Bioinformatic Analyses to Detect Signature of Genetic Drift and Adaptive Evolution in Whole Genome Sequence Data.

By

Samuel Andrew Speak

100162318

Master of Science by Research (MSc(R))

University of East Anglia (UEA)

School of Environmental Science

September 2020

Abstract

The environmental and biological crisis currently threatening a vast quantity of endangered species globally. Saving species from extinction requires a step change in conservation that includes genomics and integrates this with more traditional conservation actions such as habitat restoration and species conservation. This thesis focuses on gaining experience in bioinformatic techniques through three unique and independent projects with the aim to apply these skills in a conservation genomic setting. Firstly, “Investigating the Adaptive Evolution in A Diatom’s Genome in Response To Extreme Temperature Selection”, an analysis of genetic Single Nucleotide Polymorphism (SNP) data of the polar diatom *Thalassiosira pseudonana* under differing temperature regimes and its potential adaptations to survive extreme heat stress. Secondly, “A Phylogenetic Analysis of the Heterotrophic Diatom *Nitzschia putrida”,* which investigates the process underpinning gene family expansion of transporter gene families and aims to determine whether these expansions were a pre- or post-adaptation to a heterotrophic lifestyle. Finally, “Identifying the Deleterious Mutational Load Within the Passenger Pigeon Genome”, which highlights the importance of measuring genetic load within populations and the role it may have played in the rapid extinction of one of the world’s most numerous vertebrates over a century ago. Together these studies on microorganisms to vertebrates aim to illustrate the importance of genomics and bioinformatics as tools to investigate a variety of evolutionary questions with relevance to conservation. The tools explored and developed in these studies may help conservation biologists in their studies and efforts to protect endangered species, and these approaches will become instrumental in out fight to stop the sixth mass extinction.

Contents

Figure list.

Table list.

Acknowledgements

I would like to thank my supervisors Professor Cock van Oosterhout and Professor Thomas Mock, for the all their guidance and insightful discussions throughout my project. I am grateful to Dr Andrew Toseland, Camilla Ryan, Dr Mark McMullen and Dr Ben Ward for all of their advice and assistance with the bioinformatic techniques and methods used. I would also like to thank Professor Ryoma Kamikawa for the sequence data and investigations that allowed my analysis of the non-photosynthetic *Nitzschia putrida* to occur and Sarah Nicholls, for her work and assistance throughout the investigations into the adaptive evolution in a diatom's genome in response to extreme temperature selection. Lastly, I would like to thank Esme Peers, my friends, and my family for all their support and encouragement throughout my study.

Chapter 1.

Investigating the Adaptive Evolution in a Diatom's Genome in Response to Extreme Temperature Selection.

Background

Marine phytoplankton are responsible for 50% of annual global primary production, yet despite their biological significance, there has been little research into the genetics of diatoms. The first whole diatom genome to be sequenced was the polar *Fragilariopsis cylindrus*. The genome of *F.cylindrus* showed some unique features distinct from traditional "model" species, and in particular many of its alleles are diverged from one another due to positive selection. *F.cylundrus* expresses these alternative sets of alleles depending on the environment it encounters in a phenomenon known as differential allelic expression (Mock et al. 2017). The centric diatom *Thalassiosira pseudonana* is a "model" species that is currently responsible for roughly 20% of global primary production. However, with climate change this figure is expected to decrease due to the increase in sea temperatures above the thermal tolerance (Schaum et al., 2018).

Sea temperatures have been rising at a rate of 0.18°C per decade since 1981 (NOAA, 2020), an increase that is predicted to be catastrophic to global marine organisms and especially diatoms (Lazarus et al., 2014). It is therefore critical to understand the response that *T. pseudonona*'s genome undergoes when heat stressed to ensure that genes selected for in such environments can be identified to help understand the future effects of heat stress. Experimental evolution models have shown that *T. pseudonana* can adapt to increases in temperatures due to changes in metabolic rate, with the phenotypic expression of a genetic divergence between ancestral and evolved lineages showing the rapid adaptability of *T. pseudonana* to fluctuating temperature through major genomic changes (Schaum et al., 2018). Through this it has been shown that fluctuations in sea temperature, possibly due to the temporary restoration of control conditions allowing for a rapid increase in population size and thus a fixation of beneficial mutations allowing for the rapid response. However, investigations into permanent heat stress with no relief to the control conditions will help to determine the survival of diatoms in the future and identify genes needed for key phenotypic changes for survival.

Sewall Wright developed the well-known means of determining selection F- statistics, which can describe population structure in both livestock and natural populations (Wright, 1951). These describe the levels of heterozygosity within a population and are divided into three

indices: the inbreeding coefficient or F_{IT} measuring the individual heterozygosity relative to the total population; the F_{ST} is the fixation index and thereby measured as ratio of variance in gene frequency within the subpopulation compared to the total population and F_{IS} is the inbreeding coefficient as a ratio of the variance between the subpopulation and the individual (Weir, 2012). For this investigation we will focus on the F_{ST}, which was further developed (Weir and Cockerham, 1984) (Nei,1977) (Reynolds, Weir and Cockerham, 1983) to the statistic that is used today as a means of measuring genetic polymorphism within genomic data. F_{ST} values allow the investigation of single nucleotide polymorphism (SNP) fixation within a genome and therefore determine if alleles are under selection and if levels of heterozygosity are at particular positions or genes within the genome.

This study will use genomic data collected from experimental evolution studies conducted by this laboratory (Mock et al., in prep) to assess the impact of future increases or decreases in ocean temperature due to climate change on phytoplankton species *T. pseudonana.* The study used three temperature regimes (cold (9 °C), control (22 °C) and heat stress (32 °C)) repeated across five lines per regime to determine if *T. pseudonana* can successfully adapt to a changing climate. They also aimed to derive any relatioships between mutations selected for across the temperature regimes and selection lines.

Methods

T. pseudonana was cultured into 15 experimentally evolved lines across three temperature regimes (9°C, 22°C and 32°C) and the genomes were sequenced at time intervals from between 0 and 450 generations. Previous investigations into the F_{ST} of all chromosomes (Toseland unpublished) showed that Chromosome 8 and 15 showed higher F_{ST} when cultured at heat stress (Figure 13, Appendix). Hence, all analyses in this thesis were carried out on chromosome 8 and 15 to identify SNPs that are under selection in the heat stressed or cold stressed environments, and to determine potential large changes in allele frequency.

Identification and analysis of SNPS under selection across temperature lines:

Filtered SNP calls for both chromosome 8 and chromosome 15 of the diatom *T. pseudonana* were tested for F_{ST} score using the tool VCFtools fst (Danecek et al., 2011). Note that annotated scripts used for analysis are available in the Appendix. For this analysis, the samples used were paired testing the F_{ST} between Hot and Control, Hot and Cold and Cold and Control. SNPs that were identified to have high F_{ST} scores compared to the average across the chromosome were then used to identify regions to investigate further. This analysis was repeated using a window of size 10000 and step size of 100 to indicate areas with a high

F_{ST}. The nucleotide diversity and heterozygosity (calculated as according to Equation 1) scores were calculated over chromosome 15 and 8, using VCFtools (Danecek et al., 2011) and BCFtools(Li, 2011) respectively, using a sliding window approach with a window size of 10000 and step of 1000.

Equation 1

$$\sqrt{(\frac{final\ frequency - initial\ frequency}{initial\ frequency})^2}$$

The allele frequency was calculated for SNPs that were previously identified as under selection using BCFtools query (Li, 2011) tracking the allele frequency of individual SNPs over time for each of the five lines for each temperature regime. For areas of the chromosomes with high SNP allele frequency changes, sequences were pulled out of the files using SAMtools faidx (Li et al., 2009) and Bcftools consensus (Li, 2011) and were searched using NCIBI BLAST (available at https://www.ncbi.nlm.nih.gov/) (Zheng et al., 2000), to identify if the regions were coding, non-coding and potential genes that were being differentially expressed.

Results

To determine if the polar diatom *T. pseudonana* showed successful adaptation to temperature stress through genetic adaptation and differential allelic expression, SNP call data for chromosome 15 and chromosome 8 of diatoms cultured under three different temperature regimes; extreme heat stress (32°C), control (22°C) and cold stressed (9°C), were investigated. Identifying regions with high F_{ST} values (Fig. 1), Nucleotide Diversity (Fig. 2) and Heterozygosity (Fig. 3) values that indicate selection. Further investigations were then carried out to analyse the change in allele frequency for SNPs in the regions identified (Fig. 4) and the change in allele frequency over time (Fig. 5) on an individual per SNP basis.

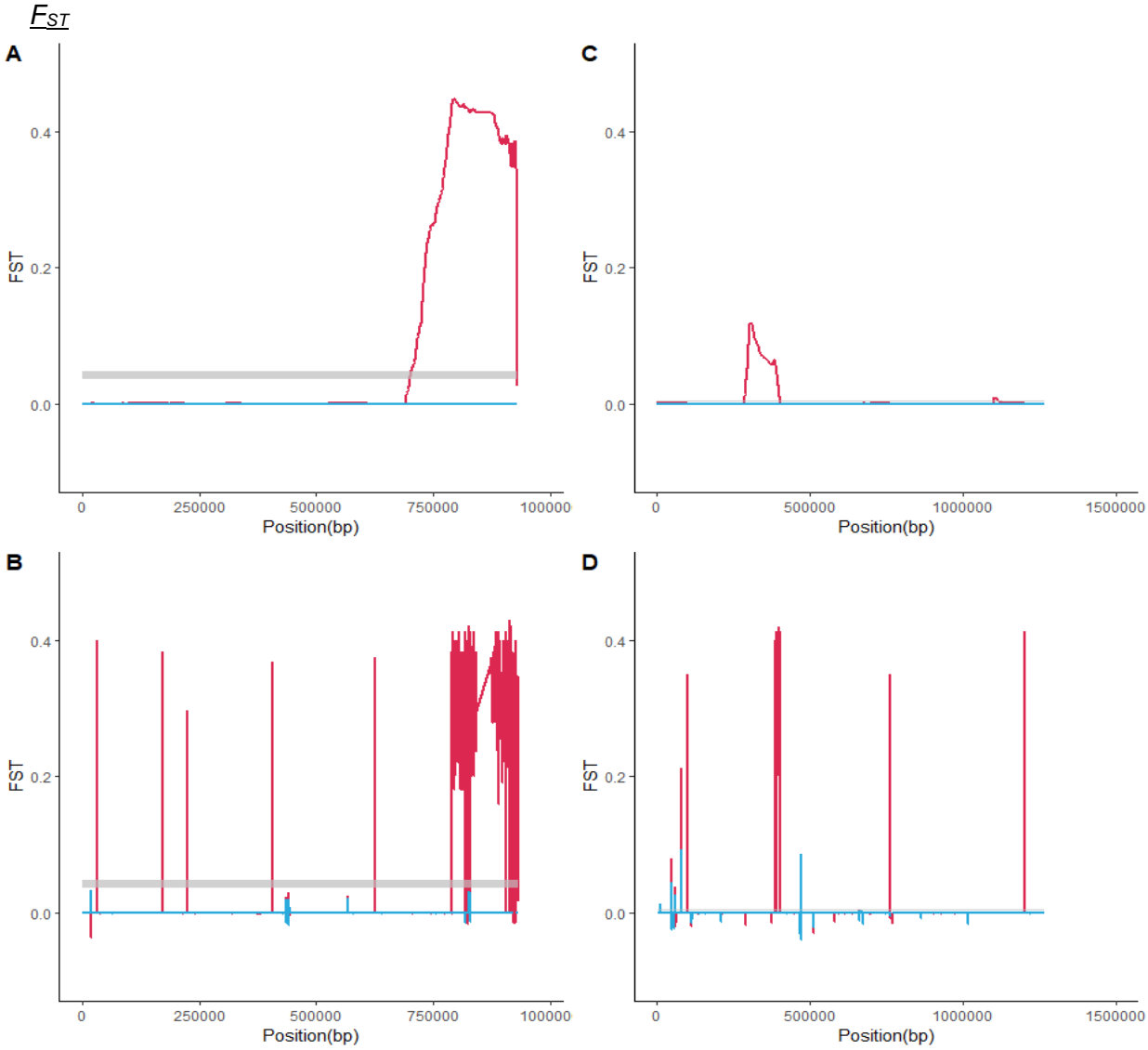

Figure 1 – F_{ST} plotted along chromosome 15 (A&B) and chromosome 8 (C&D) of *T. pseudonana* showing the mean F_{ST} (A&C) on a windowed basis with a window size of 10000 and step of 100 and individual SNP basis (B&D) showing the F_{ST} value per point. F_{ST} values calculated using VCFtools (Danecek et al., 2011) between the Hot and Control lines (Red) and Cold and Control lines (Blue), Mean F_{ST} with standard errors of 95% confidence limits shown (Grey) calculated across each chromosome, respectively.

There is a strong evidence of higher F_{ST} values in the heat stressed lines compared to the cold stressed lines (Fig. 1), across both chromosome 15 and chromosome 8. With a large mean F_{ST} calculated towards the end of chromosome 15 (Fig. 1A) in the heat stressed regime with many SNPs showing high values. With multiple areas along chromosome 15 showing high peaks around 10000 bp, 190000 bp and 400000 bp, indicating that these areas are

differentially expressed when *T. pseudonana* is heat stressed. When analysed using BLAST (Zheng et al., 2000) the region between 430000 and 450000 bp on chromosome 15 in the heat stressed lines coded for retrotransposon CoDi5.5 (Maumus et al., 2009). There were multiple SNPs in the region 800000 – 900000 bp on chromosome 15 that showed high F_{ST} values.

Nucleotide Diversity

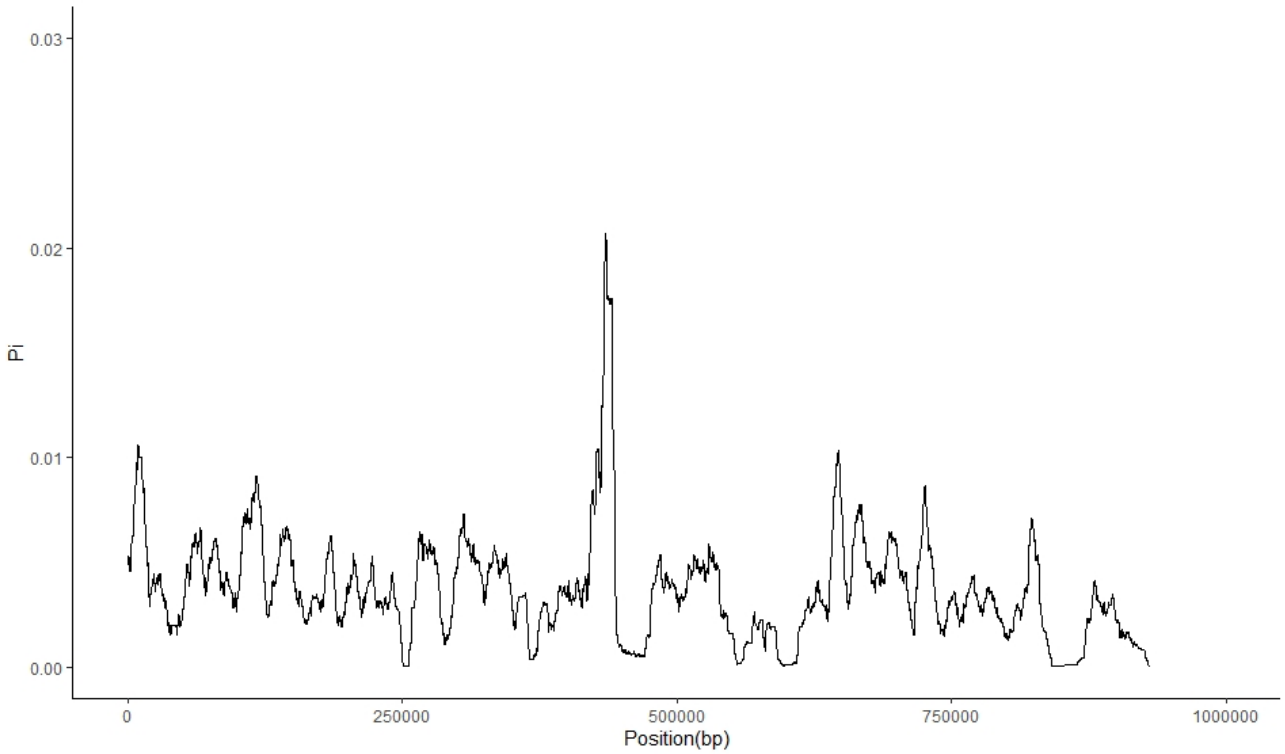

Figure 2 – Nucleotide diversity (pi) calculated along chromosome 15 of the diatom *T. pseudonana*. Calculated using VCFtools (Danecek et al., 2011) with a window size of 1000 and step of 100.

Although the F_{ST} was highest on average towards the end of chromosome 15 (Fig. 1) there were multiple SNPs shown to have high F_{ST} scores at a series of points. After investigating this further, it was identified that there was a high nucleotide diversity around 400000-450000 bp along chromosome 15 (Fig. 2). This suggests that there are a diverse number of alleles present in the *T. pseudonana* genome around this area therefore these SNPs may be under selection between the different temperature regimes.

