

**Extreme Lewontin’s paradox in ubiquitous marine phytoplankton species**

Dmitry A. Filatov

Email for correspondence: [Dmitry.Filatov@plants.ox.ac.uk](mailto:Dmitry.Filatov@plants.ox.ac.uk)

Address:

Department of Plant Sciences,

University of Oxford,

South Parks Road,

Oxford OX1 3RB

United Kingdom

## Abstract

Larger populations are expected to have larger genetic diversity. However, as pointed out by Lewontin in 1974, the range of population sizes exceeds the range of genetic diversity by many orders of magnitude (aka “Lewontin’s paradox”, LP). The reasons for LP remain obscure. Here I report an extreme case of LP in astronomically large populations of the ubiquitous unicellular marine phytoplankton species *Emiliania huxleyi* (Haptophyta) – the species that accounts for 10 to 20% of primary productivity in the oceans and its blooms are so extensive that they are visible from space. I demonstrate that despite the wide distribution and enormous population size, the world-wide sample of *E. huxleyi* strains with sequenced genomes represents a single cohesive species and contains surprisingly limited genetic diversity ( $\pi \sim 0.006$  per silent site). The patterns of polymorphism reveal even larger populations in the past, and frequent recombination ( $\rho \sim 0.006$ ) throughout the genome, ruling out demographic history and asexual reproduction as possible causes of low polymorphism in *E. huxleyi*. Natural selection wiping out genetic diversity at linked sites (aka ‘genetic draft’) must be strong and frequent to account for low polymorphism in *E. huxleyi*. This study sheds the first light on poorly understood evolutionary genetic processes in astronomically large populations of marine microplankton.

## Introduction

Extant genetic diversity in a population reflects the balance between the appearance of new genetic variants and their loss. The larger the population, the more mutations can arise in it every generation and the slower is the loss of variation due to genetic drift. Thus, larger populations are expected to contain more genetic diversity. However, as noted by Lewontin (Lewontin 1974) and others (Ellegren and Galtier 2016), the range of population sizes exceeds the range of genetic diversity by many orders of magnitude. The causes of this discrepancy between population size and genetic diversity are still debated (Leffler, et al. 2012; Corbett-Detig, et al. 2015; Coop 2016) and this long-standing problem in evolutionary biology is often referred to as Lewontin’s paradox. The potential solutions to Lewontin’s paradox include (i) demographic factors, such as variation in population size that keep effective population size ( $N_e$ )  $\ll$  census size ( $N_c$ ) (Banks, et al. 2013), (ii) ‘genetic draft’ – loss of diversity due to frequent adaptive evolution (Neher 2013), or, more generally, natural selection acting throughout the genome (Corbett-Detig, et al. 2015) and (iii) lower mutation rate in larger populations, where selection is more effective in reducing mutation rates [the “mutation barrier” hypothesis (Lynch, et al. 2016)]. Yet, it remains unclear which of these

factors play the predominant role in limiting genetic diversity in species with large population sizes (Leffler, et al. 2012).

The previous studies of Lewontin's paradox were focusing on macroscopic multicellular organisms (Corbett-Detig, et al. 2015), which necessarily limited population size of the species and the parameter space of the models analysed. The populations of microscopic eukaryotes remain poorly studied, particularly so for marine microorganisms. To address this knowledge gap I analysed genome-wide genetic diversity in an organism with astronomically large populations – the ubiquitous unicellular marine phytoplankton species *Emiliania huxleyi* (Haptophyta), where the factors, responsible for Lewontin's paradox may be particularly pronounced.

*E. huxleyi* is the most abundant coccolithophore species in modern oceans and is thought to be the main calcite producer on Earth (Paasche 2001; Monteiro, et al. 2016). This species is ubiquitous, but more abundant at high latitudes, forming annual blooms that are so extensive that they are clearly visible from space (Brown and Yoder 1994; Iglesias-Rodriguez, et al. 2002). The termination of blooms results in 'boom-and-bust' population dynamics likely typical for many phytoplankton species. The population sizes during and between the *E. huxleyi* blooms are not known and difficult to assess even to the order of magnitude. Upon cell death its highly distinct calcite 'shields' (coccoliths) sink to the ocean floor, leaving a detailed fossil record and serving as a long-term storage of sequestered carbon (Morse and Mackenzie 1990). The calcite sediment produced by coccolithophores over millions of years have significantly shaped geological features of our planet (e.g. coccoliths are the major component of the White Cliffs of Dover in southeast England) and likely contributed to the global carbon cycle (Westbroek, et al. 1993). Due to its ecological importance, high abundance in modern oceans and ease of culturing in laboratory conditions, *E. huxleyi* serves as a *de facto* model species for numerous studies focusing on the effects of climatic changes, such as ocean acidification, on coccolithophores (Iglesias-Rodriguez, et al. 2008) and the implications of such effects for the global carbon cycle (Rickaby, et al. 2007). However, surprisingly little is known about evolutionary genetic processes in populations of marine plankton generally (Blanc-Mathieu, et al. 2017; Rengefors, et al. 2017) and *E. huxleyi* populations in particular (Bendif, et al. 2016). Here I report the first genome-wide analysis of genetic diversity in this ubiquitous and globally important phytoplankton species.

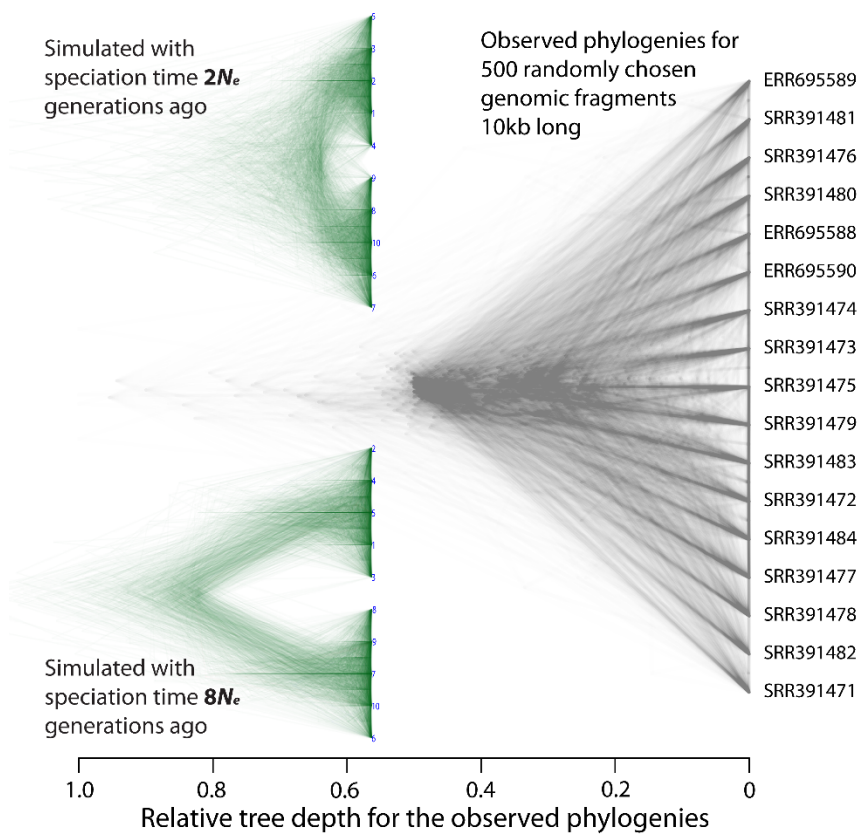
## Results and Discussion

The raw genome sequence data available for 17 *E. huxleyi* strains sampled throughout the world (Table 1) was used to generate single nucleotide polymorphisms (SNPs) dataset as

described in the methods. That polymorphism data was used to address the following questions.

