR, Keeling PJ. 2015. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proc Natl Acad Sci U S A*. 112(33):10177–10184.
- Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. *Nat Genet*. 41(4):393–395.
- Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, et al. 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol*. 1(2):e45.
- Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci U S A*. 109(45):18488–18492.
- Sung W, Tucker AE, Doak TG, Choi E, Thomas WK, Lynch M. 2012. Extraordinary genome stability in the ciliate *Paramecium tetraurelia*. *Proc Natl Acad Sci U S A*. 109(47):19339–19344.
- Terekhanova NV, Seplyarskiy VB, Soldatov RA, Bazykin GA. 2017. Evolution of local mutation rate and its determinants. *Mol Biol Evol*. 34(5):1100–1109.
- Thanos D, Li W, Provata A. 2018. Entropic fluctuations in DNA sequences. *Physica A*. 493:444–457.
- Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 14(2):178–192.
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 43(1):11.10.1–11.10.33.
- Vinga S. 2014. Information theory applications for biological sequence analysis. *Brief Bioinform*. 15(3):376–389.
- Walsh B. 2004. Population- and quantitative-genetic models of selection limits. In: Janick, editor. Plant breeding reviews: part 1: long-term selection: maize, Vol. 24. Hoboken (NJ): John Wiley & Sons, Inc. p. 177–225.
- Walsh B, Lynch M. 2018. Evolution and selection of quantitative traits. Oxford: Oxford University Press.
- Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* 8(1):77–80.
- Wells RD. 1988. Unusual DNA structures. *J Biol Chem*. 263(3):1095–1098.
- Weng ML, Becker C, Hildebrandt J, Neumann M, Rutter M, Shaw RG, Weigel D, Fenster CB. 2019. Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. *Genetics* 211(2):703–714.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capi P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 8(12):973–982.
- Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25(21):2865–2871.
- Zhang L, Vijg J. 2018. Somatic mutagenesis in mammals and its implications for human disease and aging. *Annu Rev Genet*. 52(1):397–419.
- Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. *Proc Natl Acad Sci U S A*. 111(22):E2310–E2318.