Heterozygosity graphs

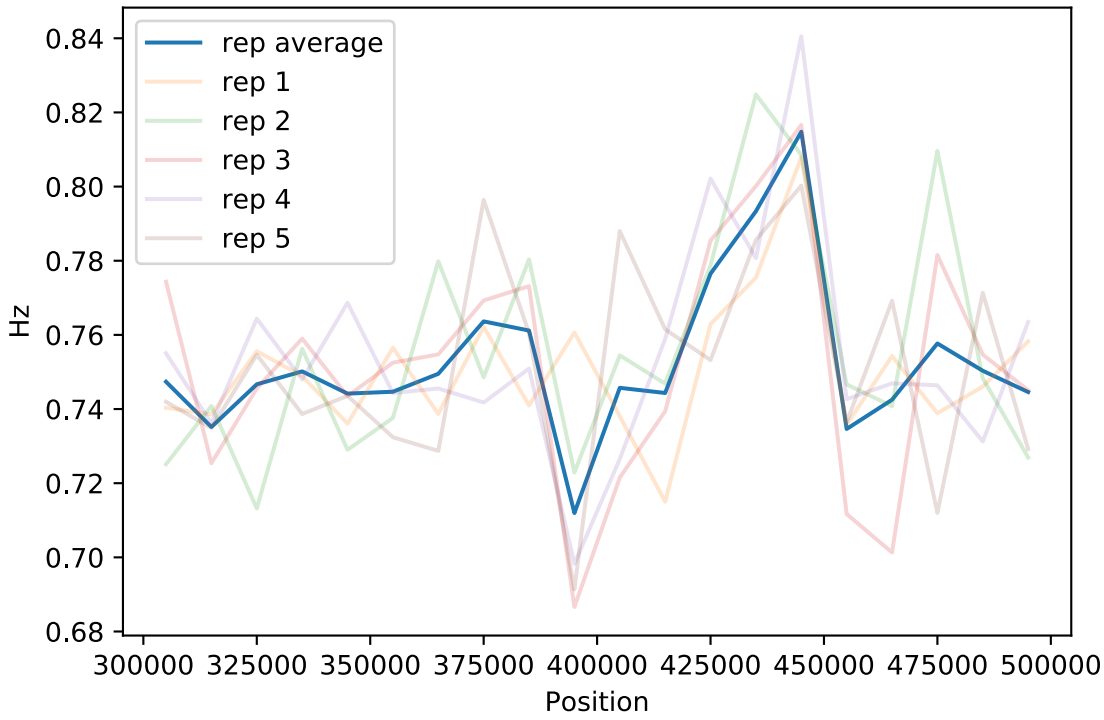

Figure 3- The heterozygosity (calculated as H = 1- p^2 (p = allele frequency) along chromosome 15 of the heat stressed replicates of the diatom *T. pseudonana* (all replicates at 32°C, 210 generations). The average is shown in blue. Window size: 10000, (graph produced by Sarah Nicholls).

After investigating the area on Chromosome 15 between 300000- 500000 bp in more detail it was apparent that there was a clear spike in the heterozygosity of SNPs across all 5 of the heat stressed temperature lines (Fig. 3) suggesting that SNPs in this area were under selection.

Allele frequency change graphs

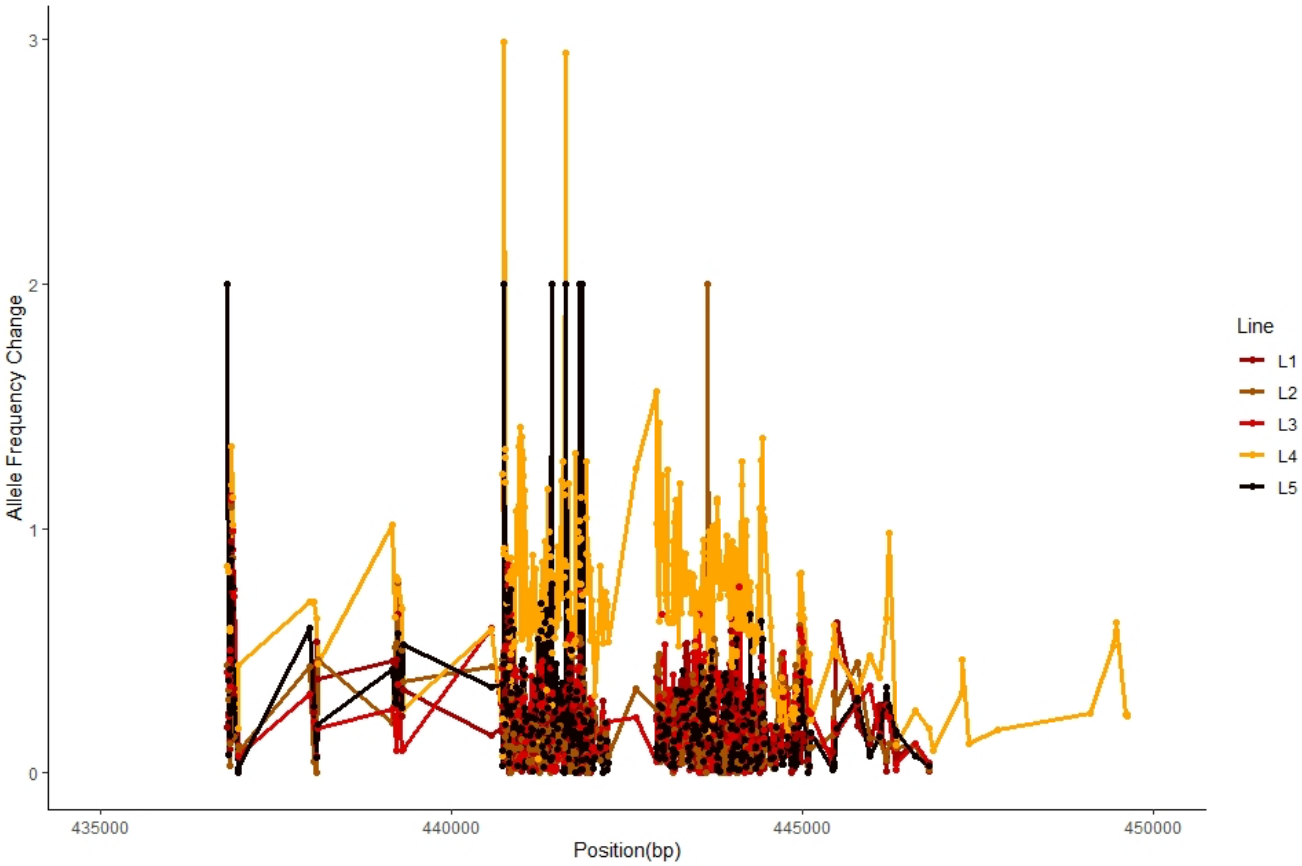

Figure 4 – The allele frequency change in the heat stressed lines for SNPs identified in the region 435000 to 450000 bp along chromosome 15 of the diatom *T. pseudonana.* The change in allele frequency was calculated as the normalized value for the difference in allele frequency for the alternative alleles between the Control line at zero generations and each of the heat stressed lines at 210 generations. Multiple lines show SNPs with high frequency change around the SNP at 440744, 441817, 441858, 443642. (The peaks at 441817 bp, 441858 bp were found at the same difference in both lines 3 and 5).

There is evidence (Fig. 4) that there are multiple SNPs in the region 435000 to 450000 bp along chromosome 15 of the diatom *T. pseudonana* that show high changes in allele frequency switching from the reference allele to an alternative allele. This suggests that this area undergoes high differential allelic frequency when heat stressed compared to the control environment.

Allele frequency difference over time

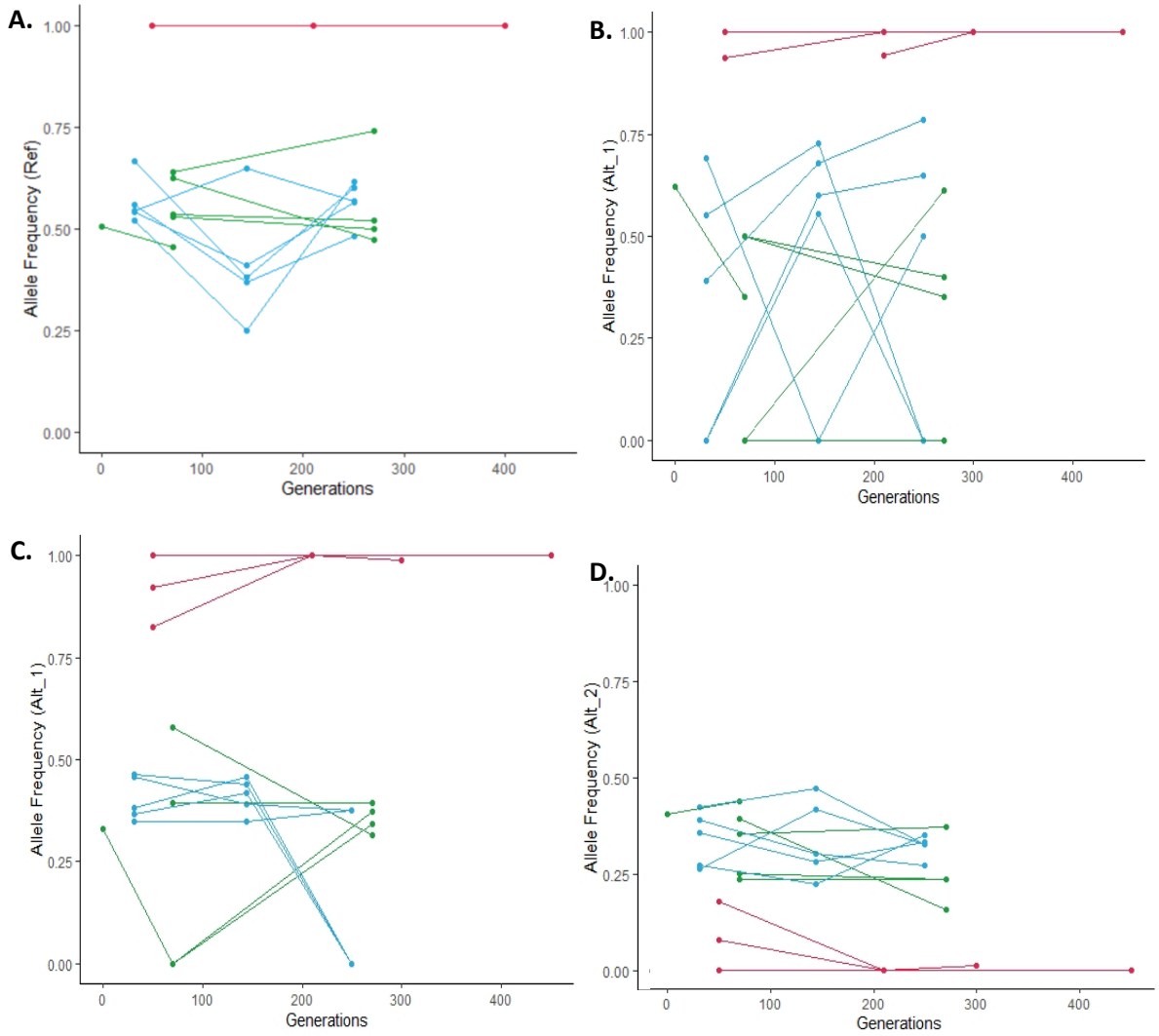

Figure 5 - The allele frequency of four SNP at positions: 406651 (A), 819182 (B), 829037 alternate allele 1 (C) and 829037 alternate allele 2 (D) on chromosome 15. Samples cultured in heat stressed conditions (32ºC) in red, control (22ºC) in green and cold stressed conditions (9ºC) in blue. Showing an increase in Allele frequency in the heat stressed lines of *Thalassiosira pseudonana* over the 400 generations.

Analysis identified a region between 800000 - 900000bp on chromosome 15 that showed high F_{ST} scores (Fig. 1). Further investigations into the allele frequency values for individual SNPs in this region on chromosome 15 of *T. pseudonana*, suggests that SNPs within the heat (32ºC) stressed lines with high F_{ST} values were reaching fixation rapidly (Fig. 5). They also showed that although there was an increase in allele frequency in the heat stressed lines the cold stressed lines were not reacting showing little change when compared to the allele frequency of the control lines.

Discussion

Multiple regions along chromosome 15 of the *T. pseudonana* genome showed high F_{ST} values compared to the rest of the chromosome (Fig. 1). Which suggests that these areas have been

subject to selection when heat stressed. The same areas were observed to respond across the heat stressed (32^{o}C) lines, whilst the cold stressed lines did not show this response. Altogether, this suggests that selection may have acted on standing genetic variation that was already present in the founder population. Upon further investigation of these sites there was evidence of high nucleotide diversity at between 435000 and 450000 bp along chromosome 15 (Fig. 2), this increase in polymorphism suggests an increase in genetic diversity around this region when lineages were cultured under heat stress and identified multiple SNPs with high allele frequency change values (Fig. 4). This region was then identified to code for the retrotransposon CoDi5.5 (Maumus et al., 2009), part of the *CoDi-II* clade of the *Ty1/Copia* family (Llorens et al., 2011). Transposable elements are mobile DNA sequences thought to be important contributors in genome evolution due to their ability to insert themselves into genetic elements of their host genomes. In particular retrotransposons transpose through reverse-transcribed RNA intermediates in a “Copy-Paste” mechanism (Lescot et al., 2016), resulting in addition or reduction in the hosts physical genome size.

Retrotransposons are found in the genomes of all eukaryotes, however their frequency in *T. pseudonana* only accounts for a relatively low 1.1% of the genome, in comparison to 5.8% of the pennate diatom *Phaeodactylum tricornutum* (Bowler et al., 2008) or the 76% in maize (Schnable et al., 2009). The identification of this 5758bp transposon in the area of high genetic diversity (Fig. 2) and heterozygosity (Fig. 3) along with the multiple SNPs with high allele frequency change (Fig. 4) supports the “epi-tranposon” hypothesis (Zeh, Zeh and Ishida, 2009). This proposes that retrotransposons, which are normally suppressed due to their deleterious properties (Slotkin and Martienssen, 2007), are instead reactivated when diatoms are under environmental stress, altering gene expression, structure, gene duplication and novel forms and expression (Pargana et al., 2020). In this way *T. pseudonana* uses transposable elements to stimulate mutations that increase genetic and phenotypic variation allowing for a rapid survival response to heat stress.

There is a large increase in the F_{ST} values towards the end of chromosome 15 (Fig. 1). This was split into two areas with high F_{ST} values; the latter of these is potentially an artefact. Commonly the ends of chromosomes have high F_{ST} due to the telomeres that are located in these regions. When the start of the region around 800000 bp was analysed using BLAST (Zheng et al., 2000) there was a hit for a currently unnamed potential protein for the diatom *T. pseudonana.* When tracked over the 400 generations of the experiment alleles that were identified show a rapid response to heat stress in SNPs with some alleles reaching fixation before the first sampling point at 32 generations (Fig. 5 A&B). In many of the SNPs that showed a large increase in allele frequency with the heat stressed lines reaching fixation, there was limited change between cold stressed lines showing similar responses to that of the

control lines. Many of the alleles identified to have a high allele frequency change (Fig. 4) did not when plotted over time show fixation in the hot lines for the first alternate allele. However, these SNPs appear to have multiple alternate alleles present at each site with some SNPs having up to three alternate alleles present. Some SNPs identified to have multiple alternate alleles on chromosome 15 did show a difference in frequency across the three temperature regimes (Fig. 5 C&D).

In conclusion there were multiple regions along chromosome 15 and chromosome 8 that showed high F_{ST} values in the heat stressed (32^{o}C) lines of the polar centric diatom *T. pseudanona*. When investigated for genetic diversity along chromosome 15 (Fig. 2), there was a large spike in genetic diversity and heterozygosity (Fig. 3) around 430,000-450,000bp and further investigation showed that this region codes for the retrotransposon CoDi5.5 (Maumus et al., 2009), suggesting that under heat stress *T. pseudonana* undergoes epistatic transposon response. This led to an increase in genetic and phenotypic variation which, when in a natural environment the temperature returns to baseline, would cause a population size increase providing opportunity for a rapid response to environmental stress. Future investigations into the presence of retrotransposons along the *T. pseudonana* genome and the genetic variation that they create will provide insight into how the globe's largest primary producers can cope with the pressures of rising sea temperatures caused by climate change.

Chapter 2.

