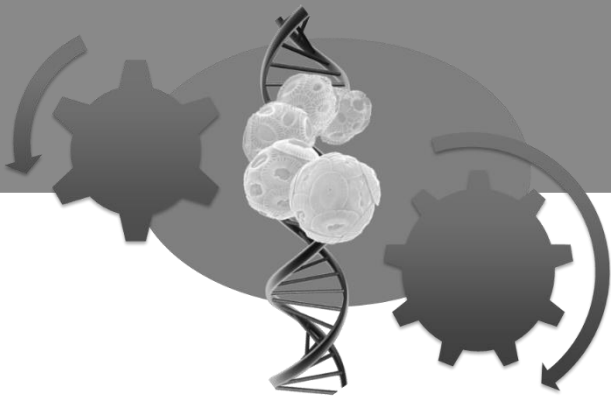# Detection of genome wide mirco-evolutionary processes in *Emiliania huxleyi* ((Lohm) Hayard Mohler)

Jennifer Hülskötter
Bremen, 28.09.2015

First examiner:
Dr. Uwe John

Second examiner:
Prof. Dr. Kai Bischof

University Bremen

Faculty 2

Master of Science Marine Biology

Detection of genome wide micro-evolutionary processes in Emiliania huxleyi ((Lohm) Hayard and Mohler)

Statutory Declaration

Herewith I declare that I have written the enclosed Master thesis completely by myself. Further I declare that I have not used any other sources as the ones cited in this thesis and any thoughts from other literal quotations is marked. Written and electronical versions of the enclosed Master thesis are equivalent and were not used to achieve any other academic grading.

Date, City   Signature (Jennifer Hülskötter)

CONTENT

I. LIST OF FIGURES

II. LISTE OF TABLES

III. ABSTRACT

Phytoplankton is responsible for approximately 50 % of the annual global primary production and *Emiliania huxleyi* ((Lohm.) Hayard and Mohler) is an important part of the phytoplankton community. It contributes to the biological- and carbon counter-pump and therefore species success of *E. huxleyi* might play an important role on vertical carbon fluxes in the ocean. Several studies have shown that *E. huxleyi* is susceptible for Global change. *Emiliania huxleyi* has a high adaptive capacity and can thrive in highly diverse but also stable environments from the subtropics to subpolar regions. *Emiliania huxleyi* has a biphasic life cycle with phenotypically distinct haploid and diploid cell stages interconnected by syngamy, a form of sexual reproduction. A recent study unveiled that *E. huxleyi* populations subjected to stable environments had a significant proportion of strains which have lost key genes associated to haploid cell formation. The loss of sex implies several evolutionary constraints for species survival and adaptive capacity, since asexual reproducing lineages tend to possess a loss of heterozygosity. However with an increase of homozygosity deleterious alleles can be unmasked and the fitness of a given species can be reduced. As a counterpart acts here ameiotic recombination, which can retain adaptive capacity in asexual lineages, by balancing the emergence of deleterious alleles. As asexual and sexual populations of *E. huxleyi* from the same geographic origin, we further expect sympatric evolution of the two subpopulations, due to separation of their gene pools. Aim of this study was to unravel the influence of demographic, ecological as well as environmental factors on the population differentiation as well as to uncover mechanisms which influence micro-evolutionary processes in Asexual *E. huxleyi* lineages. As method of choice we used a reduced representation sequencing approach called double digest Restriction site Associated DNA (ddRAD) sequencing, where only genomic regions between two different restriction enzyme cutting sites were sequenced. For 1 population from Peru and 3 populations from Chile basic population genetic parameters were assessed as well as genomic erosion and Gene Diversity in asexual Chile strains within and between populations. Based on basic population genetic parameters, moderate genetic differentiation between Peru and Chile populations was detected, most likely explained due to separation of the populations by different water masses and currents. However indications for the presence of sympatric evolution of asexual and sexual gene pools were given as well. The observed heterozygosity in Asexual *E. huxleyi* strains was not reduced although inbreeding indices were elevated, which could indicate a loss of heterozygosity. Therefore it remains unresolved which micro-evolutionary processes are dominating the genomic evolution of asexual *E. huxleyi* strains. The overall exceptionally high interspecific genetic diversity of E huxleyi might mitigate effects of Asexual reproduction and helps to sustain a high adaptive capacity.

IV. ZUSAMMENFASSUNG

Phytoplankton ist verantwortlich für rund 50 % der jährlichen, globalen Primärproduktion und *Emiliania huxleyi* ((Lohm) Hayard und Mohler) ist ein wichtiger Bestandteil des Phytoplanktons. *Emiliania huxleyi* Arterfolg kann einen direkten Einfluss auf die vertikalen Kohlenstoffflüsse im Ozean haben, da es sowohl zur biologischen- als auch zur Kohlenstoff – Gegenpumpe einen Beitrag leistet. Diverse Studien haben gezeigt, dass *E. huxleyi* für den Globalen Klimawandel anfällig ist. *Emiliania huxleyi* hat eine hohes Adaptationspotenzial und kann sich erfolgreich in hoch diversen wie auch stabilen von den Subtropen bis hin zu sub-polaren Ökosystemen vermehren. *Emiliania huxleyi* besitzt einen zweiphasigen Lebenszyklus mit phänotypisch unterschiedlichen haploiden und diploiden Zelltypen. Die beiden Lebensphasen des Lebenszyklus sind durch Syngamie, einer Form von sexueller Reproduktion, miteinander verbunden. Eine aktuelle Studie konnte zeigen, dass *E. huxleyi* Populationen aus stabilen Ökosystemen einen signifikant erhöhten Anteil an Stämmen haben, die die Fähigkeit verloren haben einen haploiden Zelltyp zu bilden. Dies wird mit dem Verlust von sexueller Vermehrung assoziiert, was weitreichende evolutionäre Einschränkungen für den Arterhalt und das Adaptationspotenzial der Art zur Folge haben kann. Asexuelle Stämme zeigen häufig einen Verlust von Heterozygotie, der wiederum dazu führt, dass letale Allele demaskiert werden und somit die Fitness der betreffenden Art. Als Gegenspieler hierzu ist ameiotische Rekombination zu verstehen, da diese Art von Rekombination hilft letale Allele aus dem Genom zu entfernen und somit ein hohes Adaptationspotenzial zu bewahren. Da asexuelle und sexuelle Populationen von *E. huxleyi* den gleichen geographischen Ursprung haben, erwarten wir eine sympatrische Evolution der beiden Subpopulationen durch Trennung der Genpools. Ziel dieser Studie war es, den Einfluss von demographischen, ökologischen wie auch Umweltfaktoren auf die Differenzierung von Population zu untersuchen, wie auch Mechanismen aufzudecken, die die genomische, micro-evolutionäre Entwicklung in asexuellen *E. huxleyi* Stämmen beeinflussen. Als Methode der Wahl haben wir uns für double „*digest Restriction site Associated DNA sequencing* (ddRAD)" entschieden. Bei ddRAD werden nur genomische Sequenzen zwischen zwei Restriktionsschnittstellen sequenziert. Populationsgenetische Parameter wurden für eine Population aus Peru sowie 3 Populationen aus Chile bestimmt. Basierend auf populationsgenetischen Analysen wurde eine moderate genetische Differenzierung zwischen Peru und Chile Populationen nachgewiesen. Ursächlich für eine Differenzierung der Populationen ist wahrscheinlich die Auftrennung der Populationen in verschiedene Wassermassen und Meeresströmungen. Es wurden jedoch auch Anzeichen für eine sympatrische Evolution von asexuellen und sexuellen Genpools gefunden. Trotz dass die beobachtete Heterozygotie in asexuellen *E. huxleyi* Stämmen nicht erniedrigt war,

wurden erhöhte Werte für den „*inbreeding coefficient*" gefunden, was eigentlich auf eine erniedrigte Heterozygotie hinweisen würde. Daher bleibt ungeklärt, welche micro-evolutionären Prozesse die genomische Entwicklung von asexuellen *E. huxleyi* Stämmen dominieren. Möglich wäre, dass die außergewöhnliche hohe intraspezifische Diversität von *E. huxleyi* mögliche negative Auswirkungen der asexuellen Fortpflanzung mildern kann und somit eine hohe Anpassungsfähigkeit erhalten kann.

1. INTRODUCTION

1.1. GLOBAL CHANGE

Within the last years, Global Change has received more and more attention in the scientific community. Researchers try to understand the impacts of anthropogenically driven Global Change on earth´s climate, biogeochemical cycles, communities as well as the fate of species under Global Change (Walther *et al.* 2002; Parmesan 2006; Rosenzweig *et al.* 2007; Walther 2010; Masson-Delmotte, V. *et al.* 2013; Pörtner *et al.* 2014). Throughout earth´s history, climatic conditions have changed by natural means leading to altered environmental conditions and naturally induced Global Change (Masson-Delmotte, V. *et al.* 2013; Pörtner *et al.* 2014). Today´s global environmental change was induced for instants by the release of an excess of anthropogenic carbon dioxide (CO_2) into the atmosphere, which diminished changes from geological time scales to only several decades or hundreds of years (Kump *et al.* 2009; Pörtner *et al.* 2014). Ongoing changes are expected to affect phenology and physiology of single species, which might have an impact on other adjacent trophic levels, leading to an altered spatial and temporal occurrence of species and communities (Walther 2010).
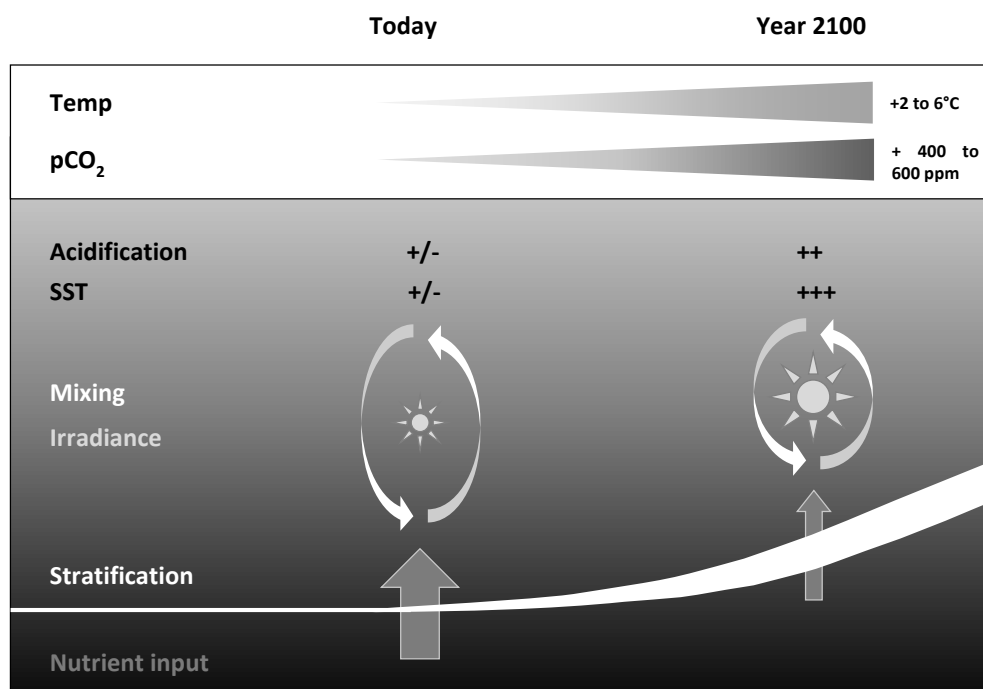
Figure 1: Changes induced by global change in the atmosphere (white area) and in the sunlight area of the open ocean (blue area) for Ocean Acidification, sea surface temperature (SST), mixing, irradiance, stratification and nutrient input from deeper layers until the year 2100.

In the course of changing climatic conditions sea surface temperatures (SST) in the mixed layer will rise, stratification will strengthen between ocean surface and deeper layers, and consequently nutrient input from nutrient rich deep water masses will decline due to lower frequency and intensity of upwelling events (Steinacher *et al.* 2010) (Figure 1). Further, the oceans will be subjected to the recently described phenomenon of Ocean Acidification (OA) (Figure 1) (Caldeira & Wickett 2003; Rost & Riebesell 2004). All these changes can have implications for a variety of physiological processes like growth in general or photosynthesis of phytoplankton (Orr *et al.* 2005; Halpern *et al.* 2008), and ultimately for the success of species or complete ecosystems in a changing ocean.

A global increase of SST will not only lead to decreased nutrient concentrations, but global ocean circulation models also predict a decrease of global phytoplankton productivity and carbon export to deep sea (Bopp *et al.* 2001; Steinacher *et al.* 2010) with rising SST. Phytoplankton species, as other ectothermic species, are adapted to a certain thermal window. Their growth rates will respond to changes in temperatures by increasing growth until the optimal temperature is attained. Beyond optimal temperatures growth will experience a unimodal decline or is negatively skewed above optimum temperatures (Eppley 1972; Kingsolver 2009). Therefore SST rise is especially challenging for species already living at their optimal growth temperatures. Model predictions unveiled a poleward shift of equatorial distribution boundaries, whereas polar range boundaries remain widely unaffected (Thomas *et al.* 2012). Tropical species are affected most by SST as their optima are already close to present mean temperatures, as those populations have not experience much temperature variations. Hence, a small increase of the SST might thrive in particular tropical species to extinction or lead to a poleward shift of species ranges in general. Possible mechanisms to mitigate effects of temperature rise are high intraspecific genetic diversity and thus a high adaptive potential of species (Thomas *et al.* 2012).

Although coastal areas are experiencing changes in the nutrient stoichiometry e.g. N:P ratio with its potential species composition shifts, large areas of the open ocean are limited in nutrients, especially in the macronutrients nitrogen and phosphorous. However, the development of algal blooms i.e. a high net primary production of many phytoplankton species is dependent on a sufficient supply with macronutrients (Smayda & Reynolds 2003; Falkowski, Paul G., Raven 2007; Hallegraeff 2010; Cavallo

et al. 2011; Pörtner *et al.* 2014). In the course of Global Change the supply with nutrients is subjected to changes and will most likely decrease as open-ocean upwelling events might decline or fail to appear completely on a regional basis (Steinacher *et al.* 2010). Despite ongoing research and modelling it is still elusive which influence global warming and the attendant stratification of the oceanic upper surface layer will have on the nutrient composition in the sunlit layer of the ocean. Predictions range from increased upwelling of nutrients to decreased upwelling with all their consequences for the environment (Lluch-Cota *et al.* 2014).

The ocean contributes crucially to the burial of atmospheric carbon and has taken up about 155 Gt carbon of the anthropogenic carbon released between 1750 and 2011. Therefore, the ocean is an important instrument in regulating the global atmospheric CO_2 concentration (Masson-Delmotte, V. *et al.* 2013; Pörtner *et al.* 2014). Past and future increase of atmospheric CO_2 led to dissolution of excess CO_2 in the ocean, which ultimately decreased oceanic pH values causing OA. Two carbon pumps are known to contribute to the global biogeochemical carbon cycle: the biological carbon pump and the carbon counter pump. Depending on the relative strength of these two pumps to each other (rain ratio) a net burial or outgassing of CO_2 from the ocean occurs (Rost & Riebesell 2004). Phytoplankton is an important part of the global biogeochemical carbon cycle as it contributes with up to 50 % to the global primary production. Therefore phytoplankton productivity is an important regulator of atmospheric CO_2 concentrations.

1.2. ADAPTATION TO GLOBAL CHANGE

Climate change will impact the marine realm in many ways. The success of existing species might get challenging and whole ecosystems are threatened unless species succeed to adapt to altered environmental conditions or will successfully migrate to other more suitable regions. As an ultimate consequence if species fail to adapt to changing conditions they will go extinct.

Fast adaptation processes at micro-evolutionary level can benefit from a large standing stock of genetic variation instead of having only spontaneous new mutations as the basis for adaptation in a population. Beneficious alleles are more likely to be present in a standing stock than from new mutations (Innan & Kim 2004). Further they are already pre-selected in previous generations and hence might be present in higher

frequencies (Rieseberg *et al.* 2003; Innan & Kim 2004). High genetic variation gives an evolutionary advantage in environments with strong selective pressures (increased CO_2 concentration, oligotrophic conditions or increased SST). This can lead to plastic changes in populations or sorting of advantageous genotypes in terms of adaptation to the given conditions (Yoshida et al. 2003; Lohbeck et al. 2012; Collins et al. 2014; Schlüter et al. 2014). Rapid adaptation substantial for species survival in a changing ocean (Hoffmann & Sgro 2011). Adaptive evolution can lead to altered coding DNA regions, directly resulting in a better adapted allele of a given protein, but on short-time scale adaptation processes might be facilitated on the level of epigenetic modifications (Jaenisch & Bird 2003; Jablonka & Raz 2009; Yaish *et al.* 2011). To some extent ecological and evolutionary relevant variation cannot be explained on the basis of genetic variation, but by gene expression changes due to selective methylation of different genomic regions. The regulation of protein expression on the basis of methylation is one part in the field of epigenetics. Selective processes acting on non-coding DNA regions are also important drivers of adaptive evolution. Andolfatto (2005) could show that large proportions of non-coding DNA have undergone positive selection were evolutionary constraint. Mutations in non-coding DNA as for example in promotors or cis-regulatory elements can lead to expression increase or decrease of the specific gene.

Furthermore also reproductive modes can have an impact on the adaptive capacity of species. Species reproducing exclusively by mitosis will tend to accumulate positive as well as negative mutations in a stepwise manner due to missing homologous recombination and purifying selection (Felsenstein 1974). Furthermore, as most mutations tend to be deleterious when homozygosity is inflated, mutation load (Lynch & Gabriel 1990; Lynch *et al.* 1990) of asexual lineages increases and could ultimately lead to extinction of certain lineages (Lynch & Gabriel 1990; Lynch *et al.* 1990; Xu *et al.* 2011). Another important factor in genome evolution of asexual lineages is the increase of hemizygosity due to nonallelic homologous recombination (unequal crossover) (Helleday 2003; Hastings *et al.* 2009) and therefore deletion of one allele of a diploid loci. Increase in hemizygosity, e.g. under relaxed selection pressures, can further lead to segmentational deletion of for e.g. complete genes or parts of genes (Muller 1964; Felsenstein 1974). In contrast to nonallelic homologous recombination, which leads to further unmasking of deleterious mutations, ameiotic recombination events can help

to overcome the accumulation of possible deleterious mutations, when present in sufficiently higher rates than mutation rates. In general, populations with sexual reproduction have an adaptive advantage in highly diverse environments compared to asexual lineages. Ameiotic recombination can serve as an important mechanism for asexual populations to sustain an adaptive capacity as high as in sexual populations, even in highly variable environments (Mandegar & Otto 2007).

1.3. COCCOLITHOPHORIDS AND THE EVOLUTION OF *EMILIANIA HUXLEYI*

Coccolithophores first appeared in the world´s ocean in the early Mesozoic and started to diversify in the late Mesozoic (Rost & Riebesell 2004; de Vargas *et al.* 2007; Kump *et al.* 2009) (Figure 2). During the Mesozoic, the ocean was highly acidic and atmospheric CO_2 levels were as high as 4000 ppm (Kump *et al.* 2009) (Figure 2). In the course of coccolithophorid evolution, the species has been subjected to highly variable environments on relatively short geological time scales, which underpins the high adaptive capacity of coccolithophores.

Figure 2: Haptophyte evolution and atmospheric CO_2 concentration for the past 500 million years. Haptophyte evolution redrawn from (de Vargas *et al.* 2007) and atmospheric CO_2 concentration redrawn from Boyten et al. (2004). Grey area on the graph for atmospheric CO_2 concentrations represents standard deviation

The coccolithophore *Emiliania huxleyi* ((Lohm.) Hay and Mohler), which arose about 290,000 years ago (Raffi *et al.* 2006), is an important part of the global phytoplankton community, being responsible for up to 50 % of the global calcite production. Coccolithophores in general contribute up to 10 – 20% to the regional total carbon fixation (Poulton *et al.* 2007). *Emiliania huxleyi* is not only fixing carbon via photosynthesis into organic carbon but also into calcite, making it a major contributor to both carbon pumps. Thus, *E. huxleyi´s* species success has a direct influence on the strength of the vertical carbon fluxes in the ocean and ultimately on the carbon uptake capacity of the ocean (Rost & Riebesell 2004). Since first appearance, *E. huxleyi* has colonized most ocean surface waters and is able to form dense blooms (10^3 - 10^5 cells

mL^{-1}) with nearly infinite numbers in a wide range of ecosystems (Paasche 2002). *Emiliania huxleyi* is found in oligotrophic and eutrophic ecosystems as well as in sub-polar or sub-tropical regions (Paasche 2002) which underpins *E. huxleyi´s* high physiological and genomic diversity (Kegel *et al.* 2013; Read *et al.* 2013) and its high adaptive capacity. Read *et al.* (2013) could show that only roughly 2/3 of the genome of strain CCMP1516 was shared species wide, reflecting a core genome which stores genes for important housekeeping functions. Whereas the highly variable parts of the genome are strain specific, explaining a possibility why *E. huxleyi* strains can inhabit so many different niches. Variable parts of the genome can further serve as an important prerequisite for fast adaptation to Global Change, as better adapted strains can get dominant via lineage sorting (e.g. Lohbeck *et al.* (2012); Schlüter *et al.* (2014)). Besides genomic diversity, the species has a high phenotypic diversity reflected by several different morphotypes (Type A, - A-overcalcified, -B, -C, -R, -O and *var.* corona) (Young & Westbroek 1991; Medlin *et al.* 1996; Young 2003; Hagino *et al.* 2011) differing in their coccolith appearance and expression of different functional traits in reaction to parameters influenced by Global Change (e.g. nutrient depletion and OA) (Houdan *et al.* 2005; Rokitta 2012; Rokitta & Rost 2012; Rokitta *et al.* 2014).

Emiliania. huxleyi possesses a biphasic life cycle divided into two physiologically distinct phases: non-flagellated, calcified diploid (2N) cells and flagellated, non-calcified haploid (1N) cells (Klaveness 1972; Green *et al.* 1996; von Dassow *et al.* 2009, 2014; Rokitta 2012). The two life cycle stages are supposed to be interconnected via syngamy which enables strains to reproduce sexually, although this has never been observed in *E. huxleyi*. It is discussed and partly demonstrated that species having a biphasic life cycle are capable to inhabit different niches (Hughes & Otto 1999; Houdan *et al.* 2005; Coelho *et al.* 2007; von Dassow *et al.* 2009; Rokitta *et al.* 2011, 2012, 2014) and that formation of haploid cells act as an escape mechanisms from biological pressures like viruses (Frada *et al.* 2008). Based on the current knowledge viruses are believed to act as an influencing variable, terminating *E. huxleyi* blooms (Brussaard *et al.* 1996; Wilson *et al.* 2002; Frada *et al.* 2008; Coolen 2011).

1.4. EFFECTS OF GLOBAL CHANGE ON *E. HUXLEYI*

The coccolithophore *E. huxleyi* is influenced by a variety of factors altered due to Global Change like surface temperature increase, OA and reduced nutrient concentrations itself but also by synergistic effects of all environmental changes (Riebesell *et al.* 2000;

Zondervan *et al.* 2001, 2002; Rost *et al.* 2003; Sciandra *et al.* 2003; Rost & Riebesell 2004; Engel *et al.* 2005; Langer *et al.* 2006, 2009, 2011; Liu & Pang 2010; Müller *et al.* 2010; Kaffes *et al.* 2010; Rokitta & Rost 2012; Kottmeier *et al.* 2014; Schlüter *et al.* 2014). Among other reactions to OA by far the most impacting reaction for biogeochemical cycles might be the reduction of particulate inorganic carbon (PIC) to particulate organic carbon (POC) ratio. In addition to the factor OA, the intensity of PIC production decrease is also dependent on other factors like e.g. light intensity (Rokitta & Rost 2012). A reduced calcite to biomass ratio (PIC:POC) as a reaction to Global Change might ultimately affect the rain ratio and also the amount of CO_2 buried by the carbon cycle.

Certain members of the species-complex of *E. huxleyi* are known to tolerate extremely low levels of the macronutrients N and particularly P while still being able to form blooms (Iglesias-Rodriguez *et al.* 2002; Tyrrell & Merico 2004; Loebl *et al.* 2010; Read *et al.* 2013). *Emiliania huxleyi* strains exposed to oligotrophic more stable environments tend to form less dense populations and bloom formation was rarely or never observed (e.g. Hagino & Okada 2004; Beaufort *et al.* 2008). Under nitrogen limitation several studies have shown that growth rate of *E. huxleyi* ceases, the calcite to organic carbon ration increases (Raven & Crawfurd 2012), photosynthetic performance decreases (Loebl *et al.* 2010) and photosynthetic exudation becomes increasingly important up to the point that it dominates over biomass production (Borchard *et al.* 2011; Van Oostende *et al.* 2013). Recently, von Dassow *et al.* (2014) showed that strains constantly subjected to stable environments (such as oligotrophic regions) have undergone some fundamental adaptation processes leading to genomic erosion of genes which play an important role in haploid cell formation. It was assumed that the release of biotic pressures, e.g. less competition for nutrients as well as a lower probability of virus infections in highly diverse but diluted populations triggered the loss of genes associated to haploid cell formation and ultimately to the loss of sex. By definition populations are "potentially inbreeding group{s} of individuals that belong to the same species and live within a restricted geographical area" (Freeland *et al.* 2011). Characteristic for sexually reproducing populations is a frequent exchange and recombination of genetic material. Therefore populations inhabiting the same geographic range are supposed to be genetically more similar than populations from distinct geographical ranges. However, the gene pool of a population divided into sub-

populations by changes in reproduction modes (asexual vs. sexual) will also be split. As both sub-populations inhabit the same geographic range without interbreeding with each other the two gene pools of asexual and sexual *E. huxleyi* strains are subjected to sympatric evolution. Planktonic organisms, such as *E. huxleyi*, have short generation times and one might expect fast differentiation between the gene pools. In contrast high intraspecific and intrapopulation genetic diversity might diminish effects of sympatric differentiation of gene pools. Still elusive remains, if genetic isolation due to altered reproduction modes might ultimately lead to formation of new species as it would be expected in the course of sympatric evolution over time.

