Multi-level genetic variation and somatic genetic drift in the model seagrass species, eelgrass (*Zostera marina* L.)

Dissertation
zur Erlangung des akademischen Grades

Doctor rerum naturalium

der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Lei Yu

Kiel, März 2022

Erster Gutachter: Prof. Thorsten Reusch
Zweite Gutachterin: Prof. Tal Dagan
Tag der Disputation: 03.05.2022

Summary

Clonal species, such as seagrass *Zostera marina*, have two levels of "population", an asexual population of ramets belonging to a single clone/genet, which is nested into the population of clones/genets, the "classical", sexually reproducing population. Although earlier theoretical and conceptual work has repeatedly suggested that somatically generated variation plays an important role in modular species, population genetics of such species has been focused on the "classical" germline genetic variation. This thesis extends the knowledge on the multi-level genetic variation in the model seagrass species *Z. marina* using whole-genome resequencing.

In Chapter 1, I studied the somatic genetic variation (SoGV) in a *Z. marina* meadow tracing back to one single sexual event. Twenty-four leaf shoots (=ramets) were collected along an inshore-offshore transect by SCUBA from a *Z. marina* meadow located in the Archipelago Sea, southern Finland. High-coverage whole-genome resequencing data (81x) were produced for each ramet, and mapped against a high-quality reference genome. The genome-wide SNPs confirmed that the collected ramets were derived from one single sexual event, and showed 7,054 SoGV with different fixed genotypes among the 24 ramets (i.e., fixed SoGV). This indicated that SoGV is an important source of genetic variation for clonal species. Ultra-deep sequencing (1370x) of 3 ramets showed very high levels of SoGV with different genotypes among cells within ramet (i.e., mosaic SoGV). My results also showed signatures of selection on both intra-ramet and inter-ramet levels. Moreover, it has been neglected that a mechanism is needed to turn an original mosaic SoGV to a fixed SoGV. I proposed that a stochastic process during the formation of new ramets/modules, *somatic genetic drift*, can lead to fixation of SoGV in the newly formed ramets.

In chapter 2, I studied population genomics of *Z. marina*, based on the worldwide distribution of the "classical" germline genetic variation. A total of 190 ramets were collected from 16 populations, spanning both sides of the Pacific and the Atlantic, respectively. High-coverage whole-genome resequencing data (53.73x) were produced for each ramet, and mapped against a chromosome-level reference genome. Using the estimated divergence time between *Z. marina* and an outgroup as the calibration point, and the extended multi-species coalescent (MSC) method based on putatively neutral and non-linked SNPs, I was able to construct a phylogenetic tree and estimate the time of the major divergence events. The time-calibrated phylogeny showed frequent and swift colonization waves across entire ocean basins, which may explain a lack of speciation events across the seagrasses in general. The evolutionary history in the Pacific was complex and involved three trans-Pacific dispersal events. The genetic input from multiple dispersal events in Willapa Bay (Washington State, USA) and Bodega Bay (California, USA) was probably the reason for the admixture there. Using demographic reconstruction based on Multiple Sequentially Markovian Coalescent (MSMC), I also demonstrated the influence of historical glacial periods on *Z. marina* populations, and identified the refugia during the Last Glacial Maximum (LGM). The Atlantic clade and the Mediterranean clade had their own refugia

during the Last Glacial Maximum (LGM), respectively. The current Northeast Atlantic was colonized by a recent trans-Atlantic dispersal event from the Northwest Atlantic at 15.8 kya (95% HPD: 19.7-12.3 kya). All these findings are completely novel for marine plants, and such a study on phylogeography is also a prerequisite for making inferences on selection.

For the study of clonal species, it is necessary to accurately distinguish ramet vs. genet. The currently available methods cannot work on the case in chapter 1 where all collected ramets belong to one single clone. In chapter 3, I proposed the concept of identity by heterozygosity (IBH), meaning that two diploid genomes are identically heterozygous at some genetic markers such as SNPs (single-nucleotide polymorphisms). Based on IBH, I developed IBH similarity (IBH_s), a novel measure of genetic similarity between two diploid genomes. Two samples with $IBH_s > 0.95$ were considered members of the same clone, while two samples with $IBH_s <= 0.95$ were considered to represent different clones. I expect that my method will be increasingly applied in situations where entire sites are comprised of a few or only one clonal lineage, in species ranging from fungi, clonal invertebrates such as corals, to plants.

In summary, two levels of SoGV, namely among cell populations, and among ramets/modules, are nested within “classical” germline genetic variation in clonal species, owing to somatic mutation and somatic genetic drift. This thesis opens up new avenues of studying multi-level genetic variation by using clonal species as model. Further modeling and theoretical work based on clonal species will broaden the current population genetics theory.

Zusammenfassung

Klonale Arten wie das Seegras *Zostera marina* haben zwei "Populationsebenen", die ungeschlechtliche Population von Ramets, die zu einem einzigen Klon/Genet gehören, und die Population von Klonen/Genets, die "klassische", sich sexuell fortpflanzende Population. Obwohl frühere theoretische und konzeptionelle Arbeiten wiederholt darauf hingewiesen haben, dass die somatisch erzeugte Variation bei modularen Arten eine wichtige Rolle spielt, hat sich die Populationsgenetik solcher Arten auf die "klassische" genetische Keimbahnvariation konzentriert. In dieser Arbeit wird das Wissen über die mehrstufige genetische Variation in der Modell-Seegrasart *Z. marina* mit Hilfe von Ganzgenom-Resequenzierungen erweitert.

In Kapitel 1 habe ich die somatische genetische Variation (SoGV) in einer *Z. marina*-Wiese untersucht, die auf ein einziges sexuelles Ereignis zurückgeht. Vierundzwanzig Blattsprossen (=Ramets) wurden entlang eines Küsten-Transekts von einer Z. marina-Wiese im Schärenmeer, Südfinnland, durch Tauchen gesammelt. Für jede Ramete wurden flächendeckende Ganzgenom-Resequenzierungsdaten (81x) erstellt und gegen ein hochwertiges Referenzgenom abgeglichen. Die genomweiten SNPs bestätigten, dass die gesammelten Rambler von einem einzigen sexuellen Ereignis abstammen, und zeigten 7.054 SoGV mit verschiedenen festen Genotypen unter den 24 Ramblerpflanzen (d. h. feste SoGV). Dies deutet darauf hin, dass SoGV eine wichtige Quelle der genetischen Variation bei klonalen Arten ist. Die ultratiefe Sequenzierung (1370x) von 3 Rambussen zeigte eine sehr hohe Konzentration von SoGV mit unterschiedlichen Genotypen zwischen den Zellen innerhalb des Rambusses (d. h. Mosaik-SoGV). Meine Ergebnisse zeigten auch Anzeichen von Selektion sowohl innerhalb als auch zwischen den Rämmern. Außerdem wurde vernachlässigt, dass ein Mechanismus erforderlich ist, um einen ursprünglichen Mosaik-SoGV in einen festen SoGV zu verwandeln. Ich habe vorgeschlagen, dass ein stochastischer Prozess während der Bildung neuer Rambets/Module, die somatische genetische Drift, zur Fixierung von SoGV in den neu gebildeten Rambets führen kann.

In Kapitel 2 untersuchte ich die Populationsgenomik von *Z. marina* auf der Grundlage der weltweiten Verteilung der "klassischen" genetischen Keimbahnvariation. Insgesamt wurden 190 Verzweigungen aus 16 Populationen auf beiden Seiten des Pazifiks bzw. des Atlantiks gesammelt. Für jeden Ramet wurden flächendeckende Ganzgenom-Resequenzierungsdaten (53,73x) erstellt und gegen ein Referenzgenom auf Chromosomenebene abgeglichen. Unter Verwendung der geschätzten Divergenzzeit zwischen Z. marina und einer Außengruppe als Kalibrierungspunkt und der erweiterten Multi-Spezies-Koaleszenz-Methode (MSC), die auf vermeintlich neutralen und nicht verknüpften SNPs basiert, konnte ich einen phylogenetischen Baum konstruieren und den Zeitpunkt der wichtigsten Divergenzereignisse schätzen. Die zeitkalibrierte Phylogenie zeigte häufige und schnelle Kolonisierungswellen über ganze Ozeanbecken hinweg, was das Fehlen von Speziationsereignissen bei den Seegräsern im

Allgemeinen erklären könnte. Die Evolutionsgeschichte im Pazifik war komplex und umfasste drei trans-pazifische Ausbreitungsereignisse. Der genetische Input aus mehreren Ausbreitungsereignissen in Willapa Bay (Washington State, USA) und Bodega Bay (Kalifornien, USA) war wahrscheinlich der Grund für die dortige Vermischung. Mit Hilfe der demografischen Rekonstruktion auf der Grundlage des Multiple Sequential Markovian Coalescent (MSMC) konnte ich auch den Einfluss historischer Eiszeiten auf die Populationen von *Z. marina* nachweisen und die Refugien während des letzten glazialen Maximums (LGM) identifizieren. Die atlantische Gruppe und die mediterrane Gruppe hatten während des letzten glazialen Maximums (LGM) jeweils ihre eigenen Refugien. Der heutige Nordostatlantik wurde durch eine transatlantische Ausbreitung aus dem Nordwestatlantik um 15,8 kya (95% HPD: 19,7-12,3 kya) kolonisiert. All diese Ergebnisse sind für Meerespflanzen völlig neu, und eine solche Studie zur Phylogeographie ist auch eine Voraussetzung für Rückschlüsse auf die Selektion.

Für die Untersuchung klonaler Arten ist eine genaue Unterscheidung zwischen Ramet und Genet erforderlich. Die derzeit verfügbaren Methoden können nicht auf den in Kapitel 1 beschriebenen Fall angewendet werden, bei dem alle gesammelten Rambets zu einem einzigen Klon gehören. In Kapitel 3 schlug ich das Konzept der Identität durch Heterozygotie (IBH) vor, was bedeutet, dass zwei diploide Genome bei einigen genetischen Markern wie SNPs (Single-Nucleotide Polymorphisms) identisch heterozygot sind. Auf der Grundlage der IBH habe ich die IBH-Ähnlichkeit (IBHs) entwickelt, ein neues Maß für die genetische Ähnlichkeit zwischen zwei diploiden Genomen. Zwei Proben mit IBH_s >0,95 wurden als Mitglieder desselben Klons betrachtet, während zwei Proben mit IBH_s <=0,95 als unterschiedliche Klone angesehen wurden. Ich erwarte, dass meine Methode zunehmend in Situationen angewendet wird, in denen ganze Standorte aus wenigen oder nur einer klonalen Linie bestehen, und zwar bei Arten wie Pilzen, klonalen wirbellosen Tieren wie Korallen und Pflanzen.

Zusammenfassend lässt sich sagen, dass zwei Ebenen der SoGV, nämlich die zwischen Zellpopulationen und zwischen Rameten/Modulen, innerhalb der "klassischen" genetischen Keimbahnvariation bei klonalen Arten aufgrund von somatischer Mutation und somatischer genetischer Drift verschachtelt sind. Diese Arbeit eröffnet neue Wege zur Untersuchung der genetischen Variation auf mehreren Ebenen, indem klonale Arten als Modell verwendet werden. Weitere Modellierungen und theoretische Arbeiten auf der Grundlage klonaler Arten werden die derzeitige Theorie der Populationsgenetik erweitern.

Table of Contents

Introduction

Framework of this dissertation

Modular species are composed of iterative units of similar parts (modules) that either stay together (e.g., tree) or become independent (clonal species). When independent, the modules are called ramets (Harper & White, 1974), a terminology that is followed throughout my thesis. Seagrasses belong to modular or clonal plant species, and seagrass meadows are composed of iterative shoots (=ramets). While the overwhelming literature on seagrasses deals with their ecology, the ecosystem services they provide, or threats they are imposed to, this thesis takes a look at their multi-level genetic variation.

Under sexual reproduction, segregation of homologous chromosomes and mutation during the process of meiosis lead to genetic variation segregating among gametes, upon which successful fertilization produces an almost infinite number of potentially different zygotes (i.e., germline genetic variation). During the development of the zygote, somatic mutation leads to genetic variation among somatic cell lineages in multicellular life, which is usually selectively neutral. If somatic genetic variation (hereafter SoGV) is under positive selection, however, it is most often detrimental such as the proliferation of defective cell lineages producing cancer tumors. Here, I also study whether there is an additional intermediate level of genetic variation in clonal species that can distinguish ramets derived from one single sexual event (i.e., clonemates), using seagrass as a model. In particular, I am interested whether there is a mechanism that can segregate genetic mosaicism into different fixed genotypes among clonemates.

Eelgrass *Zostera marina* is a model seagrass species. With this doctoral thesis, I extend the knowledge on the molecular evolution of *Z. marina* into four directions (i) by studying entire genomes rather than limited number of markers (ii) by studying SoGV among cell populations within a ramet (iii) by studying the existence and emergence of SoGV among ramets tracing back to the same zygote (iv) by studying genetic variation among populations spanning the entire worldwide distribution range in order to reconstruct demography and (re)colonization history.

Marine evolutionary processes and model species

More than 70% of the Earth's surface is covered by ocean (Ribeiro et al., 2017). Many marine species have a pelagic larvae stage or widely dispersing seeds or spores when it comes to marine plants and algae, which can disperse via ocean currents. Another feature of many marine species is their large effective population size, which decreases the effects of genetic drift (Palumbi, 1992, 1994). As a consequence, the population differentiation in the marine environment is slowed down (Carr et al., 2003; Kinlan & Gaines, 2003; Schiebelhut & Dawson, 2018). The scale of population structure in many marine species is on the order of thousands to tens of thousands of kilometers (Palumbi, 1992, 1994). The dispersal distance of the seeds in seagrasses

can be quite large, in the forms of floating reproductive fragments, buoyant fruits with viable seeds, or buoyant seedlings (Larkum et al., 2006). This may slow down the population differentiation. However, the large effective population size may not apply to some seagrass populations when there is a large degree of clonality, as there are very few individuals locally to mate with.

The evolution of life on the Earth started in the ocean at least 3.7 billion years ago (Ohtomo et al., 2014), while the colonization of the land by eukaryotes was estimated to be only a few hundred million years ago (Heckman et al., 2001). Despite of the larger spatial extent and longer evolutionary history in the ocean, the focus of evolutionary research is terrestrial, best demonstrated through a number of "model organism" such as mouse, the nematode *Caenorhabditis elegans* and *Drosophila* for animals, and *Arabidopsis* for plants. Their small size and short generation time make them ideal for experimental laboratory research. However, with the increase in the number of genome-sequencing projects, the definition of "model organism" has broadened (Hedges, 2002). Still, the species receiving an unusually large amount of attention from the research are mostly terrestrial, and include for example, the agriculturally important species (for example, rice), and those related to human health (for example, the malarial parasite *Plasmodium*) (Hedges, 2002). However, marine genomic models are emerging these days, such as the three-spined stickleback (*Gasterosteus aculeatus*) (Jones et al., 2012), as sequencing the genome of any species is now technically feasible and increasingly affordable (Ribeiro et al., 2017).

Model seagrass species (eelgrass *Zostera marina*)

Seagrasses are a polyphyletic group of marine angiosperms (~14 genera and ~65 species) which evolved in parallel three to four times from freshwater sister taxa (Les et al., 1997). They are widely distributed along temperate and tropical coastlines of all continents except Antarctica (Short et al., 2007), functioning as the foundation of one of the most important coastal ecosystems.

Eelgrass (*Z. marina*) is the most widespread seagrass species in temperate waters of the Northern Hemisphere (Larkum et al., 2018), and is an emerging model species for seagrass research. The analysis of the *Z. marina* genome shows that this species, possibly along with the other seagrasses, has regained functions enabling them to adjust to full salinity. One key trait is thought to be the cell walls that contain polyanionic, low-methylated pectins and sulfated galactans, which are important for ion homoeostasis, nutrient uptake and O_2/CO_2 exchange through leaf epidermal cells (Olsen et al., 2016). Previous studies based on microsatellite markers have promoted our understanding on the population genetics of *Z. marina*. Genetic mosaicism is widespread in *Z. marina* (Reusch & Boström, 2011), as is in other seagrass species (Digiantonio et al., 2020). Allelic richness is almost two-fold higher in the Pacific than that in the Atlantic, indicating Pacific origin, and diversity hotspots have been detected in both the Pacific

and the Atlantic, respectively (Olsen et al., 2004). Different geographic populations show different levels of sexual vs. asexual reproduction (Reusch et al., 2000; Olsen et al., 2004). In extreme cases, *Z. marina* meadows may be dominated by descendants tracing back to one or a few zygotes (Reusch et al., 1999; Reusch et al., 2000). Significant isolation-by-distance is found from ~150 to 5,000 km (Olsen et al., 2004).

The availability of a high-quality reference genome of *Z. marina* (Olsen et al., 2016; Ma et al., 2021) promotes a transition from genetics to genomics. Whole-genome information leads to much higher resolution and accuracy than using limited number of microsatellites, and it also goes beyond the power of microsatellites to accurately estimate the history of effective population size (Li & Durbin, 2011; Schiffels & Durbin, 2014), construct time-calibrated phylogenies (Bryant et al., 2012; Matschiner et al., 2020) and detect subtle genetic differences between modules of modular species (Schmid-Siegert et al., 2017; Wang et al., 2019). The genome of *Z. marina* is relatively compact (260.5 Mb), making it possible to study seagrass genomics with a relatively low cost.

SoGV in unitary and modular species

For both unitary and modular species, some specific somatic cells can divide via mitosis to increase their number. Mutations emerging during mitosis (i.e., somatic mutations) lead to SoGV among somatic cells during the development of one single zygote. SoGV has been thought to have no evolutionary consequences, due to the assumption of strict germline-soma separation at an early stage of the development (Weismann, 1892). However, this is only true for some animal species, which is a tiny fraction in the tree of life (Lee et al., 2012). For instance, in flowering plants (i.e., angiosperms), gametes are generated by flower meristems (FMs), which are converted from proliferating somatic cells in shoot apical meristems (SAMs) (Prunet et al., 2009). The transfer of SoGV from vegetative tissue to gametes and ultimately seeds has now been demonstrated in several plant species using genomics (Wang et al., 2019). Even animals may segregate germline cells from somatic cells late in development (Juliano & Wessel, 2010), such as sponges, cnidarians and flatworms (Kumano, 2015). It has also been demonstrated that SoGV can be transferred into gametes in corals (Vasquez-Kuntz et al., 2020).

Modular species may have another level of SoGV, i.e., genetic variation among modules, known as genetic mosaicism hypothesis (GMH) (Whitham, 1981; Gill, 1986). The key idea is that genetic heterogeneity makes it difficult for plant pests or herbivores to adapt to an entire tree. Although recent studies have shown very low levels of fixed genetic differences among leaves within single trees (Schmid-Siegert et al., 2017; Wang et al., 2019), the currently available data are still too little to make any generalization.

Clonal species belong to modular species, the modules of which are detached and live independently. Clonal species thus has two levels of “population”: the asexual population of

ramets belonging to a genet (Noble et al., 1979), and the population of genets, the "classical", sexually reproducing population. It is estimated that 66.5% of flora could perform clonal reproduction (Honnay & Bossuyt, 2005). This group also includes algae, fungi and basal animals such as reef-building corals, hydrozoans, bryozoans and colonial ascidians (Smith et al., 1992; Van Oppen et al., 2011).

SoGV in *Zostera marina*

In *Z. marina*, shoot apical meristems (SAMs) contain the somatic cells that can divide via mitosis to increase their number, which further differentiate to form tissues such as leaves. Once a somatic mutation occurs during mitosis, the direct consequence is that one single meristematic cell would obtain a different genotype, leading to a mosaic SoGV. It is still an open question whether a mosaic SoGV in the original ramet can get fixed in the descendant ramets (i.e., shared by all the cells within the ramet). What has been overlooked is that a mechanism is needed to turn a mosaic SoGV to a fixed one.

Since it is impossible to separate the SAM and sequence it, we will conduct bulk sequencing of the plant tissue containing the SAM and its descendant tissue. Bulk sequencing is reasonable, as the descendant tissue mirrors the genetic composition of the SAM. Fixed genotypes will show intra-ramet allele frequency of 0.5 for heterozygotes, and 1 for homozygotes. We expect mosaic SoGVs to show intermediate intra-ramet allele frequencies. To detect mosaic SoGV, we will borrow the methods in cancer research, as this is the study system most closely related. We will try to demonstrate the mechanism turning mosaic SoGV in the original SAM into a fixed state in the newly formed SAMs by analyzing the existence and emergence of fixed SoGV among *Z. marina* ramets tracing back to a single zygote.

Influence of ice ages on marine species

The population of genets (i.e., the "classical", sexually reproducing population), contain the information of demography and (re)colonization history. We study the worldwide evolutionary history of *Z. marina* by sequencing populations throughout the species' range. The earth has been through alternating glacial and interglacial periods. The effects of the glacial periods have profoundly influenced almost all shallow-water, coastal marine habitats, including seagrasses (Olsen et al., 2004). During glacial periods, the sea level dropped, which created coastal barriers or even separated previously connected ocean basins. For example, the Last Glacial Maximum (LGM) caused exposure of most East China Sea continental shelves, closure of the Taiwan Strait, and constriction of the Tsushima Strait (Shirota et al., 2021). Moreover, lower sea level could separate the previously connected sea basins, generating transient geographic barriers. For instance, sea level records indicate that the Bering Land Bridge opened and closed multiple times during glacial and inter-glacial periods, respectively (Hopkins, 1973; Hu et al., 2010).

Plant phylogenetics as a combination of nuclear and chloroplast phylogenetics

With the development of the sequencing technology, it has become much easier to obtain many independent SNPs distributed along the full genome. To construct a phylogenetic tree, one way is to generate a concatenated sequence based on the genome-wide SNPs (Gadagkar et al., 2005). However, different areas of the genome may undergo different sorting processes and support different topologies, and thus phylogenetic analyses based on the concatenation of loci may exaggerate the bias (Fang et al., 2020). Moreover, the phylogenetic tree is also affected by allele choice in the case of heterozygotes in the concatenation process (Wielstra et al., 2014). Stange et al. (2018) extended the multispecies coalescent (MSC) method implemented in SNAPP (Bryant et al., 2012), making it possible to conduct time calibration of the phylogenetic tree directly based on SNP (single nucleotide polymorphism) data. MSC-based methods are computationally prohibitive to run on large number SNPs, but a few thousand SNPs are sufficient to achieve reliable results (Fang et al., 2020). This method has been successfully applied to population-level phylogenetics (Fang et al., 2020).

Flowering plants have three separate genomes, located in the nucleus, mitochondria and chloroplasts. There are rearrangements of genetic materials on homologous chromosomes during meiosis (Smith & Nicolas, 1998), which breaks the nuclear DNA into short units that are inherited independently. By contrast, chloroplast DNA (cpDNA) and mitochondrial DNA (mtDNA) are inherited as haplotypes without recombination (Birky, 1995). Plants exhibit a significant evolutionary plasticity in their mtDNA, which contrasts with the more conservative evolution of their cpDNA (Knoop, 2004). The mtDNA of plants show frequent genomic rearrangements, the incorporation of foreign DNA from the nuclear and chloroplast genomes, an ongoing transfer of genes to the nucleus (Knoop, 2004). Thus, plant phylogenetic studies are mostly based on nuclear and chloroplast markers, and such an approach will also be followed here when analyzing the worldwide evolutionary history of *Z. marina*.

Genetic admixture (i.e., secondary contact of genetically diverged lineages) is common on population level, which affects the inference of evolutionary history. Admixture zones will show co-existence of multiple chloroplast types at the beginning of the secondary contact. However, as one “supergene”, eventually chloroplast genome will reach a fixed state of one type due to genetic drift. Thus, populations with similar admixture history may show very different chloroplast types. Although nuclear genome can reveal genetic admixture by taking into account the many independent units distributed along the genome, the phylogeny based on nuclear genome is inevitably affected by genetic admixture due to the assumption of a bifurcating evolutionary model. To reconstruct the evolutionary history of *Z. marina* populations distributed all over the world, I first need to construct a bifurcating tree that shows the evolutionary relationships of different populations, and then date these bifurcations. Genetic admixture violates the assumption of the bifurcating history, which leads to biased topology and wrong

estimates of divergence time. Thus, genetic admixture must be considered in population-level phylogenetics.

Hence, I will use both nuclear and chloroplast genomes to reconstruct the worldwide evolutionary history of *Z. marina*. As a relatively long DNA sequence (115 to 165 kb for angiosperms) (Sun et al., 2020) without recombination, chloroplast genome can show the stepwise accumulation of mutations. However, inherited as a haplotype, chloroplast genome is technically a "supergene", where a genetic variant will be either fixed or completely lost in the population. On the contrary, owing to recombination, fractions of the nuclear genome can show diverging evolutionary relationships, each representing an evolutionary possibility under the same evolutionary history. The real evolutionary history can be reconstructed by analyzing these independent units. Yet, nuclear genome cannot show stepwise accumulation of mutations. Combining nuclear and chloroplast genomes can provide better understanding of the evolutionary history.

Dissertation objective and outline

This doctoral dissertation uses *Z. marina* as a model to study multi-level genetic variation of clonal species. Firstly, by looking at SNPs distributed along the full genome, I extend the research on *Z. marina* from genetics to genomics. Secondly, I extend the research to fixed SoGV (i.e., fixed genetic differences among ramets tracing back to the same zygote) by sequencing 24 ramets of a large *Z. marina* clone. Thirdly, I extend the research to mosaic SoGV (i.e., genetic variation among cells within a ramet) by sequencing 3 *Z. marina* ramets at ultra-deep coverage (~1370x). Finally, based on the genets distributed among *Z. marina* populations (i.e., the "classical", sexually reproducing population) collected from all over the world, I reconstruct the worldwide evolutionary history of *Z. marina*.

This dissertation contains three chapters based on three separate manuscripts:
Chapter 1: Yu, L., Boström, C., Franzenburg, S., Bayer, T., Dagan, T., & Reusch, T. B. (2020). Somatic genetic drift and multilevel selection in a clonal seagrass. Nature ecology & evolution, 4(7), 952-962.
The first manuscript investigated the level of mosaic and fixed SoGVs in an asexual population of *Z. marina*. I tested whether an asexual population has the potential to adapt to the changing environments, and demonstrated how mosaic SoGV could turn to fixed SoGV.

Chapter 2: Yu, L. et al. Frequent and swift trans-ocean dispersal events of a widespread marine angiosperm, the seagrass *Zostera marina* L., to be submitted to Nature Plants
In the second manuscript, I reconstructed the worldwide evolutionary history of *Z. marina*, based on *Z. marina* populations collected throughout the species' range.

Chapter 3: Yu, L. et al. Using "identity by heterozygosity (IBH)" to detect clonemates under prevalent clonal reproduction in multicellular diploids, to be submitted to Molecular Ecology Resources

The last manuscript described a method based on shared heterozygous genotypes, to detect clonemates for any pair of samples.

Glossary

Clonal species: Multicellular organism that can perform vegetative (asexual) reproduction using single or multiple somatic cell(s) as precursor.

Fixed genotype: Genotype that is shared by all the cells within a ramet in clonal species, or within a module in modular species.

Fixed somatic genetic variation (fixed SoGV): Different fixed genotypes among ramets of the same clone /genet in clonal species, or among modules of the same organism in modular species.

Genet /Clone: Complete collection of all the ramets tracing back to the same zygote in clonal species.

Germline genetic variation: Genetic variation among zygotes.

Germline mutation: DNA mutation occurring in gametes during meiosis.

Germline: Complete collection of cells with ability to form gametes in multicellular organisms.

Modular species: Multicellular organism composed of iterative units of similar parts (modules) that either stay together (e.g., tree) or become independent (clonal species).

Mosaic somatic genetic variation (mosaic SoGV): Different genotypes among cells within a ramet in clonal species, or within a module in modular species.

Ramet: Basic unit of clonal species, which has full ability to live independently. Eventually a ramet will become completely independent, but there may be a transient stage when ramets are connected to each other.

Read frequency: Relative frequency of the reads (of next-generation sequencing) supporting a particular allele.

Reference read frequency (RRF): Read frequency of the "REF" allele in a vcf format file.

Soma: Complete collection of cells without ability to form gametes in multicellular organisms.

Somatic genetic variation (SoGV): Genetic variation caused by somatic mutations.

Somatic mutation: DNA mutation occurring in somatic tissues during mitosis.

Variant read frequency (VRF): Read frequency of the "ALT" allele in a vcf format file.

Zygote: Eukaryotic cell formed by fusion of male and female gametes, which is the initial ancestor of all the cells within a multicellular organism.

References

Birky, C. W. (1995). Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proc. Natl. Acad. Sci. U.S.A., 92*(25), 11331-11338. doi:10.1073/pnas.92.25.11331

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A. & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol., 29*(8), 1917-1932. doi:10.1093/molbev/mss086

Carr, M. H., Neigel, J. E., Estes, J. A., Andelman, S., Warner, R. R. & Largier, J. L. (2003). Comparing marine and terrestrial ecosystems: implications for the design of coastal marine reserves. *Ecol. Appl., 13*(sp1), 90-107. doi:10.1890/1051-0761(2003)013[0090:CMATEI]2.0.CO;2

Digiantonio, G., Blum, L., McGlathery, K. & Waycott, M. (2020). Genetic mosaicism and population connectivity of edge-of-range Halodule wrightii populations. *Aquat. Bot., 161*, 103161. doi:10.1016/j.aquabot.2019.103161

Fang, B., Merilä, J., Matschiner, M. & Momigliano, P. (2020). Estimating uncertainty in divergence times among three-spined stickleback clades using the multispecies coalescent. *Mol. Phylogenet. Evol., 142*, 106646. doi:10.1016/j.ympev.2019.106646

Gadagkar, S. R., Rosenberg, M. S. & Kumar, S. (2005). Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J. Exp. Zool. B Mol. Dev. Evol., 304*(1), 64-74. doi:10.1002/jez.b.21026

Gill, D. E. (1986). Individual plants as genetic mosaics: Ecological organisms versus evolutionary individuals. In C. M. J (Ed.), *Plant Ecology* (pp. 321-343). Oxford: Blackwell.

Harper, J. L. & White, J. (1974). The demography of plants. *Annu. Rev. Ecol. Evol. Syst., 5*(1), 419-463. doi:10.1146/annurev.es.05.110174.002223

Heckman, D. S., Geiser, D. M., Eidell, B. R., Stauffer, R. L., Kardos, N. L. & Hedges, S. B. (2001). Molecular evidence for the early colonization of land by fungi and plants. *science, 293*(5532), 1129-1133. doi:10.1126/science.1061457

Hedges, S. B. (2002). The origin and evolution of model organisms. *Nat. Rev. Genet., 3*(11), 838-849. doi:10.1038/nrg929

Honnay, O. & Bossuyt, B. (2005). Prolonged clonal growth: escape route or route to extinction? *Oikos, 108*(2), 427-432. doi:10.1111/j.0030-1299.2005.13569.x

Hopkins, D. M. (1973). Sea level history in beringia during the past 250,000 years1. *Quat. Res., 3*(4), 520-540. doi:10.1016/0033-5894(73)90029-X

Hu, A., Meehl, G. A., Otto-Bliesner, B. L., Waelbroeck, C., Han, W., Loutre, M.-F., . . . Rosenbloom, N. (2010). Influence of Bering Strait flow and North Atlantic circulation on glacial sea-level changes. *Nat. Geosci., 3*(2), 118-121. doi:10.1038/ngeo729

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., . . . White, S. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature, 484*(7392), 55-61. doi:10.1038/nature10944

Juliano, C. & Wessel, G. (2010). Versatile germline genes. *science, 329*(5992), 640-641. doi:10.1126/science.1194037

Kinlan, B. P. & Gaines, S. D. (2003). Propagule dispersal in marine and terrestrial environments: a community perspective. *Ecology, 84*(8), 2007-2020. doi:10.1890/01-0622

Knoop, V. (2004). The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Curr. Genet., 46*(3), 123-139. doi:10.1007/s00294-004-0522-8

Kumano, G. (2015). Evolution of germline segregation processes in animal development. *Dev. Growth Differ. , 57*(4), 324-332. doi:10.1111/dgd.12211

Larkum, A. W. D., Kendrick, G. A. & Ralph, P. J. (2018). *Seagrasses of Australia: Structure, Ecology and Conservation*: Springer.

Larkum, A. W. D., Orth, R. J. & Duarte, C. M. (2006). *Seagrasses: Biology, Ecology, and Conservation*: Springer.