A Phylogenetic Analysis of the Heterotrophic Diatom *Nitzschia putrida*

Background

Diatoms are one of the most abundant primary producers on the planet and therefore understanding their evolution and genome is key to their use to help reduce the effects of climate change as well as their extensive biotechnological uses (Daboussi et al., 2014). The capability to photosynthesise is estimated to have evolved in eukaryotes at around 12 billion years ago (Benoiston et al., 2012). There have been multiple independent losses of photosynthetic capabilities within the plastids of diatoms (Kamikawa et al., 2015) (Roger, Muñoz-Gómez and Kamikawa, 2017) the diatom *Nitzschia putrida* NIES-4239 circa 6.67 million years ago (MYA). Interestingly, although *N. putrida* has lost its photosynthetic capabilities, the genome of this diatom appeared to be of comparable size to similar photosynthetic diatom species. Compared to *F. cylindrus, Pseudo-nitzschia multistria, Phaeodactylum tricornutum* and *T. pseudonana*, *N. putrida* had retained genes relating to cell cycle proteins but lost genes coding for photoreception (Kamikawa et al., 2017). This means that *N. putrida* is no longer regulated by the cycles of light and dark unlike most photosynthetic species of diatoms (Depauw et al., 2012). Interestingly, *N. putrida* had also shown a significant expansion in the relative number of gene families coding for transporters, mainly through gene duplication.

One of the key gene families that has undergone an expansion in the genome of *N. putrida* is the Silicon Transporter gene family (SIT). Genes in this family facilitate the creation of silica cell walls (frustule) that is distinct to diatoms, allowing them to produce the wide range of frustule morphologies present within the taxa (Shrestha and Hildebrand, 2014). To produce this protective and low energy costing cell wall (Brembu et al., 2017), diatoms transport silicic acid from sea water into their cells using SIT proteins. It has been hypothesised that the cellular location of the SIT proteins, Si binding affinities and the transport rates are all influenced by this gene family and are specific to each species. This is supported by the observation that the phylogeny of SIT gene families and the phylogeny of diatoms can be mapped closely.

The SIT gene family is comprised of four clades across diatoms and bacteria (Durkin et al., 2016) and are thought to have evolved in response to the currently low levels of silicic acid in oceans compared to 140 million years ago when the frustules of diatoms likely evolved (Gersonde and Harwood, 1990) due to their presence in the fossil records around this time. The presence of SIT genes amongst all major diatom lineages suggests that diatoms, which

could at first transport the silicic acid across cell membranes through diffusion due to its high concentrations (Siever, 1991) (Tréguer et al. 1995) (Maldonado et al. 1999), evolved the ability to actively transport silicic acid to maintain frustule production at the lower concentrations present in the modern ocean (Tréguer et al. 1995). Evidence suggests that low levels of silicic acid are a main driver in the termination of diatom blooms (Krause et al., 2019). Therefore, SIT genes and the frustule that they enable are both crucial for the survival, growth, and reproduction of diatoms. Therefore, the expansion of the SIT gene family within the *N. putrida* suggests that it provides an advantage to survival in a new heterotrophic niche.

Phylogenetics is the evolutionary study of the relationships between species, individuals or genes, becoming an indispensable tool for the comparison of genomes (Yang et al., 2012) through the recreation of the phylogenetic tree for the taxa under investigation. The principles of phylogeny, a term invented by Ernst Haeckel in 1866 (Pace, Sapp and Goldenfeld, 2012), derives from the theory of cladistics the study of evolutionary trees and the arrangement of tax into clades based on their relatedness under coalescent theory (Kingman, 1982), whereby all taxa in a tree coalesce to a single common ancestor. This idea has been developed from its inception by Darwin's theory of descent with modification (Darwin, 1859), through advances in our understating of genetics and technological innovations to be able to accurately categorise organisms based on their genomes rather than purely phenotypic observations. Phylogenetic trees of organisms based upon their genetic relatedness are produced by comparing the aligned gene sequences under investigation to each other and score them based on insertions, deletions and mutations. In this way all potential clades formed can be ranked based on the likelihood of their occurrence and therefore produce a phylogenetic tree of the sequences or organisms. Bayesian Evolutionary Analysis by Sampling Trees (BEAST) (Drummond and Rambaut, 2007) and BEASTv2 (Bouckaert et al., 2019) were designed to reconstruct a phylogeny. The algorithms analyse aligned nucleotide sequences and produce likelihood trees based on the known sequences dates and genetic information (mutation rates and molecular clock) to reconstruct a Bayesian Markov Chain Monte Carlo (MCMC) tree is the most likely estimation of the phylogeny from the sequences input, providing branch lengths for estimation of divergence time.

This investigation aims to determine if the gene duplication and expansion of gene families in the non-photosynthetic diatom *N. putrida* occurred pre or post adaptation to its heterotrophic lifestyle. The secondary aim is to analyse the rate of the gene family expansions to determine if their rate of expansions accelerated over time. I hypothesise that the loss of photosynthetic capabilities has resulted in the adaptive expansion and neofunctionalization of certain multigene families. Alternatively, certain multigene families may have been expanded before the loss of photosynthesis, enabling the adaptation to the new lifestyle. This study will test

between both hypotheses, which may provide an insight into the analysis of *N.putrida*'s loss of photosynthetic capabilities and adaptation to a heterotrophic lifestyle.

Methods

The sequences for gene families of the corresponding 7 expanded gene families and 2 control gene families (Table 1.) from the non-photosynthetic diatom *Nitzschia putrida* were combined with the respective gene sequences from related photosynthetic diatom species *Fragilariopsis cylindrus*, *Pseudo-nitzschia multiseries* and the non-photosynthetic *Nitzschia alba* using SAMtools (Li et al., 2009). Where possible the maximum number of species were used depending upon the availability of annotated sequences.

Table 1 – List of the 9 expanded gene families and 2 non-expanded (control)(*) gene families with the number of sequences found in *Nitszchia putrida* and availability of sequences in annotated genomes from closely related diatom species.

			Species annotated		
Gene Family	Acronym	Number of sequences	Nitzschia alba	*Fragilariopsis cylindrus*	*Pseudo-nitzschia multiseries*
ABC transporter	ABC	90	N	Y	Y
Drag/Metabolite transporters	DMT	30	N	N	N
Major Facilitator Superfamily	MFS	66	N	Y	Y
Resistance-Nodulation-Cell division superfamily	RND	19	N	N	N
Silicon ion transporters	SIT	22	Y	Y	Y
Solute:Sodium symporters	SST	10	N	Y	Y
Myosin *	-	11	N	Y	Y
Ammonium Transport *	-	8	N	Y	Y
Glycosyl Transferase	GT2	17	N	N	N
	GT31	25	N	N	N
	GT32	19	N	N	N
	GT49	13	N	N	N
	GT8	11	N	N	N
Glycosyl Hydrolase	GH114	8	N	N	N
	GH16	15	N	N	N
	GH28	7	N	N	N
	GH72	5	N	N	N
	GH99	14	N	N	N

Alignments of the gene families.

All sequences from their corresponding gene families were first clustered using cd-hit/4.6.8 (Li, Jaroszewski and Godzik, 2001), removing sequences of 100% alignment and clustering at low identity. The largest clusters for each gene family were then aligned using Prank/170427 (Löytynoja, 2014). Poorly aligned sequences were removed with Trimal 1.2 (Capella-Gutiérrez, Silla-Martínez and Gabaldón, 2009) and put into gene blocks to remove gapped columns with Gblocks v0.91b (Castresana, 2000). Aligned sequences containing the maximum number of species were then analysed using phylogenetic analyses.

Phylogenetic analysis.

Divergence estimates were obtained using Bayesian Markov Chain Monte Carlo (MCMC) analysis is implemented in BEAST 2.6.1 (Bouckaert et al., 2019). The analysis was carried out using a relaxed molecular clock approach with an uncorrected log-normal distribution model of rate of variation, the HKY substitution model, four gamma categories and a Yule model of speciation. For each gene family five runs were carried out with 20 million MCMC generations sampled every 1000th generation. The results from all runs were combined and summarised using LOGcombiner v1.10.4, which were checked for convergence using Tracer v1.7.1 (Rambaut, 2009). A maximum clade credibility tree was generated using Tree Annotator v1.10.4 and graphically visualised using FigTreev1.4.3 software (Rambaut, 2012).

Expansion rate analysis across gene families.

The Phylogenetic Maximum Clade Credibility (MCC) trees were used to determine the rate of "speciation" (or more accurately, the expansion rate) within a given gene family. The R package Tree Evolution Simulation Software (TESS) (Höhna, May and Moore, 2016) was used for this analysis with the mass extinction turned off, measuring only for the speciation rate (expansion rate), using the phylogenetic trees produced in the divergence time estimations.

Results

Of the eight expanded transporter gene families, only the SIT gene family (Fig. 6) had annotated genomes from all four species, including both non-photosynthetic species *N. putrida* and *N. alba,* as well as the photosynthetic model species *P. multiseries* and *F. cylindrus.* In this way all gene family's phylogenetic trees were dated using the mean speciation date of 10 MYA (Nakov, Beaulieu and Alverson, 2018) between *F. cylindrus* and *P. multiseries* where annotations were present for the gene family under investigation within the genomes of both species.

Phylogenetic analysis

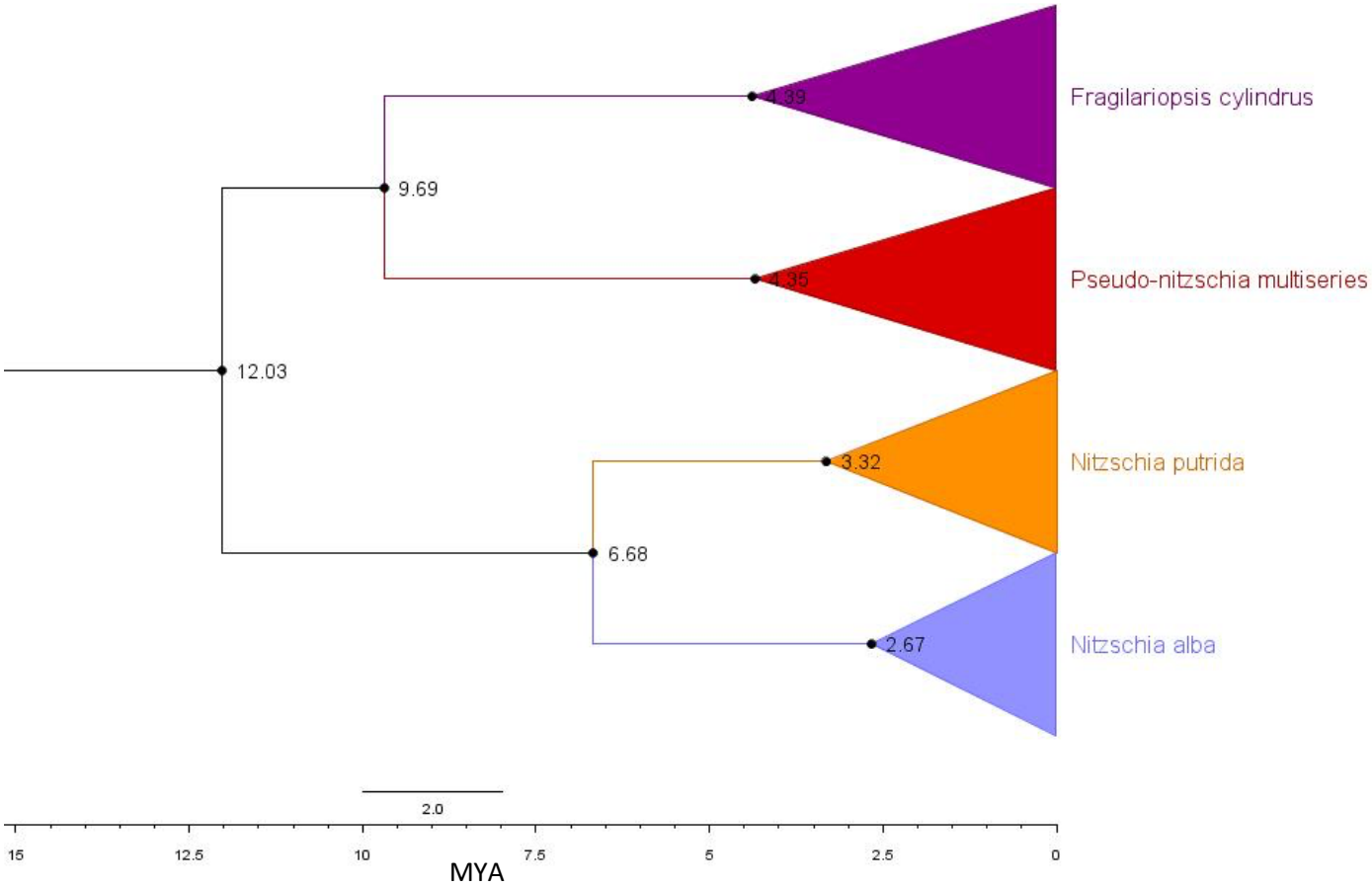

Figure 6 - Phylogenetic Maximum Clade Credibility (MCC) tree summarised by TreeAnnotator v2.6.1 for the Clade A Silicon Transporter (SIT) genes for Fragilariopsis cylindrus (n=3), Pseudo-nitzschia multiseries (n=2), Nitzschia alba (n=4) and the expanded SIT gene family of *Nitzschia putrida* NIES-4235 (n=20). Divergence estimates were obtained using Bayesian Markov Chain Monte Carlo (MCMC) analysis is implemented in BEAST v2.6.1. (Bouckaert et al., 2019), phylogenetic trees created using FigTree v1.4.3. Divergence estimates for all nodes given in millions of years ago (MYA). Fragilariopsis cylindrus, Pseudo-nitzschia multiseries and Nitzschia alba SIT gene sequences taken from published data (Durkin et al., 2016).

Divergence estimates show that *Nitzschia putrida* SIT gene family expansions occurred at around 3.3 MYA(Fig. 6), this date is later than speciation estimate between *Nitzschia putrida* and *Nitzschia alba* (6.67 MYA), suggesting SIT gene family expansions were adaptions to a heterotrophic lifestyle rather than pre-adaptation.

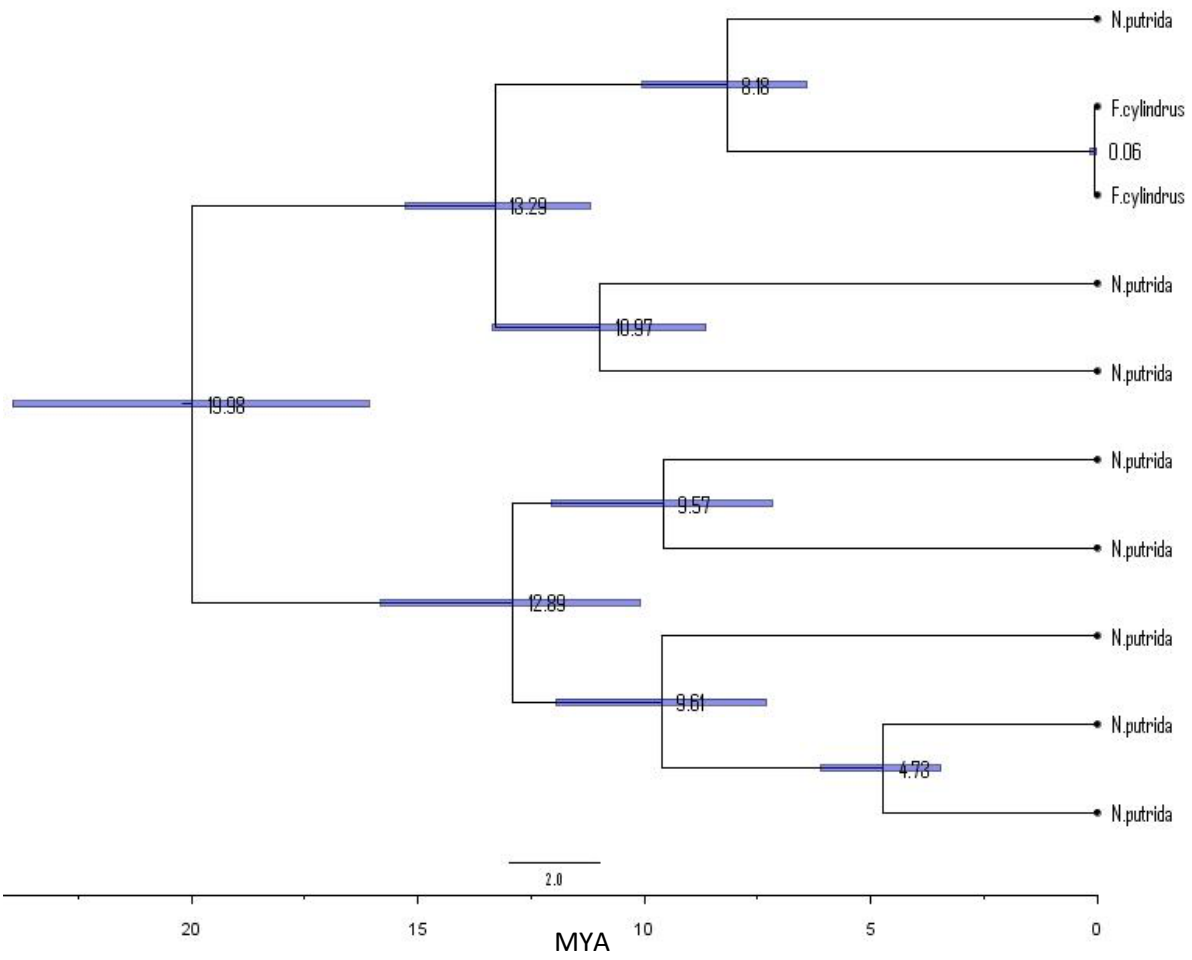

Figure 7 - Phylogenetic Maximum Clade Credibility (MCC) tree summarised by TreeAnnotator v2.6.1 for the Myosin genes for *Fragilariopsis cylindrus* (n=2) and the largest cluster of expanded Myosin gene family of Nitzschia putrida NIES-4235 (n=8). Divergence estimates were obtained using Bayesian Markov Chain Monte Carlo (MCMC) analysis using corresponding sequences from related species implemented in Beast v2.6.1. Bouckaert et al., 2019), phylogenetic trees created using FigTree v1.4.3. Divergence estimates for all nodes given in millions of years ago (MYA) with scale bar shown.