The loss of sex implies major evolutionary changes for *E. huxleyi* populations as they cannot perform homologous recombination anymore. Von Dassow et al. (2014) saw first indications for major evolutionary consequences on the genome level as homologous chromosomes in asexual strains showed considerable amounts of hemizygosity. Most likely these differences are the results of long-term absence of meiosis (von Dassow *et al.* 2014). However, *E. huxleyi* has several prerequisites to adapt on short time scales to changing factors. It has the ability to form dense blooms in a wide range of ecosystems, their populations are of nearly infinite population size (Paasche 2002), furthermore, *E. huxleyi* strains have been shown to exhibit an exceptionally high genetic diversity, which predicts a potential infinite resource for adaptation (Kegel *et al.* 2013; Read *et al.* 2013). *Emiliania huxleyi* strains from lower latitudes, growing in more oligotrophic conditions or strains recently dispersed to more oligotrophic waters already proved to be able to undergo fast adaptation (von Dassow *et al.* 2014). To what extend sexual and asexual reproduction play a role in shaping adaptive capacity of this species is still uncertain (Collins *et al.* 2014). Besides reproduction also genetic exchange between (immigration/emigration) populations has a great impact on genetic variation. Although rates of dispersal for *E. huxleyi* are still unknown, strains from different locations show considerable variation and differentiation, indicating that local adaptation within a few generations is possible(Lohbeck *et al.* 2014; Schlüter *et al.* 2014).

With ongoing improvement of next generation sequencing methods studies on genomic and transcriptomic level of single and multiple strains gets more feasible. In order to gain a better understanding why and how of processes visible on physiological level are observable, studies on transcriptomic and genomic level are an important link

to a better understanding. Several studies were conducted to uncover up- and down regulation of putative genes which might be affected by ocean acidification and nutrient limitation on the different life cycle stages of *E. huxleyi*. Outcomes of OA studies varied greatly, which on the one hand can be due to treatment conditions but on the other hand underpins the great intraspecific diversity of *E. huxleyi* (MacKinder *et al.* 2011; Richier *et al.* 2011; Benner *et al.* 2013; Bach *et al.* 2013). A recent study on the transcriptomic response of *E. huxleyi* on nitrogen limitation unveiled that as Diatoms, *E. huxleyi* switches to an alternative metabolism mainly using the ornithine-urea cycle (Rokitta *et al.* 2014). Under nitrogen limitation switching to ornithine-urea cycle but also increase proline-oxidation might act as an control for cellular nitrogen budgets (Rokitta *et al.* 2014).

1.5. NEXT-GENERATION SEQUENCING AND ANALYTICAL CONSTRAINTS

Traditionally molecular ecological studies were done with RFLPs, AFLPs, microsatellite markers and SNPs in single genes to study population genetic processes or phylogeographic distribution of species. With emerging next-generation techniques, researchers got the ability to sequence complete genomes of species for reasonable prices and in a reasonable time. Although, sequencing of complete genomes of a high number of individuals, especially for species with a large and highly diverse genome, is still not feasible. To study adaptive selection or even speciation, large portions of the genome can now be analysed generating results with a high statistical power due to a high number of loci. Double Digest Restriction site Associated DNA sequencing (ddRAD) (Peterson *et al.* 2012) is a newly implemented technique which enables researchers to analyse multiple individuals at a time by simultaneously sampling tremendous amounts of genomic regions called loci. Genome wide random sampling of loci is done by analysing stretches of DNA between two restriction enzymes. The inferred loci can then be used to study intra- and interpopulation relationships, but also to identify genomic regions involved in local adaptation (microevolution) or even regions responsible for evolution of new species (Lewontin & Krakauer 1973; Luikart *et al.* 2003; Beaumont & Balding 2004; Nielsen 2005; Nielsen *et al.* 2007; Foll & Gaggiotti 2008; Strasburg *et al.* 2012; Catchen *et al.* 2013a; b). Due to the so called "reduced representation" sequencing (Altshuler *et al.* 2000) applied in ddRAD even organisms with a highly complex genome such as *E. huxleyi* can be sequenced and analysed in a cost and time effective manner.

1.6. OBJECTIVES AND HYPOTHESES

The marine species *E. huxleyi* has always been subjected to highly dynamic environments shaping the evolution and success of the species. However, still elusive are the mechanistics and process understanding why this species was so successful in the past and how the species will succeed under predicted changes. Therefore, we aimed to unravel the influence of demographic, ecological as well as environmental factor on the population differentiation of *E. huxleyi* especially provoked by e.g. nutrient availability. Further focus of the study was to compare the genomic evolution of asexual vs. sexual reproduction of *E. huxleyi* strains to gain a better understanding of the impact of different life history traits as well as on genome wide micro-evolutionary changes. To address the main research question of this study we used strains isolated from populations with different life history traits and populations experiencing natural variability in physiochemical parameters.

The *E. huxleyi* strains used in this study were subdivided into 4 distinct populations according to (1) geographic distance and (2) differential life strategies. Strains from highly dynamic regions (upwelling) tend to form dense populations and blooms on a regular basis. Contrasting, strains isolated offshore from more stable environments tend to form less dense populations although still being an important player of the prevailing plankton community. Previous studies unveiled the loss of sex of strains isolated from stable environments (von Dassow *et al.* 2014). As a response to altered life cycling, these strains have already experienced adaptation on genomic scale by erosion of genes with released selection pressure and an increased genomic heterogeneity (von Dassow *et al.* 2014). Thus strains from these populations are the ideal candidates to unravel micro-evolutionary processes on genomic scale and study the influence of demographic, ecological and environmental factors on population differentiation and evoltution.

Based on the underlying assumptions the main hypotheses of the study were:

(1) Population genetic differentiation is higher in geographically distinct populations than intrapopulation differentiation

(2) Changes in life-history traits, particularly the loss of sex, leads to detectable intrapopulation differentiation. We suppose to detect elevated inbreeding (F_{IS}) and fixation indices (F_{ST}) between asexually and sexually reproducing populations.

(3) Higher polymorphism in the genomes of Asexual strains, therefore we suppose to sample a higher proportion of polymorphic loci and detect a higher mean Gene Diversity in these populations.

(4) Genomic erosion of haploid and diploid specific genes is detectable in Asexual populations.

(5) Differences in prevailing environmental parameters lead to non-synonymous single nucleotide polymorphism (SNPs).

(6) With the use of methylation sensitive restriction enzymes for ddRAD library production, coding regions preferably subjected to methylation modifications (epigenetic signatures) can be identified.

2. MATERIALS AND METHODS

2.1. SAMPLE COLLECTION

Emiliania huxleyi ((Lohm.) Hay and Mohler) strains (were isolated during research cruises between 2011 and 2013 of the coast of Chile and Peru in the Southeast Pacific (Figure 3, Table S I) and further cultivated as clonal stock cultures in Roscoff Culture Collection (URL: http://roscoff-culture-collection.org/). All strains used in this study were derived from either the Roscoff Culture Collection or from Peter von Dassow (Instituto Milenio de Oceanografia, Chile).
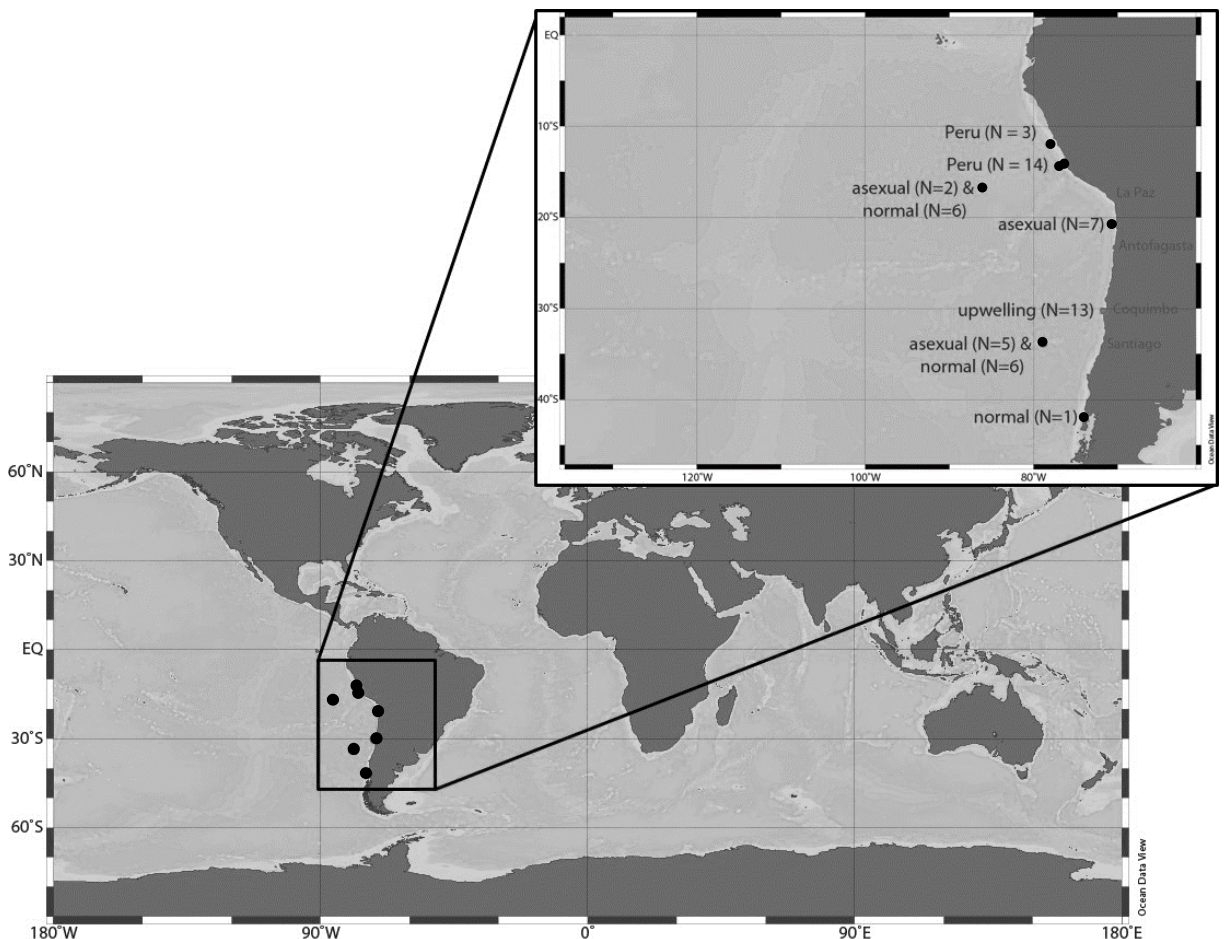
Figure 3 Sampling sites for Upwelling Chile, Upwelling Peru, Normal and Asexual population.

2.2. CULTURING CONDITIONS

Clonal *E. huxleyi* stock cultures were grown in half concentrated K-Media (Keller *et al.* 1987) (K/2) prepared with autoclaved (20 min, 121 °C, 2 bar) and 0.1 μm filtered Antarctic seawater. Media was enriched with vitamins and trace metals (Guillard & Ryther 1962). Light levels were adjusted to 70 – 90 μmol * cm^{-2} * s^{-1} under a 16:8 h

light:dark cycle and an ambient temperature of 15 °C. The stock cultures were transferred on a weekly basis to keep them in exponential growth phase.

Prior to inoculation of experiment batches, all cultures were grown in an antibiotic enriched K/2-Media (Table 1) for 4 days. After inoculation with antibiotics, cells transferred into antibiotics free K/2-Media until cell densities of about 175,000 – 325,000 cells * mL^{-1} were reached after 3 – 7 days. Cell density was assessed on harvesting day using a Multisizer III hemocytometer (Beckman-Coulter). The growth rate was calculated using the following formula (Guillard 1973), where µ is the growth rate (d^{-1}), t is the sampling day, and N_1 and N_2 are the abundances of culture at t^1 and t^2, respectively:

$$\mu = \frac{1}{t_2 - t_1} ln \frac{N_2}{N_1}$$

Table 1 Antibiotis stock solutions, working concentrations and required volumes of antibiotic stock solution to prepare 1L of 2x concentrated antiobitic enriched K/2-Media

Antibiotica	Stock concentration [mg/mL]	Working concentration in culture [µg/mL]	Volume stock solution for 2x concentrated antibiotics medium [mL/L]
Ampicillin	50	165.0	6.66
Gentamycin	10	33.3	6.66
Streptomycin	25	25.0	8.0
Chloramphenicol	34	34.0	0.06
Ciprofloxacin	10	10.0	2

2.3. DNA EXTRACTION

Cells were harvested at cell concentrations of about 175,000 – 325,000 cells * mL^{-1}. For harvesting cells were collected on 1.2 µm polycarbonate filters and subsequently washed with sterile K/2-Media (Keller *et al.* 1987) to reduce any potentially remaining bacterial background after the antibiotic treatment in the extracted DNA. Filters with cells were rolled and transferred to sterile reaction tubes containing ~ 100 µL glass beads (212 – 300 µm, Sigma). Cells were re-suspended in 600 µL lysis buffer (PL1) and lysed in the MagNA Lyser (Roche Diagnostics) 2 times for 20 seconds at 6500 rpm. To avoid possible RNA contaminations in the DNA extract, 15 µL RNAse A (0.01 mg/µL) were added and samples were incubated at 65 °C for 10 min. Before applying the lysate to the NucleoSpin® filters, crude lysate was centrifuged for 5 min at 11,000 x g in an Eppendorf 5810 R (Eppendorf, Hamburg Germany) to remove solid particles as for example calcareous coccoliths. All the other steps were performed as described in the manufacturer´s manual (Macherey-Nagel 2014a,Table S V) part 5.1 – Genomic DNA

from plant (page 18-21). Genomic DNA was eluted in two times 25 µL elution buffer resulting in a total volume of 50 µL.

2.4. DNA PURITY, CONCENTRATION AND INTEGRITY CHECK

DNA concentrations and -quality of each sample were determined with the NanoDrop® Spectrophotometer (nanoDrop products, Wilmington, United States of America) as well as with the Quant-iT™ PicoGreen® dsDNA Assay Kit (life technologies, Paisley, United Kingdom). For NanoDrop® measurements 1 µL of undiluted DNA was used. For PicoGreen® quantification, 1 µL sample was diluted 1: 1000 prior to the measurement. Integrity of isolated DNA was investigated using Agarose gel electrophoresis.

2.4.1. Quant-iT™ PicoGreen® dsDNA Assay Kit

Lambda DNA standard, provided with the kit, was diluted in three steps: 1:50 (dilution 1 = 2 µg/mL), then 1:40 (dilution 2 = 50 ng/mL) and finally 1:100 (dilution 3 = 500 pg/mL). With the diluted Lambda DNA a low-range standard curve was set up to cover concentrations between 25 ng/mL and 0 pg/mL (Table 2). The pure PicoGreen® solution was diluted 1:200 to obtain a PicoGreen® working solution. Standard curve sample were measured within the next 2 – 5 minutes. Genomic DNA samples were diluted in two steps: 1:1000 and 1:50. Subsequently 50 µL sample solution were mixed with 50 µL PicoGreen® working solution and the samples were measured within the next 2 to 5 minutes.

Table 2 Low-range standard curve pipetting scheme with volumes for 1x Tris-EDTA buffer, volumes of diluted Lambda DNA stock solution, Volume of Quant-iT™ PicoGreen® working solution and final Lambda DNA concentrations.

Volume of 1x TE [µL]	Dilution no.	Volume of diluted DNA Stock [µL]	Volume of diluted Quant-iT™ PicoGreen® Reagent [µL]	Final DNA Concentration in Quant-iT™ PicoGreen® Assay
0	2	100	100	25 ng/mL
90	2	10	100	2.5 ng/mL
99	2	1	100	250 pg/mL
90	3	10	100	25 pg/mL
100	-	0	100	0 pg/mL

2.4.2. Agarose gel electrophoresis

Furthermore integrity of DNA was verified on an 1 % Agarose gel, prepared with Tris-Acetate-EDTA (TAE) buffer (242 g Tris-HCl, 57.1 mL pure acetic acid, 100 mL 0.5 M EDTA (pH 8.0), filled up to 1L with Milli-Q® water). DNA was visualized using

Ethidiumbromide in a dilution of 1:10,000 (10 µL Ethidiumbromide in 100 mL Agarose gel). Electrophoresis was carried out at 90 V for 30 min to 1 h. The gel was subsequently viewed on a transilluminator (Herolab).

2.5. LIBRARY PREPARATION

Double digest RAD (ddRAD) libraries were prepared following a modified protocol by Peterson et al. (2012)(Figure 4). Aiming to apply two different approaches, three different restriction enzymes were used for constructing ddRAD libraries. On the one hand libraries were constructed using *Pst*I and *Msp*I (lifetechnologies). On the other hand to test the ability of ddRAD to detect potential epigenetically modified loci, the CpG methylation sensitive enzyme *Hpa*I was used instead of *Msp*I. An extended benchtop protocol for ddRAD library preparation can be found on the supplementary file CD-ROM (Table S V).

2.5.1. Restriction digestion

Approximately 400 ng genomic DNA (gDNA) were digested with 2.25 µL PstI and *Msp*I/*Hpa*II each in a total volume of 45 µL restriction digestion mix for 20 minutes at 37 °C. Restriction digested samples were cleaned with MinElute Reaction Cleanup Kit (Qiagen, Hilden) and eluted with 23 µL nuclease free water following the manufacturers manual (Qiagen 2008, Table S V).

2.5.2. Adapter ligation

The first sequencing adapter (P5) contained 24 different barcodes equivalent to the TruSeq® LT Adapters AD001 – AD027 in the P5 adapters (Table S II and Table S III). The second adapter, namely P7 adapter (Table S IV), contained a degenerated base region (DBR) (Schweyen *et al.* 2014) as well as the 4 different TruSeq® indices D701, D708, D710 and D710 (Illumina 2012). DBR regions were used to identify possible PCR duplicates and indices in the adapter were used to distinguish samples in a multiplex ddRAD approach. Further inserts of 1-3 nucleotides were used to create fragment length polymorphism and therefore different positioning of the restriction digestion site within the library. As the different nucleotides are detected in the Illumina NextSeq 500 by incorporation of single bases and creation of light signals (Figure 6C), the same nucleotide at equal positions in the whole library would lead to blinding of light signal detectors in the sequencer.
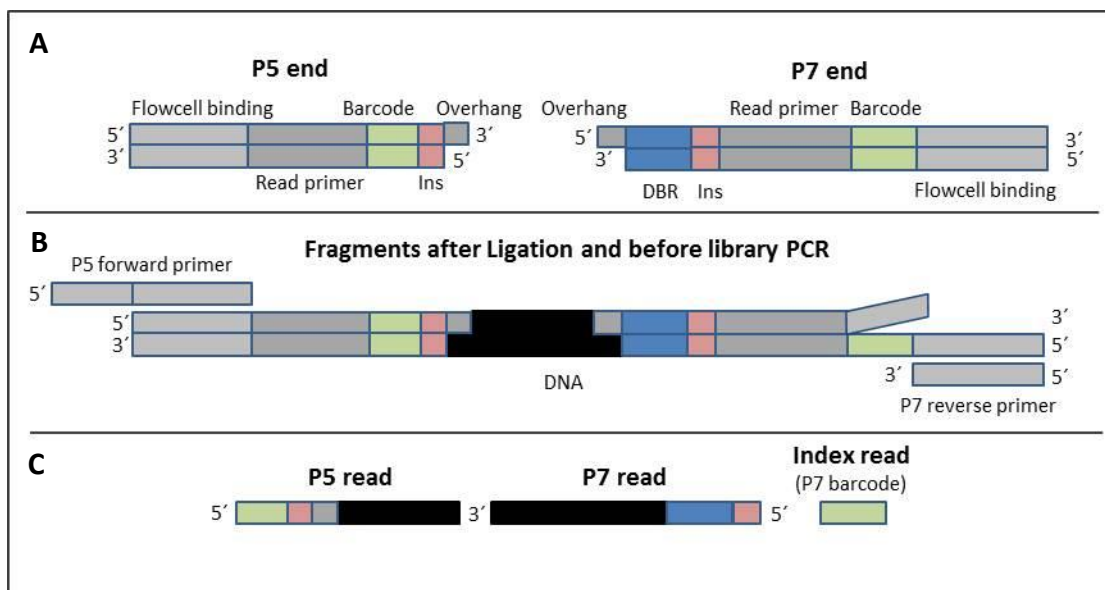
Figure 4: Library preparation overview with P5 and P7 adapters containing indices for multiplexing as well as DBR regions for detection of PCR duplicates and Inserts to create fragment length polymorphism (A), fragments after ligation and before library PCR (B) and the final raw reads after sequencing (C).Modified after (Schweyen *et al.* 2014)

P5 and P7 adapters were ligated to the restriction digested gDNA fragments with T4-Ligase (New England Biolabs Inc.) using 250 U/µLligase in a total volume of 40 µL ligation reaction mix at 16 °C for 2 hours. Subsequently ligation products were cleaned with the MinElute® ReactionCleanup Kit (Qiagen, Hilden) and eluted in 52 µL sterile, nuclease free water following the manufacturer´s manual (Qiagen 2008,Table S V).

2.5.3. Size selection with SPRIselect
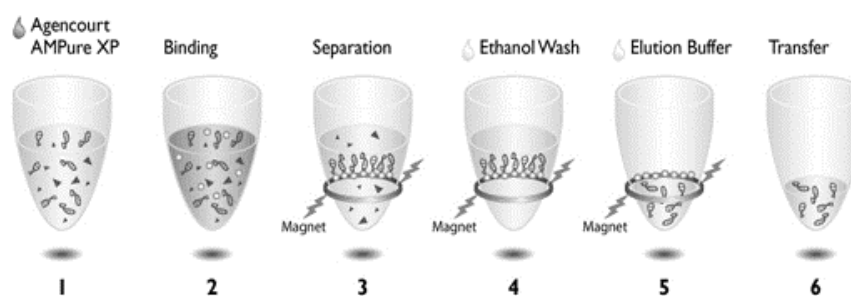
Figure 5: Principle of purification with magnetic beads (1) PCR reaction mix, (2) Binding of target DNA to magnetic beads, (3) Separation of magnetic beads with a magnetic field, (4) Washing with Ethanol, (5) Elution target DNA with water or 1x Tris-EDTA buffer, (6) Transfer of the eluted DNA sequences into a new reaction tube (Beckman Coulter 2009)

First size-selection step was carried out using SPRIselect (Beckman Coulter). With SPRIselect a desired part of the whole DNA can be isolated by selectively binding specific size classes to the SPRIselect beads, depending on the ratio of SPRIselect to sample solution. In this step 50 µL sample were added to 45 µL well mixed SPRIselect bead solution (bead to sample ratio of 0.9 (0.9x), mixed by pipetting up and down 10 times and incubated for 2 minutes. Samples were placed on a DynaMag™-96 Side (life

technologies). All the following incubation steps for binding or elution of DNA to and from the magnetic beads were carried out in 2 meter distance to the DynaMag™ magnetic plate. With the DynaMag™ magnetic plate beads can be separated from the solution to remove the supernatant and wash the samples (Figure 5). To allow magnetic beads to separate from solution and keep them separated from solution during washing steps, samples were incubated on the DynaMag™ for 2 minutes before removing the supernatant.

For washing 180 μL 85 % (v/v) ethanol were added to the beads and the cleared supernatant was removed and discarded. Residual ethanol, which would inhibit elution, was removed by incubating the open samples 5 to 10 minutes at room temperature. 28 μL sterile, nuclease-free water were added for elution and mixed by pipetting up and down 10 times. Then the cleared supernatant, now containing the size selected DNA, was pipetted into a new tube

2.5.4. PCR Amplification

Obtained fragments were amplified using Q5® Hot Start High-Fidelity DNA polymerase (New England BioLabs®*Inc.*). 0.02 U/μL Q5® polymerase, 10 μL 5x Q5®-Buffer, 1 μM P5 and P7 primers (Table 3), 0.02 μM dNTPs and 2.5 μL of fragmented DNA were used for PCR reaction and filled up to a total reaction volume of 50 μL with nuclease free water. For amplification the samples were denatured initially at 98 °C for 30 seconds, followed by 16 cycles of denaturation (10 sec, 98°C), annealing (30 sec, 65 °C) and elongation (30 sec, 72 °C) with a subsequent final elongation step of 5 min at 72 °C (Table 4).

