CellPress
OPEN ACCESS

Trends in
Microbiology

Opinion

How does evolution work in superabundant microbes?

Dmitry A. Filatov [1,*] and Mark Kirkpatrick[2]

Marine phytoplankton play crucial roles in the Earth's ecological, chemical, and geological processes. They are responsible for about half of global primary production and drive the ocean biological carbon pump. Understanding how plankton species may adapt to the Earth's rapidly changing environments is evidently an urgent priority. This problem requires evolutionary genetic approaches as evolution occurs at the level of allele frequency change within populations driven by genetic drift and natural selection (microevolution). Plankters such as the coccolithophore *Gephyrocapsa huxleyi* and the cyanobacterium *Prochlorococcus 'marinus'* are among Earth's most abundant organisms. In this opinion paper we discuss how evolution in astronomically large populations of superabundant microbes (SAMs) may act fundamentally differently than it does in the populations of more modest size found in well-studied organisms. This offers exciting opportunities to study evolution in the conditions that have yet to be explored and also leads to unique challenges. Exploring these opportunities and challenges is the goal of this article.

Highlights

Superabundant microbes (SAMs), such as marine phytoplankton, are extremely important in Earth's ecosystems.

Evolutionary processes have been mainly studied in much smaller populations, and it is poorly understood how evolution works in the huge populations of SAMs.

The standard tools of evolutionary genetics were developed for populations of smaller sizes, and many of these tools may not be suitable to study evolution in astronomically large SAM populations.

Nucleotide sequence polymorphism is surprisingly low in SAMs, corresponding to a population of a few million individuals, but the reasons for this lack of diversity are unclear.

Why is understanding evolution in SAMs so important?

The importance of SAMs, such as marine phytoplankton, is difficult to underestimate – they form the basis of the food chain and are responsible for about half of newly produced organic matter on the planet and half of the oxygen that we breathe [1]. Understanding how evolutionary processes operate in marine plankton is critical for predicting their ability to spread (e.g., polar-wards [2]), adapt to ever-changing environments (e.g., being constantly advected by currents [2–4]), and their resilience to rapid global environmental change [5,6]. Yet, surprisingly little is known about population genetic processes underpinning evolution of these microscopic but hugely important organisms [7,8]. Evolutionary genetic processes have largely been studied for organisms whose population sizes are relatively small (e.g., primates) to relatively large (e.g., *Drosophila*). The population sizes of SAMs, however, are yet larger by many orders of magnitude. For example, the census sizes of marine phytoplankton coccolithophore *Gephyrocapsa* (ex- *Emiliania*) *huxleyi* and cyanobacterium *Prochlorococcus* are of the order 10^{22} [9] and 10^{28} [10] cells, respectively – truly astronomical values that are comparable with the number of stars in the Universe. During its seasonal blooms, *G. huxleyi* can be so abundant as to be visible from space despite the microscopic cell size (~5 μm). In the following text we discuss why evolutionary processes in these vast populations may operate differently compared with much smaller populations [11–13].

Over the past 50 years, geneticists have developed an extensive statistical toolbox to study evolution at the levels of genes and populations (e.g., [14,15]). These powerful evolutionary approaches can be informative about many aspects of biology and evolution in marine microbes, such as presence/absence of sexual reproduction in non-model organisms [16–18], genome evolution [19], including the role of accessory genes in pangenomes of microorganisms [20],

[1]Department of Biology, University of Oxford, Oxford, OX1 3RB, UK
[2]Department of Integrative Biology, University of Texas, Austin, TX 78712, USA

*Correspondence:
dmitry.filatov@biology.ox.ac.uk
(D.A. Filatov).

adaptation to local environmental conditions [21–23], and speciation in marine microorganisms [13,24]. Evolutionary genetics could also be used to infer past environmental conditions [25] and to study the evolutionary processes underpinning patterns seen in the fossil record. For example, a recent paleontological study reported that climatic changes associated with variation in the Earth's orbit are driving cyclic changes in morphology in the dominant Cenozoic Noelaerhabdaceae family of coccolithophores [26]. An evolutionary genetic analysis provided complementary insight: oscillations in size and abundance of coccolithophore fossils were caused by consecutive radiations and extinctions of species rather than variation in the abundance of species with different cell sizes [27].

Evolutionary genetic reconstructions of past population size changes of marine phytoplankton species (e.g., [24,27]) may be important for inferences in ocean biogeochemistry; for example, foraminifera are widely used as proxies for ocean surface temperatures [28] and reconstructing their abundance through time would be valuable for paleoclimatic modelling. Global warming has raised concerns about the resilience of marine phytoplankton to rising sea temperature, ocean acidification [6] and plankton feedback to climate change [29]. The studies of phytoplankton performance in different conditions [30,31], revealed that strains of the same species isolated from different locations are well adapted to local environmental conditions [32,33], which implies surprisingly rapid adaptation, given that they are constantly advected by currents [3,4]. Without understanding the evolutionary genetic processes underpinning this adaptation it is difficult to predict the ability of phytoplankton species to adapt to rapidly changing environment [34] and to affect global climate [35].

In this article we explore how the extreme population sizes of superabundant microbes may cause them to evolve in unusual, perhaps unique, ways. We discuss which of the standard methods of molecular evolution are applicable to SAMs and which can go wrong or need modifications to accommodate the unusual biology of these organisms. We identify the methods that appear to be unsuitable to SAMs, and suggest that they are priorities for the development of new theory and statistics.

Why does population size matter?