Clade A Silicon Ion Transporter (SIT) gene family of *N. putrida* SIT expanded around 3.3 MYA, significantly after the speciation event between *N. putrida* and *Nitzschia alba* (6.67 MYA) (Fig. 6). This suggests that the SIT gene family expansions were adaptions in response to the changing heterotrophic lifestyle, rather than a pre-adaptation that enabled the change. In contrast to the SIT gene family, the divergence of the Myosin gene family of *N. putrida* appears to have taken place before the speciation event (Fig. 7).

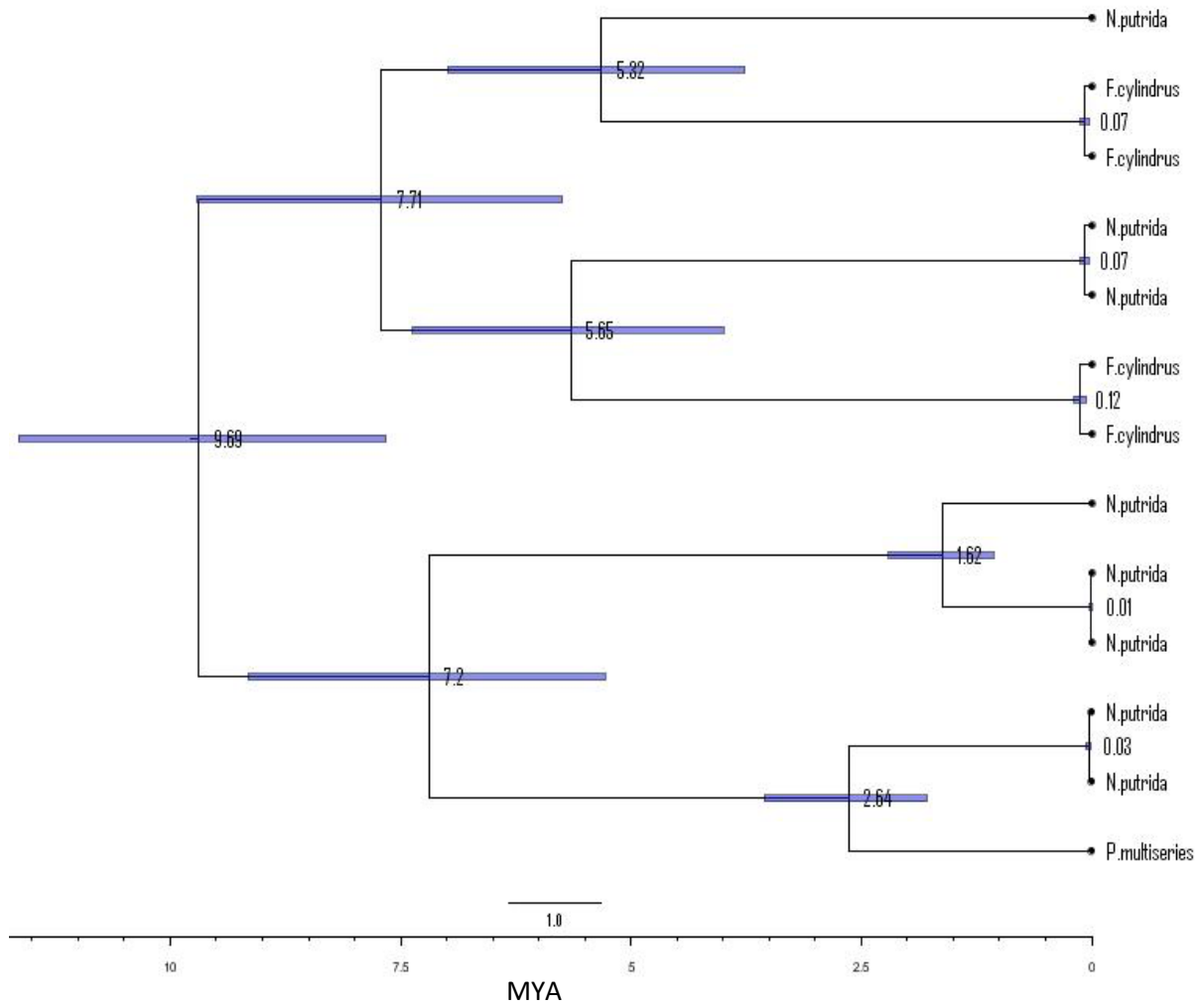

Figure 8 Phylogenetic Maximum Clade Credibility (MCC) tree summarised by TreeAnnotator v2.6.1 for the Ammonium Transporter (NH^+) genes for Pseudo-nitzschia multiseries (n=1), Fragilariopsis cylindrus (n=4) and the largest cluster of expanded NH^+ gene family of Nitzschia putrida NIES-4235 (n=8). Divergence estimates were obtained using Bayesian Markov Chain Monte Carlo (MCMC) analysis using corresponding sequences from related species implemented in BEAST v2.6.1. (Bouckaert et al., 2019), phylogenetic trees created using FigTree v1.4.3 Divergence estimates for all nodes given in millions of years ago (MYA) with scale bar shown.

The phylogenetic tree of the Ammonium Transporter (NH+) genes shows that some gene family members of *N. putrida* appear to be only recently diverged from genes of *P. multiseries* (Fig. 8), even more recent than the estimated speciation date (Fig. 6). This is likely due to high conservation of these genes by purifying selection which can cause the divergence time to be underestimated.

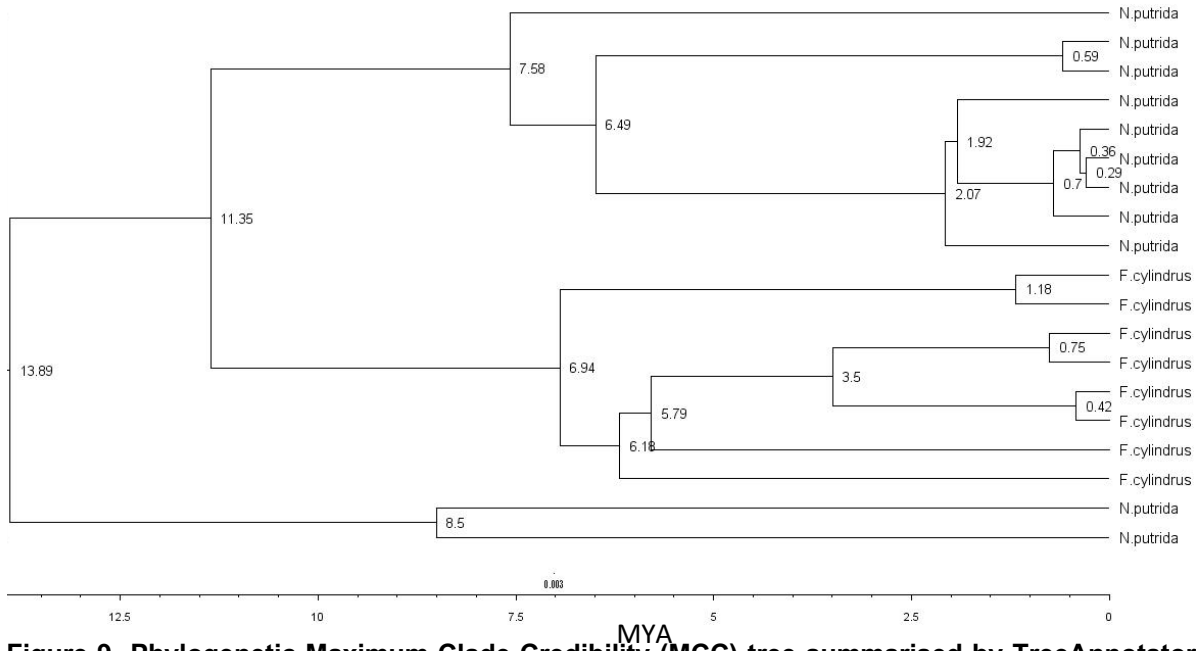

Figure 9 -Phylogenetic Maximum Clade Credibility (MCC) tree summarised by TreeAnnotator v2.6.1 for the Major Facilitator Superfamily (MFS) genes for Fragilariopsis cylindrus (n=8) and the largest cluster of expanded MFS gene family of Nitzschia putrida NIES-4235 (n=11). Divergence estimates were obtained using Bayesian Markov Chain Monte Carlo (MCMC) analysis using corresponding sequences from related species implemented in BEAST v2.6.1. (Bouckaert et al., 2019), phylogenetic trees created using FigTree v1.4.3. Divergence estimates for all nodes given in millions of years ago (MYA) with scale bar shown.

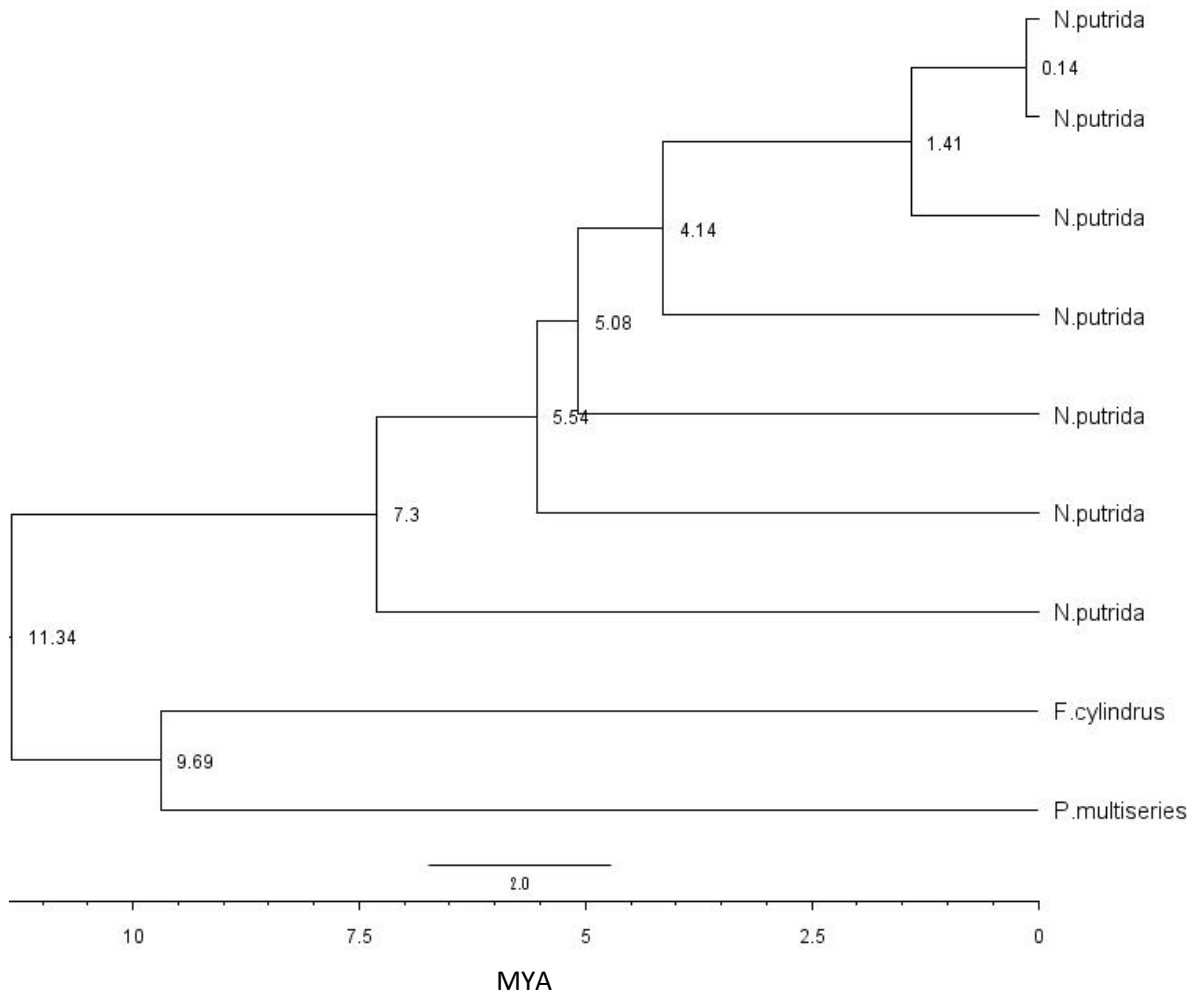

Figure 10 - Phylogenetic Maximum Clade Credibility (MCC) tree summarised by TreeAnnotator v2.6.1 for the Solute:Sodium symporters (SST) genes for *Pseudo-nitzschia multiseries* (n=1), Fragilariopsis cylindrus (n=1) and the largest cluster of expanded SST gene family of *Nitzschia putrida* NIES-4235 (n=7). Divergence estimates were obtained using Bayesian Markov Chain Monte Carlo (MCMC) analysis using corresponding sequences from related species implemented in BEAST v2.6.1. (Bouckaert et al., 2019), phylogenetic trees created using FigTree v1.4.3. Divergence estimates for all nodes given in Millions of years (MYA) before present.

Expansion rate analysis

When analysing the phylogeny of the expanded gene families of *N. putrida* it appeared that in some of the gene families there was a perceived increase in the rate of branching. Analysis was then carried out to determine the rate of gene family expansion, to determine if the rate of expansion was consistent over time and to investigate if there was any difference in the rate of expansions between the expanded and non-expanded gene families of *N. putrida.*

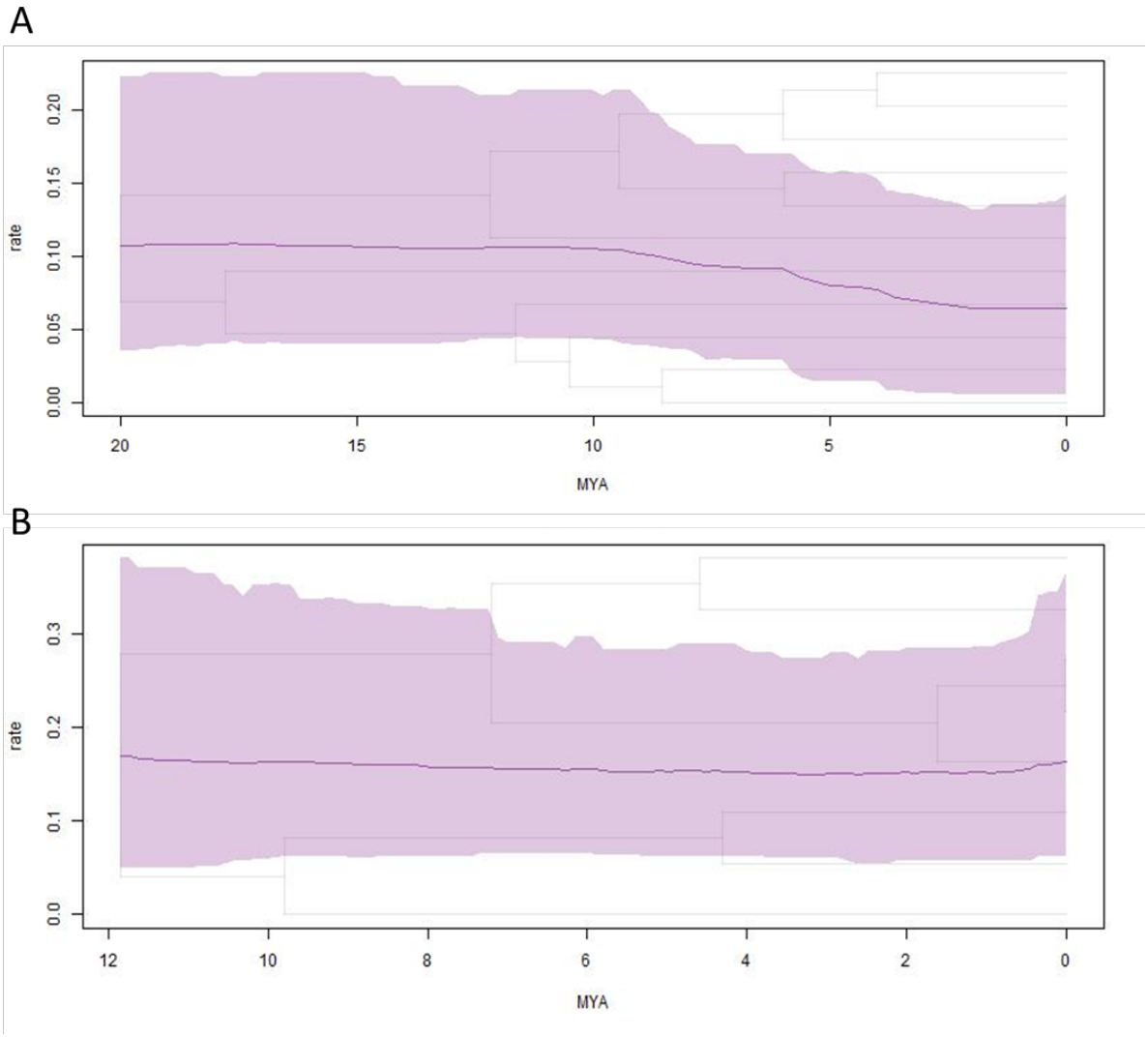

Figure 11 - Expansion rate of gene sequences within the control non-expanded Myosin (A) and the Ammonium transporter (B) gene families of the non-photosynthesising diatom *Nitzschia putrida* NIES-4235. Expansion rate was calculated using the Bayesian inference lineage diversification rate analysis tool TESS for R (Höhna, May and Moore, 2016) using the MCC phylogenetic tree produced by BEAST v2.6.1 (Bouckaert et al., 2019). With phylogenetic trees shown for reference.