Table 3 P5- and P7 Primer sequences used for amplification of ddRAD library fragments with ligated P5- and P7-adapters.

Primer	Sequence
P5 Primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
P7 Primer	CAAGCAGAAGACGGCATACGAGAT

Table 4 PCR program for amplification of ddRAD fragments after ligation of P5- and P7-adapters.

	Duration	Temperature	
Initial denaturation	30 sec	98 °C	
Denaturation	10 sec	98 °C	
Annealing	30 sec	65 °C	16 cycles
Elongation	30 sec	72 °C	
Final elongation	5 min	72 °C	
Hold 4°C			

2.5.5.   Amplicon purification with Agencourt® Ampure® XP

In order to clean amplicons after PCR amplification of ddRAD fragments, Agencourt® Ampure® XP (Agencourt Bioscience Cooperation 2009) bead solution was added to the PCR product in a ratio of 1.8 x and mixed carefully by pipetting up and down 10 times. The samples were incubated for 2 minutes to allow the DNA to bind to the magnetic beads. For washing of magnetic beads, 200 µL of freshly prepared 70 % (v/v) ethanol were applied to the beads and incubated for 30 seconds on the DynaMag™. Afterwards the cleared supernatant was discarded and the washing step repeated once more. Subsequently residual ethanol was removed by incubating the samples with an open lid for 5 to 10 minutes at room temperature. To elute DNA from the magnetic beads, 30 µL nuclease-free water were added and incubated for 1 minute. The cleared supernatant was then transferred into a new tube and stored at 4 °C until further use.

2.5.6.   Dual size selection with SPRIselect

For the first step of the dual size selection protocol 28 µL of purified PCR products were mixed with 15.4 µL SPRIselect bead solution (ratio 0.55 x). After mixing 10 times by pipetting up and down the tubes containing magnetic beads and sample mixture were placed on the DynaMag™ and incubated for 3 minutes. The cleared supernatant was pipetted into a new tube.

8.4 µL SPRIselect beads solution (ratio 0.19x) were added to the supernatant from the first step, mixed by pipetting 10 times and incubated for 2 minutes. Allow beads to settle for 2 minutes on the DynaMag™ plated and discard the cleared supernatant. Subsequently two washing steps with 180 µL 85 % (v/v) ethanol were performed and the clear ethanol was discarded. Residual ethanol was removed by incubating the open samples at room temperature for 5 to 10 minutes. 28 µL nuclease free water were added to the beads and beads were resuspended by pipetting up and down 10 times.

After incubation on the magnet for 1 minute, cleared DNA solution was pipetted into a new tube.

2.5.7.   Pooling

In order to provide equal sequencing success for all samples in the ddRAD sequencing library, samples had to be pooled with equal molarities. If molarities of single samples

is considerably higher or lower as the rest of the sequencing library, the cluster density will be biased towards or away from these samples and decreasing the amount of data produced for the sample. DNA concentrations were measured via Qbit ds DNA HS Kit following all recommendations form the manufacturer´s manual (life technologies 2015, Table S V). After calculation of sample molarities, equal amounts of DNA were pooled and cleaned with MinElute Reaction Cleanup Kit (Qiagen, Hilden) according to manufacturer´s (Qiagen 2008, Table S V) manual and eluted in nuclease free water.

2.5.8. Final size selection on LabChipXT

Final size selection of the whole library was carried out with the LabChipXT (Caliper Life Sciences, Massachusetts) following the manufacturer´s manual (Caliper Life Sciences 2011). Fragments in the size range of 308 to 462 basepair (bp) were retained for sequencing (Figure S I).

2.6. SEQUENCING

ddRAD libraries were sequenced on an Illumina NextSeq® 500 sequencer (Illumina) using NextSeq 500® Mid Output Kit. After ligation of sequencing adapters (Figure 4A and Figure 6A) and loading of equi-molar pooled samples on to the flow cell, clusters were generated by bridge amplification. Due to this large clusters of homologous sequences were generated to increase the intensity of light signal for a given cluster (Figure 6B). Libraries were sequenced paired end with 300 cycles total and 4 different, 8 bp long TruSeq® indices (D701, D708, D710 and D711) (Illumina 2012). To avoid sequencing bias towards single samples, it is important to pool samples with equal molarities.
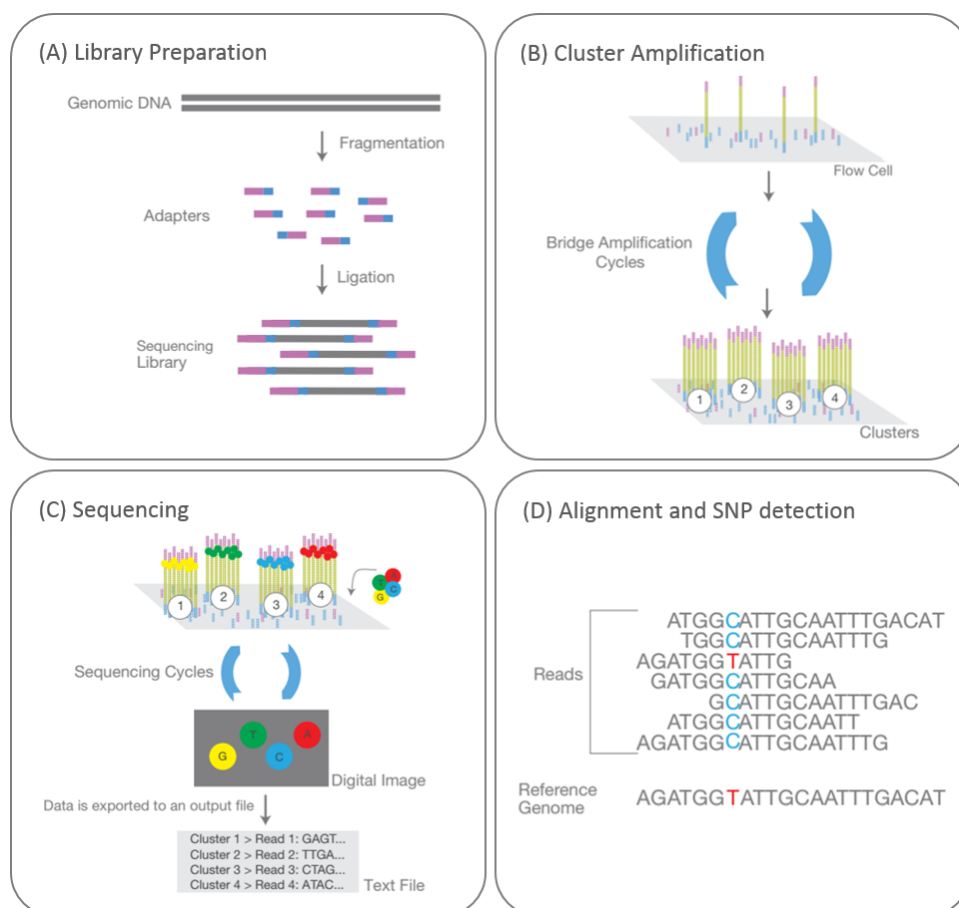
Figure 6: Overview of the sequencing process from (A) library production, (B) Cluster generation, (C) Sequencing to (D) first steps in analysis, namely alignment of the obtained sequences and identification of SNPs between the different Reads. Modified from (Illumina 2015b)

2.6.1. Calculation of library molarity

As library concentrations are given in nanogram per microliter [ng/µL] sample concentration has to be converted into molar concentrations in dependence to the final fragment size in the library using the following formula:

$$\frac{\left(concentration\ in\ {}^{ng}/_{\mu L}\right)}{\left({}^{660\ g}/_{mol} * average\ library\ size\right)} * 10^6 = library\ concentration\ in\ nM$$

In case of the prepared ddRAD library the calculation of molarity was as follows:

$$\frac{\left(21\ {}^{ng}/_{\mu L}\right)}{\left({}^{660\ g}/_{mol} * 390\right)} * 10^6 = 81.58\ nM$$

2.6.2. Denature and dilute ddRAD- and PhiX library

Before loading samples on the flowcell for sequencing, samples have to be denatured and diluted to a final loading concentration of 1.5 pM instead of 1.8 pM which was recommended from Illumina. Lower library concentrations were used to avoid an over clustereing of the flowcell. As sequencing control and for better sequencing quality, 15

% PhiX library were used as spike-in. By spiking PhiX DNA into the sequenced library blinding of the signal detection cameras can get avoided, if the same nucleotide is at the same position in all reads of the ddRAD library. For the subsequent dilution steps the ddRAD library was first diluted from 81.58 nM to 4 nM. All further dilutions of the ddRAD- and PhiX library were prepared as described in the manufacturer´s manual (Illumina 2015, Table S V). In the last step 195 µL of denatured and diluted PhiX control (1.5 pM) were combined with 1105 µL of denatured and diluted ddRAD library (1.5 pM) and subsequently loaded onto the reagent cartridge.

2.7. CYTOCHROME C OXIDASE SUBUNIT 3 PCR AND SEQUENCING

To infer informations about the phylogenetic relationship and population structuring of the analysed strains a Maximum Likelihood tree was calculated, using sequences of the *cytochrome c oxidase subunit 3* (*cox*3). Further emphasis of tree construction with *cox*3 was to identify possible sympatric evolution of asexual strains in populations of *E. huxleyi*. 10 ng of gDNA were used for PCR amplification with EG*cox3*F1/R1 primer(Bendif *et al.* 2014) with Phusion High-Fidelity DNA-Polymerase (Biozym, F-530S) following the standard protocol for one reaction in a total reaction volume of 25 µL (Table 5 and Table 6).

Table 5 PCR mix for cox3 *PCR following the standard protocol of the Phusion High-Fidelity DNA Polymerase (Biozym)*

Component	Stock concentration	Final concentration
5x Phusion Green HF Buffer	5x	1x
dNTPs	10 mM	200 µM
EG*cox3*F1 (forward Primer)	10 µM	0.5 µM
EG*cox3*R1 (reverse Primer)	10 µM	0.5 µM
Phusion Polymerase	2 U/µL	0.02 U/µL
DMSO	100 %	3 %
H_2O	-	Fill up to 25 µL

Table 6 PCR program for cox3 sequence amplification

	Duration	Temperature	
Initial denaturation	2 min	98 °C	
Denaturation	10 sec	98 °C	
Annealing	30 sec	55 °C	35 cycles
Elongation	30 sec	72 °C	
Final elongation	10 min	72 °C	
Hold 4°C			

Expected *cox3* fragment size of about 810 bp (Bendif *et al.* 2014) was checked on a 1% Agarose gel prepared with 1x TAE buffer. PCR products of all successful *cox3* amplifications were purified with the Macherey-Nagel NucleoSpin® Gel and PCR Clean-up kit following the manufacturer´s manual (Macherey-Nagel 2014b, Table S V).

Cleaned PCR products were eluted in 15 µL sterile, nuclease free water in a two-step elution summing up to a total volume of 15 µL.

1µL of the purified PCR product was used in the subsequent BigDye® Terminator v. 3.1 Cycle Sequencing (Applied Biosystems) reaction. Concentrations and volumes for sequencing reaction can be found in Table 7 and the program for the subsequent PCR amplification in Table 8.

Table 7 Sequencing reaction mix for BigDye® Terminator v 3.1 Cycle Sequencing kit.

Component	Stock concentration	Final concentration
BigDye Buffer	-	1.5
BigDye	-	0.3
Primer EGcox3F1 or EGcox3R1	1 µM	1
Purified PCR product	-	1
H$_2$O	-	6.25

Table 8 Cycler program for BigDye® Terminator v 3.1 Cycle Sequencing PCR amplification.

	Duration	Temperature	
Initial denaturation	1 min	96 °C	
Denaturation	10 sec	96 °C	
Annealing	5 sec	50 °C	25 cycles
Elongation	4 min	60 °C	
	hold	10 °C	

After PCR amplification with BigDye v 3.1 Cycle Sequencing Kit obtained sequences were purified using the Agencourt CleanSeq Dye-Terminator removal kit (Beckman Coulter, Brea CA) following the manufacturer´s manual (Beckman Coulter 2014, Table S V). The final sequencing step was carried out on the Applied Biosystems 3130xl capillary sequencer, using a 50 cm capillary.

Using CLC Main workbench v. 7 forward and reverse *cox*3 sequences were manually trimmed, reverse complemented and forward and reverse sequences were assembled forming a unique consensus sequence which was used further for phylogenetic analyses. In Mega6 (Tamura *et al.* 2013) obtained *cox*3 sequences were aligned with references sequences obtained from (Bendif *et al.* 2014). A phylogenetic analyses was performed using the Maximum Likelihood approach, settings were General Time Reversible model (Nei & Kumar 2000) with discrete Gamma distributions of 5, initial tree was inferred using the Neighbour-Joining method. Clade affiliation for strains unassigned to one of the four populations (CHC341, -349, -316, -347, 342, 343, 346, and -448) was done by El Mahdi Bendif (unpublished data) and used to identify clade

affiliation of the analysed strains. As outgroup strain RCC1320 (*Gephyrocapsa oceanica)* from the γ clade was used for tree construction.

2.8. PROCESSING OF DDRAD SEQUENCES/DATA

2.8.1. Cleaning the raw reads and de-multiplexing

Raw sequence reads obtained from sequencing on Illumina NextSeq® 500 sequencer were filtered for PCR duplicates with the perl script double_indxing (Rozenberg 2015). PCR duplicates were indicated by reads with identical DBR regions and further reads with errors in the index read of the P7 adapter were excluded from further analyses as well. Maximum 8 barcode shifts, 3 mismatches in DBR sequences as well as 1 mismatch in Inline-Barcode (P7) were set as tolerance level, when executing the double_indexing script.

Subsequently a bash script was executed, which runs Trimmomatic v 0.32 (Bolger *et al.* 2014), the perl script preprocess_ddradtags (Rozenberg 2014) as well as FastQC v 0.11.3 (Andrews 2010) after one another. The master script (i) searches for intact barcode sequences in the P5-read, (ii) identifies DBR sequences in the P7-read, (iii) discards possible PCR duplicates, (iv) removes the inserts, and (v) trims the resulting P5 and P7 reads to an equal length of 90 bp and (v) sorts reads into separate files, one for each strain.

2.8.2. Stacks analyses
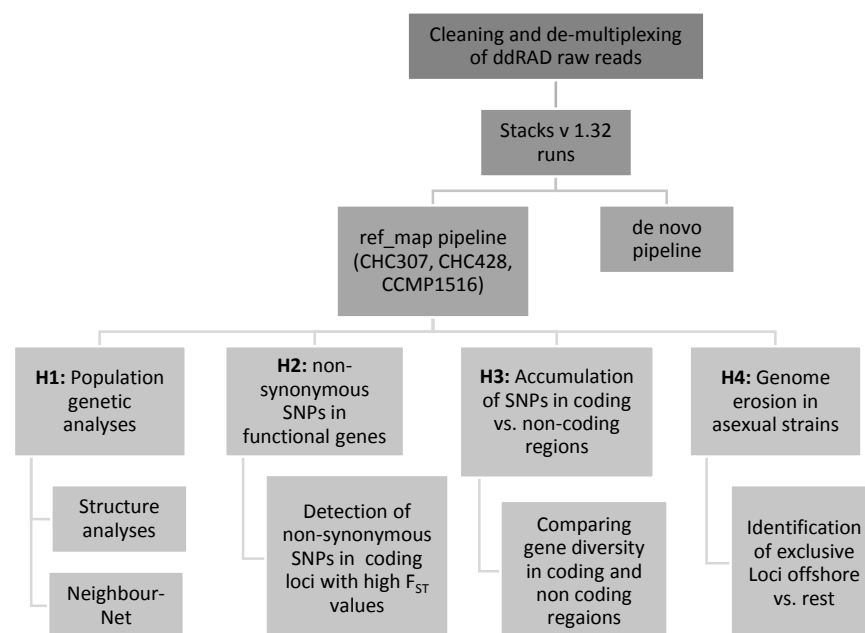
Figure 7 Schematic overview of Data analyses

To build ddRAD loci and perform basic population genomic analyses, filtered and sorted reads were analysed with the freely available software pipeline Stacks v. 1.32 (Catchen *et al.* 2011, 2013b).

In order to identify the most reasonable approach of loci formation, loci were built (i) de novo and (ii) with reference genome aligned reads (ref-map). Reads were aligned against three different reference genomes: CHC307, CHC428 (von Dassow et al., 2014; European Nucleotide Archive (ENA) study accession number PRJEB7726) and CCMP1516 (Read et al., 2013; downloaded from http://genomeportal.jgi.doe.gov). The first two genomes are obtained from strains of the normal (CHC307) and asexual (CHC428) population tested in this study. For all runs 4 mismatches between stacks (n = 4) were allowed and a minimum depth of coverage of 6 (m = 6) was required to build a loci. A loci considered for population genomic analyses had to be present in minimum 75 % of a single population and in minimum 3 out of 4 populations. For the populations part of the pipeline the following program parameters were used: kernel-smoothing (k) was turned on, F-statistics (inbreeding coefficient (F_{IS}) and fixation index (F_{ST})) were calculated as well as Structure- and fasta-files were generated as output. To test for statistical significance of F-statistics, bootstrap sampling option was turned on as well with 100 bootstrap replications.

Quality trimmed and de-multiplexed reads were mapped against two reference genomes: CHC307 (Normal population), CHC428 (Asexual population) and CCMP1516 in CLC v8. Mapping parameters were kept at default values (Mismatch costs = 2, Linear gap costs, Map randomly). Mapped reads were exported as SAM-files and used as input files for further reference map (ref-map) calculations in stacks.

2.8.3. Population genetic analyses

In order to calculate population genetic parameters, populations have to be assigned to different populations in advance. One major purpose of this study was to unravel micro-evolutionary processes induced by the loss of sex. In this context strains were divided into different populations based on their mating type (sexual vs. asexual) if strains were from the same geographical origin. Upwelling Chile and –Peru populations were separated from the Sexual and Asexual population on the base of different geographical origin (see Table S I).

Structure

To resolve geographical clustering between populations as well as subdivision by reproductive isolation within populations from the same geographical origin analyses with Structure (Pritchard *et al.* 2000; Falush *et al.* 2003, 2007; Hubisz *et al.* 2009) were carried out. Analyses were performed for both Stacks data sets generated after mapping data against the reference genome of CHC307 and CHC428 to resolve geographical clustering the no admixture model was used and alleles were treated independent among populations. For every analysis we used a burn-in length of 5000 and 5000 MCM repeats after burn-in. Further we tested 4 different predetermined cluster (K) ranging from 1 to 4 in 10 repeats per Stacks run.

Structure results were then evaluated visually as well as with the Structure harvester (Earl & vonHoldt 2012) by uploading a zip file containing all results to the Structure harvester page. ΔK values according to (Evanno *et al.* 2005) and L(K) values according to Pritchard et al., (2000) were calculated by the Structure harvester. When analysing structure results by visual inspection, indications for the "true" K-value are given, when in the bar plot diagrams individuals are strongly assigned to one population or another and not admixed (Pritchard 2010).

Neighbour-Net

Phylogenetic neighbour-net were inferred from haplotypes of Stacks loci (CHC307 and CHC428 ref-map Stacks run), by using the software SplitsTree (Huson & Bryant 2006). Input files were generated from Stacks SQL-databases and converted to Nexus files with the Perl script stacks2fasta (Macher *et al.* 2015). Data were exported as haplotypes with minimum one SNP per haplotype and only variable positions were written to the Nexus (*.nex) file. Subsequently the created Nexus-file were imported to SplitsTree (Huson & Bryant 2006) and a neighbour-net was constructed using the default values of the program.

2.8.4.   Identification of the influence of asexual reproduction on Gene Diversity

Following the assumption that asexual populations have a higher chance to accumulate mutations, Gene diversity in asexual populations is supposed to be higher compared to the other tested populations. Aiming to analyze this assumption, Gene diversities for every loci were extracted from the "batch_x.hapstats.tsv" file created by Stacks. Coding and non-coding loci were identified by mapping Stacks loci against the combined EST

sequences of CCMP1216/1217 (von Dassow et al., 2009; ENA study accession number ERP008543) and ESTs from CCMP1516 (Read et al., 2013; downloaded from http://genomeportal.jgi.doe.gov) (further called GJ-transcriptome). All subsequent calculations were performed in R-Studio (R Core Team 2015) using the packages plyr (Wickham 2011a) and data.table (R Core Team 2015). The corresponding R-Script called "Gene_diversity.R" can be found on the supplementary CD-ROM (Table S V).

The data were plotted using the package beanplot (Kampstra 2008) and for further statistical analysis, only statistically significant loci with a p-value < 0.05 were included. To test for statistical significant difference between mean Gene Diversities in the different populations a Kruskal-Wallis rank sum test and a subsequent post-hoc test according to Nemenyi for pairwise multiple comparison of ranked data was applied in R-Studio using the stats (R Core Team 2015) and PCMCR-package (Sachs 1997; Demšar 2006).

2.8.5. SNP analysis
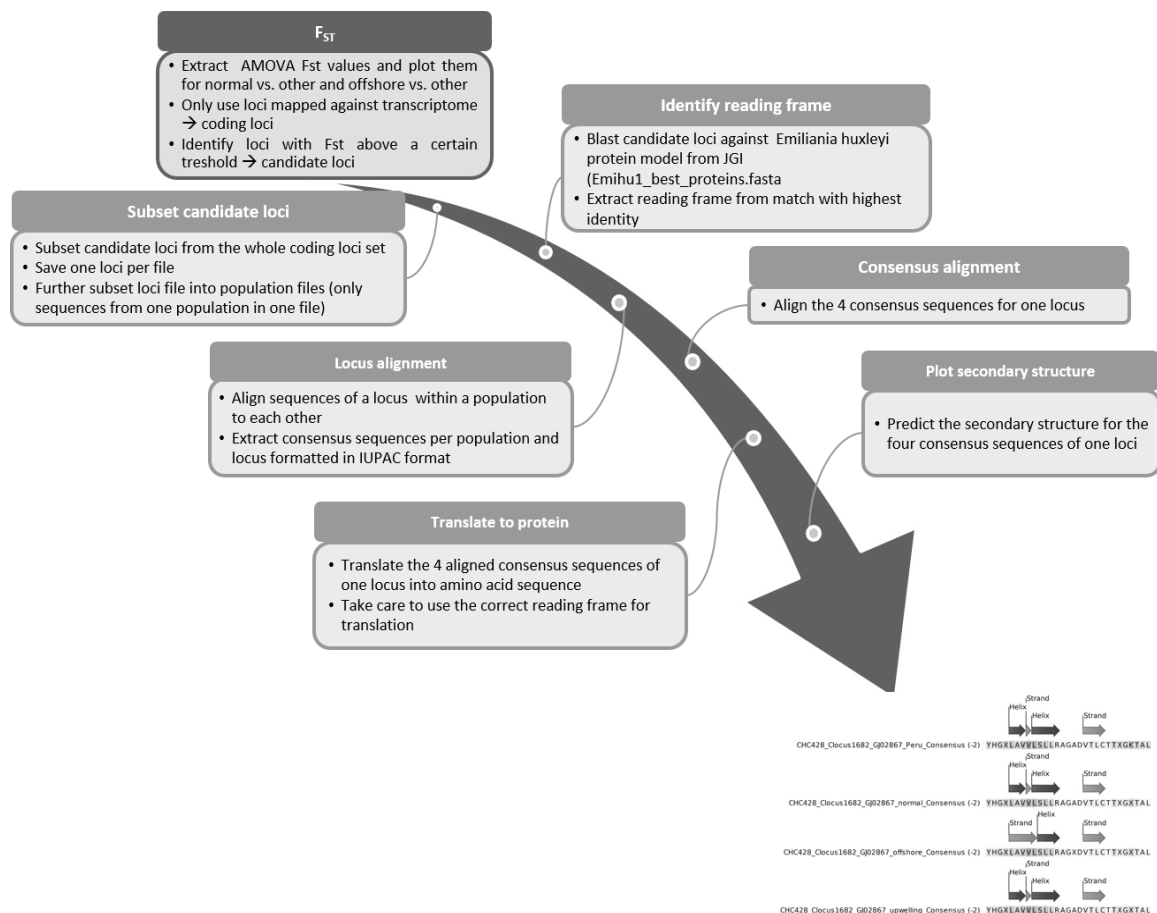
Figure 8: Schematic overview oft he different steps for the analysis of synonymous and non-synonymous SNPs in ddRAD loci of Emiliania huxleyi*.*

In order to identify synonymous and non-synonymous SNPs in the loci dataset, the first step was to identify possible evolutionary interesting loci experiencing selection pressure. For this purpose AMOVA F_{ST} values from Stack were extracted from the F_{ST} pairwise population comparison files. For this analysis it was important that loci were preferably present in as many populations as possible. Therefore FST calculation in Stacks were carried out with a dataset where loci had to be present in at least 75 % of one population and minimum in 3 out of 4 populations. Loci matching the following critera were then identified as possibly interesting and subsequently kept for analyses: F_{ST} p-value ≤ 0.05, F_{ST} values ≥ 0.2, only for Normal against Upwelling Chile ≥ 0.15.

