Rates and spectra of de novo structural mutations in *Chlamydomonas reinhardtii*

Eugenio López-Cortegano,[1,5] Rory J. Craig,[1,2,5] Jobran Chebib,[1] Eniolaye J. Balogun,[3,4] and Peter D. Keightley[1]

[1]*Institute of Ecology and Evolution, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom;* [2]*California Institute for Quantitative Biosciences, UC Berkeley, Berkeley, California 94720, USA;* [3]*Department of Ecology and Evolutionary Biology, University of Toronto, Ontario ON M5S 3B2, Canada;* [4]*Department of Biology, University of Toronto Mississauga, Mississauga ON L5L 1C6, Canada*

Genetic variation originates from several types of spontaneous mutation, including single-nucleotide substitutions, short insertions and deletions (indels), and larger structural changes. Structural mutations (SMs) drive genome evolution and are thought to play major roles in evolutionary adaptation, speciation, and genetic disease, including cancers. Sequencing of mutation accumulation (MA) lines has provided estimates of rates and spectra of single-nucleotide and indel mutations in many species, yet the rate of new SMs is largely unknown. Here, we use long-read sequencing to determine the full mutation spectrum in MA lines derived from two strains (CC-1952 and CC-2931) of the green alga *Chlamydomonas reinhardtii*. The SM rate is highly variable between strains and between MA lines, and SMs represent a substantial proportion of all mutations in both strains (CC-1952 6%; CC-2931 12%). The SM spectra differ considerably between the two strains, with almost all inversions and translocations occurring in CC-2931 MA lines. This variation is associated with heterogeneity in the number and type of active transposable elements (TEs), which comprise major proportions of SMs in both strains (CC-1952 22%; CC-2931 38%). In CC-2931, a Crypton and a previously undescribed type of DNA element have caused 71% of chromosomal rearrangements, whereas in CC-1952, a *Dualen* LINE is associated with 87% of duplications. Other SMs, notably large duplications in CC-2931, are likely products of various double-strand break repair pathways. Our results show that diverse types of SMs occur at substantial rates, and support prominent roles for SMs and TEs in evolution.

[Supplemental material is available for this article.]

Since the development of the modern synthesis in evolutionary biology, the existence of chromosomal changes visualized in cytogenetic studies led to the hypothesis that structural mutations (SMs) could be an important source of variation, leading to evolutionary change by natural selection (Dobzhansky and Epling 1948; McClintock 1950; Ohno 1970). All genetic variation has its origin in new mutations, and efforts to estimate the rate of mutations started early in the twentieth century by analysis of mutation accumulation (MA) experiments, in which spontaneous mutations are allowed to accumulate in lines of small effective population size where natural selection is ineffective (Muller 1928; Bateman 1959; Mukai 1964). The advent of whole-genome sequencing technology led to the possibility of directly estimating the rate, spectra, and genomic distribution of mutations by sequencing MA lines and, later, by the sequencing of parents and their offspring. Although studies of these kinds have been performed in many species (Halligan and Keightley 2009; Yoder and Tiley 2021), the short-read sequencing technology that has been applied reliably detects only single-nucleotide mutations (SNMs) and short insertions and deletions (indels), and little is known about the rates at which SMs occur de novo.

SMs include larger indels (often defined as those >50 bp), duplications, transposable element (TE) insertions and excisions, and chromosomal rearrangements, such as inversions and translocations. Such large structural changes are expected to have larger fitness effects than SNMs and indels, and the structural variation that arises from SMs has been implicated in many evolutionary phenomena. For example, duplications provide the raw material for gene family evolution via processes that include neo- and subfunctionalization (Kuzmin et al. 2022). Inversions may result in recombination suppression, and the subsequent evolutionary divergence of ancestral and inverted haplotypes has been implicated in local adaptation, speciation, and sex chromosome evolution (Kirkpatrick and Barton 2006; Kirkpatrick 2010). Inversions may also give rise to "supergenes," which preserve the linkage of multiple coadapted loci and can underlie complex phenotypes (Joron et al. 2011; Küpper et al. 2016). Translocations and other major rearrangements can similarly suppress recombination and may directly cause reproductive isolation (Faria and Navarro 2010; Potter et al. 2017), and deletions have also been linked to genomic differentiation during speciation (Zhang et al. 2022).

Appreciation for the diverse evolutionary roles of TEs is ever increasing. TE insertions in functional sequences can drive rapid phenotypic adaptation (van't Hof et al. 2016), and TEs can contribute substantially to regulatory sequences over evolutionary timescales (Chuong et al. 2017; Zhao et al. 2018). TE activity has been linked to diverse phenomena, including genome size evolution (Gregory 2005), genomic "turnover" (gain and loss of DNA) (Kapusta et al. 2017), and speciation (Ricci et al. 2018; Tusso et al. 2022). TEs may also mediate large deletions, duplications, and chromosomal rearrangements (Gray 2000), either directly as

[5]**These authors contributed equally to this work.**
Corresponding author: peter.keightley@ed.ac.uk
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.276957.122. Freely available online through the *Genome Research* Open Access option.

a by-product of their transposition machinery (e.g., as in the classic *Ac*/*Ds* system in maize) (Zhang et al. 2009) or by nonallelic homologous recombination between interspersed TE copies (Konkel and Batzer 2010).

Aside from their evolutionary importance, SMs are generally associated with deleterious effects and have been implicated in several human diseases and cancers (Inaki and Liu 2012; Weischenfeldt et al. 2013). Finally, SMs are also likely to become increasingly relevant in applied fields, such as selective breeding in agriculture, because structural variation has been associated with commercially important traits (Jayakodi et al. 2020; Song et al. 2020).

Increasing our understanding of the evolutionary importance of structural variation requires better knowledge of the rates at which the various types of SMs occur. Some studies have used short-read sequencing to estimate rates for particular SMs, often limited to deletions and duplications (Katju and Bergthorsson 2013; Konrad et al. 2018; Belyeu et al. 2021; Ho and Schaack 2021; Villalba de la Peña et al. 2022) but also to TE insertions and excisions (Adrion et al. 2017; Ho et al. 2021). Although results were complex and variable, SMs have been observed to occur at nonnegligible rates relative to other mutations, and heterogeneity in SM rates among genotypes was often present. Recently, advances in long-read sequencing technology have led to substantial improvements in the ability to assemble near-complete eukaryotic genomes and have stimulated the development of bioinformatic tools for the discovery of structural variation (Rhoads and Au 2015; Jain et al. 2018; Mahmoud et al. 2019; Miga et al. 2020; De Coster et al. 2021). These advances now enable the study of the complete spectra of SMs in MA lines. The single-celled green alga *Chlamydomonas reinhardtii* is an excellent model for mutation research because its relatively large genome (~111 Mb) and short generation time (~2.5 generations per day) enable the rapid accumulation of large numbers of new mutations in a short time, and the species has been used to explore diverse genomic properties of mutations (Ness et al. 2012; Sung et al. 2012; López-Cortegano et al. 2021; Böndel et al. 2022). Here, we use Pacific Biosciences (PacBio) HiFi to identify SMs in replicate MA lines of two divergent strains of *C. reinhardtii* generated in a previous study (Morgan et al. 2014). We aim to characterize the full spectra of SMs in *C. reinhardtii* to compare the rates of SMs to the rates of SNMs and indels and to investigate interstrain heterogeneity in SM rates and spectra.

Results

SM detection

We performed PacBio HiFi or continuous long-read (CLR) sequencing of MA lines of the *C. reinhardtii* strains CC-1952 ($N = 4$ MA lines, all HiFi) and CC-2931 ($N = 8$, six HiFi and two CLR). These geographically distinct strains were selected on the basis of their relatively low (CC-1952, overall $\mu = 4.05 \times 10^{-10}$ per site per generation) and high (CC-2931, $\mu = 15.6 \times 10^{-10}$) SNM and indel rates in the MA lines (Ness et al. 2015), which had been maintained for approximately 1050 generations by single-cell descent (Morgan et al. 2014). To perform strain-specific detection of SMs, we first produced near-complete reference assemblies for the two ancestral strains. The 17 *C. reinhardtii* chromosomes were assembled into totals of 50 and 39 contigs with N50s of 4.25 and 3.81 Mb, for CC-1952 and CC-2931, respectively (Supplemental Fig. S1; Supplemental Table S1). We subsequently defined ~98% of each ~111 Mb ancestor genome as "callable" (i.e., sites where SMs could be called with high confidence; see Methods). This represents a substantial increase over the ~71% obtained in our previous study using short-read technology (Supplemental Fig. S2; Ness et al. 2015). We sequenced MA lines at sufficient depth of coverage to produce highly contiguous assemblies (Supplemental Fig. S1; Supplemental Table S1), enabling us to call SMs using three approaches: directly from MA line PacBio read alignments against the appropriate ancestral reference using Sniffles (Sedlazeck et al. 2018), from MA line PacBio assembly alignments against the reference using MUM&Co (O'Donnell and Fischer 2020), and from Cactus pangenomes using vg (Garrison et al. 2018; Armstrong et al. 2020). Sniffles and MUM&Co were run individually for each MA line, whereas all MA lines and the ancestor for each strain were analyzed collectively with vg from a single Cactus alignment. We subsequently collated and manually curated all variant calls using a combination of read visualization and mapping approaches (Fig. 1A).

We classified our curated data set of SMs (a total of 120 in CC-1952 and 443 in CC-2931; Supplemental Dataset S1) into eight categories: expansions and contractions of tandemly repeated sequence (e.g., in microsatellites or satellite DNA, collectively termed tandem repeat mutations [TRMs]), duplications, deletions, insertions and excisions of mobile elements, and inversions and translocations (Fig. 1B). The different callers varied substantially in their ability to identify different SM types (Fig. 1C). Only 19.2% of SMs were called by all three tools, highlighting the importance of combining approaches. vg was most successful overall, calling 79.4% of SMs and 22.4% uniquely, although it called only 20.0% of inversions and translocations. MUM&Co called only 16.7% of TRMs, although it identified 64.5% of SMs in other categories. Sniffles called 47.2% of SMs and was generally outperformed by the assembly-based approaches, although it was superior in identifying duplications (calling 62.7% of duplications and 39.0% uniquely). The assembly-based methods failed to call duplications >30 kb (i.e., longer than the reads) because they were generally collapsed and absent in the assemblies. We also called 7.3% of SMs manually from alignment files and read visualization, most of which were TE insertions located at the breakpoints of other complex SMs (e.g., inversions and translocations).

The three approaches also differed markedly in the proportions of rejected calls (i.e., variants that could not be classified as genuine SMs), and all three returned more rejected calls than confirmed SMs (reaching 3× as many for vg) (Supplemental Fig. S3). Rejected calls generally fell in three categories: calls made within regions that we had defined as uncallable, which were not considered further; MUM&Co or vg calls that received assembly support but not read support (i.e., "valid" calls introduced by assembly errors); or calls that received support from neither assemblies nor reads (unsupported variants). Considering the ratio of rejected calls to confirmed SMs in CC-2931, Sniffles performed best (1.14), followed by MUM&Co (1.81) and vg (3.42). However, these ratios varied considerably depending on the type and sequence context of the rejected call. All categories of rejected calls were associated with repetitive sequence (Supplemental Fig. S4). Uncallable regions correspond to the most repetitive regions of the genome, generally long satellite arrays, including some centromeric and subtelomeric regions (Supplemental Dataset S2). Although they contain only ~2% of sites, a substantial proportion of all variant calls were made in these regions (26.4%). Although we expect that the majority of calls in the uncallable regions were false positives, the enrichment of these regions for tandem repeats may have led to an underestimation in the rates of