Lee, S. C., Ristaino, J. B. & Heitman, J. (2012). Parallels in intercellular communication in oomycete and fungal pathogens of plants and humans. *PLoS Pathog., 8*(12), e1003028. doi:10.1371/journal.ppat.1003028

Les, D. H., Cleland, M. A. & Waycott, M. (1997). Phylogenetic studies in Alismatidae, II: evolution of marine angiosperms (seagrasses) and hydrophily. *Syst. Bot.*, 443-463. doi:10.2307/2419820

Li, H. & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature, 475*(7357), 493-496. doi:10.1038/nature10231

Ma, X., Olsen, J. L., Reusch, T. B. H., Procaccini, G., Kudrna, D., Williams, M., . . . Van de Peer, Y. (2021). Improved chromosome-level genome assembly and annotation of the seagrass, *Zostera marina* (eelgrass). *F1000research, 10*. doi:10.12688/f1000research.38156.1

Matschiner, M., Böhne, A., Ronco, F. & Salzburger, W. (2020). The genomic timeline of cichlid fish diversification across continents. *Nat. Commun., 11*(1), 1-8. doi:10.1038/s41467-020-17827-9

Noble, J. C., Bell, A. D. & Harper, J. L. (1979). The population biology of plants with clonal growth: I. The morphology and structural demography of *Carex arenaria*. *J. Ecol., 67*, 983-1008. doi:10.2307/2259224

Ohtomo, Y., Kakegawa, T., Ishida, A., Nagase, T. & Rosing, M. T. (2014). Evidence for biogenic graphite in early Archaean Isua metasedimentary rocks. *Nat. Geosci., 7*(1), 25-28. doi:10.1038/ngeo2025

Olsen, J. L., Rouzé, P., Verhelst, B., Lin, Y.-C., Bayer, T., Collen, J., . . . Maumus, F. (2016). The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature, 530*(7590), 331-335. doi:10.1038/nature16548

Olsen, J. L., Stam, W. T., Coyer, J. A., Reusch, T. B., Billingham, M., Bostrom, C., . . . Wyllie-Echeverria, S. (2004). North Atlantic phylogeography and large-scale population differentiation of the seagrass *Zostera marina* L. *Mol. Ecol., 13*(7), 1923-1941. doi:10.1111/j.1365-294X.2004.02205.x

Palumbi, S. R. (1992). Marine speciation on a small planet. *Trends Ecol. Evol., 7*(4), 114-118. doi:10.1016/0169-5347(92)90144-Z

Palumbi, S. R. (1994). Genetic divergence, reproductive isolation, and marine speciation. *Annu. Rev. Ecol. Syst., 25*(1), 547-572. doi:10.1146/annurev.es.25.110194.002555

Prunet, N., Morel, P., Negrutiu, I. & Trehin, C. (2009). Time to stop: flower meristem termination. *Plant Physiol., 150*(4), 1764-1772. doi:10.1104/pp.109.141812

Reusch, T. B. H. & Boström, C. (2011). Widespread genetic mosaicism in the marine angiosperm *Zostera marina* is correlated with clonal reproduction. *Evol. Ecol., 25*(4), 899-913. doi:10.1007/s10682-010-9436-8

Reusch, T. B. H., Boström, C., Stam, W. T. & Olsen, J. L. (1999). An ancient eelgrass clone in the Baltic. *Mar. Ecol. Prog. Ser., 183*, 301-304. doi:10.3354/meps183301

Reusch, T. B. H., Stam, W. T. & Olsen, J. L. (2000). A microsatellite-based estimation of clonal diversity and population subdivision in *Zostera marina*, a marine flowering plant. *Mol. Ecol., 9*(2), 127-140. doi:10.1046/j.1365-294x.2000.00839.x

Ribeiro, Â. M., Foote, A. D., Kupczok, A., Frazão, B., Limborg, M. T., Piñeiro, R., . . . da Fonseca, R. R. (2017). Marine genomics: News and views. *Mar Genomics, 31*, 1-8. doi:10.1016/j.margen.2016.09.002

Schiebelhut, L. M. & Dawson, M. N. (2018). Correlates of population genetic differentiation in marine and terrestrial environments. *J. Biogeogr., 45*(11), 2427-2441. doi:10.1111/jbi.13437

Schiffels, S. & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat. Genet., 46*(8), 919-925. doi:10.1038/ng.3015

Schmid-Siegert, E., Sarkar, N., Iseli, C., Calderon, S., Gouhier-Darimont, C., Chrast, J., . . . Pagni, M. (2017). Low number of fixed somatic mutations in a long-lived oak tree. *Nat. Plants, 3*(12), 926-929. doi:10.1038/s41477-017-0066-9

Shirota, K., Okazaki, Y., Konno, S., Miyairi, Y., Yokoyama, Y. & Kubota, Y. (2021). Changes in surface water masses in the northern East China Sea since the Last Glacial Maximum based on diatom assemblages. *Prog. Earth Planet. Sci. , 8*(1), 1-17. doi:10.1186/s40645-021-00456-1

Short, F., Carruthers, T., Dennison, W. & Waycott, M. (2007). Global seagrass distribution and diversity: a bioregional model. *J. Exp. Mar. Biol. Ecol., 350*(1-2), 3-20. doi:10.1016/j.jembe.2007.06.012

Smith, K. N. & Nicolas, A. (1998). Recombination at work for meiosis. *Curr. Opin. Genet. Dev., 8*(2), 200-211. doi:10.1016/S0959-437X(98)80142-1

Smith, M. L., Bruhn, J. N. & Anderson, J. B. (1992). The fungus *Armillaria bulbosa* is among the largest and oldest living organisms. *Nature, 356*(6368), 428-431. doi:10.1038/356428a0

Stange, M., Sánchez-Villagra, M. R., Salzburger, W. & Matschiner, M. (2018). Bayesian divergence-time estimation with genome-wide single-nucleotide polymorphism data of

sea catfishes (Ariidae) supports Miocene closure of the Panamanian Isthmus. *Syst. Biol., 67*(4), 681-699. doi:10.1093/sysbio/syy006

Sun, J., Wang, Y., Liu, Y., Xu, C., Yuan, Q., Guo, L. & Huang, L. (2020). Evolutionary and phylogenetic aspects of the chloroplast genome of Chaenomeles species. *Sci. Rep., 10*, 11466. doi:10.1038/s41598-020-67943-1

Van Oppen, M. J. H., Souter, P., Howells, E. J., Heyward, A. & Berkelmans, R. (2011). Novel genetic diversity through somatic mutations: fuel for adaptation of reef corals? *Divers., 3*(3), 405-423. doi:10.3390/d3030405

Vasquez-Kuntz, K. L., Kitchen, S. A., Conn, T. L., Vohsen, S. A., Chan, A. N., Vermeij, M. J., . . . Baums, I. B. (2020). Juvenile corals inherit mutations acquired during the parents lifespan. *bioRxiv*. doi:10.1101/2020.10.19.345538

Wang, L., Ji, Y., Hu, Y., Hu, H., Jia, X., Jiang, M., . . . Jia, Y. (2019). The architecture of intra-organism mutation rate variation in plants. *PLoS Biol., 17*(4), e3000191. doi:10.1371/journal.pbio.3000191

Weismann, A. (1892). *Das Keimplasma: eine theorie der Vererbung*: G. Fischer.

Whitham, T. G. (1981). Individual trees as heterogeneous environments: adaptation to herbivory or epigenetic noise? In *Insect life history patterns* (pp. 9-27): Springer.

Wielstra, B., Arntzen, J. W., van der Gaag, K. J., Pabijan, M. & Babik, W. (2014). Data concatenation, Bayesian concordance and coalescent-based analyses of the species tree for the rapid radiation of *Triturus newts*. *PLoS ONE, 9*(10), e111011. doi:10.1371/journal.pone.0111011

Chapter 1: Somatic genetic drift and multi-level selection in a clonal seagrass

Lei Yu[1]; Christoffer Boström[2]; Sören Franzenburg[3]; Till Bayer[1]; Tal Dagan[4]; Thorsten B.H. Reusch[1*]

1 Marine Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
2 Environmental and Marine Biology, Åbo Akademi University, Turku, Finland
3 Institute for Clinical Molecular Biology, University of Kiel, Kiel, Germany
4 Institute of Microbiology, University of Kiel, Kiel, Germany
*corresponding author, treusch@geomar.de, phone +49-431-600-4550

Abstract

All multicellular organisms are genetic mosaics owing to somatic mutations. The accumulation of somatic genetic variation in clonal species undergoing asexual (or clonal) reproduction may lead to phenotypic heterogeneity among autonomous modules (termed ramets). However, the abundance and dynamics of somatic genetic variation under clonal reproduction remain poorly understood. Here we show that branching events in a seagrass (*Zostera marina*) clone or genet lead to population bottlenecks of tissue that result in the evolution of genetically differentiated ramets in a process of somatic genetic drift. Studying inter-ramet somatic genetic variation, we uncovered thousands of single-nucleotide polymorphisms that segregated among ramets. Ultra-deep resequencing of single ramets revealed that the strength of purifying selection on mosaic genetic variation was greater within compared to among ramets. Our study provides evidence for multiple levels of selection during the evolution of seagrass genets. Somatic genetic drift during clonal propagation leads to the emergence of genetically unique modules that constitute an elementary level of selection and individuality in long-lived clonal species.

Introduction

All multicellular species, from plants to humans, are genetic mosaics, owing to mitotic errors (somatic mutations) during growth and development[1,2]. The resulting genetic heterogeneity may lead to genomic conflict among cell populations and is often associated with degenerative disease such as cancer[3]. This applies, in particular, to non-clonal, unitary species that rely on organismal integration of specialized and non-redundant organs (but see ref [4]). Somatic genetic variation may play a different, more positive role in multicellular species undergoing asexual reproduction, hereafter called clonal species, featured by 65% of all plant families[5] and 35% of all animal phyla[6]. Clonal species have a simple body plan, and often indeterminate growth, during which iterative units (modules) emerge by asexual proliferation through fission, budding or branching[7]. When modules achieve physiological and morphological autonomy, the emerging ramets[8] may be subject to selection. Populations of ramets collectively form the clone or genet[8], founded by a single sexually produced genotype (Supplementary Table 1, glossary). Importantly, when clonal offspring is produced, somatic genetic variation may be unequally distributed. Inevitably, a subset of the original proliferating cell population are precursors for any new ramet, and the resulting segregation of somatic genetic variation may result in genetically differentiated ramets (Fig. 1). This process of *somatic genetic drift* at the level of proliferating cell populations is fundamental and inevitable. As any potentially important neutral population genetic process, somatic genetic drift needs to be fully understood before addressing the effect of selection[9-11]. Here, we hypothesize that recurrent events of somatic genetic drift create an intermediate layer of genetic variation in between the cell population level (as genetic mosaicism) and sexually recombining genotypes (populations of genets), namely at the ramet level in clonal species[6,12,13] (Fig. 1, Extended Data Fig. 1). In line with this notion, population genetic quantifications demonstrated that with increasing longevity, there may be far more mitotic cell divisions during one zygote cycle than meiotic ones[14], potentially providing large mutational input for both, selection and somatic genetic drift to operate[15,16]. Accordingly, several studies demonstrated the emergence of phenotypic differences among ramets of the same genet[10,17,18]. However, the corresponding genome-level assessments of frequency, dynamics and possible functional consequences of somatic genetic variation are currently lacking for clonal species[19,20].

Here we examined a clonal species with facultative asexual proliferation, the seagrass *Zostera marina*, belonging to a group of marine angiosperms well known for extensive clonal proliferation and longevity[21-23]. The relatively compact genome of *Z. marina* (204 Mbp) has recently been characterized[24], allowing us to re-sequence replicated ramets of a large genet with high sequencing coverage. Our objective was thus to obtain whole-genome-level quantification of inter- and intra-ramet genetic heterogeneity of a single, large seagrass genet or clone, as a baseline to assess the potential for intra- and inter-ramet selection. With this data, we tested the hypotheses that (i) branching events would result in somatic genetic drift, detectable as fully heterozygous (fixed) polymorphisms segregating among ramets; (ii) mosaic genetic polymorphism would change in allele frequency among ramets in accordance with the inferred

genealogy of a growing clone, and (iii) intra- and inter-ramet levels differ in terms of their molecular selection regime, providing evidence for multi-level selection.

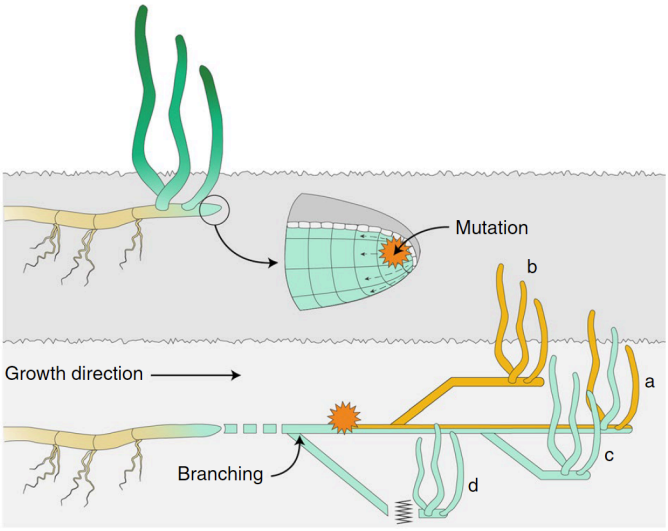


Fig. 1 | Somatic genetic drift among ramets of a seagrass genet causes segregation of genetic variation. During each branching event, a finite subsample of proliferating cells forms each new ramet. At a given locus, somatic genetic variation at a locus may remain in the mosaic state, become fixed for the mutation (orange), or lose the novel somatic genetic variation. In *Z. marina*, ramets become physically independent after ≈2 yrs.

Results

Abundant somatic genetic variation in a single plant genet

In order to describe genomic patterns of a putative large seagrass genet in space, twenty-four ramets were collected by SCUBA in 2016 along two transects at the site Angsö in SW Finland, Baltic Sea (Fig. 2a and Supplementary Table 2). This is one of the many sites that feature mega-clones of hundreds of meters in spatial extension[21]. The selected meadow (covering an area of at least 6 ha) has been previously characterized as "clonal" using microsatellite genotyping[21]. Assuming a linear model for clonal vegetation expansion after initial patch establishment[25], we estimate that between 1,190-2,381 branching events preceded each of the contemporary living leaf shoot or ramet (see Methods). Bulk DNA from basal leaf shoot tissue encompassing the meristematic region and the base of the leaves was sequenced with an average coverage of 81x per sample (Supplementary Tables 3 & 4). Sequence reads were mapped against the *Z. marina* reference genome[24] with an average mapping rate of 94%. Note that the reference genome was from a clonal meadow only 22 km away from the samples used in the present study. We focused first on fixed genetic variants that were present in all the cells within a given ramet. After removing 25,998 homozygous SNPs that differed to the reference, a total of 38,831 SNPs were detected as polymorphic compared to the reference genome after stringent filtering (Methods, Extended Data Fig. 2).

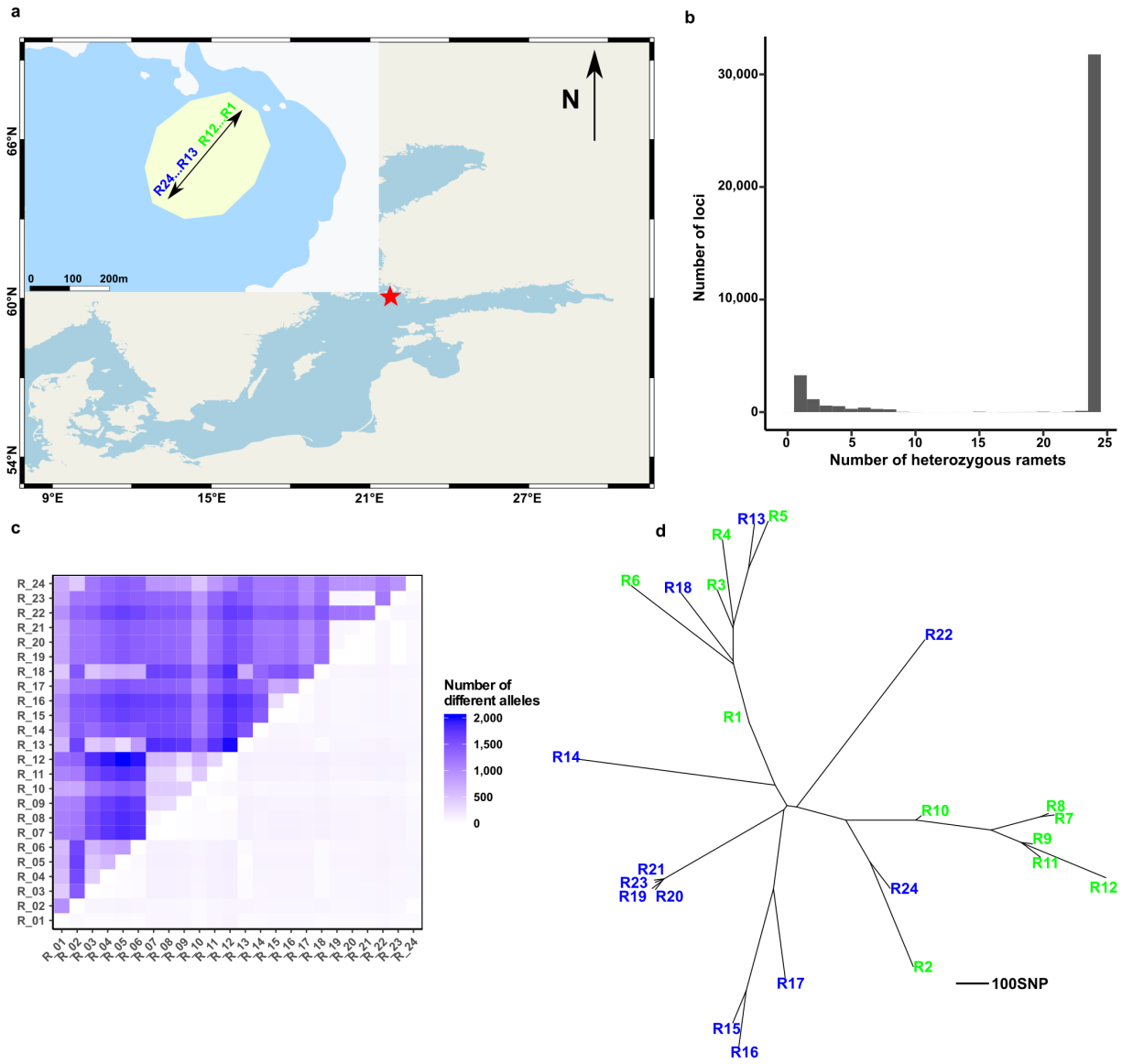


Fig. 2 | Inter-ramet genetic differentiation among the 24 seagrass ramets belonging to a single genet (=clone). a, Location of the sampled seagrass population in southern Finland and sampling points along a transect with sample numbers running from shallow to deep **b**, Histogram of shared SNP polymorphisms among 24 ramets. The x-axis depicts the number of heterozygous ramets, the y-axis represents the number of shared SNPs among samples. **c**, Pairwise number of heterozygous SNPs differing among ramets. Above the diagonal: all SNPs; below the diagonal: only non-synonymous SNPs. **d**, Neighbor-joining tree based on 24 sampled ramets (green are shallow, blue are deep sampling locations). The tree is based on 7,054 SNPs and calculated using 1,000 bootstrap replicates. Most nodes had a bootstrap support >95%, see Supplementary Figure 4.

The resulting distribution among SNP genotypes was first used to assess whether or not all 24 sampled ramets were of the same genet, i.e. originated exclusively by vegetative propagation. Members of the same genet are expected to share many heterozygous loci because under mitosis, all SNPs of the ancestral zygote will remain heterozygous and completely linked in offspring. The majority of heterozygous loci (31,777; 81.83%) were indeed shared and thus in complete genetic linkage among all 24 ramets (Fig. 2b, right hand bar), suggesting that all samples belong to the same genet. Moreover, we only found 10 homozygous SNPs differing among any pairs of ramets, which likely represented double mutations at the same genomic location and further supports that the only differentiation among ramets occurred via somatic mutations. In order to assess the likelihood that a rare recombination event may have gone undetected, we simulated a scenario where one of the 24 sampled ramets originated from a single sexual selfing event. The resulting distribution of shared heterozygous loci (Supplementary Figure 1) was very different from the empirical one. Importantly, our analysis also revealed a substantial number of SNPs that were heterozygous in a subset of ramets (Fig. 2b, left-hand side). These SNPs are best explained by somatic mutations that emerged initially as mosaics but later increased in frequency to become fully heterozygous genotypes via somatic genetic drift; these SNPs are the focus of our investigation.

Inter-ramet genetic diversity suggests widespread somatic genetic drift

After subtracting 31,777 loci that were shared among all 24 ramets (blue SNPs in Extended Data Fig. 1), we identified 7,054 SNPs that originated by somatic mutations (Fig. 2b). Experimental validation of a subset of the inferred SNPs by genotyping all combinations of 14 SNP loci tested in 24 ramets revealed that all genotypes were accurately annotated (Supplementary Figure 2 & 3, Supplementary Table 5). The average pairwise distance among ramets calculated from the observed polymorphic loci was 1,216 SNPs (Fig. 2c and Supplementary Tables 6 & 7), and no SNPs were exclusively shared among distant ramets (Supplementary Table 6). Based on the genome annotation[24] we identified 1,672 SNPs located in genic regions, and 597 in coding regions, of which 432 were non-synonymous SNPs (Supplementary Dataset 1). We then examined how the genetic similarities would correspond to the spatial location of ramets. A neighbor-joining tree of the sampled ramets revealed that geographically ones clustered together (Fig. 2d and Supplementary Figure 4), confirming the mutation accumulation hypothesis predicted for large vegetative genets. There were, however, several notable exceptions (e.g. ramets R02, R13, and R18). The incongruence between the branching topology and sampling location of these ramets indicates that they were recently introduced, most likely by uprooting and drifting to the current location as shown for another seagrass species[22].

Our analysis further revealed 1,654 insertion-deletion polymorphisms (indels) ranging between 1-201 bp in size (median 4 bp, mean 8.53 bp). The frequency of indels in our data is in line with other assessments of the expected ratio between point mutations leading to SNPs and insertion / deletion polymorphisms[26]. The distribution of indel-based genetic diversity revealed largely similar patterns to that of SNPs with respect to the pairwise inter-ramet divergence and the

resulting topology (Extended Data Fig. 3, Supplementary Figure 5). This is consistent with a view that many of the somatic polymorphisms are neutral to the ramet fitness, regardless of whether these are of the indel or point mutation type. We also identified one larger structural polymorphism of 200 bp in size, and 7 microsatellite polymorphisms, which occurred along the branching history reflected in the phylogenetic tree topology. All observed somatic variants independently confirmed the ramet genealogy based on SNPs (Extended Data Fig. 4 & Supplementary Figure 6).

Intra-ramet somatic mutations and genetic mosaicism

Any somatic mutational input enters the ramet at a low frequency through a single, proliferating cell and stays as genetic mosaic unless it is lost or rising to fixation (Fig. 1). Quantifying such somatic genetic polymorphism requires an appropriate resolution to detect intra-ramet allele frequencies that are $<<0.5$. Hence, we re-sequenced three ramets (R08, R10, and R12) to a very high coverage of 1370x (Extended Data Fig. 2, Supplementary Table 8). Using a restriction enzyme based method (see Methods) we validated 13/15 tested mosaic polymorphisms (Supplementary Figure 7, Supplementary Table 9 & 10). Both non-confirmed SNPs were at the lower range of the variant read frequency (f=0.05 and 0.06, respectively). This ultra-high coverage dataset was first used as a standard to obtain an alternative estimate of the proportion of true, fully heterozygous genotypes as described in the previous section (see Methods). To assess the performance of GATK pipeline, we used the larger sample of 38,831 SNPs detected based on the 81x resequencing data (i.e. before removing those 31,777 polymorphisms that were shared among all 24 ramets and had not originated by somatic mutation, Extended Data Fig.2). We first calculated the total number of identified heterozygous genotypes in the three ramets (96,291). We then used the 1370x data to construct a confidence interval. Reassuringly, 81.09% of the locus * ramet combinations (78,087/96,291) were confirmed as being fully heterozygous, demonstrating that our genotyping pipeline was largely successful in detecting fixed SNPs among ramets of a single genet and distinguishing them from the mosaic ones. We plotted the histogram of VRF (variant read frequency) for the three ramets based on the 81x data (7,054 SNPs for further analysis), and found that most of the SNPs were centered at VRF=0.5 (Supplementary Figure 8).

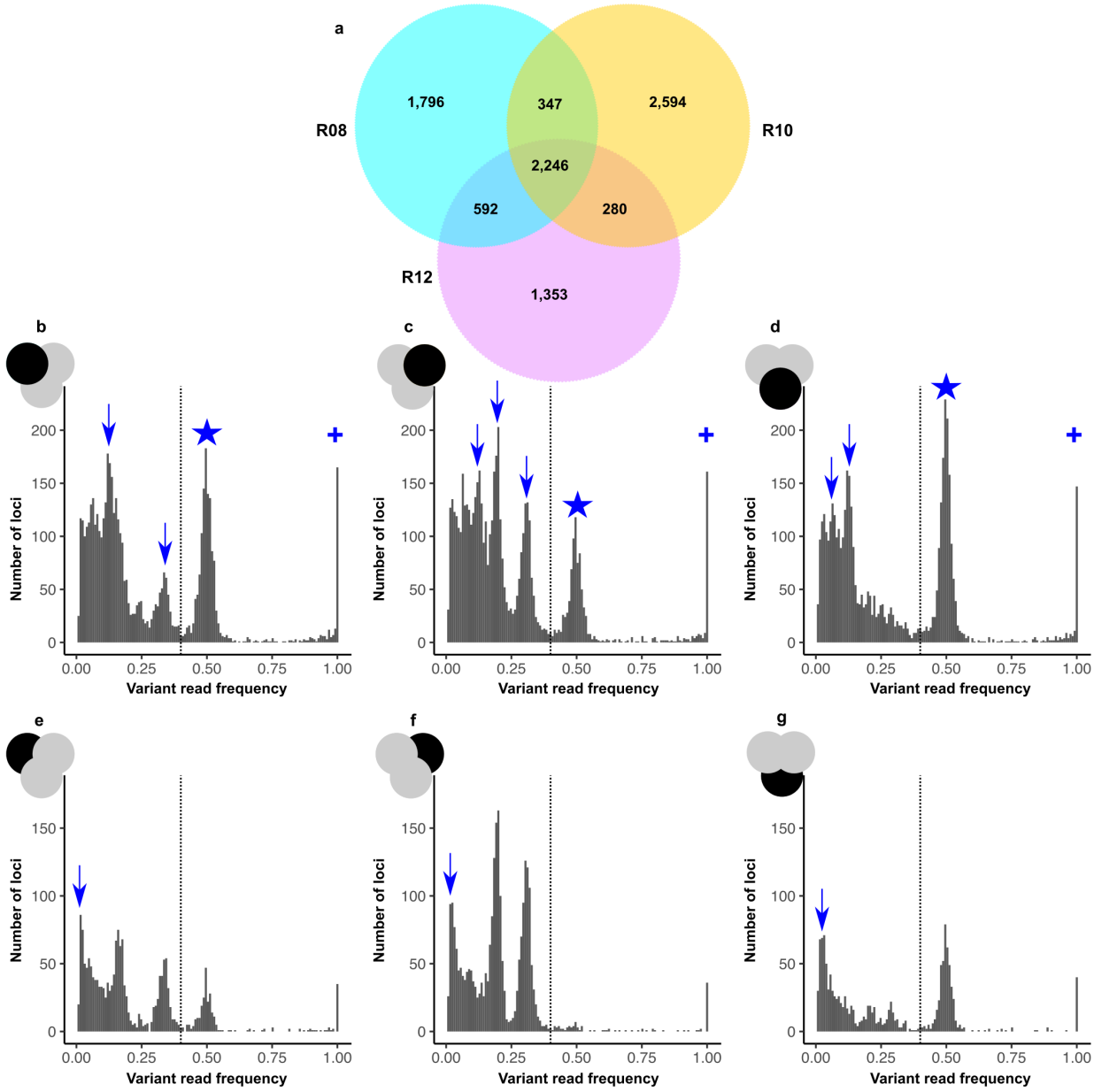


Fig. 3 | Intra-ramet somatic genetic variation. a, Venn diagram depicting overlap of mosaic and non-mosaic SNP variants in three ultra-deep sequenced ramets (1370x) that emerged via somatic mutation. In all panels, a miniature Venn diagram indicates which sample of SNPs was considered. **b-d,** Histogram of intra-ramet variant read frequency (VRF) for ramets 08, 10, and 12. The x-axis represents the VRF as a proxy for the allele frequency, and the y-axis represents the number of SNPs. SNPs were categorized into three types: (i) non-fixed, intra-ramet mosaic-type SNPs in which arrows depict differentiated cell populations (ii) fixed heterozygotes (peak at f=0.5, star); and fixed homozygotes ("+", somatic mutation 0/1- >1/1; peak at f=1). **e-g**, VRF histograms for the private variants in each ramet. Here, arrows indicate the mutational input, following a power law accumulation. In all panels, the dashed line depicts the threshold of VRF below which a SNP was considered to be in a mosaic, non-fixed state (see Methods).

On average, among the three ultra-deeply sequenced ramets, we found 4,973 intra-ramet somatic SNPs (Fig. 3a). Of these, 71.98% were in a mosaic state, i.e. had a frequency $f<0.5$. We first checked the overlap of fixed heterozygous genotypes between the 81x and 1370x coverage data. Most of the heterozygous genotypes detected under 81x coverage were identified as fixed ones by the 1370x data set (74.76%, 2,115/2,829, Supplementary Table 11). Conversely, none of the low-VRF ($f<0.4$) SNPs identified in the 1370x data set were detected as heterozygous genotype using the 81x data (Supplementary Table 12). Among the mosaic SNPs, 820 were found in genic regions with 301 in coding regions, of which 198 were non-synonymous. Among the non-synonymous polymorphisms, none of those intra-ramet, mosaic ones overlapped with the fixed SNPs detected among any of the 24 ramets. The distribution of variant read frequency (VRF), our proxy for intra-ramet allele frequency, showed concordant patterns with multiple peaks within the low-frequency regions for each ramet, indicating the coexistence of different proliferating cell populations (arrows; Fig. 3b-d). All apical shoot meristems examined thus far in flowering plants are organized into different cell populations or layers[27,28]. Hence, the modes at $f<0.5$ likely correspond to somatic genetic variants that have come to sub-fixation in separate meristematic layers[27-29]. We also infer from the VRF distribution that cells within one meristematic layer date back to a few initial cells[30], which explains the rapid sub-fixation of somatic mutations. A second non-exclusive explanation for the pronounced modes of the VRF distribution would be intra-ramet positive selection, which may favor a particular cell lineage possessing a driver mutation with which an entire cohort of neutral background mutations would "hitchhike"[31].

Under exponential clonal expansion of ramets, the accumulation of neutral genetic variants should follow a power-law distribution, leading to a left-hand peak of rare variants[32,33]. Assuming that many of the intermediate modes of the VRF distribution represent somatic genetic variants that were fixed within confined meristematic layers, we excluded those SNPs shared among ramets. When plotting the distribution of ramet-specific somatic genetic variants only, the expected left-hand peak of variants appeared, indicative of neutral mutational accumulation proportional to $1/f$ as described in modeling studies (Fig. 3e-g).

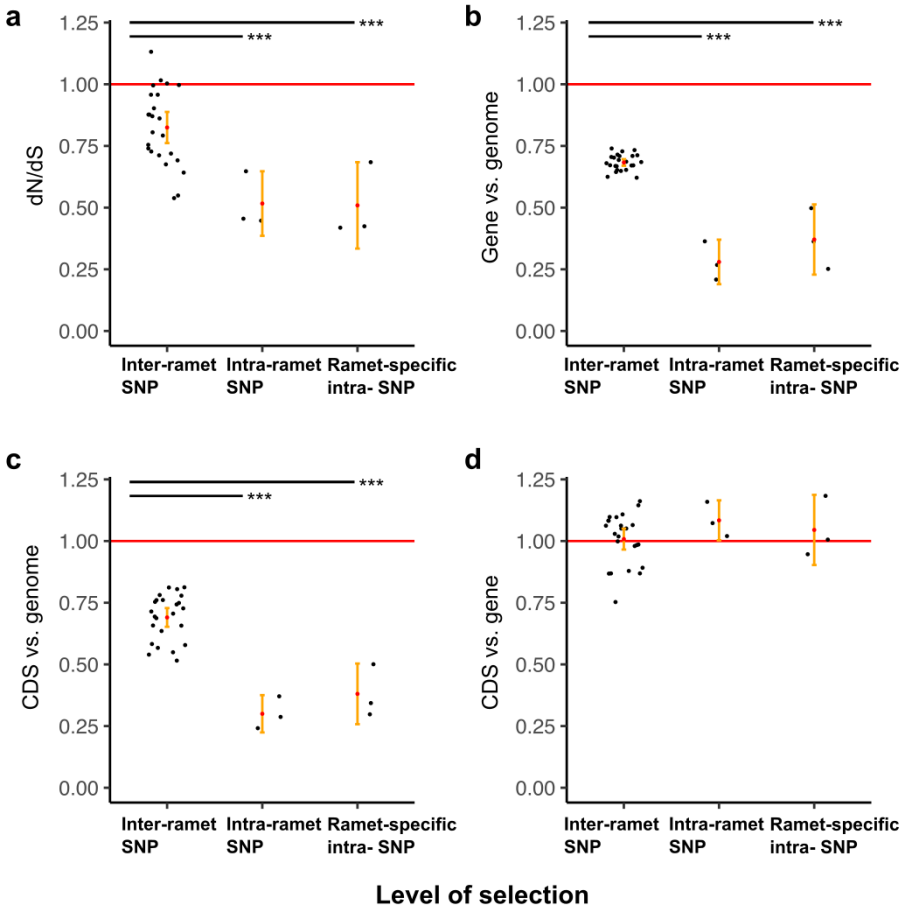


Fig. 4 | Comparison of purifying selection regimes at the inter- and intra-ramet level. In each panel, three levels of selection are distinguished, (i) among ramets based on inter-ramet SNPs (81x dataset for all 24 ramets); (ii) within ramets R08, R10 and R12 (i.e. in mosaic status) based on all intra-ramet SNPs; and (iii) specific to a single ramet (cf. Venn Diagram, Fig. 3a). In all panels, means ±1SD are given; the red line represents the expected value based on selective neutrality. Significant differences among the levels of selection are indicated with asterisks (using ANOVA and Tukey-Kramer for post-hoc comparison; ***$p<0.001$). **a**, dN/dS calculated based on intra-ramet somatic mutations. All three values significantly deviated from neutrality. **b**, Fraction of genic SNPs compared to the entire genome, standardized for the abundance of genic regions. **c**, Fraction of SNPs in coding regions compared to the entire genome, standardized for the proportion of coding regions. **d**, Fraction of coding SNPs within genic regions (intron and exon).

Differences in purifying selection at the intra- versus inter-ramet level

Given the predominance of homozygous sites in the ancestral zygote (99.91%, see Methods), most somatic mutations will result in a homozygous-to-heterozygous transition. The fitness effects of any mutation depend upon their degree of dominance and tissue-specific expression[19]. While many deleterious mutations will be fully recessive and only expressed in haploid gametes[34], a fraction has been shown to be partially dominant and can be subject to selection

even when heterozygous[35]. Hence, we compared the strength of purifying selection at the intra-ramet (i.e., cell population) and inter-ramet levels. The estimation of intra-ramet selection was based on mosaic SNPs identified in three ramets (1370x data; ramet R08, R10, and R12), while assessments of inter-ramet selection were based on fully heterozygous SNPs among all 24 ramets (81x data). We found significant signals of purifying selection at both inter- and intra-ramet level (Fig. 4). The dN/dS ratio indicated weak purifying selection at the inter-ramet level, while this signal was much enhanced at the intra-ramet level (both for ramet-specific mosaic SNPs and all mosaic SNPs) (Fig. 4a). The frequency of somatic genetic variation was considerably depleted in genic loci (Fig. 4b) and within coding sequences (Fig. 4c) relative to the entire genome. Furthermore, there was no significant difference comparing the frequency of SNPs in coding sequences (CDS) relative to all genic locations at both selection levels (Fig. 4d). This indicates that several genic mutations may also be subject to purifying selection when located in untranslated regions or introns. In none of the variables tested did we find a difference among all mosaic SNPs vs. those that were ramet-specific.

Our results confirm our initial hypothesis that selection in clonal seagrass operates at multiple levels. First, we found that purifying selection was stronger among somatic genetic variation while still at the low-frequency mosaic state, i.e., when it occurs among cells within ramets. In contrast, it appeared to be relaxed when present among ramets (Fig. 4a-c). We note that since we infer the selection regime through dN/dS tests, we are only able to address relative differences among plant organizational levels. Notwithstanding, a difference in selection operating at the within- and among ramet level was predicted in earlier models suggesting that selection at the cell level within the module is effective in purging the asexual population from deleterious mutations[11,36-38], especially when the meristem has a layered structure[28]. Additional evidence for multi-level selection came from the comparison of gene function. The fully heterozygous SNPs compared to intra-ramet, mosaic SNPs were each associated with different non-overlapping functions. The former polymorphisms were located within genes enriched (among others) for the molecular function “protein binding”, the cellular component “nucleus” and the biological process “cellular protein modification”. The latter mosaic SNPs were situated in genes significantly enriched for the non-overlapping GO terms “structure of cytoskeleton” (molecular function), “protein microtubule” (cellular component), “protein phosphorylation” (biological process) (all *P*-values <0.005; Supplementary Table 13).