Of this subset, coding loci were identified by filtering out only those loci, which could be mapped against the CCMP1516 transcriptome (Emihu1_best_transcripts, downloaded from http://genomeportal.jgi.doe.gov/). In CLC Genomics workbench 8.0 (http://www.clcbio.com) loci sequences for every strain was subsetted, separated into populations and consensus sequence (after IUPAC nomenclature) of every population was saved. In the next step, consensus sequences per locus were aligned and translated into amino acid sequences. The appropriate reading frame was identified by blasting loci sequences against the best model proteins identified in CCMP1516 (Emihu1_best_proteins, from JGI webpage). Further the secondary structure for every loci and consensus sequence was predicted in CLC Genomics Workbench 8.0 (http://www.clcbio.com).

All other analyses were carried out using R-Studio (R Core Team 2015) and the packages plyr (Wickham 2011a), data.table (R Core Team 2015) and ggplot2 (Wickham 2009).

2.8.6. Identification of genomic erosion

The number of exclusive and shared loci within and between populations was visualized using the Venn diagram technique. Further exclusive and shared loci were plotted per KOG class to gain an advanced understanding of genomic erosion especially in the Asexual population. As completeness of the data set was more important here, for Venn diagram construction no distinction between coding- and non-coding was done. Loci in this dataset had to be present in at least 75 % of one population and in minimum 1 out of 4 populations. All plots and calculations were produced in R-Studio using the packages plyr (Wickham 2011b), ggplot2 (Wickham 2009) and data.table (R

Core Team 2015). A detailed R-Script called "exclusieLoci.R" can be found on the supplementary CD-ROM (Table S V).

3. RESULTS

3.1. DDRAD LIBRARY PREPARATION

In order to obtain a "reduced representation sequencing" library of the 57 different *E. huxleyi* strains, genomic DNA was digested with two different restriction enzymes (*Eco*RI and *Msp*I/*Hpa*II) according to the principle of the double digest Restriction site Associated DNA sequencing (ddRAD) (Peterson *et al.* 2012). Size distribution of obtained fragments ranged from approximately 100 bp up to 10,000 bp. Most fragmets were expected between 100 and 1000 bp. After ligation of sequencing adapters, fragments were amplified with adapter specific primers to obtain enough starting material for sequencing. Verification of successful amplification of the ddRAD library fragments was done by agarose gel electrophoresis and revealed successful amplification for all samples (Figure 1 A-D). Main size range of the obtained fragments was as expected between 200 and 700 bp and fitted into the expected size ranges.

Figure 9: PCR amplicons of ddRAD samples separated on agarose gel (1% (w/v) agarose in 1x Tris-Acetate-EDTA buffer, Ladder O´Gene ruler 100 bp (thermofisher)). Amplification was successful when DNA fragments were visible between 200 and 700 bp.

3.2. SEQUENCING OUTPUT ILLUMINA NEXTSEQ 500

ddRAD libraries were sequenced with an Illumina NextSeq 500 sequencer. With next generation sequencing methods, such as Illumina NextSeq 500, massive amounts of sequences can be sequenced in parallel on a special flow cell. With the use of special index sequences in the sequence of interest even different individuals and populations can be analysed in parallel on one flow cell/sequence run. Sequencing of the ddRAD library containing 57 individuals from four populations on Illumina NextSeq 500 was successful and revealed a total number of 84,4 Mio. reads. Base calling accuracy, which

is measured as the Phred quality score (Q-score), is an important parameter to assess the accuracy of the sequencing process. The Q-score can be divided into different categories with increasing accuracy. A Q-score of 10 means, that 1 out of 10 bases was probably incorrect leading to a base call accuracy of 90 %, a Q-score of 20 indicates a base calling accuracy of 99 % and a Q-score of 30 a base calling accuracy of 99.9 %. In the present ddRAD library more than 90 % of the obtained reads had a Q-score of 30 or higher (Figure 10A). Mean cluster density (clusters/mm²) lane was in the range from 136,000 to 142,000 reads per square millimetre (Figure 10B) with lowest cluster densities in lane 2.
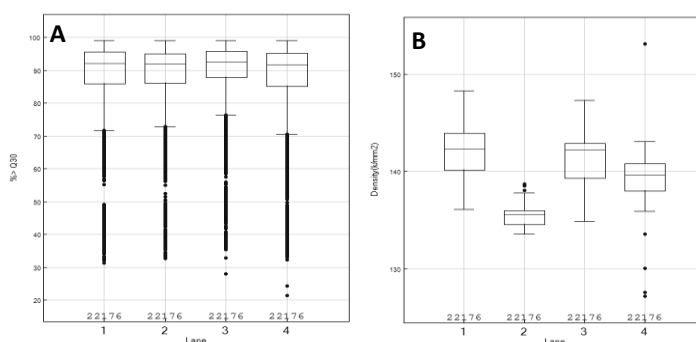
Figure 10: Percentage of data with a Q30 score or higher per lane (A) and cluster density in clusters per mm² per lane.

After first quality filtering steps, reads with incorrect or missing degenerated base region (DBRs) as well as sequences without indices were removed from the dataset revealing a similar range of raw reads between 16.49 and 18.8 Mio. reads per index.

Table 9: Number of reads before (total) and after (Correct) filtering for incorrect or missing DBR regions as well as for reads with no index splitted by the four indices 1, 8, 10 and 11.

Index	Total	Correct	Incorrect	No DBR	No index
ATCACG	19288776	16496641	15279	2683931	92925
ACTTGA	19802917	16633487	2059	2837490	329881
TAGCTT	22550143	18823621	4264	3299754	422504
GGCTAC	21769537	16551283	5365	2770582	2442307

3.3. BASIC SUMMARY STATISTICS

After sequencing of the ddRAD library several quality filtering and trimming steps are needed to start with further analyses in Stacks or other programs. On average 539,942 ± 42,293 high quality reads per population could be retained after quality filtering steps of the raw sequence reads. 83 % ± 6 % of the raw sequence reads could be aligned against the reference genome CHC307, 77 % ± 4 % against CHC428 and 80 % ± 4 %d against the genome of strain CCMP1516 (Table 10). On average 20.75 % (CHC307), 20.25 % (CHC428) and 13.75 % (CCMP1516) of the total number of loci

could be retained after running the Stacks pipeline, yielding about 349,369 ± 21,263 loci for CHC307, 321,276 ± 20,996 for CHC428 and 285,388 ± 26,144 for CCMP1516 reference map run.

Table 10: Basic sequencing output with Number of sequenced individuals (N), read length after trimming, Number of high quality read counts after quality filtering steps and percentages of reads aligned to reference genomes in table A. Total number of loci, number of retained loci and percentage of retained loci after Stacks run for the different reference map runs in table B.

A	N	Read length	High qual read count	% aligned reads CHC307	% aligned reads CHC428	% aligned reads CCMP1516
Upwelling Chile	14	90	561,881	92.14	84.14	87.61
Normal	13	90	545,402	84.05	74.41	77.13
Asexual	14	90	582,226	80.88	76.85	79.89
Upwelling Peru	17	90	470,260	76.91	73.56	76.81

B	Total # of Loci CHC307	Retained # of Loci CHC307	Percentage of retained loci after Stacks run CHC307	Total # of Loci CHC428	Retained # of Loci CHC428	Percentage of retained loci after Stacks run CHC428	Total # of Loci CCMP1516	Retained # of Loci CCMP1516	Percentage of retained loci after Stacks run CCMP1516
Upwelling Chile	1,545,978	369,172	24%	1,422,688	334,264	23%	1,901,589	304,309	16%
Normal	1,558,093	341,005	22%	1,391,667	297,269	21%	1,860,010	260,787	14%
Asexual	1,837,743	368,957	20%	1,756,988	348,666	20%	2,298,010	317,848	14%
Upwelling Peru	1,909,203	318,343	17%	1,842,274	304,904	17%	2,388,824	258,608	11%

To test whether sampling of loci was equally between the different populations, coding loci were identified and annotated. The number of loci in the different populations was equally in the different populations in all classes of eukaryotic clusters of orthologous groups (KOG classes). Most abundant KOG classes in the dataset were "Signal and transduction mechanism" and "Posttranslation modification, protein turnover, chaperons", although frequency of single loci was rather low (Figure 11). KOG classes with highest observed frequency of loci's were "Translation, ribosomal structure and biogenesis" and "Transcription" (Figure 11).
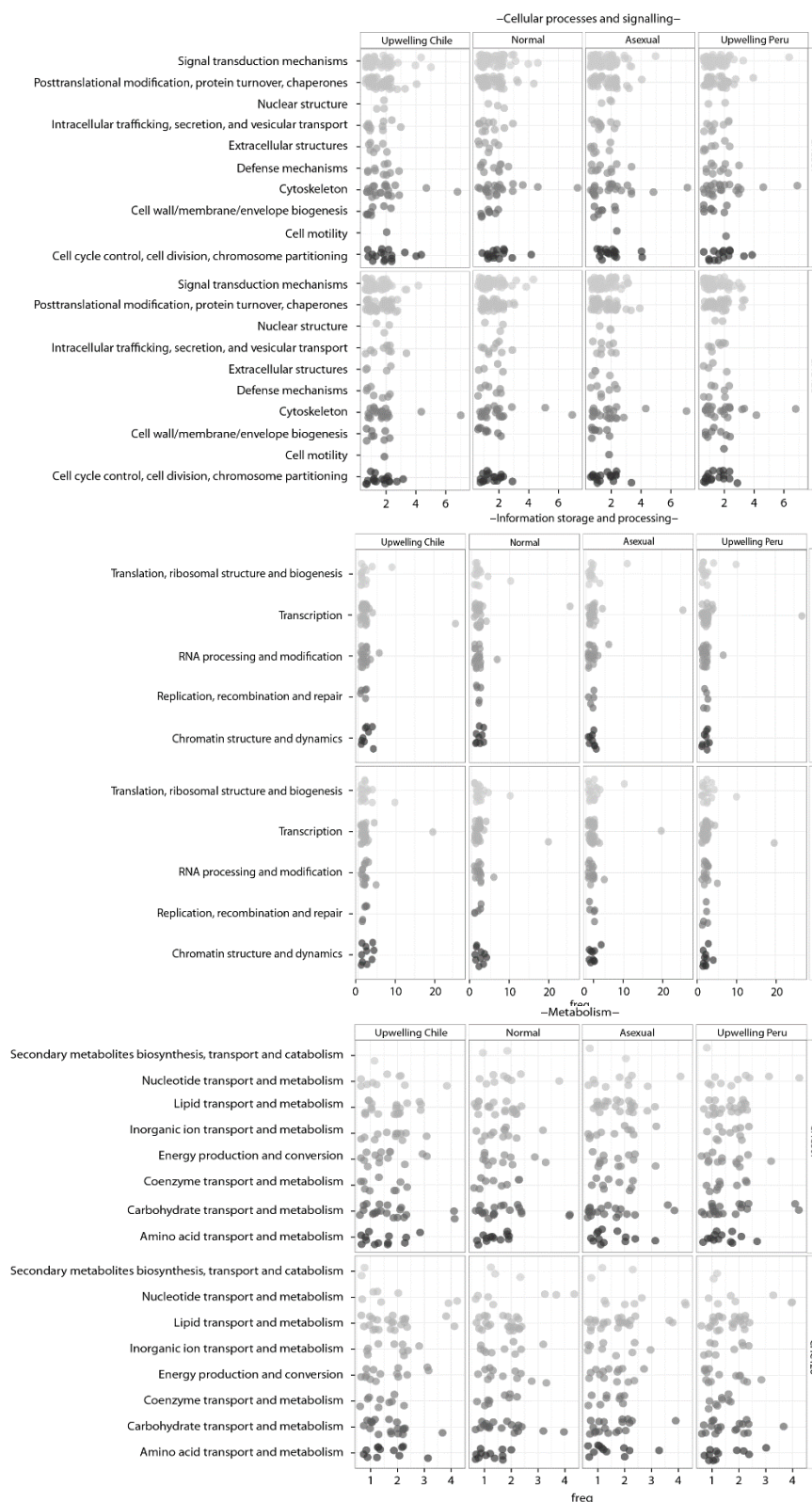
Figure 11: Frequency of loci subdivided into classes of clusters of orthologous groups (KOG classes) for both reference map Stacks runs CHC307 and CHC428 with (A) showing KOG classes for Cellular processes and signalling, (B) showing KOG classes for Information storage and processing and (C) Metabolism.

3.4. POPULATION GENETICS

Assessment of population genetic parameters can help to understand genetic differentiation of populations and processes which lead to genetic differentiation. In most natural population maintenance of genetic diversity is important to sustain a healthy and viable population. Processes like bottlenecks and reproductive isolation lead to substantial changes in the genepool. They have a great impact on the overall population wide diversity, but also on the rate of genetic exchange between populations. Population genetic parameters like observed heterozygosity (Het_{obs}), nucleotide diversity (π) but also F-statistics like the inbreeding coefficient (F_{IS}) or the fixation index (F_{ST}) can help to understand the amount and direction of genetic exchange within and between populations. For the present study strains were split into four different populations according to reproduction modes (sexual vs. asexual population), geographic origin (Peru vs. Chile) and environmental parameters (upwelling vs. offshore) (see Table S I).

When comparing the different basic population genetic parameters of the different Stacks runs, highest values for all parameters were obtained for CHC307 and CHC428 Stacks runs, except for F_{IS} values (Table 12). Parameters between CHC307 and CHC428 Stacks runs were comparable and did no differ substantially, although no statistical significance could be tested. Lowest values for all parameters were obtained when generating loci in Stacks *de novo* (Table 12). As basic population genetic parameters for CHC307 and CHC428 Stacks runs resolved the highest diversity parameters (Het_{obs}, Variant sites, % Poly and) and included the most individuals per population all further analyses were carried out only with loci from these Stacks runs.

Upwelling Peru population had the highest number of analysed individuals and number of private alleles in variant as well as in all positions. Lowest number of private alleles as well as percentage of polymorphic sites for variant as well as for all positions was observed in the Normal population (Table 12). Observed Heterozygosity (Het_{Obs}), the average frequency of the major allele (P) and the average nucleotide diversity varied only slightly between the populations, although test for statistical significance were not possible. However, differences between populations were detected for the average F_{IS} in variant positions (Table 12). F_{IS} values in variant positions were highest in Asexual population, lowest F_{IS} values were observed in the normal population (Table 12). Also the number of private alleles showed differences between the populations with the

highest number in Upwelling Peru population. Within the other geographical group of Upwelling Chile, Normal and Asexual, Asexual population had the highest number of private alleles, whereas the Normal population had the lowest (Table 12).

Mean pairwise F_{ST} values showed only moderate to low values indicating low population genetic differentiation. The Upwelling Peru population had the highest pairwise F_{ST} value. Although Normal and Asexual population are in most places geographically from the same origin, also low levels of population differentiation were detected (F_{ST} =0.031 (Table 11). Population differentiation between Normal and Asexual populations was even higher than between Normal/Asexual and Upwelling Chile population (0.023– 0.027) (Table 11).

Table 11 Mean pairwise F_{ST} values calculated after Nielsen et al. (2009) and (Weir & Cockerham 1984). Values for the CHC307 reference map run are on the top right and values for the CHC428 reference map run are on the bottom left

	Upwelling Chile	Normal	Asexual	Upwelling Peru
Upwelling Chile	-	0.0235949	0.0269056	0.0339724
Normal	0.0231851	-	0.030709	0.0422593
Asexual	0.0271388	0.031201	-	0.0274806
Upwelling Peru	0.0335927	0.0427716	0.0276066	-

It has been shown that single-gene based phylogenetics of *cox*3 marker is able to illustrate the three different morphotype groups within *E. huxleyi* (Bendif *et al.* 2014; von Dassow *et al.* 2014). It was aimed to compare this approach with a multi locus phylogenetic analyses (ddRAD) and therefore the generated *cox*3 sequences were aligned and a Maximum Likelihood tree calculated. The inferred phylogeny from *cox*3 sequences revealed a subdivision into two groups. Subdivision into the β-clade was highly supported with bootstrap values above 96%, however, the grouping into α-clade was not supported as bootstrap values were below 50% (Figure 12). Strain RCC1320, which was previously shown to belong to the γ-clade, was correctly sorted into a separate branch when inferring the phylogeny, although no bootstrap values were available here (Figure 12).

Table 12: Summary of basic genetic statistics for all populations and reference Stacks runs with CHC307, CHC428 and CCMP1516 as reference genome as well as denovo Stacks run. Parameters are calculated seperately for all polymorphic nucleotide positions ("Variant positions") as well as for "All positions", regardless they are variable or fixed. The statistics include Number of strains genotyped per population (N), the number of sites unique per population (Private), the number of variant and all nucleotide sites for Stacks runs (Sites), the percentage of polymorphic loci for variant and all sites (% Poly), the average frequency of the major allele (P), the average observed heterozygosity per population (Het_{obs}), the average nucleotide diversity (π) and the average inbreeding coefficient calculated after Wright (1965) (F_{IS}).

Variant positions	N				Private				Variant sites				% Poly			
# Pop ID	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo
Upwelling Chile	12.0	12.1	11.9	11.9	438	432	167	100	9317	9711	3777	2208	45.92	44.79	40.85	34.38
Normal	12.0	12.0	11.8	11.7	201	197	76	40	9317	9711	3777	2208	36.71	35.34	32.51	23.96
Asexual	13.0	13.0	12.8	12.9	1023	1074	462	331	9317	9711	3777	2208	62.09	62.34	61.72	59.92
Upwelling Peru	15.3	15.4	14.1	14.0	3237	3403	1399	943	9317	9711	3777	2208	50.50	51.30	47.21	42.89

	P				Het_{obs}				π				F_{IS}			
# Pop ID	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo
Upwelling Chile	0.92	0.93	0.94	0.96	0.14	0.13	0.10	0.07	0.11	0.11	0.09	0.07	-0.051	-0.046	0.001	0.012
Normal	0.93	0.93	0.94	0.97	0.13	0.12	0.09	0.06	0.10	0.09	0.08	0.05	-0.069	-0.062	-0.021	-0.003
Asexual	0.92	0.92	0.93	0.95	0.14	0.14	0.10	0.09	0.12	0.12	0.11	0.09	-0.026	-0.026	0.042	0.049
Upwelling Peru	0.92	0.93	0.94	0.96	0.13	0.13	0.10	0.07	0.11	0.11	0.10	0.07	-0.046	-0.044	0.015	0.030

All positions	N				Private				All Sites				% Poly			
# Pop ID	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo
Upwelling Chile	12.2	12.2	12.0	11.9	438	432	167	100	330340	342644	192877	181611	1.3	1.3	0.8	0.4
Normal	12.0	12.0	11.9	11.3	201	197	76	40	327286	339712	190869	257751	1.0	1.0	0.6	0.2
Asexual	13.2	13.1	13.0	12.6	1023	1074	462	331	329864	341641	191845	261531	1.8	1.8	1.2	0.5
Upwelling Peru	15.5	15.5	14.3	14.6	3237	3403	1399	943	315895	326781	181749	248751	1.5	1.5	1.0	0.4

	P				Het_{obs}				π				F_{IS}			
# Pop ID	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo	CHC428	CHC307	CCMP1516	deNovo
Upwelling Chile	0.998	0.998	0.999	1.000	0.004	0.004	0.002	0.001	0.003	0.003	0.002	0.001	-0.001	-0.001	0.000	0.000
Normal	0.998	0.998	0.999	1.000	0.004	0.004	0.002	0.001	0.003	0.003	0.002	0.000	-0.002	-0.002	0.000	0.000
Asexual	0.998	0.998	0.999	1.000	0.004	0.004	0.002	0.001	0.003	0.003	0.002	0.001	-0.001	-0.001	0.001	0.000
Upwelling Peru	0.998	0.998	0.999	1.000	0.004	0.004	0.002	0.001	0.003	0.003	0.002	0.001	-0.001	-0.001	0.000	0.000

F_{ST} and F_{IS} measures do need *a priori* information of the population structure to be able to calculate these values. Errors in population assignment of individuals or groups can result in biased population genetic parameters. However, defining boundaries of populations might be challenging in some cases additional analysis with non-*a priori* methods might be useful. A popular method used to infer population structure without *a priori* information is the Bayesian approach implemented in the program STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003, 2007; Hubisz *et al.* 2009). The optimal population subdivision inferred with Structure was with K=2 inferred from visual inspection as well as by analysis with Structure harvester (Earl & vonHoldt 2012) (Figure 13). Analyses with K=3 clusters already revealed cluster admixture with more than two clusters per strain indicating, that K=3 clusters were already too many clusters for the analysed populations (Figure 14). Two alternative results were produced when running structure with K=2 predefined clusters. One output clearly assigned strains to one cluster or the other in most cases, whereas the other output assigned most strains admixed to both clusters (Figure 13). Taking previous knowledge of population origin into account output A for K=2 clusters in both Stacks datasets is more reasonable (Figure 13A).