The role of random genetic drift and selection

The role that population size and genetic drift play in evolution (Box 1) was first worked out in the early 20th century by Fisher, Wright, and Haldane; further important advances were later made by

Box 1. Drift, population size, and genetic diversity

Population size determines the number of new mutations occurring in the population and the rate of loss of mutations by random genetic drift – stochastic variation in allele frequency over generations. Drift, which is stronger in smaller populations and weaker in larger ones, is typically quantified by the 'effective population size', symbolised as N_e [71,90]. While there are several definitions of N_e [90], for our purposes it can be thought of as the size of an imaginary 'ideal' population that has allele frequency fluctuations as large as those in the population of interest. The size of the ideal population is constant, and all individuals make equal contributions to the next generation [71].

The larger the population, the more genes mutate every generation, increasing the influx of genetic diversity. These mutations experience genetic drift and can be lost by chance, which limits overall genetic diversity in the population. Genetic diversity (π) is defined as the chance that two copies of a gene randomly chosen from a population will carry a different nucleotide at a given site. At a site in the genome that is free of selection and in a population of constant size, the balance between the influx of new mutations and their loss by genetic drift is determined by the product of the effective population size (N_e) and the mutation rate (μ):

$$\pi = 4N_e\mu \quad [I]$$

The factor of 4 in Equation I pertains to diploids, and is replaced by a factor of 2 for haploids.

Kimura [36]. Their theory shows that, for genomic sites that are free from selection ('neutral'), nucleotide diversity (π) is proportional to the 'effective size' of the population, N_e, and the mutation rate, μ (see Equation I in Box 1). When an estimate of μ is available, this simple relation can be used to estimate N_e from molecular data. At fourfold degenerate sites (thought to be the most neutral sites in the genome [37]), nucleotide diversity in the coccolithophore *G. huxleyi* is $\pi \sim 0.005$ [38], while in the cyanobacterium *Prochlorococcus* $\pi = 0.005$–0.041 [10]. Estimates for their mutation rates are respectively $\mu = 5 \times 10^{-10}$–6×10^{-10} [39] and $\mu = 2 \times 10^{-10}$–5×10^{-10} [10] per site per cell division. From Equation I (Box 1), these data imply that the effective population sizes for these microbes are of the order $N_e = 10^6$–10^7. Similar estimates come from other abundant marine plankton: $N_e \sim 10^7$ for the unicellular green alga *Osteococcus taurii* [18] and the dinoflagellate *Alexandrium ostenfeldii* [40]. What is so striking about these results is that they are wildly at variance with direct observations of the population sizes of these microbes: these numbers of cells can be found in just 500 ml of seawater.

The huge disparities between the expected and observed nucleotide diversities in these species represent extreme instances of a general phenomenon called 'Lewontin's paradox' [38] – the unexpectedly low genetic diversities in very large populations that was described by R.C. Lewontin back in 1974 [41]. A variety of hypotheses have been offered to account for this paradox [42], but a general consensus has yet to be reached [43]. Some marine microbes have largely clonal (asexual) reproduction, which could, at least partly, account for reduced genetic diversity [16], because the loss of polymorphism due to genetic hitch-hiking, such as selective sweeps (Figure 1), is much more extensive in clonal populations. However, this explanation does not apply to species, such as diatoms, in which sex and recombination are frequent. A leading hypothesis is that genetic diversity is reduced by population bottlenecks [42]. Indeed, populations of marine plankton tend to follow bloom-and-bust dynamics. Following a bottleneck, π is slow to grow back to its equilibrium value, and so N_e estimated from π is expected to be close to the population size during the bottleneck for a long period afterwards [42]. It is hard to imagine, however, that even during a bottleneck the total number of cells of a globally distributed plankton species such as *G. huxleyi* was as low as only ~ten million. Any population subdivision, for example, caused by

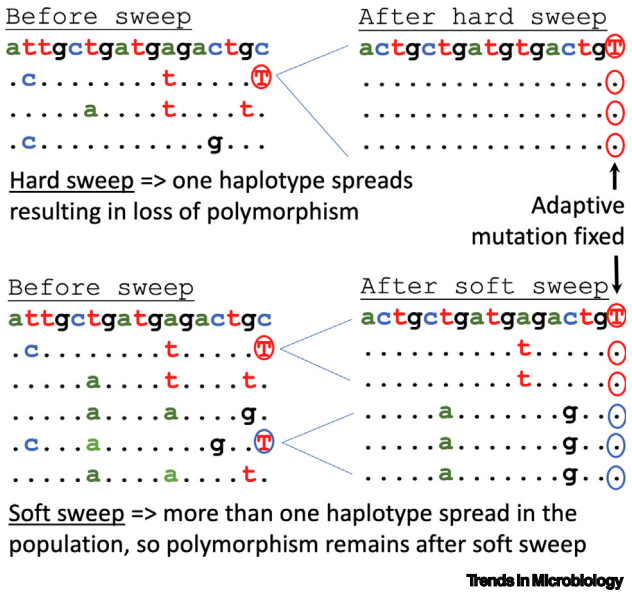

Figure 1. The effects of adaptation on genetic diversity. Spread and fixation of a single adaptive mutation (circled) results in loss of genetic diversity – a so-called 'selective sweep' [71] or 'hard sweep', as shown in the two upper panels. Conversely, if the adaptive mutation arises more than once independently, their spread does not eliminate all genetic variation in that region of the genome, making adaptation difficult to detect – a so-called 'soft sweep' [84], shown in the two bottom panels. Each panel shows a sequence alignment with dots standing for the same nucleotide as in the top row to highlight the nucleotide polymorphisms. Red and blue circles around the adaptive allele show the same mutation that arose independently in different haplotypes.

distinct environmental conditions [44] or partial isolation between water bodies [45], would only increase overall genetic diversity (and N_e estimated from species-wide π) due to divergence between subpopulations, limiting the effect of bottlenecks on genetic diversity.