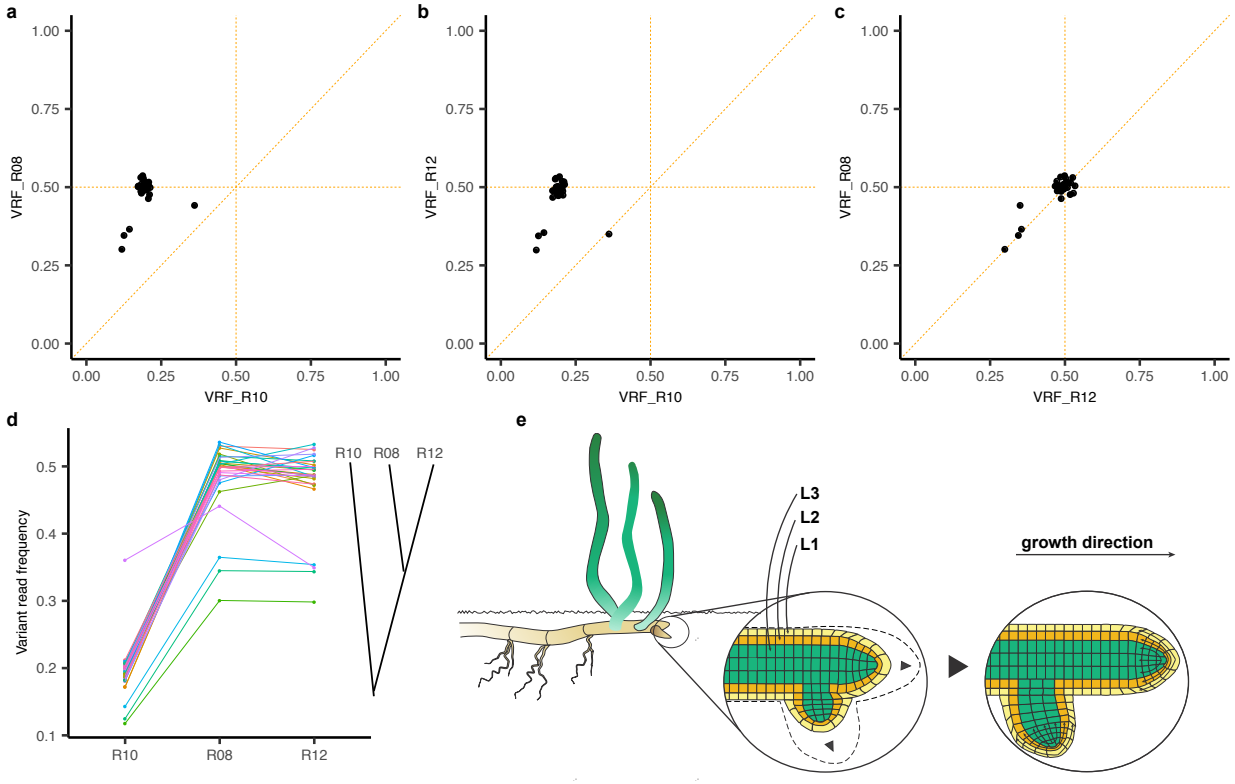


Fig. 5 | Dynamics of intra-ramet somatic polymorphism based on ultra-deep 1,370x re-sequencing. Among a total number of 2,246 SNPs shared among R08, R10, and R12, 1,768 met our coverage threshold of 500x in all three samples. Their variant read frequency (VRF) across the three ramets was calculated and divided into 6 clusters using a Dirichlet process (Extended Data Fig. 5). **a-c**, Pairwise VRFs of Dirichlet cluster #2 encompassing 30 SNPs that display different VRFs among the three ramets. **d**, VRFs of 30 SNPs belonging to Cluster 2 along the inferred genealogy (insert) of the three ramets. **e**, Diagram illustrating how the frequency of intra-module mosaics is transmitted across the hypothetical cell population layers L1… L3 during branching.

Dynamics of intra-ramet allele frequencies reveal somatic genetic drift

We used the inferred genealogy to assess the dynamics of intra-ramet allele frequency among ramets R08, R10 and R12 (Fig. 2d). Under asexual propagation, the somatic mutations accumulate along the rhizome growth path. Accordingly, the more recently two branches diverged, the more similar intra-ramet allele frequencies will be at a given locus. Of a total of 9,208 unique loci, 62% (5,743) were specific to one ramet, followed by 24% (2,246) that were shared among all three (Fig. 3a), and 14% shared by pairs of ramets. The latter two data categories provide an additional independent verification that our SNP calling within ramets was accurate, as the detection of somatic genetic variants was technically independent among the three samples. With a deep coverage of ~1370x, variant read frequency (VRF) should be a reasonable proxy for corresponding allele frequencies (AF), following a binomial distribution

determined by both intra-ramet allele frequency and read coverage at the given locus[39]. We therefore set a conservative coverage threshold of 500x as prerequisite to estimate allele frequencies via VRFs. We focused on the 1,768 somatic genetic variants shared by all 3 ramets and >500x coverage in all three ramets, most of which displayed stable frequencies (Extended Data Fig. 5), a finding which is in accordance with long-term stability of periclinal genetic mosaics (or chimeras) in horticultural plants[40]. They were divided into 6 clusters using a Dirichlet process based on their VRF in the three ramets (Extended Data Fig. 5). Upon visual inspection, cluster 2 encompassing 30 SNPs displayed different intra-ramet allele frequencies among ramets ((Fig. 5a-c), Methods), a pattern that was in concordance with the topology of the NJ-tree (Fig. 2d). Based on their phylogeny (Fig. 2d), ramet R08 and R12 diverged recently; accordingly, they revealed similar frequency dynamics that jointly differed from those of ramet R10 (Fig. 5d). More importantly, 26/30 of VRFs shared by R08 and R12 went from a mosaic to the fully heterozygous state (i.e. $f \approx 0.5$), tracing back to lower frequency alleles in the ancestral R10 where the same SNPs were still in a low-frequency state. Since ancestral (R10) and derived frequencies (R08 and R12) of these 26 loci are characterized by similar allele dynamics, the most parsimonious explanation is that they became fixed in a single asexual sweep driven by drift, selection or both.

While cell populations in meristematic layers in angiosperms are separated (Fig. 5e)[27-29], sometimes cells migrate to other layers[40]. We here provide genomic evidence that this happens at an appreciable frequency in a population of asexually propagating ramets of a single seagrass genet. Results of 347, 592, and 280 SNPs shared among pairs of ramets (R08 – R10, R08 – R12 and R10 – R12), respectively, show a similar pattern where most VRFs remain constant, while only a small fraction shows differences in allele frequency (Supplementary Figure 9a-c). Our results demonstrate the rise of mosaic somatic genetic variation to a fully heterozygous state within a ramet and hence the workings of *somatic genetic drift*.

Discussion

Modular organization, along with clonal reproduction including asexual budding, branching or fission is a widespread life-history trait shared by plants, animals and fungi[5,6,41]. In this study, we trace how somatically generated variation that is initially in a mosaic state is segregated into differentiated ramet genotypes within a large genet of a clonally growing seagrass (Extended Data Fig. 6), a group of marine plants that is known for its very large clonal expansion[21-23]. The amount of somatic genetic variation detected among ramets is roughly 100 times higher than that found among branches of long-lived trees such as oak[42,43] (Supplementary Table 14). Recently, it has been proposed that tree species possess stem-cell like subpopulations of founder cells that undergo a limited number of divisions before branching so as to minimize the total number of cell divisions[42,43]. In the clonal plant *Z. marina* we have no evidence for cell subpopulations that may divide less than the bulk of the meristem. Instead, the most parsimonious explanation for the substantial amount of somatic genetic variation found here is the high number of branching

events in the seagrass that are not restricted by the maximal individual size (e.g., a tree). In contrast to solitary trees, clonal species with indeterminate growth such as corals, reed, ferns, aspen, fungi or seagrasses, can grow, branch, and fragment as long as space for expansion is available. Each branching, in turn, presents an opportunity for somatic genetic drift, eventually resulting in strong genetic heterogeneity among different ramets. It has been speculated as to whether and how long-term vegetative growth may lead to senescence in clonal plants[44,45]. In the northern Baltic Sea, *Z. marina* usually does not complete its sexual life cycle, but it is unclear whether this is driven by too short a growing season or genetic factors. Notwithstanding, given that a conservative age estimate for the genet studied here ranges from 750 - 1,500 yrs (with an upper bound of 6,000 yrs) our results are in agreement with other studies that only found a significant decline in male fertility of aspen genets >4,000 yrs old[45] (yet, we estimate our genet to be younger than that (see Methods)).

Our findings confirm the predicted presence of multiple levels of selection in modular species[16,29,37]. This was evident through different strength of purifying selection acting on mosaic polymorphism within ramets versus somatic genetic differentiation among ramets. Considering that 432 of the fixed SNPs were non-synonymous, our study suggests the potential of somatic genetic variation as a contributor to molecular adaptation (Supplementary Dataset 1). Detecting asexual selective sweeps is difficult as any favorable mutation will be dragged along with its entire mutational background behaving as one large genetic linkage group in the absence of recombination[31,33]. A critical next step will be to test for positive selection[33] at the molecular level and to study possible adaptive consequences of somatic genetic variation in genets of *Z. marina* by controlled experimentation and measurement of physiological performance.
Once fixed, under diploidy mostly as heterozygote polymorphisms, somatic genetic variation will also enter the sexual cycle in flowering plants[2,19], while this is debated in basal metazoan species[46]. Notwithstanding the sexual cycle, our data support the view that under asexuality, ramets are the appropriate elementary level of selection and individuality as the rapid accumulation of genetic variation through somatic genetic drift makes them genetically unique[47]. Whether and how the identified inter-ramet genetic differentiation translates into different phenotypes is an open and highly intriguing question[4,12,13,19,48]. Experimental studies revealed differences in physiological performance among asexually propagating ramets belonging to the same genet in clonal plants and algal thalli[10,17,18,20], suggesting the principal significance of inter-ramet phenotypic selection. We suggest that somatic genetic drift producing inter- ramet genetic differentiation will apply to many clonally proliferating species, providing an additional source of genetic variation for adaptation. This includes, e.g. fungi, clonal trees such as aspen and basal animals such as reef-building corals, hydrozoans, bryozoans, colonial ascidians and any other invertebrate species with indeterminate growth and clonal propagation[6,41,48]. Our study opens up avenues for a comparative approach to study the abundance, fate and selective consequences of somatic genetic variation across the animal, plant and fungal kingdom.

Methods

Study species, site and sampling design. Our study species belongs to the seagrasses, a polyphyletic group of angiosperms (flowering plants) that secondarily returned to the sea[24]. *Zostera marina* L. (eelgrass) is the most widespread marine angiosperm of the northern hemisphere, occurring from subtropical areas to subarctic areas in the Atlantic and Pacific Ocean. The studied eelgrass *Z. marina* bed is located off Angsö Island in the Archipelago Sea, southern Finland (60°06.2' N 21°42.6' E)(Fig. 2a), covering an area of 300 m x 200 m (=6 ha). As in all seagrasses, in *Z. marina*, physiologically independent modules (or "ramets" sensu ref [8]) emerge naturally since rhizome connections are disintegrating in nature after a few years (Fig. 1). Initial microsatellite screening with 8 markers revealed 3 multi-locus genotypes that were only distinguishable at one locus each[21]. However, using whole genome re-sequencing in this study we found only one genet (Fig. 2b), indicating that the results based on limited number of microsatellite markers can be biased by high level of somatic mutations (Extended Data Fig. 4). Taking *Z. marina* shoot density of 311 ramets m^{-2} (ref[49]), a conservative estimate of areal extent of 60,000 m^2 yields 18,660,000 ramets. Assuming a horizontal spread of about 20 cm yr^{-1} (ref [50]) and no intermittent mortality, an expansion to 300 m can be reached in between 750-1,500 years, assuming the ancestral ramet in the center /peripheral to the meadow, respectively. Given a branching rate of 1.587 branches $yr^{-1}apex^{-1}$ (ref[25]), the number of branching events preceding each living ramet should be in between 1,190-2,381. While we acknowledge that initial seagrass patch growth is non-linear, vegetation expansion after patch establishment occurs at a constant rate perpendicular to the vegetation edge[50]. Since in fast growing *Z. marina*, the transition to linear happens within the first 5-10 yrs at a radial patch size of only 1 m (see Fig. 1, ref[25]) we consider this source of error much smaller than the many other uncertainties of estimating the number of past branching events. For an absolute upper age limit we assume 6,000 yrs, the time when the current salinity levels were reached after the re-connection of the then Baltic freshwater lake to the Atlantic Ocean[51]. Twenty-four ramets were collected by SCUBA in 2016 along transect with a shallow and a deep section (Fig. 2a, Supplementary Table 2). Twelve ramets each were collected approximately 10 m distant from each other at approximately 2 m and 4 m water depth, respectively.

Whole-genome resequencing. Bulk DNA of the meristematic region and the basal portions of the leaves, weighing approximately 20-50 mg (fresh mass) was extracted using NucleoSpin Plant II kit (Macherey-Nagel, Germany). We assume that we have sequenced more tissue than the meristematic zone, yet the emerging leaves and shoot parts should simply amplify the genotypic pattern observed among the meristematic cells themselves. DNA concentrations were determined using a Qubit Fluorometer (Thermo Fisher Scientific) and Nanodrop Spectrophotometer (Thermo Fisher Scientific). We checked the integrity of the genomic DNA by agarose gel electrophoresis against a molecular weight (mw) standard and always detected a crisp band at >20 kb in mw. DNA was sent to IKMB (Institute of Clinical Molecular Biology, University of Kiel, Kiel, Germany) on dry ice for library construction and sequencing preparation. TruSeq Nano DNA libraries were constructed for each of the 24 samples with insert size ranging from

479 bp to 515 bp. All the 24 libraries were sequenced on Illumina HiSeq4000 platform at a targeted coverage of ~80x. Based on the topology of a NJ tree, three ramets were selected (Ramets 08, 10, and 12) to be ultra-deeply re-sequenced on Illumina NovaSeq6000 platform at a targeted coverage of ~1000x.

Sequence processing and filtering. The quality of Hiseq paired-end reads (2 x 151 bp) and Novaseq paired-end reads (2 x 151 bp) was assessed by FastQC v0.11.7 (Ref[52]). Reads were then filtered and trimmed using Trimmomatic v0.36 (Ref[53]). Adaptor contaminations were trimmed, and reads with leading or trailing Phred score <20 were removed. The reads were also filtered with a sliding window of size 3; the threshold of average Phred score was set to 15. Trimmed reads with length <36 bp were also removed. FastQC was used for a second round of quality evaluation on the clean reads. After having been trimmed and filtered, the clean Hiseq reads had the average coverage of 81x per sample, and the clean Novaseq reads had the average coverage of 1370x per sample. Clean reads were then mapped against the *Z. marina* reference genome v2.1 (ref[24]) using BWA-MEM v0.7.17 (Ref [54]) with default parameters. The aligned reads were sorted using SAMtools v1.7 (Ref [55]), and duplicated reads were marked using MarkDuplicates tool in GATK v4.0.1.2 (Ref [56]).

Detection of inter-ramet SNPs. *Z. marina* is a diploid plant with no evidence for recent genome duplication[24]. Given the predominance of homozygous sites in the ancestral zygote (proportion of heterozygous loci in the ancestral zygote, 0.09%, see calculation below, "**Estimation of the number of heterozygous loci in the ancestral zygote**"), most somatic mutations will change a homozygous to a heterozygous state, notwithstanding the rare case where a heterozygous locus will change to a homozygous state (Extended Data Fig. 1). Another case would be even rarer where the same mutation double hits the same homozygous locus and change it to a different homozygote, which we detected in 10 cases. Note that we cannot make any inferences with respect to homozygously different loci shared among all 24 samples, compared to the reference genome because these sites are the result of the initial recombination event. Thus, these were not further considered for the somatic mutation detection, based on the consideration above. Once a variant allele is present in all the cells of a ramet, we considered the somatic polymorphism to be fixed heterozygous (Fig. 1, Extended Data Fig. 1), which will result in an expected intra-ramet allele frequency of exactly 0.5. SNPs were called by GATK v4.0.1.2 package. HaplotypeCaller was used to generate general variant calling files (gvcf), which were then combined by CombineGVCFs into a single gvcf file. The merged gvcf file included 81x coverage data for all 24 ramets and 1370x data for ramets 08, 10, and 12. GenotypeGVCFs was used to run joint genotyping.

We first applied GATK hard filtering to filter the SNPs (Extended Data Fig. 2). The density plots of the recommended parameters were drawn with ggplot2 (ref[57]). Based on these plots the following thresholds were applied: Coverage normalized variant depth (QD) < 15.0, strand bias (Fisher test, FS) > 10.0, root mean square of all-sample mapping quality (MQ) < 40.0, reference versus variant reads mapping quality rank sum test (MQRankSum) < -1.5, and depth of coverage (DP) > 8000. The latter filter was implemented to remove the SNPs potentially caused by

genomic duplications, which displayed higher coverage. In a final custom filtering step, we used VCFtools v0.1.15 (Ref [58]) to remove genotypes with genotype quality < 30 (option "--minGQ 30") or depth of coverage < 20 (option "--minDP 20"). After removing these genotypes, VCFtools was used to remove SNPs with more than 2 alleles (options "--min-alleles 2 --max-alleles 2"), SNPs with minor allele frequency < 0.01 (option "--maf 0.01"), and SNPs with more than 4 missing genotypes (option "--max-missing-count 4"). During this step, we also removed SNPs that were homozygous and different to the reference genome, while shared among all sampled ramets, since these are uninformative to assess somatic mutations (Extended Data Fig. 2).

Detection of fixed indels. Indels were also called by GATK v4.0.1.2 package. We first applied GATK hard filtering to filter the indels: Coverage normalized variant depth (QD) < 15.0, strand bias (Fisher test, FS) >10.0, and depth of coverage (DP) > 8000. We then used VCFtools v0.1.15 to remove genotypes with genotype quality < 30 (option "--minGQ 30") or depth of coverage < 20 (option "--minDP 20"). After removing these genotypes, VCFtools was used to remove SNPs with more than 2 alleles (options "--min-alleles 2 --max-alleles 2"), SNPs with minor allele frequency < 0.01 (option "--maf 0.01"), and SNPs with more than 4 missing genotypes (option "--max-missing-count 4").

Verification of inter-ramet SNPs. We developed an independent, non-sequencing based SNP verification method using restriction enzymes with motifs specific to the polymorphic site (Supplementary Figure 2). Accordingly, we searched the reference genome for RE motifs encompassing SNPs. Upon designing fluorescence-labelled primer pairs, a ~400 bp PCR fragment surrounding the target SNP was amplified and subsequently cut by restriction enzymes. Fragments carrying identical sequences to the reference genome possessed intact restriction sites, while fragments with a variant allele lost the restriction site recognition and remained uncut. Fluorescently labelled PCR amplicons were run on Applied Biosystems 3130xl sequencer using the fragment analysis module. If heterozygous, a genotype displayed two different fluorescent signals. All fourteen selected SNPs in combination with all 24 ramets were successfully validated (Supplementary Figure 3). In some instances, we could not obtain clean PCR products using the designed primer pairs which prevented the subsequent SNP verification.

We then verified if inter-ramet SNPs were truly fixed heterozygous (i.e. not occurring in some lower, mosaic frequencies), using the three 1370x sequenced ramets R08, R10 and R12. Under fixation, the intra-ramet allele frequency of both the variant and the ancestral allele is $f = 0.5$. With a coverage of 1370x, the read frequency calculated based on the Novaseq dataset was taken as proxy for intra-ramet allele frequency. Our analysis was restricted to those 96,291 heterozygous genotypes (i.e. ramet * SNP combination) where the coverage was >500x. Confidence intervals were calculated according to a binomial distribution of the variant read frequency centered at 0.5, and with a standard deviation (SD) determined by the coverage (Ref[59]). As an example, when coverage=500, SD would be SD=0.022, two times which was set as the threshold to obtain the ±95% confidence interval. Accordingly, any heterozygous genotype with a VRF in the confidence range [0.5±2*0.022] was considered fixed.

Verification of ramets belonging to a single genet. Based on the high-quality SNPs, we plotted a histogram to assess whether or not the ramets belong to the same genet, i.e. originated from a single zygote via vegetative propagation. Under mitosis the heterozygous loci in the ancestral zygote will remain heterozygous in all offspring cells and ramets, and become visible as one predominant mode of heterozygous loci shared by all 24 ramets. We plotted the number of loci against the number of ramets sharing a particular heterozygous locus. The few missing genotypes (four missing genotypes at most, see filtering step above) were counted as being heterozygous, so that the number of heterozygous ramets would be 24 as long as all the available genotypes were heterozygous. The same analyses were also done based on small insertion /deletion polymorphism (indels). We also simulated a single sexual event among the 24 sampled ramets (Supplementary Figure 1). We ran nine independent simulations where 23 ramets were asexual descendants of the same founder genotype, while ramet R24 was produced by self-fertilization. We assumed 38,831 heterozygous loci, the number of loci that was the basis in Fig. 2b, and no somatic mutation. Under asexual reproduction, all heterozygous genotypes would be passed on to ramet R1-23. However, R24 was produced via sexual reproduction and all 38,831 loci will segregate into three possible genotypes, with a probability of 0.5 for being heterozygous. We generated the genotype of ramet R24 at each of the 38,831 loci using the "random" module in python3, and plotted the resulting frequency distribution of shared heterozygous genotypes the same way as in Fig. 2b.

Estimation of the number of heterozygous loci in the ancestral founding genotype. This approach intended to be conservative with respect to the detectable heterozygosity in the genet, in order to provide an estimate on the likelihood of a heterozygous-to-homozygous mutational transition, and was done prior to hard filtering. Under mitosis, the heterozygous loci in the ancestral zygote will be passed on to all the offspring. The raw SNP panel (before any filtering) from the GATK calling pipeline was used to estimate the number of heterozygous loci in the ancestral zygote (Extended Data Fig. 2). Any SNP with missing genotypes <13 and all the available genotypes were heterozygous, were considered as heterozygous loci in the founder genotype.

Inter-ramet genetic heterogeneity and spatial genetic structure. All SNPs with only one common genotype shared by all 24 ramets were removed as these represented the genetic differences between the reference genome and the ancestral zygote which were passed on to all analyzed ramets (Extended Data Fig. 1, blue). The remaining SNPs represented the somatically derived polymorphisms within the genet. The somatically derived indels were obtained in the same way. Based on the SNPs, pairwise genetic distances between each pair of ramets were calculated and visualized using ggplot2 (heatmap plot). The genetic distance was quantified by the number of different alleles. SnpEff (Ref[60]) was used to annotate the SNPs. Based on the annotated nonsynonymous SNPs, the pairwise genetic distances between each pair of ramets were also calculated and visualized. Based on the fully inter-ramet SNP panel, plot.phylo function in the ape package (Ref[61]) for R (Ref[62]) was used to construct a neighbor-joining tree to examine how the genetic similarity corresponds to spatial sampling pattern. Bootstrap support

(1,000 times) was obtained using the aboot function in the poppr package (Ref[63]) for R. The same analyses were also done based on indels. We calculated the number of SNPs exclusively shared by two ramets (Supplementary Table 6), in order to verify whether distant ramet pairs based on the NJ-tree tree never exclusively share SNPs. This should apply if the constructed NJ tree depicts the true ramet topology.

Detection of structural variation. In order to detect larger structural variants, we applied CNVnator v0.3.3 (ref[64]) to find target loci followed by IGV v2.7.0 (ref [65]) for a visual check. We first selected 3 samples (ramets R05, R07, and R20) representing the largest distances within the NJ tree. Subsequently, CNVnator was used to call structural variations for each sample. We focused on deletions relative to the reference genome, considering that duplications were much more difficult to verify. A custom-made Python3 script was used to convert the CNVnator output to bed file format. Bedtools v2.27.1 (ref [66]) was used to find the overlap between R05 and the other two samples. The locus was marked as target if it was called as a deletion in R05, and was missing in at least one of the other two samples. IGV was used to check if the deletion (relative to the reference genome) was true and showed polymorphism among the three ramets. If both requirements were met, it was further checked among all 24 ramets.

Detection of intra-ramet SNPs. In order to accurately assess the level of intra-ramet somatic polymorphisms, we used the 1370x ultra-deep resequencing in ramets R08, R10 and R12 to quantify intra-ramet SNPs. This type of SNPs is only present in a subset of cells within the ramet (Extended Data Fig. 1). Hence, the novel variant allele has the intra-ramet allele frequency $f < 0.5$, depending on how many cells possess the variant allele. As reference, we took the reference genome that was sampled ~22 km distance from the studied location. The SNP calling was run independently for each of the three ramets using Mutect2 (Ref[67]). The SNPs were filtered by FilterMutectCalls following the recommendations of the authors. Since the control sample is not equal with the ancestral zygote, the calling results included the genetic differences between the control sample and the ancestral zygote which should be shared by all 24 analyzed ramets. These loci were located based on the raw SNP panel (before any filtering) from the GATK calling pipeline (Extended Data Fig. 2). Any SNP with missing genotypes <13 and all the available genotypes were same, were considered as genetic differences between the ancestral zygote and the control sample, thus 215,518 loci were checked and removed, if present in the Mutect2 calling results.

Verification of intra-ramet SNPs. We also verified the intra-ramet variation using a sequencing-free, restriction enzyme based method. During the verification of inter-ramet SNPs above we found that often the wild-type genotypes (both alleles identical to reference genome) were not digested completely, leaving a small fluorescence signal. In order to reliably detect mosaic SNPs, incomplete digestion would hence spuriously suggest low-frequency variants. Therefore, we searched for mosaic SNPs encompassing RE motifs where the heterozygous variant allele converted a non-restriction site to a restriction site. Thus, after digestion shorter fragments will only appear if the variant allele is present, an observation that cannot be biased by

undigested wild-type alleles. We verified 15 cases in total (3 cases for R08, 4 cases for R10, and 8 cases for R12) (Supplementary Figure 7, Supplementary Table 9 & 10).

Fixed vs. mosaic somatic polymorphisms. We estimated intra-ramet allele frequency, using the variant read frequency (VRF) calculated from the ultra-deep Novaseq resequencing dataset. Under 1370x coverage VRF should be a reasonable proxy for allele frequencies, and we plotted the histogram of VRF for each ramet. To detect the pattern of recent mutational input at low allele frequencies, we also plotted histograms only with ramet-specific somatic genetic variation (Venn diagram, cf. Fig. 3a). We then distinguished mosaic from fixed somatic genetic variation. For the mosaic loci, we focused on those with frequency <0.5, i.e. those below fixation in one haplotype among all ramet cells, based on above argument that somatic mutations will most likely change a homozygous to a heterozygous site. As expected, VRF histograms of fixed SNPs centered around f= 0.5, while there was a clear break at f~0.4 that separated the low-frequency SNPs, which was set as the threshold for distinguishing between intra-ramet and fixed SNPs (Fig. 3b-g; dashed line).

Detection of molecular selection at the intra- and inter-ramet level. Clonal species such as *Z. marina* may be subject to two different levels of selection, among cells within ramets, and among differentiated ramets. We took the 7,054 inter-ramet SNPs identified using the 81x coverage dataset to study inter-ramet selection. Given the predominance of homozygous sites in the ancestral founder genotype (99.91%), most somatic mutations cause a homozygous-to-heterozygous change (see "**Estimation of the number of heterozygous loci in the ancestral founding genotype**"). We first removed the 10 SNPs displaying two different homozygous genotypes (0/0 + 1/1) among the 24 samples. For the remaining loci, we assumed that they originated from homozygous-to-heterozygous mutations of the founder genotype. This data set was compared to the intra-ramet SNPs identified using ultra-deep (1370x) sequencing that are representative of selection among different cell lineages within the ramet. SnpEff was used to identify nonsynonymous and synonymous changes. To calculate the dN/dS ratio, we determined the total number of nonsynonymous sites and synonymous sites in the genome, which were calculated based on the CDS sequences of the reference genome[24]. Although applying dN/dS ratio at population level has been questioned[68], the focus of our analysis are the different intensities of selection which were significantly different at the inter- and intra-ramet level. The same SNP panels were also used to analyze the distribution of mutations in the genome. They were categorized into being located in CDS, gene, or other locations in the genome. We calculated (i) the ratio of mutations distributed in genes relative to the whole genome (ii) the ratio of mutations distributed in CDS relative to the whole genome (iii) the ratio of mutations distributed in CDS relative to genes, and compared the levels of selection using pairwise *t*-tests for unequal variances. Quantities were normalized based on the total length proportion of each of the genomic categories, e.g. for the ratio of mutations distributed in CDS relative to the whole genome, we calculated the ((number of SNPs in CDS)/(total number of base pairs in the CDS)) / ((number of SNPs in the whole genome)/(total number of base pairs in the whole genome)). A custom-made Python3 script was used to calculate the transition:transversion ratio between the

reference allele and variant allele based on the combination of fixed polymorphisms among the 24 ramets. At each level, the combination of the nonsynonymous mutations (inter-ramet, combination of SNPs of 24 ramets; intra-ramet, combination of SNPs of 3 ramets) served as input for the GO enrichment analysis which was performed using R 3.6.1 software and the topGO 2.36.0 package[69]. The *weight* algorithm and Fisher's exact test were used to detect significant enriched GO terms and show GO graph topology. We followed the recommendation of topGO authors to interpret the *p*-values as corrected or not affected by multiple testing. The GO term analysis was explorative, hence the threshold for statistical significance was set at α=0.01 based on recommendations[69]. The required gene universe was obtained from the *Z. marina* genome[24].

Dynamics of somatic genetic variation among ramets. We intended to demonstrate somatic genetic drift directly using the phylogenetic history of ramets. To do so, we followed the inferred genealogy of shared somatic polymorphisms, in particular whether and how often they rise to fixation. Although all three 1370x sequenced ramets R08, R10 and R12 share the same zygotic ancestor, the NJ tree topology revealed ramet-specific paths of specific and shared somatic polymorphism. First, we identified somatic SNPs shared by all three ramets, setting a minimal sequence coverage of 500x. Each SNP contained three-dimensional information, i.e., its VRFs in R08, R10, and R12, respectively. We used unsupervised machine learning implemented into a Dirichlet process[70] to divide shared SNPs into different clusters (Extended Data Fig. 5), based on both sharedness and frequency of VRFs among the three ramets. The analysis was done using R package "dirichletprocess". The "DirichletProcessMvnormal" function was used to create a dp object. A fit function was used on this object to infer the cluster parameters. Visual inspection revealed that only cluster 2 displayed different VRFs among the three ramets. Accordingly, the VRFs of these loci across ramets R08, R10 and R12 were plotted (Fig. 5a-c), and frequency change along the inferred genealogy was qualitatively examined (Fig. 5d). The pairwise VRF difference was also analyzed for somatic polymorphisms shared by two ramets (Venn diagram, cf. Fig. 3a, Supplementary Figure 9).

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data Availability

DNA sequence data are available in the NCBI short read archive, BioProject no. PRJNA557092, SRA accession no. SRR9879327- SRR9879353. Data on genes / putative gene functions, and on SNP verification are included into tables of the Supplementary Material. If not included into the supplementary material, data visualized in figures 3, 4, 5 are deposited in PANGAEA[71].

Code Availability

Custom-made computer code is available at Github https://github.com/leiyu37/Finnish_eelgrass_milleniumClone.git.

References

1 Behjati, S. *et al.* Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422 (2014).

2 Wang, L. *et al.* The architecture of intra-organism mutation rate variation in plants. *PLOS Biology* **17**, e3000191 (2019).

3 Frank, S. A. Somatic evolutionary genomics: Mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proceedings of the National Academy of Sciences* **107**, 1725 (2010).

4 Pineda-Krch, M. & Lehtilä, K. Costs and benefits of genetic heterogeneity within organisms. *Journal of Evolutionary Biology* **17**, 1167-1177 (2004).

5 Honnay, O. & Bossuyt, B. Prolonged clonal growth: escape route or route to extinction? *Oikos* **108**, 427-432 (2005).

6 Buss, L. W. Evolution, development, and the units of selection. *Proceedings of the National Academy of Sciences* **80**, 1387-1391 (1983).

7 Jackson, J. B. C., Buss, L. W. & Cook, R. E. *Population biology and evolution of clonal organisms*. (Yale University Press, 1985).

8 Harper, J. L. *Population Biology of Plants*. (Academic Press, 1977).

9 Gaul, H. Die verschiedenen Bezugssysteme der Mutationshäufigkeit bei Pflanzen, angewendet auf Dosis-Effektkurven. *Zeitschrift für Pflanzenzüchtung* **38**, 63-76 (1957).

10 Larkin, P. J. & Scowcroft, W. R. Somaclonal variation — a novel source of variability from cell cultures for plant improvement. *Theoretical and Applied Genetics* **60**, 197-214 (1981).

11 Klekowski, E. J. & Kazarinovafukshansky, N. Shoot Apical Meristems and Mutation - Selective Loss of Disadvantageous Cell Genotypes. *American Journal of Botany* **71**, 28-34 (1984).

12 Sutherland, W. J. & Watkinson, A. R. Somatic mutation: Do plants evolve differently? *Nature* **320**, 305-305 (1986).

13 Fagerström, T., Briscoe, D. A. & Sunnucks, P. Evolution of mitotic cell-lineages in multicellular organisms. *Trends in Ecology and Evolution* **13**, 117-120 (1998).

14 Lynch, M. Evolution of the mutation rate. *Trends in Genetics* **26**, 345-352 (2010).

15 Gill, D. E., Chao, L., Perkins, S. L. & Wolf, J. B. Genetic mosaicism in plants and clonal animals. *Annual Review in Ecology and Systematics* **26**, 423-444 (1995).

16 Antolin, M. F. & Strobeck, C. The population genetics of somatic mutations. *The American Naturalist* **126**, 52-62 (1985).

17 Breese, E. L., Hayward, M. D. & Thomas, A. C. Somatic selection in perennial ryegrass. *Heredity* **20**, 367 (1965).

18 Santelices, B., Gallegos Sánchez, C. & González, A. V. Intraorganismal genetic heterogeneity as a source of genetic variation in modular macroalgae. *Journal of Phycology* **54**, 767-771 (2018).

19 Schoen, D. J. & Schultz, S. T. Somatic Mutation and Evolution in Plants. *Annual Review of Ecology, Evolution, and Systematics* **50**, 49-73 (2019).
20 Simberloff, D. & Leppanen, C. Plant somatic mutations in nature conferring insect and herbicide resistance. *Pest Management Science* **75**, 14-17 (2019).
21 Reusch, T. B. H. & Boström, C. Widespread genetic mosaicism in the marine angiosperm *Zostera marina* is correlated with clonal reproduction. *Evol Ecol* **25**, 899-913 (2011).
22 Arnaud-Haond, S. *et al.* Implications of Extreme Life Span in Clonal Organisms: Millenary Clones in Meadows of the Threatened Seagrass Posidonia oceanica. *PLOS One* **7**, e30454 (2012).
23 Bricker, E., Calladine, A., Virnstein, R. & Waycott, M. Mega Clonality in an Aquatic Plant—A Potential Survival Strategy in a Changing Environment. *Frontiers in Plant Science* **9**(2018).
24 Olsen, J. L. *et al.* The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature* **530**, 331–335 (2016).
25 Sintes, T., Marbà, N. & Duarte, C. M. Modeling nonlinear seagrass clonal growth: Assessing the efficiency of space occupation across the seagrass flora. *Estuaries and Coasts* **29**, 72-80 (2006).
26 Sung, W. *et al.* Evolution of the Insertion-Deletion Mutation Rate Across the Tree of Life. *G3: Genes|Genomes|Genetics* **6**, 2583 (2016).
27 Poethig, S. Genetic mosaics and cell lineage analysis in plants. *Trends in Genetics* **5**, 273-277 (1989).
28 Pineda-Krch, M. & Lehtilä, K. Cell Lineage Dynamics in Stratified Shoot Apical Meristems. *Journal of Theoretical Biology* **219**, 495-505 (2002).
29 Klekowski, E. J. Plant clonality, mutation, diplontic selection and mutational meltdown. *Biological Journal of the Linnean Society* **79**, 61-67 (2003).
30 Burian, A., Barbier de Reuille, P. & Kuhlemeier, C. Patterns of Stem Cell Divisions Contribute to Plant Longevity. *Current Biology* **26**, 1385-1394 (2016).
31 Lang, G. I. *et al.* Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500**, 571-574 (2013).
32 Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nature Genetics* **48**, 238 (2016).
33 Williams, M. J. *et al.* Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics* **50**, 895-903 (2018).
34 Schultz, Stewart T. & Scofield, Douglas G. Mutation Accumulation in Real Branches: Fitness Assays for Genomic Deleterious Mutation Rate and Effect in Large-Statured Plants. *The American Naturalist* **174**, 163-175 (2009).
35 Willis, J. H. Inbreeding Load, Average Dominance and the Mutation Rate for Mildly Deleterious Alleles in *Mimulus guttatus*. *Genetics* **153**, 1885-1898 (1999).
36 Otto, S. P. & Orive, M. E. Evolutionary consequences of mutation and selection within an individual. *Genetics* **141**, 1173 (1995).