As comparative analysis to single-locus based phylogenetic analysis Neighbour-net networks were inferred from all loci of the CHC307 and CHC428 reference map dataset. Obtained neighbour-net networks showed no population specific clustering neither according to the different populations (isolation origin or physiological classification), nor into structure clusters or clades according to the *cox*3 analyses (Figure 15 and Figure S II).
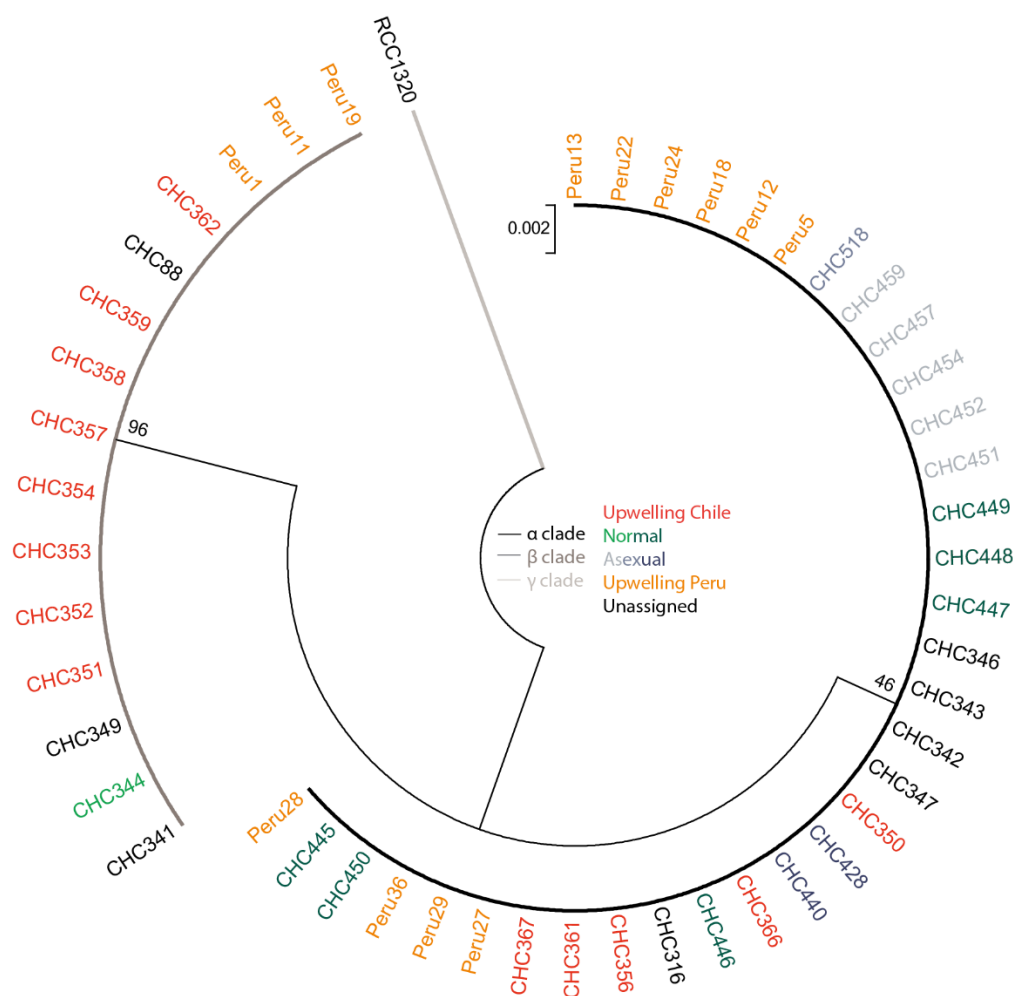
Figure 12 The evolutionary history of cox3 sequence alignment of a subset of Emiliania huxleyi *strains used in this study was inferred by using the Maximum Likelihood method based on the General Time Reversible model* (Nei & Kumar 2000). *The tree with the highest log likelihood (-188.6839) is shown. Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.1000)). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 38 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 124 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 (Tamura et al. 2013). Clades α and β correspond to the clades used in the study of Bendif et al. (2014) and correspond to the clades defined originally in (Hagino et al. 2011). The outgroup organism* Gephyrocapsa oceanica *(RCC1320) was assigned to clade γ previously described in Bendif et al. (2014)*
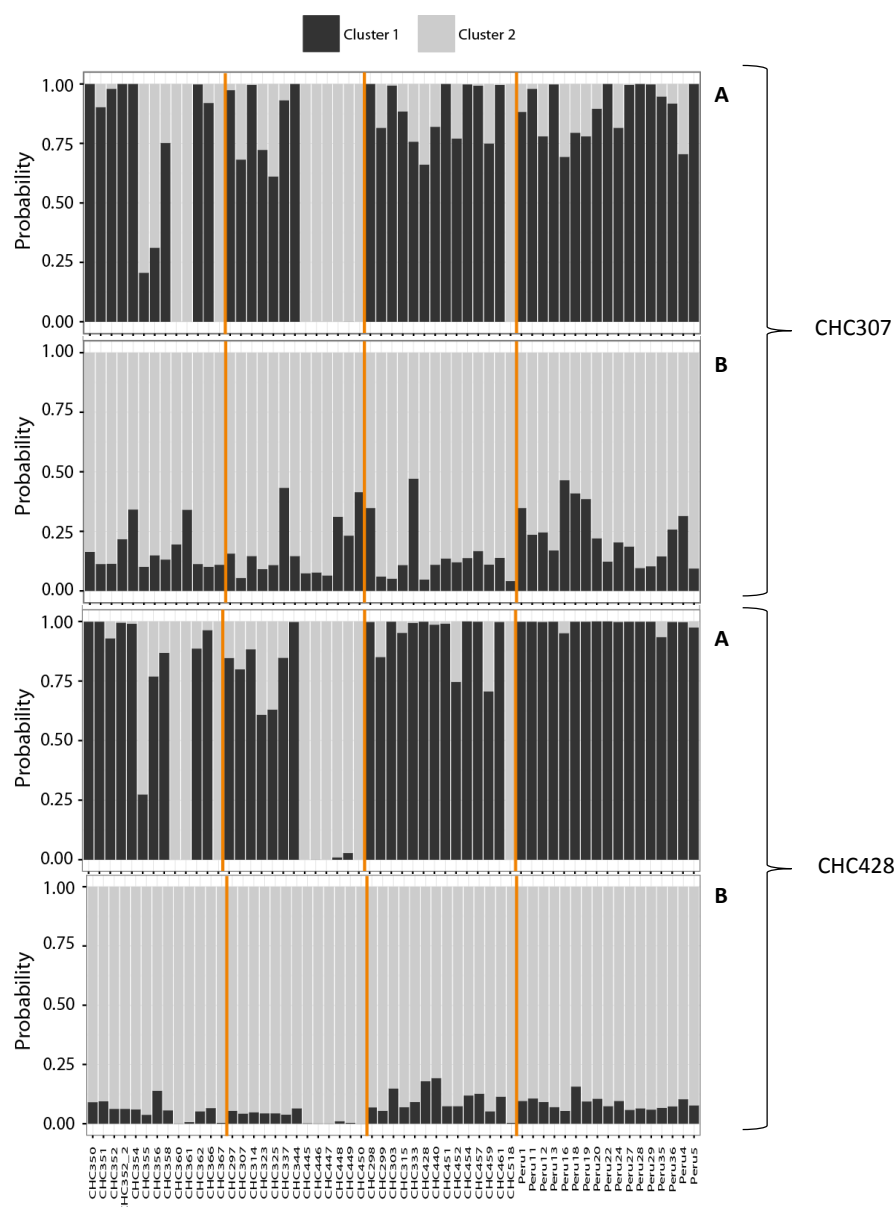
Figure 13 Probability of inferred clusters 1 and -2 obtained by running Structure v. 2.3.4 (Pritchard *et al.* 2000; Falush *et al.* 2003; Hubisz *et al.* 2009)*. with two predefined clusters,the admixture model, a burn in length of 5000 and 5000 MCM repeats after burn-in. Approximately 50 % of the iterations have resulted in plot (A) whereas the other half resulted in cluster assignment probabilities of plot (B)*
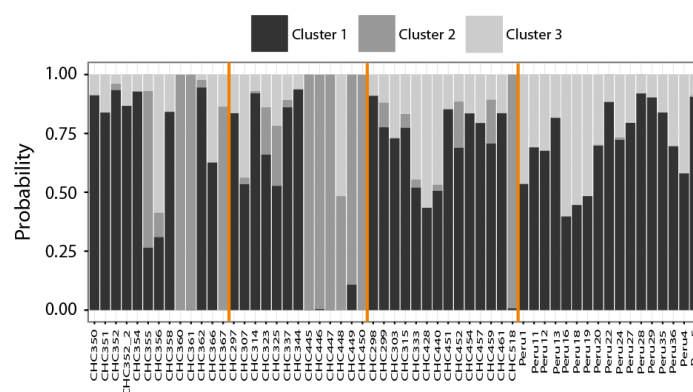
Figure 14 Probability of inferred clusters 1, -2 and -3 obtained after running Structure v. 2.3.4. (*(Pritchard *et al.* 2000; Falush *et al.* 2003; Hubisz *et al.* 2009) with three predifend clusters (K=3), the admixture model, a burn in length of 5000 and 5000 MCM repeats after burn in. As input data loci from the CHC307 stacks run were used.*
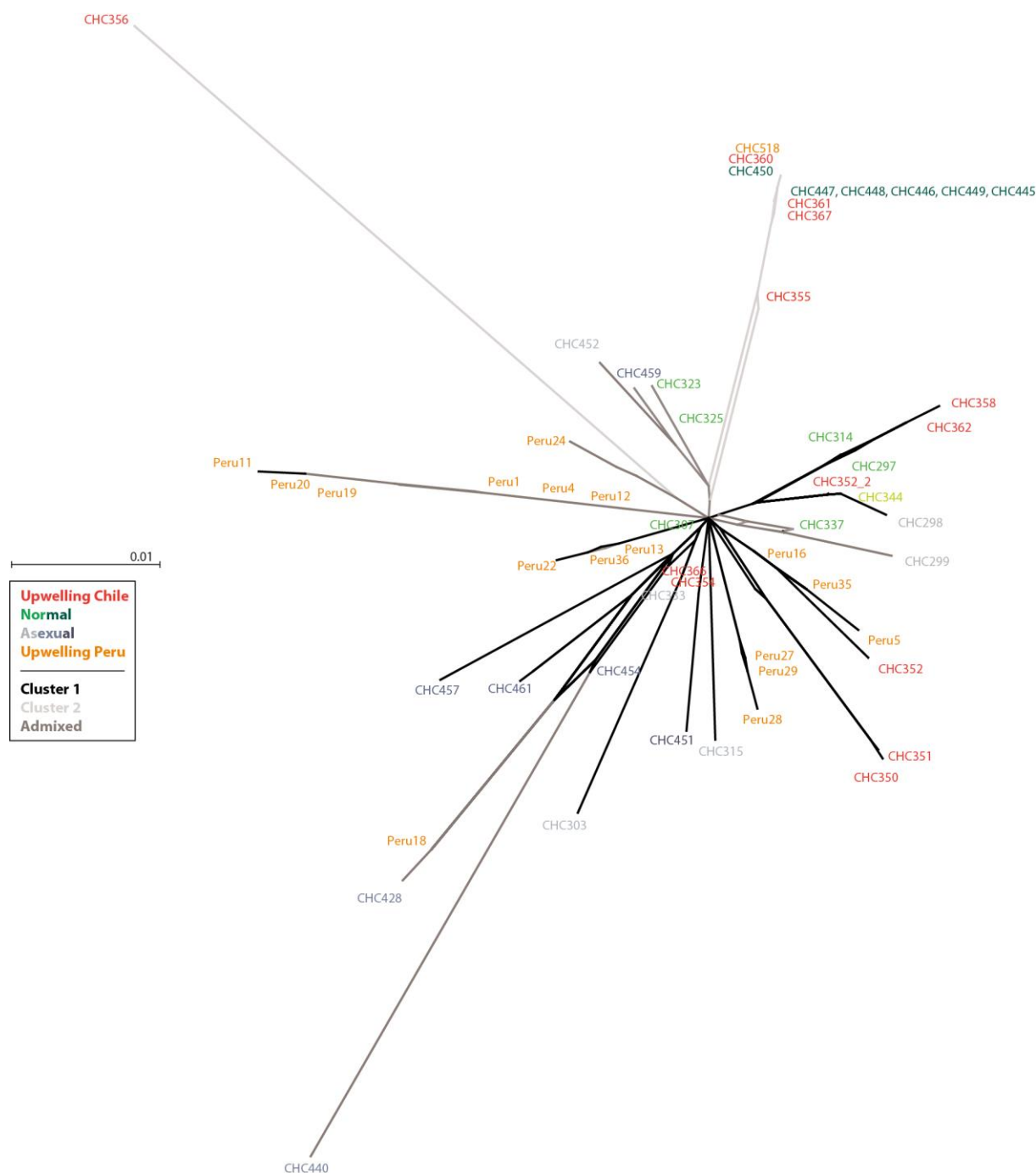
Figure 15: Neighbor-Net networks inferred with ddRAD loci of the CHC307 dataset containing minimum 1 variable SNP per locus and are present in minimum 12 individuals of one population. Neighbor networks were inferred using the software SplitsTree v. 4.13.1 (Huson & Bryant 2006) *using the default parameters and displaying only branches with a bootstrap value above 80.*

3.5. INFLUENCE OF ASEXUAL REPRODUCTION ON GENE DIVERSITY

Due to possible loss of sex in strains of the Asexual population strains were trapped in their diploid life cycle stages (von Dassow *et al.* 2014). Without recombination and only clonal cell division, certain processes, as for example accumulation of SNPs on the genome and missing purifying selection, could lead to more polymorphic genomes. As shown previously the percentage of polymorphic loci and the number of private alleles

were elevated in Asexual populations (Table 10). In order to answer this question, further analyses were done. Population specific mean Gene Diversities from various loci distributed throughout the whole genome might shed light on the question whether Asexual strains accumulate SNPs on the genome or not.
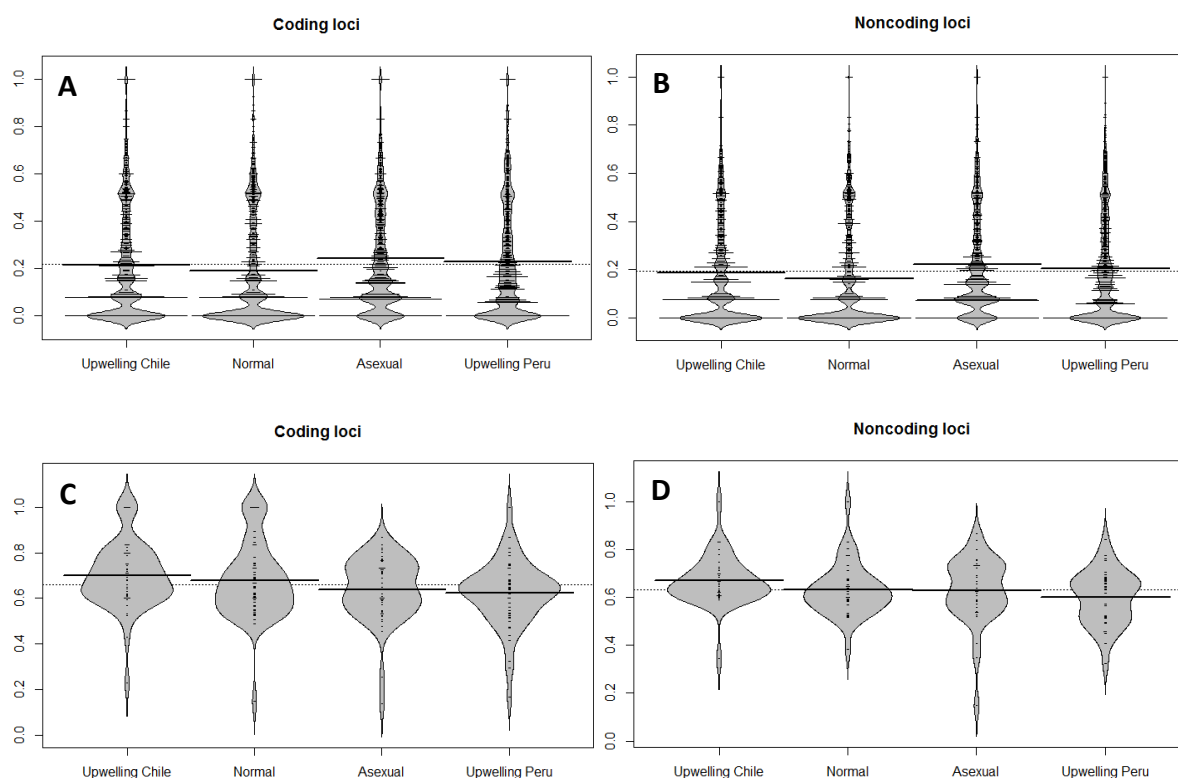
Figure 16 Gene Diversity for coding and non-coding loci from the CHC428 stacks run with (A&B) the full dataset (N$_{mean}$ = 1381.25 for coding loci, N$_{mean}$ = 1322.5 for noncoding loci) and (C&D) only loci with significant Gene Diversities with a p-value ≤ 0.05, where loci are present in all populations (N = 54 coding loci, N = 37 for noncoding loci)

Gene Diversities for every loci and population calculated with Stacks, showed clusters of Gene Diversities and no continuous distribution of Gene Diversity between 0 and 1 when including all observed Gene Diversities (Figure 16A&B). With clustered Gene Diversities calculation of population wide mean Gene Diversity was not meaningful. Furthermore, the applied Kruskal-Wallis rank sum test revealed grossly overstated significane niveaus with p-values as low as 2.2 x 10^{-16} due to the high number of sampled loci. The reduced dataset containing only significant Gene Diversities with a p-value ≤ 0.05 resulted in 54 coding and 37 non-coding loci (Figure 16C&D). Kruskal-Wallis rank sum test revealed a p-value of 0.05021 (chi-squared = 7.8054, df = 3) leading to acception of the null-hypothesis that Gene Diversities of the populations are not significantly different from each other. Population wide mean values of Gene Diversity showed significant differences between Upwelling Chile and Upwelling Peru

populations. All other populations had no significant differences in their mean Gene Diversity (Table 13). However both datasets contained ties leading to difficulties in p-value correction for the Kruskal-Wallis posthoc test after Nemenyi.

Table 13 Pairwise p-values from posthoc test after Nemenyi for Kruskal-Wallis rank sum test (p-value = 0.05021, chi-squared = 7.8054, derees of freedom = 3)

	normal	offshore	Peru
offshore	0.908	-	-
Peru	0.897	0.517	-
upwelling	0.197	0.559	0.037

3.6. IDENTIFICATION OF GENOMIC EROSION IN ASEXUAL STRAINS

Loss of flagella genes and therefore the ability form haploid stages presumably combined with the loss of sex in *E. huxleyi* was shown to be accompanied with the loss of several genes or part of genes in strains of the Asexual population (von Dassow *et al.* 2014). To test if this holds true for the tested strains or not, unique loci for the different populations were assessed. In order to understand which KOG classes are subjected to genomic erosion unique loci in the different KOG classes were identified.
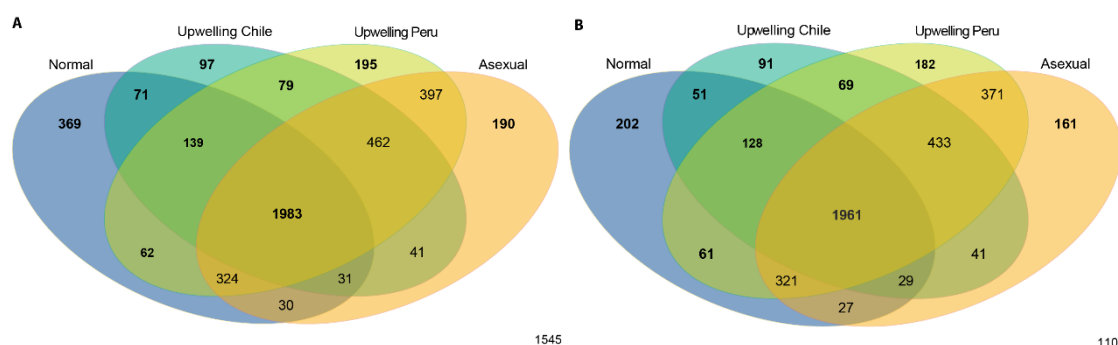
Figure 17: Venn diagram of ddRADloci from coding and non-coding regions for (A) CHC307 and (B) CHC428 Stacks run.

From the total loci set in the CHC307 Stacks dataset, 4470 loci were present in at least 12 individuals of every population. Between 2.2 and 8.3 % of the loci were population specific with the highest percentage of unique loci in the Normal population and the lowest in Upwelling Chile population. When comparing the possible Asexual population with all other populations 4.25 % of the loci were unique for the Asexual population and 22.6 % for the sexual populations (Normal, Upwelling Chile and –Peru) (Figure 17A).

For the CHC428 Stacks dataset a total number of 4128 was present in at least 12 individuals of every population. About 2.2 to 4.9 % the overall percentage of the loci were unique for the different populations. The highest percentage of unique loci had

Normal population and the lowest the Upwelling Chile population. Normal, Upwelling Chile and –Peru populations which represent possible sexually reproducing populations had 18.9 % of unique loci, which were not found in the Asexual population. Unique loci from the Asexual population contributed with 3.9 % to the total number of loci found for the CHC428 Stacks dataset (Figure 17B).

For both Stacks runs unique coding loci from Asexual populations were mainly found in the KOG classes "Unknown Function" or "General function predicted only". For most of the loci no homolog to currently known genes could be found by blast2go and were therefore excluded from Figure 18. Unique loci in sexual and asexual strains were found in nearly the same KOG classes as those who are present in all populations. Apparently no unique loci at all were found in the KOG classes associated to replication, nucleotide transport, -metabolism and –structure as well as in intracellular trafficking, coenzyme transport and cell motility. In the CHC428 Stacks run, more unique loci were detected in the Asexual population than in the CHC307 Stacks run.

3.7. SNP ANALYSES

Species subjected to highly different and diverse environments additionally with short generation times have a high possibility for mutations. Different environmental and abiotic conditions might lead to differential selection of phenotypes and genotypes. Due to that, high selective pressures not only lead to synonymous SNPs, but also to non-synonymous SNPs with potential functional impact in the protein. Most likely SNPs with non-synonymous nucleotide exchanges are present in loci which have a high fixation index (FST) in pairwise comparisons of the different populations and are therefore candidate loci for SNP analyses.

In total 72 candidate loci were identified in the CHC307 dataset and 76 candidate loci in the CHC428 dataset. From the total set of loci 24 and 22 respectively had a hit in the JGI database and a KOG class function assigned to the loci. All other loci were excluded from the SNP analyses due to either no assigned JGI hit, KOG class or the two KOG classes "General function predicted only" or "Function unknown". Due to overlaps in the two different stacks datasets (CHC307 and CHC428) a total number 32 out of 46 loci were assigned for further SNP analyses.

The detected SNPs in candidate loci were split into three categories: (1) Loci with synonymous SNPs, (2) non-synonymous SNPs and (3) loci with non-synonymous SNPs

leading to predicted changes in the secondary structure of the associated protein. About 34 % of the analysed loci showed only synonymous SNPs and for 47 % of the loci non-synonymous SNPs were detected. However 19 % of the analysed loci showed non-synonymous SNPs as well as a predicted change in the secondary structure of the predicted protein.
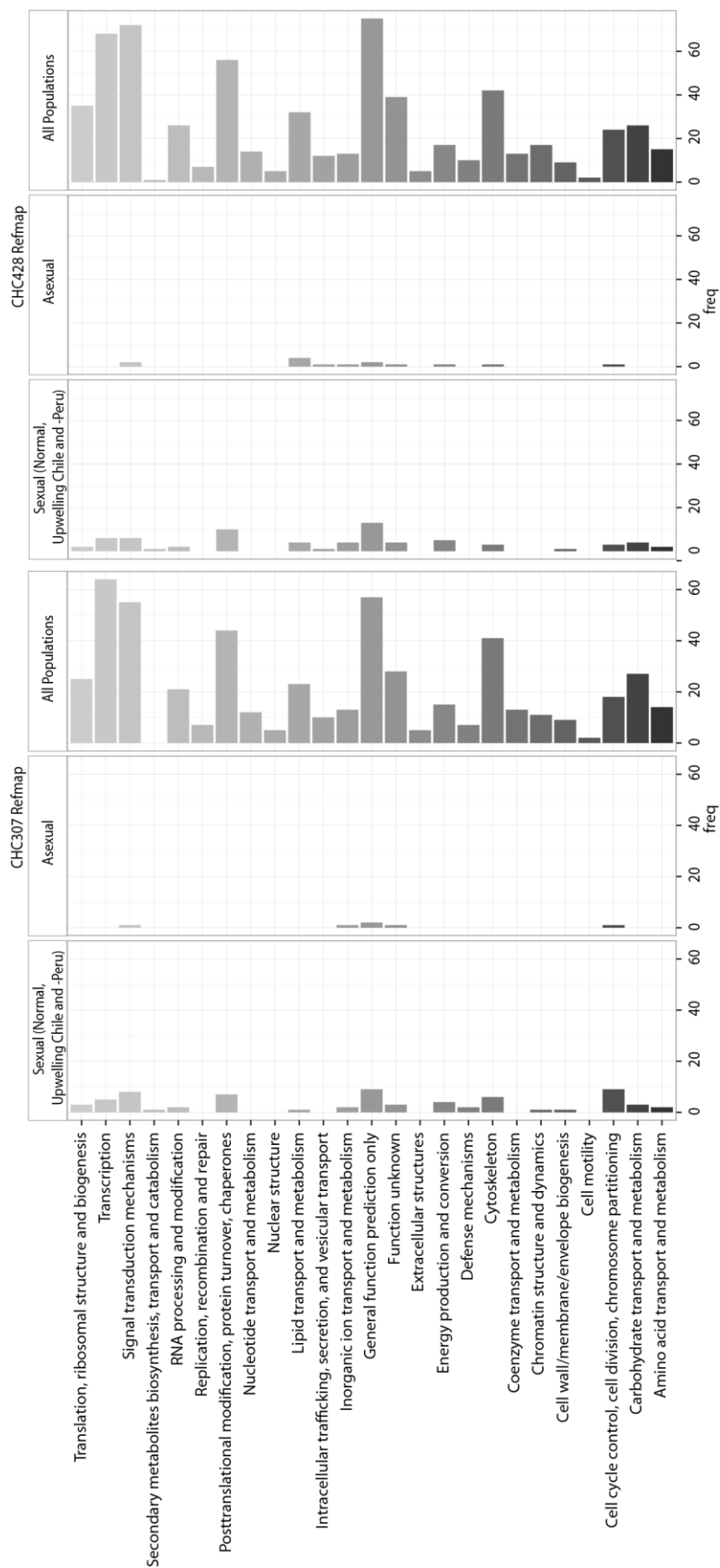
Figure 18 Exclusive loci in coding regions separated by KOG classes for(A) CHC307 and (B) CHC428 Stacks run. Loci are either exclusively found in **Upwelling, Normal and Peru** but not in Asexual population, just in **Asexual** population or in **All populations** together
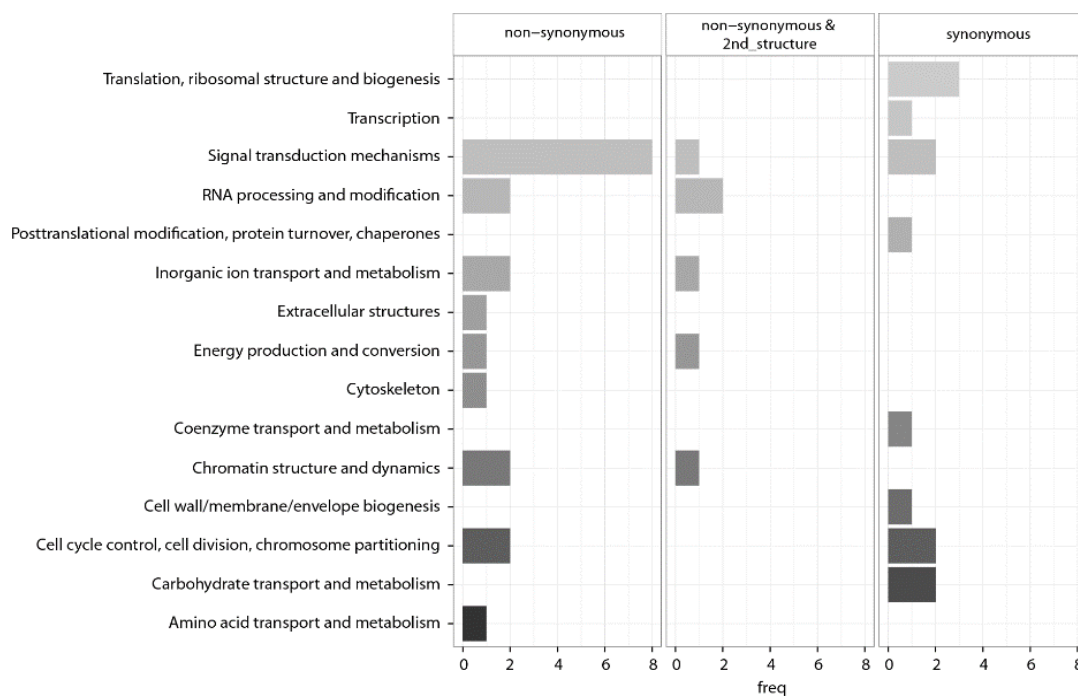
Figure 19 Number of loci with non-synonymous and synonymous SNPs per KOG class as well as number of loci with non-synonymous SNPs leading to possible secondary structure changes in the resulting protein.