There is no evidence for a significant change in the expansion rate in the past 20 million years for the Non-expanded myosin gene family or over the last 12 million years for the Ammonium transporter gene family (Fig. 11).

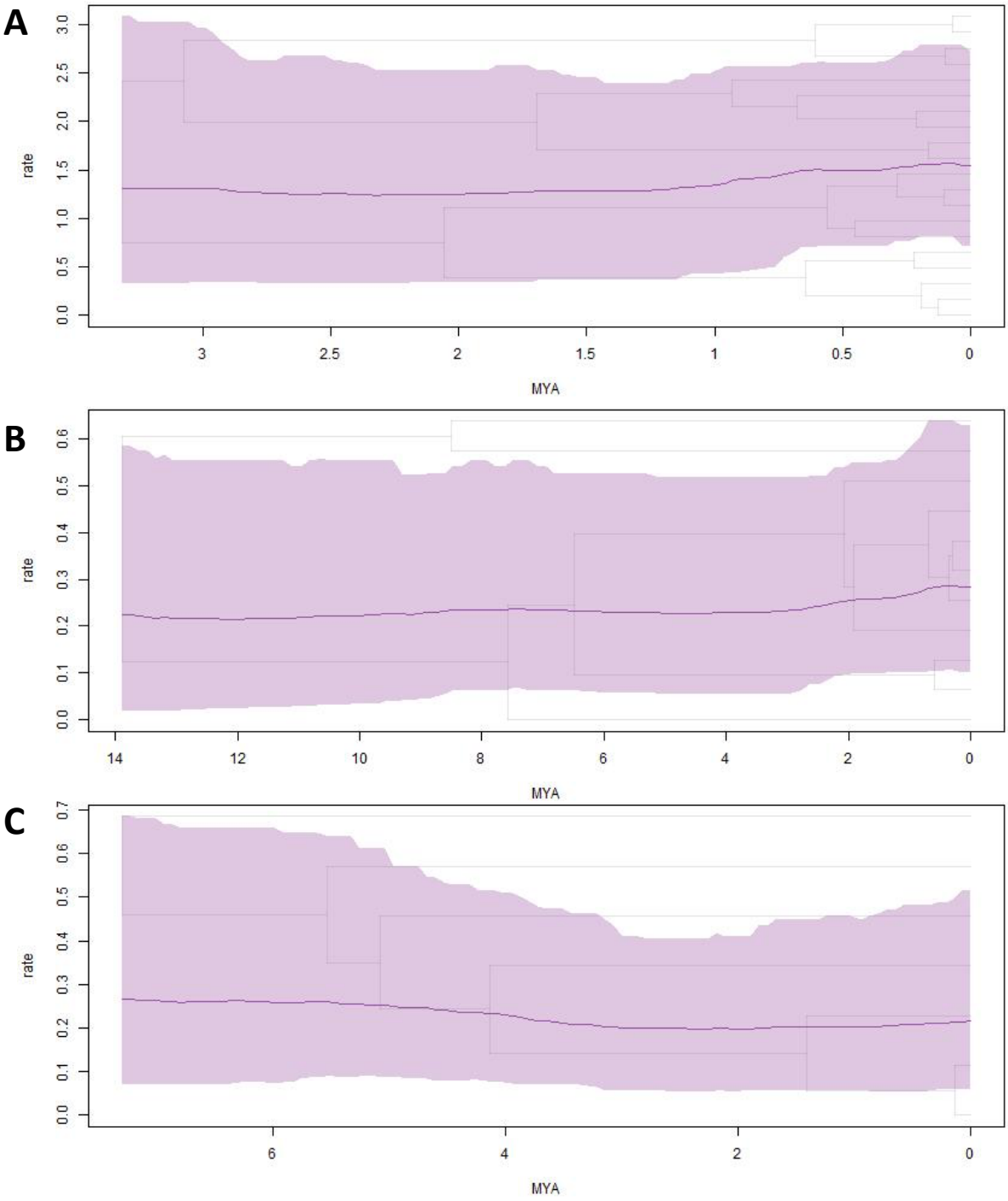

Figure 12 – Expansion rate of gene sequences within the expanded Silicon ion transporters gene family (A), Major Facilitator gene superfamily (B) and Solute:Sodium symporters gene family (C) the of the non-photosynthesising diatom *Nitzschia putrida* NIES-4235. Expansion rate was calculated using the Bayesian inference lineage diversification rate analysis tool TESS for R (Höhna, May and Moore, 2016) using the MCC phylogenetic tree produced by BEAST v2.6.1(Bouckaert et al., 2019). Phylogenetic tree shown for reference. There is no evidence for a significant change in the expansion rate in the past 3 million years, 14 million years and 7.3 million years respectively.

There is an increased rate of expansion for the SIT gene family compared to both the control gene families (Fig. 11) and the other expanded gene families (Fig. 12). However, there is no evidence for an acceleration in the gene family expansions (or gene speciation rate) in the past 20 million years for this gene family, or for any of the other gene families analysed in this study. There is also no evidence of significant acceleration of gene family expansion before, during or after speciation of *N.putrida*. For the SIT gene family, the expansions may have been adaptions in response to the changing heterotrophic lifestyle, whilst for non-expanded (Myosin gene family), the gene families diverged before the speciation event ~6.67MYA.

Discussion

The non-expanded Ammonium (Fig. 7) and Myosin (Fig. 8) gene families that were analysed showed that the gene-families were maintained for >19 MYA and >9 MYA respectively. As expected, this showed that the control genes were conserved. In comparison, there is evidence from the analysis performed (Fig. 6), that there was a speciation event between *N. putrida* and *N. alba* around 6.7 MYA. From this it can be inferred that as there have been multiple gene families within the *N. putrida* genome that have undergone significant expansion, and that these expansions were subsequent to 6.7 MYA. It is therefore likely that these gene families had expanded post adaptation to a heterotrophic lifestyle and that the recent gene expansion is due to neo-functionalisation rather than potential pre-adaptation to the switch to heterotrophy. The loss of photosynthesis within *N. putrida,* but retention of genome size suggests that the expanded gene families are a product of adaptation to the new niche that *N. putrida* is surviving in.

There is evidence that the SIT genes that had recently expanded post speciation between *N. alba* and *N. putrida* with the expansion date of the SIT genes occurring after the speciation date of the non-photosynthetic diatom species *N. putrida* and *N. alba*. This significant expansion within the SIT gene families was supported by the signs of diversifying selection within the SIT gene family which showed multiple sites under positive selection within different members of the SIT gene family (Kamikawa et al., 2021). This expansion is likely due to adaptation to the new heterotrophic lifestyle, with a possible driver being to increase uptake of organic food. This could be due to SIT genes' main functions relating to regulating the cell wall formation based on sensing silicic acid, as under most conditions silicic acid is at a sufficient concentration in the environment to diffuse into the cells passively. However, there is evidence that upregulation of SIT genes in diatoms is linked to rapidly dividing populations (Zielinski et al., 2016), therefore an alternate explanation may be that the expansion in SIT genes is a result of the change in lifestyle. This would cause rapid population increase, reducing local silicic acid concentrations and therefore causing selection pressures to act on

the SIT genes present inducing the recent gene duplication and gene family expansion seen in *N.putrida*. Another hypothesis is that the expansion is due to neo-functionalisation within the SIT genes. Increasing the survival of *N. putrida* through a new function in response to the shift in life-style through nutrient uptake as a result of the parasitism or feeding on macro algae. In this way, modification to the frustule that surrounds the diatom may provide it with a potential advantage in nutrient acquisition either through increased mobility or nutrient uptake.

The two control gene families, myosin and ammonium transporters, showed low expansion rates of around 0.1 and 0.18 respectively across the 20 MYA and 12 MYA of their trees (Fig. 11), these rates stayed fairly low across the time with limited branching occurring in either trees. In comparison to the control gene families, the SIT gene family has a higher expansion rate of 1.4 rising to 1.5 (Fig. 12) which suggests that there has been a consistently higher rate of expansions across the last 3 million years compared to the control gene families. In comparison to this both the MFS (Fig. 12B) and SST (Fig. 12C) have only slightly raised expansion rates with both maintaining a rate of between 0.2 and 0.3 across their 14 and 6 million years of expansion respectively. Although the SIT gene families had a higher rate of gene family expansion, this was consistent across time. This was similar in all the gene families with little change in rate of expansion over time. This suggests that the SIT gene families which show a speciation event occurring around 6.68 MYA have continued to expand at a constant rate reaching the 22 gene sequences currently present. This increased expansion rate supports the hypothesis that the SIT genes are under selection and has been occurring for the last 3 million years, demonstrated by a lack of any change in expansion rate over time.

In conclusion the analyses show that there is no significant acceleration of gene family expansion since speciation, although for the SIT gene families (Fig. 12) the rate of expansion was higher than that of the control gene families (Fig. 11). This work supports the hypothesis that the expansions were adaptions in response to the changing heterotrophic lifestyle that potentially occurred around 6.68 MY (Fig. 6), rather than a pre-adaptation that enabled the change. Future work into the loss of photosynthesis in *N. putridas* should investigate how a heterotrophic diatom could be used to re-introduce photosynthetic abilities to broaden the understanding of how photosynthesis potentially evolved in diatoms.

Chapter 3.

Identifying the Deleterious Mutational Load Within the Passenger Pigeon Genome.

Background

Pigeons and doves (*Columbiformes*) are one of the most diverse orders of birds containing over 300 species (Gill and Donsker, 2016). Recent work suggests that the majority of diversification within *Columbidae* occurred during the late Oligocene and into the Miocene (Selvatti et al., 2015) (Soares et al., 2016) around a period of widespread global cooling. The passenger pigeon (*Ectopistes migratorius*) was once the most abundant vertebrate species on earth with an estimated 3-5 billion individuals. However, by 1914 the species became extinct. The extinction is classically linked to the introduction of more efficient hunting techniques (Blockstein and Tordoff, 1985) and habitat destruction (Bucher 1992). The impact of natural selection was investigated and suggested that the genetic diversity of the passenger pigeon, when compared to the extant Ban-tailed pigeon, was much lower than expected for a population of 3-5 billion individuals (Murray et al., 2017). Recent investigations also indicate that habitat destruction was not a major influence on the rapid extinction due to the dietary plasticity of the species (Guiry et al., 2020), allowing for adaptation to exploit the expanding farmland. An alternative (mutually non-exclusive) hypothesis is that extinction of the passenger pigeon may have been a result of a combination of mutational load present in genome that when exposed under the external pressure of hunting led to an increase in inbreeding, which in turn accelerated the extinction of the species.

To understand how a species with such large census population size can become extinct at the rapid rate seen we must first understand the evolutionary forces acting upon the population. The five key evolutionary forces are selection, genetic drift, gene flow, recombination, and mutation, all acting upon the alleles and individuals within a population. The size of a population is often considered to greatly influence the effect of these forces, known as the “small population paradigm” it suggests that extinction disproportionately effects small populations. This is because it is assumed that large populations are able to respond and adapt due to their larger population size and gene pool. Under this theory, small populations are most at risk from the effects of genetic drift and inbreeding depression (Whitlock, 2000), both of which can lead to the rapid extinction of small and declining populations. Paired to this is the “declining population paradigm” (Caughley, 1994), which focuses on discovering the source of population decline and halting it. Recent investigations

show that although population size is important to species survival, consistently small populations have been shown to stay healthy due to purging of deleterious alleles from the genome (Valk et al., 2019). In this light, the extinction of the passenger pigeon shows that a very large population of individuals can become rapidly extinct, and this potentially reflects that it possessed a high load of recessive deleterious mutations (i.e., genetic load).

Indeed, just like the neutral genetic variation increases with effective population size, so does the genetic load of recessive deleterious mutations. The genetic load of completely recessive deleterious mutations could furthermore have been elevated in the passenger pigeon because of recurrent selective sweeps. The large population size of the species would have made it very responsive to natural selection (Fisher, 1930), and deleterious mutations that are linked to a genetic variant that is positively selected could have hitchhiked with this beneficial variant. This is known as “genetic draft” and it can resemble the signature of genetic drift by eroding genetic variation of nearby loci under selection (Gillespie, 2000) (Skipper, 2004) (Neher, 2013). This is also consistent with the identification of a relatively low genetic diversity within the passenger pigeon genome (Murray et al., 2017); although, other processes such population fluctuations prior to its extinction (Hung et al., 2018) could have contributed to this as well. One of the biggest enigmas in conservation biology is the extinction of the passenger pigeon; how could a population of so many individuals become extinct so quickly?

Mutational load is a mathematical calculation of the burden of deleterious variants within a population (Ohta and Gillespie, 1996) (Kimura, Maruyama and Crow, 1963). As mutations occur over evolutionary time, many deleterious or dominant variants are removed, but less deleterious recessive mutations are able to remain in the population in heterozygote condition (Henn et al., 2015). Therefore, with limited external pressures on the population due to the large size of the passenger pigeon flocks, prior to industrialised hunting, the passenger pigeon genome would have been accruing many recessive deleterious mutations. This mutational load, along with inbreeding depression, could have caused the low genetic diversity seen (Bouzat, 2010). The identification of deleterious genetic mutations is a key aspect of measuring the mutational load within a genome. A means of determining if the mutations that are present within the genome of a species are deleterious or not, is to focus the investigation on mutations that occur within the Ultra Conserved Elements (UCE) of the genome.

The UCE are the areas of the genome that are shared across individuals with an ancestral background often used in phylogenetics (Pierce, 2019). As these genes are highly conserved across multiple lineages and over a long evolutionary history, it can be assumed that genes in the UCE have a high level of sequence conservation and therefore any mutations that occur

within these UCEs are likely deleterious. Investigations into the human genome have identified a relatively large number of polymorphisms present within UCEs, which are estimated to occur at a frequency of one SNP per 21bp. Some of these polymorphisms are associated with genetic diseases or phenotypic traits (Habic et al., 2019). Therefore, investigations of mutations within the identified UCE can be undertaken using comparative analysis of species genomes, a common method is called the Genomic Evolutionary Rate Profiling (GERP) score (Davydov et al. 2010) (Cooper et al. 2005). This measures how the reduction in the number of substitutions in a multi-species sequence alignment compares to that expected. In this way the mutational load can be analysed to calculate the increase or decrease in variation across multi-species alignments (Huber, Kim and Lohmueller, 2020) and to determine if these variations are potentially deleterious (high GERP scores). GERP scores have been used in a range of studies including: investigations into the human genome, comparing Mayan Native American genomes and average San Sub-Saharan African genome (Henn et al., 2016); identifying deleterious amino acid-changing mutations occurring in dogs by comparisons of GERP scores between dogs and wolves (Marsden et al., 2016) and showing how purging deleterious alleles can be used to decrease the genetic load of small populations through GERP analysis across a range mammal species with historically large and small populations (Valk et al., 2019).

Another widely used method for analysis of deleterious alleles is Combined Annotation-Dependant Depletion (CADD) (Rentzsch et al., 2019), which ranks genetic variants, including SNPs as well as insertions and deletions (InDels), throughout the genome. CADD scores integrate the surrounding sequence context, gene model annotation, evolutionary constraints, epigenetic measurements, and functional predictions into the values produced. In this way the location of the mutation within the genome is assessed for the calculation. CADD has been developed for investigating conserved elements into the chicken Combined Annotation-Dependent Depletion (chCADD) (Gross et al., 2020) to identify regions within the chicken genome that are associated with known diseases. Due to the medical applications of the methodology, CADD analysis allows for both the identification of deleterious allele as well as the measurement of mutational load within the genome of a species, making it more accurate at determining the potential phenotypic negative effects of the mutations on the individuals and population.

Although the passenger pigeon went extinct with the death of the last individual, Martha, in 1914, understanding the causes of the population's rapid extinction and the effects of genetic mutational load that it possessed will help us to learn from historical failures, and improve the conservation-threatened species such as the pink pigeon (*Nesoenas mayeri*)

and echo parakeet (*Psittacula eques)*. The aims of this investigation are to determine if the passenger pigeon suffered from a significant deleterious mutational load prior to its extinction. When population size decreased this high load would have been expressed in homozygote condition leading to an accelerated extinction vortex, explaining the rapid extinction of one of the largest population of vertebrates globally. This investigation will also highlight the importance of genetic mutational load over population size as a measure of population fitness and conservation success and thereby provide a potential method for preserving genetic health of endangered species in the future.