37 Orive, M. E. Somatic Mutations in Organisms with Complex Life Histories. *Theoretical Population Biology* **59**, 235-249 (2001).
38 Otto, S. P. & Hastings, I. M. Mutation and selection within the individual. *Genetica* **102**, 507 (1998).
39 Tarabichi, M. *et al.* Neutral tumor evolution? *Nature Genetics* **50**, 1630-1633 (2018).
40 Frank, M. H. & Chitwood, D. H. Plant chimeras: The good, the bad, and the 'Bizzaria'. *Developmental Biology* **419**, 41-53 (2016).
41 Smith, M. L., Bruhn, J. N. & Anderson, J. B. The fungus *Armillaria bulbosa* is among the largest and oldest living organisms. *Nature* **356**, 428-431 (1992).
42 Schmid-Siegert, E. *et al.* Low number of fixed somatic mutations in a long-lived oak tree. *Nature Plants* **3**, 926-929 (2017).
43 Plomion, C. *et al.* Oak genome reveals facets of long lifespan. *Nature Plants* **4**, 440-452 (2018).
44 de Witte, L. C. & Stöcklin, J. Longevity of clonal plants: why it matters and how to measure it. *Annals of Botany* **106**, 859-870 (2010).
45 Ally, D., Ritland, K. & Otto, S. P. Aging in a Long-Lived Clonal Tree. *PLOS Biology* **8**, e1000454 (2010).
46 Buss, L. W. *The Evolution of Individuality*. (Princeton University Press, 1987).
47 Santelices, B. How many kinds of individuals are there? *Trends in Ecology and Evolution* **14**, 152-155 (1999).
48 Van Oppen, M. J. H., Souter, P., Howells, E. J., Heyward, A. & Berkelmans, R. Novel Genetic Diversity Through Somatic Mutations: Fuel for Adaptation of Reef Corals? *Diversity* **3**, 405-423 (2011).
49 Gustafsson, C. & Boström, C. Algal mats reduce eelgrass (*Zostera marina* L.) growth in mixed and monospecific meadows. *J Exp Mar Biol Ecol* **461**, 85-92 (2014).
50 Reusch, T. B. H., Chapman, A. R. O. & Gröger, J. P. Blue mussels (*Mytilus edulis*) do not interfere with eelgrass (*Zostera marina*) but fertilize shoot growth through biodeposition. *Mar. Ecol. Prog. Ser.* **108**, 265-282 (1994).
51 Gustafsson, B. G. & Westman, P. On the causes for salinity variations in the Baltic Sea during the last 8500 years. *Paleoceanography* **17**, 12-11-12-14 (2002).
52 Andrews, S. *FastQC: a quality control tool for high throughput sequence data*. (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> 2010).
53 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
54 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
55 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

56 Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics* **43**, 11.10.11-11.10.33 (2013).
57 Ginestet, C. ggplot2: Elegant Graphics for Data Analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**, 245-246 (2011).
58 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
59 Noorbakhsh, J. & Chuang, J. H. Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. *Nature Genetics* **49**, 1288 (2017).
60 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80-92 (2012).
61 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526-528 (2018).
62 Core Team, R. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, Vienna, Austria, URL https://www.r-project.org/. 2019).
63 Kamvar, Z. N., Brooks, J. C. & Grünwald, N. J. Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics* **6**, 208 (2015).
64 Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research* **21**, 974-984 (2011).
65 Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant Review with the Integrative Genomics Viewer. *Cancer Research* **77**, e31 (2017).
66 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
67 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31**, 213 (2013).
68 Bataillon, T. *et al.* Inference of Purifying and Positive Selection in Three Subspecies of Chimpanzees (Pan troglodytes) from Exome Sequencing. *Genome Biology and Evolution* **7**, 1122-1132 (2015).
69 Alexa, A. & Rahnenfuhrer, J. *topGO: Enrichment Analysis for Gene Ontology*. (R package version 2.36.0.a, 2019).
70 Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353-357 (2015).
71 Yu, L. *et al*. Data from: Genomic data of marine flowering plant *Zostera marina*. PANGAEA fileset. https://doi.org/10.1594/PANGAEA.910541 (2020).

Acknowledgements
This study was supported by a 4-year PhD scholarship from the China Scholarship Council to L.Y. and by the Åbo Akademi University Foundation to C.B. We thank Jeanine L. Olsen and

Benjamin Werner for helpful comments on earlier versions of the manuscript and in particular Iliana Baums for discussing methods for sequence-independent SNP verification. We thank the Archipelago Centre Korposström (Finland) for excellent working facilities and Susanne Landis for creating some of the illustrations. Sampling permit # MH 5448/2015 was granted through Parks and Wildlife Finland (Metsähallitus).

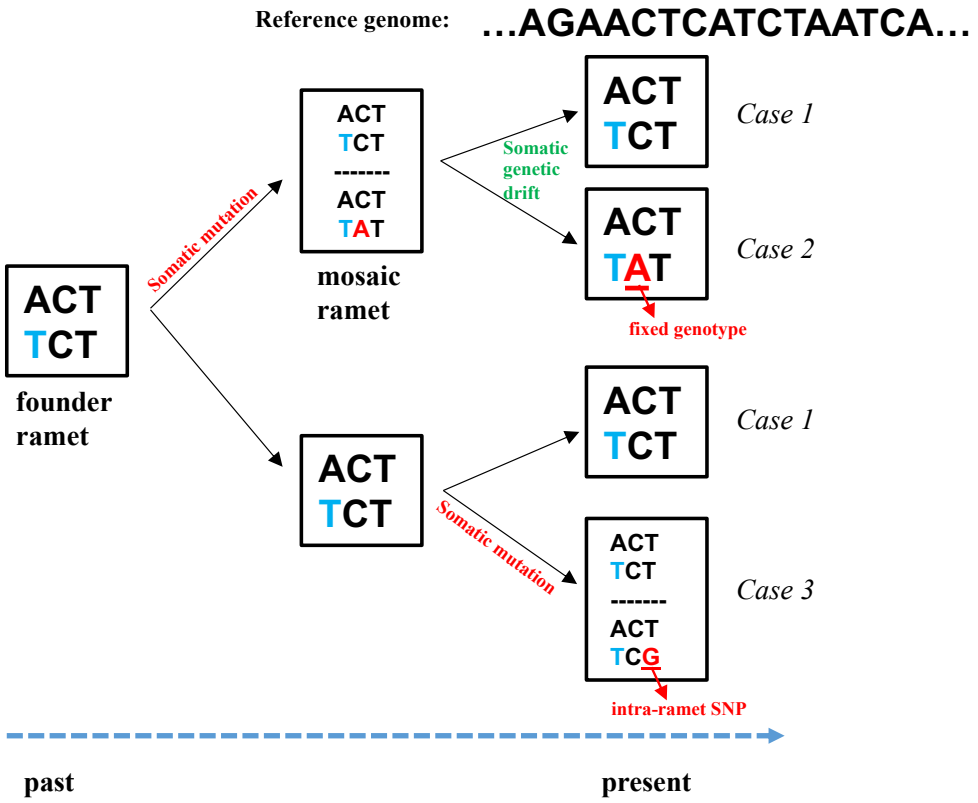
Author contributions

T.B.H.R. and L.Y. designed the study, C.B. provided access to the biological samples and conducted the sampling, T.B. and T.D. assisted with the bioinformatic analyses, S.F. prepared the libraries and performed the re-sequencing, L.Y., T.D. and T.B.H.R. analyzed, interpreted and discussed the results, T.B.H.R. and L.Y. wrote the paper, with input from all authors.

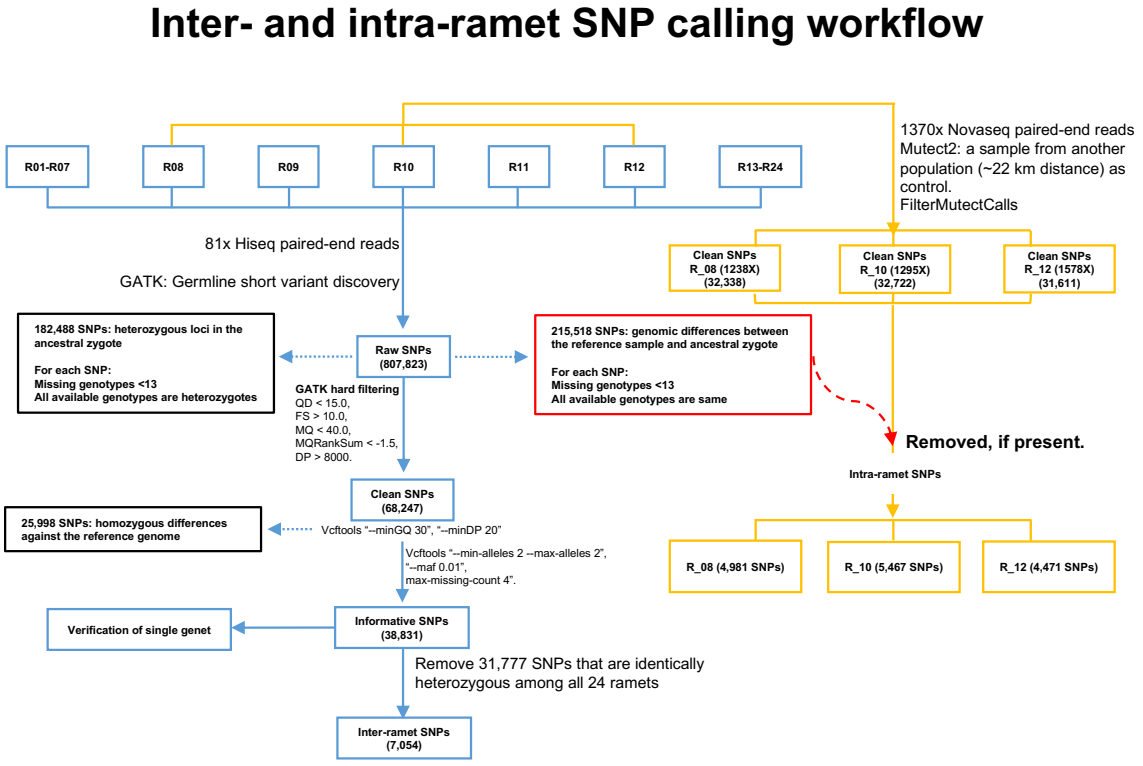
Competing interests

The authors declare no competing interests.

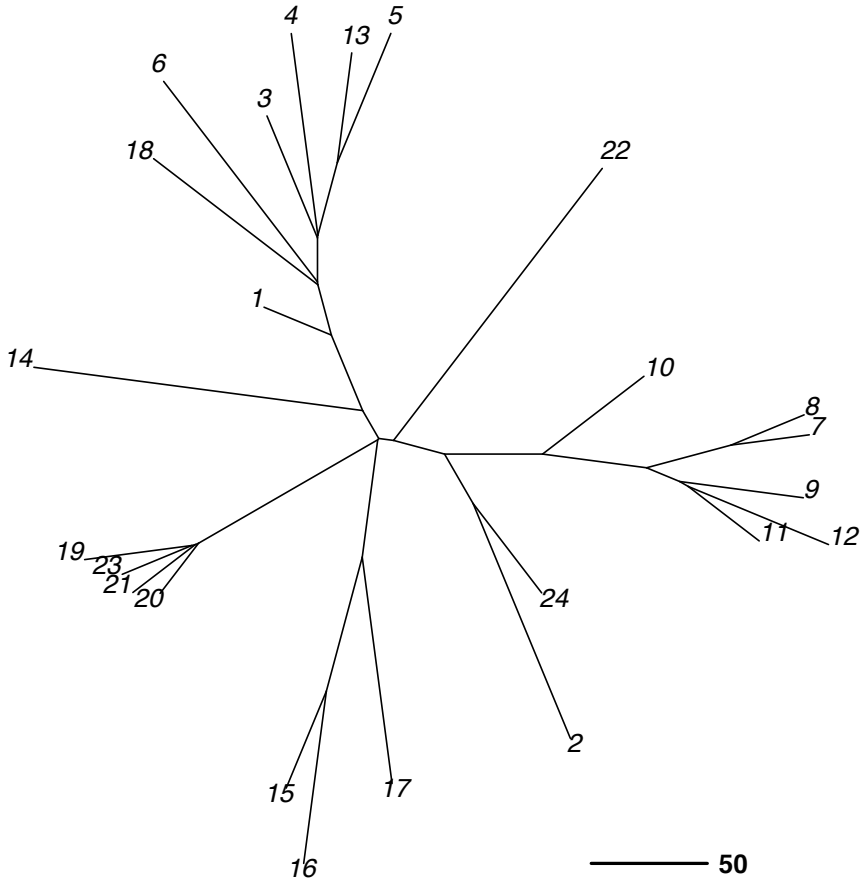
Extended Data



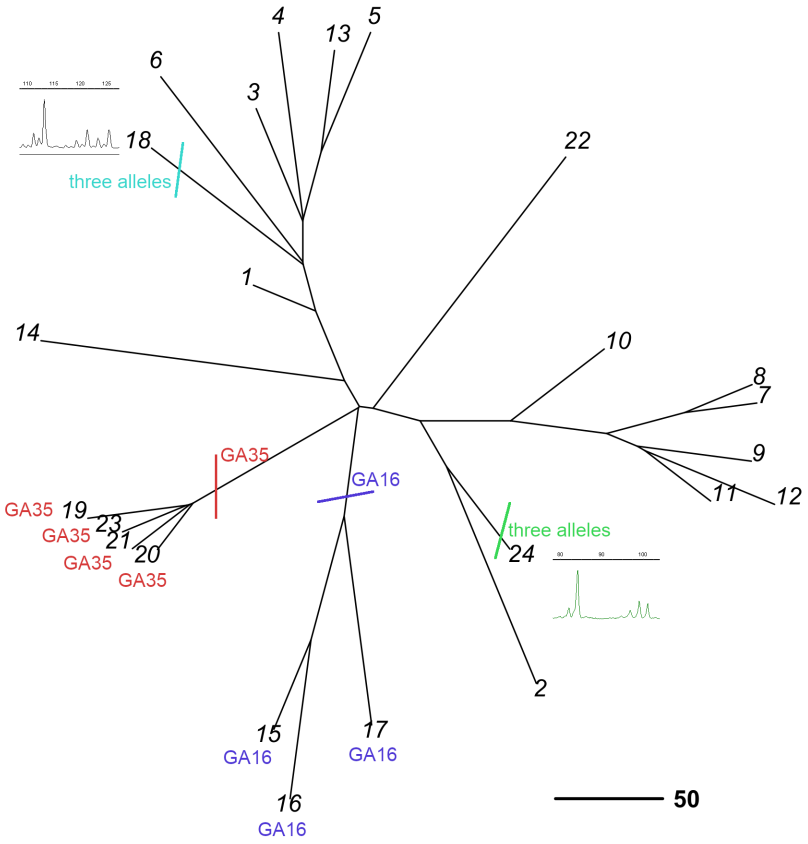
Extended Data Fig. 1 | Somatic genetic drift in clonal organisms and possible categories of single nucleotide polymorphisms (SNPs). Multicellular organisms originate from a single zygote and achieve growth and development via mitosis. The diagram indicates possible genotypes at a locus within the ramet (=box). Heterozygous polymorphisms present in the zygote (blue) are inherited in all offspring cells, and result in consistent heterozygous calls across all ramets. This class of SNPs needs to be excluded before making inferences on somatically derived polymorphisms (cf. Extended Data Fig. 2, 'Remove 31,777 SNPs that are identically heterozygous among all 24 ramets'). During growth, cells acquire somatic mutations (red) owing to mitotic errors that initially emerge as genetic mosaics, that is the somatic polymorphisms are present in only a subset of cells. When clonal organisms form another iterative unit (=ramet), a subpopulation of cells in the parent ramet are progenitors for the asexual offspring. The genetic bottleneck accompanying this random sampling process determines the segregation of somatic polymorphisms into the new ramets through somatic genetic drift. This process may convert intra-ramet SNP into different fixed ramet genotypes (cf. Fig. 5d). In case 1, somatic genetic drift restores the wild-type genotype present in the zygote. In cases 2, somatic genetic drift separates the intra-ramet SNP at that locus into a fixed state. In case 3, a novel mutation emerges as an intra-ramet SNP, which would require additional rounds of somatic genetic drift to either being fixed or lost.



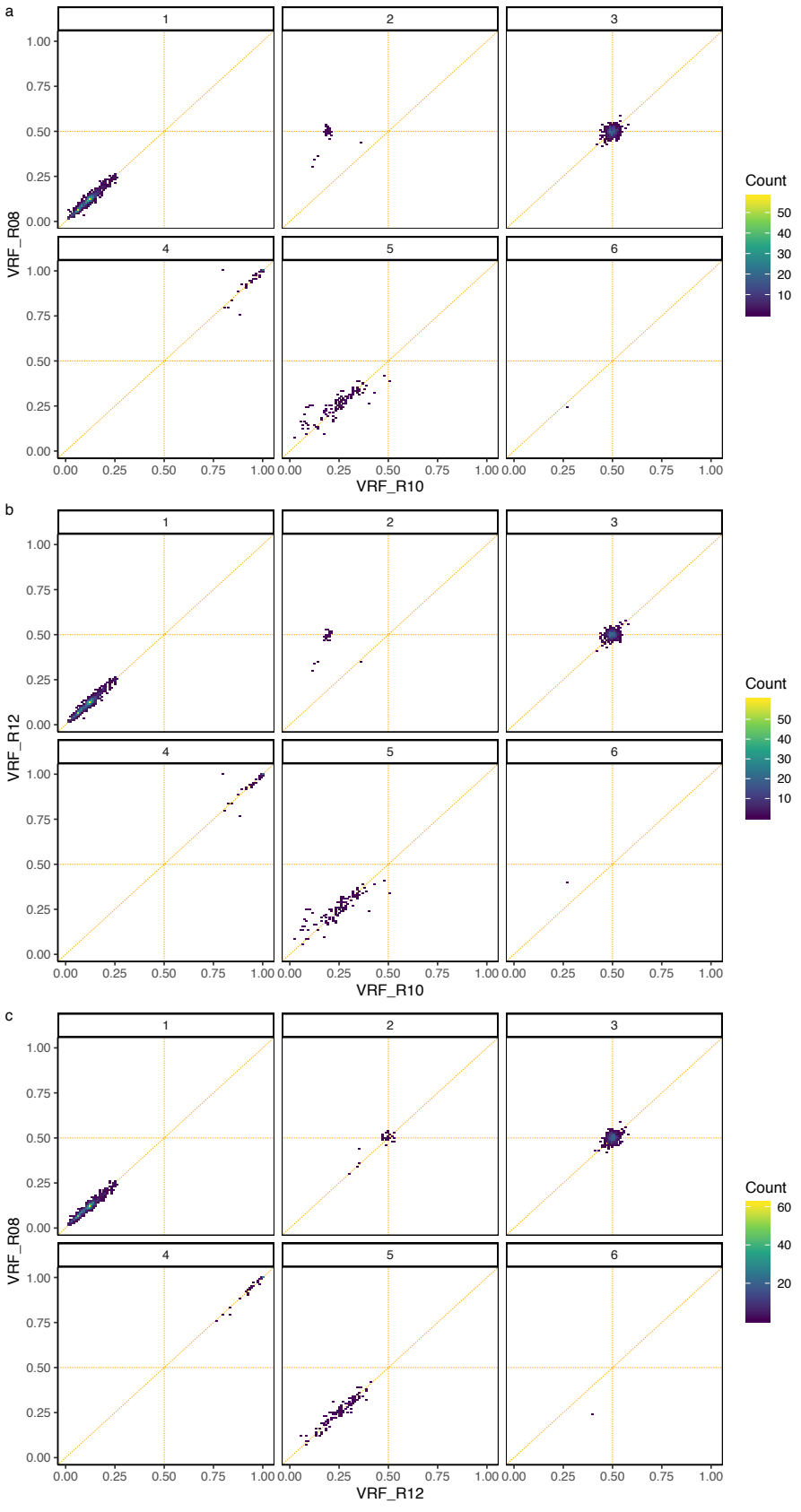
Extended Data Fig. 2 | Workflow for inter- and intra-ramet SNP calling. All 24 ramets were sequenced at an average coverage of 81x on an Illumina Hiseq 4000 platform (left, for details see Supplementary Table 3), while a subset of three ramets (R08, R10, and R12) were further sequenced at an average coverage of 1370x on Novaseq 6000 platform (right, for details see Supplementary Table 8). The filtered 81x dataset was used to (i) estimate the number of heterozygous loci in the initial zygote in a conservative way with respect to not underestimate the true nucleotide diversity, (ii) call fixed SNPs among ramets, and to (iii) identify SNPs that were genetically different between the reference genome and the initial zygote. Here, 'informative' SNPs with respect to a somatic generation of polymorphism refer to all SNPs but those homozygously different to the reference. Informative SNPs were used to verify the clonality of the 24 samples, and to make inferences on somatic mutations after subtracting the identical heterozygous loci shared among R01-24. The 1370x dataset was used to call intra-ramet somatic polymorphisms (composed of both, fixed heterozygous genotypes and mosaic ones with $f < 0.5$), which allowed us to check the proportion of mosaic intra-ramet SNPs vs. fixed heterozygous genotypes. In the 81x data set, custom filtering steps (hard filtering) using the GATK pipeline with conservative parameters intended to minimize the error of detecting false positive SNP calls. During subsequent custom filtering using the Vcftools, identical homozygous SNPs with respect to the reference were removed. In the ultra-deep sequenced data set, each of the three SNP calling approaches were technically independent and thus verified. Note that for the intra-ramet level, the analysis focused on SNPs that were co-occurring in two or three ramets.



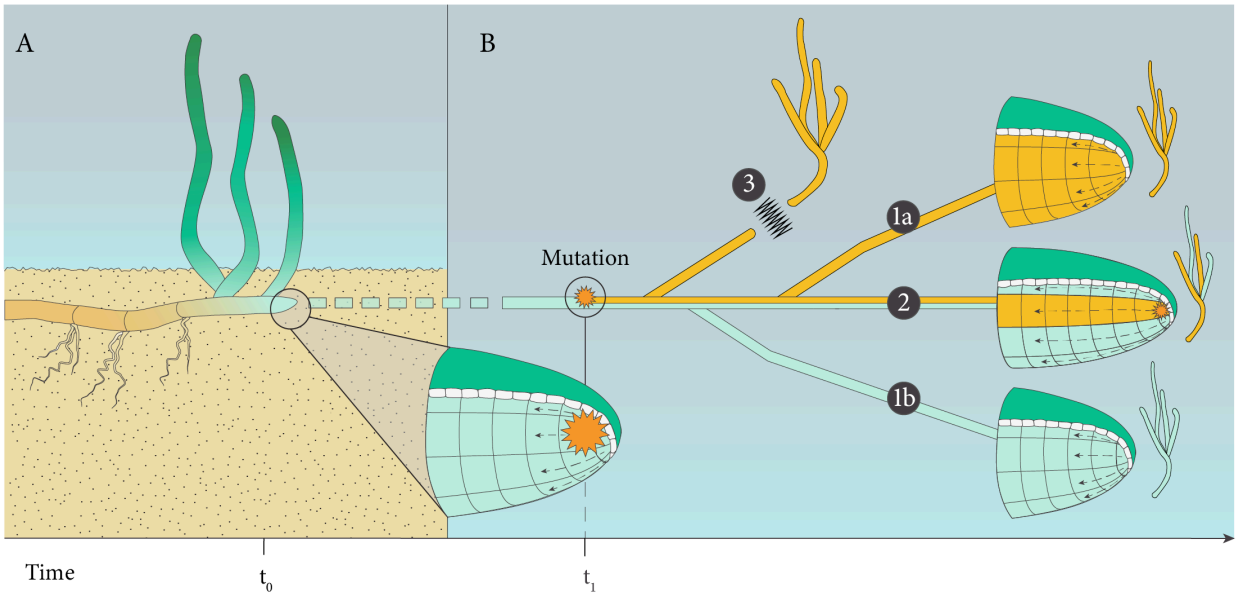
Extended Data Fig. 3 | Neighbor-joining tree using insertion/deletion (indel) based genetic differences among 24 seagrass *Zostera marina* ramets. The 1,654 inter-ramet indel polymorphisms among the 24 ramets were used to quantify the pairwise genetic differences as number of different alleles. The genetic distance matrix was used to construct a neighbor-joining tree, which displays a similar topology to that revealed by inter-ramet SNPs (Fig. 2d). Scale bar = 50 differences.



Extended Data Fig. 4 | Microsatellite polymorphisms among 24 seagrass *Zostera marina* ramets. We genotyped all ramets for 8 microsatellite loci (Genbank accession numbers: AJ009904.1, AJ009901.1, AJ249304.1, AJ249306.1, AJ009898.1, AJ009900.1, AJ249307.1, and AJ249305.1) and detected a total of 7 somatic polymorphisms as microsatellite alleles not present in the predominant genotype. Their presence/absence among the 24 ramets is consistent with the NJ tree based on 7,054 SNPs (Fig. 2d) and 1,652 indels (Extended Data Fig. 3). Two polymorphisms were present in more than one ramet and their appearance within one branch is indicated by the microsatellite locus designation. Small insets show two identified three-allelic, mosaic genotypes with novel microsatellite alleles that were only present in one ramet each.


a
1
2
3
4
5
6
VRF_R08
VRF_R10
Count
50
40
30
20
10
b
VRF_R12
VRF_R10
Count
50
40
30
20
10
c
VRF_R08
VRF_R12
Count
60
40
20
0.00
0.25
0.50
0.75
1.00

Extended Data Fig. 5 | Dirichlet clustering of variant read frequency of SNPs shared by 3 ramets based on 1370x resequencing. The 2,246 SNPs shared among all three ramets were filtered by a minimum coverage of 500x after which 1,768 SNPs were retained. These were divided into clusters using a Dirichlet process. All pairwise VRF among ramets R08, R10 and R12 of the resulting six Dirichlet clusters (top number) are depicted. Each panel was divided into grids, a color scale depicts the density of points in each grid point representing SNP counts. Lines indicate expected allele frequencies for fixed heterozygote SNPs at x = 0.5 and y = 0.5, respectively, while the line x = y depicts equal allele frequencies among ramets. a, VRF in R08 vs. R10. b, VRF in R12 vs. R10. c, VRF in R08 vs. R12.



Extended Data Fig. 6 | *Zostera marina* as a clonal plant. Ramets are physiologically independent and genetically unique 'individuals'. Somatic mutations emerge as intra-ramet SNPs. Along with growth and branching of the genet, somatic genetic drift separates the intra-ramet SNP into different fixed genotypes among ramets (case 1a or 1b), while in others, they may persist as mosaic, intra-ramet polymorphism (case 2). Note that ramets quickly become physiologically independent, as the rhizome connection among them will rot away within one or two years (case 3).

Supplementary Tables

Supplementary Table 1: Glossary of terms.

Term	Definition
Clonal species	Modular species composed of modules that become detached and achieve autonomy in a mode of asexual reproduction
Fixed genotype	Genotype that is shared among all cells within a ramet
Genet	Asexual population of ramets that are descendants of the same zygote, often used synonymously to “clone” in clonal species.
Genetic mosaic	Multicellular organism composed of cells displaying different genotypes, either caused by somatic mutation or chimerism/fusion.
Inter-ramet SNP	Genomic polymorphism displaying different fixed genotypes among an asexual population of ramets belonging to the same genet.
Intra-ramet SNP	Genomic polymorphism displaying different genotypes among cells within the ramet.
Modular species	Multicellular organism composed of iterative units of similar parts (modules), that either stay together (tree), or become independent (clonal species)
Module	Iterative morphological unit of a modular species
Ramet	Physiologically autonomous module of clonal species
Somatic mutation	Mutation emerging via mitosis during development or growth of multicellular species
Unitary species	Multicellular organism with determinate body plan, composed of non-redundant organs with specific functions
Zygote	Eukaryotic cell formed by fusion of male and female gametes, which is the initial ancestor of all the cells within a multicellular organism

Supplementary Table 2: Sampling design in the Finish Archipelago Sea off Angsö island. Twenty-four ramets of the seagrass *Z. marina* were collected along a predetermined inshore-offshore transect consisting of two sections, one at ~2 m depth and the other one at ~4 m depth. Twelve ramets were collected in each water depth zone, respectively.

SHALLOW			DEEP		
Sample	**Depth (m)**	**Distance from starting point (m)**	**Sample**	**Depth (m)**	**Distance from starting point (m)**
R01	2.1	0	R13	4.1	0
R02	2.4	10	R14	3.9	10
R03	2.4	20	R15	3.8	20
R04	2.0	30	R16	4.1	30
R05	2.0	44	R17	4.2	40
R06	2.3	54	R18	4.4	55
R07	2.4	65	R19	4.4	68
R08	2.4	75	R20	4.5	78
R09	1.9	88	R21	4.5	89
R10	2.3	100	R22	4.6	99
R11	2.5	111	R23	4.8	109
R12	2.6	120	R24	5.1	119

Supplementary Table 3: Basic summary statistics on whole-genome resequencing of 24 *Z. marina* ramets on an Illumina Hiseq 4000 platform. All 24 samples were sequenced in a paired-end (2 × 151bp) mode to 81× coverage on average.

Sample	Read	Raw reads (bases)	Trimmed reads (bases)	Trimmed reads (coverage)	Mapping rate (%)
R01	r1	6,514,773,143	4,784,239,699	48.05	93.45
	r2	6,514,773,143	5,012,644,377		
R02	r1	9,126,143,285	8,239,834,760	76.87	94.90
	r2	9,126,143,285	7,435,005,872		
R03	r1	10,248,138,970	9,146,298,916	85.69	95.01
	r2	10,248,138,970	8,327,087,256		
R04	r1	8,712,513,364	7,797,060,447	72.56	92.80
	r2	8,712,513,364	6,999,344,576		
R05	r1	12,247,961,075	9,794,025,318	94.53	94.37
	r2	12,247,961,075	9,480,772,227		
R06	r1	10,066,487,933	9,104,901,065	84.89	95.48
	r2	10,066,487,933	8,205,080,108		
R07	r1	10,018,128,975	9,114,359,843	85.37	94.30
	r2	10,018,128,975	8,293,245,178		
R08	r1	10,666,580,808	9,713,578,994	90.71	93.65
	r2	10,666,580,808	8,782,113,457		
R09	r1	9,414,017,839	8,582,514,949	80.60	92.59
	r2	9,414,017,839	7,851,937,650		
R10	r1	6,528,798,325	5,940,882,792	55.49	94.98
	r2	6,528,798,325	5,373,273,275		
R11	r1	9,783,097,626	8,990,162,245	84.53	94.84
	r2	9,783,097,626	8,246,472,161		
R12	r1	11,572,405,799	10,554,499,766	98.95	95.24
	r2	11,572,405,799	9,622,152,935		
R13	r1	9,083,441,089	8,283,428,213	77.78	93.34

	r2	9,083,441,089	7,576,753,589		
R14	r1	10,169,361,968	9,277,920,776	87.14	94.09
	r2	10,169,361,968	8,489,958,466		
R15	r1	9,313,684,228	8,483,003,078	79.84	94.99
	r2	9,313,684,228	7,798,061,512		
R16	r1	10,977,353,002	10,050,369,523	94.66	92.11
	r2	10,977,353,002	9,252,291,537		
R17	r1	11,060,730,068	10,028,027,166	93.81	94.17
	r2	11,060,730,068	9,101,583,955		
R18	r1	7,425,409,296	6,727,127,750	63.06	93.98
	r2	7,425,409,296	6,132,131,423		
R19	r1	9,126,426,108	8,363,753,419	78.78	92.75
	r2	9,126,426,108	7,699,515,554		
R20	r1	10,223,403,509	8,959,008,457	85.95	94.09
	r2	10,223,403,509	8,566,982,853		
R21	r1	9,069,893,671	8,321,208,742	78.74	93.77
	r2	9,069,893,671	7,733,756,450		
R22	r1	9,349,276,438	8,484,251,652	79.70	93.44
	r2	9,349,276,438	7,767,904,874		
R23	r1	9,959,155,472	9,137,826,475	86.27	93.31
	r2	9,959,155,472	8,453,672,321		
R24	r1	10,101,766,969	9,250,283,727	87.08	94.74
	r2	10,101,766,969	8,506,391,862		

Supplementary Table 4: NCBI accession numbers of the sequence data sets.

Sample	BioSample accession number	BioProject accession number
R01	SAMN12389300	PRJNA557092
R02	SAMN12389301	PRJNA557092
R03	SAMN12389302	PRJNA557092
R04	SAMN12389303	PRJNA557092
R05	SAMN12389304	PRJNA557092
R06	SAMN12389305	PRJNA557092
R07	SAMN12389306	PRJNA557092
R08	SAMN12389307	PRJNA557092
R09	SAMN12389308	PRJNA557092
R10	SAMN12389309	PRJNA557092
R11	SAMN12389310	PRJNA557092
R12	SAMN12389311	PRJNA557092
R13	SAMN12389312	PRJNA557092
R14	SAMN12389313	PRJNA557092
R15	SAMN12389314	PRJNA557092
R16	SAMN12389315	PRJNA557092
R17	SAMN12389316	PRJNA557092
R18	SAMN12389317	PRJNA557092
R19	SAMN12389318	PRJNA557092
R20	SAMN12389319	PRJNA557092
R21	SAMN12389320	PRJNA557092
R22	SAMN12389321	PRJNA557092
R23	SAMN12389322	PRJNA557092
R24	SAMN12389323	PRJNA557092

Supplementary Table 5: Details on the verification approach of 14 different inter-ramet SNPs. Each SNP was verified across all 24 ramets. Example electropherograms of verification results can be found in Supplementary Figure 5.

Locus	Contig No.	Position	Enzyme	Restriction site	5' Fluorescent dye	Forward primer	Reverse primer
1st_05	LFYR01001079.1	70,478	BamHI	G^GATC_C	FAM	AGGTGGTTTGACTTCCGTCC	ACCATGCAAGAGCCCCTAAC
1st_07	LFYR01000619.1	331,575	HaeIII	GG^_CC	HEX	AGGGGTTTTTGTTCGGTCGA	TTTCAGTCGAGACAGGGTGC
1st_08	LFYR01000671.1	881,966	HaeIII	GG^_CC	FAM	TCCAGAACCAGCATTGACGA	AAAGATGCAGCCAGGGGAAG
1st_10	LFYR01001213.1	1,012,778	HaeIII	GG^_CC	FAM	TGATCTTTGGTCGTGGAGCA	TCGGGAGGTTGGTCTCTCAT
1st_11	LFYR01001351.1	610,127	HaeIII	GG^_CC	FAM	AAGTCCATCAAGAGCACGCA	CGTCGCCGGATCCTATGAAA
1st_15	LFYR01001054.1	847,248	NcoI	C^CATG_G	HEX	GTGTGGGTTCGCCAGAGTTA	GCGTTTGTGAAATGGCGACT
2nd_03	LFYR01000655.1	208,328	HaeIII	GG^_CC	FAM	TGGTTCGCAGCATGAATCCT	GCAGTCAAGAGGTGTGGTGT
2nd_04	LFYR01000709.1	341,678	HaeIII	GG^_CC	FAM	TTGGATTGAGCCCGATGTCC	CGCTTTATGCTGGACCCTGA
2nd_05	LFYR01000781.1	174,973	HaeIII	GG^_CC	FAM	CCCCTTTGTTGGATCTTCTTCC	TCGGCTCGGCTTGATAAACT
2nd_06	LFYR01000915.1	494,331	HaeIII	GG^_CC	FAM	TGACTTGGAACCTCCCCAGA	ACGGCGCATTTCTGTCAAAA
2nd_08	LFYR01001661.1	436,971	HaeIII	GG^_CC	FAM	ACACCAAGAGCAATCAACCT	CACATTCCAGGCGCAAAACA
2nd_09	LFYR01001977.1	179,811	HaeIII	GG^_CC	FAM	TTTCAGCAAAGGCAGCCAAG	GCCAGCCCTTCCTCTTTGAT
2nd_10	LFYR01002125.1	147,686	HaeIII	GG^_CC	FAM	CGTAGGTTTGCCTCGTCACT	CGCTCCATAGTCTGCCGATT
2nd_TC	LFYR01002184.1	54,937	HaeIII	GG^_CC	FAM	ACAACGCTAGGAGACATGTTCT	CGATTCTACATTACCGGCCCA

Supplementary Table 6: Inter-ramet SNPs exclusively shared by two ramets. In the 81x re-sequencing dataset, we identified 7,054 inter-ramet SNPs with different fixed genotypes among the 24 sampled ramets (cf. Extended Data Fig. 2). The first column gives the number of loci exclusively shared by the two ramets, i.e., at these loci, the two ramets have the same genotypes, which are different from those in other ramets.