Loci showing non-synonymous SNPs were mainly found in KOG classes related to modification and processing of nucleic acids (DNA and RNA), signal transduction, cell cycle control and transport processes (Figure 19). Loci which also show changes in the secondary structure were from similar KOG classes as those only showing non-synonymous SNPs, although no loci were found in KOG classes for extracellular structures, cytoskeleton, cell cycle control and amino acid transport and metabolism (Figure 19). Loci only showing synonymous SNPs were mainly found in KOG classes related to translation and transcription as well as carbohydrate transport, signal transduction and cell cycle control (Figure 19). Non-synonymous SNPs were not found specifically in one population with no apparent trends towards specific adaptation processes in one population (Figure 21 and Figure 20).
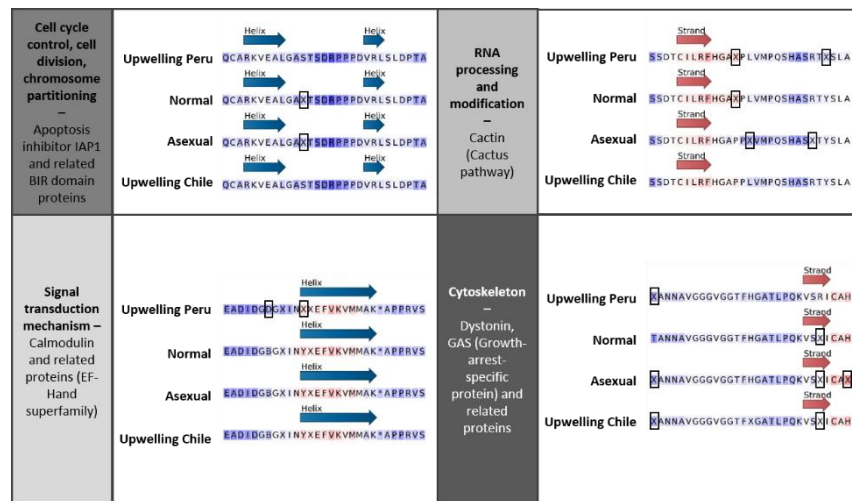
Figure 20: Amino acid sequences and KOG classes of loci with mutaions in the DNA sequence leading to non-synonmous changes. Background colour of amino acid sequence giving the hydrophobicity scale after Cornette *et al. (*1987) *(blue = min, red = max hydrophobicity)*
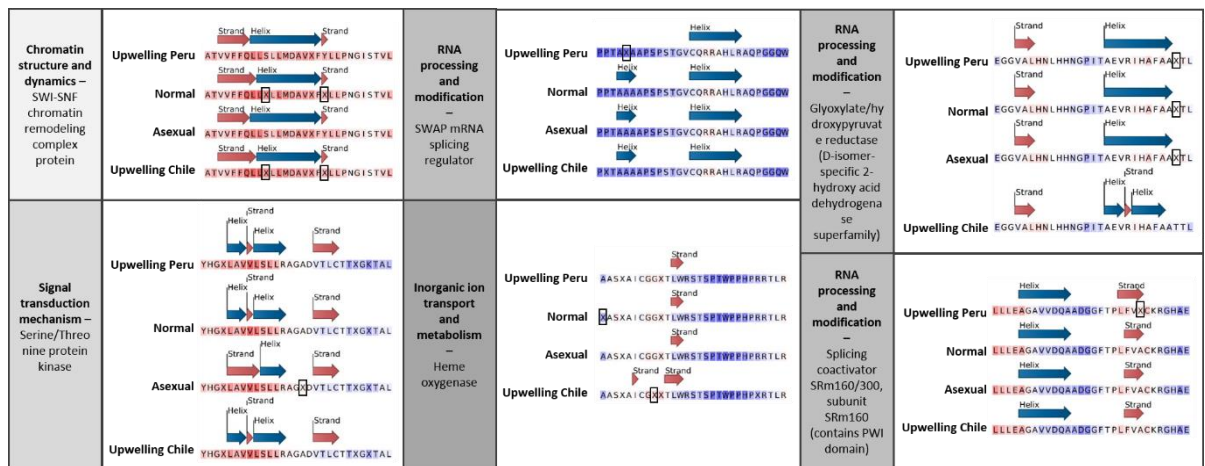
Figure 21: Amino acid sequences and KOG classes for loci with non-synonymous SNPs as well as possible predicted change in secondary structure. Background colour of amino acid sequence giving the hydrophobicity scale after Cornette et al. (1987) (blue = min, red = max hydrophobicity)

3.8. ANALSIS FOR EPIGENETIC MODIFICATIONS

Besides heritable genetic differences due to for e. g. non-synonymous SNPs also epigenetic modifications on the genome can result in changes on gene expression level and could contribute to trait variations, as it has been shown in several studies (e.g. Jaenisch & Bird 2003; Jablonka & Raz 2009; Yaish *et al.* 2011) With a modified ddRAD protocol, where additionally a methylation sensitive restriction enzyme was used, possible KOG classes subjected to epigenetic modifications could be detected.
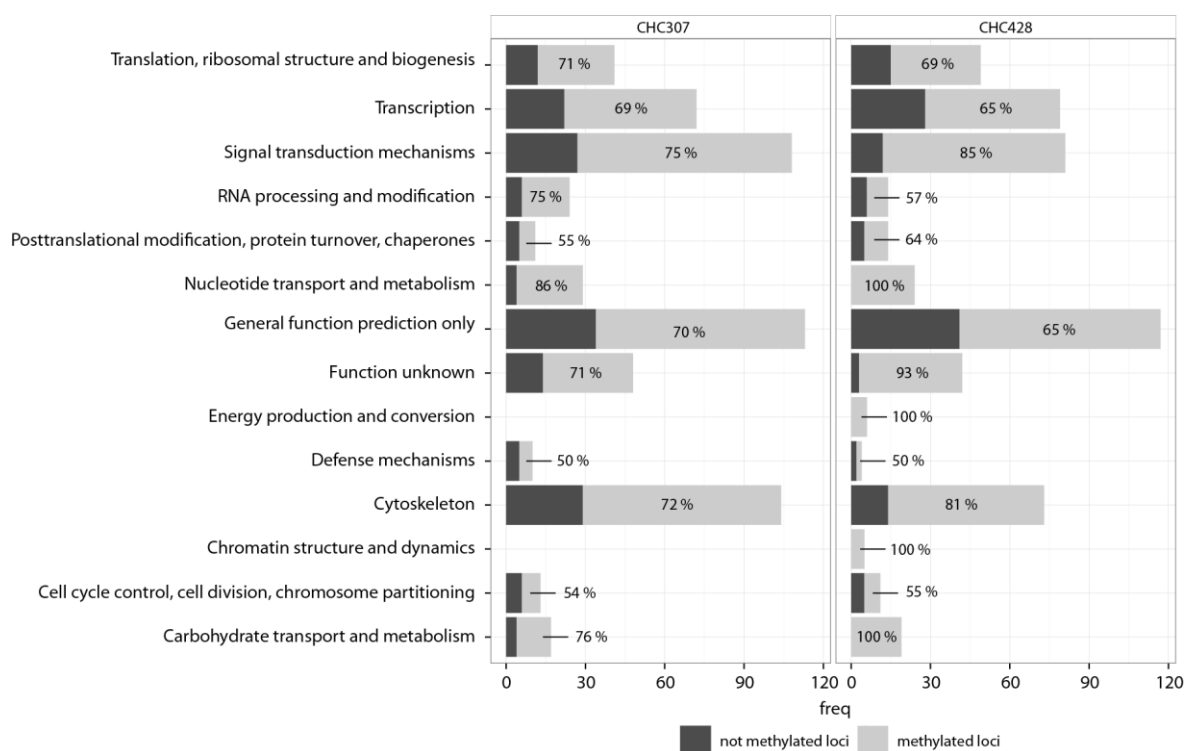
Figure 22: Number of ddRAD loci in the different KOG classes obtained after digestion with EcoRI and either MspI (light grey) or HpaII (dark grey) for the CHC307 and CHC428 Stacks runs.

Generally the percentage of methylated loci per KOG class was between 50 % and 100 %. KOG classes with more than 70 % of methylated loci were attributed to translation, signal transduction, RNA processing, nucleotide transport and metabolism, cytoskeleton and carbohydrate transport (Figure 22). In both Stacks runs methylated loci in the KOG classes with general function and function unknown had a high proportion. 100 % of the loci sampled in "Energy production and conversion" and "Chromatin structure and dynamics" were methylated. However when ddRAD sequences were mapped against CHC307 genome before Stacks analyses, loci in Energy production and chromatin structure KOG classes fail to be sampled completely (Figure 22).

4. DISCUSSION

4.1. METHODOLOGICAL DISCUSSION

Restriction site associated DNA sequencing (RAD; Miller *et al.* 2007) has become a powerful tool in molecular ecological studies. Especially further developments of the basic RAD technique by using two restriction enzymes (ddRAD; Peterson *et al.* 2012) enabled researchers to investigate population genetic relationships (e.g. Catchen et al., 2013; Emerson et al., 2010; Macher et al., 2015), genetic mappings (e.g. Zhou et al., 2014) as well as test for selection acting on genomic regions (e.g. Emerson *et al.* 2010b; Hohenlohe *et al.* 2010, 2012, 2013) on the basis of genome wide distributed SNP markers.

Although with increasing number of marker loci, especially when distributed throughout the genome, analyses reveal statistically more robust results, ddRAD analyses can have certain drawbacks. One of the most discussed drawback is the phenomenon of the so called "Allele dropout (ADO; Arnold *et al.* 2013; Gautier *et al.* 2013; McCormack *et al.* 2013). Due to mutations in enzyme cutting sites, null alleles can be obtained which leads to a bias of certain alleles. Further unequal PCR success produces a bias towards smaller ddRAD fragments as well as an increase of fragments with higher GC contents (Davey *et al.* 2012) which ultimately leads to an underrepresentation of other alleles. Diversity indices as well as overall genetic diversity in the dataset are affected by ADO which leads to inflation of homozygosity. Most sensitive parameters for reduction of diversity were observed heterozygosity (Het_{obs}) and nucleotide diversity (π) whereas ADO had only minor effects on F_{ST} values inferred with ddRAD.

However also populations with high effective population sizes are particularly susceptible for ADO. In case of *E. huxleyi*, which has a high effective population size, ADO might have led to overall decreased genetic diversity in the obtained dataset and population genetic results have to be handled with care when interpreting values for Het_{obs}, π, F_{IS} but also F_{ST}, as these might be substantially reduced due to ADO.

Besides ADO also mistakes in further analyses steps can introduce a bias to the data. When mapping ddRAD raw reads against a reference genome without disabling terminal alignments, non-mapping regions will get soft-masked in the exported SAM or BAM file. The program Stacks which was used to build loci out of raw sequence

ddRAD reads will then replace soft-masked regions with N´s. If this occurs extensively within one locus, polymorphisms and haplotypes will not be calculated from soft masked loci. The program CLC genomics workbench v. 8, which was used in the present study for mapping raw reads against reference genomes, appeared to have created terminal alignments. Unfortunately this was noticed rather late during analysis of ddRAD loci. For further ddRAD studies it is highly recommended to either turn off soft-masking in CLC or rather use other mapping programs such as GSnap (Wu & Watanabe 2005; Wu & Nacu 2010), BWA (Li & Durbin 2009) or Bowtie2 (Langmead & Salzberg 2012). However only 6% of the reference mapped loci contained a high proportion of N´s after mapping against the reference genomes CHC307 and CHC428, thus comprised only a minor part of all obtained loci and were expected to have only a minor influence on the obtained basic population genetic parameters. This is indicated by the fact, that basic population genetic parameters calculated from loci built *de novo* showed even lower levels of polymorphisms. Detected overall patterns in basic population genetic parameters were consistent in all Stacks runs, indicating, that loci drop-out due to terminal alignments played a minor role and was most likely not affecting the overall output. However to increase statistical significance and robustness of ddRAD results, for future ddRAD studies with *E. huxleyi* terminal alignments should be avoided.

4.2. POPULATION GENETIC ANALYSES

Substantial to understand the relationships between populations is calculating and evaluating basic population genetic parameters, as they describe important measures of population differentiation. Important measures are here observed heterozygosity (Het_{obs}), nucleotide diversity (π) as well as estimates for fixation of different alleles (F_{ST}) between and within populations. Het_{obs} indicates the proportion of heterozygous loci and together with observed homozygosity deviations from the Hardy-Weinberg equilibrium can be calculated. A deviation from Hardy-Weinberg equilibrium indicates, that populations experience non-random mating, selection of specific genotypes, migration and mutation processes are influencing allele frequencies, the population size is not infinite or alleles do not segregate according to normal Mendelian inheritance (Freeland *et al.* 2011). For *E. huxleyi* levels of observed heterozygosity in microsatellite studies were in the range of 0.14 and 0.73, highly dependent on which microsatellite loci was used but also depend on the geographic origin of the analysed strains (Debora Iglesias-Rodriguez *et al.* 2006). Observed heterozygosity calculated

with ddRAD loci was somewhat at the lower edge of previous studies with values between 0.13 and 0.14 for CHC307 and CHC428 Stacks runs (Table 12).

Population differentiation is influenced by a range of different factors. By far most important factors for intra-population diversity and differentiation of geographically separated populations are the rate of genetic drift but also the amount of gene flow between populations (Freeland *et al.* 2011). Genetic drift tends to reduce genetic diversity by driving alleles either to fixation or extinction in the absence of selection (Freeland *et al.* 2011). Contrary to genetic drift, gene flow increases genetic diversity due to exchange of genes between populations (Freeland *et al.* 2011). In populations with sexually and asexually reproducing strains gene flow between strains is restricted which leads to separation of gene pools within a population. Restricted gene flow between these two gene pools might favour sympatric evolution which can ultimately lead to speciation or at least to gene pools with profound differences. As homologous recombination during meiosis is supposed to be not present in asexual populations, a loss of heterozygosity is expected for Asexual populations, especially due to an increase of hemizygosity. This would lead to increase of inbreeding, detected by elevated inbreeding coefficient (F_{IS}) values. F_{IS} values calculated with ddRAD loci revealed an increase in Asexual vs. Sexual populations as highest F_{IS} values were detected in the Asexual population. Strains of the Asexual population were from the same geographic origin as strains which have retained flagella genes (Normal population), however these populations show differences in their inbreeding coefficients, leading to the assumption that different evolutionary processes, especially related to reproduction modes, were influencing genome evolution.

Loss of sex can have further evolutionary consequences for the evolution of asexual genomes. On the one hand asexual populations have a higher possibility for stepwise accumulation of mutations, but also the lack of further purifying selection of new alleles. In combination with an increase of homozygosity in asexual populations, deleterious mutations can be unmasked (Lynch & Gabriel 1990; Lynch *et al.* 1990). However as a counterpart to accumulation of deleterious mutations acts ameiotic recombination thus helps to sustain a high fitness in Asexual population, which is particularly important to sustain adaptive capacity in diverse environments. To assess the influence of asexual reproduction on the genetic diversity, percentages of polymorphic loci were calculated. Asexual strains indeed showed a higher percentage

of polymorphic loci compared to all other populations, which is in line with the underlying assumptions. However F_{IS} values and the percentage of polymorphic loci revelead contrasting results, making it difficult to make assumptions whether nonallelic homologous recombination or ameiotic recombination is shaping the genomic evolution of Asexual populations.

Assessment of phylogenetic relationships between sexual and asexual strains in previous studies (Beaufort *et al.* 2011; Hagino *et al.* 2011; Bendif *et al.* 2014) unveiled that the loss of sex occurred recently and appeared to be an independent event in different populations and strains. Although the loss sex occurred just recently, evolutionary forces already had an impact on the genome of asexual strains indicating that sympatric evolution might play a role in shaping differentiation of the two gene pools.

A measure for genetic diversity is nucleotide diversity (π). Here the mean sequence divergence among haplotypes is calculated by taking into account frequencies and pairwise divergence of sequences (Freeland *et al.* 2011). The fixation index (F_{ST}) is used to calculate the genetic differentiation of populations on the basis of correlation between random gametes (Freeland *et al.* 2011). Higher values for π and F_{ST} indicate a higher genetic diversity as well as a higher genetic divergence within and between populations. F_{ST} values calculated from microsatellite datasets in previous studies (Debora Iglesias-Rodriguez *et al.* 2006) showed moderate to low (F_{ST} = 0.099) differentiation between the different tested populations. Mean pairwise F_{ST} values calculated with ddRAD loci were apparently smaller ranging from only 0.023 to 0.043 (Table 11). On the one hand methodological bias might have been introduced by ADO, although F_{ST} values were shown to be less sensitive to a bias introduced by ADO (Arnold *et al.* 2013). Even though F_{ST} values with ddRAD loci were fairly low they still showed moderate differentiation between the populations with the highest differentiation between Peruvian and Chilean population for both Stacks runs (Table 11). In concordance with high F_{ST} values, strains from the Peruvian population also showed the highest number of private alleles for all Stacks runs (Table 12). Geographic distances between the Chilean and Peruvian populations were rather small compared to the whole sampling area (Figure 3 and Figure 23), but Peruvian and Chilean populations appear to originate from different water masses (Peruvian in Subtropical Waters (SW) and Chilean in Subantarctic Water (SAAW)) which are not interconnected

by the different prevailing currents (Silva *et al.* 2009) (Figure 23). In contrast to this, Normal, Asexual and Upwelling Chile populations are from the same water masses (Silva *et al.* 2009) and therefore gene flow between these populations might be only partially restricted and F_{ST} values remain low between these populations. This is further support by Structure as well as Neighbour Nets results (Figure 13 and Figure 15).

Especially the neighbour-net networks underpinned that differentiation was rather based on individual strains than on population based levels. Most of the strains were separated on individual branches instead being clustered on population based branches. Strains assigned to cluster 2 in Structure analyses were grouped separate from other strains in neighbour – net networks as well. Apparently these strains belong to the same geographic population except strains from Upwelling Chile population and CHC518. This might indicate genetic exchange between populations due to the Humboldt Current flowing northwards. All in all, distances between branches were small and they arose all from the same origin underpinning low genetic structuring according to populations in the dataset (Figure 15). As ddRAD samples had a high number of loci, results were suspected to show more detailed but comparable findings of population structuring as if inferred from *cox*3 sequences. This was not the case in the present study. The observed clusters found in cox3 analyses were not detected with the ddRAD dataset. However other ddRAD studies (Macher *et al.* 2015) were able to confirm patterns of population division drawn from mitochondrial markers (*cox*1).
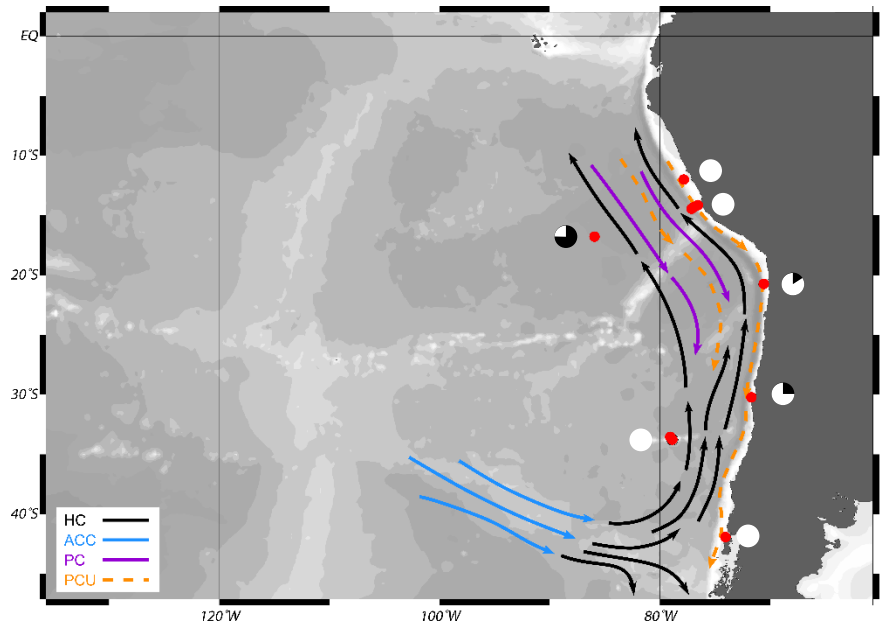
Figure 23 Prevailing Humboldt current (HC) system off the coast of Peru and Chile flowing northwards with the Peruvian countercurrent (PC) flowing southwards as well as the Peru-Chile Undercurrent (PCU) flowing southwards in about 200 m depth (Silva *et al.* 2009). *Water masses in the Humboldt Current originate from the Antarctic circumpolar current (ACC). Red dots are sampling locations of strains used in this study. The white-black circles next to the populations represent percentages of individuals per population assigned to structure cluster 1 (white) and -2 (black) with a probability of more than 90 %.*

In previous studies population genetic distinction of asexual *E. huxleyi* strains was not possible based on mitochondrial markers and von Dassow *et al.* (2014), suggested that a much higher number of strains need to be sampled to be able to distinguish asexual strains from others with population genetic analyses. The particular advantage of using ddRAD is, that with a higher number of loci, fine scale population genetic differentiation should detectable, even a few individuals as it has been shown in previous studies (for e.g. Macher *et al.* 2015). However with ddRAD population genetic discrimination of asexual strains in neighbour-net networks was not possible. No separation into populations or phenotypes/genotypes has been detected with ddRAD

4.3. IDENTIFICATION OF POLYMORPHISM INCREASE DUE TO THE LOSS OF SEX

The lack of meiosis and recombination as a result of sexual reproduction was shown to have an impact on the genomic evolution of species (Xu *et al.* 2011; Flot *et al.* 2013; von Dassow *et al.* 2014). In the diploid genome of the CCMP1516 strain of *E. huxleyi* it could be shown that chromosomes showed genetic differences leading to the conclusion, that long term absence of meiosis had led to rearrangements, inversions and hemizygous deletions in the genome of CCMP1516 (von Dassow *et al.* 2014). Similar processes were detected in the asexually reproducing species of *Daphnia* (Xu *et al.* 2011) and bdelloid

rotifers (Flot *et al.* 2013). Further in asexual species mutations tend to increase and accumulate in a stepwise manner, leading to an increased overall polymorphism on the genome. In this study we aimed to address this phenomenon by analysing genome wide Gene Diversity of asexual populations in comparison to populations still being able to reproduce sexually.

Unfiltered Gene Diversities of *E. huxleyi* calculated from ddRAD loci showed clustered values for both Stacks runs. Due to this calculation of mean genome wide Gene Diversities was not possible. After reducing the dataset that only Gene Diversities, previously identified as significant in Stacks, were included as well as only Gene diversities of loci present in all populations, the data set contained only 54 loci in coding regions and 37 in non-coding regions. Although Gene Diversities showed a more Gaussian distribution sorting and significant differences between mean Gene Diversities could be detected, we were not able to answer the underlying assumptions.

Still elusive is, what caused Gene Diversities to show a clustered distribution in the present ddRAD library (Figure 16). Gene diversities of ddRAD loci inferred from oysters (*Crassostrea gigas)* showed a normal, Gaussian like distribution (Figure 24), although ddRAD libraries were prepared the same way and sequenced in the same sequencing run on the Illumina NextSeq 500. Differences between the *C. gigas* and *E. huxleyi* ddRAD datasets appeared to be the restriction enzymes used to digest the genomes of the two different species. Read et al. (2013) unveiled that tandem repeats (TRs) in *E. huxleyi* had an exceptionally high genomic coverage compared to other eukaryotes. TR density was highest in introns and a high percentage of TRs from low complexity regions contained a trinucleotide pattern CCG or variations of this motif. The overall genomic GC content in *E. huxleyi* was found to be about 65 %. In TRs can the GC content can even exceed 65 %, possibly due to the frequent occurrence of CCG patterns in TRs (Read *et al.* 2013). *Msp*I and *Hpa*II selectively cut DNA at sites with the nucleotide sequence CCGG. The high abundance of TRs and low complexity regions (64%) as well as the high percentage of G, C nucleotides, CCG pattern in particular, could have led to a sampling bias towards TRs and low complexity regions in the dataset and a clustered Gene Diversity.

A similar pattern of clustered Gene Diversity has been observed for *Alexandrium* spp. The species complex of *Alexandrium* spp. exhibits comparable levels of repetitive sequences (51 %) and GC content (60%) as well as a high overall genetic diversity within and among populations (Alpermann *et al.* 2009, 2010; Jaeckisch *et al.* 2011). In contrast to this, Zhang *et al.* (2012) could show that the genome of *C. gigas* has only about 36 % of repetitive sequences and a GC content of 45.2 %.
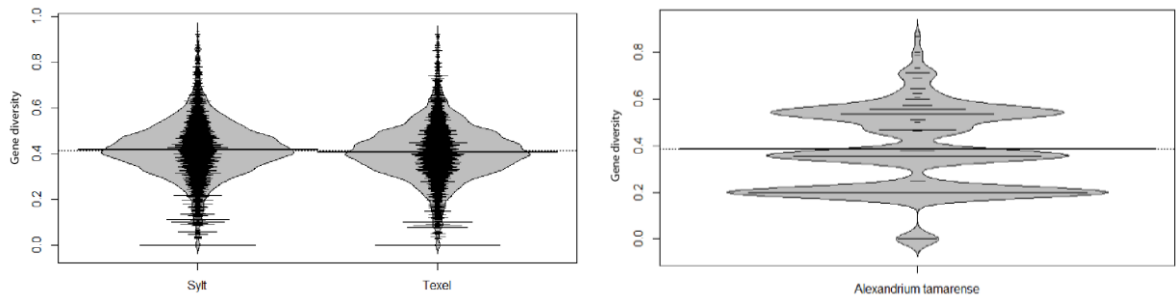
Figure 24: Gene diversity for (A) two Oyster populations and (B) six different Alexandrium spp. strains from one population inferred with Stacks v. 2.3.4. from ddRAD loci.