### *Is E. huxleyi a species complex?*

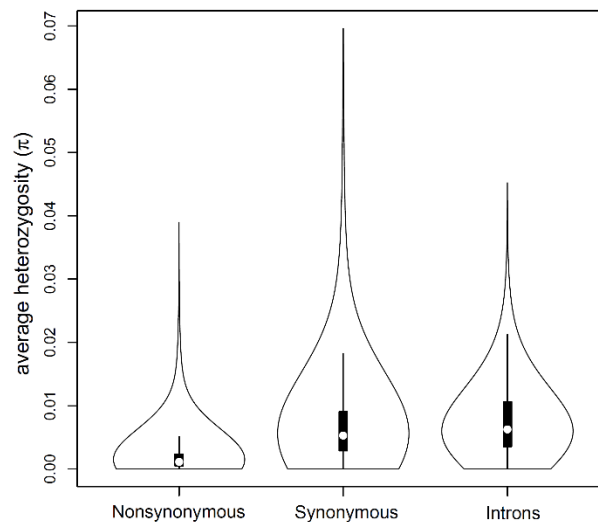
Before embarking on an evolutionary genomic study of a species it is essential to test whether the sample in hand comes from a single cohesive species, or a collection of isolated cryptic species. This is particularly important for microplankton, where relating morphological forms to real biological species can represent a significant problem (Saez, et al. 2003; Seears, et al. 2012; Van den Wyngaert, et al. 2015).



**Figure 1.** *E. huxleyi* samples represent a single cohesive species, as indicated by densiTree plot (Heled and Bouckaert 2013) showing maximum likelihood phylogenies for 500 randomly chosen 10kb fragments of the genome (bigger tree on the right). Lack of sample clustering in this plot is inconsistent with the presence of cryptic species within the *E. huxleyi* set of strains sampled around the world. To make this more apparent, the phylogenies for observed data can be compared to densiTree plots for simulated datasets (smaller trees on the left) corresponding to very recent and older speciation events (species split  $2N_e$  and  $8N_e$  generations ago, respectively). The clustering of samples by species is clearly visible in simulated

data even for very recent species divergence ( $2N_e$  generations ago). The scale bar at the bottom shows the relative depth of individual trees for the observed data.

If the world-wide sample of 17 *E. huxleyi* strains represents two or more cryptic species divergent from each other, one would expect genetic diversity data to reveal the clustering of strains by biological species. To test this I constructed maximum likelihood phylogenies for 500 randomly sampled 10kb fragments from the *E. huxleyi* genome. Using more than 500 genomic fragments for this analysis leads to the same result, but yields too cluttered densitree plot. The resulting 500 phylogenies shown on Figure 1 in the form of densiTree plot (Heled and Bouckaert 2013), reveal no clustering, which is not consistent with the presence of cryptic species in this sample of *E. huxleyi* strains. To illustrate this further I used coalescent simulations to create densiTree plots for samples coming from two different populations (simulated “cryptic species”) that diverged at different points back in time (Figure 1 insets). The clusters corresponding to two simulated populations are clearly visible even for very recently diverged populations. In the extreme case, when all 17 strains represent 17 separate species, the densiTree plot would look like a phylogenetic tree, with gene trees from different loci consistent with each other and the species phylogeny. Thus, if two or more sets of divergent strains were present in the *E. huxleyi* sample, this would be readily apparent in the densiTree plot; lack of such clustering reveals no evidence for the presence of cryptic species among the 17 *E. huxleyi* strains analysed here. Given these strains were sampled all over the world oceans (Table 1), it appears likely that *E. huxleyi* represents a single genetically cohesive species. However, this analysis cannot rule out the existence of yet un-sampled cryptic species within the *E. huxleyi* morphospecies.



**Figure 2.** The distributions of per-nucleotide average heterozygosity ( $\pi$ ) at different types of sites in 14,573 *E. huxleyi* protein coding genes. All pairwise comparisons between these distributions were significant (Wilcoxon rank sum test;  $P < 0.0001$ ).

### *How genetically diverse is E. huxleyi?*

Based on the polymorphism data from 17 sequenced *E. huxleyi* strains (Table 1), single nucleotide polymorphism ( $\pi$ ) across *E. huxleyi* genome is surprisingly low (Figure 2). At synonymous sites, that are regarded to be neutral or nearly neutral (Andolfatto 2005), per nucleotide  $\pi_s = 0.0053 \pm 0.00771$  (median  $\pm$  SD). The polymorphism in introns is slightly, but significantly (Wilcoxon rank sum test;  $P < 0.0001$ ) higher:  $\pi_i = 0.0063 \pm 0.00561$  (median  $\pm$  SD). Polymorphism at non-synonymous sites is five times lower ( $\pi_n = 0.0012 \pm 0.00319$ ), which likely reflects the action of purifying selection that eliminates deleterious mutations affecting protein sequence. Assuming *E. huxleyi* per-nucleotide mutation rate ( $\mu$ ) is of the order of  $\mu \sim 10^{-10}$ , the observed synonymous genetic diversity corresponds to a species-wide effective population size  $N_e \sim \pi/(4\mu) = 25$  million – the number of individuals contained in  $\sim 100$ ml of laboratory culture of that species. This illustrates the extreme disparity between the amount of species-wide genetic diversity and global distribution of that species in world oceans.

Our low estimate of genetic diversity in *E. huxleyi* may appear surprising given the previous report of high diversity in this species (Read, et al. 2013). However, that paper revealed high variation in genome size between *E. huxleyi* strains rather than high single nucleotide polymorphisms analysed here. The genomic bases and the evolutionary consequences of that genome size variation in *E. huxleyi* is beyond the scope of the current study that focuses on single nucleotide variation in regions present in all the sequenced *E. huxleyi* strains (core genome). Large insertion-deletion polymorphism described previously (Read, et al. 2013) may be driven by extremely high rates of chromosomal rearrangements (e.g. due to abundance of repetitive elements in the genome) or by adaptive significance of some of those insertions/deletions. As neither the rates for insertions and deletions nor their adaptive significance are known, it is very difficult to infer anything about population genetic processes driving *E. huxleyi* evolution based on high insertion-deletion polymorphism reported previously (Read, et al. 2013). The only information in (Read, et al. 2013) related to single nucleotide level polymorphism in *E. huxleyi* was the phylogeny shown on figure 3c of that paper that was based on 32 fast evolving genes rather than genome-wide sequence diversity [see supplementary section 2.8 in (Read, et al. 2013)]. The re-analysis of the published (Read, et al. 2013; von Dassow, et al. 2015) genome sequence data revealed much lower per-nucleotide genetic diversity throughout the *E. huxleyi* core genome (Figure 2).

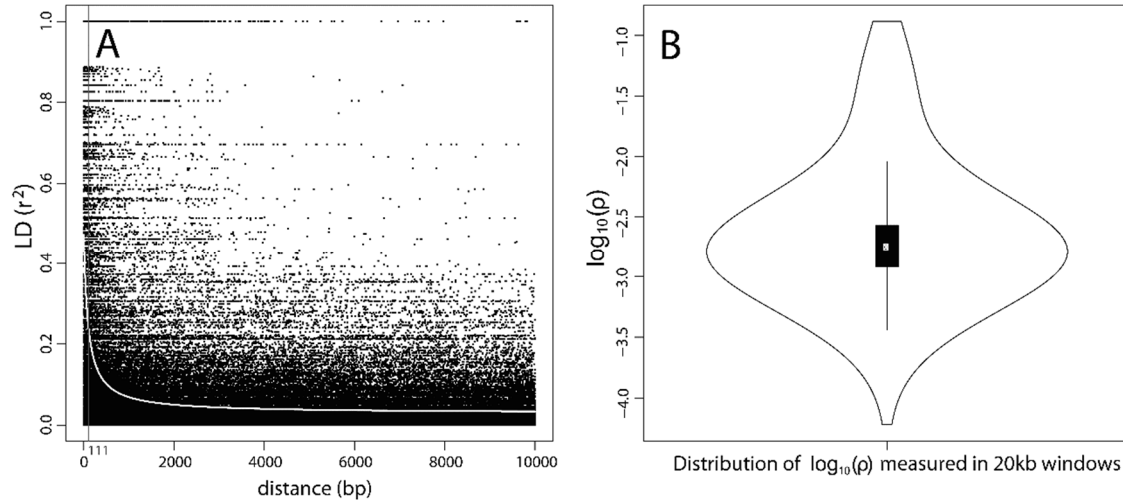
### *Is E. huxleyi predominantly clonal?*

If *E. huxleyi* is predominantly asexual, this may explain its surprisingly low genetic diversity. Asexual reproduction results in high clonality in the population and lack of recombination between the genes, exacerbating loss of genetic diversity throughout the genome due to linked selection, such as selective sweeps affecting the entire genome. Little is known about *E. huxleyi*'s mode of reproduction in the wild. Sexual reproduction has rarely been observed in coccolithophores, though their lifecycle is known to include haploid and diploid phases (Frada, et al. 2012). Hence, I undertook the analysis of recombination and linkage disequilibrium (LD) throughout the *E. huxleyi* genome to assess the contribution of sexual reproduction to the evolution of this species.