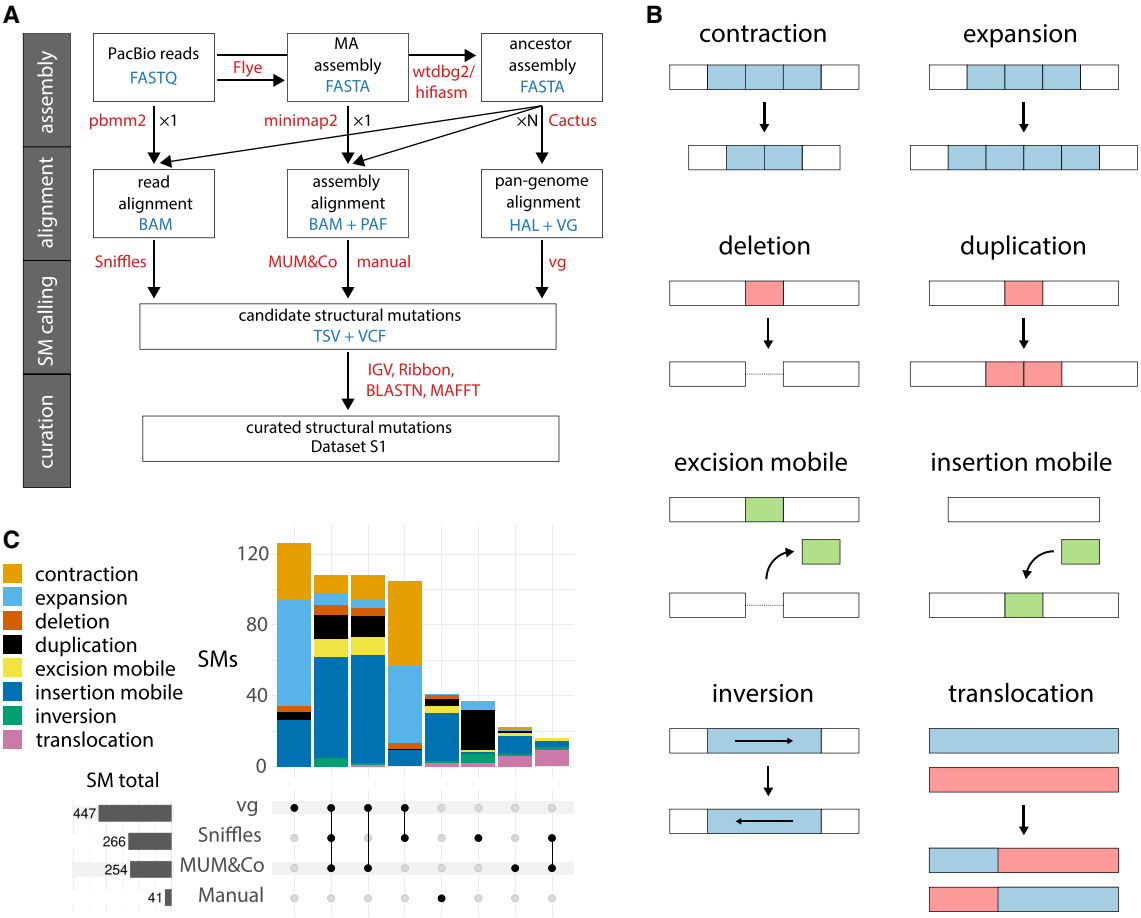

Figure 1. Structural mutation (SM) detection. (*A*) Flowchart of the SM calling pipeline. Steps are organized from *top* to *bottom* in four stages: genome assembly, mapping and alignment, SM calling, and SM curation. The software used in each step is shown in red text, and file formats are in blue text. "Manual" indicates variants that were curated directly from alignment files. "×1" indicates that each data set (reads or assemblies) was analyzed individually; "×N" indicates that all MA lines for a given strain were analyzed collectively. (*B*) Schematics illustrating the eight different types of SM called. The ancestral state is shown *above*; the mutated state, *below*. (*C*) Intersection of the number of curated SMs identified by each calling method across all MA lines (for the two strains, CC-1952 and CC-2931, combined). In vertical bars, the numbers of SMs are colored by SM type. Horizontal bars (in gray) show the total numbers of SMs called by each method.

expansions and contractions in our analysis. Considering callable regions, tandem repeats of all lengths were a major source of both assembly errors and unsupported variants. After excluding all tandem repeats, the ratio of unsupported variants to confirmed SMs improved for all callers, falling from 0.21 to 0.17 for MUM&Co and from 0.91 to 0.20 for vg, although substantial proportions of calls attributed to assembly errors remained (Supplemental Fig. S3). Consequently, although our results show that aligners and callers are capable of detecting SMs in tandemly repeated regions (expansions and contractions) (Fig. 1C), attempting to do so risks introducing many false positives in fully automated pipelines. Rejected calls are further discussed in the Supplemental Material.

Rates and spectra of SMs

The rates and spectra of SMs were markedly different between the two strains (Fig. 2A,B). CC-2931 MA lines experienced ~85% more SMs than CC-1952 MA lines, and overall SM rates were significantly different between the strains ($\mu_{SM\ (CC\text{-}1952)} = 2.58 \times 10^{-10}$ and $\mu_{SM\ (CC\text{-}2931)} = 4.30 \times 10^{-10}$ per site per generation; *W*-test, $P = 4 \times 10^{-3}$). However, the within-strain variance in μ_{SM} among MA lines was ~12% higher than between them (ANOVA test, $F = 10.8$, $P = 8 \times 10^{-3}$). In terms of the number of bases affected, CC-2931 MA lines experienced larger SMs than CC-1952 MA lines and also experienced a greater variety of SM types (Fig. 2A). We observed only three SMs >20 kb in CC-1952 MA lines, whereas almost all large chromosomal rearrangements were found in CC-2931; that is, there were 1.75 inversions (median ~243 kb) and 2.50 translocations per CC-2931 MA line (collectively 4.05×10^{-3} rearrangements per genome per generation) (Supplemental Fig. S5), compared with a single 5.2-kb inversion in the CC-1952 MA lines. Deletions were also rare in CC-1952 (less than one per MA line on average), as were mobile excisions, explained by a relatively low frequency of active *cut-and-paste* DNA transposons (see below).

Although there were clear differences between the strains, some SM properties were shared by CC-1952 and CC-2931. TRMs were the most common category of SMs <3 kb in length in both strains (representing 60.8% and 35.0% of SMs in CC-1952 and CC-2931, respectively) and occurred at similar rates in the two strains (median $\mu_{TRM\ (CC\text{-}1952)} = 1.68 \times 10^{-10}$, $\mu_{TRM\ (CC\text{-}2931)} = 1.60 \times 10^{-10}$; *W*-test, $P = 0.93$). Mobile insertions dominated the spectra of SMs >3 kb in length (representing 21.7% and 37.9% of

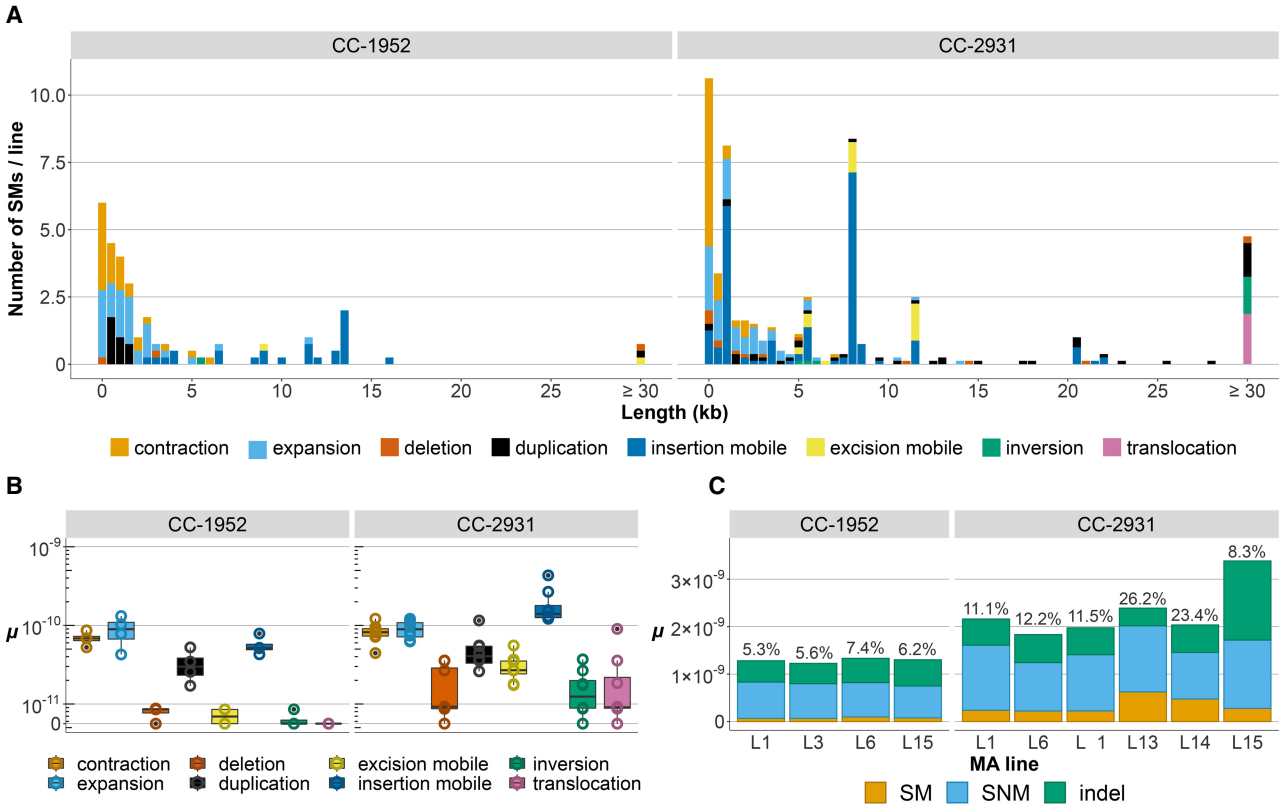

Figure 2. Spectra and rates of SMs. (*A*) Mean number of SMs by type (in colors) and length (in kilobases, rounded to 0.5 kb) per MA line. A total of five uncharacterized rearrangements of unknown length in CC-2931 MA lines are excluded. (*B*) The SM rate per site per generation (μ, on a $\log_{10}$ scale) is plotted as open points and boxplots for different types of SMs. Data points represent individual MA lines. Supplemental Figure S5 provides an analogous plot showing per genome per generation SM rates. (*C*) Mutation rates for different mutation types across the CC-1952 and CC-2931 MA lines, after excluding mutations in tandem repeat annotations. Only lines sequenced by PacBio HiFi are included. The percentage at the *top* of the bar indicates the proportion of SMs relative to all mutation types (SMs, SNMs, and indels).

all SMs in CC-1952 and CC-2931, respectively), although the median rate of mobile insertions was ~2.5× higher in CC-2931 (14.08×10^{-11} per site and generation) than in CC-1952 (5.15×10^{-11}). Duplications were also relatively common in both strains (12.5% and 9.9% of all SMs in CC-1952 and CC-2931, respectively), although duplications in CC-1952 (median length 0.9 kb) were significantly shorter than in CC-2931 (median length 12.9 kb; *W*-test, $P = 3 \times 10^{-6}$). Duplications were more frequent than deletions in both strains (*W*-test, $P < 1.5 \times 10^{-2}$) and also significantly longer in CC-2931 (median length, 12.9 kb vs. 1.7 kb, respectively; *W*-test, $P = 1.81 \times 10^{-2}$). Across all SM types, we found a bias toward increasing genome size: CC-2931 MA lines gained 171.7 bp, and CC-1952 MA lines gained 50.4 bp, per generation, on average. After excluding TEs, the rate of genome size expansion fell by approximately half in CC-2931 and approached zero in CC-1952, highlighting the role of TEs in driving genome size expansion in *C. reinhardtii* MA lines.

In addition to identifying SMs, we used the PacBio HiFi reads to call SNMs and indels <50 bp in length. This analysis suggested that SMs represent ~6% of the total mutations in CC-1952 and ~12% in CC-2931. However, most mutations affecting short lengths of sequence were called in tandem repeats (~60%). Given the uncertainty of variant calling in these sequences, we recalculated mutation rates, excluding these regions. This had a minor effect on the relative contributions of SMs to the total rate (Fig. 2C). We did not find a significant correlation between μ_{SM} and either SNM or indel mutation rates among MA lines in either strain. Disregarding covariance terms between SM, SNM, and indel mutation rates, variance in μ_{SM} explained 4.3% of total variance attributed to the total mutation rate in CC-1952 and 9.7% in CC-2931. Hence, our results suggest that SMs, including those that could have functional consequences, occur at a rate approximately 10-fold lower than the rate of SNMs and indels combined, yet their rates of occurrence are highly variable between strains. The broader improvements to short mutation detection brought by PacBio HiFi sequencing are briefly discussed in the Supplemental Material.

Genomic distribution of SMs

To explore the distribution of SMs across the genome and, in particular, their distribution relative to functional sequences, we generated ~20 Gb of RNA-seq data for the CC-2931 ancestor and annotated coding sequences, introns, and untranslated regions (UTRs). Because >70% of genes are separated by <1 kb in the compact *C. reinhardtii* genome, we divided intergenic regions into "proximal" (within 500 bp of a gene) and "distal" sequences (>500 bp from a gene), the latter category largely capturing long, highly repetitive intergenic regions, including the centromeres and subtelomeres. We then used two approaches to compare the observed distribution of SM coordinates to the expected distribution based on random sampling. First, we considered the entire

span of each SM. Second, we considered the specific coordinates of the SM breakpoints. Although the former approach takes into account the wider possible effects of mutation (e.g., the duplication of entire genes), the latter approach explores whether the breakpoints of SMs were enriched relative to any particular genomics features.

