Mitotic recombination between homologous chromosomes drives genomic diversity in diatoms

Petra Bulánková[1,2,10]*, Mirna Sekulić[1,2,3,12], Denis Jallet[4,12], Charlotte Nef[5,12], Cock van Oosterhout[6,] Tom O. Delmont[7], Ilse Vercauteren[1,2], Cristina Maria Osuna-Cruz[1,2,8], Emmelien Vancaester[1,2,8,9], Thomas Mock[6], Koen Sabbe[3], Fayza Daboussi[4], Chris Bowler[5], Wim Vyverman[3], Klaas Vandepoele[1,2,8] and Lieven De Veylder[1,2,10,11]*

Affiliations:

[1] VIB Center for Plant Systems Biology, Technologiepark 71, 9052, Ghent, Belgium.

Petra Bulánková, Mirna Sekulić, Ilse Vercauteren, Cristina Maria Osuna-Cruz, Emmelien Vancaester, Klaas Vandepoele and Lieven De Veylder

[2] Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 71, 9052, Ghent, Belgium.

Petra Bulánková, Mirna Sekulić, Ilse Vercauteren, Cristina Maria Osuna-Cruz, Emmelien Vancaester, Klaas Vandepoele and Lieven De Veylder

[3] Protistology and Aquatic Ecology, Department of Biology, Ghent University, 9000, Ghent, Belgium.

Mirna Sekulić, Koen Sabbe and Wim Vyverman

[4] TBI , Université de Toulouse, CNRS, INRAE, INSA , 135 avenue de Rangueil F-31077, Toulouse , France.

Denis Jallet, Fayza Daboussi

[5] Institut de Biologie de l'École Normale Supérieure (IBENS), École Normale Supérieure, CNRS, INSERM, PSL Université Paris, 75005 Paris, France.

Charlotte Nef, Chris Bowler

[6] School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK.

Cock Van Oosterhout, Thomas Mock

[7] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91000 Evry, France.

Tom O. Delmont

[8] Bioinformatics Institute Ghent, Ghent University, Technologiepark 71, 9052, Ghent, Belgium.

Cristina Maria Osuna-Cruz, Emmelien Vancaester and Klaas Vandepoele

[9] Present address: Tree of Life, Wellcome Sanger Institute, Cambridge, CB10 1SA, UK

[10] Senior author

[11] Lead contact

[12] These authors contributed equally to this work

*Correspondence to: lieven.deveylder@psb.vib-ugent.be or petra.bulankova@gmail.com

SUMMARY

Diatoms, an evolutionarily successful group of microalgae, display high levels of intraspecific genetic variability in natural populations. However, the contribution of various mechanisms generating such diversity is unknown. Here we estimated the genetic micro-diversity within a natural diatom population and mapped the genomic changes arising within clonally propagated diatom cell cultures. Through quantification of haplotype diversity by next-generation sequencing and amplicon re-sequencing of selected loci, we documented a rapid accumulation of multiple haplotypes accompanied by the appearance of novel protein variants in cell cultures initiated from a single founder cell. Comparison of the genomic changes between mother and daughter cells revealed copy number variation and copy-neutral loss of heterozygosity leading to the fixation of alleles within individual daughter cells. The loss of heterozygosity can be accomplished by recombination between homologous chromosomes. To test this hypothesis, we established an endogenous read-out system and estimated that the frequency of interhomolog mitotic recombination to be under standard growth conditions 4.2 events per 100 cell divisions. This frequency is increased under environmental stress conditions, including treatment with hydrogen peroxide and cadmium. These data demonstrate that copy number variation and mitotic recombination between homologous chromosomes underlie clonal variability in diatom populations. We discuss the potential adaptive evolutionary benefits of the plastic response in the interhomolog mitotic

recombination rate, and we propose that this may have contributed to the ecological success of diatoms.

INTRODUCTION

Diatoms, with as many as 100 000 estimated species, colonize a wide range of marine, freshwater and terrestrial environments[1]. Given their often vast census population size, diatoms possess high intraspecific genetic variation. Natural diatom population samples comprise between 87 to 100% clonal diversity as measured by microsatellite markers and a gene diversity ranging from 39 to 88%, suggesting that clonal lineages are significantly diverged [2]. However, many species reproduce asexually for long periods, and hence, some of this gene diversity is present as clonal diversity[2]. For example, the model diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* have never been observed to produce F1 progeny[3,4] and in many other diatom species, sex is restricted by cell size[5]. Due to the mechanics of diatom cell wall formation from inorganic silica, the cell size of mitotically dividing cells in a population decreases over time with only cells under a species-specific size threshold being sexually competent. The period necessary to reach this threshold can last months to years[3,6-8]. So how do diatoms generate novel genetic variation that is required for adaptive evolution?

Genetic variation is generated by three evolutionary forces: mutation, recombination and gene flow. Ultimately, all novel genetic variation stems from mutations, whereas the two other forces merely shuffle this variation between haplotypes or genotypes (i.e. recombination), or between populations (gene flow)[9,10]. Mutation rates in diatoms appear to be at a similarly low level as in green algae[11] and coccoliths[12], but given their often large population size, considerable variation can be generated by the input of new mutations[13,14]. However, not all variation generated in a clonal lineage is readily available to selection due to recessivity and clonal interference, i.e. the competition between different beneficial mutations that occur in individuals within the same clonal lineage[15,16]. Potentially beneficial variation can become “locked inside” in a poorly adapted genotype, and sexual reproduction can relax this evolutionary constrained. The random assortment of alleles from both parents during meiotic recombination[17,18] generates novel genotypes, which form the substrate for natural selection.

In the absence of sex, mitotic interhomolog recombination can also generate novel genotypes. Here, we define mitotic interhomolog recombination as the genetic exchange between the haplotypes of the same individual that occurs in vegetative cells, when the homologous chromosome is used as a template for homologous recombination, including both crossing over and gene conversion events. As this can result in potentially harmful loss of heterozygosity (LOH), chromosome rearrangements and consecutively mosaicism leading to the onset of cancer in multicellular organisms, such recombination is typically strongly suppressed in nearly all studied eukaryotes[19,20]. However, mitotic recombination is common to many asexual and facultatively sexual species, including yeast[21,22], ascomycete fungi[23], and oomycetes[24].

Here, we show that novel variation can be generated in diatom clonal populations through mitotic recombination. We discuss how such variation may benefit adaptive evolution under exposure to stress, and we hypothesize about the role this could play in the evolutionary dynamics during clonal competition.

RESULTS

Intraspecific SNV variability within a natural diatom population

Previously, microsatellite-based approaches have demonstrated a high level of intraspecific genetic variability in natural diatom populations [2]. However, as there is no good evidence of genome-wide intraspecific diversity, we first sought to quantify the extent of genome-wide intraspecific variability within natural diatom populations in situ using genome-wide metagenomic read recruitments from the Tara Oceans expeditions [1] to expand on previous knowledge. Diatom models commonly considered for fundamental genomic analyses and that are easily transformable, such as *Phaeodactylum tricornutum*, are however poorly retrieved in global environmental datasets such as the *Tara* Oceans metagenomes[1]. The most abundant diatom genera in terms of assigned 18S rRNA V9 rDNA reads in this dataset belong to *Chaetoceros* and *Fragilariopsis*, which account, respectively, for 23.1% and 15.5% of the total number of reads. While there are currently no whole-genome sequences from any *Chaetoceros* species, the genome of *Fragilariopsis cylindrus* is available and presents elevated genomic variability with around 25% of its diploid genome corresponding to highly divergent loci[25]. This genome was therefore chosen to explore diatom genomic variability in situ in the environment. We recruited metagenomic reads from *Tara* Oceans metagenomes using the *F. cylindrus* reference genome (mapping stringency >95% identity) and examined micro-diversity traits at the level of single nucleotide variants (SNVs)[26-28]. We were able to retrieve a large number of environmental sequences to this genome from Station 86 in the Southern Ocean (near the Antarctic peninsula, 64°30'88" S, 53°05'75" W), from both the surface (5m depth; mean coverage of 51.7X over genome) and deep chlorophyll maximum (DCM; 35m depth; mean coverage of 58.46X) layers. Overall, 89.64% (24,326) of *F. cylindrus* genes (total of 30.95 Mb) displayed coverage values similar to the entire genome in these two metagenomes and were considered for downstream analyses (see Methods). Within the scope of these genes, we identified 619,947 and 592,929 SNVs in the surface and DCM metagenomes, respectively (Data S1), which corresponds to an SNV density of ~2% (i.e., one SNV every 50 nucleotides). All the genes contained at least one SNV, with SNV density ranging from ~0.02% to ~10% (Data S1). Among these, 3,822 of these genes displayed SNV density <1% in both metagenomes. This analysis suggests that the average nucleotide identity of the genomes considered is about 98%, supportive of the existence of a single population displaying thousands of micro-diversity genomic traits. Parallel metabarcoding-based surveys based on 18S rRNA revealed sequences most homologous to *F. cylindrus* in the samples from Station 86. Among the competing nucleotides in the metagenomes, A-G and C-T transitions each contributed 30% of SNVs, followed by the transversions A-C (13%), G-T (13%), A-T (10%) and C-G (4%). These statistics were highly similar for the two metagenomes, with a comparable transition to transversion ratio of around 1.5. Yet, of all the SNVs identified, only 429,530 (54.83%) were common to the two metagenomes (Figure S1). We, therefore, conclude from this analysis that natural populations of diatoms can harbour a large amount of micro-diversity, which is not restricted to microsatellites but is present genome-wide.

Genome-wide haplotype diversity

In natural populations, the impact of sexual reproduction cannot be ruled out, and given the vast census population size, their high nucleotide diversity is perhaps not surprising. However, previous studies have hinted at genomic variability within laboratory clones, such as extensive allelic diversity in the pennate polar diatom *F. cylindrus*[25] and differences between *P. tricornutum* cultures belonging to the same strain derived originally from a single cell[29]. Correspondingly, we noticed the presence of multiple haplotypes instead of the expected two when sequencing various genomic loci in cultures clonally grown from a single cell of *P. tricornutum* as well as of *Seminavis robusta*, grown under conditions that preclude sexual reproduction. To map the distribution of loci with multiple haplotypes in these diatom species, we took advantage of two available genome-wide datasets: short-read Illumina sequencing was used to identify a set of reliable SNPs (single-nucleotide polymorphism, present in at least 20% of reads) and PacBio and MinION long-read sequencing were used to identify the number of haplotypes in *S. robusta* and *P. tricornutum,* respectively (Figure S2). To decrease the error rate in long read sequencing, we used PacBio Circular Consensus Sequences (CCS) reads and canu[30] for self-correction of both PacBio and MinION reads. Next, we removed repeat regions and counted the number of combinations formed by confident SNPs in individual reads in 1kb windows and selected loci with at least three haplotypes supported each by a minimum of two reads. This analysis uncovered 1,405 of such loci in *S. robusta* (125.6 Mb genome size) and 3,380 loci in *P. tricornutum* (27.4 Mb) (Figure 1A-D, Table 1, Figure S3, Data S2). To examine whether the number of uncovered haplotypes could be caused by a high error rate in long-read sequencing datasets, we performed an equivalent counting in available datasets from a haploid *Saccharomyces cerevisiae* (12 Mb) culture derived from a single cell[31] and diploid *Arabidopsis thaliana* (135 Mb) datasets derived from multiple inbred plants[32,33], where loci with multiple haplotypes are not expected. This yielded only 3 and 83 loci with multiple haplotypes in *S. cerevisiae* and *A. thaliana*, respectively (Data S2).

Accumulation of novel haplotypes

Genome-wide haplotype counting via long-read sequencing can suffer from increased sequencing noise and as the datasets were derived from cultures with different cultivation history, we could not conclude on the rate of the appearance of novel haplotypes. We, therefore, validated the genome-wide data by observing selected loci from newly isolated, single-cell cultures (Figure S4). For *S. robusta*, we profiled three loci identified in the genome-wide haplotype analysis in three independent cultures, four months after single diploid cell isolation. To overcome the potential problem of artefact generation during DNA amplification, we used emulsion PCR followed by Sanger sequencing of cloned PCR products. While the control mixture of two different alleles returned the two original haplotypes after PCR, we observed 2 to 6 haplotypes for the endogenous *S. robusta* loci (Figure 2A, Table 2) by manually examining the combinations of reliable SNPs in individual Sanger sequencing reads. Due to the low efficiency of emulsion PCR reactions, we were not able to sequence the founder cell. However, in every case, two prominent haplotypes were supported by a higher number of reads, possibly representing the haplotypes present in the founder cell, whereas the additional haplotypes presumably appeared during the four months in culture.