Methods

Due to time constraints the proposed analysis was unable to be completed, therefore the following is a summary of the proposed methods and analysis. The methodology aims to investigate the mutational load within the UCE of the passenger pigeon genome assembly, built using the methods supplied by Murray et al., 2017, using the chCADD methodology described by Gross et al., 2020.

Genome assembly

The genome was to be assembled using the de novo band-tailed pigeon genome (Murray et al., 2017) as a reference, with the aim to assemble the nuclear genomes for four passenger pigeons (reads downloaded from the NCBI SRA). For all five pigeons, the sequencing adapters were removed using SeqPrep ("-M 0.05 -N 0.75 -m 0.8 -n 0.02 -X 0.25 -Z 26"),and mapped to the draft genome of the band-tailed pigeon using BWA-MEM 0.7.10 with default parameters (Li and Durbin, 2009). The produced SAM files were then sorted, indexed, and duplicates were removed using SAMtools (Li et al., 2009).

Read groups were then added to the analysis-ready reads using Picard tools AddOrReplaceReadGroups (http://broadinstitute.github.io/picard/). The reads were then processed using the Genome Analysis Toolkit GATK4.18 as described in the workflow according to GATK Best Practices recommendations (DePristo et al., 2011) (Van der Auwera et al., 2013) (Poplin et al. 2017) and the SNP call methodology used by Murray et al., 2017 to create five reference-based genomes for further analysis.

ChCADD analysis

Once a reference-based genome assembly had been completed for the passenger pigeon a multiple genome alignment would be performed along with the genomes of 23 sauropsid. Once completed the aligned UCE would be used to calculate the chCADD score calculated following the methodology of Gross et al., 2020, for the passenger pigeon genome determining the potential genetic load that is present in the genome.

Discussion

Causes of the rapid extinction

With a population as large as the Passenger pigeon, it has been questioned whether hunting alone would have been capable of its rapid extinction. Indirect effects of human activities, in particular habitat destruction, may also have contributed to its demise. Deforestation may have had severe effects because of the highly specialized diet and foraging strategy of the species. Mast (tree nuts) formed an important part of the passenger pigeon diet, which were depleted in the deforested habitats. A recent study used stable isotope and ancient DNA analysis to find evidence of two dietary trends within the population (Guiry et al., 2020). One passenger pigeon group showed evidence of a specialised diet of mast whilst another group specialised in maize foraging. This suggests that the passenger pigeon had a flexible diet and that individuals intentionally exploited the change in resources over their lives (Guiry et al., 2020). An alternative hypothesis, that will be examined further in future planned investigations, is that the population's rapid decline was a result of a poor "genetic heath". The low genetic diversity and population fluctuations prior to extinction indicate a high mutational load present within the genome. This likely would have exacerbated the population decline due to hunting and made it impossible for the population to recover. To understand how this would have occurred we must understand the hypothesis in turn.

The first hypothesis for the species' rapid decline is the classical view of over-hunting (Blockstein and Tordoff, 1985). This hypothesis is similar to over-fishing, which has been suggested to have caused extinctions in marine species with previously large populations through range contraction resulting in population collapse (Burgess et al., 2017). Although it has been suggested that due to high fecundity and low parental care, traits not common in birds, extinctions are considered rare (Pape et al., 2017). With the technological advances occurring during the 19th century and the expansion westward across America, there is evidence that populations can be hunted to extinction however, the rate of the extinction seems too impressive even for humans. This hypothesis can therefore be developed by adding the effects of a poor genetic health of the species. Once the population began to decrease suddenly, due to hunting, the recessive deleterious mutations present in the population would occur in homozygous condition at higher frequency due to increased inbreeding within the population and the fewer alleles present due to the decreased size and poor variation. With the deleterious alleles expressed and the decreasing population size, genetic drift would have accelerated the extinction vortex. An interesting hypothesis is that the harsh environmental factors of hunting limited the effects of selection (Trask et al., 2019). In this way, all individuals' reproductive fitness was equally low; thereby the deleterious alleles

permeated throughout the population as a neutral allele would. This high frequency would in turn result in an increase in individuals with homozygosity, therefore suffering from the genetic diseases. As these diseases become more common, their effect on the selection coefficient changes, increasing the selection pressure on individuals with the deleterious alleles, which by now encompasses the majority of the population. This would then lead to rapid extinction through inbreeding depression as well as the hunting occurring (made worse by the genetic disorders that they suffer from). Another hypothesis is that the extinction was caused by the allee effect (Allee and Bowen, 1932). Due to the low genetic diversity and consistently large population size for a long evolutionary time, many alleles were fixed in states that were beneficial for large population sizes but detrimental when in smaller flocks (Murray et al., 2017). In this way the decreasing population size led to a greater population decrease as individuals were poorly adapted to the selection pressures of a small population, potentially making them easier prey for hunting.

Each hypothesis individually may provide an adequate explanation for how the passenger pigeon became extinct; however, a more likely answer is that they are not mutually exclusive, but rather all interact resulting in the population collapse seen. The deleterious alleles constituting the high mutational load of the population could have become linked to alleles beneficial to large population size through genetic hitchhiking either as genetic draft or by the means of epistatic selection. Hence the deleterious alleles are conserved due to the direct benefit of the genes to which they are linked, in the case of the passenger pigeon genes beneficial to a large population size, thereby maintaining these alleles within the population hidden from both recombination and purifying selection. Therefore, when the population size decreased due to industrialisation and increased hunting, the recessive deleterious mutations became expressed in a homozygote condition and permeated though the populations resulting in severe inbreeding depression. When combined with the negative allee effects of behavioural adaptation to large population sizes, the additive effect of both of these results of the species's genetic history and mutational load amplify the presence of the industrialised hunting and would have led the rapidity of the extinction witnessed. In this hypothesis, even in the absence of hunting, the high mutational load present and the small founder population, the inbreeding depression and genetic drift would have resulted in the species's eventual extinction.

Implications for currently endangered species

If a large population size is ineffective at maintaining a population through population bottlenecks, hunting, invasive species or natural disasters when the effects of inbreeding and

high genetic mutational load are revealed. Therefore, many of the current conservation success stories are at a severe risk of extinction. The pink pigeon, is often considered one of the greatest success stories of modern conservation, rescuing the population of just 11 individuals from extinction through intense management and breeding, resulting in the current population of over 450 birds. Although no longer one of the rarest birds on the planet it is not "out of the woods yet". Vortex models have shown that without genetic reintroduction, as well as population and environmental management strategies, the population will become extinct within the next 100 years (Jackson et al., in prep). Similar threats are faced by other species, which suggests that conserved populations managed through primarily habitat or demographic rescue alone are at much greater threat of extinction than management strategies using both approaches (Trask et al., 2019). This suggests the need for a new man-made natural selection to act as the currently not present selection pressure upon the populations. This would be achieved by removing individuals with deleterious genetic traits from the breeding population and creating breeding strategies to promote the purging of genetic deleterious load from the gene pool of conserved or endangered species. It has been shown that environment and habitat harshness can negate the effects of selection (Trask et al., 2019). If the habitat is either too harsh or not harsh enough, populations accrue deleterious genetic mutations as all individuals have the same chance of survival and therefore reproductive success, meaning that deleterious alleles remain hidden from natural selection. It is therefore critical that conservation strategies should incorporate the testing and analysis of mutation load within populations and determine breeding strategies to minimise its effects, in this way maintaining the genetic health of the population and providing it with the ability to survive future potential population bottlenecks.

Future work

The current global climate crisis has meant that many species are threatened with extinction. It is therefore critical that an evaluation of the risk to each species be performed. With this goal at its forefront, the International Union for Conservation of Nature (IUCN) red list (IUCN, 2020) was designed. The IUCN red list aims to quantify the risk of extinction faced by 160,000 species before the end of 2020, a goal likely to fall short. Of the species currently identified 27% are currently threatened with extinction. The IUCN red list however does not currently take the measurement of genetic risk into the calculation of the risk to a species (Díez-del-Molino et al. 2018). If the passenger pigeon, a species of 3 - 5 billion individuals, was able to rapidly go extinct due to a poor underlying "genetic health" stemming from a high deleterious mutational load within the population, then far more species may currently be at risk of extinction than previously thought.

In conclusion, it is therefore critical that analysis of the mutational load within the genomes of endangered species becomes a commonly used tool for conservation. This allows for both assessments of species risk as well as management strategies and breeding regimes designed not to maximise genetic diversity within a population but minimise the genetic load of the population, thereby ensuring the survival of endangered species through the likely future population bottlenecks that they will face.

Definitions

Alternate allele – ALT

ATP-binding cassette transporter - ABC

Basic Local Alignment Search Tool - BLAST

Bayesian Evolutionary Analysis by Sampling Trees - BEAST

Binary sequence Alignment/Map -BAM

Binary variant Call Format - BCF

Chicken Combined Annotation-Dependant Depletion - chCADD

Combined Annotation-Dependant Depletion - CADD

Drag/Metabolite transporters- DMT

Genome Analysis Toolkit – GATK

Genomic Evolutionary Rate Profiling - GERP

Glycosyl Hydrolase - GH

Glycosyl Transferase - GT

Heterozygosity – HZ

Insertions and Deletions - InDels

International Union for Conservation of Nature - IUCN

Major Facilitator Superfamily - MFS

Maximum Clade Credibility - MCC

Million Years Ago -MYA

Nucleotide diversity – pi

Reference allele – REF

Resistance-Nodulation-Cell division superfamily - RND

Sequence Alignment/Map - SAM

Silicon Ion Transporter – SIT

Single Nucleotide Polymorphism - SNP

Solute:Sodium symporters - SST

Tree Evolution Simulation Software -TESS

Ultra-Conserved Elements – UCE

Variant Call Format – VCF

References

Allee, W.C. and Bowen, E.S., 1932. Studies in animal aggregations: mass protection against colloidal silver among goldfishes. Journal of Experimental Zoology, 61(2), pp.185-207.

Benoiston, A.S., Ibarbalz, F.M., Bittner, L., Guidi, L., Jahn, O., Dutkiewicz, S. and Bowler, C., 2017. The evolution of diatoms and their biogeochemical functions. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1728). doi:/10.1098/rstb.2016.0397.

Blockstein, D.E., Tordoff, H.B., 1985. Gone forever: a contemporary look at the extinction of the passenger pigeon Am. Birds, 39 (5), pp. 845-851

Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N. and Matschiner, M., 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS computational biology, 15(4), p.e1006650.

Bouzat, J.L., 2010. Conservation genetics of population bottlenecks: the role of chance, selection, and history. Conservation Genetics, 11(2), pp.463-478.

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P. and Rayko, E., 2008. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature, 456(7219), pp.239-244.

Brembu, T., Chauton, M.S., Winge, P., Bones, A.M. and Vadstein, O., 2017. Dynamic responses to silicon in Thalasiossira pseudonana-Identification, characterisation and classification of signature genes and their corresponding protein motifs. Scientific reports, 7(1), pp.1-14.

Bucher, E.H., 1992. The causes of extinction of the passenger pigeon. In Current ornithology, Springer, Boston, pp. 1-36

Burgess, M.G., Costello, C., Fredston-Hermann, A., Pinsky, M.L., Gaines, S.D., Tilman, D. and Polasky, S., 2017. Range contraction enables harvesting to extinction. Proceedings of the National Academy of Sciences, 114(15), pp.3945-3950.

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25(15), pp.1972-1973.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution, 17(4), pp.540-552.

Caughley, G., 1994. Directions in conservation biology. Journal of animal ecology, pp.215-244.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S. and Sidow, A., 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome research, 15(7), pp.901-913.

Daboussi, F., Leduc, S., Maréchal, A., Dubois, G., Guyot, V., Perez-Michaut, C., Amato, A., Falciatore, A., Juillerat, A., Beurdeley, M. and Voytas, D.F., 2014. Genome engineering empowers the diatom Phaeodactylum tricornutum for biotechnology. Nature communications, 5(1), pp.1-7.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. doi:10.1093/bioinformatics/btr330

Darwin's, C., 1859. On the origin of species.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A. and Batzoglou, S., 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol, 6(12), p.e1001025.

Depauw, F.A., Rogato, A., Ribera d'Alcalá, M. and Falciatore, A., 2012. Exploring the molecular basis of responses to light in marine diatoms. Journal of experimental botany, 63(4), pp.1575-1591.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M. and McKenna, A., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics, 43(5), pp.491-498.

Díez-del-Molino, D., Sánchez-Barreiro, F., Barnes, I., Gilbert, M.T.P. and Dalén, L., 2018. Quantifying temporal genomic erosion in endangered species. Trends in ecology & evolution, 33(3), pp.176-185.

Drummond, A.J. and Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology, 7(1), pp.1-8.

Durkin, C.A., Koester, J.A., Bender, S.J., Armbrust, E.V., 2016. The evolution of silicon transporters in diatoms. J. Phycol. 52, 716–731. doi:10.1111/jpy.12441

Fisher, R. A., 1930. The Genetical Theory of Natural Selection. Oxford: Oxford University Press.

Gersonde, R. & Harwood, D. M. 1990. Lower Cretaceous diatoms from ODP Leg 113 Site 693 (Weddell Sea). Part 1. Vegetative cells. In Proceedings of the Ocean Drilling Program, Scientific Results. Ocean Drilling Program, College Station, Texas, pp. 365–402.

Gillespie, J.H., 2000. Genetic drift in an infinite population: the pseudohitchhiking model. Genetics, 155(2), pp.909-919.

Gross, C., Bortoluzzi, C., de Ridder, D., Megens, H.J., Groenen, M., Reinders, M. and Bosse, M., 2020. Evolutionarily conserved non-protein-coding regions in the chicken genome harbor functionally important variation. bioRxiv.

Guiry, E.J., Orchard, T.J., Royle, T.C., Cheung, C. and Yang, D.Y., 2020. Dietary plasticity and the extinction of the passenger pigeon (Ectopistes migratorius). Quaternary Science Reviews, 233. doi:/10.1016/j.quascirev.2020.106225

Habic, A., Mattick, J.S., Calin, G.A., Krese, R., Konc, J. and Kunej, T., 2019. Genetic Variations of Ultraconserved Elements in the Human Genome. Omics: a journal of integrative biology, 23(11), pp.549-559.

Henn, B.M., Botigué, L.R., Bustamante, C.D., Clark, A.G. and Gravel, S., 2015. Estimating the mutation load in human genomes. Nature Reviews Genetics, 16(6), pp.333-343.

Henn, B.M., Botigué, L.R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B.K., Martin, A.R., Musharoff, S., Cann, H., Snyder, M.P. and Excoffier, L., 2016. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proceedings of the National Academy of Sciences, 113(4), pp.E440-E449.

Höhna, S., May, M.R., Moore, B.R., 2016. TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. Bioinformatics 32, 789–791. doi:10.1093/bioinformatics/btv651

Huber, C.D., Kim, B.Y., Lohmueller, K.E., 2020. Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. BioRxiv. doi:10.1101/836858

Hung, C.-M., Zink, R.M., Li, S.-H., 2018. Can genomic variation explain the extinction of the passenger pigeon? J. Avian Biol. 49, e01858. doi:10.1111/jav.01858

IUCN 2020. The IUCN Red List of Threatened Species. Version 2020-2. https://www.iucnredlist.org

Jackson, H., Percival-Alwyn, L., Ryan, C., Albeshr, M.F., Venturi, L., Mathers, T.C., Cocker, J., Speak, S.A., Accinelli, G.G., Willman, F., Dawson, D., Ward, L., Tatayah, V., Zuël, N., Young, R., Concannon, L., Bunbury, N., Tyler, K.M., Ruhomaun, K., Bruford, M.W., Jones, C.G., Tollington, S., Bell, D., Groombridge, J.J., Clark, M., van Oosterhout, C., [in prep]. A blind spot in the IUCN Red List criteria; not all is rosy for the Pink Pigeon.

Kamikawa, R., Mochizuki, T., Sakamoto, M., Tanizawa, Y., Nakayama, T., Onuma, R., Cenci, U., Moog, D., Speak, S., Sarkozi, K. and Toseland, A., 2021. Genome evolution of a non-parasitic secondary heterotroph, the diatom Nitzschia putrida. bioRxiv.

Kamikawa, R., Moog, D., Zauner, S., Tanifuji, G., Ishida, K.-I., Miyashita, H., Mayama, S., Hashimoto, T., Maier, U.G., Archibald, J.M., Inagaki, Y., 2017. A Non-photosynthetic Diatom Reveals Early Steps of Reductive Evolution in Plastids. Mol. Biol. Evol. 34, 2355–2366. doi:10.1093/molbev/msx172

Kamikawa, R., Yubuki, N., Yoshida, M., Taira, M., Nakamura, N., Ishida, K., Leander, B.S., Miyashita, H., Hashimoto, T., Mayama, S., Inagaki, Y., 2015. Multiple losses of photosynthesis in Nitzschia (Bacillariophyceae). Phycological Res. 63, 19–28. doi:10.1111/pre.12072

Kimura, M., Maruyama, T. and Crow, J.F., 1963. The mutation load in small populations. Genetics, 48(10), pp.1303-1312.