Number of loci	Ramet pairs
335	R15&R16
182	R5&R13
155	R7&R8
145	R2&R24
8	R6&R19
4	R6&R9
3	R18&R20
3	R3&R22
3	R4&R18
3	R7&R20
2	R10&R13, R10&R19, R13&R16, R15&R18, R1&R18, R2&R4, R3&R13, R6&R17, R6&R21, R7&R16, R8&R13, R8&R17, R8&R22, R9&R14
1	R10&R14, R10&R20, R10&R21, R10&R24, R11&R12, R11&R15, R11&R16, R12&R16, R13&R14, R13&R15, R16&R18, R17&R18, R17&R21, R19&R20, R1&R11, R1&R17, R1&R24, R1&R3, R1&R6, R21&R22, R22&R24, R2&R11, R2&R18, R2&R22, R2&R3, R3&R10, R3&R12, R3&R8, R4&R11, R4&R16, R5&R15, R5&R7, R6&R20, R6&R22, R6&R8, R7&R12, R7&R13, R7&R23, R9&R13, R9&R17, R9&R23

Supplementary Table 7: Pairwise genetic distance matrix of fixed differences IN SNPs among *Z. marina* ramets. The genetic distance is quantified as number of different alleles. The upper half was calculated based on 7,054 inter-ramet SNPs, and the lower half was calculated based on the 432 nonsynonymous ones.

	R01	R02	R03	R04	R05	R06	R07	R08	R09	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24
R01	-	988	452	583	665	594	1,134	1,118	1,071	757	1,092	1,291	642	850	996	1,049	840	481	807	795	784	979	783	740
R02	55	-	1,446	1,588	1,664	1,610	1,185	1,171	1,117	761	1,149	1,362	1,649	1,432	1,429	1,487	1,257	1,472	1,208	1,211	1,200	1,343	1,188	466
R03	31	88	-	413	490	649	1,588	1,583	1,535	1,162	1,575	1,799	471	1,288	1,442	1,523	1,289	507	1,240	1,224	1,220	1,436	1,220	1,189
R04	32	88	24	-	611	804	1,756	1,731	1,678	1,315	1,724	1,946	605	1,456	1,604	1,668	1,439	662	1,377	1,376	1,377	1,593	1,374	1,341
R05	44	100	36	36	-	873	1,838	1,820	1,764	1,408	1,807	2,020	305	1,522	1,670	1,740	1,512	727	1,465	1,459	1,452	1,669	1,445	1,420
R06	34	90	44	44	56	-	1,776	1,763	1,695	1,333	1,747	1,956	861	1,462	1,611	1,683	1,440	697	1,390	1,399	1,387	1,605	1,383	1,356
R07	83	81	118	123	135	125	-	72	345	464	391	609	1,816	1,572	1,578	1,657	1,427	1,645	1,376	1,353	1,346	1,493	1,349	926
R08	83	81	118	123	135	125	0	-	321	447	374	596	1,792	1,559	1,571	1,640	1,403	1,627	1,353	1,350	1,326	1,472	1,326	914
R09	78	70	112	112	124	114	27	27	-	412	114	330	1,734	1,521	1,505	1,582	1,351	1,573	1,292	1,295	1,287	1,441	1,283	855
R10	48	45	82	87	99	89	36	36	43	-	450	663	1,380	1,143	1,149	1,217	996	1,212	938	937	913	1,070	925	508
R11	82	75	117	117	129	119	32	32	5	48	-	363	1,773	1,562	1,550	1,624	1,387	1,612	1,344	1,338	1,326	1,486	1,325	892
R12	99	91	133	133	145	135	48	48	21	64	26	-	1,995	1,781	1,772	1,833	1,604	1,822	1,558	1,554	1,537	1,699	1,543	1,103
R13	39	96	32	32	22	52	131	131	120	95	125	141	-	1,506	1,654	1,721	1,496	718	1,446	1,443	1,429	1,644	1,423	1,399
R14	50	84	83	88	100	90	101	101	108	65	113	129	96	-	1,436	1,505	1,268	1,339	1,224	1,218	1,189	1,410	1,200	1,171
R15	61	81	93	93	105	95	116	116	105	80	110	126	101	89	-	291	735	1,479	1,187	1,177	1,176	1,424	1,176	1,167
R16	60	81	93	93	105	95	116	116	105	80	110	126	99	89	12	-	793	1,543	1,258	1,255	1,242	1,484	1,239	1,234
R17	57	76	88	88	100	90	111	111	100	75	105	121	96	84	43	43	-	1,326	1,023	1,024	1,008	1,258	1,003	996
R18	33	88	39	39	51	44	123	123	112	87	117	133	47	88	93	93	88	-	1,271	1,255	1,249	1,469	1,251	1,219
R19	48	68	80	80	92	82	103	103	92	67	97	113	88	76	71	71	66	80	-	75	81	1,212	81	947
R20	48	68	80	80	92	82	103	103	92	67	97	113	88	76	71	71	66	80	0	-	72	1,212	77	937
R21	49	74	81	86	98	88	94	94	98	58	103	119	94	67	77	77	72	86	6	6	-	1,176	59	932
R22	61	86	95	102	114	104	103	103	110	67	115	131	110	80	95	95	90	102	82	82	73	-	1,172	1,098
R23	48	71	81	83	95	85	100	100	95	64	100	116	91	73	74	74	69	83	3	3	3	79	-	939
R24	45	19	77	77	89	79	70	70	59	34	64	80	85	73	70	70	65	77	57	57	63	75	60	-

Supplementary Table 8: Basic information on ultra-deep whole-genome resequencing of 3 seagrass *Z. marina* ramets. Ramets R08, 10 and 12 were re- sequenced on Illumina Novaseq 6000 platform at 1370× coverage on average (paired-end 2 × 151bp).

Sample	Read	Raw_reads (bases)	Trimmed_reads (bases)	Trimmed_reads (coverage)	Mapping rate (%)
R08	r1	133,813,282,831	126,375,440,334	1,238	93.63
	r2	133,813,282,831	126,134,322,303		
R10	r1	139,403,148,056	131,822,147,904	1,295	95.01
	r2	139,403,148,056	132,180,830,941		
R12	r1	170,779,135,150	160,947,684,601	1,578	95.19
	r2	170,779,135,150	160,737,274,523		

Supplementary Table 9: Verification results for 9 intra-ramet SNPs. Intra-ramet SNPs were verified in one to three ramets, depending on their presence based on the SNP calling results. Note that both non-verified SNPs had low variant read frequencies (f<0.06). The corresponding electropherograms can be found in Supplementary Figure 11.

SNP	Coverage	Variant read frequency	Sample	Verification
Mu_02	2,348	0.1	R08	√
	2,386	0.1	R10	√
	3,097	0.1	R12	√
Mu_03	834	0.08	R08	√
	949	0.08	R10	√
	1,162	0.06	R12	√
Mu_04	919	0.19	R10	√
Mu_05	726	0.43	R08	√
	730	0.42	R10	√
	916	0.41	R12	√
Mu_07	556	0.06	R12	√
Mu_09	1,055	0.05	R12	×
Mu_10	966	0.06	R12	×
Mu_11	34	0.29	R12	√
Mu_12	1,604	0.21	R12	√

Supplementary Table 10: Detailed information on the verification approach of 9 intra-ramet SNPs. Verification was possible in one to three ramets, depending on the SNP calling results. For verification results see Supplementary Figure 11 and Supplementary Table 9.

Locus	Contig No.	Position	Enzyme	Restriction site	5' Fluorescent dye	Forward primer	Reverse primer
Mu_02	LFYR01000036.1	459,138	HaeIII	GG^_CC	FAM	CAGGGTAACACACAACCTGGA	TGCTTTTTCTTCAACCTGTTCCT
Mu_03	LFYR01000337.1	71,847	HaeIII	GG^_CC	FAM	TGGAACACCGAATCTGCACA	GCGGAACACATAAAGGGGGA
Mu_04	LFYR01000584.1	263,636	HaeIII	GG^_CC	FAM	AATGCAAAGCCCCAACATCG	ATTTGAACCCATTGCGCTGG
Mu_05	LFYR01000601.1	320,610	HaeIII	GG^_CC	FAM	CGGTGCGAAAATTCAGGTGG	AGGACTTTGTGTGACAGACTTCT
Mu_07	LFYR01000850.1	181,739	HaeIII	GG^_CC	FAM	CCAGGGAGAGACAAGAACGAC	AGTGTTGCTGTTGCTGCATG
Mu_09	LFYR01000907.1	145,314	HaeIII	GG^_CC	FAM	CTTGCCTTCCCCATGACCAT	TTCTGTCACCAAATCGCCGA
Mu_10	LFYR01000907.1	188,608	HaeIII	GG^_CC	FAM	CTTGCCTTCCCCATGACCAT	GCCAGCTCGCGTCATTAATG
Mu_11	LFYR01001228.1	68,748	HaeIII	GG^_CC	FAM	CGATTCACACAGGCAATCGC	CGGCGACTGGACTTAGGTTT
Mu_12	LFYR01001279.1	783,253	HaeIII	GG^_CC	FAM	AGTGTGATGTTGGGTGAGACA	ACAGACTTGGACAGCATGGG

Supplementary Table 11: Number of heterozygous SNPs detected by GATK pipeline (81x data). Numbers in brackets represent those categorized as fixed ones according to the 1,370x resequencing data.

(81x data)	
Sample	Number of heterozygous SNPs
R08	996(746)
R10	602(403)
R12	1231(966)

Supplementary Table 12: Mosaic vs. fixed SNPs detected by the 1370x resequencing data. The number in the brackets represent the number of SNPs identified as heterozygous genotypes in GATK pipeline (81x data). Note that for the assessment of frequency change of SNPs among ramets, our coverage threshold was 500x.

1370x data	coverage ≥500x			coverage <500x
Sample	VRF<0.4	Fixed	VRF>0.6	
R08	2763(0)	892(746)	80(1)	1246(26)
R10	3759(0)	500(403)	75(2)	1133(4)
R12	2325(0)	1174(966)	81(2)	891(4)

Supplementary Table 13: Significantly enriched GO terms for intra- and inter-ramet SNPs relative to the entire annotated gene universe of *Z. marina*. Tested gene samples for the intra-ramet level were the combination of all mosaic (intra-ramet allele frequency <0.5) non-synonymous SNPs detected based on three 1370x-sequenced ramets; for the inter-ramet level it was the combination of inter-ramet SNPs detected based on twenty-four 81x-sequenced samples.

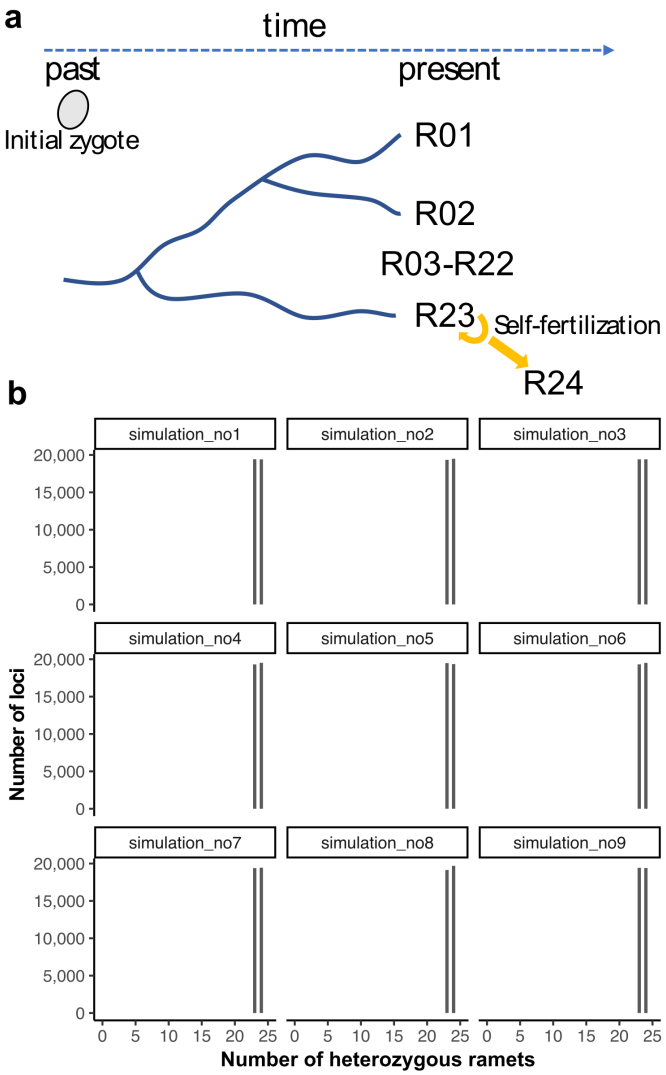
Intra-ramet SNPs					
BP (biological process)					
GO.ID	Term	Annotated	Significant	Expected	Unjusted P
GO:0006468	protein phosphorylation	785	14	5.68	0.0013
GO:0006075	1,3-beta-D-glucan biosynthetic process	17	2	0.12	0.0065
GO:0006422	aspartyl-tRNA aminoacylation	1	1	0.01	0.0072
CC (cellular component)					
GO.ID	Term	Annotated	Significant	Expected	Unjusted P
GO:0098797	plasma membrane protein complex	35	3	0.19	0.00083
GO:0005874	microtubule	21	2	0.11	0.00563
MF (molecular function)					
GO.ID	Term	Annotated	Significant	Expected	Unjusted P
GO:0016165	linoleate 13S-lipoxygenase activity	5	2	0.04	0.00056
GO:0004674	protein serine/threonine kinase activity	582	11	4.42	0.00463
GO:0003843	1,3-beta-D-glucan synthase activity	17	2	0.13	0.00721
GO:0005200	structural constituent of cytoskeleton	17	2	0.13	0.00721
GO:0004815	aspartate-tRNA ligase activity	1	1	0.01	0.00759
GO:0035639	purine ribonucleoside triphosphate binding	2262	27	17.17	0.00933
Inter-ramet SNPs					
BP (biological process)					
GO.ID	Term	Annotated	Significant	Expected	Unjusted P
GO:0006464	cellular protein modification process	1011	35	16.81	0.00012
CC (cellular component)					
GO.ID	Term	Annotated	Significant	Expected	Unjusted P
GO:0005634	nucleus	542	18	8.11	0.0015
MF (molecular function)					
GO.ID	Term	Annotated	Significant	Expected	Unjusted P
GO:0005515	protein binding	3074	93	58.9	4.10E-06
GO:0003830	beta-1,4-mannosylglycoprotein 4-beta-N-acetylglucosaminyltransferase activity	4	2	0.08	0.0021
GO:0046914	transition metal ion binding	1161	36	22.25	0.0027
GO:0015450	P-P-bond-hydrolysis-driven protein transmembrane transporter activity	7	2	0.13	0.0072
GO:0008236	serine-type peptidase activity	148	8	2.84	0.0077

Supplementary Table 14: Abundance of somatic genetic polymorphisms (as SNPs) in plant species.

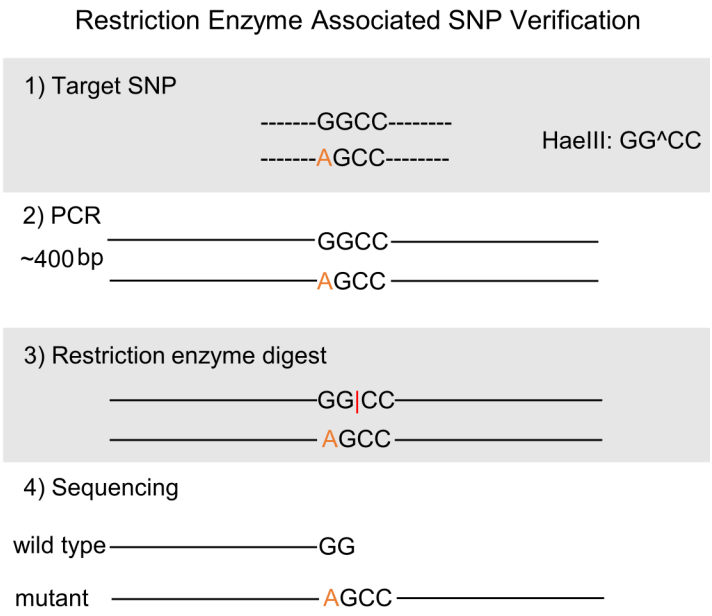
Species (genome size)	Sample source	Estimated age (years)	tissue	Sequenced samples	Average SNP per sample	Pairwise-sample differentiation	Data source
Prunus mira (225 Mb)	one tree	600	Leaf	32	12.7	-	Wang et al. 2019
	one tree	550	Leaf	12	23.9	-	
	one tree	420	Leaf	23	17.7	-	
	one tree	300	Leaf	9	12.8	-	
Prunus persica (225Mb)	one tree	21	Leaf	23	3.74	-	
			Root	13	29.8	-	
	one tree	25	Leaf	16	6.19	-	
			Petal	13	11.31	-	
	one tree	30	Leaf	26	6.46	-	
	one tree	50	Leaf	8	6.25	-	
	one tree	40	Leaf	16	3.56	-	
	one tree	2	Leaf	75	1.97	-	
Prunus mune (220 Mb)	one tree	20	Leaf	25	12.9	-	
			Root	32	25.4	-	
	one tree	8	Leaf	33	5.7	-	
Salix suchowensis (480 Mb)	one tree	1	Leaf	19	1.26	-	
			Root	21	2.86	-	
Brachypodium distachyon (272 Mb)	one plant	1	Leaf	29	3.17	-	
			Root	8	4.75	-	
			Lemma	7	2.57	-	
Fragaria vesca (210 Mb)	one plant	1	Leaf	45	1.93	-	
			Stem	4	4.75	-	
Arabidopsis thaliana (119 Mb)	two plants	1	Leaf	64	0.69	-	
Oryza sativa (373 Mb)	four plants	1	Leaf (Tiller)	29	4.79	-	
			Leaf (Callus)	13	194.8	-	
Quercus robur (720 Mb)	one tree	234	Leaf	2	-	38-47	Schmid-Siegert et al. 2017
Zostera marina (204 Mb)	one genet	750 – 1,500	Meristematic region and the basal portions of the leaves	24	-	1,216 (on average)	Present study

Schmid-Siegert, Emanuel, et al. "Low number of fixed somatic mutations in a long-lived oak tree." Nature plants 3.12 (2017): 926.
Wang, Long, et al. "The architecture of intra-organism mutation rate variation in plants." PLoS biology 17.4 (2019): e3000191.

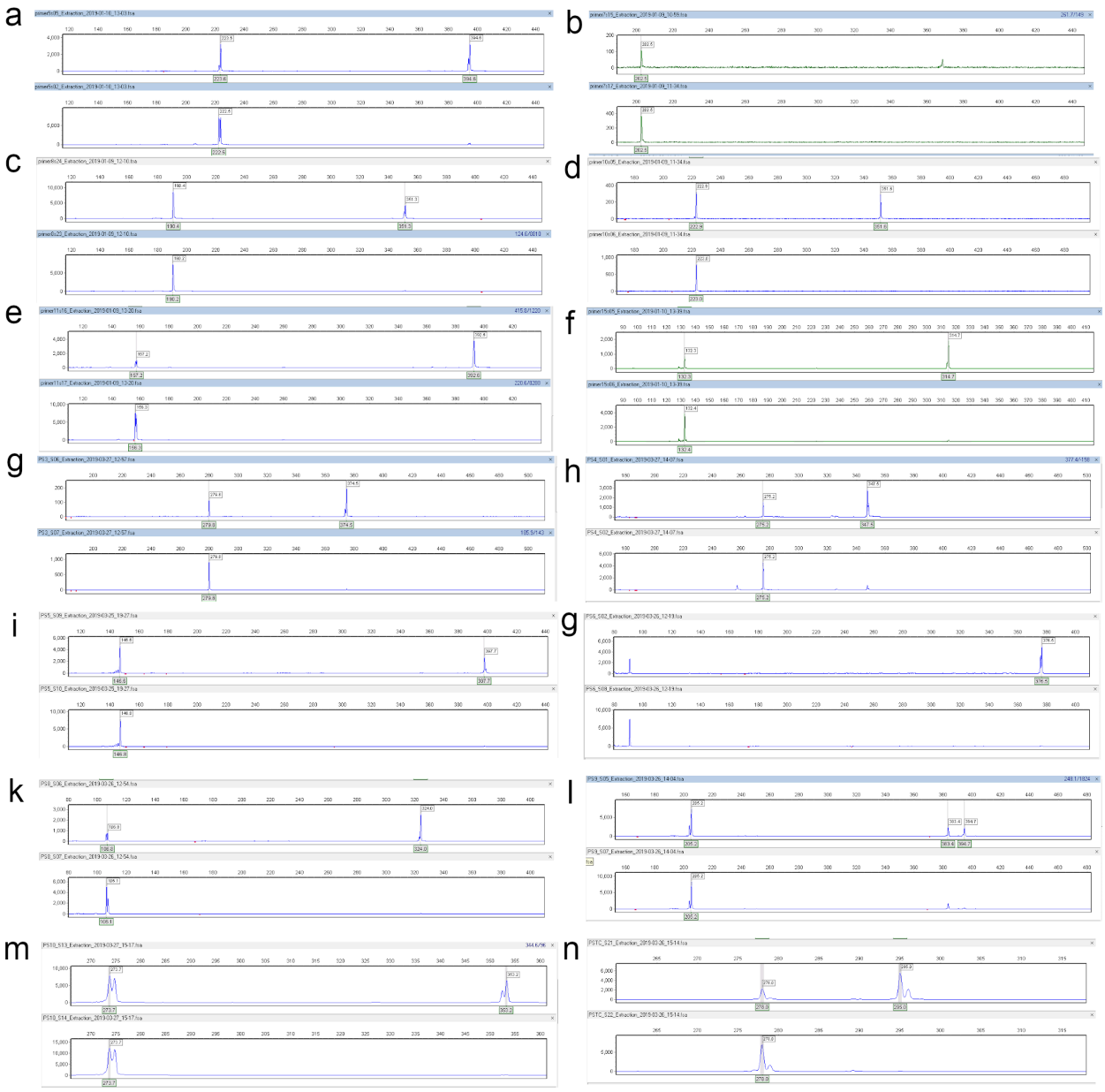
Supplementary Figures



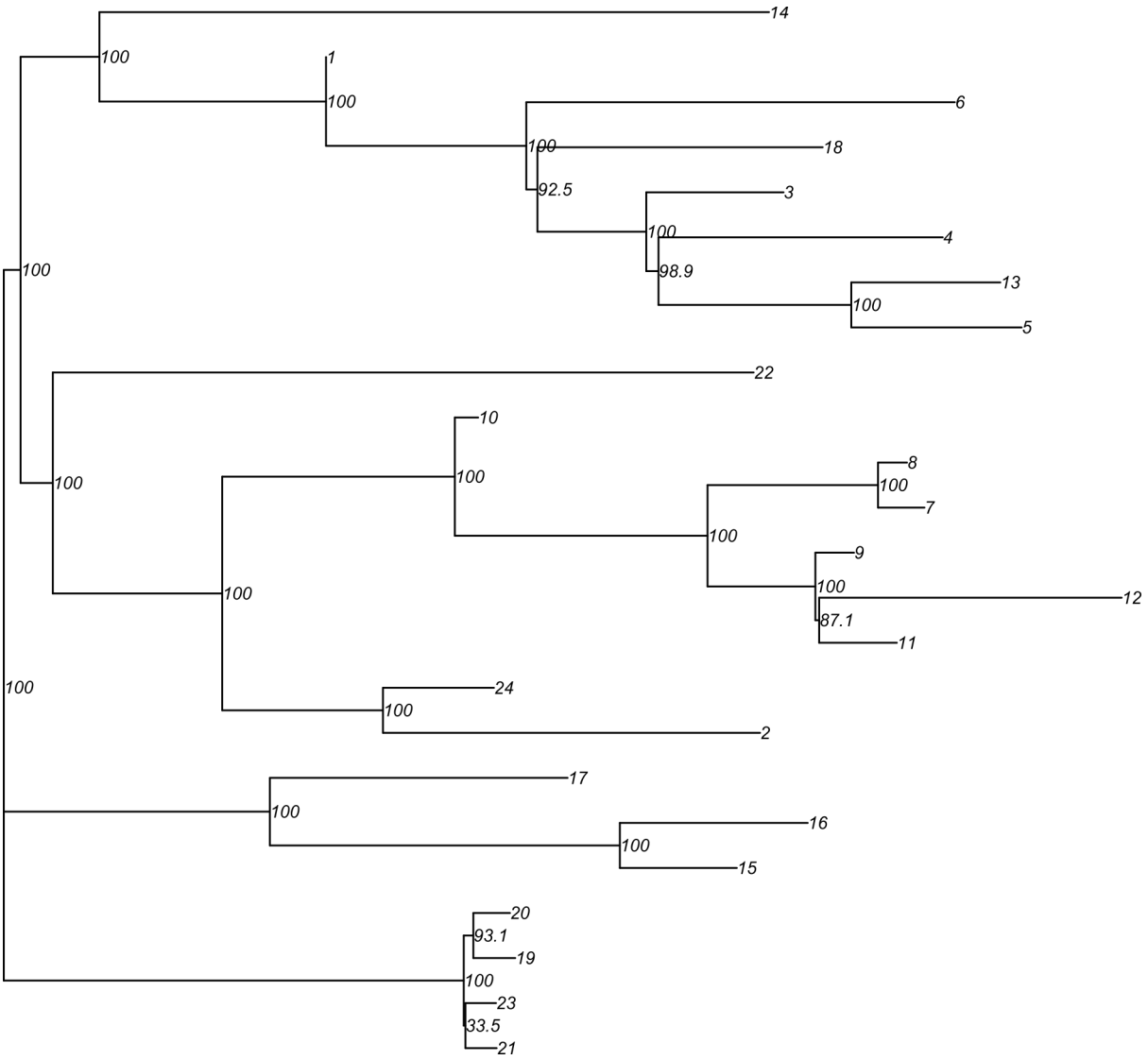
Supplementary Figure 1: Simulation of a single event of sexual reproduction among the 24 sampled ramets. a, depiction of the scenario that simulated 23 ramets to be asexual descendants of the founder genotype, while R24 is formed by self-fertilization of any of the other ramets. We assumed 31,777 heterozygous loci in the initial zygote, which is the number of SNPs used for plotting Fig. 2b, and no additional somatic mutations. Under asexual reproduction, all heterozygous loci in the initial zygote will be inherited in complete linkage, i.e. are all shared among 23 ramets (R1…R23). In contrast, for each of the 31,777 loci, R24 will have one of the three possible genotypes (i.e., f=2pq=0.5 for heterozygous genotype). **b**, we ran 9 independent simulations, and plotted the resulting histograms in the same way as Fig. 2b. Simulations always resulted in two peaks (x=23 and x=24) of almost similar genotype abundance. Thus, even one sexual reproduction will result in a pattern very different to the empirical pattern found in the field (cf. Fig. 2b). The single dominant peak in Fig. 2b can only be explained by complete asexual reproduction.



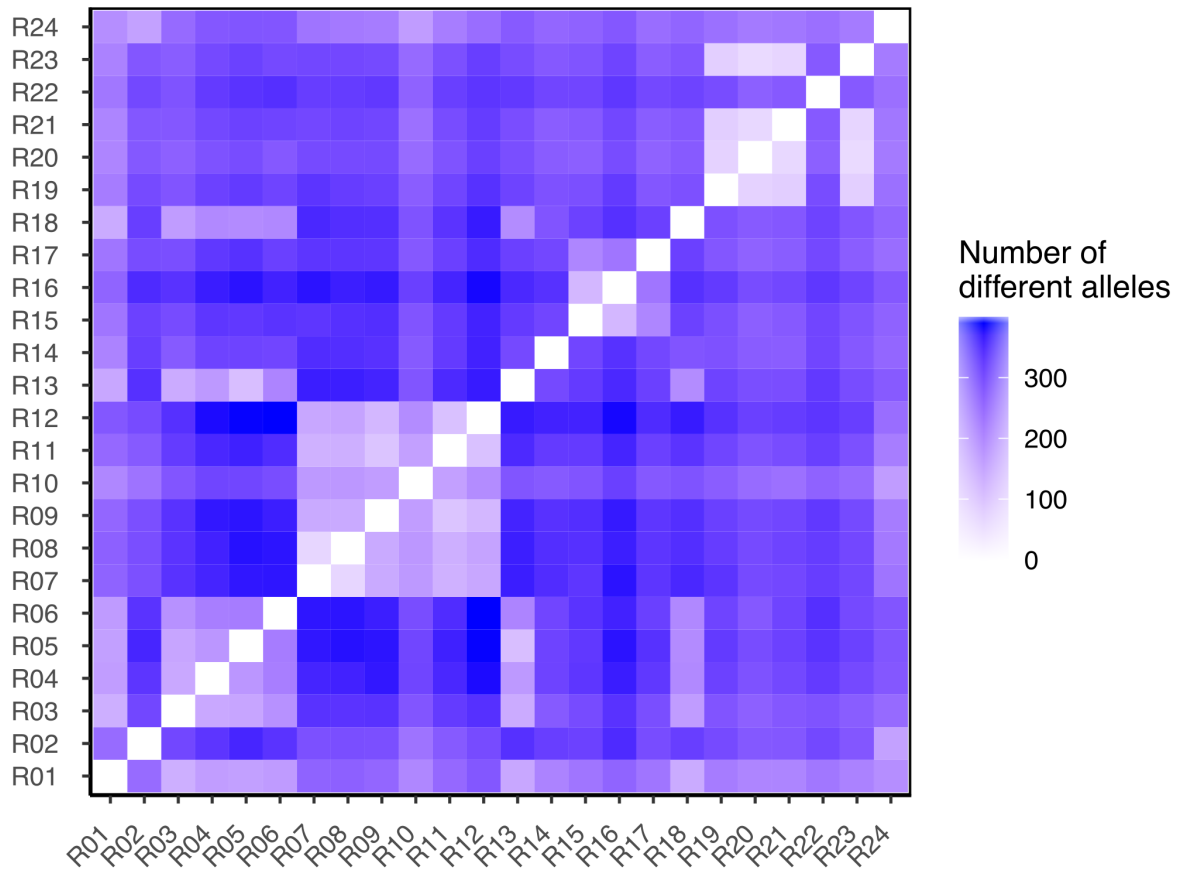
Supplementary Figure 2: Restriction-enzyme based SNP verification method. In order to verify SNP loci independent of any sequencing technology, we employed a restriction enzyme based approach. To this end, restriction enzyme motifs covering the target SNPs were identified on the reference genome. The genotype call to be verified only displayed an intact digestible motif when homozygous while the variant allele changes the recognition sequence. Fluorescence-labelled primer pairs were designed to amplify the ~400 bp amplicon encompassing the target SNP, which was subsequently digested with an appropriate restriction enzyme. Upon electrophoretic separation on an ABI 3130xL genetic analyzer, heterozygous samples display fluorescent peaks at two different positions, while the homozygous genotype had only one signal (i.e. both homologous sites are cut at the same position).



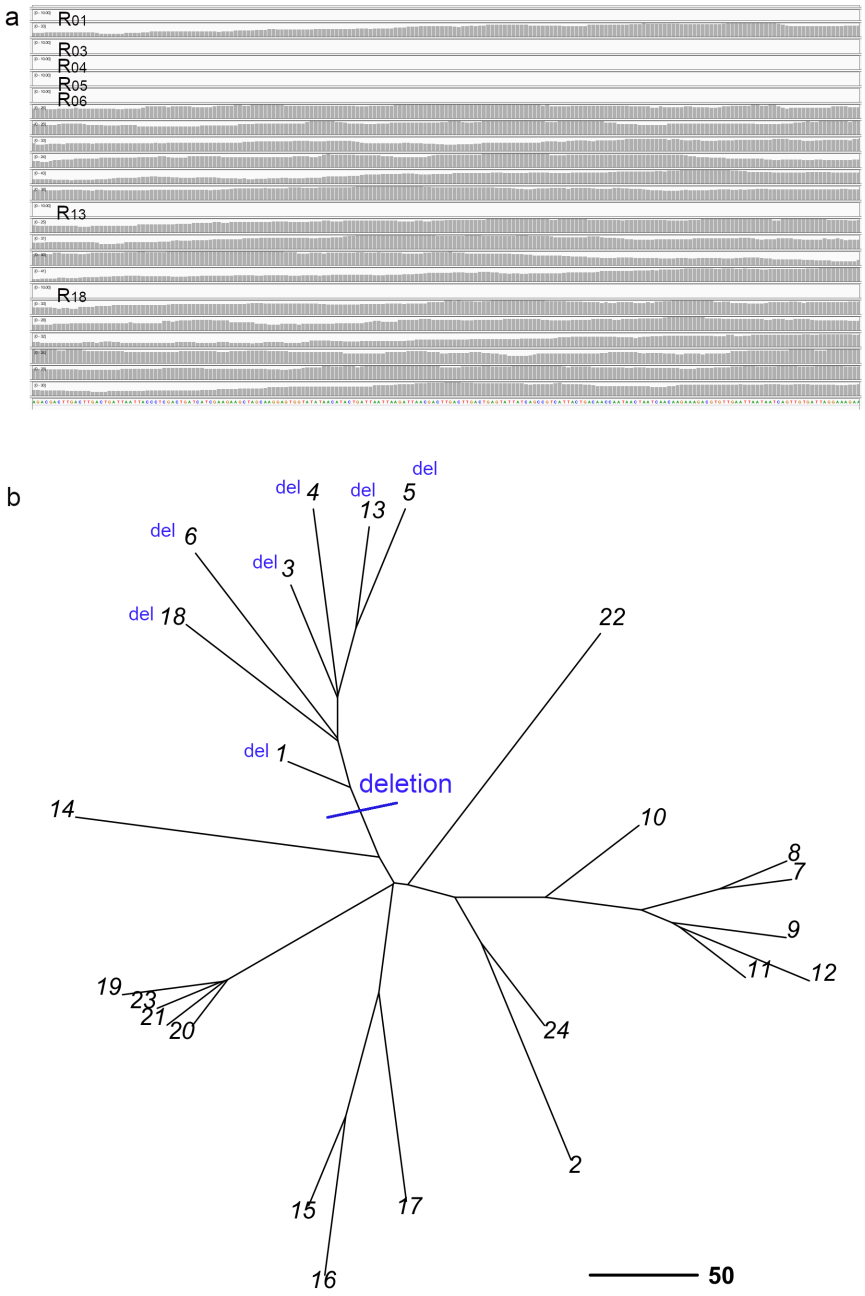
Supplementary Figure 3: Experimental verification of inter-ramet SNPs in a population of *Z. marina* ramets. Analyses were performed on an ABI3100xL genetic analyzer using one fluorescently labelled primer and the fragment analysis mode. In all panels (a-n), the upper graph with fluorescent peaks at two different positions represents the heterozygous genotype. The panels (a-n) indicate the following loci: 1st_05, 1st_07, 1st_08, 1st_10, 1st_11, 1st_15, 2nd_03, 2nd_04, 2nd_05, 2nd_06, 2nd_08, 2nd_09, 2nd_10, and 2nd_TC. The detailed information for each locus can be found in Supplementary Table 6. In some cases, such as panel (m), the fluorescent signal strength was too high (>10,000) which led to a spurious twin peak, which however does not interfere with a proper interpretation. All 14 loci in combination with 24 ramets were confirmed.