When considering the whole length of duplications, deletions, and inversions, these mutations intersected with genomic annotations, as expected under a random distribution (Fig. 3A). As a result, their proportions of overlap with each functional annotation resembled the actual proportion of the corresponding annotation in the genome. This included coding sequences, suggesting that many of these SMs could have large fitness effects. Furthermore, the breakpoints of duplications and deletions were also distributed randomly (Fig. 3B). These results are consistent with a near absence of selection in our MA experiment, as expected for populations subject to regular bottlenecks. As detailed below, mobile insertions were underrepresented in coding sequences and introns and were enriched in 5′ UTRs and intergenic sequences (Fig. 3B). The coordinates of inversion and translocation breakpoints had qualitatively similar distributions, a result that we attribute to the association of these rearrangements with TEs (see below).

Active TEs

With the exception of one putative mobile satellite in CC-2931, all mobile elements were TEs. We found that the two strains differed markedly in the number and diversity of active TE families. There were 12 active TE families from seven subclasses in CC-2931 and only three active families from two subclasses in CC-1952. We also observed considerable heterogeneity in insertion rates among MA lines and among TE families (Figs. 2B, 4A).

The most active retrotransposons in CC-2931 were an autonomous (*Chlamys-9_cRei*) and nonautonomous (*Chlamys-N4_cRei*) pair of *Penelope*-like elements (PLEs), which generally caused very short insertions of median length 128 bp as a consequence of 5′ truncation. All *Chlamys* insertions were intronic, a pattern that is broadly consistent with the underlying distribution of their $(C)_n$ microsatellite target (Fig. 4B,C; Craig et al. 2021b). The remaining CC-2931 TEs comprised various types of DNA transposons. The *cut-and-paste* DD(E/D) transposons *EnSpm-3_cRei* (autonomous) and *EnSpm-N3_cRei* (nonautonomous) were active in all lines. *EnSpm* insertions were significantly enriched in intergenic regions distant from genes, although unlike other TE families, they were not underrepresented in coding sequences (Fig. 4C). Consistent with a higher efficiency of nonautonomous transposons relative to their longer autonomous counterparts (Han et al. 2013), we observed a net increase in copies of *EnSpm-N3_cRei* (i.e., three ancestral copies compared with a mean of 3.6 copies in the MA lines) and a net decrease in copies of *EnSpm-3_cRei* (i.e., three ancestral copies compared to a mean of 2.4 copies in the MA lines). The increase in *EnSpm-N3_cRei* copies is presumably a consequence of transposition occurring during DNA replication (Feschotte and Pritham 2007). It should be noted that the estimated number of *EnSpm* insertions is a minimum estimate (Fig. 4A) because several *cut-and-paste* transpositions may have occurred during MA that were not captured.

We observed 12 insertions of *copy-and-paste* Helitrons, including the particularly long 20.4-kb autonomous *Helitron2-7_cRei* (Fig. 4B). We also observed an unusual pair of *copy-and-paste* transposons, the autonomous *Replitron-1* and nonautonomous *Replitron-N1*, which are the founding elements of a new group of eukaryotic transposons named Replitrons that will be described elsewhere (Craig 2022). Like Helitrons, *Replitron-1* encodes an HUH endonuclease of the Rep class, although it features only one catalytic tyrosine (i.e., Y1 rather than Y2) in the Rep domain and does not feature a C-terminal helicase domain (Fig. 4B). The Replitrons inserted upstream of "RG" target sequences (R indicating purine), causing variable length target site duplications. *Replitron-N1* was the second most active TE, causing insertions in all lines and a maximum of 22 insertions in L13 (Fig. 3B). *Replitron-1* and *Replitron-N1* insertions were significantly underrepresented within coding sequence and introns (Fig. 4C).

We observed two families of autonomous Cryptons, including *CryptonF-1_cRei,* which was the most active TE in CC-2931. To our knowledge, these are the first observations of active Cryptons. Goodwin et al. (2003) proposed a model of Crypton insertion via site-specific recombination between a short donor sequence at one terminus and a near-identical target sequence at the integration site, catalyzed by the Crypton-encoded tyrosine recombinase. Our data were consistent with this model, because *CryptonF-1_cRei* terminates in the motif "CACCG" and targeted "CAYCG" (Y indicating pyrimidine) (Fig. 4B). However, it was also proposed that Cryptons would undergo excision. Although we observed clean Crypton excisions that left behind only the "CAYCG" target, Cryptons increased in copy number, suggesting a complex mode of transposition. *CryptonF-1_cRei* insertions were enriched at gene-proximal intergenic sequences and 5′ UTRs.

In contrast to CC-2931, only the LINE retrotransposon *Dualen-4b_cRei* was active in multiple CC-1952 MA lines, and the number of insertions ranged from four insertions in MA line 1 (L1) to nine in L15 (Fig. 4A). This family was also active in CC-2931, but it only caused one insertion in L11 (not shown in Fig. 4A). We also observed two *cut-and-*

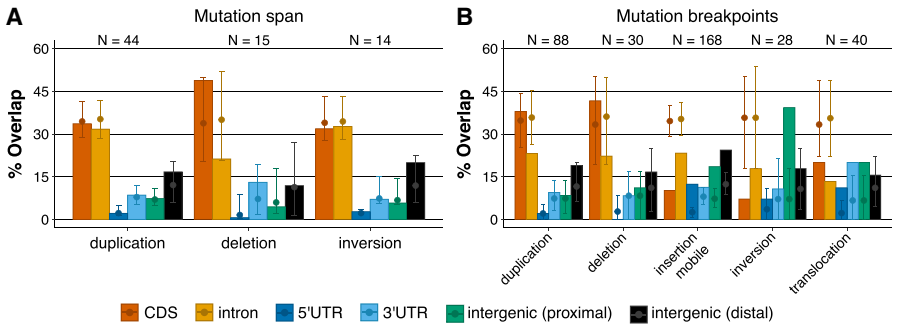

Figure 3. SM genomic distribution. (*A*) Overlap of SMs with genome annotations. Colored vertical bars represent the overlap between different functional annotations (in colors) and types of SM. The expected overlap was then estimated after the sampling of randomly distributed SMs of the same length, using 1000 replicates. Closed circles represent the median expected overlap of SMs with functional annotations, and error bars represent the corresponding 95% confidence intervals. (*B*) Distribution of SM breakpoint coordinates (start and end) relative to genomic annotations. The expected overlap of SM breakpoint coordinates with functional annotation was estimated as in panel *A*, that is, from random sampling of genomic coordinates. Note that the number of observations is twice that in panel *A* because the mutations have two breakpoints (except for mobile insertions).

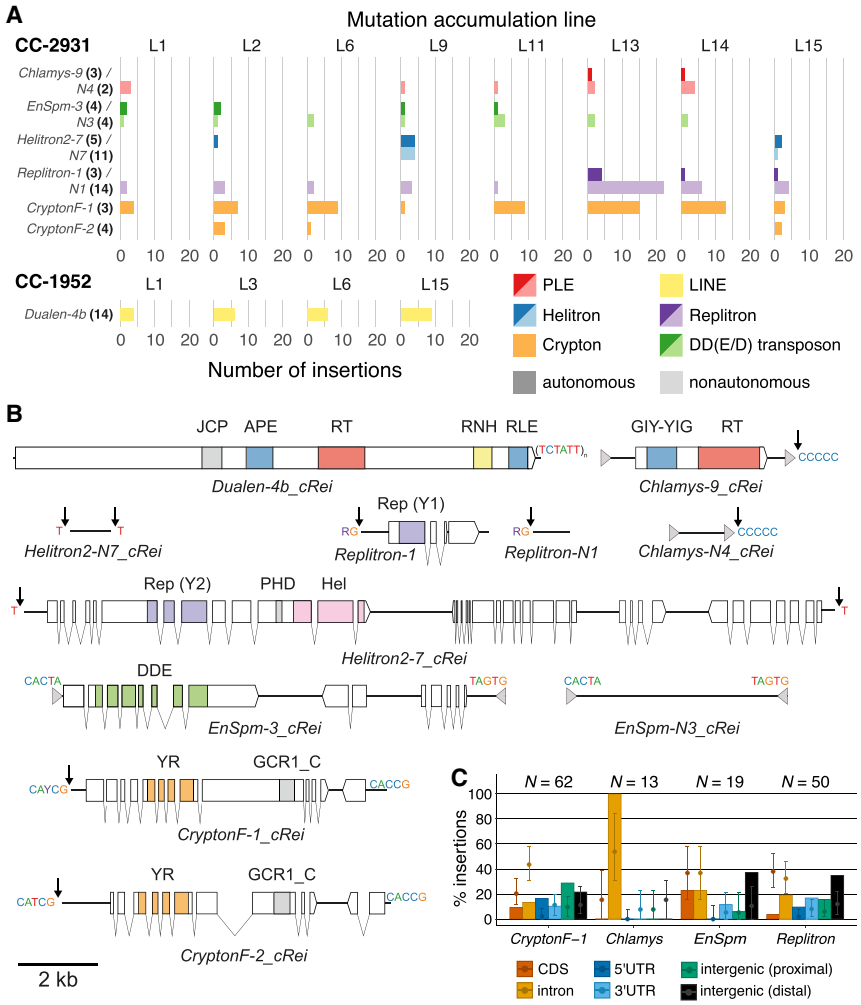

Figure 4. Active TEs. (*A*) Number of insertions per TE family per MA line. Families with less than two insertions per strain are not shown. TE subclasses are shown in colors (red: PLE; yellow: LINE; blue: Helitron; purple: Replitron; orange: Crypton; green: DD(E/D) transposon). Autonomous (darker colors) and nonautonomous (lighter colors) families putatively relying on the same transposition machinery are grouped. The prefix of each TE name denotes the superfamily. Numbers in parentheses and bold indicate the number of ancestral copies of each TE family. (*B*) Schematics of active TE families (to scale). Terminal inverted or direct repeats are shown by gray arrows, terminal sequences are shown *above* the main TE bodies (solid black lines), and insertion targets are shown next to black arrows. Coding sequence and introns of genes are shown by blocks and connecting lines, with domains colored. (RT) reverse transcriptase; (GIY-YIG) GIY-YIG endonuclease; (RNH) Ribonuclease H; (APE) apurinic/apyrimidinic endonuclease-like endonuclease; (RLE) restriction-like endonuclease; (JCP) Josephin-related cysteine protease; (Rep) replication protein (HUH endonuclease); (Hel) helicase; (PHD) plant homeodomain finger; (DDE) DDE transposase; (YR) tyrosine recombinase; (GCR1_C) DNA-binding domain. (*C*) Distribution of specific TE family insertions relative to genomic annotations in CC-2931. Autonomous and nonautonomous pairs were considered together. Error bars show the expectation based on the random sampling of TE insertions, with 1000 replicates per TE insertion (see Fig. 3). Here, the sampling process was adjusted for the genomic distribution of family-specific target sequences (see panel *B*).

paste DD(E/D) transposons in CC-1952 MA lines: a single excision and a single insertion of a *P* element (*P-2_cRei*) in L1, and an excision of a giant single-copy 31.7-kb *Zisupton* DNA transposon (*Zisupton-3_cRei*) in L3, resulting in its extinction from the genome.

Incidentally, we did not observe any gene retrocopy events in either CC-1952 or CC-2931, presumably because of an absence of active polyadenylated LINEs (e.g., *L1* elements in mammals) (Kaessmann et al. 2009) and autonomous LTRs (Tan et al. 2016). Although LTR elements are abundant in the genomes of both strains, we also observed no "deletions" of LTR elements (i.e., formation of solo LTRs by homologous recombination).

TE-mediated SMs

We found that TE insertion and excision events were associated with many other SM types. The number of TE insertions was positively correlated with the combined number of deletions, duplications, inversions, and translocations across all MA lines (Pearson's product-moment correlation, $t_{10} = 2.64$, $r = 0.64$, $P = 0.02$). More specifically, 70.6% of inversions and translocations in the CC-2931 MA lines involved either *CryptonF-1_cRei* or Replitron elements at one or both of the breakpoints (Table 1). One exception to this pattern involved truncated copies of both *Replitron-N1* and *CryptonF-1_cRei* at a single breakpoint. We also attributed a far smaller number of SMs to homology-mediated double-strand break (DSB) repair mechanisms (Table 2). These are described in more detail below.