Kingman, J.F., 1982. On the genealogy of large populations. Journal of applied probability, pp.27-43.

Krause, J.W., Schulz, I.K., Rowe, K.A., Dobbins, W., Winding, M.H., Sejr, M.K., Duarte, C.M. and Agustí, S., 2019. Silicic acid limitation drives bloom termination and potential carbon sequestration in an Arctic bloom. Scientific reports, 9(1), pp.1-11.

Lazarus, D., Barron, J., Renaudie, J., Diver, P., Türke, A., (2014). Cenozoic planktonic marine diatom diversity and correlation to climate change. PLoS ONE 9, e84857. doi:10.1371/journal.pone.0084857

Le Pape, O., Bonhommeau, S., Nieblas, A.E. and Fromentin, J.M., 2017. Overfishing causes frequent fish population collapses but rare extinctions. Proceedings of the National Academy of Sciences, 114(31), pp.6274-6274.

Lescot, M., Hingamp, P., Kojima, K.K., Villar, E., Romac, S., Veluchamy, A., Boccara, M., Jaillon, O., Iudicone, D., Bowler, C. and Wincker, P., 2016. Reverse transcriptase genes are

highly abundant and transcriptionally active in marine plankton assemblages. The ISME journal, 10(5), pp.1134-1146.

Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics, 25(14), pp.1754-1760.

Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21), pp.2987-2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), pp.2078-2079.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), pp.2078-2079.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), pp.2078-2079.

Li, W., Jaroszewski, L. and Godzik, A., 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics, 17(3), pp.282-283.

Llorens, C., Futami, R., Covelli, L., Dominguez-Escriba, L., Viu, J.M., Tamarit, D., Aguilar-Rodriguez, J. Vicente-Ripolles, M., Fuster, G., Bernet, G.P., Maumus, F., Munoz-Pomer, A., Sempere, J.M., LaTorre, A., Moya, A. (2011) The Gypsy Database (GyDB) of Mobile Genetic Elements: Release 2.0 Nucleic Acids Research (NARESE) 39 (suppl 1): D70-D74 doi: 10.1093/nar/gkq1061

Löytynoja, A., 2014. Phylogeny-aware alignment with PRANK. In Multiple sequence alignment methods, pp. 155-170.

Maldonado, M., Carmona, M. C., Uriz, M. J. & Cruzado, A. 1999. Decline in Mesozoic reef-building sponges explained by silicon limitation. Nature, 401, pp.785–788.

Marsden, C.D., Ortega-Del Vecchyo, D., O'Brien, D.P., Taylor, J.F., Ramirez, O., Vilà, C., Marques-Bonet, T., Schnabel, R.D., Wayne, R.K. and Lohmueller, K.E., 2016. Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proceedings of the National Academy of Sciences, 113(1), pp.152-157.

Maumus, F., Allen, A. E., Mhiri, C., Hu, H., Jabbari, K., Vardi, A., Grandbastien, M. A., & Bowler, C. 2009. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BMC genomics, 10, 624. https://doi-org.uea.idm.oclc.org/10.1186/1471-2164-10-624

Mock, T., Otillar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A., Sanges, R., Toseland, A., Ward, B.J., Allen, A.E., Dupont, C.L., Frickenhaus, S., Maumus, F., Veluchamy, A., Wu, T., Barry, K.W., Falciatore, A., Ferrante, M.I., Fortunato, A.E., Glöckner, G., Gruber, A., Hipkin, R., Janech, M.G., Kroth, P.G., Leese, F., Lindquist, E.A., Lyon, B.R., Martin, J., Mayer, C., Parker, M., Quesneville, H., Raymond, J.A., Uhlig, C., Valas, R.E., Valentin, K.U., Worden, A.Z., Armbrust, E.V., Clark, M.D., Bowler, C., Green, B.R., Moulton, V., van Oosterhout, C., Grigoriev, I.V., 2017. Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature 541, 536–540. doi:10.1038/nature20803

Murray, G.G.R., Soares, A.E.R., Novak, B.J., Schaefer, N.K., Cahill, J.A., Baker, A.J., Demboski, J.R., Doll, A., Da Fonseca, R.R., Fulton, T.L., Gilbert, M.T.P., Heintzman, P.D., Letts, B., McIntosh, G., O'Connell, B.L., Peck, M., Pipes, M.-L., Rice, E.S., Santos, K.M., Sohrweide, A.G., Vohr, S.H., Corbett-Detig, R.B., Green, R.E., Shapiro, B., 2017. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. Science 358, 951–954. doi:10.1126/science.aao0960

Nakov, T., Beaulieu, J.M., Alverson, A.J., 2018. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). New Phytol. 219, 462–473. doi:10.1111/nph.15137

National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2020 Sep 14]. Available from: https://www.ncbi.nlm.nih.gov/

Neher, R.A., 2013. Genetic draft, selective interference, and population genetics of rapid adaptation. Annual review of Ecology, evolution, and Systematics, 44, pp.195-215.

Nei, M., 1977. F-statistics and analysis of gene diversity in subdivided populations. Annals of human genetics, 41(2), pp.225-233.

NOAA National Centers for Environmental Information, State of the Climate: Global Climate Report for Annual 2019, published online January 2020, retrieved on September 15, 2020. Available from: https://www.ncdc.noaa.gov/sotc/global/201913.

Ohta, T. and Gillespie, J.H., 1996. Development of neutral and nearly neutral theories. Theoretical population biology, 49(2), pp.128-142.

Pace, N.R., Sapp, J. and Goldenfeld, N., 2012. Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. Proceedings of the National Academy of Sciences, 109(4), pp.1011-1018.

Pargana, A., Musacchia, F., Sanges, R., Russo, M.T., Ferrante, M.I., Bowler, C. and Zingone, A., 2020. Intraspecific Diversity in the Cold Stress Response of Transposable Elements in the Diatom Leptocylindrus aporus. Genes, 11(1), pp. https://doi.org/10.3390/genes11010009

Pierce, M., 2019. Filling in the gaps: adopting ultraconserved elements alongside COI to strengthen metabarcoding studies. Frontiers in Ecology and Evolution, 7. doi:10.3389/fevo.2019.00469

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D. and Shakir, K., 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv. doi:/10.1101/201178.

Rambaut, A., 2009. http://tree.bio.ed.ac.uk/software/tracer/

Rambaut, A., 2012. FigTree v1. 4.

Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. and Kircher, M., 2019. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research, 47(D1), pp.886-894.

Reynolds, J., Weir, B.S., Cockerham, C.C., 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. Genetics 105, pp.767–779

Roger, A.J., Muñoz-Gómez, S.A., Kamikawa, R., 2017. The origin and diversification of mitochondria. Curr. Biol. 27, R1177–R1192. doi:10.1016/j.cub.2017.09.015

Schaum, C.-E., Buckling, A., Smirnoff, N., Studholme, D.J., Yvon-Durocher, G., 2018. Environmental fluctuations accelerate molecular evolution of thermal tolerance in a marine diatom. Nat. Commun. 9, 1719. doi:10.1038/s41467-018-03906-5

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112–1115.

Selvatti, A.P., Gonzaga, L.P., Russo, C.A. de M., 2015. A Paleogene origin for crown passerines and the diversification of the Oscines in the New World. Mol. Phylogenet. Evol. 88, 1–15. doi:10.1016/j.ympev.2015.03.018

Shrestha, R.P. and Hildebrand, M., 2015. Evidence for a regulatory role of diatom silicon transporters in cellular silicon responses. Eukaryotic cell, 14(1), pp.29-40.

Siever, R., 1991. Silica in the oceans: biological-geochemical interplay. In Schneider, S.H. & Boston, P.J. [Ed.] Scientists on Gaia. MIT Press, Boston, pp. 287–295.

Skipper Jr, R.A., 2004. Stochastic evolutionary dynamics: Drift versus draft. Philosophy of Science, 73(5), pp.655-665.

Slotkin, R.K. and Martienssen, R., 2007. Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics, 8(4), pp.272-285.

Soares, A.E.R., Novak, B.J., Haile, J., Heupink, T.H., Fjeldså, J., Gilbert, M.T.P., Poinar, H., Church, G.M., Shapiro, B., 2016. Complete mitochondrial genomes of living and extinct pigeons revise the timing of the columbiform radiation. BMC Evol. Biol. 16, 230. doi:10.1186/s12862-016-0800-3

Trask, A.E., Fenn, S.R., Bignal, E.M., McCracken, D.I., Monaghan, P. and Reid, J.M., 2019. Evaluating the efficacy of independent versus simultaneous management strategies to address ecological and genetic threats to population viability. Journal of Applied Ecology, 56(10), pp.2264-2273.

Tréguer, P., Nelson, D. M., Van Bennekom, A. J., DeMaster, D. J., Leynaert, A. & Quéguiner, B., 1995. The silica balance in the world ocean: a reestimate. Science, 268, pp.375–379.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J. and Banks, E., 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics, 43(1), pp.11-10.

van der Valk, T., de Manuel, M., Marques-Bonet, T. and Guschanski, K., 2019. Estimates of genetic load in small populations suggest extensive purging of deleterious alleles. bioRxiv. doi: /10.1101/696831.

Weir, B.S. and Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. evolution, pp.1358-1370.

Weir, B.S., 2012. Estimating F-statistics: A historical view. Philos. Sci. 79, 637–643. doi:10.1086/667904

Whitlock, M.C., 2000. Fixation of new alleles and the extinction of small populations: drift load, beneficial alleles, and sexual selection. Evolution, 54(6), pp.1855-1861.

Wright, B., Farquharson, K.A., McLennan, E.A., Belov, K., Hogg, C.J., Grueber, C.E., 2019. From reference genomes to population genomics: comparing three reference-aligned

reduced-representation sequencing pipelines in two wildlife species. BMC Genomics 20, 453. doi:10.1186/s12864-019-5806-y

Yang, Z., Rannala, B., 2012. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. 13, 303–314. doi:10.1038/nrg3186

Zeh, D.W.; Zeh, J.A.; Ishida, Y. Transposable elements and an epigenetic basis for punctuated equilibria. BioEssays 2009, 31, 715–726.

Zheng Zhang, Scott Schwartz, Lukas Wagner, and Webb Miller (2000), "A greedy algorithm for aligning DNA sequences", J Comput Biol 2000; 7(1-2):203-14.

Zielinski, B.L., Allen, A.E., Carpenter, E.J., Coles, V.J., Crump, B.C., Doherty, M., Foster, R.A., Goes, J.I., Gomes, H.R., Hood, R.R. and McCrow, J.P., 2016. Patterns of transcript abundance of eukaryotic biogeochemically-relevant genes in the Amazon River Plume. PLoS One, 11(9), p.e0160929.

Appendix

Investigating the adaptive evolution in a diatom's genome in response to extreme temperature selection.

FST across regimes.

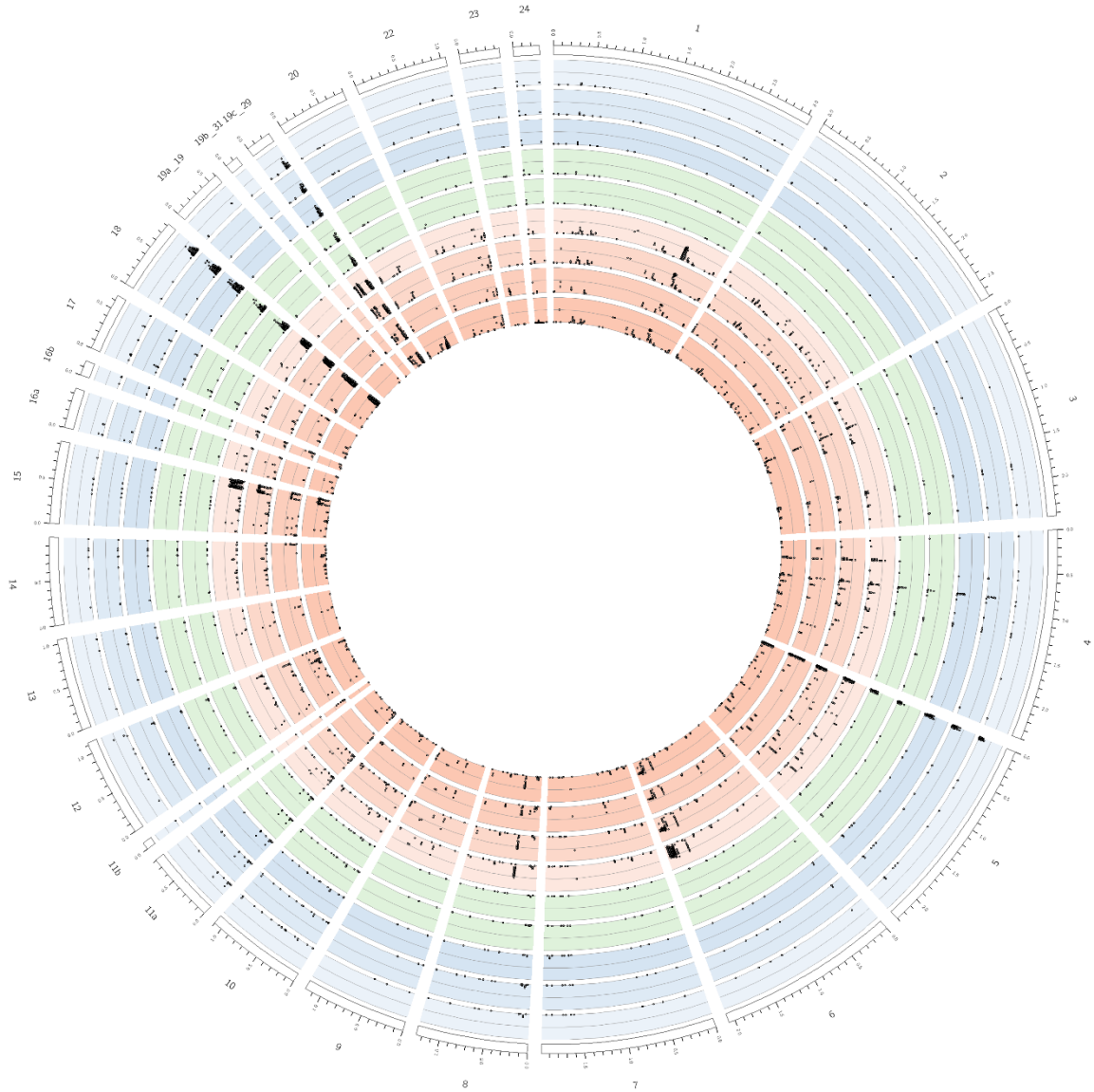

Figure 13 - Fst of each temperature regime heat stressed 32 ºC (Red), cold stressed 9 ºC (Blue) and control 22 ºC (Green) compared to the Control regime at 0 generations, with increasing colour fill corresponding to increasing generation number for each regime. Shown across all chromosomes with one chromosome per segment. Fst values >0 show. (figure produced by Toseland unpublished)

SNP allele frequency calculation.

```
while read p; do for file in Input.bcf; do bcftools query -R ListOfPositions.txt -s $p
        -f'%CHROM\t%POS\t%REF\t%ALT[\t%DP\t%AD{0}\t%AD{1}\t%AD{2}
        \t%AD{3}\t%SAMPLE]\n' $file | awk '{if ($5
        > 0) print ($10"\t"$1"\t"$2"\t"$6/$5"\t"$7/$5"\t"$8/$5"\t"$9/$5)}' >
        > ${file%.bcf}_sample_code.txt ; done; done < SampleCodes.txt
```

This process took the input filtered BCF file for the chromosome of interest, a file containing the positions that were being identified and a text file containing the sample codes (each sample code related to a sampling point time point with also indicating the temperature regimes and which of the5 lines for each regime). This would then output the sample code (this was converted using the unix commands awks and sed to display the line, generation and temperature regime), the chromosome, the base position, and the allele frequency for the reference allele, alternate allele and if present multiple alternate alleles (Table.1)

Table 2 – Table showing an example output from allele frequency call scripts Equation 1. Showing generation at the sampling point, temprature regime, Line for each regime, chromosome, position along chromosome, allele frequency for the reference, alternate allele 1, alternate allele 2 and alternate allele 3 (if present). This output shows the presence of multiple alternate alleles present across the three temperature regimes and lines, with increased allele frequency for alternate allele 1 in heat stressed lines.