We propose an explanation for Lewontin's paradox in SAMs that is often neglected in the literature. The relative contributions of drift and natural selection to the evolution of a mutation is determined by the population's effective size (N_e) and the mutation's 'selection coefficient', symbolized by s (Box 2). A key point is that when $|N_e s| < 1$, the mutation is expected to evolve as if it is neutral, even if it is not (i.e., $|s| > 0$). For SAMs such as *G. huxleyi* and *Prochlorococcus*, it is plausible that the true N_e (rather than N_e estimated from π) is so large that very few or even no mutations have fitness effects sufficiently small that $|N_e s| < 1$. That is, virtually all mutations are deleterious and are quickly removed by natural selection, causing genetic diversity to be much smaller than expected from Equation I (Box 1). Occasionally arising adaptive mutations also do not contribute much to polymorphism because they spread and fix quickly. This may be sufficient to account for the extreme examples of Lewontin's paradox in SAMs.

Putting this discussion into an historical context: the hypothesis that all mutations experience selection and that most are deleterious harkens back to the 'classical' or 'panselectionist' view during the famous debate between those who favoured Kimura's neutral model of molecular evolution and those who did not [41]. While a modified version of Kimura's model – the 'nearly neutral theory' [46] – is widely accepted for evolutionary genetic processes in populations of almost all organisms (eukaryotes and prokaryotes alike), the panselectionist view may be more suitable for the astronomical population sizes of SAMs, such as marine phytoplankton.

The input of beneficial mutations

A second crucial role that population size plays in evolution is to modulate the number of beneficial mutations that enter a population. As the census size of the population (N_c) grows, the total number of mutations entering the population each generation ($= N_c \mu$) increases. This effect allows larger populations to adapt more quickly, for example, to changing environments. *G. huxleyi* has a per-site mutation rate of $\sim 5 \times 10^{-10}$ [39]. Its census population size is conservatively estimated at $N_c = 10^{22}$ cells [9]. These divide at a rate of about once a day in laboratory culture [39], and likely at a much lower rate when the environmental conditions are not ideal. Say that cells in natural populations divide once a week on average. Then every base pair in the genome mutates somewhere in the population about 10^{11} times per day (and even more frequently when conditions are favourable, e.g., during blooms). The numbers for *Prochlorococcus*, and probably for SAMs generally, are equally striking. With this kind of mutational saturation, there is no waiting time for the adaptive mutations to arise. Adaptation is expected to proceed quickly (with timescales of months to years), and to result from the spread of adaptive alleles that arise many times independently (Figure 1). This process has been seen during the evolution of insecticide resistance in *Drosophila* [47]. Given that populations of SAMs are many orders of magnitude larger,

Box 2. Population size and selection

Unlike drift, selection causes systematic changes in allele frequencies, with positive selection increasing the frequencies of adaptive alleles and negative selection reducing frequencies of deleterious alleles. The strength of selection is quantified by the 'selection coefficient' (s). This quantity is the proportional increase or decrease that the mutation makes to an individual's 'fitness', that is, the number of offspring it expects to contribute to the next generation. The selection coefficient is negative when the mutation is deleterious and positive when advantageous. When a mutation has no effect whatever on fitness, then $s = 0$, and we say it is 'neutral'. But, as the nearly neutral theory [46] tells us, a mutation with non-zero s can evolve as if neutral when $|N_e s| < 1$ because its change in frequency over time is dominated by drift. The extremely large values of N_e in SAMs require s to be extremely close to zero to keep $|N_e s| < 1$, which may mean that none of their mutations evolve as if neutral. This situation may have no parallel in the vast majority of species on Earth.

this evolutionary regime is likely the predominant way that adaptation works in SAMs. Whether this is the case could be tested by analysing patterns of molecular variation, as discussed later.

Recombination, linkage disequilibrium, and population size

The key evolutionary effect of recombination is to break down nonrandom associations (linkage disequilibrium, LD) between alleles at different loci. This can have a profound effect on adaptation: decreased LD allows alleles at different loci to evolve independently [48]. Consider this extreme case: in a population without recombination, allele A_1 at locus A always occurs with allele B_1 at locus B. Then selection that causes A_1 to spread to fixation will also cause fixation of B_1 even if that allele is completely free of selection. If loci A and B both experience selection, 'selective interference' occurs, and neither adapts as quickly as it would in the absence of the other. Selective interference can be particularly strong in asexual species, where LD builds up across the entire genome. Further, the strength of selective interference grows with population size.

Blooms of phytoplankton, such as coccolithophore *G. huxleyi*, are thought to be dominated by the rapid growth of multiple clones that mainly reproduce asexually [49] (though, sex during blooms has been reported, e.g., in diatoms [50,51]). It is therefore surprising that *G. huxleyi* has extremely low LD [38]. How can we reconcile frequent clonal reproduction and low LD? The extent of LD depends on how much recombination occurs in the population, which is measured by population-scaled recombination rate (ρ). As $\rho (= 4N_e r)$ depends on the product of per-individual recombination rate (r) and the effective population size (N_e), in very large populations ρ can be large (and LD small) even if r is low. Thus, even if sexual reproduction is infrequent, there is enough recombination in a very large population to break down LD. This means that even in the SAM species with relatively rare sexual reproduction, LD is likely low and even the sites at short distances from each other are independent in evolutionary sense, with relatively little selective interference occurring. Low LD and selective interference in SAM genomes ensure higher efficacy of selection, which should help their adaptation.

Are the current evolutionary genetic approaches applicable to SAMs?