Independently, haplotype diversity and the rate at which new haplotypes appeared were analyzed for 62 *P. tricornutum* 2-kb loci using emulsion PCR followed by PacBio amplicon sequencing. Five loci (G32 to G36) were amplified at 1 month (T1) and 6 months (T6) after single-cell isolation, whereas the remaining 57 loci were amplified at T6 only. The heterozygosity of selected loci was profiled by SNP calling on the culture used for amplification at T1. Again, we used the short-read sequencing dataset to identify reliable SNPs and counted the number of haplotypes per locus formed by their combinations in corrected PacBio amplicon reads. The control reactions for random errors and artificial haplotype detection yielded the expected one and two haplotypes, respectively, demonstrating that the emulsion PCR, PacBio library preparation, and sequencing did not generate artefacts (Figure S4, Data S3). The number of recovered haplotypes varied between 1 and 15 (Figure 2C, Figure S4 and Data S3), with 6 loci displaying a single haplotype, 5 loci with two haplotypes, and 51 loci displaying at least three haplotypes. For four out of five loci amplified at both T1 and T6, an increase in the number of haplotypes was observed in the T6 sample (Figure 2B, Figure S4) despite deeper sequencing coverage of the T1 samples, suggesting that haplotypes accumulate over time (Table 3). We analyzed the impact of haplotype variability on protein sequence in 20 genes fully covered by amplicon sequencing and found six for which the different haplotypes resulted in more than two putative protein variants, with up to six variants in the diatom-specific gene *Phatr3_J47122* (Figure 2D, Figure S4, Table S1).

Although only one locus was identified as homozygous by SNP calling in T1, five additional loci with single haplotypes were found in T6 sequencing (Figure 2C, Figure S4). These loci were identified as being heterozygous by SNP calling in T1, suggesting a loss of heterozygosity (LOH) [34]. Moreover, the novel haplotypes that accumulated in both *S. robusta* and *P. tricornutum* cultures were recombinants lacking *de novo* mutations. Such new combinations are typically generated during sexual reproduction through the meiotic recombination of homologous chromosomes[35].

Genome-wide detection of loss of heterozygosity and copy number variation

Although interhomolog recombination is rare in vegetative cells, we tested whether it could be the source of haplotype diversity in clonal diatom populations as sexual reproduction was excluded in our cultures. We sought to detect LOH and copy number variation (CNV) events in *P. tricornutum* under controlled conditions over a defined number of cell divisions. Three independent mother cultures (MC1 - MC3) were initiated from a single cell isolate and cultivated under conditions allowing approximately a single cell division per day (Figure 3A, Figure S5). After 30 days (T1), three single cells were again isolated from each mother culture to obtain nine daughter cultures (DC1.1 - DC3.3) that were harvested 30 days later (T2). At both T1 and T2, part of the mother cultures was also harvested. Following genome resequencing and SNP calling of all cultures, a pairwise comparison between the individual daughter cultures and their respective mother cultures was performed to identify novel CNVs and tracts of at least three consecutive SNPs that were lost in the daughter culture.

Changes in comparison with the mother cultures were found in four out of nine daughter cells. One copy-neutral 8016 bp LOH, where one allele of the locus was replaced by the other allele, was observed in DC1.2, three copy-neutral LOH events (296 bp, 614 bp and 1644 bp in length) and a 31.4 kb duplication covering 14 genes were observed in DC1.3 culture, and one 30.9 kb and one 156.9 kb deletion were detected in cultures DC3.1 and DC2.1, respectively, and were confirmed by Sanger re-sequencing or qPCR (Figure 3, Figure S5, Table S2, Data S3). Besides the LOH events that were unique to a respective daughter culture, we identified several regions with reduced SNP density common to all cultures. SNP density was ten times lower than the genome average over almost the entire chromosome 19, and seventeen and thirty-eight times lower at the extremities of chromosomes 27 and 28, respectively, in comparison with the rest of the chromosome (Figure S5). These regions were not found to be SNP poor when sequencing the same *P. tricornutum* strain from other laboratories[36-38].

Profiling of the functional effect of 2914 SNPs in LOH regions revealed 59 SNPs (0.362%) with possible high impact on gene function, 650 (3.984%) with low, 702 (4.303 %) with moderate and 14,903 (91.351%) with modifier effect according to SnpEff categorization[39]. Most SNPs with a high effect on protein function were found in the 156.9 kb deletion on chromosome 26. This deletion was identified in the primary analysis as six LOH regions and confirmed as a single deletion only after Sanger resequencing of LOH border regions. Therefore, we were only able to analyze the effect of the 19 SNPs with high effect found in the regions identified in the primary LOH analysis (Data S3) and found 3 SNPs that caused a loss of function by introducing a premature stop codon in the respective gene (Data S3).

Copy-neutral loss of heterozygosity at the PtUMPS locus

While the mechanism behind the observed deletions and duplication remains difficult to interpret, the copy-neutral LOH events require an exchange of genetic information between homologous chromosomes. To estimate the rate of interhomolog recombination in *P. tricornutum*, we established a tractable endogenous readout system for copy-neutral LOH detection, based on three strains containing two different mutant alleles of the *PtUMPS* gene, generated through gene editing[40,41]. In strain *ptumps-1bp,* the 1 bp indel mutations in the two alleles occur at a position only 1 bp apart, in strain *ptumps-320bp* they are separated by 320 bp and in strain *ptumps-1368bp* by 1368 bp (Figure 4A). As the *PtUMPS* protein is required for uracil biosynthesis, cells with a wild-type (WT) allele can synthesize uracil, but also convert 5-fluoroorotic (5-FOA) acid into the toxic 5-fluorouracil (5-FU), resulting in cell death. In contrast, mutant cells are resistant to 5-FOA but are uracil auxotrophs. The *ptumps-/-* strains were cultivated under non-selective conditions for 14 days (with uracil and without 5-FOA) to permit potential recombination at the *PtUMPS* locus (Figure S6). Subsequently, 5x10^7 cells from the culture were plated on a medium without uracil to select cells that underwent recombination at the *PtUMPS* locus and restored the WT allele. We recovered no colonies in strain *ptumps-1bp*, confirming that the WT allele was not restored by a random mutation, 12 colonies in strain *ptumps-320bp* and 83 colonies in strain *ptumps-1368bp* (Figure 4B, Data S4-S5). Moreover, sequencing of *PtUMPS* alleles from ten *ptumps-1368bp* colonies and five *ptumps-320bp* colonies corroborated the restoration of the WT allele through copy-neutral LOH events (Figure 4C, Figure S6).

Next, the *PtUMPS* system was used to obtain an estimate of the interhomolog recombination frequency. A total of $2x10^7$ cells per replica from 5-FOA- and uracil-supplemented medium (preventing recombination at the *PtUMPS* locus) were directly plated onto medium without uracil to select only those cells that were in the process of interhomolog recombination during a single round of cell division. The average frequency of interhomolog recombination was 4.2 per 100 cell divisions per genome (Figure 4D, Data S4-S5), approximately ten times higher than the rate reported for *S. cerevisiae* after recalculation per cell division[42,43].

To test whether the rate of interhomolog recombination can be influenced by environmental conditions, we employed the *PtUMPS* readout system to test the effect of the DNA double-strand break inducing drug zeocin[44] and three physiologically relevant stresses: hydrogen peroxide (H_2O_2), which is produced by various phytoplankton groups and can act as a signalling molecule as well as cause oxidative damage[45], the trace metal cadmium[46], which contaminates aquatic environments, and a polyunsaturated aldehyde (E,E)-2,4-Decadienal that is involved in diatom intercellular signalling, stress surveillance, and defence against grazers, but which can trigger lethality at high concentrations[47,48]. For each mock and stress treatment, $25x10^6$ cells per replica were transferred from 5-FOA- and uracil-supplemented medium to medium containing uracil for 24 h, thus allowing a maximum of one cell division. Next, cells were plated on a selective medium without uracil to recover cells that restored the WT *PtUMPS* allele through interhomolog recombination. Only the zeocin treatment resulted in the appearance of uracil prototrophic colonies in both *ptumps-1bp* and *ptumps-1368bp* in a dose-dependent manner (Figure 4E). Sequencing of *ptumps-1bp* colonies revealed restoration of the *PtUMPS* WT allele through *de novo* mutations (Figure S6). We thus suppose that zeocin treatment induced robust DNA damage. No *ptumps-1bp* colonies were observed in the other treatments hinting at a lack of *de novo* mutations. Whereas the (E, E)-2,4-Decadienal treatment did not influence the rate of interhomolog recombination, we found a positive, concentration-dependent effect of H_2O_2 and cadmium on the number of recovered colonies (Figure 4F-G, Data S4-S5). These data illustrate that environmental stresses increase the frequency of recombination between homologous chromosomes.

DISCUSSION

Analysis of environmental samples of subpopulations of the diatom *F. cylindrus* showed that they harbour extensive genome-wide SNV diversity. However, as *F. cylindrus* is not easily accessible to genome manipulation methods, we investigated the possible underlying mechanism in commonly used diatom model species. By following the number of haplotypes in cultures of *S. robusta* and *P. tricornutum* initiated from a single cell, we documented that both diatom species rapidly accumulate recombined haplotypes throughout the genome. The resulting novel SNP combinations in protein-coding genes can give rise to novel protein variants that were not present in the founder cell, potentially contributing to the physiological divergence of individual subclones in the clonal population. Additionally, a comparison of genomic changes between mother cultures and their respective daughter cultures revealed the appearance of copy-neutral LOH and CNV events over a brief period. We hypothesize that CNVs arise from ectopic recombination or non-homologous end-joining[49-51]. In the copy-neutral LOH events, the information from the homologous chromosome either replaces the original allele in case of a gene conversion event, or it leads to reciprocal exchange in case of mitotic crossing over. Subsequent sister chromatid segregation during mitosis may cause LOH tracts in the daughter cell(s), resulting in the fixation of polymorphisms in a homozygous state[19,20], further contributing to phenotypic differences within the clonal population.

Estimating the rate of the mitotic interhomolog recombination per cell division revealed that the frequency in *P. tricornutum* exceeds by ten times the frequency in the yeast *S. cerevisiae*, the key model in mitotic recombination research [42,43]. Although this comparison does not take into account the possible differences between the diatom and yeast outcomes of the interhomolog recombination, it suggests that such recombination is highly common in *P. tricornutum* and that the constraints preventing the use of homologous chromosomes as a template for homologous recombination might be relaxed in diatoms. The capability to rapidly fix novel SNVs in a population through LOH could explain the differences observed in the metagenomes of *F. cylindrus* subpopulations of the surface and DCM samples from the same station.

We demonstrated that the rate of mitotic recombination increased under environmental stress, which suggests that it has a degree of phenotypic plasticity. A similar increase was documented for both meiotic and mitotic recombination in various organisms including yeasts[52,53], plants[54-56] and metazoans[57-59] and has important implications for evolution. Recombination related genomic changes were shown to shape the genomes of pathogenic fungi such as *Candida albicans*, where the frequency increases under stress and during host infections and contributes to the fitness advantage of resulting clones[60], as well as in the oomycete *P. ramorum* where extensive runs of homozygosity (ROH) differentiate individual invasive lineages[24]. However, in both pathogenic species, the recombination involved preferentially the repetitive regions and transposons, and the exact frequency is not known. By contrast, repetitive sequences were excluded in our analysis, and recombinant haplotypes were found to be equally dispersed throughout the *P. tricornutum* genome.

Besides *de novo* LOH events, the analysis of mother and daughter cells revealed the presence of regions with low SNP density common to all sequenced strains. These regions were not detected as being low in SNP content in *P. tricornutum* Pt1 strains from other laboratories [36,37] and a similar situation was reported for other genomic regions[29]. Low heterozygosity regions can also arise due to inbreeding or purifying selection at a linked genetic loci. However, *as P. tricornutum* has never been observed to reproduce sexually in laboratory conditions, we propose that these might represent past LOH events in the ancestor cell of the respective population. Low SNP density regions have also been observed in presumably asexual isolates of the centric diatom *Thalassiosira pseudonana*[4]. It was speculated that the loss of heterozygosity in these regions due to inbreeding resulted in the fixation of mutations in genes required for sexual reproduction. In the light of high levels of mitotic interhomolog recombination in *P. tricornutum*, an alternative cause of the decrease in heterozygosity could be LOH accompanying such recombination.

Many diatom species accomplish rapid population expansion through clonal reproduction[61]. During this phase of exponential growth, a small difference in fitness can have a large effect on the eventual population number reached by each clonal lineage. However, significant environmental changes are likely to deteriorate the fitness of any well-adapted lineage. Yet, without sexual reproduction, each clonal lineage is limited in its adaptive response by the variation contained within its genome. We hypothesize that mitotic recombination can exploit the non-additive genetic variation (i.e. dominance and epistatic variation) that is present within each genome but hidden from natural selection[62,63].