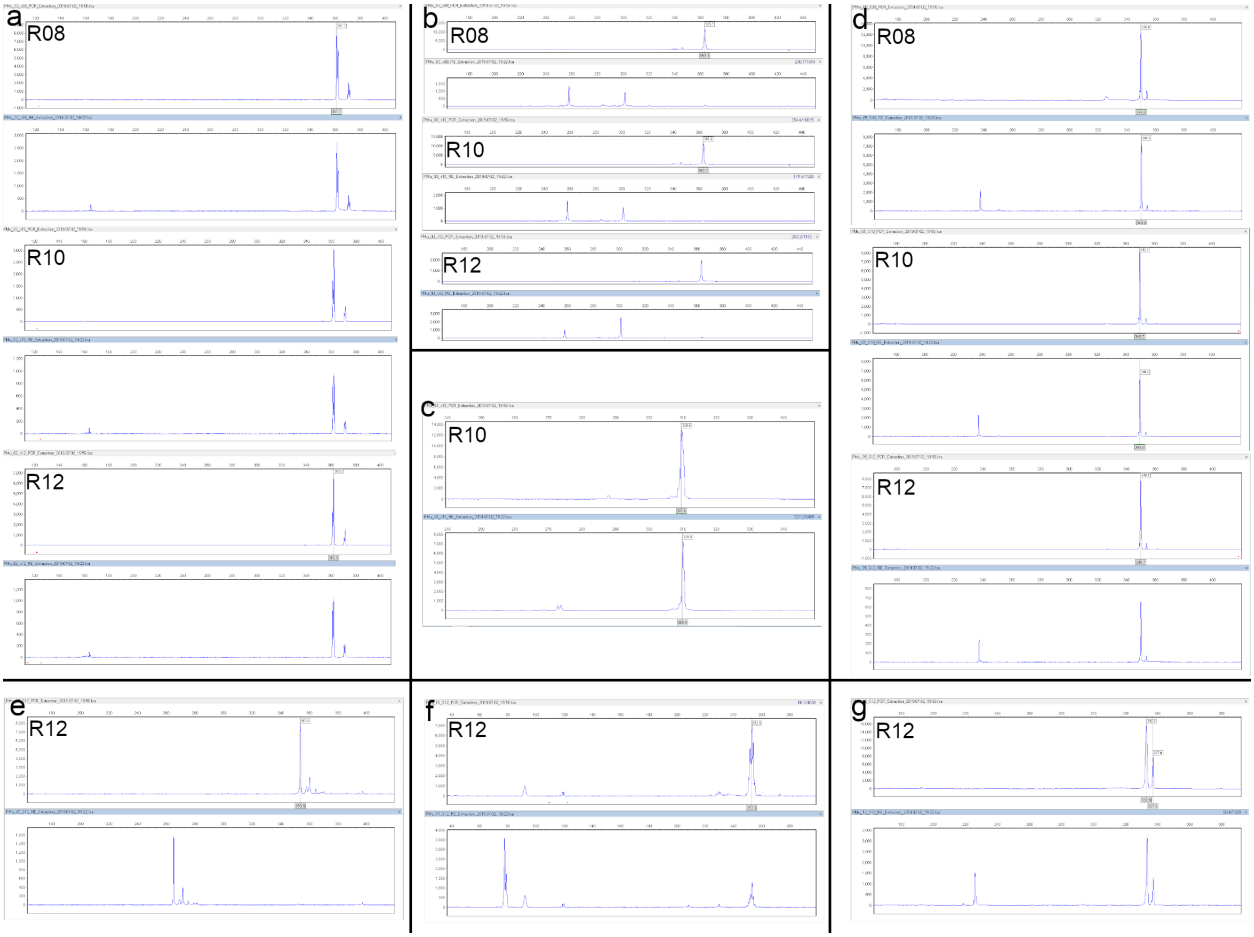
Supplementary Figure 4: Bootstrap values accompanying the NJ tree depicted in Figure 2d.



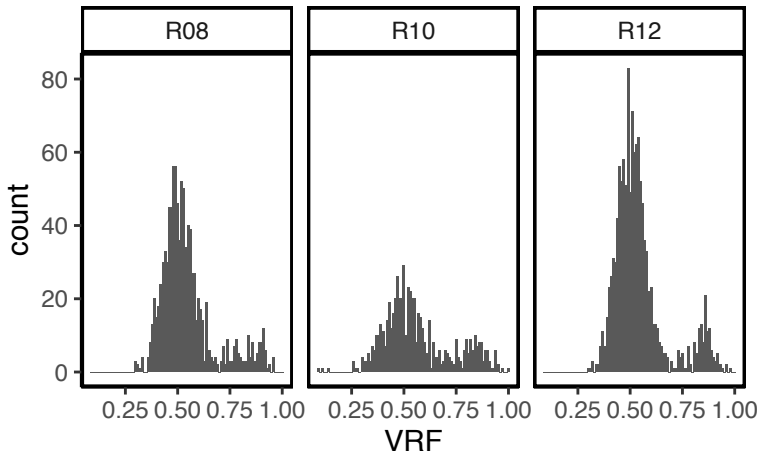
Supplementary Figure 5: Pairwise genetic differences among 24 seagrass *Z. marina* ramets based on small insertion-deletion polymorphisms (indels). 1,654 inter-ramet indels among the 24 ramets were used to calculate pairwise genetic differences based on the number of different alleles. The result is consistent to the heat map based on inter-ramet fixed SNPs (Fig. 2c).



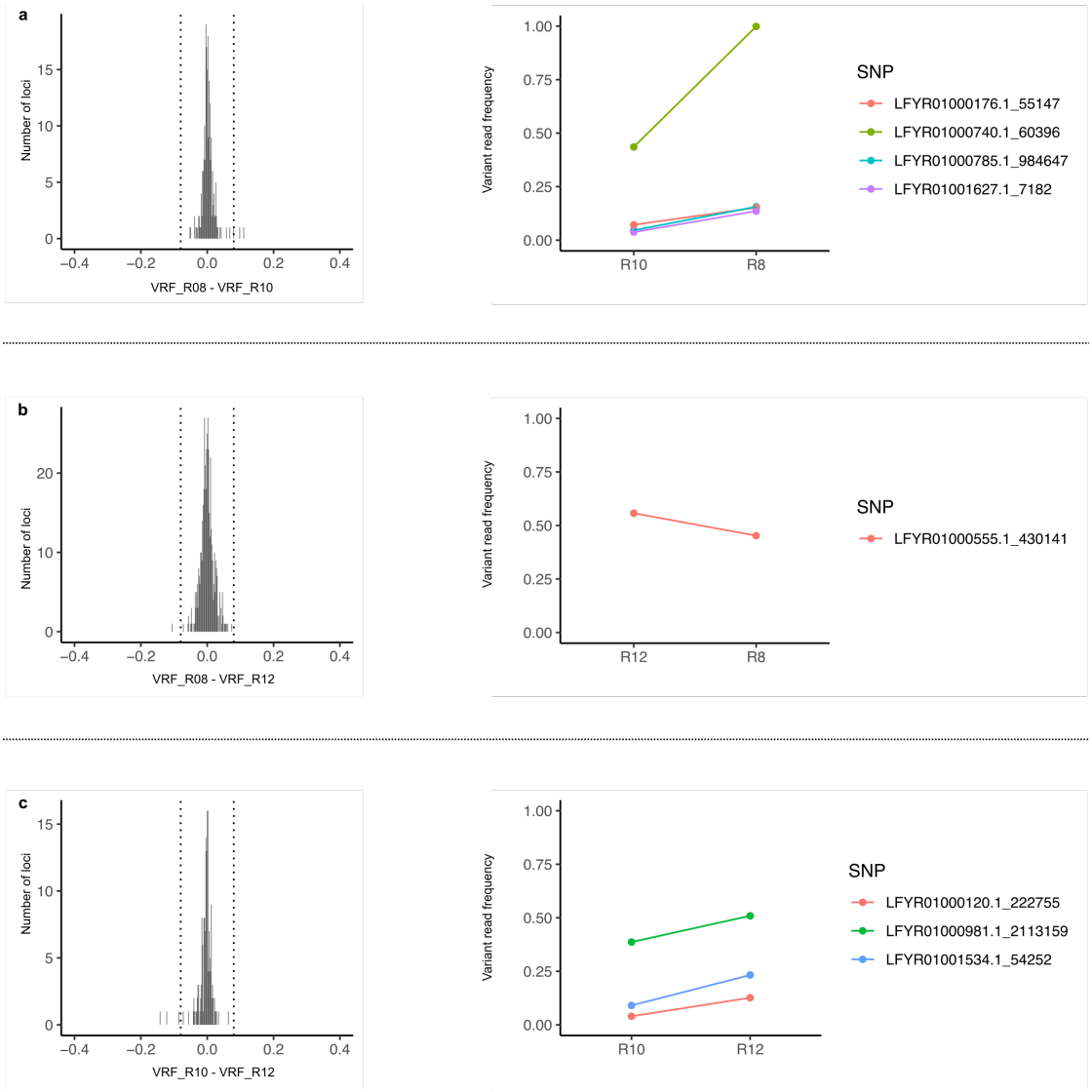
Supplementary Figure 6: A large structural polymorphism among *Z. marina* ramets. We found a >200 bp structural polymorphism among the 24 ramets. Its presence/absence among the 24 ramets is consistent with the topology depicted by the NJ tree.



Supplementary Figure 7: Verification of intra-ramet mosaic SNPs in *Z. marina*. We verified intra-ramet SNPs at selected loci where the variant allele converted a non-restriction into a restriction site. Thus, after restriction enzyme digestion one shorter PCR fragments will appear only if the variant allele is present, an observation that cannot be biased by undigested wild-type alleles. Analyses were performed on an ABI3100xL genetic analyzer using one fluorescently labelled primer and the fragment analysis mode. For each combination of locus and sample (e.g. a + R08), the upper graph and the lower graph represent size-separated PCR products before and after restriction enzyme digest, respectively. The variant allele is represented as a shorter fragment in the lower graph. Panels (a-g) represent the following loci: Mu_02, Mu_03, Mu_04, Mu_05, Mu_07, Mu_11, and Mu_12. The detailed information for each locus can be found in Supplementary Table 10. In panel (b), the amplicon contains another restriction site ("GGCC") at the downstream of the target SNP, so all the amplicons were universally cleaved into a shorter one, based on which the restriction enzyme cut another time on the target site.



Supplementary Figure 8: Histograms of variant read frequency based on 81x data. For the heterozygous genotypes in R08, R10, and R12 detected by 81x data. separate histograms of variant read frequency (VRF) are given.



Supplementary Figure 9: Mosaic somatic polymorphisms shared by two ramets show largely stable allele frequencies. At the intra-ramet section (1370x dataset, ramets R08, R10, and R12), the somatic polymorphisms were categorized into 1) being present in only one ramet, 2) being shared by two ramets, and 3) being shared by all three ramets (cf. Venn diagram Fig 3a). Plots are based on SNPs shared by two ramets, in panel **a** among R08 and R10, in **b** among R08 and R12, in **c** among R10 and R12. The minimum coverage for each locus was set to 500x. At a given locus, assume the intra-ramet allele frequency is f (i.e., the proportion of a specific allele within a ramet), the variant read frequency (VRF) of the locus would follow a binominal distribution ($\sigma=\sqrt{f*(1-f)/n}$), given sequencing coverage n. σ obtains maximum value σ_{max} when f=0.5 and n=500 (σ_{max}=0.02). Assume the two ramets have equal f, their VRFs would differ by $4*\sigma_{max}$=0.08 at most (95% confidence, x=0.08 and x=-0.08 are indicated by dotted lines). The loci where the VRFs of the two ramets differ >0.08 are shown in the right panels.

Supplementary Data

Supplementary Dataset 1: Detailed information for 432 inter-ramet non-synonymous SNPs

Contig	Position	Protein	Effects
LFYR01000729.1	70552	Pyridoxal phosphate phosphatase PHOSPHO2	missense_variant
LFYR01000729.1	554121	putative Receptor-kinase	missense_variant
LFYR01000729.1	556128	putative Receptor-kinase	missense_variant
LFYR01000729.1	1145646	UDP-glucose:glycoprotein glucosyltransferase%2C family GT24	missense_variant
LFYR01000729.1	1155852	UDP-glucose:glycoprotein glucosyltransferase%2C family GT24	missense_variant
LFYR01000729.1	1609873	DNA-directed RNA polymerase	missense_variant
LFYR01000729.1	1832693	putative Lung seven transmembrane receptor	missense_variant
LFYR01000729.1	1959437	RING finger protein 13	missense_variant
LFYR01000729.1	2221136	hypothetical protein	missense_variant
LFYR01000113.1	351246	Chaperone protein dnaJ 10	missense_variant
LFYR01000113.1	544881	hypothetical protein	missense_variant
LFYR01000113.1	867859	Protein kinase	missense_variant
LFYR01000113.1	908479	putative Squamosa promoter-binding protein	missense_variant
LFYR01000113.1	964422	AT hook motif DNA-binding family protein	missense_variant
LFYR01000113.1	1166763	Retinoblastoma-related protein	missense_variant
LFYR01000036.1	124308	hypothetical protein	missense_variant
LFYR01000047.1	272444	PHD finger family protein	missense_variant
LFYR01000047.1	532465	hypothetical protein	stop_gained
LFYR01000056.1	263523	hypothetical protein	missense_variant
LFYR01000056.1	505006	Receptor kinase	missense_variant
LFYR01000056.1	514822	hypothetical protein	missense_variant
LFYR01000079.1	185712	hypothetical protein	missense_variant
LFYR01000079.1	213877	hypothetical protein	missense_variant
LFYR01000079.1	443742	Translocase of chloroplast 34	missense_variant
LFYR01000073.1	4741	hypothetical protein	missense_variant
LFYR01000090.1	224116	Nodulin-like / Major Facilitator Superfamily protein	missense_variant
LFYR01000112.1	19115	Protein POLLEN DEFECTIVE IN GUIDANCE 1	missense_variant
LFYR01000112.1	374594	hypothetical protein	missense_variant

LFYR01000216.1	316313	Zinc finger protein CONSTANS-like protein	missense_variant
LFYR01000216.1	1126846	putative Protein arginine n-methyltransferase	missense_variant
LFYR01000120.1	531533	hypothetical protein	missense_variant
LFYR01000129.1	149715	Octicosapeptide/Phox/Bem1p domain-containing protein	missense_variant
LFYR01000176.1	135728	hypothetical protein	missense_variant
LFYR01000182.1	253395	Protein YABBY 6	missense_variant
LFYR01000182.1	419720	hypothetical protein	missense_variant
LFYR01000182.1	425215	Cationic amino acid transporter 2%2C vacuolar	missense_variant
LFYR01000338.1	228276	Peroxidase	missense_variant
LFYR01000223.1	234508	Zinc finger and SCAN domain-containing protein	missense_variant
LFYR01000235.1	218393	putative Pentatricopeptide repeat-containing protein	missense_variant
LFYR01000235.1	285789	hypothetical protein	missense_variant
LFYR01000244.1	232160	hypothetical protein	missense_variant
LFYR01000252.1	192095	hypothetical protein	missense_variant
LFYR01000252.1	309674	putative Pentatricopeptide repeat-containing protein	missense_variant
LFYR01000277.1	492849	Replication factor C subunit 1	missense_variant
LFYR01000304.1	406193	hypothetical protein	missense_variant
LFYR01000337.1	251856	laccase 11	missense_variant
LFYR01000337.1	254492	Glycosyltransferase%2C putative%2C expressed	missense_variant
LFYR01000337.1	444568	hypothetical protein	missense_variant
LFYR01000514.1	340515	putative Eukaryotic translation initiation factor 3 subunit	missense_variant
LFYR01000514.1	650645	hypothetical protein	stop_gained
LFYR01000514.1	689301	hypothetical protein	missense_variant
LFYR01000379.1	385998	Protein GAMETE EXPRESSED 2	missense_variant
LFYR01000391.1	168010	hypothetical protein	missense_variant
LFYR01000391.1	345605	putative Pre-mRNA splicing factor	missense_variant
LFYR01000391.1	362247	RNA polymerase II C-terminal domain phosphatase-like protein	missense_variant
LFYR01000429.1	200880	Pentatricopeptide repeat-containing protein%2C mitochondrial	missense_variant
LFYR01000429.1	245300	hypothetical protein	missense_variant
LFYR01000429.1	268826	putative Pentatricopeptide repeat-containing protein	missense_variant
LFYR01000448.1	313902	Subtilisin-like serine protease	missense_variant

LFYR01000468.1	119622	plastid movementimpaired 2	missense_variant
LFYR01000468.1	197286	putative Tyrosine decarboxylase (or alternatively Aromatic aldehyde synthase)	missense_variant
LFYR01000468.1	361391	Calcium dependent protein kinase-like protein	missense_variant
LFYR01000483.1	78962	RIC1-like protein	missense_variant
LFYR01000483.1	127083	putative Protein kinase	missense_variant
LFYR01000513.1	316926	putative Bystin	missense_variant
LFYR01000585.1	335268	Pleiotropic drug resistance protein 12	missense_variant
LFYR01000585.1	385957	Polygalacturonase%2C family GH28	missense_variant
LFYR01000585.1	458340	S-adenosyl-L-methionine-dependent methyltransferases superfamilyprotein	missense_variant
LFYR01000537.1	250994	Prenylated rab acceptor family protein	missense_variant
LFYR01000550.1	62555	Phosphoinositide phospholipase C	missense_variant
LFYR01000550.1	212952	hypothetical protein	missense_variant
LFYR01000562.1	179904	Sumo activating enzyme 1b	stop_gained
LFYR01000569.1	157785	hypothetical protein	missense_variant
LFYR01000574.1	247581	hypothetical protein	missense_variant
LFYR01000580.1	230737	Basic helix-loop-helix (BHLH) family transcription factor	missense_variant
LFYR01000620.1	105726	50S ribosomal protein L4	missense_variant
LFYR01000620.1	198800	WRKY transcription factor 29	missense_variant
LFYR01000620.1	198890	WRKY transcription factor 29	missense_variant
LFYR01000620.1	451510	hypothetical protein	missense_variant
LFYR01000620.1	627773	Signal recognition particle 72 kDa protein	missense_variant
LFYR01000620.1	627774	Signal recognition particle 72 kDa protein	missense_variant
LFYR01000620.1	731158	WD repeat-containing protein-like protein	missense_variant
LFYR01000620.1	1023449	hypothetical protein	missense_variant
LFYR01000620.1	1072893	hypothetical protein	stop_gained
LFYR01000593.1	196309	Premnaspirodiene oxygenase	missense_variant
LFYR01000601.1	50464	Matrix metalloproteinase-9	missense_variant
LFYR01000601.1	360584	UDP-D-apiose/UDP-D-xylose synthase 2	missense_variant
LFYR01000611.1	60466	Pleiotropic drug resistance protein 12	missense_variant
LFYR01000611.1	262069	Pentatricopeptide repeat-containing protein	missense_variant
LFYR01000611.1	319083	auxin transport protein (BIG)	missense_variant

LFYR01000611.1	346696	putative Oligopeptidase B	missense_variant
LFYR01000614.1	370462	Zinc finger protein-like protein	missense_variant
LFYR01000619.1	113391	Ubiquitin-protein ligase 4	missense_variant
LFYR01000643.1	88182	hypothetical protein	missense_variant
LFYR01000621.1	257851	hypothetical protein	missense_variant
LFYR01000624.1	12217	endo-beta-1%2C4-glucanase%2C family GH9	missense_variant
LFYR01000624.1	15491	hypothetical protein	missense_variant
LFYR01000624.1	32872	Succinate dehydrogenase [ubiquinone] iron-sulfur subunit 2%2C mitochondrial	missense_variant
LFYR01000624.1	139339	Amino acid permease I%2C putative%2C expressed	missense_variant
LFYR01000624.1	202606	Pyruvate kinase	missense_variant
LFYR01000625.1	232412	hypothetical protein	missense_variant
LFYR01000625.1	298007	Nascent polypeptide-associated complex subunit beta	missense_variant
LFYR01000633.1	261719	Alpha-sialyltransferase%2C family GT29	missense_variant
LFYR01000633.1	302336	putative Pentatricopeptide repeat-containing protein	missense_variant
LFYR01000640.1	183168	putative dolichyl-diphosphooligosaccharide--protein glycosyltransferasesubunit 3B	missense_variant
LFYR01000642.1	381130	putative glucose-6-phosphate 1-epimerase	missense_variant
LFYR01000671.1	195340	putative Protein kinase	stop_gained
LFYR01000671.1	364023	hypothetical protein	missense_variant
LFYR01000671.1	416612	S-adenosyl-L-methionine-dependent methyltransferases superfamilyprotein	missense_variant
LFYR01000671.1	424906	hypothetical protein	missense_variant
LFYR01000671.1	527751	SEC1 family transport protein SLY1	missense_variant
LFYR01000671.1	881966	Glutathione S-transferase-O-methyltransferase fusion protein 7	missense_variant
LFYR01000671.1	914145	Cyclic nucleotide-gated channel	missense_variant
LFYR01000647.1	137615	Pentatricopeptide repeat protein	stop_gained
LFYR01000650.1	34447	Early response to dehydration 15-like protein	missense_variant
LFYR01000650.1	251885	hypothetical protein	missense_variant
LFYR01000653.1	75616	light-harvesting complex I chlorophyll a/b binding protein 1	missense_variant
LFYR01000653.1	77596	hypothetical protein	missense_variant
LFYR01000655.1	300281	Protein argonaute 10	missense_variant

LFYR01000655.1	300284	Protein argonaute 10	missense_variant
LFYR01000658.1	143521	hypothetical protein	missense_variant
LFYR01000658.1	216880	hypothetical protein	missense_variant
LFYR01000661.1	303327	Exo-polygalacturonase%2C family GH28	missense_variant
LFYR01000661.1	379790	Pentatricopeptide repeat (PPR) family protein 1	missense_variant
LFYR01000664.1	66964	hypothetical protein	missense_variant
LFYR01000667.1	69327	Leucine-rich repeat receptor-like protein kinase family	missense_variant
LFYR01000670.1	175802	Argininosuccinate lyase	missense_variant
LFYR01000692.1	122720	chromosome-associated kinesin KIF4-like protein	missense_variant
LFYR01000692.1	562138	Pentatricopeptide repeat-containing protein	missense_variant
LFYR01000692.1	713651	Calcium-dependent protein kinase-like protein	stop_gained
LFYR01000692.1	989956	GDSL esterase/lipase	missense_variant
LFYR01000691.1	34561	Type I inositol-1%2C4%2C5-trisphosphate 5-phosphatase	missense_variant
LFYR01000728.1	989012	Protein STICHEL-like 4	missense_variant
LFYR01000696.1	229283	S-locus receptor kinase (SRK)	missense_variant
LFYR01000705.1	213110	ATP-dependent zinc metalloprotease FtsH 4	missense_variant
LFYR01000709.1	120595	Serine carboxypeptidase II-2	missense_variant
LFYR01000709.1	176406	hypothetical protein	missense_variant
LFYR01000718.1	202196	C2 and GRAM domain-containing protein	missense_variant
LFYR01000721.1	59330	Protein ULTRAPETALA 1	missense_variant
LFYR01000727.1	94117	ATP-dependent DNA helicase RecG	missense_variant
LFYR01000981.1	1084732	Long-chain fatty alcohol oxidase	missense_variant
LFYR01000981.1	1162741	Dipeptidyl peptidase	missense_variant
LFYR01000981.1	1280004	ABC transporter G family member	missense_variant
LFYR01000981.1	1311358	hypothetical protein	missense_variant
LFYR01000981.1	1693711	hypothetical protein	missense_variant
LFYR01000981.1	2505654	hypothetical protein	missense_variant
LFYR01000757.1	354671	Uncharacterized protein	missense_variant
LFYR01000757.1	742011	Pentatricopeptide repeat-containing protein	missense_variant
LFYR01000731.1	194102	Serine/threonine protein phosphatase 4 regulatory subunit	missense_variant

		putative agamous-like MADS-box protein AGL21-related ZmMADS4%2C	
LFYR01000740.1	138335	AGL17/ANR1 sub-family	missense_variant
LFYR01000740.1	215088	hypothetical protein	missense_variant
LFYR01000740.1	245246	hypothetical protein	missense_variant
LFYR01000740.1	306569	Histone-lysine N-methyltransferase	missense_variant
LFYR01000753.1	224748	putative DELLA protein GAI	missense_variant
LFYR01000756.1	135567	S-adenosyl-L-methionine-dependent methyltransferases superfamilyprotein	missense_variant
LFYR01000785.1	955152	Rac-like GTP-binding protein ARAC3	missense_variant
LFYR01000759.1	274057	hypothetical protein	missense_variant
LFYR01000760.1	16674	hypothetical protein	missense_variant
LFYR01000760.1	200406	hypothetical protein	missense_variant
LFYR01000769.1	57436	HEAT repeat-containing protein	missense_variant
LFYR01000769.1	57437	HEAT repeat-containing protein	missense_variant
LFYR01000769.1	260986	hypothetical protein	missense_variant
LFYR01000773.1	179496	Cyanase	stop_gained
LFYR01000777.1	158440	hypothetical protein	missense_variant
LFYR01000777.1	162462	hypothetical protein	missense_variant
LFYR01000781.1	45900	hypothetical protein	missense_variant
LFYR01000781.1	212290	Chaperone protein clpB	stop_gained
LFYR01000781.1	249743	myb domain protein 36	missense_variant
LFYR01000811.1	868520	putative 26S protease (S4) regulatory subunit	stop_gained
LFYR01000788.1	326633	hypothetical protein	start_lost
LFYR01000789.1	161234	hypothetical protein	missense_variant
LFYR01000791.1	250588	Homogentisate 1%2C2-dioxygenase	missense_variant
LFYR01000791.1	289784	Ubiquitin carboxyl-terminal hydrolase-like protein	missense_variant
LFYR01000794.1	143382	hypothetical protein	missense_variant
LFYR01000800.1	130615	CBS domain containing protein	missense_variant
LFYR01000804.1	305307	hypothetical protein	missense_variant
LFYR01000809.1	77954	U-box domain-containing protein 15	missense_variant
LFYR01000839.1	434370	hypothetical protein	stop_gained
LFYR01000839.1	596024	Lipoxygenase (potentially 13-LOX)	missense_variant

LFYR01000817.1	198328	11S seed storage globulin B	missense_variant
LFYR01000825.1	27848	hypothetical protein	missense_variant
LFYR01000834.1	88245	Cytochrome b5	missense_variant
LFYR01000838.1	45826	RNA polymerase II C-terminal domain phosphatase-like protein	missense_variant
LFYR01000864.1	140758	Beta-xylosidase%2C family GH3	missense_variant
LFYR01000864.1	159534	hypothetical protein	missense_variant
LFYR01000864.1	465408	hypothetical protein	missense_variant
LFYR01000845.1	119284	Zinc finger family protein	missense_variant
LFYR01000845.1	218136	putative LRR receptor-like serine/threonine-protein kinase	missense_variant
LFYR01000857.1	167275	Thioredoxin domain-containing protein	missense_variant
LFYR01000889.1	427938	3-ketoacyl-CoA synthase	missense_variant
LFYR01000867.1	101145	Tetratricopeptide repeat-containing protein	missense_variant
LFYR01000869.1	24956	Kinase family protein	missense_variant
LFYR01000874.1	33731	hypothetical protein	missense_variant
LFYR01000880.1	1556	Nuclear matrix constituent protein 2	stop_gained
LFYR01000880.1	154019	Subtilisin-like serine protease	stop_gained
LFYR01000884.1	6184	hypothetical protein	missense_variant
LFYR01000884.1	117508	Cyclin A-like protein	missense_variant
LFYR01000915.1	201233	hypothetical protein	missense_variant
LFYR01000915.1	348335	MATE efflux family protein	missense_variant
LFYR01000915.1	678512	Protein SCAI	missense_variant
LFYR01000915.1	798507	pectin methylesterase%2C family CE8	missense_variant
LFYR01000897.1	46022	hypothetical protein	missense_variant
LFYR01000901.1	134098	hypothetical protein	missense_variant
LFYR01000914.1	79943	Hypothetical protein	missense_variant
LFYR01000914.1	170101	hypothetical protein	missense_variant
LFYR01000932.1	306235	Lipid tranfer protein	missense_variant
LFYR01000932.1	815783	Glutamate--ammonia ligase	missense_variant
LFYR01000916.1	97081	Nucleotidyltransferase domain containing protein%2C expressed	missense_variant
LFYR01000923.1	88837	hypothetical protein	missense_variant
LFYR01000923.1	144477	Ethylene-overproduction protein 1	missense_variant

LFYR01000926.1	184478	putative Protein kinase	missense_variant
LFYR01000926.1	184484	putative Protein kinase	missense_variant
LFYR01000926.1	184486	putative Protein kinase	missense_variant
LFYR01000926.1	184496	putative Protein kinase	missense_variant
LFYR01000926.1	184501	putative Protein kinase	missense_variant
LFYR01000926.1	184502	putative Protein kinase	missense_variant
LFYR01000926.1	184525	putative Protein kinase	missense_variant
LFYR01000926.1	184526	putative Protein kinase	missense_variant
LFYR01000926.1	184543	putative Protein kinase	missense_variant
LFYR01000931.1	93924	Multidrug resistance protein ABC transporter family	missense_variant
LFYR01000958.1	852660	hypothetical protein	missense_variant
LFYR01000955.1	28227	Ubiquitin conjugation factor E4	missense_variant
LFYR01000955.1	122105	GDSL esterase/lipase	missense_variant
LFYR01000957.1	117917	Glucomannan 4-beta-mannosyltransferase%2C family GT2	missense_variant
LFYR01000980.1	328461	Beta-1%2C4-mannosyl-glycoprotein beta-1%2C4-N-acetylglucosaminyltransferase%2C family GT17	missense_variant
LFYR01000980.1	361039	Polygalacturonase%2C family GH28	missense_variant
LFYR01000980.1	427850	60S ribosomal protein L30	missense_variant
LFYR01000980.1	748609	Trehalose-6-Phosphate Synthase	missense_variant
LFYR01000980.1	768683	Myb family transcription factor-like protein	missense_variant
LFYR01000962.1	117143	hypothetical protein	missense_variant
LFYR01000962.1	138471	AP2-like ethylene-responsive transcription factor	missense_variant
LFYR01001213.1	869376	hypothetical protein	missense_variant
LFYR01001213.1	1378453	putative Origin recognition complex subunit	missense_variant
LFYR01001011.1	836634	Isoamylase%2C family GH13	missense_variant
LFYR01001011.1	841508	hypothetical protein	missense_variant
LFYR01000985.1	122940	pectin methylesterase%2C family CE8	stop_gained
LFYR01000988.1	154910	TBC1 domain family member 8B	missense_variant
LFYR01000991.1	86843	hypothetical protein	missense_variant
LFYR01000994.1	78043	hypothetical protein	missense_variant
LFYR01000997.1	166301	Pre-mRNA-splicing factor CLF1	missense_variant

LFYR01000999.1	82833	hypothetical protein	missense_variant
LFYR01000999.1	104407	hypothetical protein	stop_gained
LFYR01001005.1	54504	hypothetical protein	missense_variant
LFYR01001032.1	82542	Calcium-dependent protein kinase isoform 2	missense_variant
LFYR01001032.1	181181	hypothetical protein	missense_variant
LFYR01001032.1	308068	hypothetical protein	missense_variant
LFYR01001018.1	114843	Splicing factor 3B subunit	missense_variant
LFYR01001018.1	171045	putative Ribosomal RNA small subunit methyltransferase	missense_variant
LFYR01001054.1	205441	hypothetical protein	missense_variant
LFYR01001054.1	276100	hypothetical protein	missense_variant
LFYR01001054.1	479339	Kinesin heavy chain isolog	missense_variant
LFYR01001041.1	116413	hypothetical protein	missense_variant
LFYR01001044.1	104906	Protein TSSC1	missense_variant
LFYR01001077.1	73228	hypothetical protein	missense_variant
LFYR01001077.1	434566	hypothetical protein	missense_variant
LFYR01001077.1	620351	hypothetical protein	missense_variant
LFYR01001077.1	822775	Nodulin-like / Major Facilitator Superfamily protein	missense_variant
LFYR01001059.1	83991	NB-ARC domain-containing disease resistance protein	missense_variant
LFYR01001070.1	106890	Receptor-like kinase	stop_gained
LFYR01001099.1	331733	hypothetical protein	missense_variant
LFYR01001099.1	634699	Pentatricopeptide repeat protein 91	missense_variant
LFYR01001099.1	635522	Pentatricopeptide repeat protein 91	missense_variant
LFYR01001099.1	774587	Peroxyureidoacrylate/ureidoacrylate amidohydrolase RutB	missense_variant
LFYR01001079.1	67116	hypothetical protein	missense_variant
LFYR01001090.1	53512	hypothetical protein	missense_variant
LFYR01001090.1	67563	hypothetical protein	missense_variant
LFYR01001097.1	158959	NAD-dependent epimerase/dehydratase	missense_variant
LFYR01001125.1	323841	Pentatricopeptide repeat-containing protein	missense_variant
LFYR01001151.1	591834	hypothetical protein	missense_variant
LFYR01001173.1	163707	CCR4-NOT transcription complex subunit 3	missense_variant
LFYR01001173.1	256746	CBM43-containing protein	missense_variant

LFYR01001173.1	578028	Phytochrome-associated serine/threonine-protein phosphatase 3	missense_variant
LFYR01001161.1	94372	Nicotinate phosphoribosyltransferase	missense_variant
LFYR01001162.1	37816	Kinase	missense_variant
LFYR01001167.1	119069	ABC transporter G family member 1	missense_variant
LFYR01001172.1	50426	hypothetical protein	missense_variant
LFYR01001193.1	227094	hypothetical protein	missense_variant
LFYR01001193.1	630207	hypothetical protein	missense_variant
LFYR01001212.1	241384	hypothetical protein	missense_variant
LFYR01001195.1	86522	SNF2 helicase	missense_variant
LFYR01001197.1	85728	putative ubiquitin-conjugating enzyme E2 16	missense_variant
LFYR01001430.1	298759	Mutator-like transposase	stop_gained
LFYR01001430.1	989173	hypothetical protein	missense_variant
LFYR01001430.1	1099632	Alcohol dehydrogenase-like 6	missense_variant
LFYR01001430.1	1242274	hypothetical protein	missense_variant
LFYR01001430.1	1322500	NADP-dependent alkenal double bond reductase P2	missense_variant
LFYR01001430.1	1348555	26S protease regulatory subunit	missense_variant
LFYR01001430.1	1449872	Flavonoid glucosyltransferase%2C family GT1	missense_variant
LFYR01001430.1	1502360	Kinase	missense_variant
LFYR01001237.1	292498	Stomatal cytokinesis defective / SCD1 protein	missense_variant
LFYR01001237.1	704276	Timeless family protein	missense_variant
LFYR01001219.1	72041	Eukaryotic aspartyl protease family protein	missense_variant
LFYR01001258.1	59379	hypothetical protein	missense_variant
LFYR01001258.1	205612	Kinesin-4	missense_variant
LFYR01001258.1	617682	hypothetical protein	missense_variant
LFYR01001279.1	226808	ATP-dependent Clp protease proteolytic subunit	missense_variant
LFYR01001269.1	3994	hypothetical protein	missense_variant
LFYR01001274.1	76876	ferulate 5-hydroxylase	missense_variant
LFYR01001305.1	555252	NAC domain protein	missense_variant
LFYR01001330.1	454713	Glucosamine 6-phosphate N-acetyltransferase	missense_variant
LFYR01001315.1	62439	hypothetical protein	missense_variant
LFYR01001351.1	610127	Cellulose synthase-like CSLD%2C family GT2	missense_variant