4.4. DETECTION OF GENOME EROSION IN ASEXUAL STRAINS

Recent studies unveiled, that asexual strains of *E. huxleyi* were subjected to substantial loss of genes specific for formation of haploid cell stages as well as other haploid specific genes indicating the possibility for complete loss of sex (von Dassow *et al.* 2014). Results from comparisons of coding ddRAD loci between sexual and asexual strains indicate the loss of several coding genes, some of which previously identified in the study of von Dassow et al. (2014).

As in previous studies also in the present study ddRAD loci in genes which are related to 1N cell formation and prospectively involved in cell behaviour were missing from asexual strains. Among these were for example two Calmodulin homologs (GJ06730 and GJ10041) and a predicted tyrosine-tubulin ligase (GJ01643). Two remaining calmodulin homologs conserved in all strains showed non-synonymous SNPs and high FST values in pairwise population comparison, indicating a high selective purifying pressure on these genes. However ddRAD loci in other genes previously demonstrated to be subjected to genomic erosion remain undetected in the present study. The sampling approach of ddRAD is based on reduced representation sequencing, where the loss of genomic informations in the sequenced library was taken into account intentionally.

Although with the ddRAD approach erosion of genes directly involved in flagella formation (cDHC and DHC1b genes) could not be detected, genes which might play a role in flagella functioning were missing from asexual strains (Myosin class V heavy chain, (GJ01749 and GJ00105).

No unique loci were found in the KOG classes associated to replication, nucleotide transport, -metabolism and –structure as well as in intracellular trafficking, coenzyme transport and cell motility. Possibly genes in these KOG classes comply important housekeeping functions in diploid cells and were therefore not subjected to genomic erosion or failed to be detected due to the choice of restriction enzymes or ADO.

Differences in the sampling success of unique loci were detected between the CHC307 and CHC428 dataset. Various loci remained undetected when mapping Illumina raw reads against CHC428 prior to Stacks analysis, for e.g two Calmodulin homologs. This again underpins the pan genome concept in *E. huxleyi* (Read *et al.* 2013) as well as the presence of genome erosion in asexual strains, since flagella related genes remained undetected in CHC428 in previous studies (von Dassow *et al.* 2014). About 31 % of all unique loci didn´t show homologies to previously known genes and might be attributed to coccolithophore or even *E. huxleyi* specific functions.

Loss of genes in general was consistent for all sampled asexual *E. huxleyi* strains from more stable, oligotrophic environments which is in concordance with the results from von Dassow et al. (2014). Although no direct indications were present, that the loss of sex is attributed to oligotrophy, von Dassow et al. (2014) could show that the loss of sex is significantly increased in regions with more stable, oligotrophic waters as well as diluted *E. huxleyi* populations. Analysis of metagenomics data of sampling regions indicated that the *E. huxleyi* virus EhV was not present, at least not in high concentrations. Biotic pressures to maintain a biphasic life and in particular sexual reproduction, as sexual populations are suspected to have a higher adaptive capacity. In regions where EhV is barely detectable and *E. huxleyi* populations are highly diluted, biotic pressures are reduced or even released. The importance of biphasic life cycling diminishes and released selective pressure could lead to small scale loss of genes attributed to haploid cell formation in a high proportion of (von Dassow *et al.* 2014).

4.5. SNP ANALYSES

The presence of non-synonymous SNPs with and without possible impact at the protein function, such as secondary structure changes, could be detected for both datasets. An accumulation of non-synonymous SNP appeared to be especially in the KOG class Signal transduction mechanism. Only synonymous SNPs could be detected in "Translation, ribosomal structure and biogenesis" and "Transcription". Non-synonymous SNPs can, but don´t have to lead to substantial protein modifications, e.g. when located in the active centre. Translation and Transcription are important processes fundamental for survival of cells and seem to experience higher level of conservation and purifying selection as other classes of orthologous genes, indicated by only synonymous SNPs. Diversifying selection might play an important role in facilitating signal transduction as the percentage of non-synonymous SNPs were high. Several proteins possibly involved in the regulation of cell proliferation, apoptosis or cell differentiation were detected with non-synonymous SNPs. A putative serine-threonine kinase was detected to show a non-synonymous SNP with possible changes in secondary structure in the Asexual population. As serine/threonine kinases are thought to be involved in cell proliferation and apoptosis, relaxed selection pressures due to loss of sex, might lead to an increase of mutations in haploid cell stage related genes. However SNPs are neither specific for asexual or sexual populations nor for populations from highly diverse or stable environments. Apparently adaptation processes are not detectable based on single nucleotide polymorphism but rather on presence and absence of genes in different strains/populations according to the concept of a pan genome with highly strain specific genomic regions (Read *et al.* 2013).

4.6. ANALYSIS FOR EPIGENETIC MODIFICATIONS

Epigenetic modifications of DNA are nowadays widely accepted to have an impact on gene expression and trait variation, which was shown in numeral different studies (e.g. Jaenisch & Bird 2003; Jablonka & Raz 2009; Yaish *et al.* 2011). In the course of epigenetic modifications, DNA or histones get methylated which leads to repression or activation of certain genes. To some extent mutations cannot explain the amount trait variation we can find in for example Phytoplankton. With the presented modified ddRAD approach (EpiRADSeq), which was just recently by Schield *et al.* (2015), signatures of epigenetic modifications can be detected throughout the genome in a cost

and time effective manner. Sampling of methylated loci with ddRAD was shown to be successful in the present study and just recently in a study of Schield *et al.* (2015).

The highest number of epigenetically modified loci was found in "Signal transduction mechanism", "Cytoskeleton" and "General Function predicted only". The presence of non-synonymous SNPs and a high number of epigenetically modified loci in this KOG classes might indicate adaptation processes on short- and meso time scales and their importance for the high adaptive capacity of *E. huxleyi* strains.

Transcriptomic studies in combination with EpiRADSeq can gain an advanced understanding how asexually and sexually reproducing *E. huxleyi* strains react to environmental and biotic stressors, especially in terms of Global Change.

4.7. CONCLUSION AND OUTLOOK

The present study underpins the high intraspecific genomic diversity in *E. huxleyi*, indicated for e.g. by low inter-population F_{ST} values or population specific genes. However, sympatric genomic evolution on intra-population level induced by reproductive isolation was indicated by low, but detectable, F_{ST} values within asexual and sexual populations from the same geographic range. However based on the obtained data in this study, it remains unclear which underlying mechanisms is shaping the genomic evolution of asexual *E. huxleyi* strains. Controversal findings for inbreeding coefficients, observed heterozygosities and genome wide polymorphisms made it difficult to assess if Asexual populations are pruned to a loss of adaptive capacity or not. The underlying, exceptionally high genetic diversity of *E. huxleyi* could somehow mitigate negative effects of asexual reproduction on a short time scale.

Species success of asexual *E. huxleyi* is of particular interest as geographic ranges with stable, oligotrophic environments will increase in the course of Global Change. With the range expansion of oligotrophic environments the chance for the loss of sex in further *E. huxleyi* populations increases as well. On the other hand, global change will also impact the open ocean and subtropical- to tropical environments where species occupy niches closer to their reaction norm boundaries. Asexuall populations with an increased hemizygosity can have reduced adaptive capabilities in such areas, due to unmasking of deleterious allels, and might be particular challenged by increasing temperatures and shift and reduction in the nutrient regimes. Since *E. huxleyi* is an important player in the global, biogeochemical carbon cycle, a decreasing adaptive

capacity of *E. huxleyi* might ultimately decrease vertical carbon fluxes within and into the ocean.

In order to be able to make more reliable predictions on the fate of asexual *E. huxleyi* lineages, further parameters, as for example ameiotic recombination rates and mutation rates should be determined on whole genome scale in mutation accumulation (MA) experiments. MA experiments with a small bottleneck size eliminate the effect of natural selection and random mutations can accumulate. By comparing mutation and ameiotic recombination rates in asexual and sexual strains of *E. huxleyi* we would be able to determine if asexual *E. huxleyi* strains are pruned to accumulate deleterious mutations or not.

An advanced understanding of adaptive capacities and the underlying mechanisms in asexual and sexual populations can further be gained by transcriptomic and epigenetic studies with for example EpiRADSeq. To confirm results from the present study ddRAD experiments with more strains and other restriction enzymes are highly recommended. With the use of other restriction enzymes than *Msp*I the observerd patterns of Gene Diversity can be avoided and more robust results can be obtained.

The use of mapping tools as for example Bowtie or GSnap would avoid terminal alginments which could further increase the number of ddRAD available for analysis and therefore the statistical power of ddRAD results.

5. REFERENCES

Agencourt Bioscience Cooperation (2009) *Agencourt® AMPure® XP - PCR Purification*. Massachusetts.

Alpermann TJ, Beszteri B, John U, Tillmann U, Cembella AD (2009) Implications of life-history transitions on the population genetic structure of the toxigenic marine dinoflagellate Alexandrium tamarense. *Molecular ecology*, **18**, 2122–2133.

Alpermann TJ, Tillmann U, Beszteri B, Cembella AD, John U (2010) Phenotypic variation and genotypic diversity in a planktonic population of the toxigenic marine dinoflagellate Alexandrium tamarense (Dinophyceae). *Journal of Phycology*, **46**, 18–32.

Altshuler D, Pollara VJ, Cowles CR *et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Andrews S (2010) FastQC A quality control tool for high throughput sequence data.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

Bach LT, Mackinder LCM, Schulz KG *et al.* (2013) Dissecting the impact of CO2 and pH on the mechanisms of photosynthesis and calcification in the coccolithophore Emiliania huxleyi. *The New phytologist*, **199**, 121–34.

Beaufort L, Couapel M, Buchet N *et al.* (2008) Calcite production by coccolithophores in the south east Pacific Ocean. , 1101–1117.

Beaufort L, Probert I, de Garidel-Thoron T *et al.* (2011) Sensitivity of coccolithophores to carbonate chemistry and ocean acidification. *Nature*, **476**, 80–3.

Beaumont M a., Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Beckman Coulter I (2014) *Agencourt CleanSEQ Dye-Terminator Removal Protocol*. Brea, California.

Bendif EM, Probert I, Carmichael M *et al.* (2014) Genetic delineation between and within the widespread coccolithophore morpho-species Emiliania huxleyi and Gephyrocapsa oceanica (Haptophyta) (T Mock, Ed,). *Journal of Phycology*, **50**, 140–148.

Benner I, Diner RE, Lefebvre SC *et al.* (2013) Emiliania huxleyi increases calcification but not expression of calcification-related genes in long-term exposure to elevated temperature and p CO 2 Emiliania huxleyi increases calcification but not expression of calcification-related genes in long-term e. *Philosophical Transactions of the Royal Society B*, **368**.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bopp L, Monfray P, Aumont O *et al.* (2001) Potential impact of climate change on marine export production. *Global Biogeochemical Cycles*, **15**, 81.

Borchard C, Borges A V., Händel N, Engel A (2011) Biogeochemical response of Emiliania huxleyi (PML B92/11) to elevated CO2 and temperature under phosphorous limitation: A chemostat study. *Journal of Experimental Marine Biology and Ecology*, **410**, 61–71.

Brussaard C, Kempers R, Kop A, Riegman R, Heldal M (1996) Virus-like particles in a summer bloom of *Emiliania huxleyi* in the North Sea. *Aquatic Microbial Ecology*, **10**, 105–113.

Caldeira K, Wickett ME (2003) Anthropogenic carbon and ocean pH. *Nature*, **425**, 365.

Caliper Life Sciences (2011) *LabChip XT / XTe - User manual*. Massachusetts.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda, Md.)*, **1**, 171–82.

Catchen J, Bassham S, Wilson T *et al.* (2013a) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology*, **22**, 2864–2883.

Catchen J, Hohenlohe P a, Bassham S, Amores A, Cresko W a (2013b) Stacks: an analysis tool set for population genomics. *Molecular ecology*, **22**, 3124–40.

Cavallo F, De Giovanni C, Nanni P, Forni G, Lollini P-L (2011) 2011: the immune hallmarks of cancer. *Cancer immunology, immunotherapy : CII*, **60**, 319–326.

Coelho SM, Peters AF, Charrier B *et al.* (2007) Complex life cycles of multicellular eukaryotes: New approaches based on the use of model organisms. *Gene*, **406**, 152–170.

Collins S, Rost B, Rynearson T a (2014) Evolutionary potential of marine phytoplankton under ocean acidification. *Evolutionary applications*, **7**, 140–55.

Coolen MJL (2011) 7000 years of Emiliania huxleyi viruses in the Black Sea. *Science (New York, N.Y.)*, **333**, 451–452.

Cornette JL, Cease KB, Margalit H *et al.* (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *Journal of molecular biology*, **195**, 659–685.

Von Dassow P, John U, Ogata H *et al.* (2014) Life-cycle modification in open oceans accounts for genome variability in a cosmopolitan phytoplankton. *The ISME Journal*, 1–13.

Von Dassow P, Ogata H, Probert I *et al.* (2009) Transcriptome analysis of functional differentiation between haploid and diploid cells of Emiliania huxleyi, a globally significant photosynthetic calcifying cell. *Genome biology*, **10**, R114.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2012) Special features of RAD Sequencing data: implications for genotyping. *Molecular ecology*, n/a–n/a.

Debora Iglesias-Rodriguez M, Schofield OM, Batley J, Medlin LK, Hayes PK (2006) Intraspecific genetic diversity in the marine coccolithophore Emiliania huxleyi (Prymnesiophyceae): The use of microsatellite analysis in marine phytoplankton population studies. *Journal of Phycology*, **42**, 526–536.

Demšar J (2006) Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research*, **7**, 1–30.

Earl D a., vonHoldt BM (2012) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16196–16200.

Engel A, Zondervan I, Beaufort L *et al.* (2005) Testing the direct effect of CO 2 concentration on a bloom of the coccolithophorid Emiliania huxleyi in mesocosm experiments Marie-Dominique Pizay. *Limnology and Oceanography*, **50**, 493–507.

Eppley RW (1972) Temperature and phytoplankton growth in the sea. *Fishery Bulletin*, **70**, 1063 – 1085.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE : a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Falkowski, Paul G., Raven J a. (2007) An Introduction to Photosynthesis in Aquatic Systems. *Aquatic Photosynthesis*, 1–44.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574–578.

Felsenstein J (1974) The evolutionary advantage of recombination. *Genetics*, **78**, 737–756.

Flot J-F, Hespeels B, Li X *et al.* (2013) Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. *Nature*, **500**, 453–7.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*, **180**, 977–993.

Frada M, Probert I, Allen MJ, Wilson WH, de Vargas C (2008) The "Cheshire Cat" escape strategy of the coccolithophore Emiliania huxleyi in response to viral infection. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 1–6.

Freeland JR, Kirk H, Petersen SD (2011) *Molecular Ecology*. Wiley-Blackwell.

Gautier M, Gharbi K, Cezard T *et al.* (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.

Green JC, Course PA, Tarran GA (1996) The life cycle of Emiliania huxleyi: A brief review and a study of relative ploidy levels analysed by flow cytometry. *Journal of Marine Systems*, **9**, 33–44.

Guillard RRL (1973) Methods for microflagellates and nannoplankton. In: *Stein JR (ed) Handbook of phycological methods. Culture methods and growth measurements.*, pp. 69–85. Cambridge University Press.

Guillard RRL, Ryther J (1962) Studies on marine planktonic diatoms I. Cyclotella nana Hustedt and Detonula confervacea Cleve. *Canadian Journal of Microbiology*, **8**, 229–239.

Hagino K, Bendif EM, Young JR *et al.* (2011) New evidence for morphological and genetic variation in the cosmopolitan coccolithophore Emiliania huxleyi (Prymnesiophyceae) from the cox1b-ATP4 genes. *Journal of Phycology*, **47**, 1164–1176.

Hagino K, Okada H (2004) Floral response of coccolithophores to progressive oligotrophication in the South Equatorial Current, Pacific Ocean. *Global Environmental Change in the Ocean and on Land*, 121–132.

Hallegraeff GM (2010) Ocean climate change, phytoplankton community responses, and harmful algal blooms: A formidable predictive challenge. *Journal of Phycology*, **46**, 220–235.

Halpern BS, Walbridge S, Selkoe K a *et al.* (2008) A global map of human impact on marine ecosystems. *Science (New York, N.Y.)*, **319**, 948–52.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. *Nature reviews. Genetics*, **10**, 551–564.

Helleday T (2003) Pathways for mitotic homologous recombination in mammalian cells. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, **532**, 103–115.

Hoffmann AA, Sgro CM (2011) Climate change and evolutionary adaptation.

Hohenlohe P a., Bassham S, Currey M, Cresko W a. (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 395–408.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. *PLoS Genet*, **6**, e1000862.

Hohenlohe PA, Day MD, Amish SJ *et al.* (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. *Molecular Ecology*, **22**, 3002–3013.

Houdan A, Probert I, van Lenning K, Lefebvre S (2005) Comparison of photosynthetic responses in diploid and haploid life-cycle phases of Emiliania huxleyi (Prymnesiophyceae). *Marine Ecology Progress Series*, **292**, 139–146.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

Hughes JS, Otto SP (1999) Ecology and the Evolution of Biphasic Life Cycles. *The American Naturalist*, **154**, 306–320.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254–267.

Iglesias-Rodriguez MD, Brown CW, Doney SC *et al.* (2002) Representing key phytoplankton functional groups in ocean carbon cycle models: Coccolithophorids.

Illumina (2012) *TruSeq DNA PCR-Free Sample Preparation Guide*. San Diego, California, U.S.A.

Illumina (2015a) *NextSeq 500 System - Denature and Dilute Libraries Guide*. San Diego, California, U.S.A.

Illumina (2015b) An Introduction to Next-Generation Sequencing Technology Table of Contents.

Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 10667–72.

Jablonka E, Raz G (2009) Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *The Quarterly review of biology*, **84**, 131–176.

Jaeckisch N, Yang I, Wohlrab S *et al.* (2011) Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate Alexandrium ostenfeldii. *PLoS ONE*, **6**.

Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, **33 Suppl**, 245–254.

Kaffes A, Thoms S, Trimborn S *et al.* (2010) Carbon and nitrogen fluxes in the marine coccolithophore Emiliania huxleyi grown under different nitrate concentrations. *Journal of Experimental Marine Biology and Ecology*, **393**, 1–8.

Kampstra P (2008) Beanplot: A Boxplot Alternative for Visual\nComparison of Distributions. *Journal of Statistical Software*, **28**, 1–9.

Kegel JU, John U, Valentin K, Frickenhaus S (2013) Genome variations associated with viral susceptibility and calcification in Emiliania huxleyi. *PloS one*, **8**, e80684.

Keller MD, Seluin RC, Claus W, Guillard RRL (1987) Media for the culture of oceanic ultraphytoplankton. *Journal of Phycology*, **23**, 633–638.

Kingsolver JG (2009) The well-temperatured biologist. (American Society of Naturalists Presidential Address). *The American naturalist*, **174**, 755–768.

Klaveness D (1972) Coccolithus huxleyi (Lohm.) Kamptn II. The flagellate cell, aberrant cell types, vegetative propagation and life cycles. *British Phycological Journal*, **7**, 309–318.

Kottmeier DM, Rokitta SD, Tortell PD, Rost B (2014) Strong shift from HCO3 (-) to CO 2 uptake in Emiliania huxleyi with acidification: new approach unravels acclimation versus short-term pH effects. *Photosynthesis research*, **121**, 265–75.

Kump LR, Bralower TJ, Ridgwell A (2009) Ocean Acidification in Deep Time. *Oceanography*, **22**, 94 – 107.

Langer G, Geisen M, Baumann K-H *et al.* (2006) Species-specific responses of calcifying algae to changing seawater carbonate chemistry. *Geochemistry, Geophysics, Geosystems*, **7**, n/a–n/a.

Langer G, Nehrke G, Probert I, Ly J, Ziveri P (2009) Strain-specific responses of *Emiliania huxleyi* to changing seawater carbonate chemistry. *Biogeosciences*, **6**, 2637–2646.

Langer G, Probert I, Ziveri P (2011) The morphological response of Emiliania huxleyi to seawater carbonate chemistry changes : an inter-strain comparison. *Journal of Nannoplankton Research*, **32**, 29–34.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Meth*, **9**, 357–359.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

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

life technologies (2015) *Qubit ™ dsDNA HS Assay Kits*.

Liu F, Pang SJ (2010) Stress tolerance and antioxidant enzymatic activities in the metabolisms of the reactive oxygen species in two intertidal red algae Grateloupia turuturu and Palmaria palmata. *Journal of Experimental Marine Biology and Ecology*, **382**, 82–87.

Lluch-Cota SE, Hoegh-Guldberg O, Karl D *et al.* (2014) 2014: Cross-chapter box on uncertain trends in major upwelling ecosystem. In: *Climate Change 2014: Impacts, Adaptation and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Field CB, Barros V, Dokken D, et al.), pp. 149–151. Cambridge University Press, Cambridge, United Kingdom and New York, NY USA.

Loebl M, Cockshutt AM, Campbell DA, Finkel Z V. (2010) Physiological basis for high resistance to photoinhibition under nitrogen depletion in Emiliania huxleyi. *Limnology and Oceanography*, **55**, 2150–2160.

Lohbeck KT, Riebesell U, Reusch TBH (2012) Adaptive evolution of a key phytoplankton species to ocean acidification. *Nature Geoscience*, **5**, 346–351.

Lohbeck KT, Riebesell U, Reusch TBH (2014) Gene expression changes in the coccolithophore Emiliania huxleyi after 500 generations of selection to ocean acidification. *Proceedings. Biological sciences / The Royal Society*, **281**.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature reviews. Genetics*, **4**, 981–994.

Lynch M, Bürger R, Butcher D, Gabriel W (1990) The mutational meltdown in asexual populations. *The Journal of heredity*, **84**, 339–344.

Lynch M, Gabriel W (1990) Mutation Load and the Survival of Small Populations. *Evolution*, **44**, 1725–1737.

Macher J-N, Rozenberg A, Pauls SU *et al.* (2015) Assessing the phylogeographic history of the montane caddisfly Thremma gallicum using mitochondrial and restriction-site-associated DNA (RAD) markers. *Ecology and Evolution*, **5**, 648–662.

Macherey-Nagel (2014a) *Genomic DNA from plant - User manual (NucleoSpin® Plant II, -Midi, -Maxi).*

Macherey-Nagel (2014b) *NucleoSpin® Gel and PCR Clean-up - User manual.* Düren.

MacKinder L, Wheeler G, Schroeder D *et al.* (2011) Expression of biomineralization-related ion transport genes in Emiliania huxleyi. *Environmental Microbiology*, **13**, 3250–3265.

Mandegar MA, Otto SP (2007) Mitotic recombination counteracts the benefits of genetic segregation. *Proceedings. Biological sciences / The Royal Society*, **274**, 1301–1307.

Masson-Delmotte, V. M, Schulz A, Abe-Ouchi J *et al.* (2013) Information from Paleoclimate Archives. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Gupta AK, Rahimzadeh F, Raynaud D, Wanner H), pp. 383–464. Cambridge University Press, Cambridge, United Kingdom and New York, NY USA.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, **66**, 526–538.

Medlin LK, Barker GL a, Campbell L *et al.* (1996) Genetic characterisation of Emiliania huxleyi (Haptophyta). *Journal of Marine Systems*, **9**, 13–31.

Miller M, Dunham J, Amores a, Cresko W, Johnson E (2007) genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Muller H (1964) The relation of recombination to mutational advance. *Mutation Research*, **1**, 2–9.

Müller MN, Schulz KG, Riebesell U (2010) Effects of long-term high CO_2 exposure on two species of coccolithophores. *Biogeosciences*, **7**, 1109–1116.

Nei M, Kumar S (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

Nielsen R (2005) Molecular signatures of natural selection. *Annual review of genetics*, **39**, 197–218.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. *Nature reviews. Genetics*, **8**, 857–868.

Nielsen R, Hubisz MJ, Hellmann I *et al.* (2009) Darwinian and demographic forces affecting human protein coding genes Darwinian and demographic forces affecting human protein coding genes. , 838–849.

Van Oostende N, Moerdijk-Poortvliet TCW, Boschker HTS, Vyverman W, Sabbe K (2013) Release of dissolved carbohydrates by Emiliania huxleyi and formation of transparent exopolymer particles depend on algal life cycle and bacterial activity. *Environmental Microbiology*, **15**, 1514–1531.

Orr JC, Fabry VJ, Aumont O *et al.* (2005) Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. *Nature*, **437**, 681–6.