Plastic response in mitotic recombination could offer at least three important fitness advantages during clonal competition. Firstly, the evolution of asexual microbes is generally not limited by the number of beneficial single-point mutations, but rather, by overcoming clonal interference and combining multiple mutations into a single genotype. For example, beneficial mutations are readily available in yeast, but they compete with one another in the population for fixation[64-66]. Mitotic recombination can relax the evolutionary constraints imposed by clonal interference, by generating novel combinations of alleles. Alleles can thus be 'tried and tested' against slightly different genomic backgrounds, which increases the probability of finding a superior combination of multiple mutations. Mitotic recombination thus can not only uncover hidden dominance variation by making loci homozygous, but it can also reveal epistatic variation by creating novel allelic combinations that would otherwise not have arisen. This could be particularly important during periods of environmental change and stress, enabling the clonal lineage to discover other fitness peaks in a dynamic fitness landscape[67].

Secondly, density and frequency-dependent processes are likely to regulate clonal expansion. Lewontin put this succinctly: "a genotype is its own worst enemy, its fitness will decrease as it becomes more common"[68]. Such negative frequency dependence is likely to play an important role in asexual species, particularly during and after clonal expansion. The ability to generate novel genotypes during mitotic recombination could mitigate this effect, reducing the competition between clone-mates, and generating a more diffuse target for antagonistically coevolving species, such as pathogens.

Thirdly, generating evolutionary novelty, either through mutation or recombination, does impose a fitness cost to the individual or clonal lineage, i.e. a genetic load[69]. In a well-adapted genotype, each mitotic recombination event is more likely to reduce fitness than to increase it. However, occasionally, some mitotic recombination events could be selectively advantageous, and this is more likely to occur if the clonal genotype is not optimally adapted to its environment[22]. In diatoms such as *P. tricornutum*, natural selection thus trades off the costs of the ‘mitotic recombination load’ against the potential benefits realized by such recombination. These benefits include reducing possible negative frequency-dependent effects, and uncovering hidden dominance and epistatic variance, enabling the genotype to climb or discover a fitness peak in the adaptive landscape. In other words, there may be an optimum level of mitotic recombination, depending on the stability of the environment, the match between the phenotype and the environment, and the amount of negative frequency-dependent selection. We hypothesize that phenotypic plasticity may enable lineages to track this optimal level of mitotic recombination, with natural selection favouring an increased rate under stressful environmental conditions. Alternatively, stressful environmental conditions could increase the mitotic recombination rate, for example by impairing DNA repair mechanisms. The question not answered by our experiments is whether the observed increase in mitotic recombination during stress is adaptive. We propose this is plausible, and that this hypothesis provides an interesting avenue for future research.

ACKNOWLEDGMENTS

We thank Dr. Nicole Poulsen for providing GK-359 and GK-333 plasmids and Dr. Annick Bleys and Dr. James Matthew Watson for proofreading the manuscript. This research was supported by BOF project GOA01G01715 to L.D.V., K.V. and W.V.; Erwin Schrödinger fellowship from Austrian Science Fund (project J3692-B22) to P.B.; Gordon and Betty Moore Foundation grant (GBMF 4966), a Région Midi-Pyrénées grant (15058490 financial support for Accueil d’Equipes d’Excellence), an ANR JCJC grant (ANR-16-CE05-0006-01), and the 3BCAR Carnot Institute funding to D.J. and F.D.; C.N. and C.B. were supported by European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program through the project DIATOMIC (grant agreement No. 835067) to C.B.

AUTHOR CONTRIBUTIONS

P.B. and L.D.V. conceptualized the study. P.B. initiated, designed and performed experiments and bioinformatics analysis on *P. tricornutum* and *S. robusta*. M.S. performed sequencing of uracil prototrophic colonies. D.J. generated *ptumps* mutant strains. C.N and T.D. performed the *F. cylindrus* metagenome assembly and analysis, I.V. helped with diatom culture maintenance. C.M.O-C. and E.M. provided bioinformatic datasets. P.B. and C.V.O. wrote the manuscript, generated all figures and data visualizations. L.D.V., K.V., F.D., C.B., W.V., K.S. supervised the research. P.B., L.D.V, K.V., W.V., F.D, K.S., C.B., C.V.O., T.M. reviewed and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

MAIN FIGURE TITLES AND LEGENDS

Figure 1. Genome-wide distribution of haplotypes in *P. tricornutum* and *S. robusta*. (A-B) Distribution of the detected number of haplotypes per 1 kb loci in *S. robusta* contigs above 20 kb (A) and *P. tricornutum* chromosomes 1-33 (B). From outside to inside: loci with more than two haplotypes (red), loci with two haplotypes (blue), loci with a single haplotype (grey), gene density*, SNP density* (black), GC content*; * per 10 kb. (C-D) Example of a representative genomic region from *S. robusta* (C) and *P. tricornutum* (D). The chromosome is represented by a grey rectangle. Above the chromosome: loci with single haplotype (grey bars), loci with two haplotypes (blue bars) and loci with more than two haplotypes (red bars). Below

chromosome from top to bottom: gene density**, SNP density** (grey line), GC content**; ** per 1 kb. See also Figures S1-S3 and Data S1-S2.

Figure 2. Accumulation of novel haplotypes in cultures freshly initiated from a single cell. (A) Quantification of haplotypes on three loci in three *S. robusta* cultures (SR1-3) four months after cultivation from a single cell. (B-C) Quantification of the number of haplotypes in *P. tricornutum* at 1 month (T1) and 6 months (T6) after cultivation from a single cell. (B) Quantification of the number of haplotypes at loci G32-G36 detected at T1 and T6. (C) Quantification of the number of haplotypes detected at T6 (orange outline). Categories of the shift in the number of haplotypes from founder cell to the number of haplotypes detected at T6 are on the x-axis in the format: expected in founder cell > observed at T6, the number of haplotypes on the y-axis. The size of the circle corresponds to the number of cases in a given category. In (A-C) the expected two haplotypes in the founder cells are indicated in blue. (D) Schematic representation of predicted proteins variants in Phatr3_J47122 gene. The top line shows the position of amino acid variants on the protein indicated by red flags. Green regions depict conserved domains according to the CDD/SPARKLE database [70]. The lines below represent individual predicted variants. See also Figures S4, Table S1 and Data S3.

Figure 3. Genome-wide detection of LOH in *P. tricornutum* mother and daughter cultures after 30 days. (A) Position of copy-neutral LOHs (orange), duplication (dark blue) and deletions (red) in individual daughter cultures. Heterozygous regions are in light blue, blank space – no SNPs. Nominators at the left refer to the daughter cell (DC) culture, whereas the lowercase digit indicates the chromosome number. (B) Zoomed-in regions with detected LOH,

duplication or deletion events in the respective mother and daughter cultures (DC). Blue dots – heterozygous SNPs, orange dots – homozygous SNPs in copy-neutral LOHs, red dots - SNPs in deletions, grey area – sequence coverage. (C) Confirmation of a DNA duplication event on chromosome 23 by qPCR in daughter culture DC1.3. The position of target loci (red boxes) is shown in the upper part. The bar chart depicts the fold change in comparison to the MC1T1 sample on control loci D, E and F. Blue dots – heterozygous SNPs, dark blue - SNPs in duplication. See also Figure S5, Table S2 and Data S3.

Figure 4. Detection of LOH events at the *P. tricornutum PtUMPS* locus. (A) Schematic of alleles in the *PtUMPS* strains. Homologous chromosomes are depicted as grey bars with exons in blue and green, loss-of-function indel mutations are in red, purple and orange bars represent silent SNPs between the two original alleles. (B-C) Recombination in *PtUMPS* mutant strains during 14 days of cultivation under non-selective conditions. (B) The number of recovered uracil prototrophic colonies per strain. (C) Examples of sequenced recombinant alleles in one *ptumps-320bp* and two *ptumps-1368bp* colonies. (D) Estimation of the interhomolog recombination frequency per thousand cell divisions. Each dot represents one replica. (E–H) Recombination events in response to stress-induced by (E) zeocin; (F) H_2O_2; (G) cadmium and (H) 2,4-Decadienal. *ptumps-1bp* replicas are depicted in shades of grey, *ptumps-1368bp* replicas are depicted in shades of blue. See also Figure S6 and Data S4-S5.

TABLES

Table 1. Characteristics of loci with multiple haplotypes found in *S. robusta* and *P. tricornutum*

	Whole-genome average		Loci with multiple haplotypes	
	Total number	**Percent**	**Total number**	**Percent**
***Seminavis robusta* (1405 loci)**				
GC content		48.5 %		48.8 %
SNPs total	489799	100 %	7714	100%
SNPs in intergenic regions	149782	30.58 %	2662	34.50 %
SNPs in protein coding genes	339890	69.39 %	5050	65.47 %
Intron	17300	3.53 %	233	3.03 %
Exon	322590	65.86 %	4817	62.44 %
Functional RNAs	127	0.03 %	2	0.03 %
***Phaeodactylum tricornutum* (3380 loci)**				
GC content		48.77 %		49.04 %
SNPs total	290164	100%	22531	100%
SNPs in intergenic regions	110727	38.16 %	6767	30.03 %
SNPs in protein coding genes	178906	61.65 %	15735	69.83 %

Intron	16271	5.61 %	1242	5.51 %
Exon	162635	56.04 %	14493	64.32 %
Pseudogenes	425	0.15 %	27	0.12 %
Functional RNAs	106	0.04 %	2	0.01 %

Table 2. Verification of haplotype diversity in *S. robusta* at three selected loci in three cultures (Sr1 – Sr3) through emulsion PCR amplification, cloning and Sanger sequencing of individual clones

Locus	Number of SNPs	Culture	Number of haplotypes at 4 months after single cell isolation	Number of supporting reads for each haplotype
Sro_contig211: 7509-8241				
	11	Sr1	5	26; 15; 1; 1; 1
	11	Sr2	4	23; 18; 8; 1
	11	Sr3	3	26; 25; 2
Sro_contig2103: 8397-9162				
	5	Sr1	3	19; 12; 1
	5	Sr2	3	8; 6; 1
	5	Sr3	2	24; 13
Sro_contig872: 16034-16975				
	4	Sr1	3	24; 16; 1
	4	Sr2	5	19; 14; 1; 1; 1
	4	Sr3	6	29; 19; 3; 2; 1; 1
Sro_contig556:54453-55487 - control mix of plasmids containing two alleles of the locus				
	3	-	2	44; 19

Table 3. Change in number of recovered haplotypes over time in *P. tricornutum*

Locus name	Coordinates	Number of SNPs	Samples harvested at 1 month after single cell isolation		Samples harvested at 6 months after single cell isolation	
			Number of haplotypes	**Coverage**	**Number of haplotypes**	**Coverage**
G32	13:103145-105183	10	8	3903	12	878
G33	27:205731-207770	18	6	5712	8	988
G34	20:101923-103985	28	4	3140	8	654
G35	12:519921-521994	12	9	2992	10	706
G36	2:961633-963692	12	8	5225	8	688

STAR METHODS

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Lieven De Veylder (lieven.deveylder@psb.vib-ugent.be)

Materials Availability

Material generated in this study is available upon request from the lead contact

Data and Code Availability

Raw sequencing data were deposited to the Sequence Read Archive (SRA) under BioProject accessions PRJNA658511 and PRJNA658224. SRA accession numbers for individual samples are listed in Data S3. Processed datasets were uploaded to zenodo: Aligned and processed long-read sequencing datasets *S. robusta* PacBio, *P. tricornutum* MinION reads and SNP selected for haplotype counting for both species are available at https://doi.org/10.5281/zenodo.4005721. Aligned PacBio amplicon sequencing reads, reference file and selected biallelic SNPs used in haplotype counting are available at https://doi.org/10.5281/zenodo.4005643. Processed datasets from LOH detection in mother and daughter cultures including ILLUMINA reads aligned to the reference P. tricornutum genome used for SNP calling, SNP calls for individual samples and jointly called SNPs on all samples are available at https://doi.org/10.5281/zenodo.4006016. **Code availability**

The haplotype coding script to count haplotypes in long-read sequencing datasets is available on zenodo, at https://doi.org/10.5281/zenodo.4001752. A version of the haplotype counting script that outputs the combination of bases at selected SNP sites is available on zenodo, at https://doi.org/10.5281/zenodo.4173002. All other data are available from the authors upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Diatoms datasets and strains

Datasets and strains used in this study are summarized in Data S3. The *S. robusta* D6 reference strain (accession number DCG 0498) is available from the BCCM/DCG diatom culture collection at Ghent University (http://bccm.belspo.be/about-us/bccm-dcg). Publicly available genomes of *S. robusta* strain D6 [71] https://www.ebi.ac.uk/ena/browser/view/CAICTM010000000 and *P. tricornutum* Pt1 8.6 (CCMP2561) strain [91] https://www.ebi.ac.uk/ena/browser/view/GCA_000150955.2 and next-generation sequencing datasets were used for our analysis.