Translocations mediated by *Replitron-1* and *Replitron-N1* generally featured breakpoints coinciding precisely with the right end of the transposons and were frequently repaired so that one of the derived chromosomes had two elements in a tail-to-tail arrangement. For example, this outcome was observed in a reciprocal translocation between Chromosomes 14 and 17 in CC-2931 L9 (Fig. 5A). Although we do not yet understand the transposition mechanism of these newly discovered TEs, the presence of target site duplications suggests that they may cause DSBs, and their simultaneous insertion at different genomic regions could lead to aberrant repair and rearrangements. *CryptonF-1_cRei*-mediated rearrangements were associated with both insertions and excisions. For example, CC-2931 L11 experienced a reciprocal translocation between Chromosomes 9 and 11, involving insertions at each breakpoint (Fig. 5B). Other events involved only one *CryptonF-1_cRei* insertion, although in all cases we observed a "CAYCG" target site at the breakpoint without insertion. Notably, the high insertion rates of both *CryptonF-1_cRei* and Replitrons in CC-2931 L13 (Fig. 4A) resulted in a highly derived karyotype in this line (Supplemental Fig. S6A). This included a cluster of nonreciprocal translocations involving four chromosomes, mediated by three *Replitron-N1* and one *Replitron-1* insertions that may have occurred simultaneously.

CryptonF-1_cRei was associated with six inversions on Chromosome 16 in different MA lines (Fig. 5C). In the CC-2931

Table 1. Proportion of SMs associated with TEs in MA lines of two *C. reinhardtii* strains

Strain	SM	% associated with TEs (count/total)	TE families (count)
CC-1952	Deletion	33.3 (1/3)	*Dualen-4b_cRei* (1)
	Duplication	86.7 (13/15)	*Dualen-4b_cRei* (13)
	Inversion	0 (0/1)	NA
	Translocation	NA	NA
CC-2931	Deletion	13.3 (2/15)	*CryptonF-1_cRei* (2)
	Duplication	2.27 (1/44)	*Dualen-4b_cRei* (1)
	Inversion	64.3 (9/14)	*CryptonF-1_cRei* (8); *Replitron-1/Replitron-N1* (1)
	Translocation	75.0 (15/20)	*CryptonF-1c_Rei* (8); *Replitron-1/Replitron-N1* (8)

Note that one translocation involved both *CryptonF-1_cRei* and *Replitron-N1*, so the number of TEs does not always match the number of TE-mediated SMs.

ancestor, Chromosome 16 features two *CryptonF-1_cRei* copies in opposite orientation separated by ~233 kb. However, we did not observe this ancestral state in any of the eight MA lines. The ~233-kb region was inverted in four MA lines, with either none, one, or both of the copies excised. In L2 and L14, longer inversions were mediated between one of the *CryptonF-1_cRei* copies and independent "CAYCG" target sites elsewhere on the chromosome. Overall, there were seven different mutated states relative to the ancestor (Fig. 5C), evidencing the hypermutability of Crypton activity. The presence of target sites at each breakpoint in both inversions and translocations may suggest a role for the site-specific recombination activity of the Crypton tyrosine recombinase enzyme in mediating rearrangements.

Finally, we observed duplications associated with the LINE *Dualen-4b_cRei*, which was most active in CC-1952 (Fig. 4A) but also caused a single insertion and associated duplication in one CC-2931 MA line. More than half of the *Dualen-4b_cRei* insertions were associated with duplications >50 bp, which represented 87% of duplications in CC-1952 (Table 1). These duplications (of median length 900 bp) (Supplemental Fig. S7) resembled the variable length target site duplications that flank insertions of non-LTR elements (LINEs and PLEs); that is, the duplicated sequence flanked either side of the *Dualen* insertions. Such target site duplications are caused by resolution of the DNA nicks introduced during insertion, the distance between cleavage sites corresponding to the target site duplication length. These putative *Dualen-4b_cRei* target site duplications are considerably longer than other large target site duplications reported previously, that is, the 126-bp target site duplications observed in *R9* LINEs of rotifers (Gladyshev and Arkhipova 2009). Active *Dualens* have not previously been observed, and exceptionally long target site duplications are possibly mediated by the dual action of RLE and APE endonucleases (Fig. 4B), which are uniquely present together in the *Dualen* clade (Kojima and Fujiwara 2005).

Homology-mediated SMs

We next attempted to identify homology-based DSB repair mechanisms that may have mediated the remaining SMs. These included most deletions and duplications in CC-2931 (Table 2). Many mechanisms associated with SMs involve DSB repair, which typically proceeds via two distinct pathways: homologous recombination and canonical nonhomologous end joining. Homologous recombination can induce SMs via recombination between interspersed paralogous sequences, that is, nonallelic homologous recombination. Nonhomologous end joining can also mediate SMs. For example, two DSBs that are present on the same chromosome could be repaired aberrantly, yielding a deletion or an inversion, or two DSBs that are present on different chromosomes could yield a translocation. Several other DSB repair pathways that involve varying lengths of homology tracts also exist (So et al. 2017). For example, microhomology-mediated end joining requires ~1–16 bp of sequence homology, and single-strand annealing requires macrohomology >30 bp. Both mechanisms generate deletions at single DSBs because they involve DNA end resection, whereby homology is revealed at DNA ends when degraded to single-strand sequences, and this is followed by deletion of any sequence overhanging the homology tract (Sfeir and Symington 2015).

We found evidence of macrohomology (>30 bp) in only ~5% of duplication events, one event in CC-1952 and two in CC-2931 (Table 2; Supplemental Fig. S8). These duplications could potentially have been caused by nonallelic homologous recombination. In all these cases, the paralogous sequence did not feature TEs. We found no evidence of macrohomology-mediated deletions, suggesting little role for the single-strand annealing repair pathway causing SMs. In contrast, we identified microhomologies in 28% of deletions, four in CC-2931 and one in CC-1952 (Table 2; Supplemental Fig. S9). One of these deletions (in CC-2931 L6), was flanked by several clustered SNMs and indels (Supplemental Fig. S10), a phenomenon that has been observed flanking DSBs repaired by microhomology-mediated end joining (Sinha et al. 2017). Although similar hypermutability was not observed flanking the other microhomology-associated deletions, these deletions were generally shorter than deletions showing no evidence of homology (median length, 225 bp vs. 2,945 bp; *W*-test, $P = 3.77 \times 10^{-2}$). This is consistent with the involvement of DNA end resection in microhomology-mediated end joining.

Table 2. Proportion of SMs associated with homology-based mechanisms in MA lines of two *C. reinhardtii* strains

Strain	SM	% with macrohomology (count/total)	% with microhomology (count/total)
CC-1952	Deletion	0 (0/3)	33.3 (1/3)
	Duplication	6.7 (1/15)	0 (0/15)
CC-2931	Deletion	0 (0/15)	26.7 (4/15)
	Duplication	4.5 (2/44)	6.8 (3/44)

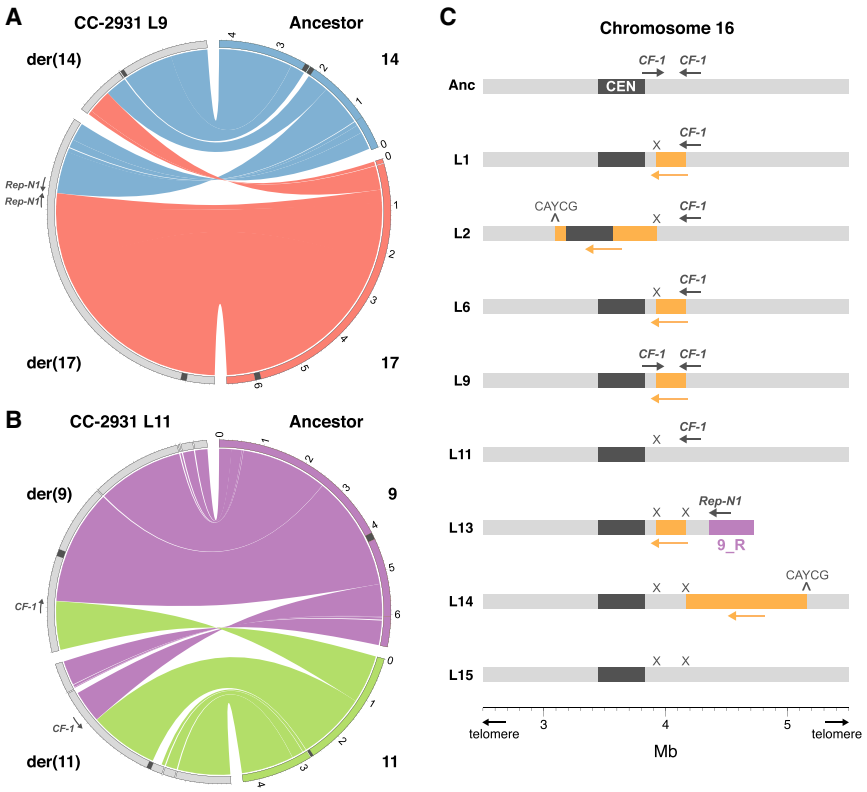

Figure 5. Genomic rearrangements mediated by TEs. Translocations in CC-2931 MA lines 9 (*A*) and 11 (*B*). The ancestor chromosomes are represented on the *right* half of each Circos plot (Krzywinski et al. 2009), whereas the MA line chromosomes (as contigs) are on the *left*. Chromosome numbers are given for the ancestor, and derived chromosomes (denoted by "der") are provided for the MA lines based on centromere annotation (dark gray regions). Scale is in megabases. *CF-1* indicates *CryptonF-1_cRei* insertions; *Rep-N1* indicates *Replitron-N1*. The direction of the arrows indicates the left to right orientation of the TE sequence. (*C*) *CryptonF-1_cRei* (*CF-1*)-mediated inversions on Chromosome 16 in CC-2931 MA lines. Ancestor and MA line genomes are shown from *top* to *bottom*. The dark gray region represents the centromere, and the orange blocks represent inversions. The purple region in L13 shows a *Replitron-N1*-mediated translocation with Chromosome 9. Gray arrows indicate the orientation of *CryptonF-1_cRei* from left to right, and "X" indicates *CryptonF-1_cRei* excisions.

We also found three duplications in CC-2931 that had five, three, and two nucleotides of sequence homology between their respective breakpoints. Although these duplications are consistent with microhomology-mediated mechanisms, the presence of such short stretches of homology might also be explained by chance. Finally, we found no evidence of either micro- or macrohomology at the breakpoints of the inversions and translocations not associated with active TEs. However, we did observe short deletions at several inversion and translocation breakpoints, which is consistent with repair via nonhomologous end joining. Notably, 40% of non-TE-associated inversions and translocations occurred in a single CC-2931 line, L15 (one inversion and three translocations) (Supplemental Fig. S6B).

Tandem repeat mutations

Tandemly repeated sequences are known to be hypermutable and evolve via mechanisms that include replication slippage and unequal exchange (Lower et al. 2018). As mentioned above, unlike most other SMs, TRMs (grouping expansions and contractions >50 bp) occurred at similar rates in the CC-2931 and CC-1952 MA lines. This could be owing to the independent underlying mechanisms of mutation compared with other SMs. Approximately 22% of TRMs occurred in microsatellites (defined as tandem repeats with monomers <10 bp), and most of the remainder occurred in satellite DNA. The *C. reinhardtii* genome contains many introns featuring tandem repeats (Craig et al. 2022), and 30.3% and 64.7% of satellite and microsatellite TRMs, respectively, were located in genic sequences. It is important to note that the rate of TRMs per site per generation is several times higher than the genome-wide rate displayed in Figure 2B, because TRMs can only occur within the ~12% of the genome that features tandem repeats, effectively resulting in a smaller "callable" genome.

We also observed numerous expansions and contractions in the centromeric and subtelomeric regions. The centromeres mostly consist of an *L1* LINE retrotransposon, *ZeppL-1_cRei* (Craig et al. 2021a), and we observed expansions and contractions of these sequences (generating 7.0% of TRMs), consistent with centromere evolution via satellite-like mechanisms such as unequal exchange rather than by active transposition. Subtelomeres feature a unique satellite called *Sultan*, and we observed expansions and contractions of the *Sultan* monomer within the same subtelomere, as hypothesized by Chaux-Jukic et al. (2021). A small number of TRMs involved the expansion or contraction of tandemly repeated gene families. Examples included 5S ribosomal RNA arrays and clusters of the large *Chlamydomonas*-specific *NCL* gene family on Chromosome 15, which encode RNA-binding proteins and appear to be experiencing rapid and ongoing evolution in *C. reinhardtii* (Boulouis et al. 2015).