Many of the evolutionary genetic models (e.g., [52]) assume infinitely large populations, which is a reasonable approximation for very large SAM populations. However, as discussed earlier, SAMs likely violate the assumptions of the nearly neutral theory [46] which serves as a foundation for many of the widely used evolutionary genetic approaches that can be problematic for SAMs. A simple (but highly conservative) statistical test for adaptation at a gene compares the rate of substitution at silent (synonymous) sites, which is denoted as K_s (or D_s), with the rate at non-silent (nonsynonymous) sites, denoted as K_a (or D_n) [53]. One typically assumes that non-silent sites evolve under selection, while silent sites evolve neutrally, which is likely incorrect for extremely large populations of SAM, as discussed earlier. Most often, selection is purifying, which decreases the substitution rate at non-silent sites, and so the K_a/K_s is less than 1. Conversely, when positive selection (adaptation) does occur at nonsynonymous sites, it increases their substitution rates. A K_a/K_s ratio that exceeds 1 is therefore taken as evidence of adaptation at a gene [53]. But what if silent sites do in fact experience selection? The prediction is then less clear. A plausible argument is that selection (both purifying and positive) will generally be weaker on silent than on non-silent sites. If so, then $K_a{:}K_s > 1$ would again suggest adaptation is occurring at the non-silent sites. The Hudson–Kreitman–Aguade (HKA) [54] and the McDonald–Kreitman (MK) [55] tests to detect selection in DNA sequence data are more powerful than the K_a/K_s ratio, but they are also more sensitive to violations of the assumption that silent sites evolve neutrally. The distribution of fitness effects (DFE)-alpha method [56], which estimates the fitness effects of new mutations and the fraction of substitutions caused by selection and by drift, will likewise fail if no sites in the genome are evolving neutrally. The suitability of these approaches for SAMs

is questionable, but the use of pseudogenes as a neutral reference [20] may make these approaches applicable to SAMs if mutations in pseudogenes are neutral.

Another theory-related problem is that SAMs may violate the assumptions of the 'coalescent' theory [53,57] that is the foundation for many of our inferences in evolutionary genetics (Box 3). A key assumption in this framework is that the evolutionary histories of genes (genealogies) are bifurcating trees. However, this assumption is likely to be violated in many marine organisms [58], including SAMs, where seasonal phytoplankton blooms can be dominated by a few actively reproducing clones [49]. This may cause multiple branches in the genealogy to descend from a single ancestor, a highly successful clone that disproportionately contributed to future generations. This would violate the classical coalescent model, resulting in more 'star-like' gene genealogies with shorter internal branches (Box 3), decreased neutral diversity (π), and altered distributions of allele frequencies. Among the consequences are that standard statistics to estimate demographic history may fail. Tajima's *D* [59] is a widely-used statistic that is based on the distribution of allele frequencies. A negative value of *D* is often taken as evidence of recent population growth [59], but negative values can also result from multiple mergers in gene genealogies [60]. The good news here is that alternatives to the standard coalescent model ['multiple merger coalescent' (MMC) models] are being developed that could be appropriate to SAMs [58,60]. An application of MMC in microbial population genetics revealed that previous conclusions based on standard coalescent process may need to be revised [61].

Box 3. Use of coalescent theory in experimental evolutionary genetic studies

Coalescent theory [53,71] can be used to create the null expectation for the patterns of polymorphism under a certain demographic scenario. For this purpose, multiple random gene trees (genealogies) are generated (Figure IA–C), and mutations randomly added to them according to a set of rules from coalescent theory [57]. This creates a set of simulated datasets of the same size and level of polymorphism as the real dataset. Statistics such as π and Tajima's *D* [59] are calculated for the simulated and the observed datasets. If the value of a statistic in the observed data falls in the tail of the distribution obtained from the simulated datasets (Figure ID), the real sample significantly deviates from the null hypothesis. For example, in a stable population (Figure IA) the distribution of the statistic is centred around zero, while in an expanding population (Figure IB) it is shifted to negative values (Figure ID). The multiple merger coalescent (MMC) (Figure IC) makes different predictions compared with the standard coalescent (Figure ID), which can lead to mis-inference. For example, a stable population evolving with multifurcating genealogies can be mistaken for an expanding population if the standard coalescent is used instead of MMC.

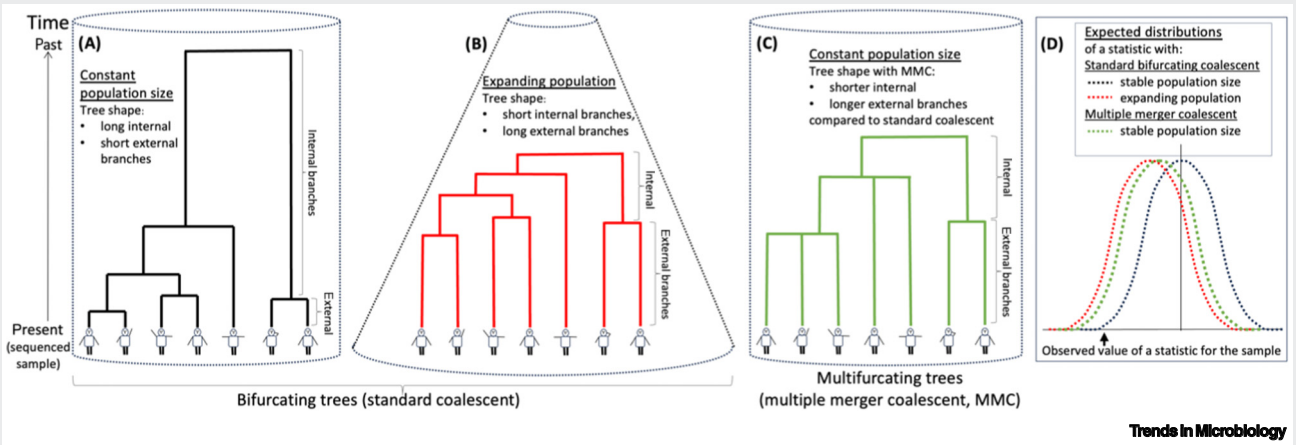

Figure I. A typical shape of gene genealogies generated with standard coalescent with constant (A) and expanding (B) populations, and with a multiple merger coalescent (MMC) (C). These models lead to different expected distributions for a statistic of interest (D). The little figures below the trees illustrate that the external branches lead to sequenced individuals in the sample from the same species.