Generation	Temp	Line	Chr	Pos	Ref	Alt_1	Alt_2	Alt_3
0	22C	L1	15	819182	0.379032	0.620968	0	0
70	22C	L1	15	819182	0.65	0.35	0	0
70	22C	L2	15	819182	0.652174	0	0.26087	0
70	22C	L3	15	819182	0.5	0.5	0	0
70	22C	L4	15	819182	0.647059	0	0.235294	0
70	22C	L5	15	819182	0.5	0.5	0	0
270	22C	L2	15	819182	0.388889	0.611111	0	0
270	22C	L3	15	819182	0.6	0.4	0	0
270	22C	L4	15	819182	0.65	0	0.25	0
270	22C	L5	15	819182	0.65	0.35	0	0
32	9C	L1	15	819182	0.611111	0.388889	0	0
32	9C	L2	15	819182	0.45	0.55	0	0
32	9C	L3	15	819182	0.578947	0	0.421053	0
32	9C	L4	15	819182	0.310345	0.689655	0	0
32	9C	L5	15	819182	0.75	0	0.25	0
144	9C	L1	15	819182	0.321429	0.678571	0	0
144	9C	L2	15	819182	0.272727	0.727273	0	0
144	9C	L3	15	819182	0.444444	0.555556	0	0
144	9C	L4	15	819182	0.6875	0	0.3125	0
144	9C	L5	15	819182	0.4	0.6	0	0
250	9C	L1	15	819182	0.214286	0.785714	0	0
250	9C	L2	15	819182	0.727273	0	0.272727	0
250	9C	L3	15	819182	0.695652	0	0.173913	0
250	9C	L4	15	819182	0.5	0.5	0	0
250	9C	L5	15	819182	0.35	0.65	0	0
50	32C	L2	15	819182	0	1	0	0
50	32C	L3	15	819182	0.0625	0.9375	0	0
50	32C	L4	15	819182	0	1	0	0
50	32C	L5	15	819182	0	1	0	0
210	32C	L1	15	819182	0.0555556	0.944444	0	0
210	32C	L3	15	819182	0	1	0	0
210	32C	L5	15	819182	0	1	0	0
300	32C	L1	15	819182	0	1	0	0
450	32C	L2	15	819182	0	1	0	0
450	32C	L3	15	819182	0	1	0	0
450	32C	L5	15	819182	0	1	0	0

Allele frequency change.

```
bcftools query -r RegionForAnalysis -S SampleCodes.txt
-f'%CHROM\t%POS\t%REF\t%ALT[\t%DP\t%AD{0}\t%AD{1}]
\n' Input.bcf | awk '{if (($5 > 0)&&($8
> 0)) print ($1"\t"$2"\t"$6/$5"\t"$7/$5"\t"$9/$8"\t"$10/$8"\t"(($10/$8)
-($6/$5))/((($6/$5)+($9/$8))/2)"\t"(($10/$8)-($7/$5))/((($7/$5)
+($10/$8))/2))}' > Output.txt
```

This script worked on a similar basis to that of the allele frequency calculation, with the added calculation to calculate the difference between the starting allele frequency for T0 of the control temperature regime to the line that is under analysis at time point 210 for the heat stressed lines. For this rather than a SNP being analysed it was calculated across a region inputted with the -r flag.

Heterozygosity

```
bcftools query -r RegionForAnalysis -S SampleCodesPairwise.txt
-f'%CHROM\t%POS\t%REF\t%ALT[\t%DP\t%AD{0}\t%AD{1}]
\n' Input.bcf | awk '{if ($8
> 0) print ($1"\t"$2"\t"1-(($9/$8)**2)"\t"1-(($10/$8)**2))}'
> Output.txt
```

This script calculated the heterozygosity of sites on a per site basis

Nucleotide Diversity.

```
vcftools --bcf Input.bcf --keep SampleCodes.txt --window-pi 1000 --window-pi
-step 100 --out OutputPrefix
```

A Phylogenetic Analysis Of The Heterotrophic Diatom *Nitzschia putrida*

Sequence alignment

```
grep "Major facilitator superfamily"
~/Nitzschia/Fc_Genes/Fc_annotations/Fracy1_domaininfo_FilteredModels1.tab | cut -f 1 >
MFS_Fc_ID.txt


for i in $(cat MFS_Fc_ID.txt); do grep $i
~/Nitzschia/Fc_Genes/Fc_genome/Fracy1_GeneModels_FilteredModels1_nt.fasta | sed
's/>//g'>> MFS_Fc_full_Id.txt | sort MFS_Fc_full_Id.txt | uniq > MFS_Fc_full_Id_sorted.txt;
done


sort MFS_Fc_full_Id.txt | uniq > MFS_Fc_full_Id_sorted.txt
```

BEAST Analysis output.

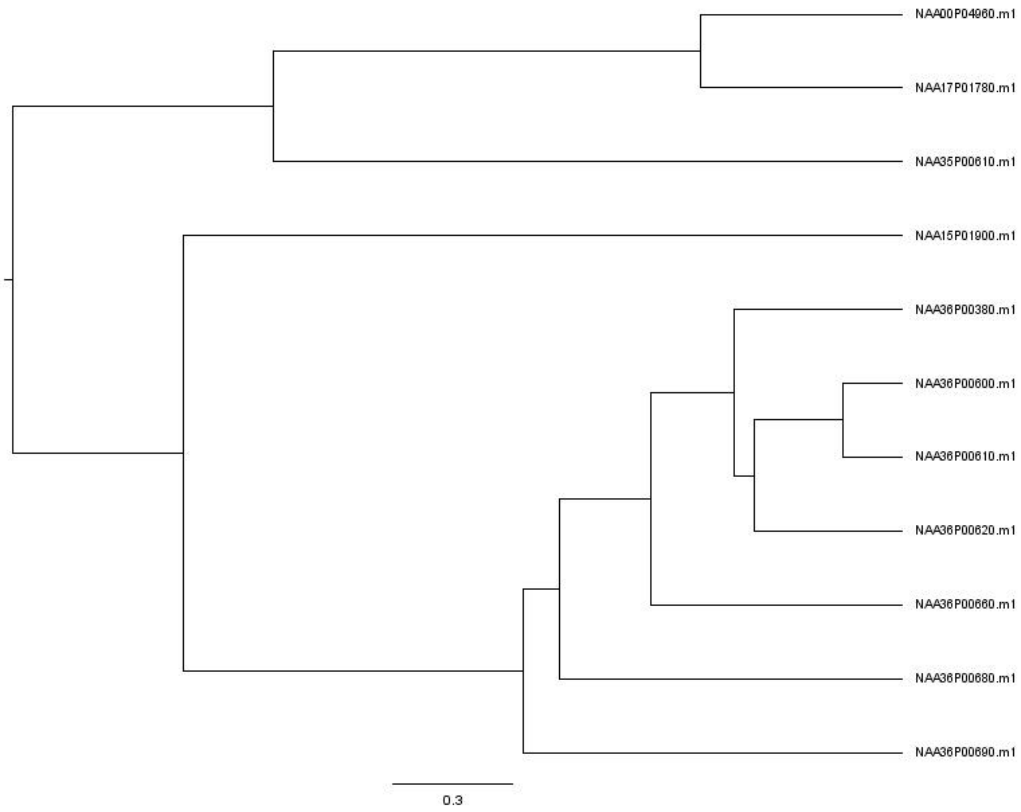

Figure 14 - Phylogenetic Maximum Clade Credibility (MCC) tree summarised by TreeAnnotator v2.6.1 for the ABC Transporter (ABC) genes for the largest cluster of expanded ABC gene family of Nitzschia putrida NIES-4235 (n=11). Divergence estimates were unable to be obtained using Bayesian Markov Chain Monte Carlo (MCMC) analyse due to poor alignment between Nitzschia sequences and corresponding sequences from related species resulting in no sequences with no time points with which to calibrate the phylogenetic analysis implemented in Beast v2.6.1. (Bouckaert et al., 2019), phylogenetic trees created using FigTree v1.4.3.

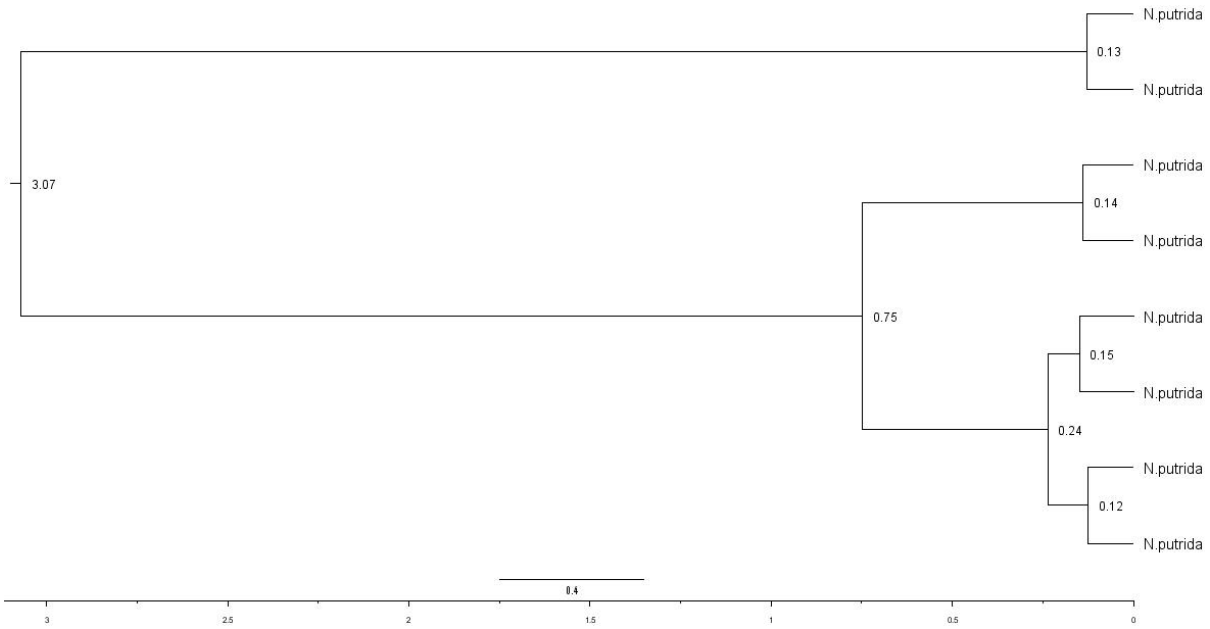

Figure 15 - Phylogenetic Maximum Clade Credibility (MCC) tree summarised by TreeAnnotator v2.6.1 for the Resistance-Nodulation-Cell division superfamily (RND) genes for the largest cluster of expanded RND gene family of *Nitzschia putrida* NIES-4235 (n=8). Divergence estimates were obtained using Bayesian Markov Chain Monte Carlo (MCMC) analyse using mutation rates, implemented in Beast v2.6.1. (Bouckaert et al., 2019), phylogenetic trees created using FigTree v1.4.3. Divergence estimates for all nodes given in Millions of years (Myr) before present.

TESS Analysis.

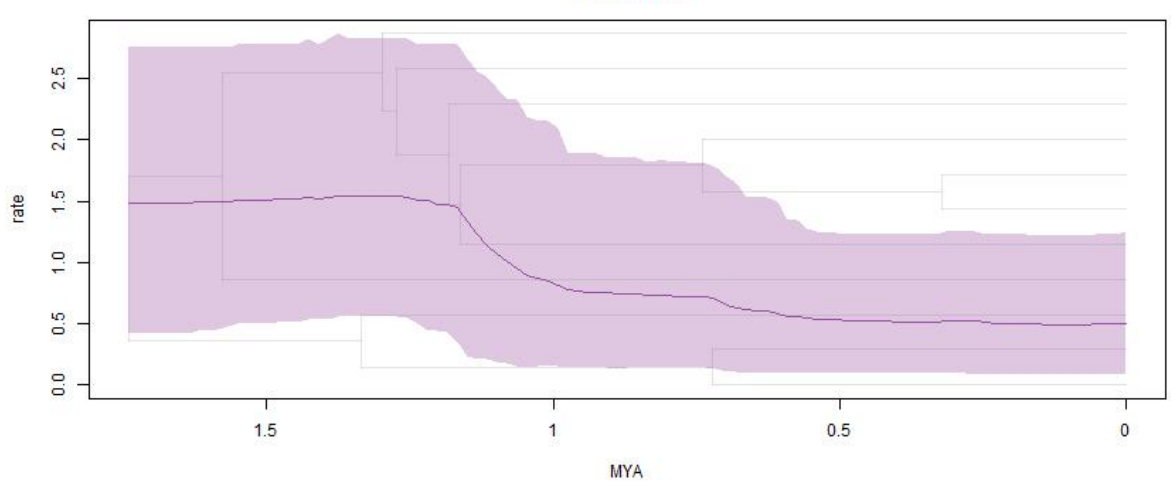

Figure 16- Expansion rate of gene sequences within the expanded ABC transporter gene family of the non-photosynthesising diatom *Nitzschia putrida* NIES-4235. Expansion rate was calculated using the Bayesian inference lineage diversification rate analysis tool TESS for R (Höhna, May and Moore, 2016) using the MCC phylogenetic tree produced by BEAST v2.6.1(Bouckaert et al., 2019). Phylogenetic tree shown for reference. There is no evidence for a significant change in the expansion rate in the past 2 million years.

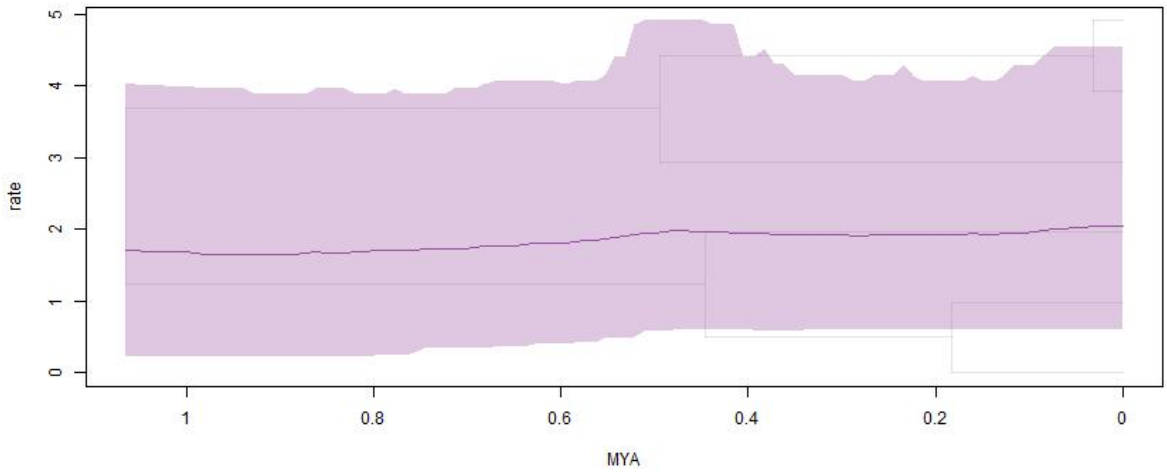

Figure 17- Expansion rate of gene sequences within the expanded Drag/Metabolite transporters gene family of the non-photosynthesising diatom *Nitzschia putrida* NIES-4235. Expansion rate was calculated using the Bayesian inference lineage diversification rate analysis tool TESS for R (Höhna, May and Moore, 2016) using the MCC phylogenetic tree produced by BEAST v2.6.1(Bouckaert et al., 2019). Phylogenetic tree shown for reference. There is no evidence for a significant change in the expansion rate in the past 1 million years

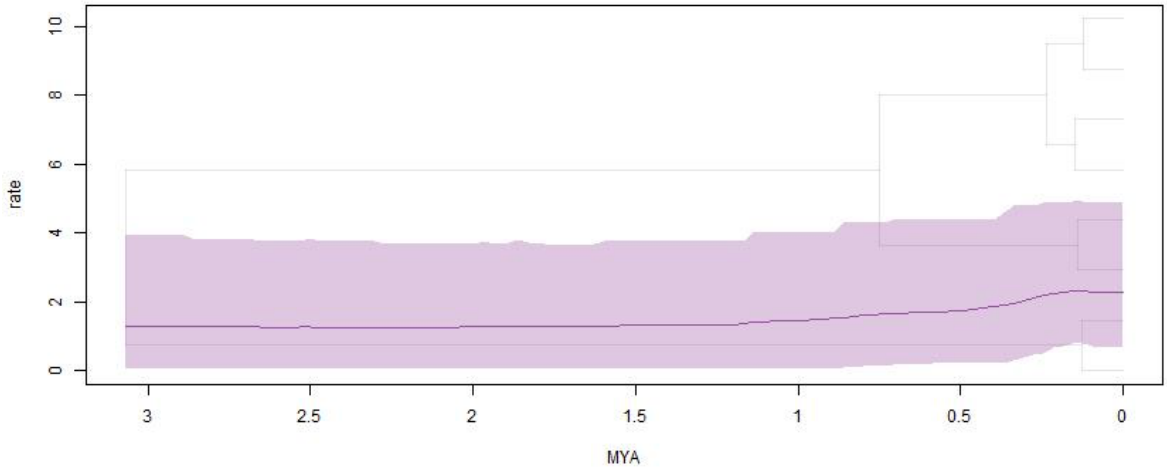

Figure 18 -Expansion rate of gene sequences within the expanded Resistance-Nodulation-Cell division superfamily of the non-photosynthesising diatom *Nitzschia putrida* NIES-4235. Expansion rate was calculated using the Bayesian inference lineage diversification rate analysis tool TESS for R (Höhna, May and Moore, 2016) using the MCC phylogenetic tree produced by BEAST v2.6.1(Bouckaert et al., 2019). Phylogenetic tree shown for reference. There is no evidence for a significant change in the expansion rate in the past 3 million years