If *E. huxleyi* reproduction were predominantly asexual, phylogenies built from different parts of the genome would be consistent with each other. However, the gene trees built from 500 randomly sampled 10kb fragments from the *E. huxleyi* core genome revealed extensive phylogenetic incongruence (Figure 1) that is not consistent with asexual reproduction and indicates that this species regularly goes through the sexual cycle. To explicitly test for the presence of recombination in *E. huxleyi* genome I applied population genetic tests for recombination (McVean, et al. 2002) to 100 randomly selected *E. huxleyi* genomic contigs with the length exceeding 100kb. These tests rejected ( $P < 0.0001$ ) the null model of no recombination in all the genomic contigs analysed. LD rapidly declines with distance; in particular, LD, measured as  $r^2$  between pairs of polymorphic sites declines to  $r^2 < 0.1$  within 1 kb and the distance of half LD decay is only 111 bases (Figure 3A). LD-based estimation of recombination rate [ $\rho$ , (Stumpf and McVean 2003)], measured in 20kb windows, indicates relatively frequent recombination throughout the genome (per nucleotide  $\rho = 0.0068 \pm 0.01514$ ; Figure 3B). Thus, recombination and sexual reproduction are likely to be common in *E. huxleyi* genome and predominantly clonal reproduction cannot be the cause of surprisingly low genetic diversity in this globally distributed species.

The finding of low LD and frequent recombination in *E. huxleyi* genome does not contradict previous report of high clonality in a study of microsatellite variation in an *E. huxleyi* bloom in North sea (Krueger-Hadfield, et al. 2014). Firstly, the blooms are transient events that follow rapid boom-and-bust dynamics and may not contribute significantly to long-term patterns of polymorphism. The LD-based analysis reported above reflects long-term species-wide patterns. Secondly, the LD-based estimates of recombination are 'population scaled', that is  $\rho = 4N_e r$ , where  $N_e$  is effective population size and  $r$  is the probability of a recombination event occurring during meiosis (Stumpf and McVean 2003). Thus, if  $N_e$  is huge, even occasional sexual reproduction (low  $r$ ) may be sufficient to break down non-random associations between the alleles, yielding observed low LD and relatively high  $\rho$  (Figure 3). Unfortunately, no estimates of  $r$  are available for *E. huxleyi* and this species is not amenable

to crosses in the lab, making it difficult to accurately estimate the extent of clonality in the wild based on the analysis of LD. Nevertheless, the  $\rho/\theta$  ratio ( $= 4N_e r / 4N_e \mu = r/\mu$ ) is rather high,  $\rho/\theta \sim 1$  (or rather  $\rho/\pi_s \sim 1$ , with  $\pi_s$  being an estimator of  $\theta$ ), implying that recombination ( $r$ ) and mutation rates ( $\mu$ ) are of the same order of magnitude ( $r/\mu \sim 1$ ) in *E. huxleyi*, which is possible if recombination and sexual reproduction are quite frequent in this species.



**Figure 3.** Linkage disequilibrium and recombination rate in *E. huxleyi* genome. A) The decline of LD ( $r^2$ ) with distance between pairs of sites. The white curve shows nonlinear regression of  $r^2$  on weighted distance using the approach described in (Remington, et al. 2001). The thin vertical line at the left shows the distance of half-decay of LD ( $=111$  nucleotides). B) The distribution of  $\log_{10}$ -transformed per nucleotide population-scaled recombination rate ( $\rho$ ) estimated with maximum likelihood (McVean, et al. 2002) in 20kb non-overlapping genomic windows.

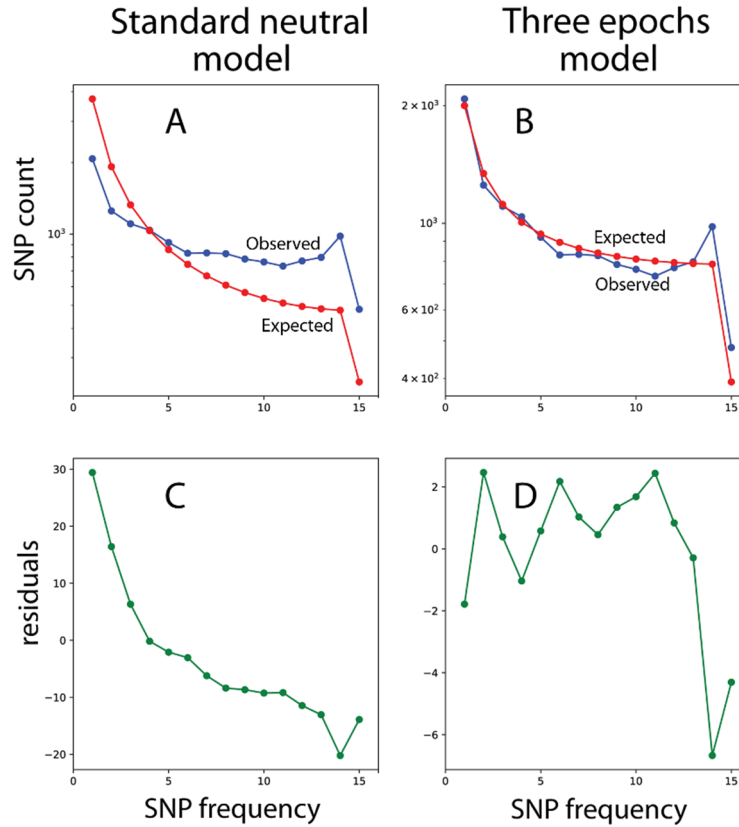
#### *Can demography account for low genetic diversity in E. huxleyi?*

Low genetic diversity in modern *E. huxleyi* populations may be explained by demographic history of this species. In particular, if a population regularly goes through fluctuation of its size, as is the case for *E. huxleyi* seasonal blooms, its effective population size is approximated by the harmonic mean of short term population sizes (Wright 1938). That is, for *E. huxleyi*,  $N_e$  should be closer to the lower population size between the blooms. Nevertheless, given that *E. huxleyi* is very common in the world oceans, its population size during the lower part of its annual fluctuation cycle must still be astronomically large and is unlikely to account for relatively low genetic diversity in this species.

On the other hand, if *E. huxleyi* had a very small population in the past and became abundant very recently, this could explain the lack of genetic diversity in this species. Recent population growth leaves a very distinctive signature in site frequency spectrum (SFS) of



genetic polymorphism – the excess of low frequency and lack of high frequency polymorphisms because a larger population contains more individuals that can mutate. This generates influx of new mutations into the population after population growth, and all the new mutations come at very low frequency, biasing SFS to low frequency variants.



**Figure 4.** Observed site frequency spectrum (SFS) at synonymous sites of *E. huxleyi* and SFS expected under A) standard neutral model and (B) 3epoch model. Panels C and D show deviation of observed from expected values for different site frequencies. The demographic model fit was done with *dadi* software (Gutenkunst, et al. 2009).