Diatom cultivation conditions

Both *S. robusta* strain D6 and *P. tricornutum* strain Pt1 subculture MC2 were cultivated in 1x TMB medium consisting of 34.5 g/L of Tropic Marin Bio-Actif sea salt (Tropic Marin, Germany) and 0.08g/L sodium bicarbonate (Sigma-Aldrich) supplemented with 1x Guillard's (F/2) Marine Water Enrichment Solution (Sigma-Aldrich), 100 µg/mL ampicillin, 50 µg/mL gentamycin and 100 µg/mL streptomycin in 12 h/12 h light/dark cycle. *P. tricornutum* cultures were cultivated at 20°C, under photosynthetic LED light with an intensity of 160 µmol photons m^{-2} s^{-1} and with

100 rpm shaking. *S. robusta* cultures were cultivated at 18°C with approximately 85 µmol photons m^{-2} s^{-1} from cool-white fluorescent lights.

METHOD DETAILS

Estimation of intra-specific variability in *Fragilariopsis cylindrus* metagenomes

Tara Oceans metagenomic reads from 0.8-5 µm, 5-20 µm, 20-180 µm, 180-2000 µm and 0.8-2000 µm size fractions were mapped against the FASTA file of the *Fragilariopsis cylindrus* CCMP 1102 genome [25] (available at http://genome.jgi-psf.org/Fracy1/Fracy1.home.html) using Bowtie2 v2.3.4.332 with a 95% identity filter. Two depths, surface (5m depth) and deep chlorophyll maximum (DCM; 35m depth), both located in the epipelagic mixed layer [92] from Station 86 situated in the Southern Ocean (near the Antarctic peninsula, 64°30'88" S, 53°05'75" W), displayed vertical sequence coverage superior or equal to 10X for three size fractions (0.8-5, 5-20 and 0.8-2000 µm) and were thus selected for further analysis. Using SAMtools v1.10 33, the resulting SAM files were converted into BAM files and for each sample the BAM files of the three different size fractions were merged to increase the coverage (final mean coverage 51.75X and 58.46X for surface and DCM respectively). Downstream analyses were performed with the anvi'o platform[73] to generate profile databases based on the BAM files that were combined into a merged profile database. Genes were imported into anvi'o at the level of individual exons. Then, the program "anvi-summarize" was used with the "init-gene-coverages" flag to characterize the mean coverage of each gene in the surface and DCM samples. Genes that were considered for downstream analyses (n = 24,326) were invariably detected within a population niche (here the metagenome) [93]. These genes had to occur in the two samples and their mean coverage in each sample had to remain within a factor 3 of the mean coverage of all 27,137 genes in the same metagenome. The filtering step based on gene level coverage values is critical to remove outlier genes that may recruit reads from other related genera or species that potentially co-occurred in the samples (e.g., the 18S rRNA gene will recruit reads from other genera due to its high evolutionary stability, and genes from closely related species will display higher coverage values compared to the species-specific genes). Additionally, it allows to remove genes with hypervariable regions that will not recruit reads, preventing the subsequent analysis of single nucleotide variants (hereafter referred to as SNVs) [94]. Finally, the intra-population variability of *F. cylindrus* was analysed across the selected genes and in the two samples using the programme "anvi-gen-variability-profile", which provided tables reporting SNVs and their nucleotide frequencies in the recruited reads (Data S1). We defined SNVs as positions displaying at least 10% variation from the consensus nucleotide and with a mean vertical coverage ≥ 20X in the two samples. The variability tables were imported into R v4.0.1 to compute the number of variable positions and SNV density (i.e. the number of positions with SNVs for each exon in the selected genes divided by the corresponding exon length) for each exon. Gene-level mean coverage, number of variable positions and SNV density were computed using the information from the individual exons.

Genome-wide haplotype counting in *S. robusta* and *P. tricornutum* next-generation sequencing data

For both *S. robusta* and *P. tricornutum* genome-wide haplotype counting, a reliable single nucleotide polymorphism (hereafter referred to as SNP; in contrast to SNVs found in metagenomes from natural populations, SNP had been supported by at least 20% of reads in the sample from laboratory single strain) set was first identified in ILLUMINA short-read sequencing datasets and then used for counting of the number of haplotypes in the PacBio RS II and MinION long reads. The ILLUMINA and PacBio data of *S. robusta* from https://www.ebi.ac.uk/ena/browser/view/PRJEB36614 and ILLUMINA and Minion data of *P. tricornutum* from https://www.ebi.ac.uk/ena/browser/view/PRJNA487263. Because in long-read sequencing the error rate for indels is higher than for SNPs, indels were ignored in our analysis.

SNP calling: SNP calling on ILLUMINA short-read sequencing was done using GATK HaplotypeCaller 3.7.0[80]. In short, adapters and reads with a quality score below 20 were removed from ILLUMINA reads using BBduk2 [75] with minlen=35 qtrim=rl trimq=20 hdist=1 tbo tpe options and custom adapter reference file. Next, the respective reads were aligned to *S. robusta* v1 assembly CAICTM010000001-CAICTM010004752 (European Nucleotide Archive) or *P. tricornutum* v2 assembly2 GCA_000150955.2 (European Nucleotide Archive) using Burrows-Wheeler Alignment Tool (BWA)[76] algorithm BWA-MEM with -M option. Unmapped and multi-mapped reads were removed using SAMtools [77] view with -h -F 4 -q 1 options. Aligned reads were then sorted using picard-tools 1.8.0 [78] SortSam and duplicate reads were marked with MarkDuplicates and indexed with BuildBamIndex. Read base quality scores were adjusted by two round of recalibration. Here, SNPs and indels were called by GATK HaplotypeCaller[79] and filtered with a set of hard filters using SelectVariants; QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0 for SNPs and QD < 2.0, FS > 200.0, ReadPosRankSum < -20.0 for indels. Recalibration table was generated with BaseRecalibrator and recalibrated reads were printed with PrintReads. After the second round of recalibration, germline SNPs were called using HaplotypeCaller with --genotyping_mode DISCOVERY. Next, reliable biallelic SNPs were selected using SelectVariants with --restrictAllelesTo BIALLELIC -selectType SNP and QD < 2.0, QUAL < 30.0, SOR > 3.0, FS > 60.2, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, AF > 0.2 and DP < 10 options.

Repeat regions and low complexity DNA sequences in *S. robusta* and *P. tricornutum* were identified using RepeatModeler 1.0.9[81] and masked using RepeatMasker 4.0.5 [82], and SNPs in these regions were removed from the dataset using BEDtools [83] subtract algorithm. Finally, selected fields (CHROM, POS, REF, ALT) from the SNP dataset were extracted from the vcf file to a table and split into independent files by contig/chromosome using awk.

S. robusta PacBio reads processing: Circular Consensus Sequences (CCS) were obtained with smrtanalysis 2.3.0 (PacBio) with minFullPasses 0 option. CCS reads were then self-corrected using canu 1.4[30] with canu_correct genomeSize=136.0m errorRate=0.035 -pacbio-raw options and trimmed with canu_trim genomeSize=136.0m errorRate=0.035 -pacbio-corrected options. Corrected reads were mapped to the reference genome using BLASR[89] with -sam -clipping soft options. The CIGAR string was corrected with samfixcigar, soft-clipped bases were removed with biostar84452 from jvarkit[84] and uniquely mapped reads with mapping quality >20 were selected using SAMtools[77]. Coverage was estimated using GATK 3.7.0 DepthOfCoverage. SAMtools view and awk were used to split the PacBio reads to separate files per contig.

P. tricornutum MinION reads processing: MinION reads were self-corrected using canu 1.4[30] with canu_correct genomeSize=30m errorRate=0.144 -nanopore-raw options. Reads were aligned to the genome using GraphMap[85] with default settings and uniquely mapped reads were selected using SAMtools view. The CIGAR string was corrected with samfixcigar, soft-clipped bases were removed with biostar84452 from jvarkit. Coverage was estimated using GATK 3.7.0 DepthOfCoverage. SAMtools view and awk were used to split the PacBio reads to separate files per contig.

Haplotype counting: The haplotype counting was done with a custom script based on bash, awk and Sam2Tsv from jvarkit (https://doi.org/10.5281/zenodo.4001752). In short, a record for every base in each processed PacBio/MinION read with the position, reference and the actual base was obtained using Sam2Tsv from jvarkit[84]. Next, only positions of SNPs selected in ILLUMINA reads were retained. The record was divided into fixed windows of max 1 kb from the first SNP and haplotypes for selected sites were written for each read separately. Reads containing an indel or another base than the reference or the alternative base at the selected SNP position or not covering the 1 kb region were removed and the number of haplotypes and number of supporting reads for each haplotype was counted using awk. Loci with multiple haplotypes were selected from the record of the number of haplotypes per 1 kb with the number of supporting reads with the following conditions: at least three haplotypes had to be supported each by at least 2 reads, the locus had to be at least 100 bp long and the coverage had to be below 100x to remove repeat regions that were not masked. Visualization of haplotype counting data was done using Circos [86] and karyoploteR [88].

Haplotype counting in control genomes: As a control for haplotype counting overcounting due to high error rate in long-read sequencing data, the genome assembly, ILLUMINA and PacBio sequencing data publicly available for haploid yeast *Saccharomyces cerevisiae* GLBRCY22-3 [31] grown from single colony https://www.ncbi.nlm.nih.gov/bioproject/PRJNA279877 and for several diploid *Arabidopsis thaliana* Ler plants [32,33] from https://www.ncbi.nlm.nih.gov/bioproject/PRJNA311266 and https://www.ncbi.nlm.nih.gov/bioproject/237120 were used. The SNP calling, PacBio data processing and haplotype counting were done as described above and are summarized in Data S2.

Calculation of the error rate in self-corrected long-read sequencing data used for haplotype counting.

The long-read sequencing technology is known to have a higher error rate than the short-read sequencing technology[95]. To improve the read quality, we generated circular consensus sequencing (CCS) reads for the PacBio datasets and used self-correction for both PacBio and MinION reads (detailed description is available in the Methods). The self-correction was preferred over the correction by ILLUMINA reads, as the ILLUMINA reads were used for calling an SNP dataset and this could introduce bias in subsequent haplotype counting. The long-sequencing reads were further processed after alignment to the genome by removal of soft clipping, correction of CIGAR string and selection of uniquely mapped reads. The error rate in corrected and aligned PacBio and MinION files was estimated using Alfred [90] (Table S8).

The haplotype counting script processes only positions of reliable SNPs selected in ILLUMINA sequencing data and removes all reads containing insertions or deletions or other bases than the expected reference and alternative allele at the selected sites. Therefore, only the mismatch rate is relevant for the haplotype counting. Each position with a mismatch can result in one of the other three bases different from the reference. If the probability of mismatch to the reference is considered identical for each possible mismatched base, it is equal to one third. As only one of these bases is accepted by the haplotype counting script, the final probability of a mismatch at positions of reliable SNPs is equal to one-third of the mismatch rate. Thus, the probability of error at selected SNP sites used for haplotype determination ranged between 1.46% for *S. robusta* PacBio genome-wide sequencing and 0.05% for *P. tricornutum* T1 amplicon re-sequencing (Table S8).

Resequencing of *S. robusta* and *P. tricornutum* loci with multiple haplotypes

DNA extraction: DNA for deep sequencing was harvested and isolated by the CTAB DNA extraction method. Cells from approximately 500 mL of exponentially growing *S. robusta* and *P. tricornutum* cultures were harvested by centrifugation at 1216 x g for 5 min. The supernatant was discarded and the cell pellet was resuspended in 400 µl of CTAB buffer (1% (w/v) CTAB, 100 mM Tris-HCl pH 7.5, 10 mM EDTA pH 8, 700 mM NaCl and freshly added 4 µg RNase A). *S. robusta* cells were disrupted by agitation with glass/zirconium beads (0.1-mm diameter; Biospec) on a bead mill (Retsch) for three times 1 min at frequency 20 Hz. Samples were incubated for 30 min at 60°C and afterwards let to cool down on the ice for 15 min. Next, 250 µl of chloroform:isoamylalcohol 24:1 was added and the samples were mixed manually for 1 min. Phases were separated by centrifugation at 20 000 x g for 10 min. The upper aqueous phase was transferred to a new tube and DNA was precipitated by the addition of an equal volume of isopropanol followed by centrifugation for 15 min at 20 000 x g. The DNA pellet was washed with 70% ethanol, air-dried and resuspended in 50 µl of 10 mM Tris-HCl pH 8.5.