The very large populations of SAMs confront phylogenetic reconstruction with two difficulties. The first is 'incomplete lineage sorting' (ILS), which occurs when the times between successive nodes on a gene tree (the 'coalescent events') are greater than the times between phylogenetic branching events (speciation) [62]. The result of ILS is that the tree inferred from a gene will often be incongruent with the true phylogenetic relationship in the species tree. As the extent of ILS is proportional to the population size at the time of speciation (see Figure 4 in [63]), ILS is expected to be extensive in SAMs, unless speciation is associated with a population bottleneck, for example, when a new species forms in a small lagoon cut off from the sea or via a genome rearrangement, such as polyploidization, that creates a reproductive barrier [13]. The second difficulty is 'mutational saturation', which blurs the phylogenetic signal when recurrent mutations occur independently in different lineages, is expected to occur in very large populations when $N_e \mu >> 1$ [64]. In principle, phylogenetic methods based on appropriate assumptions [65] can accommodate both ILS and mutational saturation, but the phylogenetic signal may be weaker than for species with modest population sizes. Recently, methods have been proposed that use genome sequences to delimit species boundaries that are explicitly suited to SAMs [64]. They do, however, forfeit the goal of finding the phylogenetic relations between the species.

What approaches can we use to study the evolution of SAMs?

Microbial evolution is often studied in microcosm [66,67] or reciprocal transplant [33] experiments in the laboratory. However, laboratory-based microcosms have limited capacity and even semi-natural mesocosm experiments [68] can only accommodate population sizes that are many orders of magnitude smaller than natural SAM populations in world oceans. Given the importance of population size for the ways evolution works (discussed earlier), it is important to study evolutionary processes in natural SAM populations. In the following text we discuss what evolutionary genetic approaches are suitable for this purpose and what tools are likely to fail in astronomically large SAM populations.

We have seen that superabundant microbes will challenge many of the evolutionary genetic methods, but some of the existing approaches may be useful and accurate. Intuitively, LD-based statistics, such as Z_{nS} [69], and the statistics based on allele frequency, such as Tajima's *D* [59], may be suitable for SAMs. These applications would, however, require care in choice of the appropriate null models, perhaps replacing standard coalescent with MMC (as discussed earlier), or using an empirical distribution of the statistic across the genome. The allele-frequency-based analyses are informative about population structure, past species demography and selective pressures (e.g., [23]). Demographic inferences based on allele frequency distributions (e.g., [70]) may also be used in SAMs with models that account for the nonneutrality of most polymorphisms in very large populations. Such approaches are useful to study past population and species dynamics, infer population size changes through time and estimate the rate of interspecific gene flow [24,26–28] after speciation. This can be very informative about the ways new species form in SAMs [13].

Other evolutionary genetic approaches based on allele frequencies can also be adapted for the analysis of SAM data. Clines, which are smooth spatial gradients in allele frequencies or phenotypes, can form when loci adapt to environmental gradients [71]. Many marine plankton show clines associated with latitude, depth, salinity, and other environmental variables [72–74]. The spatial form of these clines could be used to estimate important evolutionary quantities such as how rapidly selection varies in space and the rates of movement between populations [75,76]. It is worth noting that the deterministic models of clines [52] assume infinite population size, which is a reasonable approximation for SAMs. The genome-wide analysis of clines in SAMs could reveal the number of loci adapted to local environmental conditions. Correlations between

phenotypic traits, environmental conditions, and allele frequencies could identify the genes contributing to locally adapted phenotypes [23,77].

Codon bias occurs when different codons that correspond to the same amino acid occur at unequal frequencies across the genome. The classical explanation, which involves very weak selection acting on these synonymous (silent) alleles, is expected to produce stronger codon bias in larger populations as selection becomes more powerful relative to drift [78]. The recent discovery that phytoplankton species do not show very strong codon bias therefore came as an intriguing surprise [79]. These results may point to alternative hypotheses for codon bias [80–82]. The vast population sizes of SAMs provide unique opportunities to test these ideas.

Over the past decade, a major goal of evolutionary genomics has been to identify regions of the genome involved in recent adaptation. One approach that is widely used for this purpose is to search for regions of the genome that were depleted of diversity as beneficial mutations spread to fixation (Figure 1). Parameter-free approaches (e.g., [83]) could be used to detect selective sweeps in SAMs. But the extremely large population sizes may make sweeps of this sort very rare. As discussed earlier, in marine phytoplankton and other species with populations so large that $N\mu >> 1$, adaptive mutations are likely to arise many times independently (Figure 1). In that case, windows of low diversity are not expected to form in populations that are roughly constant in size [84] – a prediction that can be tested in SAMs.

Another useful application of DNA polymorphism data analysis is to estimate rates of recombination, sexual reproduction, and self-fertilization in natural populations [16,18,85–88]. Some of the statistical methods developed for this purpose rely on the assumption of neutral evolution at silent sites [87] so in their present form they may be inappropriate for use with SAMs. Instead, it may be safer to use heuristic methods to detect recombination in superabundant microbes (e.g., [88]) as they do not depend on explicit models of evolution. The relationship between population-scaled recombination rate (ρ) and N_e mentioned earlier ($\rho = 4N_e r$) provides a way to estimate effective population size from LD [89] independently from Equation I in Box 1, but this has not yet been done for any SAM species.