Finally, we observed one example of a "mobile" satellite, which caused four insertions ranging from ~0.5 to >21 kb in length in three different CC-2931 MA lines. This satellite, *MSAT-11_cRei*, consists of a ~1.9-kb monomer and does not feature any characteristics typical of a TE. We have recently observed mobile insertions of *MSAT-11_cRei* in other *C. reinhardtii* strains (CC-1690 and CC-4532) (Craig et al. 2022). Although the mechanisms mediating satellite dissemination are not well understood, the phenomenon has been observed in other species and may be an important mechanism of satellite evolution (Ruiz-Ruano et al. 2016).

Whole-chromosome duplication and terminal deletions

We explored the possibility of whole-chromosome duplications using a genome coverage analysis and found evidence for one whole-chromosome duplication in CC-2931 L15 (Supplemental Fig. S11), which gives an estimated rate of 1.2×10^{-4} per genome per generation in the CC-2931 genotype. The existence of the duplication is supported by nearly 2× coverage of both the

PacBio and Illumina reads relative to their median whole-genome coverage. Furthermore, we identified "heterozygous" mutations of all types along this chromosome, which presumably accumulated after the chromosome duplication event. One example of a heterozygous SM is a reciprocal translocation between one copy of the duplicated Chromosome 1 and the single copy of Chromosome 11 (Supplemental Fig. S11). Aneuploidy events, such as chromosome gain, can likely be attributed to chromosome segregation errors and are generally expected to be deleterious (Krasovec et al. 2022), although they can also facilitate adaptation in some instances (Gilchrist and Stelkens 2019). We did not observe chromosome loss, which would be lethal. However, we did observe the deletion of a chromosome terminus in two MA lines (CC-1952 L15 and CC-2931 L1), followed by de novo telomere addition (i.e., "telomere healing") (Supplemental Fig. S12). These events resulted in terminal deletions of 94 kb and 163 kb, in CC-1952 L15 and CC-2931 L1, respectively.

Discussion

In total, we identified 563 SMs in 12 MA lines that were derived from the *C. reinhardtii* strains CC-1952 and CC-2931. To our knowledge, these are the first direct estimates of the rates and spectra of de novo SMs based on long-read sequencing of MA lines. In agreement with previous results on the rates of SNMs and indels in *C. reinhardtii* (Ness et al. 2015), SM rates and spectra vary greatly between MA lines and strains. Furthermore, SMs represented a substantial proportion of the overall genomic mutation rate in both strains and affected far higher proportions of the genome at the per base level than SNMs and indels.

Calling SMs from long reads and assemblies

De novo mutations are inherently rare events, and highly accurate methods are therefore required for their detection. Our ability to detect SMs was aided by major advances in long-read sequencing technology, long-read and whole-genome aligners, and structural variant callers. Nonetheless, our results show that the detection of structural variants remains much more challenging than the detection of shorter variants, making it necessary to use a combination of approaches. Because there was only a partial overlap between variants identified by different callers, it is possible that some SMs were undetected in our study. Although a read-based caller (Sniffles) was required for the identification of many variants, especially duplications, this approach failed to detect the full range of SMs. We found that the pangenome approach implemented by Cactus and vg was particularly successful, calling ~80% of all curated SMs.

However, in all cases, our results show that structural variant callers are likely to yield high rates of false positives, even when the samples analyzed are nearly isogenic, as in our experiment. Although we were able to detect many genuine SMs in tandem repeats, we found that these regions were responsible for the majority of false-positive calls. Accurately calling SMs in tandem repeats may therefore require specific alignment and variant calling tools to be developed, and given that manual curation of variants is unlikely to be manageable in larger and more complex genomes than that of *C. reinhardtii*, masking of tandem repeats may be appropriate in automated analyses. The substantial contribution of assembly errors to rejected calls may also warrant the use of multiple assemblers in variant calling, or the development of methods that combine assembly-based structural variant detection with read-based verification. Overall, we recommend sequencing samples with sufficient coverage to enable de novo assembly, followed by both assembly and read-based SM or variant calling. If possible, manual verification with visualization tools such as IGV should also be performed.

Mechanisms underlying SM and between-strain heterogeneity of rates and spectra

Excluding expansions and contractions of tandem repeats, ~79% of SMs were associated with TEs. Of these, 84% were TE insertions and excisions, which therefore formed major components of the SM spectra in both strains. The remaining 16% of TE-associated SMs included most inversions and translocations in CC-2931 and almost all of the duplications in CC-1952. The *C. reinhardtii* genome contains more than 200 diverse TE families, almost all of which show evidence of recent activity (Craig 2021). However, none of the three TEs implicated in causing rearrangements or duplications have previously been observed as active elements. Two of these active elements, Cryptons (Goodwin et al. 2003) and *Dualen* LINEs (Kojima and Fujiwara 2005), were first described from multicopy repeats in genetic and genomic data, whereas the active Replitrons found here have not been described previously.

Beyond TEs, we found little role for homology-based mechanisms of DSB repair, with the possible exception of the microhomology-mediated end joining pathway, which putatively caused deletions of moderate lengths. This is in contrast to SM studies in yeast, in which nonallelic homologous recombination has been shown to be the predominant mechanism mediating deletions, duplications, and rearrangements (Sui et al. 2020). Similarly, 10%–20% of de novo SMs in humans are thought to be mediated by nonallelic homologous recombination (Parks et al. 2015). Instead of homology-based mechanisms, the nonhomologous end joining repair pathway may have been involved in many of the SMs, particularly those not mediated by TEs. This is consistent with the very low rates of homologous recombination observed in *C. reinhardtii* under vegetative growth, where the species is haploid and nonhomologous end joining is the dominant DSB repair pathway (Ferenczi et al. 2021). The relative rates at which species repair DSBs via either nonhomologous end joining or homologous recombination have been implicated in many aspects of genome evolution, such as the evolution of base composition (Weissman et al. 2019) and intron density (Farlow et al. 2011), and it is likely that variation in the activity of DSB repair pathways also leads to substantial variation in SM spectra among species.

We found substantial variation in the rates of SNMs and indels between the CC-1952 and CC-2931 strains, both in our previous analysis with Illumina sequencing (Ness et al. 2015) and in the reanalysis reported herein using PacBio HiFi. There is substantial nucleotide diversity among *C. reinhardtii* strains (Flowers et al. 2015; Craig et al. 2019), and some of the within-species variation in mutation rates that we have observed may be caused by the presence of mutator alleles in certain strains. Consistent with our findings for SNM and short indel mutations, the SM rate in CC-1952 MA lines was significantly lower than in CC-2931. Furthermore, the SM spectra differed substantially between the strains. In particular, CC-2931 has a higher overall rate of transposition than CC-1952, involving a more diverse array of TEs. TE suppression is not well understood in *C. reinhardtii* but likely occurs at transcriptional and post-transcriptional levels (van Dijk et al. 2006) via mechanisms including repressive histone modifications (Jeong

et al. 2002; Zhang et al. 2002) and RNA interference (Casas-Mollano et al. 2008). Because the CC-2931 and CC-1952 genomes both harbor many more potentially active TE families than we observed to be active in our study, we infer that all but a few families are silenced effectively. TE suppression may differ between CC-2931 and CC-1952 owing to genetic variation in genes involved in silencing pathways. Alternatively, environmental factors could impact TE activity. Almost nothing is known about local adaptation in *C. reinhardtii*, although natural isolates differ in their growth rates under laboratory conditions (Morgan et al. 2014; Kraemer et al. 2017). CC-1952 and CC-2931 were sampled ~1600 km apart, from Minnesota and North Carolina, respectively, and they may differ in the extent of their adaptation to the highly artificial laboratory environment. Such differences could potentially cause stress-related interactions with transposition, which can result in TE activation or repression (Horváth et al. 2017).

In addition to a higher rate of transposition, CC-2931 MA lines also appear to have experienced a higher rate of DSBs than CC-1952 MA lines, which may explain the higher rates of duplications, deletions, and rearrangements in MA lines of this strain, even after TE-mediated SMs have been accounted for. DSBs are generally considered to be the most mutagenic DNA lesions (So et al. 2017) and are induced by intrinsic cellular factors (e.g., replicative and oxidative stresses) and by exogenous sources (e.g., mutagens). The rate of DSBs in *C. reinhardtii* is, however, not well understood, and differences between the strains could arise from genetic or environmental factors. Overall, we infer that the CC-2931 genome appears to be less stable than that of CC-1952, although the reasons for this are currently obscure.

Evolutionary implications of high and variable rates of SM

Population-level long-read sequencing projects have generally found that structural variants are common but segregate at low frequencies, implying that most are strongly deleterious (Chakraborty et al. 2019; Weissensteiner et al. 2020). Although we have not explored the relationship between SMs and fitness, the genomic distribution of SMs, in many cases overlapping coding sequences, suggests that many have large fitness effects. In contrast, a growing body of evidence suggests that coding sequences are less mutable than intergenic sequences with respect to SNMs and indels (Lee et al. 2012; Krasovec et al. 2017; Belfield et al. 2018; López-Cortegano et al. 2021; Monroe et al. 2022). We also found a substantial number of tandem repeat expansions and contractions in genic regions, a class of mutations that has been implicated as a major source of human genetic disease (Hannan 2018). Furthermore, the genomic distribution of TE insertions was not random. Considering the most active elements in CC-2931, Replitron insertions were enriched in intergenic sequences distant from genes, whereas *CryptonF-1_cRei* insertions were overrepresented in 5′ UTRs and gene proximal intergenic regions. Similar insertion biases in other TE families are well documented, and for example, *P* elements in *Drosophila* and *Mu* elements in maize show biases toward transcription start sites of highly expressed genes (Zhang et al. 2020). Gene proximal TE insertions can have important effects on gene expression, via the disruption of regulatory sequences, regional effects of transcriptional silencing, or the deposition of new regulatory elements (Cridland et al. 2015; Uzunović et al. 2019; Rech et al. 2022).

Consistent with the association between TEs and chromosomal rearrangements, the breakpoints of inversions and translocations also shared a similar genomic distribution to that of TE insertions. We expect that many of these events may have large fitness effects. Although not a factor in clonally propagated MA lines, many of the observed rearrangements are likely to cause meiotic incompatibilities. Furthermore, it is very likely that we have underestimated the true rate of translocations; we only observed translocations that retained a single centromere per chromosome, likely owing to highly reduced fitness following the formation of acentric chromosomes. Overall, the high fraction of the total mutation rate explained by SMs, together with their genomic distribution, suggests that SMs contribute substantially to the mutation load but also to the potential for adaptive evolution.

Notably, we also observed a strong bias toward genome size expansion in MA lines, which was driven by TE insertions and duplications. In some cases, the ancestral copy number of a TE increased severalfold in an MA line genome over the course of the experiment. The *C. reinhardtii* genome is highly compact, and TE copy number is generally low within TE families (Craig 2021), suggesting that the observed tendency toward genome expansion in MA lines is counteracted by purifying selection in nature. Based on short-read data, similar expansion biases have been reported for duplications in MA lines of *Caenorhabditis elegans* (Konrad et al. 2018) and for TE insertions in MA lines of *Daphnia magna* (Ho et al. 2021). Briefly, it is important to note that *copy-and-paste* TEs may effectively act as mutators in an MA experiment because each new insertion has the potential to create a new active copy and subsequently increase transposition rate. Thus, our observed rates of TE insertion, which were averaged over almost 1050 generations, may be overestimated relative to those occurring in nature.

Our results highlight the prevalence and importance of TEs among SMs and support a prominent evolutionary role for TEs. The heterogeneity in the rate and identity of TE insertions between CC-2931 and CC-1952 contributed substantially to the overall differences in SM rates and spectra between the strains. Our results suggest that species, populations, and even individuals may differ considerably in their SM spectra as a result of their active TE repertoire. More work would be required to investigate the rate of new SMs in other species and to elucidate the generality of results observed here. In particular, it will be important to test whether similar within-species variation in SM rates and spectra exists in other taxa. Our results add to the weight of evidence supporting the importance and prevalence of SMs and should further encourage structural variant discovery via assembly-based methods.