To test whether demographic factors, such as recent population size changes, are responsible for low genetic diversity in *E. huxleyi*, I undertook population genetic reconstruction of past demographic history of this species using SFS-based approach implemented in *dadi* software (Gutenkunst, et al. 2009). The comparison of the SFS observed at synonymous sites genome-wide with that expected under the standard neutral model revealed excess of intermediate frequency polymorphisms and lack of low frequency polymorphisms in *E. huxleyi* (Figure 4A), which is the opposite to what would be expected if *E. huxleyi* has expanded from a relatively small population in the recent past. Thus, the data are inconsistent with the scenario that *E. huxleyi* has low genetic diversity due to recent expansion from a very small population. A more complex “three epochs” model that allows

for two population size changes, fits the data significantly better than the standard neutral model (Figure 4B). Reducing this model to two epochs does not significantly reduce model fit to data (Table 2), indicating that a simple scenario with two population sizes – the current and the ancestral population sizes, is sufficient to describe the demographic history of this species. Parameter estimates for this model indicate that ancestral *E. huxleyi* population size has likely been 10 to 20 times larger than the current size ( $n_F \sim 0.04-0.11$ ; Table 2). Sediment data from ocean floor drilling cores indicate that *E. huxleyi* abundance has been at its peak during the last ice age ~70-14 thousand years ago and somewhat diminished since then (Raffi, et al. 2006), which is consistent with our SFS-based demographic analysis.

If *E. huxleyi* population size has indeed contracted ~10- to 20-fold in the recent past, could this account for relatively low genetic diversity in this species? Population contraction is expected to result in some loss of low frequency polymorphisms, but most genetic variation with intermediate allele frequencies is expected to persist in the contracted population and overall loss of genetic diversity should be very modest (Allendorf 1986). Even after the ~10- to 20-fold contraction, the current *E. huxleyi* population size remains astronomical, as this species is very common in oceans around the world. Thus, the post-glacial *E. huxleyi* population contraction is unlikely to account for relatively low overall genetic diversity in this species.

Relative excess of intermediate frequency polymorphisms in *E. huxleyi* SFS (Figure 4) may also be caused by population structure (e.g. due to geography, currents or local adaptation) in this species. However, population structure would result in clustering of the samples by sub-population, which is not apparent in figure 1. Furthermore, population subdivision inflates observed polymorphism for the sample across several populations because the divergence between populations contributes to polymorphism [e.g. (Nei and Takahata 1993)]. Thus, the estimate of genetic diversity based on a world-wide sample is conservative with regard to population structure. That is, if there were population structure in *E. huxleyi*, this would mean that the level of intra-population diversity is even lower than reported here (Figure 2), exacerbating the “Lewontin’s paradox” for this species.

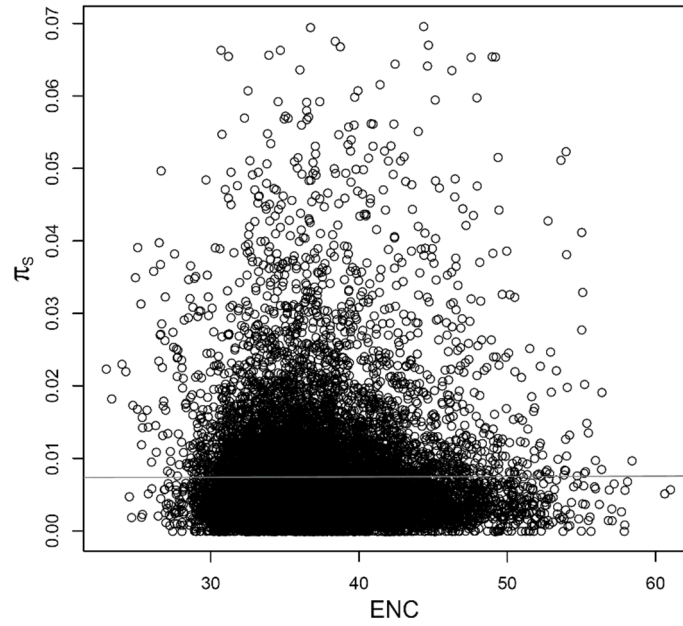
#### *Does multiple merger coalescent resolve the Lewontin’s paradox in E. huxleyi?*

Relatively recent developments in population genetic theory [e.g. see (Tellier and Lemaire 2014)] have led to realisation that classic Kingman’s coalescent (Kingman 1982) may not be the most appropriate model for evolutionary genetic analyses in species with strongly skewed distribution of offspring, as is the case for many marine organisms (Sargsyan and Wakeley 2008) and viruses (Irwin, et al. 2016). When one or few individuals contribute a

large proportion of progeny to the next generation [aka 'sweepstakes reproductive success', (Hedgecock and Pudovkin 2011)], multiple merger coalescent [MMC (Eldon and Wakeley 2006; Tellier and Lemaire 2014)], which is a generalisation of Kingsman's bifurcating coalescent, may be more appropriate. Due to merger of multiple lineages in MMC, it predicts more 'star-like' phylogenies that have shorter internal branches and longer external branches, resulting in lower expectations of neutral diversity and SFS biased towards rare alleles, compared to Kingsman's coalescent (Eldon and Wakeley 2006). Rapid proliferation of one or several clones in *E. huxleyi* blooms is equivalent to highly skewed number of offspring, when one or few individuals contribute large proportion of offspring to the next generation. Although this species does not show signature of high clonality (Figure 3), partial clonality during the plankton blooms (Krueger-Hadfield, et al. 2014), can increase the variance in fitness among individuals and contribute to the low single nucleotide polymorphism. However, this process predicts the bias of SFS towards rare alleles (Eldon and Wakeley 2006), which is the opposite to what is observed in *E. huxleyi* SFS (Figure 4). Furthermore, given that *E. huxleyi* is ubiquitous in the world oceans, the contribution of a single clone to the next generation must be very small. Once the contribution of an individual to next generation is small, the MMC effectively reduces to Kingsman's coalescent (Eldon and Wakeley 2006) and the sweepstakes reproductive success is unlikely to be a major contributing factor to low genetic diversity in *E. huxleyi*.

#### *Can low mutation rate account for low genetic diversity in E. huxleyi?*

The disparity between extremely large census population size and modest *E. huxleyi* genetic diversity may, at least partly be due to a low per generation mutation rate in this species. Although no mutation rate estimates are available for Haptophytes, species with larger population sizes tend to have lower mutation rates, likely because selection to reduce mutation rate further is less effective in smaller population – the so called, "drift-barrier" hypothesis (Lynch, et al. 2016). Thus, very low mutation rate may be a general property of marine plankton. Indeed, the mutation rates in several other unicellular marine algal species were shown to be one to two orders of magnitude lower than in metazoans or plants (Krasovec, et al. 2017). However, even if per-nucleotide mutation rate in *E. huxleyi* is as low as  $10^{-11}$  or  $10^{-12}$ , an order of magnitude lower than in any other organisms studied so far (Lynch, et al. 2016), this would not appear sufficient to explain the relatively modest genetic diversity in *E. huxleyi*, given the census population size of that ubiquitous marine microplankton species is likely many orders of magnitude higher than in macroscopic eukaryotes studied previously [e.g. (Corbett-Detig, et al. 2015)].



**Figure 5.** Genetic diversity at synonymous sites ( $\pi_s$ ) shows no significant correlation with codon bias (ENC) in *E. huxleyi* genome (Pearson  $r = 0.0031$  [95%CI: -0.01318 to 0.01929],  $P = 0.712$ ). The horizontal grey line shows linear regression of  $\pi_s$  on ENC for 14,573 genes analysed.