A very different perspective on the evolution of SAMs would come from studies that track allele frequency changes in time. These time series can be used to directly measure genetic drift in real time, and the so-called 'variance effective population size' [90]. Unlike the estimates of N_e from Equation I in Box 1, which averages over long time periods, this approach yields estimates of the current N_e that are unaffected by population bottlenecks in the past. Selection results in time series that look quite different than those caused by drift: it produces consistent directional changes in allele frequencies. Thus, time series are able to parse out the contributions of drift and selection to evolutionary change. This sort of analysis has been done with bacterial populations in the laboratory [67,91], but has not been attempted for any free-living marine microorganism. Even time series sampled over just a few years may be sufficiently long to study adaptation in SAMs, as they can go through many generations per year. Such analyses would be informative about the timescale required for adaptation in SAMs to occur – is it fast enough for SAMs to adapt to seasonal changes, or even to rapidly changing conditions during a single phytoplankton bloom? The largest and the longest (since 1931) long-term plankton sampling is conducted by the Continuous Plankton Recorder (CPR) survey [92], and the methods for high-throughput sequencing of formamide-preserved CPR samples are being developed [93]. Smaller scale time series plankton samples are also collected by various marine laboratories, but they are mainly used for metabarcoding to analyse species richness and its temporal variation [94,95]. Wider use of these serial samples for whole genome metagenomic sequencing would answer many questions regarding SAM evolution discussed earlier.

The fossil record that is available for some abundant marine plankton can provide a perspective that complements inferences drawn from molecular data. The calcium carbonate shells of coccolithophores and foraminifera are well preserved in the fossil record and can be used to estimate relative abundance through time. A recent study of the coccolithophore genus *Gephyrocapsa* revealed a good correspondence between genetic estimates of population size change through time and species abundance in the fossil data (Figure 2 D,F in [24]). Such integrated evolutionary genetic and palaeontological analyses provide a way to cross-validate the two independent lines of evidence, each of which has its own strengths and weaknesses.

Concluding remarks

Microevolutionary processes are the very foundation of evolutionary change, yet they remain woefully understudied in superabundant microbes (see Outstanding questions), including many species of marine plankton [7,8]. Population size is one of the most important parameters in evolutionary genetics [71,90], and precisely because of their astronomical abundance microevolution in SAMs may work in rather unusual ways [7,12,38]. We suggest that evolution in SAMs may conform to the panselectionist view that dominated in biology prior to the current era of the neutral and nearly neutral [46] theories. Testing this idea and (more generally) studying how microevolution works in SAMs will require new evolutionary genetic approaches suitable for astronomically large populations.

Outstanding questions

Does a panselectionist view of evolution fit superabundant microbes (SAMs) better than the widely accepted nearly neutral theory?

How does the process of adaptation work in SAMs?

In SAMs, do 'soft selective sweeps' occur to the exclusion of 'hard selective sweeps', as theory predicts in extremely large populations?

How big are effective population sizes in SAMs?

What limits genetic diversity in SAMs?

What evolutionary genetic approaches and tools are applicable to study evolution in astronomically large populations of SAMs?

Acknowledgments

This work was supported by Natural Environment Research Council (NERC) grant NE/V011049/1 and Biotechnology and Biological Sciences Research Council (BBSRC) grant BB/P009808/1 to D.A.F., and by a National Institutes of Health (NIH) grant GM116853 and National Science Foundation (NSF) grant DEB-1831730 to M.K.

Declaration of interests

No interests are declared.

References

1. Field, C.B. *et al.* (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281, 237–240
2. Oziel, L. *et al.* (2020) Faster Atlantic currents drive poleward expansion of temperate phytoplankton in the Arctic Ocean. *Nat. Commun.* 11, 1705
3. Walworth, N.G. *et al.* (2020) Microbial evolutionary strategies in a dynamic ocean. *Proc. Natl. Acad. Sci. U. S. A.* 117, 5943–5948
4. Jonsson, B.F. and Watson, J.R. (2016) The timescales of global surface-ocean connectivity. *Nat. Commun.* 7, 11239
5. Slater, S.M. *et al.* (2022) Global record of 'ghost' nannofossils reveals plankton resilience to high CO(2) and warming. *Science* 376, 853–856
6. Vargas, C.A. *et al.* (2017) Species-specific responses to ocean acidification should account for local adaptation and adaptive plasticity. *Nat. Ecol. Evol.* 1, 84
7. Rengefors, K. *et al.* (2017) Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies. *J. Plankton Res.* 39, 165–179
8. Sjöqvist, C. (2022) Evolution of phytoplankton as estimated from genetic diversity. *J. Mar. Sci. Eng.* 10, 456
9. Emiliani, C. (1993) Extinction and viruses. *Bio Systems* 31, 155–159
10. Chen, Z. *et al.* (2022) *Prochlorococcus* have low global mutation rate and small effective population size. *Nat. Ecol. Evol.* 6, 183–194
11. Barton, N. (2010) Understanding adaptation in large populations. *PLoS Genet.* 6, e1000987
12. Lynch, M. *et al.* (1991) Adaptive and demographic responses of plankton populations to environmental change. *Limnol. Oceanogr.* 36, 1301–1312
13. Filatov, D.A. (2023) How does speciation in marine plankton work? *Trends Microbiol.* 31, 989–991
14. Booker, T.R. *et al.* (2017) Detecting positive selection in the genome. *BMC Biol.* 15, 98
15. Excoffier, L. *et al.* (2013) Robust demographic inference from genomic and SNP data. *PLoS Genet.* 9, e1003905
16. Tsai, I.J. *et al.* (2008) Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proc. Natl. Acad. Sci. U. S. A.* 105, 4957–4962
17. von Dassow, P. *et al.* (2015) Life-cycle modification in open oceans accounts for genome variability in a cosmopolitan phytoplankton. *ISME J.* 9, 1365–1377
18. Blanc-Mathieu, R. *et al.* (2017) Population genomics of picophytoplankton unveils novel chromosome hypervariability. *Sci. Adv.* 3, e1700239
19. Bobay, L.M. and Ochman, H. (2018) Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol. Biol.* 18, 153
20. Douglas, G.M. and Shapiro, B.J. (2024) Pseudogenes act as a neutral reference for detecting selection in prokaryotic pangenomes. *Nat. Ecol. Evol.* 8, 304–314
21. Sjoqvist, C. *et al.* (2015) Local adaptation and oceanographic connectivity patterns explain genetic differentiation of a marine diatom across the North Sea-Baltic Sea salinity gradient. *Mol. Ecol.* 24, 2871–2885
22. Postel, U. *et al.* (2020) Adaptive divergence across Southern Ocean gradients in the pelagic diatom *Fragilariopsis kerguelensis*. *Mol. Ecol.* 29, 4913–4924
23. Pinseel, E. *et al.* (2023) Local adaptation of a marine diatom is governed by genome-wide changes in diverse metabolic processes. *bioRxiv*, Published online September 23, 2023. https://doi.org/10.1101/2023.09.22.559080