LFYR01001335.1	86420	hypothetical protein	missense_variant
LFYR01001368.1	686483	hypothetical protein	missense_variant
LFYR01001368.1	695594	putative Ring finger protein	stop_gained
LFYR01001368.1	786880	hypothetical protein	missense_variant
LFYR01001364.1	46502	Ubiquitin-40S ribosomal protein S27a-1	missense_variant
LFYR01001364.1	64237	Galactinol synthase 1	missense_variant
LFYR01001390.1	149649	MATE efflux family protein expressed	stop_gained
LFYR01001390.1	665136	hypothetical protein	missense_variant
LFYR01001389.1	61472	Flap endonuclease 1	missense_variant
LFYR01001410.1	178675	hypothetical protein	missense_variant
LFYR01001410.1	311462	hypothetical protein	missense_variant
LFYR01001410.1	338012	hypothetical protein	missense_variant
LFYR01001410.1	791373	hypothetical protein	missense_variant
LFYR01001429.1	745545	FAR1-related sequence 5	missense_variant
LFYR01001418.1	51940	hypothetical protein	missense_variant
LFYR01001623.1	122935	hypothetical protein	missense_variant
LFYR01001623.1	123368	hypothetical protein	missense_variant
LFYR01001623.1	452726	Flavonoid glucosyltransferase%2C family GT1	missense_variant
LFYR01001623.1	1161754	hypothetical protein	missense_variant
LFYR01001488.1	279632	starch phosphorylase%2C family GT35	missense_variant
LFYR01001488.1	289120	hypothetical protein	missense_variant
LFYR01001488.1	462334	hypothetical protein	missense_variant
LFYR01001488.1	624407	Protein transport protein SEC61 gamma subunit	missense_variant
LFYR01001475.1	30205	Dolichyl-diphosphooligosaccharide--protein glycotransferase	missense_variant
LFYR01001477.1	13761	Vacuolar processing enzyme 1	missense_variant
LFYR01001477.1	13769	Vacuolar processing enzyme 1	stop_gained
LFYR01001477.1	13771	Vacuolar processing enzyme 1	missense_variant
LFYR01001477.1	13772	Vacuolar processing enzyme 1	missense_variant
LFYR01001477.1	13774	Vacuolar processing enzyme 1	missense_variant
LFYR01001508.1	249327	U-box domain-containing protein	missense_variant
LFYR01001508.1	579627	Subtilisin-like serine protease	missense_variant

LFYR01001491.1	29223	putative Protein kinase	missense_variant
LFYR01001495.1	8431	Pentatricopeptide repeat-containing protein	missense_variant
LFYR01001529.1	55177	putative Magnesium and cobalt efflux protein corC	missense_variant
LFYR01001529.1	413516	Chloroplast lipocalin	missense_variant
LFYR01001529.1	493041	hypothetical protein	missense_variant
LFYR01001520.1	19301	Phosphoserine transaminase	stop_gained
LFYR01001545.1	495721	hypothetical protein	missense_variant
LFYR01001565.1	232159	Glycine-rich RNA-binding protein 4%2C mitochondrial	missense_variant
LFYR01001565.1	475179	Receptor-like protein kinase	missense_variant
LFYR01001587.1	348326	hypothetical protein	missense_variant
LFYR01001587.1	383475	hypothetical protein	missense_variant
LFYR01001587.1	552178	hypothetical protein	missense_variant
LFYR01001587.1	585332	hypothetical protein	missense_variant
LFYR01001606.1	499143	Peptidyl-prolyl cis-trans isomerase	missense_variant
LFYR01001592.1	16803	hypothetical protein	missense_variant
LFYR01001599.1	42618	S-locus receptor kinase (SRK)	missense_variant
LFYR01001622.1	594277	putative Pentatricopeptide repeat-containing protein	missense_variant
LFYR01001803.1	182636	putative protein S-acyltransferase 4	missense_variant
LFYR01001803.1	668048	WD repeat-containing protein	missense_variant
LFYR01001640.1	110321	hypothetical protein	missense_variant
LFYR01001680.1	45653	hypothetical protein	missense_variant
LFYR01001699.1	317805	hypothetical protein	stop_gained
LFYR01001714.1	75983	hypothetical protein	missense_variant
LFYR01001714.1	215981	Cell division cycle protein 27/anaphase promoting complex subunit3	missense_variant
LFYR01001714.1	319869	hypothetical protein	missense_variant
LFYR01001714.1	409475	hypothetical protein	missense_variant
LFYR01001714.1	577123	hypothetical protein	missense_variant
LFYR01001739.1	57012	hypothetical protein	missense_variant
LFYR01001739.1	213877	Zinc finger A20 and AN1 domain-containing stress-associated protein	missense_variant
LFYR01001739.1	286023	Beta-galactosidase%2C family GH35	missense_variant
LFYR01001739.1	384371	putative Myosin XI	missense_variant

LFYR01001757.1	632100	hypothetical protein	missense_variant
LFYR01001785.1	399805	1%2C3-beta-glucan synthase	missense_variant
LFYR01001785.1	375658	1%2C3-beta-glucan synthase	missense_variant
LFYR01001802.1	200645	Potassium transporter 11	missense_variant
LFYR01001802.1	446726	Adenosine kinase	missense_variant
LFYR01001978.1	214078	hypothetical protein	missense_variant
LFYR01001978.1	215049	hypothetical protein	missense_variant
LFYR01001978.1	794504	Glycosyltransferase%2C family GT4	missense_variant
LFYR01001823.1	243562	Structural maintenance of chromosomes protein	missense_variant
LFYR01001858.1	38675	Nucleoside-diphosphate kinase	missense_variant
LFYR01001858.1	52201	Early nodulin-like protein 3	missense_variant
LFYR01001858.1	182584	putative serine/threonine-protein kinase GCN2	missense_variant
LFYR01001858.1	477032	putative protein phosphatase 2C 44	missense_variant
LFYR01001858.1	561874	Cation transporter HKT1	missense_variant
LFYR01001882.1	58691	pectin acetylesterase%2C family CE13	stop_gained
LFYR01001898.1	368682	Lysine ketoglutarate reductase trans-splicing related 1	missense_variant
LFYR01001913.1	114642	hypothetical protein	missense_variant
LFYR01001913.1	171384	Alpha-taxilin%2C putative%2C expressed	missense_variant
LFYR01001913.1	614496	Calmodulin	missense_variant
LFYR01001927.1	221533	LATE ELONGATED HYPOCOTYL transcription factor	missense_variant
LFYR01001927.1	391731	Homeobox-leucine zipper protein HOX3	missense_variant
LFYR01001917.1	4939	hypothetical protein	missense_variant
LFYR01001945.1	250843	Phospholipase D delta	missense_variant
LFYR01001945.1	558711	hypothetical protein	missense_variant
LFYR01001962.1	555317	putative peptide transporter%2C Protein NRT1/ PTR FAMILY 5.14	missense_variant
LFYR01002110.1	361385	Prolyl 4-hydroxylase alpha subunit-like protein	missense_variant
LFYR01002110.1	371849	hypothetical protein	missense_variant
LFYR01002110.1	413448	hypothetical protein	missense_variant
LFYR01002110.1	1157193	Pyrroline-5-carboxylate reductase	missense_variant
LFYR01001997.1	429151	hypothetical protein	missense_variant
LFYR01002015.1	99017	G-box binding factor	missense_variant

LFYR01002015.1	282038	hypothetical protein	missense_variant
LFYR01002002.1	8918	hypothetical protein	missense_variant
LFYR01002027.1	126934	Polycomb group RING finger protein 4	missense_variant
LFYR01002027.1	615000	Glycerol-3-Phosphate Acyltransferase	missense_variant
LFYR01002037.1	1827	hypothetical protein	stop_lost
LFYR01002048.1	20455	hypothetical protein	stop_gained
LFYR01002048.1	350375	Protein TONSOKU	missense_variant
LFYR01002048.1	370603	Protein STRUBBELIG-RECEPTOR FAMILY 7	missense_variant
LFYR01002060.1	281289	hypothetical protein	missense_variant
LFYR01002060.1	574625	hypothetical protein	missense_variant
LFYR01002072.1	108712	DEAD-box ATP-dependent RNA helicase 28	missense_variant
LFYR01002072.1	152667	putative Xaa-Pro aminopeptidase P	missense_variant
LFYR01002072.1	232178	Protein kinase superfamily protein	missense_variant
LFYR01002072.1	248949	50S ribosomal protein L1	missense_variant
LFYR01002072.1	270341	Trichome birefringence	missense_variant
LFYR01002091.1	80416	ERECTA receptor like kinase	missense_variant
LFYR01002091.1	229295	hypothetical protein	missense_variant
LFYR01002091.1	300282	Rhodanese-like domain-containing protein	missense_variant
LFYR01002091.1	349809	Beta-1%2C4-mannosyl-glycoprotein beta-1%2C4-N-acetylglucosaminyltransferase%2C family GT17	missense_variant
LFYR01002091.1	498972	NB-ARC domain-containing disease resistance protein	missense_variant
LFYR01002101.1	501026	Translation initiation factor 2	missense_variant
LFYR01002109.1	148402	Oxidoreductase%2C 2OG-Fe(II) oxygenase family protein	stop_gained
LFYR01002228.1	174223	hypothetical protein	missense_variant
LFYR01002228.1	750249	Arf-GAP with GTPase%2C ANK repeat and PH domain-containing protein2	missense_variant
LFYR01002228.1	1189728	60S ribosomal protein L4/L1	missense_variant
LFYR01002228.1	1280883	hypothetical protein	missense_variant
LFYR01002125.1	80136	hypothetical protein	missense_variant
LFYR01002125.1	548239	hypothetical protein	missense_variant
LFYR01002138.1	68386	ARM repeat superfamily protein	missense_variant
LFYR01002147.1	328581	N-glycan beta-1%2C3-galactosyltransferase%2C family GT31	missense_variant

LFYR01002171.1	204609	hypothetical protein	missense_variant
LFYR01002171.1	339039	WD-40 repeat-containing protein	stop_gained
LFYR01002184.1	319030	hypothetical protein	missense_variant
LFYR01002227.1	309346	hypothetical protein	missense_variant
LFYR01002227.1	348719	Peptidyl-prolyl cis-trans isomerase	missense_variant
LFYR01002227.1	452912	hypothetical protein	stop_gained
LFYR01002227.1	473127	Alpha/beta-Hydrolases superfamily protein	missense_variant

Chapter 2: Population genomics of the model seagrass *Zostera marina* L. reveals swift and repeated trans-oceanic dispersal events

Lei Yu[1], Marina Khachaturyan[1, 2], Michael Matschiner[3, 4], Adam Healey[5], John J. Stachowicz[6, 7], Jeremy Schmutz[5], Tal Dagan[2], Jeanine L. Olsen[8], Thorsten B. H. Reusch[1*]

[1]Marine Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
[2]Institute of Microbiology, University of Kiel, Kiel, Germany
[3]Department of Palaeontology and Museum, University of Zurich, Zurich, Switzerland
[4]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, Oslo, Norway
[5]Genome Sequencing Center, HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA.
[6]Center for Population Biology, University of California, Davis, CA, USA
[7]Department of Evolution and Ecology, University of California, Davis, CA, USA
[8]Groningen Institute of Evolutionary Life Sciences, Groningen, AG, The Netherlands
[*]Corresponding author, treusch@geomar.de, phone +49-431-600-4550

Abstract

The ocean is open and features potentially unlimited genetic exchange which implies that marine organisms can disperse swiftly and cover large distribution ranges. Here, we performed a population genomic analysis of eelgrass (*Zostera marina* L.), the model seagrass species with circumglobal distribution, based on whole-genome resequencing of 190 individuals belonging to 16 locations throughout the northern hemisphere. We show the origin and circumglobal trans-oceanic colonization history of *Z. marina*. Based on genome-wide genetic diversity, we deduced the southern Japan to be the species' origin. Our data provided evidence for three trans-Pacific dispersal events and one very recent trans-Atlantic dispersal event (15.8 kya, 95% HPD: 19.7-12.3 kya). Multiple trans-Pacific colonization events are likely responsible for a complex, web-like topology within the Pacific Ocean. The populations in the Atlantic Ocean and the Mediterranean Sea showed much lower levels of genetic differentiation among populations and genetic diversity compared with the Pacific populations, which is likely due to the severe influence of the repeated glacial periods in combination with the relatively younger age. Our results indicate frequent and swift trans-oceanic dispersal events for a model seagrass with circumglobal distribution, which may explain a lack of speciation events across the seagrasses in general.

Keywords: seagrass, *Zostera marina*, whole-genome resequencing, population genomics, trans-oceanic dispersal, time-calibrated phylogeny

Introduction

Processes of colonization and dispersal that are responsible for biological diversity in the largest habitat on earth, the ocean, are very different from terrestrial habitats (Palumbi, 1992, 1994). In the marine environment, ocean currents, along with geographic/biogeographic barriers, determine the levels of connectivity among populations. These oceanographic processes must be considered for a meaningful interpretation of spatial population genetic patterns (White et al., 2010). While it is well established in terrestrial plant biogeography how ice ages have shaped demography and colonization patterns (Hewitt, 2000), much less is known for marine species. Large changes in sea level during glacial periods have also affected the marine angiosperms (or seagrasses), as they are confined to shallow water <50m depth. A range-wide sample collection including the putative area of species' origin is needed to understand the processes that have shaped the genetic structure and diversity, in order to make inferences on the past colonization routes and time scales.

Eelgrass (*Zostera marina* L.) is an emerging model species for studying seagrass genomics (Olsen et al., 2016; Ma et al., 2021). It features a circumglobal distribution in the northern hemisphere, and is also the only seagrass species with a published chromosome-level reference genome (Ma et al., 2021). Earlier assessments of its population structure were inconclusive regarding the timing of the major colonization events (Olsen et al., 2004), owing to the nature of the molecular markers used (i.e., allele-frequency based microsatellite data, and low levels of sequence variance of *mat*K and ITS). In addition, the Northwest Pacific region, the putative origin of seagrasses of the genus *Zostera* (Coyer et al., 2013), was not included (Olsen et al., 2004). Here, we sampled worldwide locations inhabited by *Z. marina* and provided a comprehensive genomic assessment using whole-genome resequencing data, in order to resolve the question on the origin of *Z. marina*, and to date major colonization events within and among Pacific and Atlantic. Given that many seagrass species are very widely distributed, including our model species that occurs around the northern hemisphere from subtropical to subarctic latitudes (Larkum et al., 2006), we were particularly interested as to how *Z. marina* had spread from its origin to eventually colonize the entire North Pacific and North Atlantic.

One major objective was obtaining a time-calibrated phylogenetic tree of our worldwide population samples. Owing to the lack of fossil calibration points, obtaining such temporal calibration was complex, but possible through the application of the recently published whole genome duplication (WGD) event of *Z. marina* (~67 mya) (Olsen et al., 2016). The molecular clock calculated from the paralogs originated from the WGD event was used to estimate the divergence time between *Z. marina* and an outgroup species (*Z. japonica*) (for details see Material and Methods), which served as a calibration point for dating the other divergence events. Stange et al. (2018) extended the multispecies coalescent (MSC) method implemented in SNAPP (Bryant et al., 2012), making it possible to construct phylogenetic tree directly based on SNP (single nucleotide polymorphism) data. This method has been successfully applied to population-level phylogenetics (Fang et al., 2020). Dating past evolutionary events requires nodes in a bifurcating tree that depict the events of population

divergence. However, algorithms to construct phylogenies are sensitive towards violations of the assumption of no secondary exchange among taxa, i.e., genetic admixture, which may actually be the rule rather than the exception in plant populations (Linder & Rieseberg, 2004). Thus, genetic admixture must be considered before constructing the phylogenetic tree.

Plant phylogenetic studies are mostly based on nuclear and chloroplast genomes due to the significant evolutionary plasticity in their mitochondrial DNA (Knoop, 2004), and such an approach will also be followed here. As a relatively long DNA sequence (115 to 165 kb for angiosperms) (Sun et al., 2020) without recombination, chloroplast genome can show the stepwise accumulation of mutations. However, inherited as a haplotype (Wicke et al., 2011), chloroplast genome is easily affected by genetic drift (Whitlock, 2003). On the contrary, the millions of independent loci in the nuclear genome can overcome the influence of genetic drift, but they don't show stepwise accumulation of mutations. Combining nuclear and chloroplast genomes can provide better understanding of the evolutionary history.

Here we performed whole-genome resequencing for comprehensive, range-wide *Z. marina* samples from 16 populations worldwide. We addressed the following main questions (1) where did *Z. marina* originate? (2) how has *Z. marina* dispersed to current distribution range? (3) what is the contemporary genetic population structure of *Z. marina*? (4) where were glacial refugia during the Last Glacial Maximum (LGM), and how have the glacial periods shaped the present genetic structure and diversity?

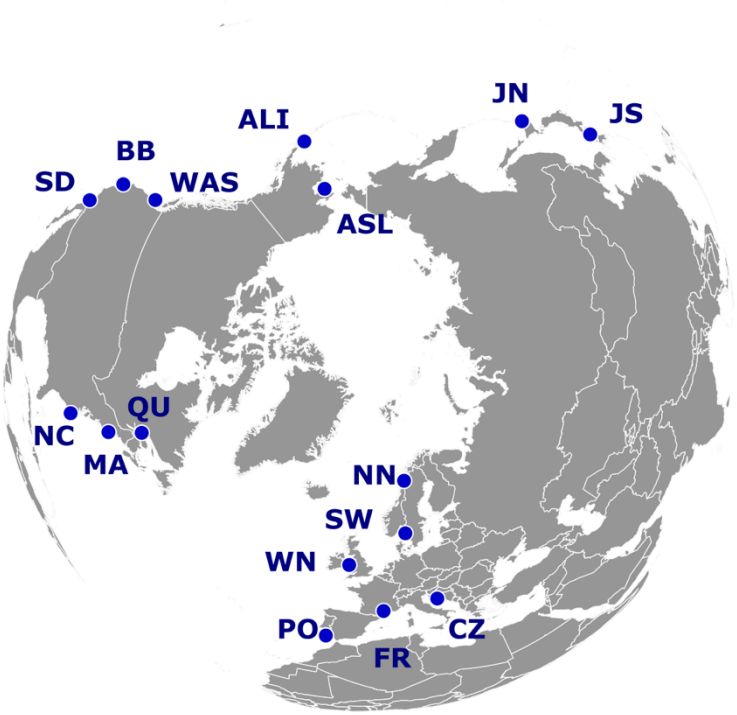


Fig. 1 | Sampling sites for *Zostera marina*. Abbreviations: San Diego, California (SD); Bodega Bay, California (BB); Washington state (WAS); Alaska-Izembek (ALI); Alaska-Safety Lagoon (ASL); Japan-North (JN); Japan-South (JS); North Carolina (NC); Massachusetts (MA); Quebec (QU); Northern Norway (NN); Sweden (SW); Wales North

(WN); Portugal (PO); Mediterranean France (FR); Croatia (CZ). Further metadata can be found in Supplementary Table 1.

Results

Whole-genome resequencing

Zostera marina is a plant featuring a mixture of clonal (=vegetative) and sexual reproduction, with varying proportions across sites. When exhibiting a clonal life-history strategy, any population consists of a series of modular units (= ramets) (Noble et al., 1979) belonging to sexually produced individuals, also called genets, which in turn form the meadow. A total of 190 *Z. marina* ramets were collected from 16 geographic locations, covering the distribution range of the species (Fig. 1, Supplementary Table 1).

After quality filtering of the whole-genome resequencing data of each sampled ramet, the filtered reads reached an average coverage of 53.73x (Supplementary Data 1). Genetic variants were called following the "Best Practice Workflow for Germline short variant discovery (SNPs + Indels)" using GATK4 (Van der Auwera & O'Connor, 2020) (Supplementary Fig. 1 & 2). We only focused on SNPs that were beyond 20 base pairs of an indel or other variant types, because indels (or other variant types) may cause artificial SNPs (Altmann et al., 2012). A total of 3,975,407 SNPs were retained after custom hard filtering. Genotypes that were outside our custom quality criteria were represented as missing data. The number of SNPs was reduced to 3,892,668 after omitting 82,739 SNPs with more than two alleles. Assuming different individuals within the same species have very similar genomes is inappropriate for a plant species with circumglobal distribution range. Genes are supposed to be more conservative, and are more likely to be shared by all individuals. We identified 18,717 genes that were on average observed in 97% of the ramets, which contained 763,580 SNPs.

For a parent-descendant pair under selfing, the descendant cannot be considered being formed by random mating. Accordingly, ten ramets possibly produced by selfing of the other collected genets were excluded (Supplementary Table 2, Supplementary Fig. 3; Yu et al., 2022). For any population genomic analysis, it is imperative to not duplicate any genet by including multiple sampled ramets (Supplementary Table 3, Supplementary Fig. 3; Yu et al. 2022). Thus, seventeen ramets representing replicated genotypes were also excluded. We further excluded 10 ramets with high missing rate (Supplementary Fig. 4), which also showed low mapping rate (Supplementary Data 2). As a result, a total of 37 ramets were excluded, and the rest 153 ramets representing unique genotypes were used for further analyses. To obtain putatively neutral SNPs, we extracted only synonymous SNPs in coding region (CDS). To assure the independency of SNPs, the dataset was further thinned using a distance (physical distance in the genome) filter of interval >3k bp, which led to the core SNP dataset containing 11,705 SNPs (ZM_Core_SNPs). We also used different filtering strategies for some specific analyses (Supplementary Fig. 1 & 2).

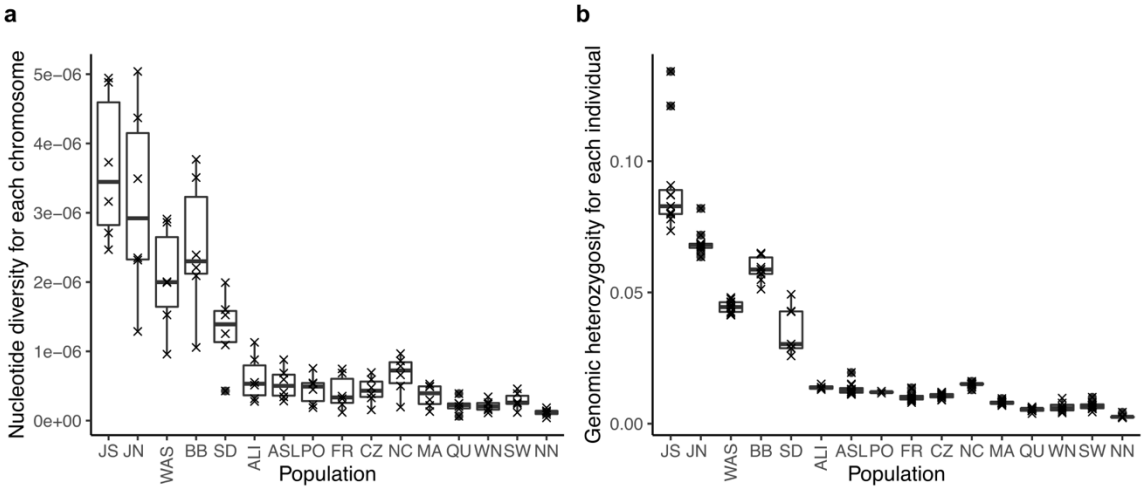


Fig. 2 | Genetic diversity across 16 worldwide eelgrass (*Zostera marina*) populations. a, Nucleotide diversity (π) was calculated for each of the six chromosomes based on the 11,705 SNP set (See Supplementary Fig. 2). Each data point indicates one chromosome. **b**, Genomic heterozygosity for an individual genome was calculated as (number of heterozygous sites) / (total number of sites with available genotype calls), based on the same dataset. Each data point indicates one unique ramet (=genotype).

Genetic diversity

As populations close to the origin of a species are expected to have the highest levels of genetic diversity, we assessed the genetic diversity for each sampled population, quantified as nucleotide diversity and genome-wide heterozygosity (Fig. 2). As expected, the Pacific side, the ocean basin of the origin of the genus *Zostera* (Coyer et al., 2013), displayed much higher genetic diversity than the Atlantic side (Atlantic + Mediterranean Sea), with the highest genetic diversity in southern Japan (JS), supporting the long-standing notion that *Z. marina* originated in the Northwest Pacific region before its spread throughout the entire northern hemisphere. This was in line with that the highest diversity of both genera and species of seagrasses was found in tropical to subtropical waters of the Indo-West Pacific (Olsen et al., 2004). Taking into account that *Z. marina* is a temperate species, we thus considered the southern Japan (JS) to be, or close to the species' origin. Alaska (ALI + ASL) showed the lowest genetic diversity among the Pacific populations, although still relatively high compared with the Atlantic side. This was probably due to the severe bottleneck caused by the Last Glacial Maximum (LGM). The Atlantic side featured a south-north gradient in genetic diversity. NC showed the highest genetic diversity; MA and the three populations from the Mediterranean Sea (PO, FR, and CZ) were intermediate, and the northern Europe featured the lowest measured in the entire dataset. The south-north gradient in genetic diversity within the Atlantic side indicated the northward colonization route after the LGM. The higher genetic diversity in the Mediterranean Sea than that in the northern Europe indicated that the Mediterranean Sea either was colonized earlier after the LGM, or had its own refugium during the LGM.

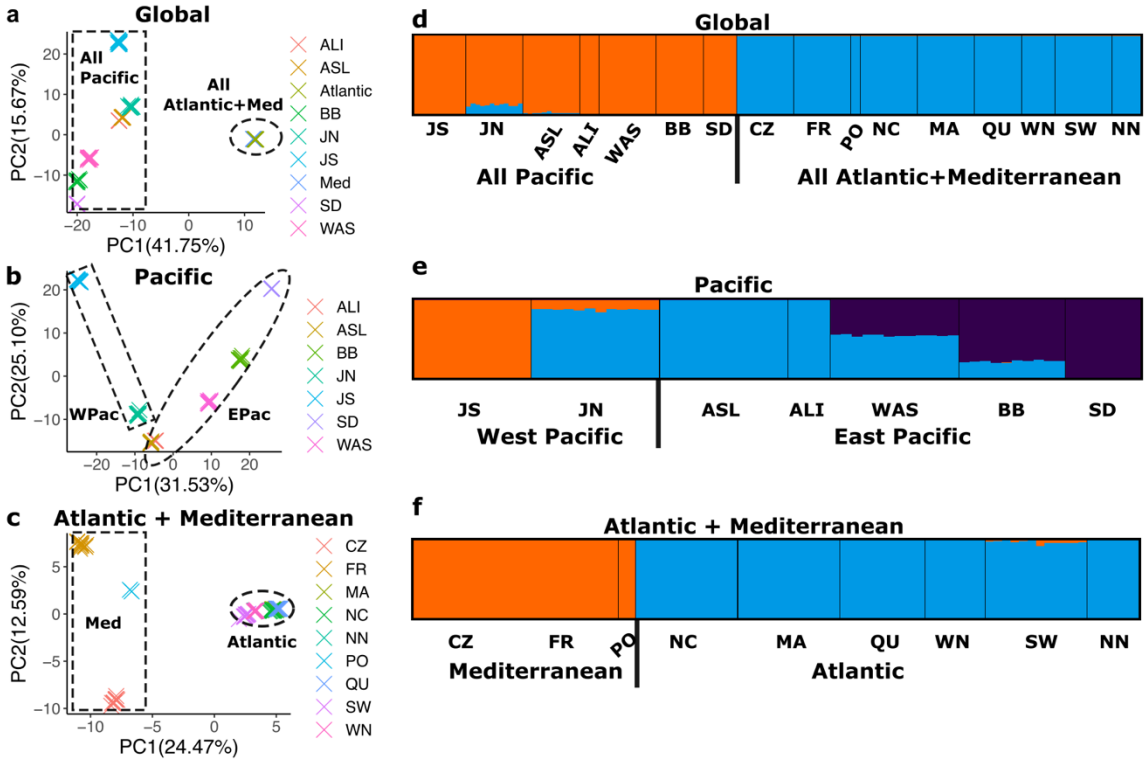


Fig. 3 | Population structure based on putatively neutral, synonymous and non-linked SNPs in the nuclear genome. Only non-redundant, non-inbred individuals were included. See Supplementary Fig. 1 for further details on the SNP sets used. **a-c**, Principal Component Analysis (PCA). **a**, Global (11,705 SNPs). Owing to their genetic similarity, all Atlantic populations are depicted with one color; likewise, all Mediterranean populations are depicted with one color. Pacific populations show much higher levels of diversity (Fig. 2). **b**, Pacific (12,514 SNPs). **c**, Atlantic and Mediterranean (8,552 SNPs). **d-f**, STRUCTURE analysis. Each individual is represented by a vertical bar partitioned into colors based on its K-group membership. Individuals consisting of more than one color indicate genetic admixture (JN, WAS and BB) in which the proportion of the color indicates the contribution of the corresponding group. Only the optimal number of genetic clusters/populations, K, as determined by delta-K method, is shown. See Supplementary Fig. 5-7 for results of all STRUCTURE results. For population abbreviations, see Fig. 1 and Supplementary Table 1. Note that colors should not be compared across the three panels, but only within a panel. **d**, Global (K = 2, 2,353 SNPs). **e**, Pacific (K = 3; 6,168 SNPs). **f**, Atlantic and Mediterranean (K = 2; 8,552 SNPs).

Genetic population structure

To reveal the basic genetic population structure among our 16 localities, we ran a Principal Component Analysis (PCA) for all populations based on the 11,705 putatively neutral, genic, synonymous and non-linked SNPs (ZM_Core_SNPs). The first principal component (PC) (41.75%) clearly separated the Pacific side and the Atlantic side, and the Pacific side showed much higher genetic variance than the Atlantic side, visible as spread among sampling locations (Fig. 3a). PCA was also run separately for the Pacific side and the Atlantic side, respectively. For the Pacific Ocean, the pattern was similar to the PCA based on all

populations (Fig. 3b). The population structure within the Atlantic side became much clearer after excluding the Pacific populations, and the first PC (24.47%) now separated the Atlantic Ocean and the Mediterranean Sea (Fig. 3c).

As an alternative method, we also analyzed the population structure with a Bayesian approach implemented in the software package STRUCTURE (Pritchard et al., 2000). Because of computing run-time constraints, STRUCTURE was run for all 16 populations based on 2,353 SNPs that were randomly selected from the 11,705 putatively neutral, genic, synonymous and non-linked SNPs (ZM_Core_SNPs). The optimal K value of the overall data set was K=2 (Supplementary Fig. 5; Evanno et al., 2005). The 16 populations fell into two well-defined genetic clusters, corresponding to locations in the Pacific side and the Atlantic side, respectively (Fig. 3d), consistent with the PCA result.

We then performed STRUCTURE analysis for each ocean basin separately, because we expected pronounced nested population structure within Atlantic and Pacific side, respectively (Janes et al., 2017). Since the separate analyses involved different sets of populations, we only used genomic loci showing polymorphism among the subset of populations. The nested population structure was confirmed, with a clear population structure within both the Pacific (6,168 SNPs) and the Atlantic side (8,552 SNPs), respectively. For the Pacific side, the optimal number of genetic clusters was 3 (Supplementary Fig. 6). JN, WAS, and BB showed signatures of admixture (Fig. 3e), visible as vertical bars displaying different colors (i.e., genetic components). JN showed admixture of JS and Alaska. WAS and BB were admixture of Alaska and SD; however, the proportions of the two genetic components were different. For the Atlantic side, the optimal number of genetic clusters was 2 (Supplementary Fig. 7). The Atlantic Ocean and the Mediterranean Sea were clearly separated (Fig. 3f), which was consistent with the PCA result.

Reticulated topology

To check whether the evolutionary history of all sampled sites was consistent with a bifurcating model, we constructed a split network based on the 11,705 putatively neutral, genic, synonymous and non-linked SNPs (ZM_Core_SNPs). A split network provided a combinatorial generalization of different tree topologies supported by different loci in the dataset (Huson & Bryant, 2006). Pacific populations were connected in a web-like fashion, particularly with populations WAS and BB being involved in alternative network edges (Fig. 4a), which was a major difference to the Atlantic side. This indicated complex evolutionary processes beyond a bifurcating model within the Pacific Ocean, possibly including secondary contact of diverged lineages (i.e., genetic admixture), in line with the admixture signatures showed by STRUCTURE (Fig. 3e). In contrast, the Atlantic side showed simple bifurcating fashion, and the branches were much shorter than those in the Pacific side, which indicated that the populations in the Atlantic side had started to diverge very recently.

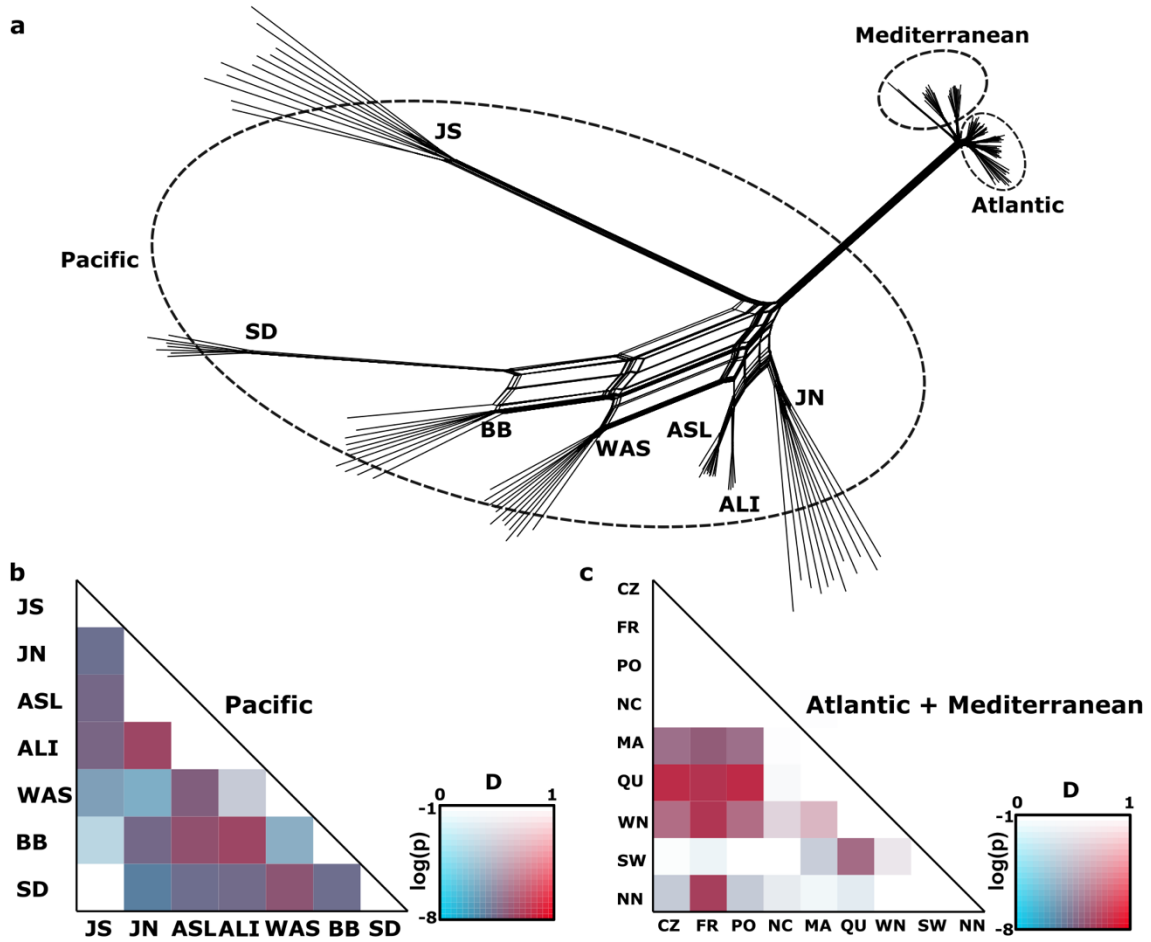


Fig. 4 | Evidence for genetic admixture among sampling sites within the Pacific Ocean and the Atlantic side, respectively. a, splitsTree network based on the 11,705 SNPs (See Supplementary Fig. 1). Each terminal branch indicates one individual. Individuals involved in more than one edge of the network indicate that portions of the genome support different evolutionary histories and are particularly strong with respect to BB and WAS. **b & c**, Patterson's D-matrix for admixture (also known as the ABBA-BABA statistic) for the Pacific Ocean and the Atlantic side, respectively. Red colors indicate higher D values, and more saturated colors indicate greater significance. Significant and high D values indicate admixture between the two populations, but direction is unknown. The admixture can be caused by direct gene flow between the two populations, or by genetic input from the same source population.