Emulsion PCR: Loci for re-sequencing of haplotypes were selected from the list of loci with multiple haplotypes obtained through genome-wide haplotype detection. Three loci were selected in *S. robusta* for Sanger sequencing verification (Table 2) and 62 loci for PacBio amplicon sequencing verification in the case of *P. tricornutum* (Table S3). Primers for amplification were designed manually (Data S3). To avoid PCR recombination artefacts [96,97], selected loci were amplified by emulsion PCR using the MICELLULA DNA Emulsion &

Purification Kit (roboklon, Germany) according to the manufacturer's instructions. The DNA concentration was measured on NanoDrop (ThermoFisher Scientific) and the number of DNA template copies per µg of DNA was calculated according to the genome size of the respective diatom. A maximum of 10^7 DNA molecules was used per single emulsion PCR reaction. The PCR reaction mix consisted of 1x OptiTaq PCR buffer B (roboklon), 200 µM dNTP mix, 2 µM of each forward and reverse primer, DNA template with 10^6-10^7 molecules, 1 mg/mL acetylated BSA and 2.5U Opti Taq DNA polymerase (roboklon) in 50 µL of total volume. The emulsion mix was prepared separately by mixing 220 µL of emulsion component 1, 20 µL of emulsion component 2 and 60 µL of emulsion component 3 per PCR reaction. The 50 µL PCR reaction was mixed with 300 µL of emulsion mix and emulsion was created by continuous vortexing at 1400 rpm at 4°C for 5 min. Each emulsion PCR reaction was split into three PCR tubes and run with the following parameters: 94°C initial denaturation for 2 min, 26 cycles of 94°C denaturation for 15 s, 56°C annealing for 30 s and 72°C extension with 1 kb/min relative to the amplified fragment length, followed by a final extension at 72 °C for 10 min. The emulsion was broken by the addition of 1 mL of isobutanol and vortexing. Next, 400 µL of Orange-DX solution was added and reactions were gently mixed and centrifuged for 2 min at 20 000 x g. The organic phase was removed and the aqueous phase was transferred to a Micellula spin column activated by 40 µL of DX buffer. Columns were centrifuged at 11 000 x g for 1 min, washed first with 500 µL of Wash-DX1 buffer, and then with 650 µL of Wash-DX2 buffer and the leftovers of buffer were removed by an additional centrifugation for 2 min. PCR products were eluted in 50 µL of Elution-DX buffer (all components: roboklon).

Sanger sequencing of *S. robusta* amplicons

S. robusta emulsion PCR products were cloned into the pGEM-T vector (Promega) according to the manufacturer's instructions. In brief, the A overhangs were added by incubation of PCR product with 10 µM dNTP mix and 1U of Taq DNA polymerase (Invitrogen) at 72°C for 10 min. Then, 3.5 µL of PCR product was mixed with 5 µL of 2x ligase buffer, 0.75 µL of pGEM-T vector and 2.25 U of T4 DNA ligase (all Promega) and incubated for 12 h at 4°C. Ligation mixtures were transformed through electroporation into *E. coli* DH5alpha cells and transformants were selected on LB supplemented with 100 µg/mL ampicillin (Duchefa). Clones containing cloned PCR products were selected by Sanger sequencing with pGEM-5 and pGEM-6 primers (Data S3). Sequencing results were aligned with the reference using Clustal Omega [98], and haplotypes were manually assembled for each clone. Two alleles of Sro_contig556:54453-55487 (Data S3) were cloned into the pGEM-T vector and an equimolar mix of these two plasmids was used for emulsion PCR as a control for artefact generation. To simulate conditions similar to emulsion PCR reactions on *S. robusta* genomic DNA, 10^6-10^7 molecules of *S. robusta* genomic DNA were added and the control samples were amplified with pGEM-3 and pGEM-4 primers (Data S3). The PCR products were again cloned into the pGEM-T vector and sequenced with pGEM-5 and pGEM-6 primers. For counting the number of found haplotypes, we considered only SNPs that were found in SNP call (so only the reference and alternative allele at a selected position). Therefore, if the base at the SNP position was neither reference nor an alternative base identified in the SNP call, the read was discarded. We considered the probability of mismatch to the reference identical for each possible mismatched base, and therefore one third. With 99.99% accuracy of the Sanger sequencing (1 error per 10,000 sequenced base pairs) and the probability of mismatch turning the base to the wrong reference/alternative allele being one third, the final probability of error at selected sites is 1: 30,000.

P. tricornutum cultures and PacBio amplicon sequencing and haplotype counting

13 intergenic loci and 59 loci overlapping with coding regions (Table S3) were selected based on an SNP call on T1 cell culture and the list of loci with more than two haplotypes for PacBio Sequel amplicon sequencing. All loci were amplified by emulsion PCR as described above with primers listed in Data S3. In the case of low amplification efficiency, the emulsion PCR was repeated. Purified PCR products were concentrated using Genomic DNA Clean & Concentrator (Zymo Research) according to the manufacturer's instructions. Amplifications of *CFP*, *GFP* and *YFP* genes were used for control reactions for random mistakes and control reactions for artificial haplotypes detection. The *CFP*, *GFP* and *YFP* sequences (Table S2 and Data S3) were amplified by emulsion PCR with primers binding to vector backbone (Data S3) from GK-333-CFP, GK-359 and GK-333-YFP plasmids respectively (GK-333, GenBank[72] accession MW934548 and GK-359, GenBank accession MW934549 were a gift from Dr. Nicole Poulsen, Center for Molecular Bioengineering at TU Dresden). Control reactions for random errors consisted of separately amplified GFP and YFP and control reactions for PCR-mediated recombination consisted of mixed amplification of CFP+GFP and CFP+YFP (Table S2). Amplicons were pooled together into two samples. Sample 1 contained 63 *P. tricornutum* endogenous 63 amplicons from DNA harvested at a T6 time point, YFP amplified separately and CFP+GFP amplified in one reaction. Sample 2 contained 5 *P. tricornutum* endogenous 5 amplicons from DNA harvested at a T1 time point, GFP amplified separately and CFP+YFP amplified in one reaction. Samples were barcoded, mixed in 9:1 ratio and sequenced on 1 PacBio Sequel SMRT cell at Novogene (UK).

Circular Consensus Sequence (CCS) reads were obtained with SMRT Link 8.0.0 software (PacBio) with --min-length 500 --max-length 2400 --min-passes 4 --by-strand and mapped to the reference loci with BLASR[89] with default settings. The CIGAR string was corrected with samfixcigar, soft-clipped bases were removed with biostar84452 from jvarkit[84] and reads with mapping quality >20 were selected using SAMtools [77]. Next, the haplotype number per locus was counted as described above. Only haplotypes supported either by at least 1% of valid reads or at least two reads if read count was lower than 200 were selected (Table S3). The genes fully covered by PacBio amplicons were manually annotated to detect putative variants.

LOH and CNV detection

Cultures started from a single cell: To isolate single cells from *P. tricornutum* cultures, an aliquot from the respective culture was diluted to 10^6, 10^9 and 10^{12} in the growth medium. 200 µl of diluted culture were spread on a 120 x 120 mm Petri dish (Corning Gosselin) with solid medium prepared with 17.25 g/L of Tropic Marin Bio-Actif sea salt solid and 10 g/L of Plant Tissue Culture Agar (Neogen) supplemented with 1x Guillard's (F/2) Marine Water Enrichment Solution (Sigma-Aldrich), 100 µg/L ampicillin, 50 µg/L gentamycin and 100 µg/L streptomycin, and the plates were incubated at room temperature with a 12-h/12-h light/dark cycle. Plates were checked for the presence of colonies after 14 days and single colonies were transferred from the plate with the highest dilution factor that contained colonies into liquid TMB medium using a pipette tip. This procedure was used to isolate and start three colonies from a single cell from the Pt1 culture to obtain mother cultures MC1, MC2 and MC3. Thirty days after mother culture isolation (T1 time point), daughter cells DC1.1-DC3.3 were isolated from the respective mother cultures (Figure S5). Cultures for deep sequencing were harvested twice; at time point T1 part of the mother cultures was harvested at time point T1, and 30 days later at time point T2 all cultures in the experiment were harvested.

Estimation of cell division rate: To pre-test the number of cell divisions per 6 days in our culture conditions, 1 x 10^6 cells were inoculated on day 1 in 6 replicas and the number of cells was counted after 6 full days). The average number of cells after 6 days was 8.03 x 10^6. Using the following formula: $N_t = N_0 2^{tf}$, where N(t) is the number of cells at time t, N_0 is the initial number of cells, t is the time in days and f is the frequency of cell cycles per unit time, the cell division rate was estimated to be around 0.501 cell division per day. The actual rate of cell division can be higher if cell mortality is counted.

Illumina sequencing, SNP calling, LOH and CNV detection: DNA for deep sequencing was harvested and isolated by the CTAB DNA extraction method as described above. Paired-end libraries were prepared with the NEBNext Ultra DNA Library Prep Kit for Illumina (NEB) with a 500-bp insert size and sequencing was performed on a 2× 150bp Illumina NextSeq500 Medium at the VIB Nucleomics Core (Leuven, Belgium). Adapters and reads with a quality score below 20 were removed using BBduk2[75] with minlen=35 qtrim=rl trimq=20 hdist=1 tbo tpe and custom adapter reference file. Trimmed reads were aligned to *P. tricornutum* v2 assembly GCA_000150955.2[91] (European Nucleotide Archive) using Burrows-Wheeler Alignment Tool (BWA)[76] algorithm BWA-MEM and processed and re-calibrated as described above. After the second round of recalibration, SNPs were called in three different ways:

Germline SNP calling with GATK HaplotypeCaller a) joint genotyping with GATK 4.2.1 and b) for individual samples with GATK 3.7.0: First, germline SNPs were called with HaplotypeCaller with either a) -ERC GVCF option and gVCF files were combined with CombineGVCFs and then jointly genotyped with GenotypeGVCFs b) or without the -ERC GVCF option and joint genotyping. SNPs were filtered with QD < 2.0, QUAL < 30.0, SOR > 3.0, FS > 60.2, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0 and DP < 10 and indels with QD < 2.0, QUAL < 30.0, FS > 200.0, ReadPosRankSum < -20.0, DP < 10 using VariantFiltration and filtered SNPs and indels were removed with SelectVariants. These set of germline SNPs were used to build a Panel of Normals for following pairwise comparison of mother and daughter cell cultures using Mutect2 and for cross-verification of LOH regions identified in Mutect2.

Pairwise comparison of mother and daughter cultures with GATK Mutect2: All GATK algorithms were version 4.2.1 if not stated otherwise. First, SNPs were called on each sample with Mutect2 in tumour-only mode. Next, a vcf file for each sample was created by moving the sample-level AF allele-fraction annotation were moved into the INFO field for each sample using VariantsToTable. The Panel of Normals was prepared using the germline SNPs dataset for individual samples generated by HaplotypeCaller as a germline-resource by first calling the SNPs for each sample with Mutect2 in tumour-only mode, then merging all files with CombineVariants (GATK3.7.0) and finally creating the Panel of Normals with CreateSomaticPanelOfNormals. A Panel of Normals file for each pairwise comparison was prepared by masking germline variants from the respective individual samples using germline SNPs called by HaplotypeCaller using CatVariants (GATK3.7.0) to merge SNPs from both samples and then masking them using SelectVariants with -XL option. Finally, LOH regions and *de novo* mutations were detected by Mutect2 SNP call with vcf file with sample-level AF allele-fraction in the INFO field used as --germline-resource, the respective masked Panel of Normals file used as –pon, --genotype-germline-sites true and either the mother culture used as tumour sample and daughter culture used as normal to detect LOH events in the daughter culture or vice-versa to detect *de novo* mutations in daughter culture (Table S5 and Data S3). Called SNPs were filtered with FilterMutectCalls and filtered SNPs were removed with SelectVariants –excludeFiltered. As a control, the T1 with T2 time point of each mother culture were compared. A minimum of three consecutive SNPs missing in the daughter or T2 mother culture was considered as a LOH event and all LOH regions were reexamined in the datasets of germline SNPs obtained either by individual SNP calling or joint genotyping. The nature of the LOH was judged by comparison of coverage of the LOH region and its surrounding heterozygous borders and through Sanger resequencing of the LOH border (Tables S5 and Data S3).