24. Filatov, D.A. *et al.* (2021) The mode of speciation during a recent radiation in open-ocean phytoplankton. *Curr. Biol.* 31, 5439–5449
25. Young, J.N. *et al.* (2012) Adaptive signals in algal Rubisco reveal a history of ancient atmospheric carbon dioxide. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 367, 483–492
26. Beaufort, L. *et al.* (2021) Cyclic evolution of phytoplankton forced by changes in tropical seasonality. *Nature* 601, 79–84
27. Bendif, M. *et al.* (2019) Repeated species radiations in the recent evolution of the key marine phytoplankton lineage *Gephyrocapsa*. *Nat. Commun.* 10, 4234
28. Hoogakker, B.A.A. *et al.* (2022) Planktonic foraminifera organic carbon isotopes as archives of upper ocean carbon cycling. *Nat. Commun.* 13, 4841
29. Boscolo-Galazzo, F. *et al.* (2018) Temperature dependency of metabolic rates in the upper ocean: a positive feedback to global climate change? *Glob. Planet. Chang.* 170, 201–212
30. Dedman, C.J. *et al.* (2023) The cellular response to ocean warming in *Emiliania huxleyi*. *Front. Microbiol.* 14, 1177349
31. Barton, S. and Yvon-Durocher, G. (2019) Quantifying the temperature dependence of growth rate in marine phytoplankton within and across species. *Limnol. Oceanogr.* 64, 2081–2091
32. Rickaby, R.E.M. *et al.* (2016) Environmental carbonate chemistry selects for phenotype of recently isolated strains of *Emiliania huxleyi*. *Deep-Sea Res. II* 127, 28–40
33. Sefbom, J. *et al.* (2023) Local adaptation through countergradient selection in northern populations of *Skeletonema marinoi*. *Evol. Appl.* 16, 311–320
34. Irwin, A.J. *et al.* (2015) Phytoplankton adapt to changing ocean environments. *Proc. Natl. Acad. Sci. U. S. A.* 112, 5762–5766
35. Ward, B.A. *et al.* (2019) Considering the role of adaptive evolution in models of the ocean and climate system. *J. Adv. Model Earth Syst.* 11, 3343–3361
36. Crow, J.F. (1987) Population genetics history: a personal view. *Annu. Rev. Genet.* 21, 1–22
37. Wright, S.I. and Andolfatto, P. (2008) The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Annu. Rev. Ecol. Evol. Syst.* 39, 193–213
38. Filatov, D.A. (2019) Extreme Lewontin's paradox in ubiquitous marine phytoplankton species. *Mol. Biol. Evol.* 36, 4–14
39. Krasovec, M. *et al.* (2020) Evolution of mutation rate in astronomically large phytoplankton populations. *Genome Biol. Evol.* 12, 1051–1059
40. Jerney, J. *et al.* (2022) Seasonal genotype dynamics of a marine dinoflagellate: Pelagic populations are homogeneous and as diverse as benthic seed banks. *Mol. Ecol.* 31, 512–528
41. Lewontin, R.C. (1974) *The Genetic Basis of Evolutionary Change*, Columbia biological series vol. 25. Columbia University Press
42. Charlesworth, B. and Jensen, J.D. (2022) How can we resolve Lewontin's paradox? *Genome Biol. Evol.* 14, evac096
43. Buffalo, V. (2021) Quantifying the relationship between genetic diversity and population size suggests natural selection cannot explain Lewontin's paradox. *eLife* 19, e67509
44. Ward, B.A. *et al.* (2021) Selective constraints on global plankton dispersal. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2007388118
45. Rynearson, T.A. and Virginia Armbrust, E. (2004) Genetic differentiation among populations of the planktonic marine diatom *Ditylum brightwellii* (Bacillariophyceae). *J. Phycol.* 40, 34–43
46. Ohta, T. (1992) The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* 23, 263–286
47. Karasov, T. *et al.* (2010) Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *PLoS Genet.* 6, e1000924
48. Otto, S.P. (2021) Selective interference and the evolution of sex. *J. Hered.* 112, 9–18
49. Krueger-Hadfield, S.A. *et al.* (2014) Genotyping an *Emiliania huxleyi* (Prymnesiophyceae) bloom event in the North Sea reveals evidence of asexual reproduction. *Biogeosciences* 11, 5215–5234
50. Crawford, R.M. (1995) The role of sex in the sedimentation of a marine diatom bloom. *Limnol. Oceanogr.* 40, 200–204
51. Sarno, D. *et al.* (2010) A massive and simultaneous sex event of two *Pseudo-nitzschia* species. *Deep-Sea Res. II Top. Stud. Oceanogr.* 57, 248–255
52. Barton, N.H. (1999) Clines in polygenic traits. *Genet. Res.* 74, 223–236
53. Yang, Z. (2014) *Molecular Evolution: A Statistical Approach*, Oxford University Press
54. Hudson, R.R. *et al.* (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116, 153–159
55. McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351, 652–654
56. Eyre-Walker, A. and Keightley, P.D. (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* 26, 2097–2108
57. Hudson, R.R. (1991) Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* 7, 1–44
58. Sargsyan, O. and 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
59. Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595
60. Tellier, A. and Lemaire, C. (2014) Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol. Ecol.* 23, 2637–2652
61. Menardo, F. *et al.* (2021) Multiple merger genealogies in outbreaks of *Mycobacterium tuberculosis*. *Mol. Biol. Evol.* 38, 290–306
62. Degnan, J.H. and Rosenberg, N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340
63. Mailund, T. *et al.* (2014) Lineage sorting in apes. *Annu. Rev. Genet.* 48, 519–535
64. Miyagi, M. *et al.* (2022) How many ecological niches are defined by the superabundant marine microbe Prochlorococcus? *bioRxiv*, Published online November 29, 2022. https://doi.org/10.1101/2022.11.29.517206
65. Mirarab, S. *et al.* (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30, i541–i548
66. Schluter, L. *et al.* (2016) Long-term dynamics of adaptive evolution in a globally important phytoplankton species to ocean acidification. *Sci. Adv.* 2, e1501660
67. Good, B.H. *et al.* (2017) The dynamics of molecular evolution over 60,000 generations. *Nature* 551, 45–50
68. Scheinin, M. *et al.* (2015) Experimental evolution gone wild. *J. R. Soc. Interface* 12, 20150056
69. Kelly, J.K. (1997) A test of neutrality based on interlocus associations. *Genetics* 146, 1197–1206
70. Gutenkunst, R.N. *et al.* (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5, e1000695
71. Coop, G. (2020) *Population and Quantitative Genetics*. GitHub
72. Ibarbalz, F.M. *et al.* (2019) Global trends in marine plankton diversity across kingdoms of life. *Cell* 179, 1084–1097
73. Thomas, M.K. *et al.* (2012) A global pattern of thermal adaptation in marine phytoplankton. *Science* 338, 1085–1088
74. Pinseel, E. *et al.* (2022) Strain-specific transcriptional responses overshadow salinity effects in a marine diatom sampled along the Baltic Sea salinity cline. *ISME J.* 16, 1776–1787
75. De Mita, S. *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol. Ecol.* 22, 1383–1399
76. Szymura, J.M. and Barton, N.H. (1986) Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution* 40, 1141–1159
77. Rellstab, C. *et al.* (2015) A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* 24, 4348–4370
78. Sharp, P.M. *et al.* (2010) Forces that influence the evolution of codon bias. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 365, 1203–1212
79. Krasovec, M. and Filatov, D.A. (2022) Codon usage bias in phytoplankton. *J. Mar. Sci. Eng.* 10, 168
80. Zeng, K. and Charlesworth, B. (2009) Estimating selection intensity on synonymous codon usage in a nonequilibrium population. *Genetics* 183, 651–662
81. Machado, H.E. *et al.* (2020) Pervasive strong selection at the level of codon usage bias in *Drosophila melanogaster*. *Genetics* 214, 511–528