#### *Can selection on codon usage explain low genetic diversity in E. huxleyi?*

Selection on codon usage could be very efficient in species with very large population size, such as *E. huxleyi*. Indeed, the effective number of codons [ENC; (Wright 1990)] – a common measure of codon bias that ranges from 61 (no codon bias) to 20 (extreme codon bias with only one codon used per amino acid) reveals fairly strong codon bias in *E. huxleyi* ( $ENC = 37.12 \pm 4.694$ ). Direct selection on synonymous variation could significantly reduce polymorphism at synonymous sites and it may partly account for low diversity in *E. huxleyi*. Indeed, synonymous nucleotide diversity ( $\pi_s$ ) is slightly, but significantly lower compared to intron diversity ( $\pi_i$ ; Figure 2). Selection at synonymous sites is expected to inflate non-synonymous to synonymous polymorphism ratio ( $\pi_n/\pi_s$ ). The  $\pi_n/\pi_s$  ratio in *E. huxleyi* is relatively high ( $\pi_n/\pi_s = 0.215$ ), though it falls well within the range  $0.1 < \pi_n/\pi_s < 0.3$  reported for previously studied species (Romiguier, et al. 2014; Chen, et al. 2017). If selection at synonymous sites is a significant contributor to the observed lack of nucleotide polymorphism in *E. huxleyi*, the genes with stronger codon bias are expected to have lower synonymous nucleotide diversity. However,  $\pi_s$  shows no significant correlation with ENC (Figure 5). Furthermore, selection on codon usage is not expected to affect non-coding

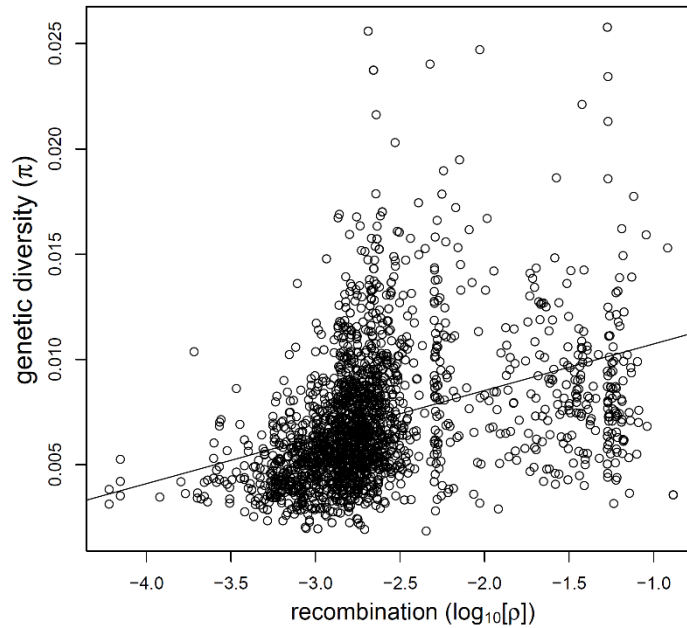
regions, such as introns, which also show relatively low nucleotide polymorphism (Figure 2). Thus, direct selection on synonymous sites does not appear to be a sufficient explanation for low nucleotide diversity throughout *E. huxleyi* genome.

*Is linked selection sufficient to explain limited diversity in astronomically large populations?*

A spread of an adaptive allele eliminates genetic diversity at linked sites – the process widely known as hitchhiking (Smith and Haigh 1974). It has been argued that ‘genetic draft’ – loss of diversity due to frequent adaptive evolution may be responsible for limited genetic diversity observed in species with large populations sizes (Gillespie 2000; Neher 2013). Furthermore, removal of deleterious mutations by purifying selection also results in loss of polymorphism [background selection (Charlesworth, et al. 1995)]. Both of these processes require linkage disequilibrium between the selected site and polymorphisms nearby and thus are often referred to as ‘linked selection’ effects. Linked selection is expected to be stronger in genomic regions with higher LD and positive correlation between recombination rate and genetic diversity was reported for many organisms (Begun and Aquadro 1992). The analysis of nucleotide diversity ( $\pi$ ) and LD-based estimates of recombination rate ( $\rho$ ) in *E. huxleyi* also reveals a positive correlation (Figure 6), indicating that linked selection is common in this species.

There is plentiful evidence that both types of linked selection are affecting the levels of polymorphism across the genomes (Cutter and Payseur 2013). The analysis of polymorphism data across a wide range of animals and plants supports the view that linked selection is a powerful force that contributes to limiting the level of genetic diversity across a wide range of organisms (Corbett-Detig, et al. 2015). However, whether this force is sufficient to explain Lewontin’s paradox remains unclear (Coop 2016).

The previous analyses related to Lewontin’s paradox were restricted to macroscopic eukaryotes with limited population sizes and it is unclear whether their results can be automatically extrapolated to astronomically large populations in microorganisms, such as the eukaryotic marine phytoplankton *E. huxleyi*. In principle, the existing models of linked selection (Kaplan, et al. 1989; Coop and Ralph 2012) indicate that even ‘genetic draft’ alone is sufficient to limit genetic diversity in very large populations. To illustrate this point I plotted the amount of genetic diversity expected in the population under the neutral theory ( $\theta = 4N\mu$ , where  $N$  is population size and  $\mu$  is mutation rate) and under the recurrent selective sweeps model [ $\pi$ ; equation 19 in (Coop and Ralph 2012)] for a range of population sizes (Figure 7).

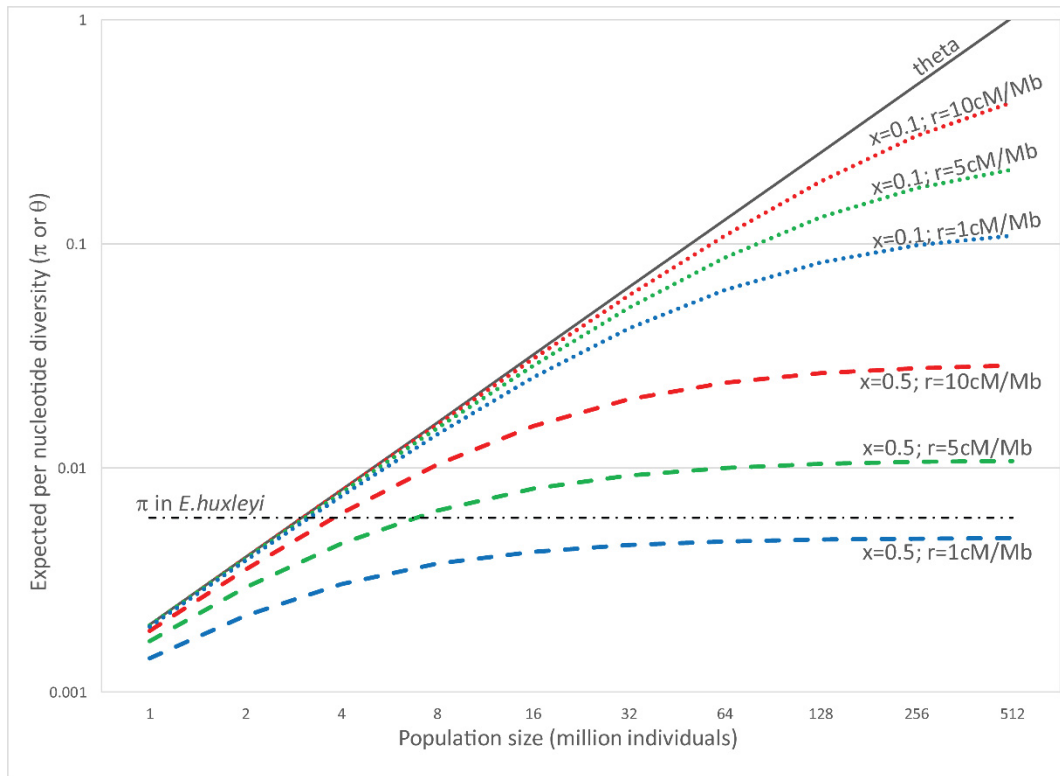


**Figure 6.** Genetic diversity ( $\pi$ ) correlates positively with recombination rate ( $\rho$ ) in *E. huxleyi* genome ( $r = 0.21$  [95%CI: 0.162 to 0.247];  $P < 0.0001$ ). The data points represent 20kb genomic windows and the line shows linear regression of polymorphism on log-transformed recombination rate.

While  $\theta$  scales linearly with population size,  $\pi$  rapidly reaches the plateau and does not increase any further with population size increase (dashed lines on Figure 7). The level of this plateau is determined by the balance between new genetic diversity entering the population and loss of diversity due to recurrent selective sweeps. The latter depends on recombination rate ( $r$ ), frequency of selective sweeps ( $v$ ) and trajectory of selective sweeps (Coop and Ralph 2012), with more frequent “harder” sweeps eliminating more genetic diversity and reducing the level of the plateau. Thus, in principle, limited diversity in *E. huxleyi* and other widespread marine plankton species, may be accounted for by frequent adaptive evolution in the genome.