Methods

Biological samples and nucleic acids extraction

The MA line ancestors were *C. reinhardtii* wild strains CC-1952 (from Minnesota, 1986) and CC-2931 (North Carolina, 1991), which were originally obtained from the *Chlamydomonas* Resource Center (https://www.chlamycollection.org/). The MA experiment was conducted by Morgan et al. (2014). Briefly, MA lines were initiated from the ancestor strains and cultured on Bold's medium agar plates under white light at 25°C. MA lines were bottlenecked at regular intervals of 3–5 d by randomly picking single colonies and transferring them from one plate to another. MA lines were maintained for estimated averages of 1066 and 1050 generations for CC-1952 and CC-2931, respectively, after which Illumina sequencing was performed. The original ancestors and MA lines from the last transfer of the experiment were cryopreserved in liquid nitrogen.

For this study, we reconditioned the CC-1952 and CC-2931 ancestors along with several MA lines after cryopreservation and grew all samples in liquid Bold's medium before transferring to agar slants to produce stock cultures. Four CC-1952 MA lines (L1, L3, L6, and L15) and eight CC-2931 MA lines (L1, L2, L6, L9, L11, L13, L14, and L15) were sequenced together with the two ancestors. MA lines were randomly selected after excluding lines with combined SNM and indel mutation rates greater than or less than 1.5 times the interquartile range for that strain (Ness et al. 2015), because these lines may have accumulated mutations that modify the ancestral mutation rate. Cells were inoculated in six-well plate liquid cultures and grown for 4 d under constant light to produce sufficient biomass for DNA extraction. High-molecular-weight genomic DNA was extracted using a cetyltrimethylammonium bromide (CTAB) and phenol:chloroform protocol, following the method of Craig et al. (2021a). RNA was extracted in triplicate from independent cultures of the CC-2931 ancestor grown in liquid Bold's medium under constant light via a Maxwell RSC 48 instrument.

Nucleic acid sequencing

All of the CC-1952 samples and six of eight CC-2931 MA lines were sequenced on the PacBio sequel II platform with a 30-h movie, using the circular consensus sequencing (CCS) mode to generate HiFi reads. Samples were multiplexed, with between four and six samples per SMRT cell. Library and barcoding preparation, sequencing, CCS analysis, and demultiplexing were performed at the Earlham Institute (Norwich, United Kingdom). The mean read length N50 was 20.2 kb per sample, and mean coverage was about 27× per sample.

The CC-2931 ancestor and CC-2931 MA lines L2 and L9 were sequenced on individual SMRT cells using the PacBio sequel I platform with a 10-h movie, using the CLR mode. Library preparation and sequencing was performed at Edinburgh Genomics (Edinburgh, United Kingdom). Mean read length N50 was 19.3 kb, and mean coverage was about 54× per sample.

CC-2931 ancestor RNA-seq library preparation was conducted with the NEB mRNA stranded library preparation kit and Illumina NovaSeq preparation. The three replicate samples were sequenced using Illumina 100-bp paired-end sequencing. Library preparation and sequencing were performed by Génome Québec (Montreal, Canada).

Genome assembly

A different assembly strategy was followed for MA lines and ancestors. MA lines were assembled de novo to the contig-level. Flye v2.8.2 (Kolmogorov et al. 2019) was selected for assembly because it produced assemblies that were most representative of the haploid state (i.e., other assemblers yielded redundant contigs in repetitive regions). A genome length of 111.1 Mb was assumed ("-g 111.1m" in Flye; command line arguments are shown here and elsewhere in double quotation marks). Postprocessing was performed with purge_dups (Guan et al. 2020). Error correction was only performed for the two CC-2931 MA lines (L2 and L9) sequenced using CLR by applying two iterative rounds of the Arrow algorithm (Hepler et al. 2016). To facilitate their use as reference genomes, the ancestor genomes were assembled de novo to the chromosome-level using a combination of assembly methods and extensive manual assembly (see Supplemental Material).

Assembly metrics were quantified using QUAST v5.0.2 (Gurevich et al. 2013). Assembly completeness was estimated using BUSCO v4.0.6 (Manni et al. 2021), which was run in genome mode using the chlorophyta_odb10 data set ("augustus_species chlamy2011"). A list of all samples with a summary of their assembly approaches and quality metrics can be found in Supplemental Table S1.

Genome annotation

Structural annotation of the CC-2931 ancestor assembly was performed using BRAKER2 v2.15 (Brůna et al. 2021). RNA-seq data were mapped to the assembly using STAR v2.7.9 ("--alignIntronMax 5000 --twopassMode Basic") (Dobin et al. 2013). BRAKER2 was first run on an assembly softmasked for all repeats (see below) using existing AUGUSTUS parameters for *C. reinhardtii* ("--species = chlamy2011 --skipAllTraining"). A second BRAKER2 run was performed with UTR prediction ("--stranded = +, --UTR = on"), where the input BAM alignments were split into forward- and reverse-strand read sets using SAMtools v1.9 (Danecek et al. 2021). Because the run without UTR prediction returned a superior BUSCO score ("protein" mode), we designated this the primary annotation but also attempted to add UTRs where possible. For gene models in which coding sequence coordinates had a one-to-one correspondence between the two BRAKER2 runs (with variation permitted at only one exon for models with more than two exons), the model with UTRs was introduced as a replacement. Gene models were filtered if their coding sequence intersected in >30% with TE sequence or >70% with simple repeats (i.e., microsatellites) identified with RepeatMasker v4.0.9 (https://www.repeatmasker.org).

Repeat annotation was performed for the CC-2931 and CC-1952 ancestor assemblies. A custom library of *C. reinhardtii* TEs and satellites (Craig 2021) was first passed to RepeatMasker. A small number of TE families newly identified in this study were manually curated and added to the library (see below). The RepeatMasker annotations were supplemented with additional microsatellites and satellites identified by Tandem Repeats Finder ("2 7 7 80 10 50 1000 -f -d -m -ngs") (Benson 1999). A final set of tandem repeats for each assembly was produced by combining simple repeats and satellites from RepeatMasker, satellites and microsatellites from Tandem Repeats Finder, and manually curated centromeric, subtelomeric, and ribosomal DNA array coordinates. Centromeres were identified based on the span of the constituent LINE *ZeppL-1_cRei* (Craig et al. 2021a), and subtelomeres were identified as sequence between the telomere and the characteristic spacer sequence, which coincides with the transition from the subtelomeric repeats (Chaux-Jukic et al. 2021). Tandem repeat sequences were considered to be microsatellites if they had a monomer length of <10 bp, and as satellite DNA if their monomer was >10 bp.

Mapping and alignment

To produce the input files for Sniffles, MA line PacBio read mapping was performed against the appropriate ancestor reference assembly using pbmm2 align (https://github.com/PacificBiosciences/pbmm2), adding the "--preset CCS" flag for HiFi samples. For input to DeepVariant (see below), additional processing with SAMtools view (Danecek et al. 2021) run with the option "-F 256" was performed to remove reads with a flag indicating secondary alignment.

For visualization and curation of SMs, MA line assemblies were aligned to the appropriate reference assembly using minimap2 ("-x asm5") (Li 2018). We also used unimap v0.1-r41 (https://github.com/lh3/unimap), a tool derived from minimap2 and optimized for assembly alignments, which provided superior alignments across many tandem repeats.

A pangenome alignment was produced for each ancestor and its MA lines using Cactus v1.3.0 (Armstrong et al. 2020). All

assemblies were first softmasked for repeats using RepeatMasker and Tandem Repeats Finder, as described above. Minigraph v0.15-r426 ("-xggs") (Li et al. 2020) was used to produce the graphical fragment assembly (GFA) file that was provided as input for cactus-graphmap together with softmasked assemblies. The resulting PAF file was then passed to cactus-align ("--pangenome --pafInput --outVG") in order to produce the final pangenome in variation graph (VG) format.

Callable sites

Callable sites were defined as genomic sites where SMs could be called with high confidence, based on visualization of the sequencing and assembly data. We previously defined callable sites based on short-read mapping parameters (Ness et al. 2015; López-Cortegano et al. 2021). Here, we used two criteria based on the de novo genome assemblies of the ancestors and MA lines. First, the ancestor assembly was aligned against itself with minimap2 ("-x asm5"), and genomic regions that were absent in the resulting PAF file (i.e., that were unmapped) were deemed uncallable. This first criterion was used because these regions are essentially unmappable even as isogenic sequences (at least with minimap2) and are hence inaccessible to variant calling. Second, a similar procedure was followed for each MA line by aligning their assemblies to the ancestor genome and extracting unmapped genomic coordinates. The unmapped coordinates extracted from all MA lines per ancestor were then intersected using BEDTools "intersect" (Quinlan and Hall 2010), and any regions that were present in at least two MA lines were defined as uncallable, because an unmapped region in a single MA line could be an SM such as a large deletion. This second criterion was adopted because these regions are prone to assembly breaks across multiple lines, even though they are assembled in the ancestor references. Taking the output from both criteria, uncallable regions separated by <30 kb were merged. Finally, coordinates corresponding to active *cut-and-paste* DNA transposons were manually reincluded as callable, because the excisions at these sites could otherwise be classified as uncallable if the same TE was excised in multiple lines. Because the uncallable regions are enriched in tandem repeats (see Supplemental Material; Supplemental Dataset S2; for the example dotplot, see Supplemental Fig. S4), we expect that we may have underestimated the mutation rate of TRMs. However, there was no overrepresentation of other SM types in callable tandem repeats, suggesting that the callable regions of the ancestor assemblies provide near-complete references for detecting most SMs genome-wide.

To compare callable sites between our previous Illumina sequencing and the PacBio sequencing described here, callable site coordinates from Ness et al. (2015) were converted to correspond to our new ancestor assemblies. A whole-genome alignment of the v5 reference assembly and the ancestor assemblies was generated using Cactus (Armstrong et al. 2020). Coordinates were then lifted over to the relevant ancestor assembly using the HAL tools command halLiftover (Hickey et al. 2013).

SM identification

SMs were called using three different variant callers, each of which relied on a different underlying alignment tool. Sniffles v1.0.12b (Sedlazeck et al. 2018) was used to call SMs based on the pbmm2 read alignments described above. BAM files were preprocessed using SAMtools-calmd to generate the MD tag, which provides information on mismatching positions (i.e., variable coordinates in the reads). Sniffles was first run on each MA line individually, and the resulting VCF files were merged using SURVIVOR v1.0.7 (Jeffares et al. 2017). Following the pipeline recommended for population calling (https://github.com/fritzsedlazeck/Sniffles/wiki/), Sniffles was then run again with the merged VCF as input and the option "--Ivcf." This population calling enables consistent presence or absence calls for SMs across all MA lines within a strain. SURVIVOR was used again to generate a multisample VCF.

MUM&Co v3 (O'Donnell and Fischer 2020) was used to call SMs from individual alignments of MA line assemblies to their ancestral reference, setting a genome size of 110 Mb ("-g 110000000"). MUM&Co calls variants based on alignments produced by MUMmer v4 (Marçais et al. 2018), which is performed as part of a single script. Variants were obtained as TSV and VCF files.

The variation graph tool (vg) (Garrison et al. 2018) was used to call variants directly from the pangenome alignments using the deconstruct command ("--path-traversals"). The resulting VCF file for each strain was reduced to variants >50 bp.

All called variants in callable regions were manually curated via visualization of read and assembly alignments using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011). SMs were rejected if they were not supported unambiguously by the read alignments. Read support for very large SMs was visualized via Ribbon v1.1 (Nattestad et al. 2021), which enables the visualization of reads mapping to discordant genomic regions. Supplemental Figures S12–S26 provide examples of SM visualization and curation. Most variants were entirely spanned by the reads, leading to simple visual confirmation in IGV, but variants >30 kb in length (approximately the upper limit of read lengths), including large inversions and translocations, required additional curation. In addition to read support from Ribbon, these rearrangements were traced in the MA line assemblies by manually assessing the discordant mapping of MA line contigs in the PAF alignment files (see Supplemental Fig. S23). Complex SMs, including large rearrangements and duplications, were further visualized using Ribbon v1.1 (Nattestad et al. 2021).