82. Charlesworth, B. (2013) Stabilizing selection, purifying selection, and mutational bias in finite populations. *Genetics* 194, 955–971
83. Alachiotis, N. and Pavlidis, P. (2018) RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Commun. Biol.* 1, 79
84. Messer, P.W. and Petrov, D.A. (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* 28, 659–669
85. Bulankova, P. *et al.* (2021) Mitotic recombination between homologous chromosomes drives genomic diversity in diatoms. *Curr. Biol.* 31, 3221–3232.e9
86. Wang, J.M. *et al.* (2018) The genome of the human pathogen *Candida albicans* is shaped by mutation and cryptic sexual recombination. *mBio* 9, e01205-18
87. Krishnan, S. *et al.* (2023) Rhometa: population recombination rate estimation from metagenomic read datasets. *PLoS Genet.* 19, e1010683
88. Croucher, N.J. *et al.* (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* 43, e15
89. Santiago, E. *et al.* (2024) Estimation of the contemporary effective population size from SNP data while accounting for mating structure. *Mol. Ecol. Resour.* 24, e13890
90. Wang, J. *et al.* (2016) Prediction and estimation of effective population size. *Heredity (Edinb)* 117, 193–206
91. Perreau, J. *et al.* (2021) Strong within-host selection in a maternally inherited obligate symbiont: *Buchnera* and aphids. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2102467118
92. Richardson, A.J. *et al.* (2006) Using continuous plankton recorder data. *Prog. Oceanogr.* 68, 27–74
93. Vezzulli, L. *et al.* (2021) Continuous Plankton Recorder in the omics era: from marine microbiome to global ocean observations. *Curr. Opin. Biotechnol.* 73, 61–66
94. Ruggiero, M.V. *et al.* (2022) Temporal changes of genetic structure and diversity in a marine diatom genus discovered via metabarcoding. *Environ. DNA* 4, 763–775
95. Fontaine, D.N. and Rynearson, T.A. (2023) Multi-year time series reveals temporally synchronous diatom communities with annual frequency of recurrence in a temperate estuary. *Limnol. Oceanogr.* 68, 1982–1994