Paasche E (2002) A review of the coccolithophorid Emiliania huxleyi (Prymnesiophyceae), with particular reference to growth, coccolith formation, and calcification-photosynthesis interactions. *Phycologia*, **40**, 503–529.

Parmesan C (2006) Ecological and Evolutionary Responses to Recent Climate Change. *Annual Review of Ecology Evolution and Systematics*, **37**, -.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE*, **7**, e37135.

Pörtner H-O, Karl DM, Boyd PW *et al.* (2014) Ocean Systems. In: *Climate Change 2014: Impacts, Adaptation and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Chagne* (eds Drinkwater KF, Polonsky A), pp. 411–484. Cambridge University Press, Cambridge, United Kingdom and New York, NY USA.

Poulton AJ, Adey TR, Balch WM, Holligan PM (2007) Relating coccolithophore calcification rates to phytoplankton community dynamics: Regional differences and implications for carbon export. *Deep-Sea Research Part II: Topical Studies in Oceanography*, **54**, 538–557.

Pritchard JK (2010) Documentation for structure software : Version 2 . 3. , 1–39.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Qiagen (2008) *MinElute® Handbook*. Hilden.

R Core Team R (2015) R: A language and environment for statistical computing.

Raffi I, Backman J, Fornaciari E *et al.* (2006) A review of calcareous nannofossil astrobiochronology encompassing the past 25 million years☆. *Quaternary Science Reviews*, **25**, 3113–3137.

Raven J a., Crawfurd K (2012) Environmental controls on coccolithophore calcification. *Marine Ecology Progress Series*, **470**, 137–166.

Read BA, Kegel J, Klute MJ *et al.* (2013) Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature*, **499**, 209–13.

Richier S, Fiorini S, Kerros M-E, von Dassow P, Gattuso J-P (2011) Response of the calcifying coccolithophore Emiliania huxleyi to low pH/high pCO2: from physiology to molecular level. *Marine biology*, **158**, 551–560.

Riebesell U, Zondervan I, Rost B *et al.* (2000) Reduced calcification of marine plankton in response to increased atmospheric CO 2. , **407**, 364–367.

Rieseberg LH, Raymond O, Rosenthal DM *et al.* (2003) Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. , **301**.

Rokitta SD (2012) Characterization of the life-cycle stages of the coccolithophore Emiliania huxleyi and their responses to Ocean Acidification.

Rokitta SD, Von Dassow P, Rost B, John U (2014) Emiliania huxleyi endures N-limitation with an efficient metabolic budgeting and effective ATP synthesis. *BMC Genomics*, **15**, 1051.

Rokitta SD, John U, Rost B (2012) Ocean acidification affects redox-balance and ion-homeostasis in the life-cycle stages of Emiliania huxleyi. *PloS one*, **7**, e52212.

Rokitta SD, de Nooijer LJ, Trimborn S *et al.* (2011) Transcriptome Analyses Reveal Differential Gene Expression Patterns Between the Life-Cycle Stages of Emiliania Huxleyi (Haptophyta) and Reflect Specialization To Different Ecological Niches1. *Journal of Phycology*, **47**, 829–838.

Rokitta SD, Rost B (2012) Effects of CO2 and their modulation by light in the life-cycle stages of the coccolithophore Emiliania huxleyi. *Limnology and Oceanography*, **57**, 607–618.

Rosenzweig C, Casassa G, Karoly DJ *et al.* (2007) Assessment of observed changes and responses in natural and managed systems. In: *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* (eds Kajfež-Bogataj L, Pretel J, Watkinson A), pp. 79–131. Cambridge University Press, Cambridge, United Kingdom.

Rost B, Riebesell U (2004) Coccolithophores and the biological pump: responses to environmental changes. In: *Coccolithophores: from molecular processes to global impact* (eds Thierstein HR, Young JR), pp. 99–125. Berlin [u.a.].

Rost B, Riebesell U, Burkhardt S (2003) Carbon acquisition of bloom-forming marine phytoplankton. *Limn*, **48**, 55–67.

Rozenberg A (2014) preprocess_ddradtags.pl.

Rozenberg A (2015) double_indexing.pl.

Sachs L (1997) Angewandte Statistik. In: *Angewandte Statistik*, pp. 668–675.

Schield DR, Walsh MR, Card DC *et al.* (2015) EpiRADseq: scalable analysis of genome-wide patterns of methylation using next-generation sequencing. *Methods in Ecology and Evolution*, n/a–n/a.

Schlüter L, Lohbeck KT, Gutowska M a. *et al.* (2014) Adaptation of a globally important coccolithophore to ocean warming and acidification. *Nature Climate Change*, **4**.

Schweyen H, Rozenberg A, Leese F (2014) Detection and Removal of PCR Duplicates in Population Genomic ddRAD Studies by Addition of a Degenerate Base Region (DBR) in Sequencing Adapters. , 146–160.

Sciandra A, Harlay J, Lefèvre D *et al.* (2003) Response of coccolithophorid Emiliania huxleyi to elevated partial pressure of CO2 under nitrogen limitation. *Marine Ecology Progress Series*, **261**, 111–122.

Silva N, Rojas N, Fedele A (2009) Water masses in the Humboldt Current System: Properties, distribution, and the nitrate deficit as a chemical water mass tracer for Equatorial

Subsurface Water off Chile. *Deep-Sea Research Part II: Topical Studies in Oceanography*, **56**, 992–1008.

Smayda TJ, Reynolds CS (2003) Strategies of marine dinoflagellate survival and some rules of assembly. *Structuring Factors of Shallow Marine Coastal Communities, Part II*, **49**, 95–106.

Steinacher M, Joos F, Frolicher T *et al.* (2010) Projected 21st century decrease in marine productivity: a multi-model analysis. *Biogeosciences*, **7**, 979–1005.

Strasburg JL, Sherman N a., Wright KM *et al.* (2012) What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 364–373.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, **30**, 2725–2729.

Thomas MK, Kremer CT, Klausmeier C a., Litchman E (2012) A Global Pattern of Thermal Adaptation in Marine Phytoplankton. *Science*, **338**, 1085–1088.

Tyrrell T, Merico A (2004) Emiliania huxleyi: Bloom observations and the conditions that induce them. In: *Coccolithophores: from molecular processes to global impact* (eds Thierstein H, Young J), pp. 75–97. Springer.

De Vargas C, Aubry M-P, Probert I, Young J (2007) Origin and evolution of coccolithophores: From coastal hunters to oceanic farmers. In: *Evolution of Primary Producers in the Sea*, pp. 251 –285.

Walther G-R (2010) Community and ecosystem responses to recent climate change. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 2019–2024.

Walther G-R, Post E, Convey P *et al.* (2002) Ecological responses to recent climate change. *Nature*, **416**, 389–95.

Weir B, Cockerham C (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Wickham H (2009) *ggplot2: elegant graphics for data analysis*. Springer, New York.

Wickham H (2011a) The Split-Apply-Combine Strategy for Data. *Journal of Statistical Software*, **40**, 1–29.

Wickham H (2011b) The Split-Apply-Combine Strategy for Data. *Journal of Statistical Software*, **40**, 1–29.

Wilson WH, Tarran G a., Schroeder D *et al.* (2002) Isolation of viruses responsible for the demise of an Emiliania huxleyi bloom in the English Channel.

Wright S (1965) The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution*, **19**, 395–420.

Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

Wu TD, Watanabe CK (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Xu S, Omilian AR, Cristescu ME (2011) High rate of large-scale hemizygous deletions in asexually propagating Daphnia: Implications for the evolution of sex. *Molecular Biology and Evolution*, **28**, 335–342.

Yaish MW, Colasanti J, Rothstein SJ (2011) The role of epigenetic processes in controlling flowering time in plants exposed to stress. *Journal of Experimental Botany*, **62**, 3727–3735.

Yoshida T, Jones LE, Ellner SP, Fussmann GF, Hairston NG (2003) Rapid evolution drives ecological dynamics in a predator-prey system. *Nature*, **424**, 303–306.

Young J (2003) A guide to extant coccolithophore taxonomy. *Journal of Nannoplankton Research*, 132.

Young JR, Westbroek P (1991) Genotypic variation in the coccolithophorid species Emiliania huxleyi. *Marine Micropaleontology*, **18**, 5–23.

Zhang G, Fang X, Guo X *et al.* (2012) The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, **490**, 49–54.

Zhou X, Xia Y, Ren X *et al.* (2014) Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *BMC genomics*, **15**, 351.

Zondervan I, Rost B, Riebesell U (2002) Effect of CO2 concentration on the PIC/POC ratio in the coccolithophore Emiliania huxleyi grown under light-limiting conditions and different daylengths. *Journal of Experimental Marine Biology and Ecology*, **272**, 55–70.

Zondervan I, Zeebe RE, Rost B, Riebesell U (2001) Decreasing marine biogenic calcification: A negative feedback on rising atmospheric pCO2. *Global Biogeochemical cycles*, **15**, 507–516.

6. SUPPLEMENT

Table S I List of strains used in this study with Population assignment, Isolation year as well as the GPS coordinates of isolation origin.

Strain	Population	Isolation year	Latitude [degrees north]	Longitude [degrees east]
CHC351	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC352	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC354	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC355	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC356	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC358	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC350	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC360	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC361	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC362	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC366	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC367	Upwelling Chile	2012	-30.2469444	-71.6938889
CHC344	normal	2012	-41.924883	-74.038211
CHC307	normal	2011	-33.7522222	-78.8458333
CHC323	normal	2011	-33.7522222	-78.8458333
CHC325	normal	2011	-33.7522222	-78.8458333
CHC297	normal	2011	-33.7522222	-78.8458333
CHC314	normal	2011	-33.6405556	-78.9994444
CHC337	normal	2011	-33.6405556	-78.9994444
CHC298	asexual	2011	-33.7522222	-78.8458333
CHC299	asexual	2011	-33.7522222	-78.8458333
CHC303	asexual	2011	-33.7522222	-78.8458333
CHC315	asexual	2011	-33.6405556	-78.9994444
CHC333	asexual	2011	-33.7522222	-78.8458333
CHC445	normal	2013	-16.7488889	-85.998
CHC446	normal	2013	-16.7488889	-85.998
CHC447	normal	2013	-16.7488889	-85.998
CHC448	normal	2013	-16.7488889	-85.998
CHC449	normal	2013	-16.7488889	-85.998
CHC450	normal	2013	-16.7488889	-85.998
CHC428	asexual	2013	-16.7488889	-85.998
CHC440	asexual	2013	-16.7488889	-85.998
CHC451	asexual	2013	-20.7688889	-70.659
CHC452	asexual	2013	-20.7688889	-70.659
CHC454	asexual	2013	-20.7688889	-70.659
CHC457	asexual	2013	-20.7688889	-70.659
CHC459	asexual	2013	-20.7688889	-70.659
CHC461	asexual	2013	-20.7688889	-70.659
CHC518	asexual	2013	-20.7688889	-70.659
Peru 1	Upwelling Peru	2014	-14.25	-76.8
Peru 4	Upwelling Peru	2014	-12	-77.833333
Peru 5	Upwelling Peru	2014	-14.116667	-76.5
Peru 11	Upwelling Peru	2014	-12	-77.833333
Peru 12	Upwelling Peru	2014	-12	-77.833333
Peru 13	Upwelling Peru	2014	-12	-77.833333

Table S I – proceeded: List of strains used in this study with Population assignment, Isolation year as well as the GPS coordinates of isolation origin.

Strain	Population	Isolation year	Latitude [degrees north]	Longitude [degrees east]
Peru 16	Upwelling Peru	2014	-12	-77.833333
Peru 18	Upwelling Peru	2014	-12	-77.833333
Peru 19	Upwelling Peru	2014	-12	-77.833333
Peru 20	Upwelling Peru	2014	-12	-77.833333
Peru 22	Upwelling Peru	2014	-12	-77.833333
Peru 24	Upwelling Peru	2014	-12	-77.833333
Peru 27	Upwelling Peru	2014	-12	-77.5
Peru 28	Upwelling Peru	2014	-12	-77.5
Peru 29	Upwelling Peru	2014	-12	-77.5
Peru 35	Upwelling Peru	2014	-14.416667	-77
Peru 36	Upwelling Peru	2014	-14.416667	-77

Table S II: List of the upper part of P5 adapters with Name, sequence of read 1 sequencing primer (Rd1SP), P5-TrueSeq Indices (TruSeq), Insert of 1-3 nucleotides, Overhang and the complete sequence

P5 upper part

Name	Rd1SP	TruSeq	Insert	Overhang	Complete Sequence
P_OL_AD001	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ATCACG	–	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATCACGTGCA
P_OL_AD002	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGATGT	GAC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATGTGACTGCA
P_OL_AD003	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TTAGGC	AC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTTAGGCACTGCA
P_OL_AD004	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TGACCA	C	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGACCACTGCA
P_OL_AD005	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACAGTG	AC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGTGACTGCA
P_OL_AD006	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GCCAAT	–	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCAATTGCA
P_OL_AD007	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CAGATC	–	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGATCTGCA
P_OL_AD008	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACTTGA	GAC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTTGAGACTGCA
P_OL_AD009	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GATCAG	GAC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATCAGGACTGCA
P_OL_AD010	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	TAGCTT	AC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGCTTACTGCA
P_OL_AD011	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GGCTAC	C	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCTACCTGCA
P_OL_AD012	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CTTGTA	GAC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTGTAGACTGCA
P_OL_AD013	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGTCAA	C	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTCAACTGCA
P_OL_AD014	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	AGTTCC	–	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTTCCTGCA
P_OL_AD015	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ATGTCA	GAC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATGTCAGACTGCA
P_OL_AD016	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CCGTCC	AC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCGTCCACTGCA
P_OL_AD018	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GTCCGC	AC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCCGCACTGCA
P_OL_AD019	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GTGAAA	C	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGAAACTGCA
P_OL_AD020	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GTGGCC	GAC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGGCCGACTGCA
P_OL_AD021	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GTTTCG	–	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTTTCGTGCA
P_OL_AD022	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	CGTACG	–	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTACGTGCA
P_OL_AD023	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	GAGTGG	AC	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAGTGGACTGCA
P_OL_AD025	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ACTGAT	C	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGATCTGCA
P_OL_AD027	ACACTCTTTCCCTACACGACGCTCTTCCGATCT	ATTCCT	C	TGCA	ACACTCTTTCCCTACACGACGCTCTTCCGATCTATTCCTCTGCA

Table S III: List of the lower part of P5 adapters with Name, Overhang, Insert of 1-3 nucleotides,P5-TruSeq Indices (TruSeq), complementary read 1 sequencing primer sequence (Rd1SP) and the complete sequence

P5 – lower part

Name	Overhang	Insert	TruSeq	Rd1SP	Complete Sequence
K_OL_AD001	–	–	CGTGAT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	CGTGATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD002	–	GTC	ACATCG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTCACATCGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD003	–	GT	GCCTAA	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTGCCTAAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD004	–	G	TGGTCA	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTGGTCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD005	–	GT	CACTGT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTCACTGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD006	–	–	ATTGGC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	ATTGGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD007	–	–	GATCTG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GATCTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD008	–	GTC	TCAAGT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTCTCAAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD009	–	GTC	CTGATC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTCCTGATCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD010	–	GT	AAGCTA	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTAAGCTAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD011	–	G	GTAGCC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GGTAGCCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD012	–	GTC	TACAAG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTCTACAAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD013	–	G	TTGACT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTTGACTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD014	–	–	GGAACT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GGAACTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD015	–	GTC	TGACAT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTCTGACATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD016	–	GT	GGACGG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTGGACGGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD018	–	GT	GCGGAC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTGCGGACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD019	–	G	TTTCAC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTTTCACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD020	–	GTC	GGCCAC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTCGGCCACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD021	–	–	CGAAAC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	CGAAACAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD022	–	–	CGTACG	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	CGTACGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD023	–	GT	CCACTC	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GTCCACTCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD025	–	G	ATCAGT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GATCAGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
K_OL_AD027	–	G	AGGAAT	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	GAGGAATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

Table S IV List of P7 adapter sequences containing the overhang, Inline Barcodes, degenerate base regions (DBR), Inserts, read 2 sequencing primer sequences (Rd2SP), and forked end for the upper part of the P7 adapter and P7 TruSeq Indices for paired end sequencening in the lower part of the P7 adapter.

P7 upper part

	Over-hang	Inline-Barcode	DBR	Insert	Rd2SP	Forked End	Complete Sequence
P7_K_01_DBR_CG	CG	CGTGAT	CGTCCIINNNNNNN		AGATCGGAAGAGCACACGTCTGAACT	AGGAGACT	CGCGTGATCGTCCIINNNNNNNAGATCGGAAGAGCACACGTCTGAACTAGGAGACT
P7_K_08_DBR_CG	CG	TCAAGT	CGTCCIINNNNNNN	C	AGATCGGAAGAGCACACGTCTGAACT	AGGAGACT	CGTCAAGTCGTCCIINNNNNNNCAGATCGGAAGAGCACACGTCTGAACTAGGAGACT
P7_K_10_DBR_CG	CG	AAGCTA	CGTCCIINNNNNNN	AT	AGATCGGAAGAGCACACGTCTGAACT	AGGAGACT	CGAAGCTACGTCCIINNNNNNNATAGATCGGAAGAGCACACGTCTGAACTAGGAGACT
P7_K_11_DBR_CG	CG	GTAGCC	CGTCCIINNNNNNN	TCG	AGATCGGAAGAGCACACGTCTGAACT	AGGAGACT	CGGTAGCCCGTCCIINNNNNNNTCGAGATCGGAAGAGCACACGTCTGAACTAGGAGACT

P7 lower part

|	FlowCell P7	TruSeq	Rd2SP	Insert	DBR	Inline Barcode	Complete Sequence
P7_F_01_DBR_CG	CAAGCAGAAGACGGCATACGAGAT	CGTGAT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT		NNNNNNNMMGGACG	ATCACG	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNNMMGGACGATCACG
P7_F_08_DBR_CG	CAAGCAGAAGACGGCATACGAGAT	TCAAGT	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	G	NNNNNNNMMGGACG	ACTTGA	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGNNNNNNNMMGGACGACTTGA
P7_F_10_DBR_CG	CAAGCAGAAGACGGCATACGAGAT	AAGCTA	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	AT	NNNNNNNMMGGACG	TAGCTT	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATNNNNNNNMMGGACGTAGCTT
P7_F_11_DBR_CG	CAAGCAGAAGACGGCATACGAGAT	GTAGCC	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	CGA	NNNNNNNMMGGACG	GGCTAC	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGANNNNNNNMMGGACGGGCTAC
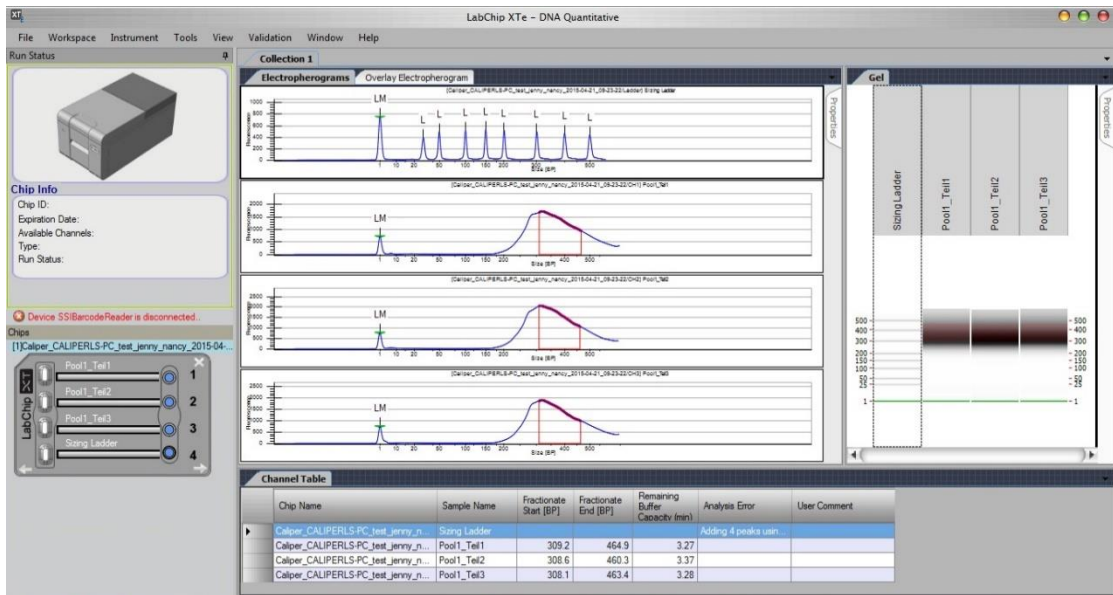
Figure S I Output of final library pool size selection step with LabChip XTe
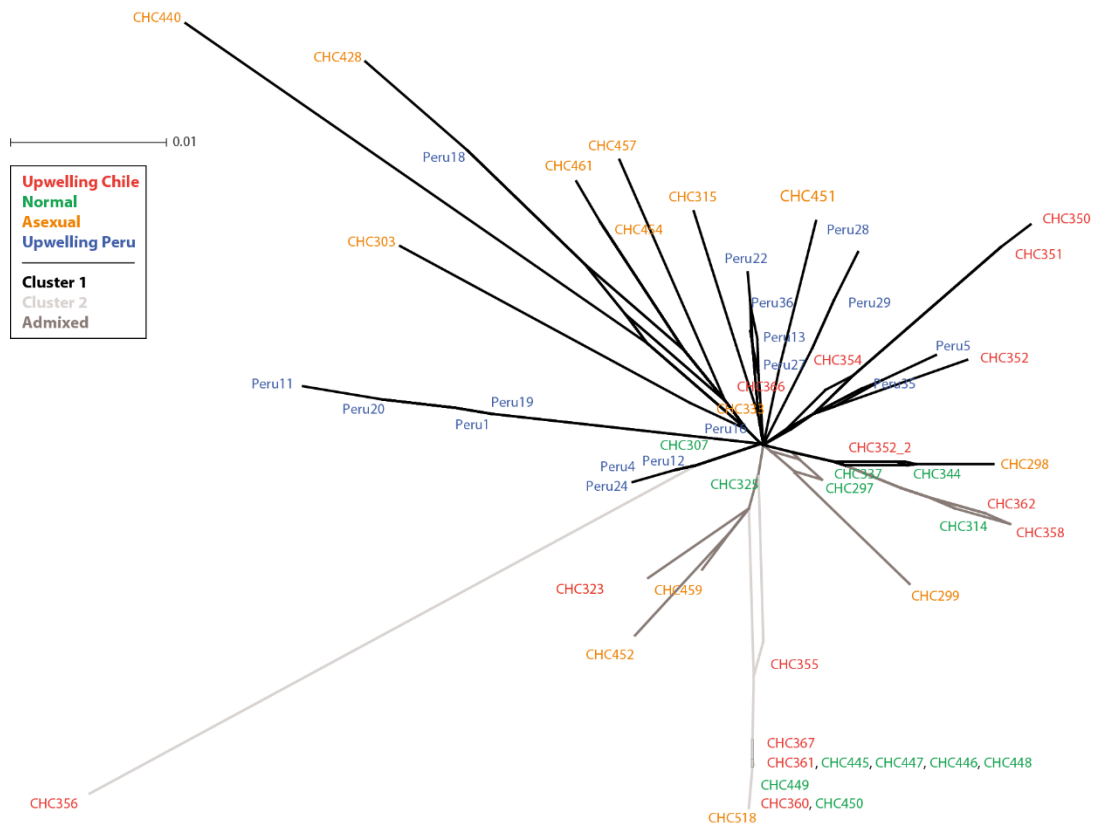

Figure S II Neighbor-Net networks inferred with ddRAD loci of the CHC428 dataset containing minimum 1 variable SNP per locus and are present in minimum 12 individuals of one population. Neighbor networks were inferred using the software SplitsTree v. 4.13.1 (Huson & Bryant 2006) using the default parameters and displaying only branches with a bootstrap value above 80.

Table S V List of supplementary files on the attached CD

Files	Location/Folder
Macherey-Nagel - 2014 - Genomic DNA from plant - User manual	Benchtop protocols
Qiagen - 2008 - MinElute® Handbook	Benchtop protocols
Macherey-Nagel 2014 – PCR clean-up Gel extraction – User Manual	Benchtop protocols
NextSeq 500 System - Denature and Dilute Libraries Guide.	Benchtop protocols
life technologies - 2015 - Qubit ™ dsDNA HS Assay Kits	Benchtop protocols
Agencourt CleanSeq Dye Terminator Removal - Manual	Benchtop protocols
Epigenetic analysis - Input files and R-script	R_and_Tables → Epigenetic
FST_plots - Input files and R-script	R_and_Tables → FST_plots
Gene diversity Analysis - Input files and R-script	R_and_Tables → Gene_diversity
Genomic erosion analysis - Input files and R-script	R_and_Tables → Genomic erosion
Table with Unique Loci	R_and_Tables → Genomic erosion → Analysis
SNP_analyses - Input files and R-script	R_and_Tables → SNP_analyses
ddRAD benchtop protocol	Benchtop_protocols
Double_indexing.pl manual	Software manuals
Master.bash manual	Software manuals
Double_indexing perl script	Software
Master bash script	Software
Preprocess_ddradtags script	Software

Table S V List of supplementary files on the attached CD