Duplications and deletions were curated as tandem repeat expansions or contractions if they involved the duplication or deletion of one or more monomers of a tandem repeat. Most fell within existing tandem repeat annotations, that is, satellites and microsatellites, whereas a small number required manual inspection of indel flanks by self-vs-self dotplots generated using the MAFFT v7 online server (Katoh et al. 2019). Deletions that perfectly intersected with TEs annotated by RepeatMasker in the ancestor genome were called as mobile excisions. Mobile insertions for described TE families were identified as cases in which the inserted sequence had a near-perfect BLASTN match (Camacho et al. 2009) to the *Chlamydomonas* repeat library (Craig 2021). These hits all had expected length distributions; LINE and PLE insertions frequently only contained the 3′ end owing to 5′ truncation, whereas insertions of other TEs corresponded to the entire length of the TE. In cases in which an inserted sequence had no match to an existing TE model, we queried the insert sequence against the ancestor genome, extracted and aligned hits, and manually curated new consensus sequences following established protocols for mobile element annotation (Goubert et al. 2022). All insertions unambiguously matched either the existing or newly produced consensus sequences and could be neatly defined to specific mobile element families. The one exception to this pattern was the duplications mediated by *Dualen* LINEs, where the sequence called as an insertion partly matched *Dualen-4b_cRei* and partly matched the sequence immediately flanking the insertion. These *Dualen*-mediated duplications were manually split to two called SMs: one mobile insertion and one duplication of the appropriate lengths.

When curating inversions and translocations, we noticed that many events featured additional insertions at the

rearrangement breakpoints that were not specifically detected by the variant callers. As above, these insertions were compared to the annotated TEs and defined as mobile insertions of specific TE families. Five rearrangements could not be fully characterized because one of the breakpoints was clearly supported, but the other was in an uncallable region. These were arbitrarily classified as translocations.

SNMs and indels

DeepVariant 1.1.0 (Poplin et al. 2018) was used for calling SNMs and indels based on read alignments, using the option "--model_type=PACBIO." DeepVariant was run on individual MA line pbmm2 alignment files. The resulting VCF files were merged using GLNexus 1.3.1 (" --config Deepvariant_unfiltered") (Yun et al. 2020). The merged VCF file was further processed to retain only high-quality calls (QUAL ≥ 20) of sites called as homozygous genotype, because DeepVariant assumes diploid genomes and biallelic variants, with a minimum read depth of eight. In addition, only variant calls that were unique to a single MA line were retained as mutations. All SNMs and indels were confirmed visually in IGV.

Genomic distributions of SMs and TE insertions

The coordinates of SM breakpoints in CC-2931 were intersected with genomic annotations (coding sequence, introns, etc.) using BEDTools. Additionally, we also calculated the intersection of deletions, duplications, and inversions based on the entire span of these SMs. The observed distributions of SMs were compared against null expectations based on random sampling of the callable genome. For the analysis of individual TE families, the random expectation was adjusted to account for insertion target sequences (see Fig. 4B). The target motifs were identified using SeqKit v2.1.0 (Shen et al. 2016), and these motifs were then sampled from the callable genome. TRMs and mobile excisions were not analyzed.

Sequence homology

To explore the role of different DSB repair pathways in SM generation, we searched for patterns of micro- and macrohomology at SM breakpoints, following the method of Belyeu et al. (2021). We first searched for evidence of macrohomology between the breakpoints of deletions, duplications, inversions, and translocations that were not previously associated with TEs (≥95% sequence identity detected by the megablast BLASTN algorithm) (Camacho et al. 2009). Query and target sequences were the 100 bp upstream of and downstream from each breakpoint. In addition to macrohomology, we looked for patterns of microhomology. Given that microhomology mechanisms such as microhomology-mediated end joining can require as little as 2 bp of homology, we manually compared the 20 bp of sequence surrounding breakpoints.

Data access

The raw read and ancestor genome assembly data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA839925.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Alexander Suh and Aaron Vogan for valuable discussions on TE classification and mechanisms. PacBio HiFi sequencing and library construction were delivered via the Biotechnology and Biological Sciences Research Council (BBSRC) National Capability in Genomics and Single-Cell Analysis (BBS/E/T/000PR9816) at the Earlham Institute by members of the Genomics Pipelines Group. Analyses performed here made use of the high-performance computing resources at the Edinburgh Compute and Data Facility. 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).

References

Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S. 2017. Genome-wide estimates of transposable element insertion and deletion rates in *Drosophila melanogaster*. *Genome Biol Evol* **9:** 1329–1340. doi:10.1093/gbe/evx050

Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, Fang Q, Xie D, Feng S, Stiller J, et al. 2020. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587:** 246–251. doi:10.1038/s41586-020-2871-y

Bateman AJ. 1959. The viability of near-normal irradiated chromosomes. *Int J Radiat Biol* **1:** 170–180. doi:10.1080/09553005914550241

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:** 66–74. doi:10.1101/gr.219303.116

Belyeu JR, Brand H, Wang H, Zhao X, Pedersen BS, Feusier J, Gupta M, Nicholas TJ, Brown J, Baird L, et al. 2021. *De novo* structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am J Hum Genet* **108:** 597–607. doi:10.1016/j.ajhg.2021.02.012

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573–580. doi:10.1093/nar/27.2.573

Böndel KB, Samuels T, Craig RJ, Ness RW, Colegrave N, Keightley PD. 2022. The distribution of fitness effects of spontaneous mutations in *Chlamydomonas reinhardtii* inferred using frequency changes under experimental evolution. *PLoS Genet* **18:** e1009840. doi:10.1371/journal.pgen.1009840

Boulouis A, Drapier D, Razafimanantsoa H, Wostrikoff K, Tourasse NJ, Pascal K, Girard-Bascou J, Vallon O, Wollman FA, Choquet Y. 2015. Spontaneous dominant mutations in *Chlamydomonas* highlight ongoing evolution by gene diversification. *Plant Cell* **27:** 984–1001. doi:10.1105/tpc.15.00010

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* **3:** lqaa108. doi:10.1093/nargab/lqaa108

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10:** 421. doi:10.1186/1471-2105-10-421

Casas-Mollano JA, Rohr J, Kim EJ, Balassa E, van Dijk K, Cerutti H. 2008. Diversification of the core RNA interference machinery in *Chlamydomonas reinhardtii* and the role of DCL1 in transposon silencing. *Genetics* **179:** 69–81. doi:10.1534/genetics.107.086546

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun* **10:** 4872. doi:10.1038/s41467-019-12884-1

Chaux-Jukic F, O'Donnell S, Craig RJ, Eberhard S, Vallon O, Xu Z. 2021. Architecture and evolution of subtelomeres in the unicellular green alga *Chlamydomonas reinhardtii*. *Nucleic Acids Res* **49:** 7571–7587. doi:10.1093/nar/gkab534

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18:** 71–86. doi:10.1038/nrg.2016.139

Craig RJ. 2021. "The evolutionary genomics of *Chlamydomonas*." PhD thesis, University of Edinburgh, Edinburgh. doi:10.7488/era/1603

Craig RJ. 2022. Replitrons: a new group of eukaryotic transposons encoding HUH endonuclease. bioRxiv doi:10.1101/2022.12.15.520654v1

Craig RJ, Böndel KB, Arakawa K, Nakada T, Ito T, Bell G, Colegrave N, Keightley PD, Ness RW. 2019. Patterns of population structure and complex haplotype sharing among field isolates of the green alga

Chlamydomonas reinhardtii. *Mol Ecol* **28:** 3977–3993. doi:10.1111/mec.15193
Craig RJ, Hasan AR, Ness RW, Keightley PD. 2021a. Comparative genomics of *Chlamydomonas*. *Plant Cell* **33:** 1016–1041. doi:10.1093/plcell/koab026
Craig RJ, Yushenova IA, Rodriguez F, Arkhipova IR. 2021b. An ancient clade of *Penelope*-like retroelements with permuted domains is present in the green lineage and protists, and dominates many invertebrate genomes. *Mol Biol Evol* **38:** 5005–5020. doi:10.1093/molbev/msab225
Craig RJ,Gallaher SD, Shu S, Salomé P, Jenkins JW, Blaby-Haas CE, Purvine SO, O'Donnell S, Barry K, Grimwood J, et al. 2022. The Chlamydomonas Genome Project, version 6: reference assemblies for mating type *plus* and *minus* strains reveal extensive structural mutation in the laboratory. *Plant Cell* doi:10.1093/plcell/koac347
Cridland JM, Thornton KR, Long AD. 2015. Gene expression variation in *Drosophila melanogaster* due to rare transposable element insertion alleles of large effect. *Genetics* **199:** 85–93. doi:10.1534/genetics.114.170837
Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10**. doi:10.1093/gigascience/giab008
De Coster W, Weissensteiner MH, Sedlazeck FJ. 2021. Towards population-scale long-read sequencing. *Nat Rev Genet* **22:** 572–587. doi:10.1038/s41576-021-00367-3
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635
Dobzhansky T, Epling C. 1948. The suppression of crossing over in inversion heterozygotes of *Drosophila pseudoobscura*. *Proc Natl Acad Sci* **34:** 137–141. doi:10.1073/pnas.34.4.137
Faria R, Navarro A. 2010. Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends Ecol Evol* **25:** 660–669. doi:10.1016/j.tree.2010.07.008
Farlow A, Meduri E, Schlötterer C. 2011. DNA double-strand break repair and the evolution of intron density. *Trends Genet* **27:** 1–6. doi:10.1016/j.tig.2010.10.004
Ferenczi A, Chew YP, Kroll E, von Koppenfels C, Hudson A, Molnar A. 2021. Mechanistic and genetic basis of single-strand templated repair at Cas12a-induced DNA breaks in *Chlamydomonas reinhardtii*. *Nat Commun* **12:** 6751. doi:10.1038/s41467-021-27004-1
Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41:** 331–368. doi:10.1146/annurev.genet.40.110405.090448
Flowers JM, Hazzouri KM, Pham GM, Rosas U, Bahmani T, Khraiwesh B, Nelson DR, Jijakli K, Abdrabu R, Harris EH, et al. 2015. Whole-genome resequencing reveals extensive natural variation in the model green alga *Chlamydomonas reinhardtii*. *Plant Cell* **27:** 2353–2369. doi:10.1105/tpc.15.00492
Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, et al. 2018. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* **36:** 875–879. doi:10.1038/nbt.4227
Gilchrist C, Stelkens R. 2019. Aneuploidy in yeast: segregation error or adaptation mechanism? *Yeast* **36:** 525–539. doi:10.1002/yea.3427
Gladyshev EA, Arkhipova IR. 2009. Rotifer rDNA-specific R9 retrotransposable elements generate an exceptionally long target site duplication upon insertion. *Gene* **448:** 145–150. doi:10.1016/j.gene.2009.08.016
Goodwin TJ, Butler MI, Poulter RT. 2003. Cryptons: a group of tyrosine-recombinase-encoding DNA transposons from pathogenic fungi. *Microbiology* **149:** 3099–3109. doi:10.1099/mic.0.26529-0
Goubert C, Craig RJ, Bilat A, Peona V, Vogan AA, Protasio AV. 2022. A beginner's guide to the manual curation of transposable elements. *Mob DNA* **13:** 7. doi:10.1186/s13100-021-00259-7
Gray YH. 2000. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet* **16:** 461–468. doi:10.1016/S0168-9525(00)02104-1
Gregory TR. 2005. Synergy between sequence and size in large-scale genomics. *Nat Rev Genet* **6:** 699–708. doi:10.1038/nrg1674
Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36:** 2896–2898. doi:10.1093/bioinformatics/btaa025
Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29:** 1072–1075. doi:10.1093/bioinformatics/btt086
Halligan DL, Keightley PD. 2009. Spontaneous mutation accumulation studies in evolutionary genetics. *Annu Rev Ecol Evol Syst* **40:** 151–172. doi:10.1146/annurev.ecolsys.39.110707.173437
Han Y, Qin S, Wessler SR. 2013. Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. *BMC Genomics* **14:** 71. doi:10.1186/1471-2164-14-71
Hannan AJ. 2018. Tandem repeats mediating plasticity in health and disease. *Nat Rev Genet* **19:** 286–298. doi:10.1038/nrg.2017.115
Hepler NL, Delaney N, Brown M, Smith ML, Katzenstein D, Paxinos EE, Alexander D. 2016. An improved circular consensus algorithm with an application to detect HIV-1 drug resistance associated mutations (DRAMs). In *Conference on Advances in Genome Biology and Technology*. PacBio, Menlo Park, CA.
Hickey G, Paten B, Earl D, Zerbino D, Haussler D. 2013. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29:** 1341–1342. doi:10.1093/bioinformatics/btt128
Ho EKH, Schaack S. 2021. Intraspecific variation in the rates of mutations causing structural variation in *Daphnia magna*. *Genom Biol Evol* **13:** evab241. doi:10.1093/gbe/evab241
Ho EKH, Bellis ES, Calkins J, Adrion JR, Latta IV LC, Schaack S. 2021. Engines of change: transposable element mutation rates are high and variable within *Daphnia magna*. *PLoS Genet* **17:** e1009827. doi:10.1371/journal.pgen.1009827
Horváth V, Merenciano M, Gonzalez J. 2017. Revisiting the relationship between transposable elements and the eukaryotic stress response. *Trends Genet* **33:** 832–841. doi:10.1016/j.tig.2017.08.007
Inaki K, Liu ET. 2012. Structural mutations in cancer: mechanistic and functional insights. *Trends Genet* **28:** 550–559. doi:10.1016/j.tig.2012.07.002
Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36:** 338–345. doi:10.1038/nbt.4060
Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, et al. 2020. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* **588:** 284–289. doi:10.1038/s41586-020-2947-8
Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bahler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* **8:** 14061. doi:10.1038/ncomms14061
Jeong BR, Wu-Scharf D, Zhang C, Cerutti H. 2002. Suppressors of transcriptional transgenic silencing in *Chlamydomonas* are sensitive to DNA-damaging agents and reactivate transposable elements. *Proc Natl Acad Sci* **99:** 1076–1081. doi:10.1073/pnas.022392999
Joron M, Frezal L, Jones RT, Chamberlain NL, Lee SF, Haag CR, Whibley A, Becuwe M, Baxter SW, Ferguson L, et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477:** 203–206. doi:10.1038/nature10341
Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10:** 19–31. doi:10.1038/nrg2487
Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. *Proc Natl Acad Sci* **114:** E1460–E1469. doi:10.1073/pnas.1616702114
Katju V, Bergthorsson U. 2013. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front Genet* **10:** 273. doi:10.3389/fgene.2013.00273
Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform* **20:** 1160–1166. doi:10.1093/bib/bbx108
Kirkpatrick M. 2010. How and why chromosome inversions evolve. *PLoS Biol* **8**. doi:10.1371/journal.pbio.1000501
Kirkpatrick M, Barton N. 2006. Chromosome inversions, local adaptation and speciation. *Genetics* **173:** 419–434. doi:10.1534/genetics.105.047985
Kojima KK, Fujiwara H. 2005. An extraordinary retrotransposon family encoding dual endonucleases. *Genome Res* **15:** 1106–1117. doi:10.1101/gr.3271405
Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37:** 540–546. doi:10.1038/s41587-019-0072-8
Konkel MK, Batzer MA. 2010. A mobile threat to genome stability: the impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol* **20:** 211–221. doi:10.1016/j.semcancer.2010.03.001
Konrad A, Flibotte S, Taylor J, Waterston RH, Moerman DG, Bergthorsson U, Katju V. 2018. Mutational and transcriptional landscape of spontaneous gene duplications and deletions in *Caenorhabditis elegans*. *Proc Natl Acad Sci* **115:** 7386–7391. doi:10.1073/pnas.1801930115
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 (N Y)* **71:** 2918–2929. doi:10.1111/evo.13360