CNV detection: The copy-number variation was detected using GATK 4.1.7 CNV detection pipeline. First, intervals list with bin length set to 100 bp and -interval-merging-rule OVERLAPPING_ONLY was prepared with PreprocessIntervals and collect raw counts were collected using CollectReadCounts and CNV panel of normals was generated by CreateReadCountPanelOfNormal with --minimum-interval-median-percentile 5.0 setting. Standardized copy ratios and denoised copy ratios were obtained using DenoiseReadCounts against a panel of normals. Reference and alternative allele counts at common germline sites called on all samples by HaplotypeCaller in GVCF mode and jointly genotyped were obtained using CollectAllelicCounts for each sample. Segments of contiguous copy ratios were acquired by ModelSegments in a paired analysis with --denoised-copy-ratios and --allelic-counts from the daughter culture and --normal-allelic-counts from the respective mother culture. Amplified, deleted and copy-neutral segments were called with CallCopyRatioSegments with default settings and plotted using PlotModeledSegments. Detected CNV events were cross-verified in the datasets of germline SNPs obtained by SNP calling and joint genotyping and in estimated coverage counts per 10 bp obtained using bedtools 2.2.28 coverage. Further, identified LOH in tandem repeat on chromosome 5 in DC1.2 was verified by Sanger sequencing and duplication on chromosome 23 in DC1.3 was confirmed by qPCR quantification (see below).

Re-sequencing of identified LOH events in mother versus daughter cell culture comparison and CNV analysis: The nature of LOH events identified by pairwise comparison between mother and daughter cell culture was first judged by visual comparison of the sample coverage at the LOH region and the neighbouring region. Next, the region was amplified by emulsion PCR as described in the Methods, PCR products were cloned into the pGEM-T vector and individual clones were sequenced. If the LOH region sequence was found together with both alleles of neighbouring heterozygous SNP(s), the region was considered as copy-neutral LOH. If only a single allele was found and the coverage data corresponded, the region was considered as a deletion. Predicted heterozygous SNPs in regions between identified LOH events on chromosome 26 in DC2.1 culture were amplified by a standard PCR using Phusion High-Fidelity DNA Polymerases (Thermo Scientific) according to the manufacturers' instructions and sequenced by Sanger sequencing to verify their heterozygosity (Data S3). The same approach was used to verify the hetero/homozygosity of three SNPs on chromosome 27 that were predicted as LOH in culture DC2.2, but was called as homozygous by independent SNP calling also in the mother culture MC2. The LOH on chromosome 5 in DC1.2 was identified as CNV, but PCR amplification confirmed the presence of two alleles with different length (differing in 1960 bp). The longer allele contained a tandemly duplicated region while in the short allele the duplication was missing. Subsequent Sanger sequencing of alleles showed that culture DC1.2 contains two short alleles with LOH of SNPs in the surrounding region (Figure S5 and Data S3). The effect of SNPs found in was profiled using SnpEff[39]. First, pre-build SnpEff *P. tricornutum* database was downloaded and records for SNPs in LOH regions were selected using bedtools intersect[83]. The selected SNP variants were annotated using snpEff with default settings. SNPs annotated with “HIGH” effect were selected using grep and their exact effect was manually annotated.

qPCR quantification of duplication at chromosome 23 in DC1.4 culture:

The relative copy number variation on chromosome 23 was examined in daughter cultures DC1.1, DC1.2 and DC1.3 in comparison with mother culture MC1 by quantitative real-time PCR (qPCR). DNA was extracted as described above and concentration was adjusted to 48 pg/µL for each sample. Two primer pairs were designed into the region containing putative duplication on chromosome 23 in DC1.3, two primer pairs were located on chromosome 23 outside the duplicated region and two primers pairs were targeting loci outside of chromosome 23, one on chromosome 6 and one on chromosome 22 (Data S3). qPCR was performed using the SYBR Green kit (Roche) with 100 nM primers and 0.125 µL DNA in a total volume of 5 µL per reaction. qPCR amplification reactions were run and analyzed on the LightCycler 480 (Roche) with following cycling conditions: 10 min polymerase pre-incubation at 95°C and 45 cycles of amplification at 95°C for 10 s, 60°C for 15 s, and 72°C for 15 s. Melting curves were recorder after the last cycle by heating from 65 to 95°C. All qPCR amplicons were sequenced by Sanger sequencing to confirm their integrity. For each reaction, three technical repeats were performed. Data were analyzed using qbase+ (Biogazelle)[99] with a copy number analysis option and with the mother culture MC1 as a reference sample and loci on the distal arm of chromosome 23 (locus D) chromosome 6 (locus E) and chromosome 22 (locus F) as reference targets.

PtUMPS read-out system

PtUMPS cultures: Strains *ptumps-1bp* and *ptumps-1368bp* were generated based on a previously described protocol[41]. Briefly, Cas9 ribonucleoprotein (RNP) complexes were assembled to target the *PtUMPS* locus, at either the gUMPS1 site or the gUMPS4 site (Data S3). An equimolar mixture of RNP gUMPS1 and RNP gUMPS4 (4 µg each) was bombarded into wild-type *P. tricornutum* Pt1 8.6 (CCMP2561) cells. Two rounds of selection were made on silicate-free F/2 medium (Sigma) plates supplemented with 50 µg/mL uracil (Sigma) and 300 µg/mL 5-fluoroorotic acid (5-FOA; ThermoFisher). Cell lysates were then prepared to serve as a template for genotyping. PCRs using the Q5 High Fidelity DNA polymerase (New England Biolabs) and primers UMPS_5UTR_F and UMPS_3UTR_R (Data S3) were performed to amplify the *PtUMPS* locus. The generated amplicons were subcloned employing the CloneJET PCR cloning kit (Thermo Scientific) and analyzed through Sanger sequencing. The *ptumps-320bp* strain was prepared and described previously under the name UA17[41]. *PtUMPS* mutant strains were maintained in conditions described above in 1xTMB medium supplemented with 50 µg/mL uracil and 100 µg/mL 5-FOA to prevent the restoration of the wild-type allele.

PtUMPS 14 day cultivation in non-selective conditions: Cells densities of *ptumps-1bp*, *ptumps-320bp* and *ptumps-1368bp* were estimated using Bürker counting chamber and 20x10^6 cells from were harvested by centrifugation at 1216 x g for 5 minutes. The cell pellet was washed four times with 50 mL of 1xTMB medium, then resuspended in 750 mL of 1x TMB medium supplemented with 50 µg/mL uracil and resulting cultures were cultivated in 12h/12h light/dark cycle. After 7 days, another 750 mL of fresh medium supplemented with uracil were added. After 14 days, cultures were harvested by centrifugation, cell density was estimated and 50x10^6 cells from the culture were washed four times with TMB medium and plated on 1% agar ½ TMB medium 245 x 245 mm Nunc™ Square BioAssay plates (ThermoFisher) and incubated in 18h/6h light/dark cycle at 20°C for 6 weeks. The resulting colonies were manually counted (Data S4 and S5). Colonies selected for sequencing of *PtUMPS* locus were transferred to fresh 1xTMB medium and grown cell cultures were harvested as described above. The *PtUMPS* locus was amplified with primers PtUMPS-1 and PtUMPS-2 (Data S3) using two consecutive rounds of emulsion PCR. PCR products were cloned to pGEM-T vector and sequenced by Sanger sequencing.

Estimation of interhomolog recombination frequency:

~50x10^6 cells of three *ptumps-1bp* and five independent *ptumps-1368bp* cell subcultures started from a single cell were harvested by centrifugation at 1216 x g for 5 minutes. The cell pellet was washed four times with 50 mL of 1xTMB medium, then resuspended in 1xTMB medium. The cell density was determined and 2 to 6 replicas of either 25x10^6 or 20x10^6 cells were immediately plated on 1% agar ½ TMB medium and incubated as described above and incubated in 18h/6h light/dark cycle at 20°C for 6 weeks. The resulting colonies were manually counted (Data S4 and S5).

The frequency if interhomolog recombination was calculated based on the number of uracil prototrophic colonies after the immediate transfer of *ptumps-1368bp* strain from 5-FOA- and uracil-supplemented medium (only mutant cells survive) on plates without uracil (only cells that restored the wild-type allele survive). As the recombination that restored the wild-type *PtUMPS* allele had to occur within 1368-bp region, first, the incidence of LOH events per 1000 bp was estimated for each replica by dividing the number of colonies by 1.38. The exact nuclear genome size of *P. tricornutum* was determined from the genome fasta file using awk as 27,450,724 bp. Taking into account the genome length and number of plated cells, the recombination rate was recalculated per whole genome of 100 cells. As the copy-neutral LOH is detectable in half of the cases of mitotic interhomolog recombination due to random segregation of sister chromatids during mitotic metaphase, the number was multiplied by 2 to obtain the rate of mitotic recombination. The average value obtained from data for all replicas was 4.2 x 10^{-2}. The rate of reciprocal cross-overs on a 120-kb region in *S. cerevisiae* was estimated at 4 x 10^{-5} per cell division and the rate of gene conversion events as 3.5 × 10^{-6} per cell division [42] or 3.3 x 10^{-3} per cell division for interstitial LOH, 1.4 x 10^{-3} for terminal LOH genome-wide [43]. The frequency of reciprocal cross-overs and gene conversion on a 120-kb region were combined and recalculated first per 1 kb and subsequently per 11.89 Mb *S. cerevisiae* genome. The resulting rate of interhomolog recombination in *S. cerevisiae* was calculated as 4.3 x 10^{-3} in the case of 120 kb region or 4.7 x 10^{-3} in the case of the genome-wide studies.

Effect of treatment with cadmium, H_2O_2 and (E,E)-2,4-decadienal on interhomolog recombination at PtUMPS locus

The ranges of concentrations of chemicals used for treatments were first surveyed in literature, then selected concentrations were tested by treatment of Pt1 *P. tricornutum* strain for seven days. Afterwards, the cell survival of mock and treated cells was compared and the maximal dose that did not cause a decrease in cell density was selected as the maximal dose in the respective experiment. *ptumps-1bp* and *ptumps-1368bp* cell cultures were harvested by centrifugation at 1216 x g for 5 minutes and the cell pellet was washed four times with 50 mL of 1xTMB medium, then resuspended in 50 mL of 1xTMB medium. Cell density per mL was estimated and 25x10^6 cells per replica were transferred to 200 mL of 1xTMB medium supplemented with 50 µg/mL uracil and respective treatment or mock treatment. For zeocin treatment, zeocin was added to the final concentration (InvivoGen) was added to final concentrations of 1 µg/mL and 10 µg/mL from 1000× stock solution. For cadmium treatment, $CdCl_2$ (Sigma-Aldrich) was added to final Cd^{2+} concentrations of 5 µg/L and 50 µg/L from 4000× stock solution. For H_2O_2 treatment, a 30% solution of H_2O_2 (Merck) was added to a final concentration of 5 µM or 50 µM H_2O_2. For (E,E)-2,4-Decadienal treatment, 200 µL of DMSO was added to the mock-treated cell cultures and (E,E)-2,4-Decadienal (Sigma-Aldrich) was added to 0.1 µM and 1 µM final concentration from 1000x and 100x concentrated stock solution respectively. Cultures were incubated for 24h in 12h/12h light/dark cycle at 20°C, under photosynthetic LED light with an intensity of 160 µmol photons m^{-2} s^{-1} and with 100 rpm shaking. Afterwards, all samples were harvested by centrifugation at 1216 x g for 5 minutes and cell pellets were washed four times with 50 mL of 1xTMB medium and plated on 1% agar ½ TMB medium as described above and incubated in 18h/6h light/dark cycle at 20°C for 6 weeks. The resulting colonies were manually counted (Data S4 and S5).

Quantification and statistical analysis

No sample-size calculations were performed. Sample sizes were determined to be adequate based on preliminary experiments and feasibility. The interhomolog mitotic recombination rate measured in three ptumps-1bp and five ptumps-1368bp independent biological replicates, each with two to six technical replicas. The influence of environmental stresses on interhomolog mitotic recombination rate was performed in three biological replicas. The number of biological replicates for each data panel is indicated in the figure panel, in Supplementary data and the source data files. No randomization was performed (not applicable). Data exclusions: A threshold was set to count haplotypes in long-read sequencing for genome-wide and re-sequencing experiments to remove false-positive data as described above. No other data were excluded from this study.

DATA SX TITLES AND LEGENDS:

Data S1. Details of the SNV positions and densities in the selected genes in *F. cylindrus* metagenomes. Related to STAR Methods and Figure S1.