Unfortunately, very little is known about the adaptation process in oceanic plankton, which makes it difficult to assess the applicability of that model. For example, it may be argued that all adaptation in extremely large populations may happen via incomplete “soft” selective sweeps because  $\theta$  is so large that any possible mutation occurs multiple times every generation (Karasov, et al. 2010; Messer and Petrov 2013). This would substantially limit the power of recurrent selective sweeps to eliminate genetic diversity (dotted lines on Figure 7). For example, in the latter case, as much as 20 soft selective sweeps per megabase per generation would be required to reduce genetic diversity to the level observed in *E. huxleyi*, which seems unrealistically frequent adaptive evolution. Unfortunately, given the paucity of our knowledge about evolutionary genetic processes in marine plankton, it is very difficult to

be certain what parameter values are realistic. In particular, the values in Figure 7 are based on “simple” Wright-Fisher models and it is possible that draft effects are amplified in highly fluctuating populations with strong skew in offspring distribution. Clearly, more work is needed to understand population genetic processes generally and adaptation process specifically in astronomically large and fluctuating populations of microscopic eukaryotes.



**Figure 7.** The level of neutral genetic diversity per nucleotide expected for populations of different size ( $N$ ). Expected genetic diversity under neutrality,  $\theta = 4N\mu$ , is shown with solid straight black line. Expected genetic diversity under the recurrent selective sweeps model [ $\pi$ ; equation 19 in (Coop and Ralph 2012)] for different parameter values is shown with dashed and dotted lines. The more frequent is recombination ( $r$ ) and the ‘softer’ are the sweeps (with lower value of  $x$  - frequency of selected allele after incomplete sweep) the less diversity is lost from the population due to linked selection, resulting in  $\pi$  values closer to that expected under neutrality ( $\theta$ ). Per nucleotide mutation rate ( $\mu$ ) was assumed to be  $\mu = 5 \times 10^{-10}$ ; increasing or decreasing this rate simply shifts the entire plot up or down without changing the shape of the curves. Horizontal dash-dotted line shows the observed level of genetic diversity at silent sites across *E. huxleyi* genome.

## Conclusions

The results of analyses presented above indicate that *E. huxleyi* is a genetically cohesive species that frequently undergoes sexual reproduction. Despite the ‘common wisdom’ of high genetic diversity in ubiquitous marine plankton species, such as *E. huxleyi* [e.g. (Read, et al. 2013)], the actual level of single nucleotide polymorphism is surprisingly low. However, the analyses in the current paper were restricted to the ‘core’ part of *E. huxleyi* genome present in all isolates analysed. It is possible that high morphological and ecological diversity across the range of this species is largely determined by the presence/absence polymorphism of the ‘flexible’ part of *E. huxleyi* genome present only in some strains (Read, et al. 2013). The extent and importance of variation in the ‘flexible’ part of *E. huxleyi* genome will have to be addressed in future studies.

Without extending this work to other species it is not possible to conclude whether low genetic diversity is typical for other marine plankton species. There is very little data on genetic diversity in microscopic eukaryotic plankton (Rengefors, et al. 2017), with most papers focusing on interspecific divergence (de Vargas, et al. 2015). A recent evolutionary genetic study of microscopic marine green algae *Osteococcus tauri* also revealed a relatively modest per-nucleotide diversity level,  $\pi_s \sim 1\%$  (Blanc-Mathieu, et al. 2017). However, it is difficult to build parallels between *O. tauri* and *E. huxleyi* given the former species was described and sampled only from the French Mediterranean coast, while the latter is globally distributed, likely resulting in enormous census population size difference between these species.

The causes of the extreme case of Lewontin’s paradox in *E. huxleyi* remain unclear, as the most obvious explanations – clonality and recent massive population expansion – appear unlikely for *E. huxleyi*. Potentially, low mutation rate in organisms with larger population size (Lynch, et al. 2016) can account for some reduction in genetic diversity in *E. huxleyi*, however, as argued above, this is unlikely to be a sufficient factor. Linked selection appears a plausible explanation for the Lewontin’s paradox in *E. huxleyi*, however the critical parameters, such as the rate and trajectory of selective sweeps remain unknown for marine plankton populations, which makes it difficult to assess whether linked selection is sufficient to fully account for lack of diversity in *E. huxleyi*.

With most evolutionary genetic work focusing on terrestrial organisms, relatively little is known about the microevolutionary processes in marine plankton species (Rengefors, et al. 2017). Population genetic forces may work in rather different ways in relatively small and often subdivided populations of terrestrial organisms and astronomically large populations of marine phytoplankton inhabiting a fairly homogenous environment. Accurate estimates of



even the very basic parameters, such as average silent genetic diversity across the genome, are typically unavailable for marine plankton species (Rengefors, et al. 2017), which is unfortunate, given the huge ecological importance of these tiny but ubiquitous organisms (Westbroek, et al. 1993; Bolton, et al. 2016; Monteiro, et al. 2016). Thus, more evolutionary genetic work is needed in order to understand the microevolutionary processes in microscopic eukaryotes generally and marine phytoplankton in particular.

## Methods

### *Data processing*

The genome sequence data for 14 and 3 *E. huxleyi* previously sequenced strains were obtained from (Read, et al. 2013) and (von Dassow, et al. 2015), respectively. These strains originate from samples collected all over the world oceans (Table 1). The sequence reads from these samples were mapped to *E. huxleyi* reference genome (Read, et al. 2013) with *bwa* v0.7 (Li and Durbin 2009). Duplicate reads were removed with *samtools* v1.2 (<http://samtools.sourceforge.net>), and regions around indels were realigned with *GATK* v3.4 (McKenna, et al. 2010). SNP calling was done for each strain separately, with *samtools* and *bcftools* v 1.2 (part of *samtools* package) using the alternative multiallelic variant caller (-m option) and including homozygous blocks with minimum depth of 8 (-g 8), after excluding reads with mapping quality below 20 (-q 20) and bases with base quality below 20 (-Q 20). SNPs with fewer than 8 reads supporting it, within 3 bp of an indel, with quality below 10, or with fewer than 2 reads supporting each allele (for heterozygous calls) were marked as low-quality and excluded from further analysis. Resulting vcf files including confident SNP calls and homozygous blocks were converted to fasta format using *vcf2fas* (available from <https://github.com/brunonevado/vcf2fas>), with heterozygous SNPs coded with iupac symbols. Coding regions (CDS), annotated in the “Emihu1\_best\_genes.gff” file for the reference genome (Read, et al. 2013), were extracted from the alignments of 17 strains to reference genomic scaffolds using a tool in proSeq software (Filatov 2009). Only complete genes with intact start and stop codons, containing at least 200 synonymous positions were used in the analyses of genetic diversity at synonymous and non-synonymous sites.

### *The analyses of genetic variation*

Average heterozygosity ( $\pi$ ) at different types of sites was calculated using *mstatspop* (Ramos-Onsins, et al. 2018). To test for presence of recombination in *E. huxleyi* genetic diversity data I used coalescent-based likelihood permutation tests for recombination

implemented in LDhat program (McVean, et al. 2002). Pairwise  $r^2$  for pairs of polymorphic sites (Figure 3A) were calculated with proSeq (Filatov 2009) and decay of LD with distance between sites was evaluated by nonlinear regression in R using the approach described in (Remington, et al. 2001). The distribution of population-scaled recombination rate ( $\rho$ ; Figure 3B) across the *E. huxleyi* genome was estimated in 20kb non-overlapping windows with maximum likelihood approach in LDhat (McVean, et al. 2002).

To visualise phylogenetic discordances between loci and possible clustering of samples into cryptic species I used DensiTree plot (Figure 1) based on 500 Maximum Likelihood (ML) trees constructed for the randomly chosen 10 kb fragments of *E. huxleyi* genome. The ML trees were reconstructed using the GTRGAMMA model and 100 bootstrap replicates in RAXML version 8 (Stamatakis 2014). For each ML tree, I used the *pruneTree* function in the R phangorn package (Schliep 2011) in R version 3.1.2 (Team 2014) to collapse nodes with bootstrap support <75%. Trees with no nodes over 75% bootstrap support were discarded. Then each of the pruned trees were made ultrametric using the *chronos* function with default settings in the R *ape* package (Paradis, et al. 2004). Resulting trees were then loaded into DensiTree version 2.2.1 (Heled and Bouckaert 2013) to generate the plot shown on figure 1.