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

Krasovec M, Lipinska AP, Coelho SM. 2022. Low spontaneous mutation rate in a complex multicellular eukaryote with a haploid-diploid life cycle. bioRxiv doi:10.1101/2022.05.13.491831

Krzywinski M, Schein I, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19:** 1639–1645. doi:10.1101/gr.092759.109

Küpper C, Stocks M, Risse JE, Dos Remedios N, Farrell LL, McRae SB, Morgan TC, Karlionova N, Pinchuk P, Verkuil YI, et al. 2016. A supergene determines highly divergent male reproductive morphs in the ruff. *Nat Genet* **48:** 79–83. doi:10.1038/ng.3443

Kuzmin E, Taylor JS, Boone C. 2022. Retention of duplicated genes in evolution. *Trends Genet* **38:** 59–72. doi:10.1016/j.tig.2021.06.016

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* **109:** E2774–E2783. doi:10.1073/pnas.1210309109

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Li H, Feng X, Chu C. 2020. The design and construction of reference pangenome graphs with minigraph. *Genome Biol* **21:** 265. doi:10.1186/s13059-020-02168-z

López-Cortegano E, Craig RJ, Chebib J, Samuels T, Morgan AD, Kraemer SA, Böndel KB, Ness RW, Colegrave N, Keightley PD. 2021. De novo mutation rate variation and its determinants in *Chlamydomonas*. *Mol Biol Evol* **38:** 3709–3723. doi:10.1093/molbev/msab140

Lower SS, McGurk MP, Clark AG, Barbash DA. 2018. Satellite DNA evolution: old ideas, new approaches. *Curr Opin Genet Dev* **49:** 70–78. doi:10.1016/j.gde.2018.03.003

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biol* **20:** 246. doi:10.1186/s13059-019-1828-7

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* **38:** 4647–4654. doi:10.1093/molbev/msab199

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14:** e1005944. doi:10.1371/journal.pcbi.1005944

McClintock B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci* **36:** 344–355. doi:10.1073/pnas.36.6.344

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585:** 79–84. doi:10.1038/s41586-020-2547-7

Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M, Klein M, Hildebrandt J, Neumann M, Kliebenstein D, et al. 2022. Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature* **602:** 101–105. doi:10.1038/s41586-021-04269-6

Morgan AD, Ness RW, Keightley PD, Colegrave N. 2014. Spontaneous mutation accumulation in multiple strains of the green alga, *Chlamydomonas reinhardtii*. *Evolution (N Y)* **68:** 2589–2602. doi:10.1111/evo.12448

Mukai T. 1964. The genetic structure of natural populations of *Drosophila melanogaster*. I: spontaneous mutation rate of polygenes controlling viability. *Genetics* **50:** 1–19. doi:10.1093/genetics/50.1.1

Muller HJ. 1928. The measurement of gene mutation rate in *Drosophila*, its high variability, and its dependence upon temperature. *Genetics* **13:** 279–357. doi:10.1093/genetics/13.4.279

Nattestad M, Aboukhalil R, Chin CS, Schatz MC. 2021. Ribbon: intuitive visualization for complex genomic variation. *Bioinformatics* **37:** 413–415. doi:10.1093/bioinformatics/btaa680

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

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:** 1739–1749. doi:10.1101/gr.191494.115

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

Ohno S. 1970. *Evolution by gene duplication*. Springer-Verlag, London.

Parks MM, Lawrence CE, Raphael BJ. 2015. Detecting non-allelic homologous recombination from high-throughput sequencing data. *Genome Biol* **16:** 72. doi:10.1186/s13059-015-0633-1

Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36:** 983–987. doi:10.1038/nbt.4235

Potter S, Bragg JG, Blom MP, Deakin JE, Kirkpatrick M, Eldridge MD, Moritz C. 2017. Chromosomal speciation in the genomics era: disentangling phylogenetic evolution of rock-wallabies. *Front Genet* **8:** 10. doi:10.3389/fgene.2017.00010

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Rech GE, Radio S, Guirao-Rico S, Aguilera L, Horvath V, Green L, Lindstadt H, Jamilloux V, Quesneville H, González J. 2022. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*. *Nat Commun* **13:** 1948. doi:10.1038/s41467-022-29518-8

Rhoads A, Au KF. 2015. PacBio sequencing and its applications. *Genom Proteom Bioinform* **13:** 278–289. doi:10.1016/j.gpb.2015.08.002

Ricci M, Peona V, Guichard E, Taccioli C, Boattini A. 2018. Transposable elements activity is positively related to rate of speciation in mammals. *J Mol Evol* **86:** 303–310. doi:10.1007/s00239-018-9847-7

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26. doi:10.1038/nbt.1754

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. 2016. High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Sci Rep* **6:** 28333. doi:10.1038/srep28333

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15:** 461–468. doi:10.1038/s41592-018-0001-7

Sfeir A, Symington LS. 2015. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? *Trends Biochem Sci* **40:** 701–714. doi:10.1016/j.tibs.2015.08.006

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11:** e0163962. doi:10.1371/journal.pone.0163962

Sinha S, Li F, Villarreal D, Shim JH, Yoon S, Myung K, Shim EY, Lee SE. 2017. Microhomology-mediated end joining induces hypermutagenesis at breakpoint junctions. *PLoS Genet* **13:** e1006714. doi:10.1371/journal.pgen.1006714

So A, Le Guen T, Lopez BS, Guirouilh-Barbat J. 2017. Genomic rearrangements induced by unscheduled DNA double strand breaks in somatic mammalian cells. *FEBS J* **284:** 2324–2344. doi:10.1111/febs.14053

Song JM, Guan Z, Hu J, Guo C, Yang Z, Wang S, Liu D, Wang B, Lu S, Zhou R, et al. 2020. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat Plants* **6:** 34–45. doi:10.1038/s41477-019-0577-7

Sui Y, Qi L, Wu JK, Wen XP, Tang XX, Ma ZJ, Wu XC, Zhang K, Kokoska RJ, Zheng DQ, et al. 2020. Genome-wide mapping of spontaneous genetic alterations in diploid yeast cells. *Proc Natl Acad Sci* **117:** 28191–28200. doi:10.1073/pnas.2018633117

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci* **109:** 18488–18492. doi:10.1073/pnas.1216223109

Tan S, Cardoso-Moreira M, Shi W, Zhang D, Huang J, Mao Y, Jia H, Zhang Y, Chen C, Shao Y, et al. 2016. LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Res* **26:** 1663–1675. doi:10.1101/gr.204925.116

Tusso S, Suo F, Liang Y, Du LL, Wolf JBW. 2022. Reactivation of transposable elements following hybridization in fission yeast. *Genome Res* **32:** 324–336. doi:10.1101/gr.276056.121

Uzunović J, Josephs EB, Stinchcombe JR, Wright SI. 2019. Transposable elements are important contributors to standing variation in gene expression in *Capsella grandiflora*. *Mol Biol Evol* **36:** 1734–1745. doi:10.1093/molbev/msz098

van Dijk K, Xu H, Cerutti H. 2006. Epigenetic silencing of transposons in the green alga *chlamydomonas reinhardtii*. In *Small RNAs: analysis and regulatory functions* (ed. Nellen W, Hammann C), pp. 159–178. Springer, Berlin, Germany. doi:10.1007/978-3-540-28130-6_8

van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ. 2016. The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534:** 102–105. doi:10.1038/nature17951

Villalba de la Peña M, Summanen PAM, Liukkonen M, Kronholm I. 2022. Chromatin structure influences rate and spectrum of spontaneous mutations in *Neurospora crassa*. bioRxiv doi:10.1101/2022.03.13.484164

Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14:** 125–138. doi:10.1038/nrg3373

Weissensteiner MH, Bunikis I, Catalán A, Francoijs KJ, Knief U, Heim W, Peona V, Pophaly SD, Sedlazeck FJ, Suh A, et al. 2020. Discovery and

population genomics of structural variation in a songbird genus. *Nat Commun* **11:** 3403. doi:10.1038/s41467-020-17195-4
Weissman JL, Fagan WF, Johnson PLF. 2019. Linking high GC content to the repair of double strand breaks in prokaryotic genomes. *PLoS Genet* **15:** e1008493. doi:10.1371/journal.pgen.1008493
Yoder AD, Tiley GP. 2021. The challenge and promise of estimating the de novo mutation rate from whole-genome comparisons among closely related individuals. *Mol Ecol* **30:** 6087–6100. doi:10.1111/mec.16007
Yun T, Li H, Chang PC, Lin MF, Carroll A, McLean CY. 2020. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36:** 5582–5589. doi:10.1093/bioinformatics/btaa1081
Zhang C, Wu-Scharf D, Jeong BR, Cerutti H. 2002. A WD40-repeat containing protein, similar to a fungal co-repressor, is required for transcriptional gene silencing in *Chlamydomonas*. *Plant J* **31:** 25–36. doi:10.1046/j.1365-313X.2002.01331.x
Zhang J, Yu C, Pulletikurti V, Lamb J, Danilova T, Weber DF, Birchler J, Peterson T. 2009. Alternative *Ac/Ds* transposition induces major chromosomal rearrangements in maize. *Genes Dev* **23:** 755–765. doi:10.1101/gad.1776909
Zhang X, Zhao M, McCarty DR, Lisch D. 2020. Transposable elements employ distinct integration strategies with respect to transcriptional landscapes in eukaryotic genomes. *Nucleic Acids Res* **48:** 6685–6698. doi:10.1093/nar/gkaa370
Zhang L, Chaturvedi S, Nice CC, Lucas LK, Gompert Z. 2022. Population genomic evidence of selection on structural variants in a natural hybrid zone. *Mol Ecol* doi:10.1111/mec.16469
Zhao H, Zhang W, Chen L, Wang L, Marand AP, Wu Y, Jiang J. 2018. Proliferation of regulatory DNA elements derived from transposable elements in the maize genome. *Plant Physiol* **176:** 2789–2803. doi:10.1104/pp.17.01467

Received May 23, 2022; accepted in revised form December 6, 2022.