A) SNV positions of the exons present in the core genes, for the surface metagenome (10% divergence from consensus, at least 20X coverage in both samples). B) SNV positions of the exons present in the core genes, for the DCM metagenome (10% divergence from consensus, at least 20X coverage in both samples). C) SNV densities (%) of the exons present in the core genes. D) SNV densities (%) of the core genes (filtered by mean coverage).

Data S2. List of loci with more than two haplotypes detected genome-wide in profiled organisms. Related to STAR Methods and Figures 1 and S2-S3.

A) List of loci with more than two haplotypes detected in genome-wide analysis in *S. robusta*. B) List of loci with more than two haplotypes detected in genome-wide analysis in *P. tricornutum*. C) List of loci with more than two haplotypes detected in genome-wide analysis in *A. thaliana*. and D) List of loci with more than one haplotype detected in the genome-wide haplotype analysis in *S. cerevisiae*. E) Overview of genome-wide haplotype counting in diatoms *S. robusta* and *P. tricornutum*, yeast *S. cerevisiae* and plant *A. thaliana*. F) Error rate in processed long-read sequencing data used for haplotype counting estimated by Alfred.

Data S3. Details of re-sequencing of loci with multiple haplotypes, analysis of genetic changes in daughter cells and materials used throughout the study. Related to STAR Methods and Figures 1-4 and S2-6.

A) List of *S. robusta* loci used for Sanger sequencing verification of haplotype diversity. B) List of *P. tricornutum* 2-kb loci used for PacBio amplicon sequencing verification of haplotype diversity. C) The number of haplotypes at control amplicons obtained by PacBio amplicon sequencing. Only haplotypes supported by at least 1% of reads from the total read count or at least 2 reads were added to the haplotype count. D) The number of haplotypes of *P.*

tricornutum loci used for the verification of haplotype diversity by PacBio amplicon sequencing, six months after single-cell isolation. Only haplotypes supported by at least 1% of reads from the total read count or at least 2 reads were added to the haplotype count. E) Position of alleles in genes used as for a control amplification in emulsion PCR experiments. F) De novo SNPs identified by pairwise comparison of mother and daughter cultures in the genome-wide LOH analysis. G) SNPs with high effect on protein function fixed by LOHs and deletions. H) Genes covered by duplication on chromosome 23 in DC1.3. I) List of diatom datasets and strains. J) List of oligonucleotides used throughout the study.

Data S4. Photographs of plates with uracil prototrophic colonies resulting from mitotic interhomolog recombination. Related to STAR Methods and Figures 4 and S6.

Data S5. Counts of *ptumps-1bp* and *ptumps-1368bp* uracil prototrophic colonies. Related to STAR Methods and Figures 4 and S6.

A) Counts of *ptumps-1bp*, *ptumps-320bp* and *ptumps-1368bp* uracil prototrophic colonies after a 14-day cultivation in non-selective conditions. B) Counts of *ptumps-1bp* and *ptumps-1368bp* uracil prototrophic colonies after immediate transfer from 5-FOA- and uracil-supplemented medium to medium without uracil. C) Counts of *ptumps-1bp* and *ptumps-1368bp* uracil prototrophic colonies after 24-h zeocin treatment in non-selective conditions. D) Counts of *ptumps-1bp* and *ptumps-1368bp* uracil prototrophic colonies after 24-h H2O2 treatment in non-selective conditions. E) Counts of *ptumps-1bp* and *ptumps-1368bp* uracil prototrophic colonies after 24-h cadmium treatment in non-selective conditions. F) Counts of *ptumps-1bp* and *ptumps-1368bp* uracil prototrophic colonies after 24-h (E,E)-2,4-Decadienal treatment in non-selective conditions.

REFERENCES

1. Malviya, S., Scalco, E., Audic, S., Vincenta, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., de Vargas, C., Bittner, L., et al. (2016). Insights into global diatom distribution and diversity in the world's ocean. P Natl Acad Sci USA *113*, E1516-E1525. 10.1073/pnas.1509523113.
2. Godhe, A., and Rynearson, T. (2017). The role of intraspecific variation in the ecological and evolutionary success of diatoms in changing environments. Philos T R Soc B *372*. ARTN 20160399 10.1098/rstb.2016.0399.
3. Chepurnov, V.A., Mann, D.G., von Dassow, P., Vanormelingen, P., Gillard, J., Inze, D., Sabbe, K., and Vyverman, W. (2008). In search of new tractable diatoms for experimental biology. Bioessays *30*, 692-702. 10.1002/bies.20773.
4. Koester, J.A., Berthiaume, C.T., Hiranuma, N., Parker, M.S., Iverson, V., Morales, R., Ruzzo, W.L., and Armbrust, E.V. (2018). Sexual ancestors generated an obligate asexual and globally dispersed clone within the model diatom species Thalassiosira pseudonana. Sci Rep-Uk *8*. ARTN 10492 10.1038/s41598-018-28630-4.
5. Lewis, W.M. (1984). The Diatom Sex Clock and Its Evolutionary Significance. American Naturalist *123*, 73-80. Doi 10.1086/284187.
6. Davidovich, N.A., Davidovich, O.I., Podunay, Y.A., Gastineau, R., Kaczmarska, I., Poulickova, A., and Witkowski, A. (2017). Ardissonea crystallina has a type of sexual reproduction that is unusual for centric diatoms. Scientific Reports *7*. ARTN 14670 10.1038/s41598-017-15301-z.
7. Jewson, D.H. (1992). Size-Reduction, Reproductive Strategy and the Life-Cycle of a Centric Diatom. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences *336*, 191-213. DOI 10.1098/rstb.1992.0056.

8. Fuchs, N., Scalco, E., Kooistra, W.H.C.F., Assmy, P., and Montresor, M. (2013). Genetic characterization and life cycle of the diatom Fragilariopsis kerguelensis. European Journal of Phycology *48*, 411-426. 10.1080/09670262.2013.849360.

9. Lynch, M., Ackerman, M.S., Gout, J.F., Long, H., Sung, W., Thomas, W.K., and Foster, P.L. (2016). Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet *17*, 704-714. 10.1038/nrg.2016.104.

10. Hedrick, P.W. (2010). Genetics of populations.

11. Krasovec, M., Sanchez-Brosseau, S., and Piganeau, G. (2019). First Estimation of the Spontaneous Mutation Rate in Diatoms. Genome Biol Evol *11*, 1829-1837. 10.1093/gbe/evz130.

12. Krasovec, M., Rickaby, R.E.M., and Filatov, D.A. (2020). Evolution of Mutation Rate in Astronomically Large Phytoplankton Populations. Genome Biol Evol *12*, 1051-1059. 10.1093/gbe/evaa131.

13. Bürger, R. (2000). The mathematical theory of selection, recombination, and mutation (Wiley).

14. Ewens Warren, J. (2004). Mathematical population genetics [Texte imprimé] . I, Theoretical introduction / Warren J. Ewens, Second edition Edition (Springer).

15. Lang, G.I., Rice, D.P., Hickman, M.J., Sodergren, E., Weinstock, G.M., Botstein, D., and Desai, M.M. (2013). Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature *500*, 571-574. 10.1038/nature12344.

16. Maddamsetti, R., Lenski, R.E., and Barrick, J.E. (2015). Adaptation, Clonal Interference, and Frequency-Dependent Interactions in a Long-Term Evolution Experiment with Escherichia coli. Genetics *200*, 619-631. 10.1534/genetics.115.176677.

17. Fisher, R.A. (1930). The genetical theory of natural selection (Clarendon Press). 10.5962/bhl.title.27468.

18. Muller, H.J. (1932). Some Genetic Aspects of Sex. The American Naturalist *66*, 118-138. 10.1086/280418.

19. Johnson, R.D., and Jasin, M. (2001). Double-strand-break-induced homologous recombination in mammalian cells. Biochemical Society Transactions *29*, 196-201. Doi 10.1042/Bst0290196.

20. Kadyk, L.C., and Hartwell, L.H. (1992). Sister Chromatids Are Preferred over Homologs as Substrates for Recombinational Repair in Saccharomyces-Cerevisiae. Genetics *132*, 387-402.

21. Aguilera, A., Chavez, S., and Malagon, F. (2000). Mitotic recombination in yeast: elements controlling its incidence. Yeast *16*, 731-754. Doi 10.1002/1097-0061(20000615)16:8<731::Aid-Yea586>3.0.Co;2-L.

22. James, T.Y., Michelotti, L.A., Glasco, A.D., Clemons, R.A., Powers, R.A., James, E.S., Simmons, D.R., Bei, F.Y., and Ge, S.H. (2019). Adaptation by Loss of Heterozygosity in Saccharomyces cerevisiae Clones Under Divergent Selection. Genetics *213*, 665-683. 10.1534/genetics.119.302411.

23. Schoustra, S.E., Debets, A.J.M., Slakhorst, M., and Hoekstra, R.F. (2007). Mitotic recombination accelerates adaptation in the fungus Aspergillus nidulans. Plos Genet *3*. ARTN e68 10.1371/journal.pgen.0030068.

24. Dale, A.L., Feau, N., Everhart, S.E., Dhillon, B., Wong, B., Sheppard, J., Bilodeau, G.J., Brar, A., Tabima, J.F., Shen, D., et al. (2019). Mitotic Recombination and Rapid Genome Evolution in the Invasive Forest Pathogen Phytophthora ramorum. Mbio *10*. ARTN e02452-18 10.1128/mBio.02452-18.

25. Mock, T., Otillar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., Ward, B.J., et al. (2017). Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature *541*, 536-540. 10.1038/nature20803.

26. Madoui, M.A., Poulain, J., Sugier, K., Wessner, M., Noel, B., Berline, L., Labadie, K., Cornils, A., Blanco-Bercial, L., Stemmann, L., et al. (2017). New insights into global biogeography,

population structure and natural selection from the genome of the epipelagic copepod Oithona. Molecular Ecology *26*, 4467-4482. 10.1111/mec.14214.

27. Olm, M.R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B.A., Morowitz, M.J., and Banfield, J.F. (2021). inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol. 10.1038/s41587-020-00797-0.

28. Quince, C., Delmont, T.O., Raguideau, S., Alneberg, J., Darling, A.E., Collins, G., and Eren, A.M. (2017). DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol *18*. ARTN 181 10.1186/s13059-017-1309-9.

29. Russo, M.T., Cigliano, R.A., Sanseverino, W., and Ferrante, M.I. (2018). Assessment of genomic changes in a CRISPR/Cas9 Phaeodactylum tricornutum mutant through whole genome resequencing. Peerj *6*. ARTN e5507 10.7717/peerj.5507.

30. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res *27*, 722-736. 10.1101/gr.215087.116.

31. McIlwain, S.J., Peris, D., Sardi, M., Moskvin, O.V., Zhan, F.J., Myers, K.S., Riley, N.M., Buzzell, A., Parreiras, L.S., Ong, I.M., et al. (2016). Genome Sequence and Analysis of a Stress-Tolerant, Wild-Derived Strain of Saccharomyces cerevisiae Used in Biofuels Research. G3-Genes Genomes Genetics *6*, 1757-1766. 10.1534/g3.116.029389.

32. Kim, K.E., Peluso, P., Babayan, P., Yeadon, P.J., Yu, C., Fisher, W.W., Chin, C.S., Rapicavoli, N.A., Rank, D.R., Li, J., et al. (2014). Long-read, whole-genome shotgun sequence data for five model organisms. Scientific Data *1*. ARTN 140045 10.1038/sdata.2014.45.

33. Zapata, L., Ding, J., Willing, E.M., Hartwig, B., Bezdan, D., Jiao, W.B., Patel, V., James, G.V., Koornneef, M., Ossowski, S., and Schneeberger, K. (2016). Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. P Natl Acad Sci USA *113*, E4052-E4060. 10.1073/pnas.1607532113.

34. Hiraoka, M., Watanabe, K., Umezu, K., and Maki, H. (2000). Spontaneous loss of heterozygosity in diploid Saccharomyces cerevisiae cells. Genetics *156*, 1531-1548.

35. Hunter, N. (2015). Meiotic Recombination: The Essence of Heredity. Cold Spring Harb Perspect Biol *7*. 10.1101/cshperspect.a016618.

36. Rastogi, A., Vieira, F.R.J., Deton-Cabanillas, A.F., Veluchamy, A., Cantrel, C., Wang, G.H., Vanormelingen, P., Bowler, C., Piganeau, G., Hu, H.H., and Tirichine, L. (2020). A genomics approach reveals the global genetic polymorphism, structure, and functional diversity of ten accessions of the marine model diatom Phaeodactylum tricornutum. Isme Journal *14*, 347-363. 10.1038/s41396-019-0528-3.