To generate the densiTree plots with simulated datasets (green insets on figure 1) I ran coalescent simulations using program ms (Hudson 1992) with the sample size of 10 (5 per population) and level of polymorphism in the total sample identical to that in the observed dataset. The simulations were run for population split model with no migration. The population split (“speciation”) times ranged from 2 to 8 effective population sizes ( $N_e$ ), representing very recent and older speciation events, respectively. For every set of parameters I generated 500 simulated phylogenies that were plotted with densiTree.

For demographic inference and visualisation of 1-dimensional site frequency spectrum (SFS) I used the *dadi* package (Gutenkunst, et al. 2009). In the absence of an outgroup to establish the ancestral state for each SNP, I used a ‘folded’ frequency spectrum for which all SNPs with frequency  $x$  and  $n - x$  (where  $x$  is SNP frequency and  $n$  is sample size) were pooled together. Following the recommendation in the *dadi* manual, the sample size was ‘projected down’ to 15 samples to account for missing data. To ensure the analyses reach the global maximum I used replicate runs with perturbed starting parameters (*perturb\_params* function in *dadi*), taking the run with the maximal likelihood as the best estimate of the parameter values. Plotting of observed and modelled SFSs (Figure 3) was done with *plot\_1d\_comp\_multinom* command in *dadi* package.

Initially I conducted an exploratory analysis using a set of models of different complexity to find the simplest model that adequately describes the data. The analysis started with the

‘three epochs’ model allowing for a bottleneck and population expansion following the bottleneck. Then the model was progressively simplified excluding free parameters as long as there was no significant reduction in the fit of the model to data. In particular, fixing the  $n_B$  and/or  $n_F$  parameters to 1 the *3epochs* model was reduced to *1epoch*, *2epochs* and *bottleneck* models, which are nested within the *3epochs* model. The significance of the parameters accounting for the bottleneck ( $n_B$ ) and the post-bottleneck population expansion ( $n_F$ ) was tested with likelihood ratio tests (LRTs) comparing the more general model *3epochs* with each of the less parameter rich nested models (Table 2). To compare the models that are not nested I used the Akaike’s Information Criterion (AIC) to rank the different models and used the best fitting model as a reference to calculate the relative likelihood ( $\text{ReLik} = \exp((\text{AIC}_{\text{best}} - \text{AIC}_i)/2)$ , Table 2) that can be interpreted as the probability that the model is the best.

## Acknowledgments

This work was supported by a grant from the John Fell Fund (Grant 152/079 to DAF). The author thanks Mahdi Bendif and Ros Rickaby for insightful discussions regarding the biology and ecology of *E. huxleyi* and Aris Katzourakis for advice and proofreading of the manuscript.

## References

- Allendorf FW. 1986. Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology* 5:181-190.
- Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437:1149-1152.
- Banks SC, Cary GJ, Smith AL, Davies ID, Driscoll DA, Gill AM, Lindenmayer DB, Peakall R. 2013. How does ecological disturbance influence genetic diversity? *Trends Ecol Evol* 28:670-679.
- Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356:519-520.
- Bendif EM, Probert I, Díaz-Rosas F, Thomas D, van den Engh G, Young JR, von Dassow P. 2016. Recent reticulate evolution in the ecologically dominant lineage of Coccolithophores. *Frontiers in Microbiology* 7:784.
- Blanc-Mathieu R, Krasovec M, Hebrard M, Yau S, Desgranges E, Martin J, Schackwitz W, Kuo A, Salin G, Donnadiou C, et al. 2017. Population genomics of picophytoplankton unveils novel chromosome hypervariability. *Science Advances* 3:e1700239.

- Bolton CT, Hernandez-Sanchez MT, Fuertes MA, Gonzalez-Lemos S, Abrevaya L, Mendez-Vicente A, Flores JA, Probert I, Giosan L, Johnson J, et al. 2016. Decrease in coccolithophore calcification and CO<sub>2</sub> since the middle Miocene. *Nature Communications* 7:10284.
- Brown CW, Yoder JA. 1994. Coccolithophorid blooms in the global ocean. *Journal of Geophysical Research - Oceans* 99:7467-7482.
- Charlesworth D, Charlesworth B, Morgan MT. 1995. The pattern of neutral molecular variation under the background selection model. *Genetics* 141:1619-1632.
- Chen J, Glemin S, Lascoux M. 2017. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Mol Biol Evol* 34:1417-1428.
- Coop G. 2016. Does linked selection explain the narrow range of genetic diversity across species? *BioArXiv*. doi: 10.1101/042598
- Coop G, Ralph P. 2012. Patterns of neutral diversity under general models of selective sweeps. *Genetics* 192:205-224.
- Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol* 13:e1002112.
- Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet* 14:262-274.
- de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, et al. 2015. Eukaryotic plankton diversity in the sunlit ocean. *Science* 348:1261605.
- Eldon B, Wakeley J. 2006. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* 172:2621-2633.
- Ellegren H, Galtier N. 2016. Determinants of genetic diversity. *Nat Rev Genet* 17:422-433.
- Filatov DA. 2009. Processing and population genetic analysis of multigenic datasets with ProSeq3 software. *Bioinformatics* 25:3189-3190.
- Frada MJ, Bidle KD, Probert I, de Vargas C. 2012. In situ survey of life cycle phases of the coccolithophore *Emiliana huxleyi* (Haptophyta). *Environ Microbiol* 14:1558-1569.
- Gillespie JH. 2000. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* 155:909-919.
- Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* 5:e1000695.
- Hedgecock D, Pudovkin AI. 2011. Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. *Bulletin of Marine Science* 87:971-1002.

- Heled J, Bouckaert RR. 2013. Looking for trees in the forest: summary tree from posterior samples. *BMC Evolutionary Biology* 13:221.
- Hudson RR. 1992. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337-338.
- Iglesias-Rodriguez MD, Brown CW, Doney SC, Kleypas J, Kolber D, Kolber Z, Hayes PK, Falkowski PG. 2002. Representing key phytoplankton functional groups in ocean carbon cycle models: Coccolithophorids. *Global Biogeochemical Cycles* 16:47-41–47-20.
- Iglesias-Rodriguez MD, Halloran PR, Rickaby REM, Hall IR, Colmenero-Hidalgo E, Gittins JR, Green DRH, Tyrrell T, Gibbs SJ, von Dassow P, et al. 2008. Phytoplankton calcification in a high-CO<sub>2</sub> world. *Science* 320:336-340.
- Irwin KK, Laurent S, Matuszewski S, Vuilleumier S, Ormond L, Shim H, Bank C, Jensen JD. 2016. On the importance of skewed offspring distributions and background selection in virus population genetics. *Heredity* 117:393-399.
- Kaplan NL, Hudson RR, Langley CH. 1989. The "hitchhiking effect" revisited. *Genetics* 123:887-899.
- Karasov T, Messer PW, Petrov DA. 2010. Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *Plos Genetics* 6:e1000924.
- Kingman JFC. 1982. The coalescent. *Stoch Proc Appl* 13:235-248.
- Krasovec M, Eyre-Walker A, Sanchez-Ferandin S, Piganeau G. 2017. Spontaneous mutation rate in the smallest photosynthetic eukaryotes. *Mol Biol Evol* 34:1770-1779.
- Krueger-Hadfield SA, Balestreri C, Schroeder J, Highfield A, Helaouët P, Allum J, Moate R, Lohbeck KT, Miller PI, 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.
- Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol* 10:e1001388.
- Lewontin RC. 1974. The genetic basis of evolutionary change. New York,: Columbia University Press.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760.
- Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* 17:704-714.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297-1303.

- McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160:1231-1241.
- Messer PW, Petrov DA. 2013. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol* 28:659-669.
- Monteiro FM, Bach LT, Brownlee C, Bown P, Rickaby REM, Poulton AJ, Tyrrell T, Beaufort L, Dutkiewicz S, Gibbs S, et al. 2016. Why marine phytoplankton calcify. *Science Advances* 2.
- Morse JW, Mackenzie FT. 1990. Geochemistry of sedimentary carbonates. Amsterdam ; NY, U.S.A.: Elsevier Science Pub. Co.
- Neher RA. 2013. Genetic draft, selective interference, and population genetics of rapid adaptation. *Annu Rev Ecol Evol Syst* 44:195-215.
- Nei M, Takahata N. 1993. Effective population size, genetic diversity, and coalescence time in subdivided populations. *J Mol Evol* 37:240-244.
- Paasche E. 2001. A review of the coccolithophorid *Emiliana huxleyi* (*Prymnesiophyceae*), with particular reference to growth, coccolith formation, and calcification-photosynthesis interactions. *Phycologia* 40:503-529.
- Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289-290.
- Raffi I, Backman J, Fornaciari E, Palike H, Rio D, Lourens L, Hilgen F. 2006. A review of calcareous nannofossil astrobiochronology encompassing the past 25 million years. *Quaternary Science Reviews* 25:3113-3137.
- Ramos-Onsins SE, Ferretti L, Raineri E, Jené J, Marmorini G, Burgos W, Vera G. 2018. *mstatspop*: statistical analysis using multiple populations for genomic data. Available: <https://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html>
- Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J, Monier A, Salamov A, et al. 2013. Pan genome of the phytoplankton *Emiliana* underpins its global distribution. *Nature* 499:209-213.
- Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ESt. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci U S A* 98:11479-11484.
- Rengefors K, Kremp A, Reusch TBH, Wood AM. 2017. Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies. *Journal of Plankton Research* 39:165-179.
- Rickaby REM, Bard E, Sonzogni C, Rostek F, Beaufort L, Barker S, Rees G, Schrag DP. 2007. Coccolith chemistry reveals secular variations in the global ocean carbon cycle? *Earth and Planetary Science Letters* 253:83-95.

- Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, et al. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* 515:261-263.
- Saez AG, Probert I, Geisen M, Quinn P, Young JR, Medlin LK. 2003. Pseudo-cryptic speciation in coccolithophores. *Proc Natl Acad Sci U S A* 100:7163-7168.
- Sargsyan O, Wakeley J. 2008. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor Popul Biol* 74:104-114.
- Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592-593.
- Seeers HA, Darling KF, Wade CM. 2012. Ecological partitioning and diversity in tropical planktonic foraminifera. *BMC Evolutionary Biology* 12:54.
- Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genet Res* 23:23-35.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312-1313.
- Stumpf MP, McVean GA. 2003. Estimating recombination rates from population-genetic data. *Nat Rev Genet* 4:959-968.
- Team RDC. 2014. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- Tellier A, Lemaire C. 2014. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol Ecol* 23:2637-2652.
- Van den Wyngaert S, Most M, Freimann R, Ibelings BW, Spaak P. 2015. Hidden diversity in the freshwater planktonic diatom *Asterionella formosa*. *Mol Ecol* 24:2955-2972.
- von Dassow P, John U, Ogata H, Probert I, Bendif el M, Kegel JU, Audic S, Wincker P, Da Silva C, Claverie JM, et al. 2015. Life-cycle modification in open oceans accounts for genome variability in a cosmopolitan phytoplankton. *ISME J* 9:1365-1377.
- Westbroek P, Brown CW, Vanbleijswijk J, Brownlee C, Brummer GJ, Conte M, Egge J, Fernandez E, Jordan R, Knappertsbusch M, et al. 1993. A model system approach to biological climate forcing - the example of *Emiliania huxleyi*. *Global and Planetary Change* 8:27-46.
- Wright F. 1990. The "effective number of codons" used in a gene. *Gene* 87:23-29.
- Wright S. 1938. Size of a population and breeding structure in relation to evolution. *Science* 87:430-431.

**Table 1.** *E. huxleyi* strains analysed in this study.

<b>Name</b>	<b>Location</b>	<b>Reference</b>	<b>SRA acc#</b>	<b>SRA sample</b>
AWI1516	South Pacific	(Read, et al. 2013)	SRR391471	SAMN00767675
92D	English Channel	(Read, et al. 2013)	SRR391472	SAMN00767676
92E	English Channel	(Read, et al. 2013)	SRR391473	SAMN00767677
92A	English Channel	(Read, et al. 2013)	SRR391474	SAMN00767678
NZEH	New Zealand	(Read, et al. 2013)	SRR391475	SAMN00767679
L	Oslo fjord	(Read, et al. 2013)	SRR391476	SAMN00767680
12-1	Sargasso Sea	(Read, et al. 2013)	SRR391477	SAMN00767681
EH2	Australia	(Read, et al. 2013)	SRR391478	SAMN00767682
M219	New Zealand	(Read, et al. 2013)	SRR391479	SAMN00767683
B11	Bergen Sea	(Read, et al. 2013)	SRR391480	SAMN00767684
B39	Bergen Sea	(Read, et al. 2013)	SRR391481	SAMN00767685
M217	Bergen Sea	(Read, et al. 2013)	SRR391482	SAMN00767686
Van556	Vancouver, BC	(Read, et al. 2013)	SRR391483	SAMN00767687
92F	English Channel	(Read, et al. 2013)	SRR391484	SAMN00767688
CHC428	South Pacific	(von Dassow, et al. 2015)	ERR695590	SAMEA3164474
CHC350	South Pacific	(von Dassow, et al. 2015)	ERR695589	SAMEA3164475
CHC307	South Pacific	(von Dassow, et al. 2015)	ERR695588	SAMEA3164476



**Table 2.** The fit of demographic models to whole genome *E. huxleyi* diversity dataset.

models <sup>1)</sup>	lnL <sup>2)</sup>	#pars <sup>3)</sup>	LRT <sup>4)</sup>	AIC <sup>5)</sup>	RelLik <sup>6)</sup>	$n_B$ <sup>7)</sup>	$n_F$ <sup>8)</sup>	$T_B$ <sup>9)</sup>	$T_F$ <sup>10)</sup>
<i>SNM</i>	-1284.11	0	na	2568.22	0.000	-	-	-	-
<i>1epoch</i>	-1284.12	2	2361***	2572.237	0.000	1 <sup>11)</sup>	1 <sup>11)</sup>	0.150	0.199
<i>bottleneck</i>	-104.34	3	2.1, ns	214.68	0.961	0.085	1 <sup>11)</sup>	0.108	0.001
<i>2epochs</i>	-104.4	3	2.2, ns	214.8	0.905	1 <sup>11)</sup>	0.044	0.076	0.098
<i>3epochs</i>	-103.3	4	na	214.6	1.000	0.002	0.118	0.021	0.024

<sup>1)</sup> models: *SNM* – standard neutral model; *3epochs* – the model with two population size changes separating three epochs of different population size; all other models except *SNM* are nested within *3epochs* model by fixing  $n_B$  and/or  $n_F$  parameters to 1.

<sup>2)</sup> lnL: log-likelihood of model fit to data.

<sup>3)</sup> #pars: number of parameters estimated in the model.

<sup>4)</sup> LRT: likelihood ratio test for the model versus *3epochs*; done only for nested models.

<sup>5)</sup> AIC: Akaike's Information Criterion.

<sup>6)</sup> RelLik: relative likelihood of the model ( $=\exp((AIC_{best} - AIC_i)/2)$ ).

<sup>7)</sup>  $n_B$ : Ratio of bottleneck population size to ancestral population size.

<sup>8)</sup>  $n_F$ : Ratio of contemporary to ancestral population size.

<sup>9)</sup>  $T_B$ : Length of bottleneck (in units of  $2N_a$  generations, where  $N_a$  is ancestral size)

<sup>10)</sup>  $T_F$ : Time since bottleneck recovery (in units of  $2N_a$  generations)

<sup>11)</sup> Parameter fixed to 1