37. National Library of Medicine (US) (2018). Sequence Read Archive (SRA) SRR7762337. Bethesda (MD): National Center for Biotechnology Information.

38. (US), N.L.o.M. (2018). Sequence Read Archive (SRA) SRR7762336. Bethesda (MD): National Center for Biotechnology Information.

39. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X.Y., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. Fly *6*, 80-92. 10.4161/fly.19695.

40. Sakaguchi, T., Nakajima, K., and Matsuda, Y. (2011). Identification of the UMP Synthase Gene by Establishment of Uracil Auxotrophic Mutants and the Phenotypic Complementation System in the Marine Diatom Phaeodactylum tricornutum. Plant Physiol *156*, 78-89. 10.1104/pp.110.169631.

41. Serif, M., Dubois, G., Finoux, A.L., Teste, M.A., Jallet, D., and Daboussi, F. (2018). One-step generation of multiple gene knock-outs in the diatom Phaeodactylum tricornutum by DNA-free genome editing. Nat Commun *9*. ARTN 3924 10.1038/s41467-018-06378-9.

42. Barbera, M.A., and Petes, T.D. (2006). Selection and analysis of spontaneous reciprocal mitotic cross-overs in Saccharomyces cerevisiae. P Natl Acad Sci USA *103*, 12819-12824. 10.1073/pnas.0605778103.

43. Sui, Y., Qi, L., Wu, J.K., Wen, X.P., Tang, X.X., Ma, Z.J., Wu, X.C., Zhang, K., Kokoska, R.J., Zheng, D.Q., and Petes, T.D. (2020). Genome-wide mapping of spontaneous genetic alterations in diploid yeast cells. Proc Natl Acad Sci U S A. 10.1073/pnas.2018633117.

44. Povirk, L.F. (1996). DNA damage and mutagenesis by radiomimetic DNA-cleaving agents: bleomycin, neocarzinostatin and other enediynes. Mutat Res *355*, 71-89. 10.1016/0027-5107(96)00023-1.

45. D'Autreaux, B., and Toledano, M.B. (2007). ROS as signalling molecules: mechanisms that generate specificity in ROS homeostasis. Nat Rev Mol Cell Bio *8*, 813-824. 10.1038/nrm2256.

46. Brembu, T., Jorstad, M., Winge, P., Valle, K.C., and Bones, A.M. (2011). Genome-Wide Profiling of Responses to Cadmium in the Diatom Phaeodactylum tricornutum. Environ Sci Technol *45*, 7640-7647. 10.1021/es2002259.

47. Ianora, A., Miralto, A., Poulet, S.A., Carotenuto, Y., Buttino, I., Romano, G., Casotti, R., Pohnert, G., Wichard, T., Colucci-D'Amato, L., et al. (2004). Aldehyde suppression of copepod recruitment in blooms of a ubiquitous planktonic diatom. Nature *429*, 403-407. 10.1038/nature02526.

48. Vardi, A., Formiggini, F., Casotti, R., De Martino, A., Ribalet, F., Miralto, A., and Bowler, C. (2006). A stress surveillance system based on calcium and nitric oxide in marine diatoms. Plos Biol *4*, 411-419. ARTN e60 10.1371/journal.pbio.0040060.

49. Lieber, M.R. (2008). The mechanism of human nonhomologous DNA end joining. J Biol Chem *283*, 1-5. 10.1074/jbc.R700039200.

50. Alves, I., Houle, A.A., Hussin, J.G., and Awadalla, P. (2017). The impact of recombination on human mutation load and disease. Philos Trans R Soc Lond B Biol Sci *372*. 10.1098/rstb.2016.0465.

51. Symington, L.S., Rothstein, R., and Lisby, M. (2014). Mechanisms and Regulation of Mitotic Recombination in Saccharomyces cerevisiae. Genetics *198*, 795-835. 10.1534/genetics.114.166140.

52. Abdullah, M.F.F., and Borts, R.H. (2001). Meiotic recombination frequencies are affected by nutritional states in Saccharomyces cerevisiae. P Natl Acad Sci USA *98*, 14524-14529. DOI 10.1073/pnas.201529598.

53. Forche, A., Abbey, D., Pisithkul, T., Weinzierl, M.A., Ringstrom, T., Bruck, D., Petersen, K., and Berman, J. (2011). Stress Alters Rates and Types of Loss of Heterozygosity in Candida albicans. Mbio *2*. ARTN e00129-11 10.1128/mBio.00129-11.

54. Modliszewski, J.L., Wang, H.K., Albright, A.R., Lewis, S.M., Bennett, A.R., Huang, J.Y., Ma, H., Wang, Y.X., and Copenhaver, G.P. (2018). Elevated temperature increases meiotic crossover frequency via the interfering (Type I) pathway in Arabidopsis thaliana. Plos Genet *14*. ARTN e1007384 10.1371/journal.pgen.1007384.

55. Lloyd, A., Morgan, C., Franklin, F.C.H., and Bomblies, K. (2018). Plasticity of Meiotic Recombination Rates in Response to Temperature in Arabidopsis. Genetics *208*, 1409-1420. 10.1534/genetics.117.300588.

56. Lucht, J.M., Mauch-Mani, B., Steiner, H.Y., Metraux, J.P., Ryals, J., and Hohn, B. (2002). Pathogen stress increases somatic recombination frequency in Arabidopsis. Nature Genetics *30*, 311-314. 10.1038/ng846.

57. Stevison, L.S., Sefick, S., Rushton, C., and Graze, R.M. (2017). Recombination rate plasticity: revealing mechanisms by design. Philos T R Soc B *372*. ARTN 20160459 10.1098/rstb.2016.0459.

58. Jackson, S., Nielsen, D.M., and Singh, N.D. (2015). Increased exposure to acute thermal stress is associated with a non-linear increase in recombination frequency and an independent linear decrease in fitness in Drosophila. Bmc Evol Biol *15*. ARTN 175 10.1186/s12862-015-0452-8.

59. Lim, J.G.Y., Stine, R.R.W., and Yanowitz, J.L. (2008). Domain-Specific Regulation of Recombination in Caenorhabditis elegans in Response to Temperature, Age and Sex. Genetics *180*, 715-726. 10.1534/genetics.108.090142.

60. Gusa, A., and Jinks-Robertson, S. (2019). Mitotic Recombination and Adaptive Genomic Changes in Human Pathogenic Fungi. Genes (Basel) *10*. 10.3390/genes10110901.

61. Krueger-Hadfield, S.A., Balestreri, C., Schroeder, J., Highfield, A., Helaouet, P., Allum, J., Moate, R., Lohbeck, K.T., Miller, P.I., Riebesell, U., et al. (2014). Genotyping an Emiliania huxleyi (prymnesiophyceae) bloom event in the North Sea reveals evidence of asexual reproduction. Biogeosciences *11*, 5215-5234. 10.5194/bg-11-5215-2014.

62. Wright, S. (1929). Fisher's Theory of Dominance. The American Naturalist *63*, 274-279.

63. Wright, S. (1934). Physiological and Evolutionary Theories of Dominance. The American Naturalist *68*, 24-53.

64. Desai, M.M., Fisher, D.S., and Murray, A.W. (2007). The speed of evolution and maintenance of variation in asexual populations. Curr Biol *17*, 385-394. 10.1016/j.cub.2007.01.072.

65. Kao, K.C., and Sherlock, G. (2008). Molecular characterization of clonal interference during adaptive evolution in asexual populations of Saccharomyces cerevisiae. Nat Genet *40*, 1499-1504. 10.1038/ng.280.

66. Lang, G.I., Botstein, D., and Desai, M.M. (2011). Genetic variation and the fate of beneficial mutations in asexual populations. Genetics *188*, 647-661. 10.1534/genetics.111.128942.

67. Bajic, D., Vila, J.C.C., Blount, Z.D., and Sanchez, A. (2018). On the deformability of an empirical fitness landscape by microbial evolution. Proc Natl Acad Sci U S A *115*, 11286-11291. 10.1073/pnas.1808485115.

68. Lewontin, R.C. (1974). The genetic basis of evolutionary change (Columbia University Press).

69. Crow, J.F. (1970). Genetic Loads and the Cost of Natural Selection. In Mathematical Topics in Population Genetics, K.-i. Kojima, ed. (Springer Berlin Heidelberg), pp. 128-177. 10.1007/978-3-642-46244-3_5.

70. Lu, S.N., Wang, J.Y., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Marchler, G.H., Song, J.S., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Research *48*, D265-D268. 10.1093/nar/gkz991.

71. Osuna-Cruz, C.M., Bilcke, G., Vancaester, E., De Decker, S., Bones, A.M., Winge, P., Poulsen, N., Bulankova, P., Verhelst, B., Audoor, S., et al. (2020). The Seminavis robusta genome provides insights into the evolutionary adaptations of benthic diatoms. Nat Commun *11*, 3320. 10.1038/s41467-020-17191-8.

72. Sayers, E.W., Cavanaugh, M., Clark, K., Pruitt, K.D., Schoch, C.L., Sherry, S.T., and Karsch-Mizrachi, I. (2021). GenBank. Nucleic Acids Res *49*, D92-D96. 10.1093/nar/gkaa1023.

73. Eren, A.M., Esen, O.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015). Anvi'o: an advanced analysis and visualization platformfor 'omics data. Peerj *3*. ARTN e1319 10.7717/peerj.1319.

74. Team, R.C. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

75. Bushnell, B. BBMap. sourceforge.net/projects/bbmap/.

76. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760. 10.1093/bioinformatics/btp324.

77. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079. 10.1093/bioinformatics/btp352.

78. Institute, B. (Accessed: 2019, version 2.6.0). Picard Tools. Broad Institute, GitHub repository.

79. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. 201178. 10.1101/201178 %J bioRxiv.

80. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics *43*, 11 10 11-11 10 33. 10.1002/0471250953.bi1110s43.

81. Smit, A., Hubley, R. (2008-2015). RepeatModeler Open-1.0. <http://www.repeatmasker.org>.

82. Smit, A., Hubley, R & Green, P (2013-2015). RepeatMasker Open-4.0. <http://www.repeatmasker.org>.

83. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842. 10.1093/bioinformatics/btq033.

84. Lindenbaum, P. (2015). JVarkit: java-based utilities for Bioinformatics. figshare.

85. Sovic, I., Sikic, M., Wilm, A., Fenlon, S.N., Chen, S., and Nagarajan, N. (2016). Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun *7*. ARTN 11307 10.1038/ncomms11307.

86. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: An information aesthetic for comparative genomics. Genome Research *19*, 1639-1645. 10.1101/gr.092759.109.

87. Team, R. (2019). RStudio: Integrated Development for R. RStudio.

88. Gel, B., and Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics *33*, 3088-3090. 10.1093/bioinformatics/btx346.

89. Chaisson, M.J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. Bmc Bioinformatics *13*. Artn 238 10.1186/1471-2105-13-238.

90. Rausch, T., Fritz, M.H.Y., Korbel, J.O., and Benes, V. (2019). Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. Bioinformatics *35*, 2489-2491. 10.1093/bioinformatics/bty1007.

91. Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., et al. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature *456*, 239-244. 10.1038/nature07410.

92. Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Trouble, R., et al. (2015). Open science resources for the discovery and analysis of Tara Oceans data. Sci Data *2*, 150023. 10.1038/sdata.2015.23.

93. Delmont, T.O., and Eren, A.M. (2018). Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. Peerj *6*, e4320. 10.7717/peerj.4320.

94. Delmont, T.O., Kiefl, E., Kilinc, O., Esen, O.C., Uysal, I., Rappe, M.S., Giovannoni, S., and Eren, A.M. (2019). Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. Elife *8*. ARTN e46497 10.7554/eLife.46497.

95. Amarasinghe, S.L., Su, S., Dong, X.Y., Zappia, L., Ritchie, M.E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. Genome Biol *21*. ARTN 30 10.1186/s13059-020-1935-5.

96. Williams, R., Peisajovich, S.G., Miller, O.J., Magdassi, S., Tawfik, D.S., and Griffiths, A.D. (2006). Amplification of complex gene libraries by emulsion PCR. Nat Methods *3*, 545-550. 10.1038/nmeth896.

97. Kalle, E., Kubista, M., and Rensing, C. (2014). Multi-template polymerase chain reaction. Biomol Detect Quantif *2*, 11-29. 10.1016/j.bdq.2014.11.002.

98. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D., and Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res *47*, W636-W641. 10.1093/nar/gkz268.

99. Hellemans, J., Mortier, G., De Paepe, A., Speleman, F., and Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biol *8*, R19. 10.1186/gb-2007-8-2-r19.