We also used the statistic Patterson's D (Patterson et al., 2012) to evaluate admixture in a quantitative way. Since the Pacific side and the Atlantic side showed a large genomic divergence, Patterson's D was calculated for each trio within the Pacific side (Fig. 4b) and the Atlantic side (Fig. 4c), respectively. For the Pacific side, WAS & SD, BB & ALI, BB & ASL and JN & ALI, showed the most significant D values. This may be caused by direct gene flow between the two populations, or by genetic input from the same source population. For the Atlantic side, the pattern was much clearer compared with the Pacific side, indicating connection between the Atlantic and the Mediterranean Sea, which was likely driven by the North Atlantic Drift.

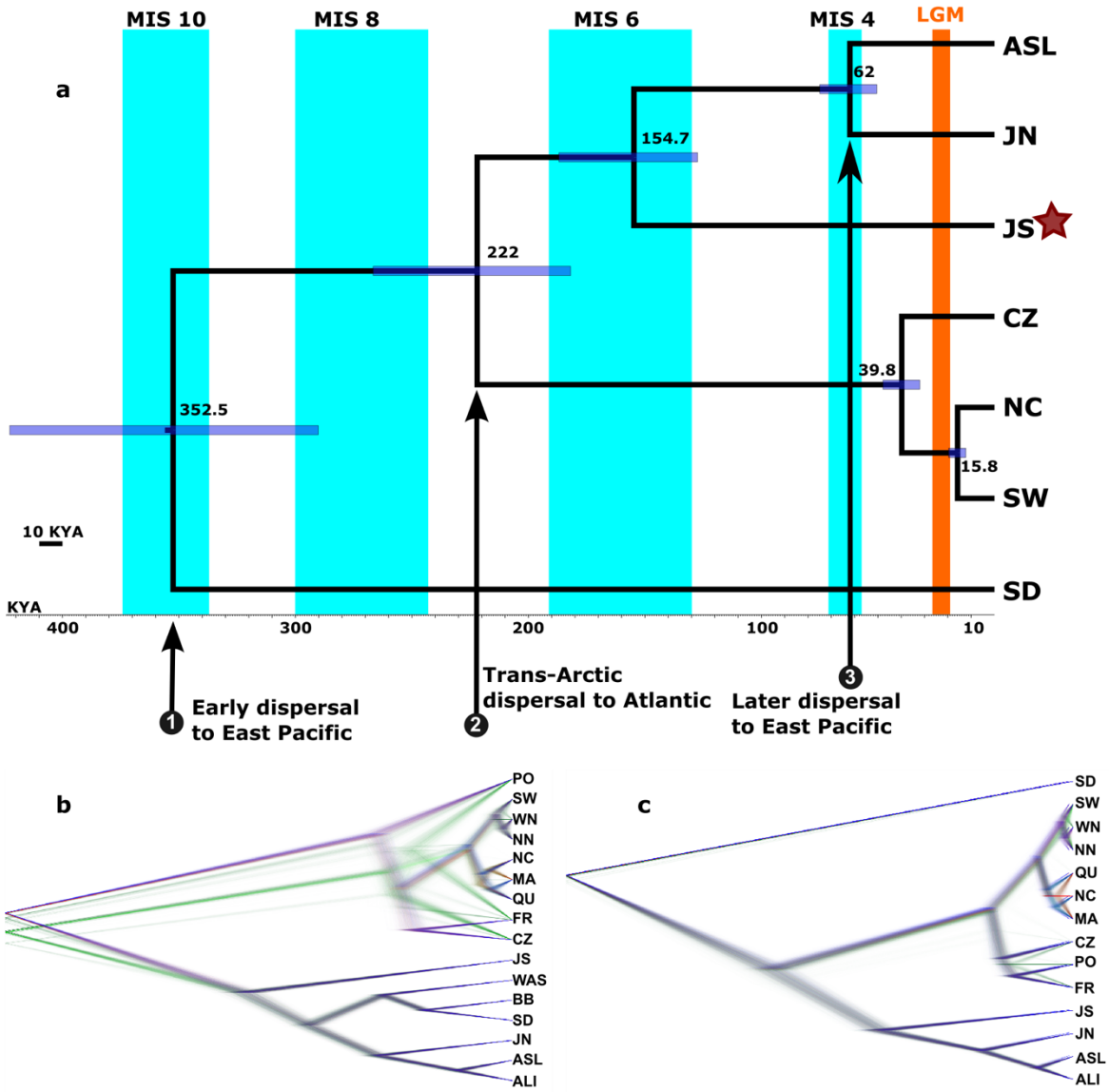


Fig. 5 | Time-calibrated phylogenetic tree constructed with the extended multi-species coalescent (MSC) method (Stange et al., 2018). a, Turquoise bars indicate glacial periods with Marine Isotope Stages (MIS) alternating with warm to cool interglacial periods. The orange bar indicates the Last Glacial Maximum (LGM, 26.5-19 kya). Blue bars across nodes indicate the 95% Highest Posterior Density (HPD) of the estimated divergence time. The star represents that *Zostera marina* and other *Zostera* species arose in the Japanese Archipelago (JS) and dispersed across the Pacific (Coyer et al., 2013). Key divergence events are shown. An accurate time estimate requires that the populations are not admixed, and thus some populations were removed due to admixture based on STRUCTURE (Fig. 3), splitsTree or D-statistics (Fig. 4). There is no strong admixture among the selected populations. **b**, DensiTree with all 16 populations, thus not accounting for admixture and creating error in the placement of SD. **c**, DensiTree with the most strongly admixed populations (Fig. 3 & 4), WAS and BB, removed.

Time-calibrated phylogenetic tree

Against the background of a marked global population structure, we next assessed when, in absolute time, major events of colonization occurred. To this end, the colonization history was analyzed with a time-calibrated phylogenetic tree using the method developed by Stange et al. (2018) which was based on coalescence model (Fig. 5). Because such an analysis assumes that all analyzed populations have diverged under a bifurcating model, and are not in exchange thereafter, some populations were excluded based on STRUCTURE (Fig. 3), splitstree, and D-statistics (Fig. 4). For the Pacific side, WAS, BB, and ALI were excluded, while for the Atlantic side, NC, SW, and CZ were selected to be representatives for the Northwest Atlantic, Northeast Atlantic, and the Mediterranean Sea, respectively. There was no strong admixture among the selected populations (Fig. 3 & 4). We will in the following only interpret the results from the reduced phylogenetic tree that only includes populations with little or no exchange (Fig. 5a).

Time calibration requires at least one calibration point, which is usually fossil evidence. To the best of our knowledge, no such information is available for the genus *Zostera*. As an alternative, the divergence time between *Z. marina* and *Z. japonica* was estimated based on a molecular clock (Supplementary Note) and used as calibration point to estimate the time of the other divergence events (for details of the time calibration see full Materials and Methods). Using the previously identified whole-genome duplication event of *Z. marina* (~67 mya) (Olsen et al., 2016), we first estimated the molecular clock of synonymous genic sites based on diverging paralogous gene sequences. Then, the divergence time between *Z. marina* and *Z. japonica* was estimated to be between 9.86 mya and 12.67 mya (Supplementary Note), which was consistent with the upper bound estimate of 11.9 mya from Coyer et al. (2013). Since most of the SNPs represented the marked genetic differences between *Z. marina* and the outgroup species *Z. japonica*, the topology within *Z. marina* was compressed. To obtain a clear relationship within *Z. marina* only, we extracted the time of the crown divergence for *Z. marina* populations as a new calibration point, and ran the analysis again for only *Z. marina* populations based on 13,732 genomic sites that were polymorphic within our target species (Supplementary Fig. 1).

Under a colonization scenario where populations have been colonized sequentially, the population in the origin of the species will first branch off from all the other populations. However, populations may also be colonized by multiple dispersal events from the origin. Under this scenario, the population located in the origin will first coalescence with the last colonized population. In our phylogeny (Fig. 5), the putative origin, Japan, did not branch off from all the other populations, and merged very recently with Alaska, which indicated the latter scenario.

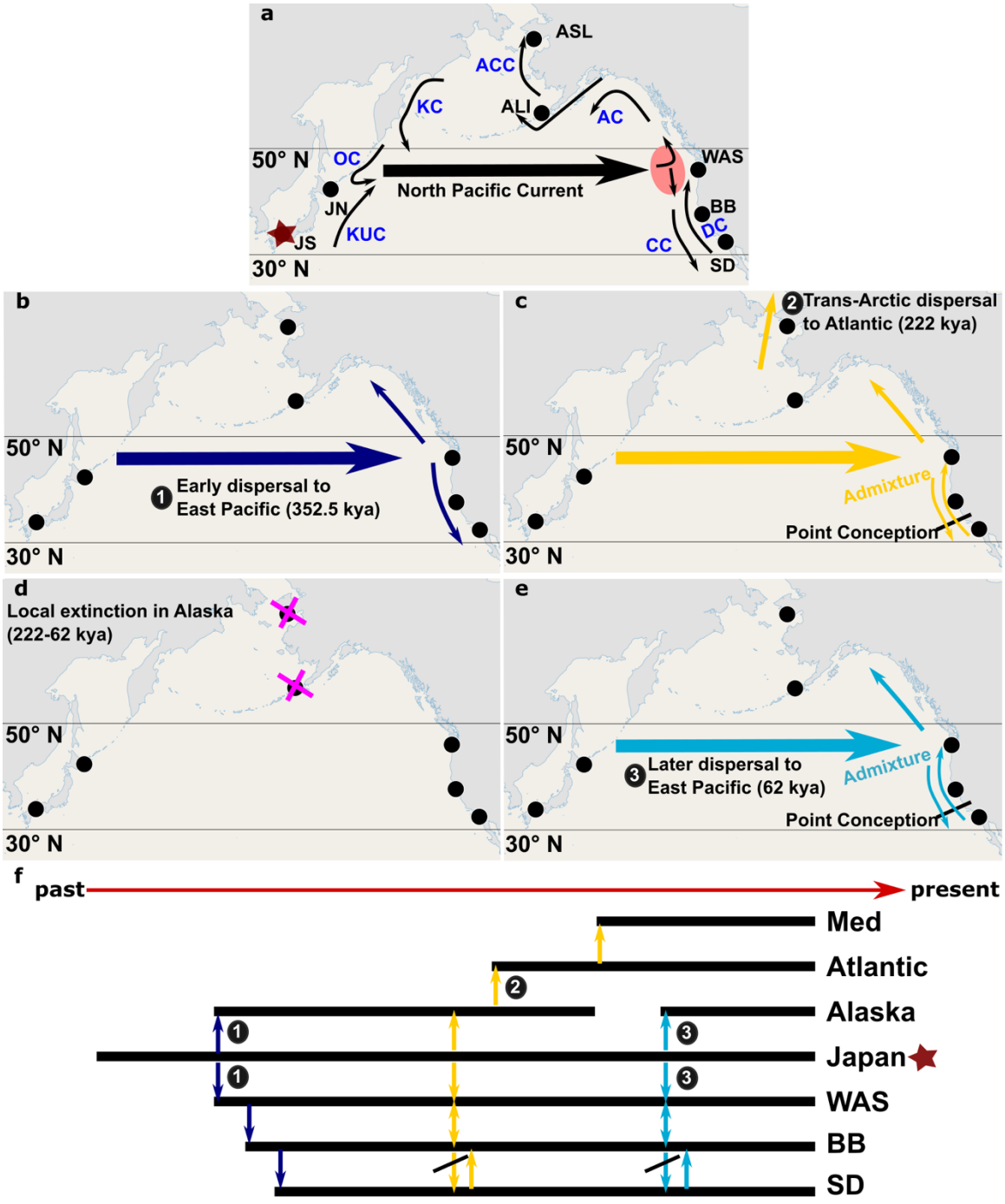


Fig. 6 | Ocean currents and colonization history. a, The star denotes the origin of *Zostera marina* and other *Zostera* species in the Japanese Archipelago. Dispersal across the Pacific is via the North Pacific Current, arriving at the "gateway", where it splits both south following the California Current, and north via the Alaska Current. The Alaska Current begins at about 48-50°N, and thus the Washington population WAS (46.47°N) is not affected, but mainly affected by the California Current. Solid arrows denote dispersal pathways along major currents and cross points: AC=Alaska Current, ACC=Alaska Coastal Current, KC=Kamchatka Current, OC=Oyashio Current, KUC=Kuroshio Current, DC= Davidson near shore current, CC=California Current. **b-e**, Major trans-Pacific dispersal events. White numbers in black circles cross reference to Fig. 5. Alaska might have maintained suitable habitats for *Z. marina* along the Beringia Bridge during the glacial periods, but our results

indicate local extinction. The current Alaska was colonized very recently by Japan (62 kya, see Fig. 5). **f**, Schematic illustration. Solid vertical arrows indicate dispersal, horizontal black lines colonization and persistence, and gaps indicate local extinctions. Med=Mediterranean. The ancient San Diego population (SD) apparently became restricted to the south (Fig. 5). SD later dispersed northwards (presumably via the Davidson Current), forming sequential admixtures with BB and WAS. Alternatively, this ancient population may have survived along the northern California coast where it mixed with new incoming dispersals from the "gateway" involving WAS and BB admixtures. This model of multiple trans-Pacific dispersal events is consistent with the phylogenetic tree (Fig. 5).

Considering the pattern of ocean currents, *Z. marina* in Japan (JS) has to first cross the Pacific Ocean to North America before the spread to other locations (Fig. 6a). Our data provide evidence for three trans-Pacific dispersal events from the origin (Fig. 6b-f). The first trans-Pacific dispersal event happened at 352.5 kya (95% HPD: 422.6-290.2 kya), and has possibly founded the San Diego population (SD). Furthermore, San Diego was free from the influence of the subsequent trans-Pacific dispersal events.

A second trans-Pacific dispersal event at 222 kya (95% HPD: 266.6-181.9 kya) led to the colonization of the Atlantic side through the Arctic Ocean, a surprisingly recent estimate given that the possibility for the colonization of the Atlantic side opened as early as 3.5 mya ago (Olsen et al., 2004). This dispersal event did not leave genomic traces in Alaska even although it was on the pathway to the Arctic Ocean, because the current Alaskan population (ASL) was derived from a subsequent additional trans-Pacific dispersal event from Japan that happened later (62 kya, 95% HPD: 50.5-74.9 kya). This indicates local extinction of *Z. marina* in Alaska, which is also supported by the genetic closeness between Japan and the Atlantic side. If there was no local extinction in Alaska, it would be the population that shows the highest genetic closeness to the Atlantic. Yet, JN showed the smallest F_{ST} with all populations from the Atlantic side compared with the other Pacific populations (Supplementary Table 4). Moreover, JN was also the only Pacific population that displayed the shared genetic component with the Atlantic side (Fig. 3d). The local extinction in Alaska might have happened during glacial period MIS 6 (Fig. 5a).

After the initial colonization (222 kya; 95% HPD: 266.6-181.9 kya), the next divergence in the Atlantic side was observed at 39.8 kya (95% HPD: 47.8 kya - 32.0 kya) when the Atlantic Ocean and the Mediterranean Sea evolved into separate clades. There were two glacial periods (MIS 6 and MIS 4) from 222 kya to 39.8 kya (Fig. 5a), and the Atlantic side might have been severely influenced. This is consistent with the much lower levels of genetic diversity (Fig. 2) and genetic differentiation among locations (Fig. 3a) compared with the Pacific side. The Northwest (NC) and Northeast Atlantic (SW) diverged very recently at 15.8 kya (95% HPD: 19.7-12.3 kya), and shared the same common ancestor during the LGM (Fig. 5a), indicating that they were derived from the same glacial refugium. The glacial refugium was likely located in the southern region in the Northwest Atlantic, because northern Carolina (NC) showed the highest genetic diversity among the populations in the Atlantic Ocean and the Mediterranean Sea (Fig. 2).

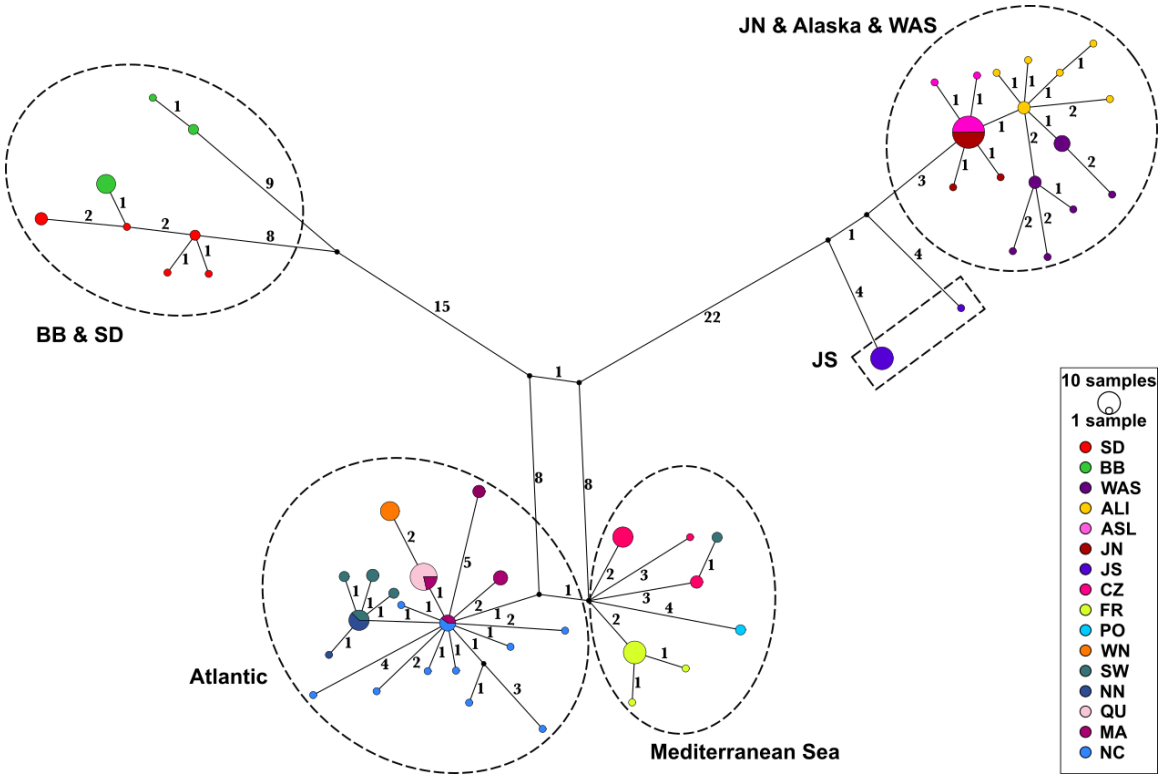


Fig. 7 | Chloroplast haplotype network based on the fixed variants within individuals in chloroplast genome. The numbers represent mutation steps. Colors correspond to the population. Split-colored circles indicate that a particular haplotype is shared between populations and color size is proportional to frequency. Haplotypes conform to three clusters, consistent with three trans-Pacific dispersal events (Fig. 5 & 6). Haplotypes from BB and SD are on average 53 mutation steps different from those of the other Pacific populations, confirming an early trans-Pacific dispersal event (Fig. 5 & 6). Alaska (ALI + ASL) displays chloroplast haplotypes that are similar to those of northern Japan (JN), in line with the third trans-Pacific dispersal that happened more recently (Fig. 5 & 6). For the Atlantic side, two clusters differ by 2 mutational steps, corresponding to the Atlantic Ocean and the Mediterranean Sea. In the Atlantic, most haplotypes are descendants of an abundant/dominant haplotype shared by NC and MA, indicating that the founding population of the Atlantic were most likely located in the West Atlantic.

Chloroplast haplotype network

Chloroplast genome is independent from the nuclear genome, and thus provides additional information on the evolutionary history. A chloroplast haplotype network was constructed based on full genome data (Fig. 7). The chloroplast haplotypes were divided into three diverged lineages, which was in line with the three trans-Pacific dispersal events (Fig. 6). Inherited as a haplotype, chloroplast genome is easily affected by genetic drift. Washington (WAS) and Bodega Bay (BB) showed very different chloroplast haplotypes. Washington (WAS) displayed chloroplast haplotypes that were similar to those of Alaska (ALI & ASL) and northern Japan (JN), while Bodega Bay (BB) displayed chloroplast haplotypes that were similar to those of San Diego (SD). According to STRUCTURE result based on nuclear

genome (Fig. 3e), Washington (WAS) and Bodega Bay (BB) were admixture of the same two genetic components. However, Washington (WAS) consisted of more component of Alaska (represented by the blue color), while Bodega Bay (BB) consisted of more component of SD (represented by the purple color). Our results show that populations with input from two source populations will eventually reach a fixed chloroplast state, showing chloroplast haplotypes similar to the source population with larger proportion.

As a relatively long DNA sequence without recombination (143,968 bp), chloroplast genome contains information in terms of mutational steps that should be proportional to the time elapsed. ASL (Alaska) and JN (northern Japan) shared the same dominant chloroplast haplotype, indicating that there was refugium in Alaska during the LGM. The contemporary *Z. marina* meadows along the coast of Alaska is likely to be the largest continuous seagrass meadows in northern North America (Talbot et al., 2016), indicating that the meadows have persisted for a long time, in line with other data from the area on brown algae (Grant & Chenoweth, 2021). Chloroplast haplotypes in BB and SD were on average 53 mutation steps different from those of the other Pacific populations, indicating an early trans-Pacific dispersal event, in line with the nuclear data (Fig. 5a). Alaska (ALI + ASL) displayed chloroplast haplotypes that were similar to those of Japan, consistent with the third trans-Pacific dispersal that happened recently (Fig. 5a).

The level of mutational steps among the chloroplast haplotypes in the Atlantic side (Atlantic Ocean + Mediterranean Sea) was low, with only two mutational steps separating the Atlantic Ocean and the Mediterranean Sea. This is consistent with the recent divergence of the Atlantic Ocean and the Mediterranean Sea at 39.8 kya (95% HPD: 47.8 kya - 32.0 kya) estimated based on nuclear genome (Fig. 5a). Within the Atlantic Ocean, the haplotypes were mostly descendants of a common haplotype shared by NC and MA, indicating that founding population of the Atlantic populations were possibly located in an area ranging from NC to MA, in line with the high genetic diversity in NC and MA based on nuclear data (Fig. 2).

Demographic history

In order to reconstruct past changes in effective population size as evidence for glaciation driven bottlenecks, the Multiple Sequentially Markovian Coalescent (MSMC) was used to infer demographic history, taking different unique genotypes within the same population as independent replicates (Fig. 8). We only focused on relatively deeper time intervals where different replicate runs per population converged, because MSMC creates unreliable estimates in recent time (Schiffels & Wang, 2020). Owing to marked differences in the degree of clonality and the relative amount of sexual vs. clonal reproduction (Olsen et al., 2004), the generation time of *Z. marina* varies across populations which prevented us to represent the x-axis in absolute time. We focused on the general pattern of the demographic plots, and tried to make inferences combining those to our other results.

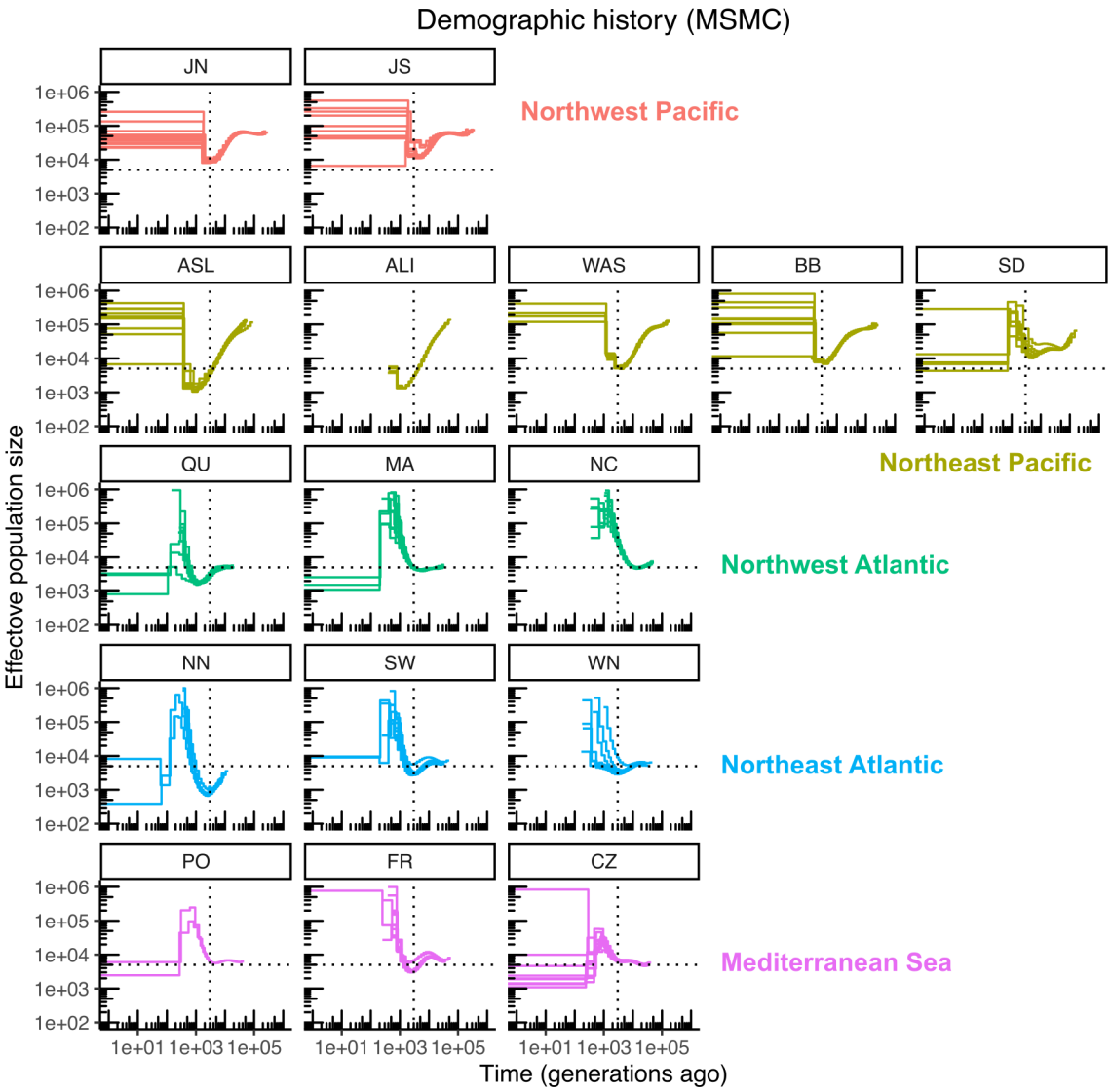


Fig. 8 | Demographic history inferred by the multiple sequentially Markovian coalescent (MSMC). Within the same population, each curve indicates the MSMC result based on one genet. Time is given in generations noting that generation time for *Zostera marina* is difficult to generalize due to the range of life histories, from annuals to decadal to centennial long-lived clones. All three populations in the Northeast Atlantic and one population in the Mediterranean Sea (FR) show a bottleneck at around 3,000 generations ago (dotted vertical line), probably indicating the Last Glacial Maximum. Effective population sizes are relative only. MSMC loses accuracy in recent time. When effective population size exceeded 1 million, which is unlikely in real world, the data point itself and the data points more recent than that were removed. For this analysis, the shape of the curves rather than actual values is the most important. With respect to the LGM, the Atlantic side was more affected than Pacific. Southern populations fared better in both oceans, and west coast of oceans fared better.

Within the Pacific Ocean, the southmost population, the San Diego population (SD), showed stable effective population size, while the other populations showed bottlenecks to different degrees. SD diverged into a unique clade very early (Fig. 5a). As the southernmost population in the Pacific Ocean, it was likely not affected by the historical glacial periods, which was in line with a stable effective population size detected here. Reassuringly, geographically closed populations showed similar demographic histories (JN & JS, ALI & ASL, WAS & BB), apparently providing natural replicate estimates. The two Alaskan populations (ALI + ASL) went through a severe genetic bottleneck, followed by a recent population expansion, consistent with the relatively low genetic diversity (Fig. 2). In the Northeast Pacific, the bottlenecks became deeper from south to north (SD -> BB -> WAS -> ALI/ASL). As for the Atlantic side, refugia during the LGM did not show bottlenecks (southern refugium in the Northwest Atlantic: NC & MA; Mediterranean refugium: PO & CZ), while northern populations did show bottlenecks (northern region in the Northwest Atlantic: QU; northern region in the Northeast Atlantic: NN & SW).

Post-LGM recolonization modeling

Finally, we used approximate Bayesian computations (ABC-modeling) to distinguish between competing alternative models of the recolonization history of *Z. marina* after the LGM. Based on our other results, the Atlantic side has been severely influenced by the LGM, and thus is the focus here. Considering that the Mediterranean Sea had its own glacial refugium, the ABC-modeling was conducted for only the Atlantic Ocean.

We constructed three plausible recolonization scenarios (Fig. 9). In the first, NC and MA were the only glacial refugium in the Atlantic, which recolonized QU and the Northeast Atlantic sequentially. In the second scenario, NC and MA were still the only glacial refugium in the Atlantic, while both QU and all Northeast Atlantic were directly recolonized by the glacial refugium. In the last scenario, NC and MA were the glacial refugium in the Northwest Atlantic. There was also refugium in the Northeast Atlantic, which was not sampled in our study. QU and Northeast Atlantic were recolonized by the refugia in the Northwest Atlantic and the Northeast Atlantic, respectively. According to model selection algorithm implemented in DIYABC-RF, the first scenario was the most likely one and provided support for a refugium in the Northwest Atlantic (represented by NC and MA) first colonizing QU as a stepping stone, while *Z. marina* then dispersed into the Northeast Atlantic.

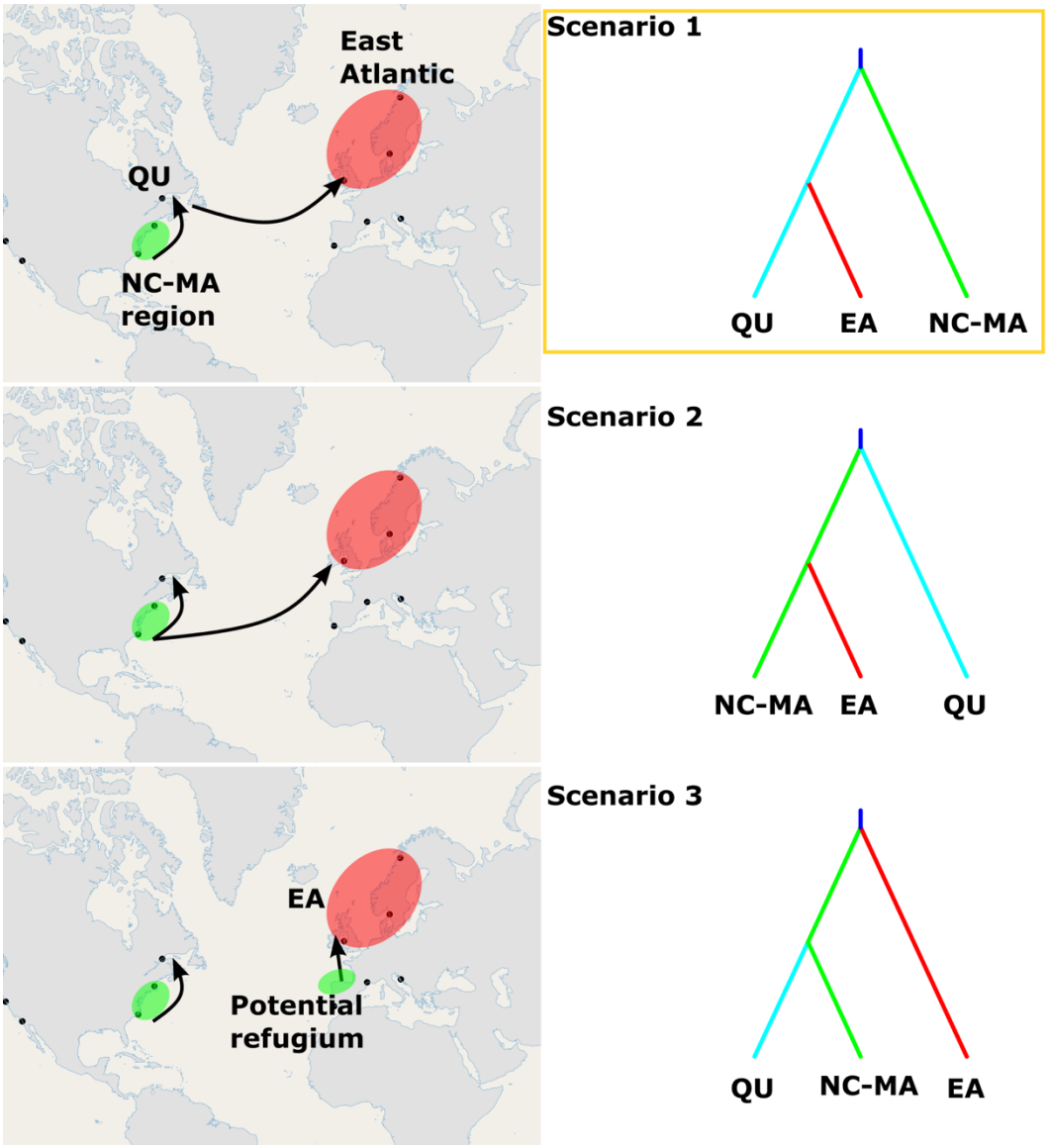


Fig. 9 | Recolonization in the Atlantic Ocean after the Last Glacial Maximum (LGM) using approximate Bayesian computation. Genetic diversity (Fig. 2), phylogenetic tree (Fig. 5) and chloroplast haplotype network (Fig. 7) show that NC-MA region was one of the glacial refugia during the LGM, and Mediterranean had its own refuge. Southern Brittany, France has been proposed as a refuge for rocky intertidal/subtidal species (Jenkins et al., 2018), but a site in Brittany was not included in the present study. Approximate Bayesian computation was used to test the possible recolonization pathways in the Atlantic Ocean after the LGM. **Scenario 1**: NC-MA region recolonized QU and the Northeast Atlantic sequentially. **Scenario 2**: NC-MA region recolonized QU and the Northeast Atlantic separately. **Scenario 3**: NC-MA region recolonized QU, and the potential refugium in southern Brittany recolonized the Northwest Atlantic. **Scenario 1** is the most likely scenario.

Discussion

The seagrass *Zostera marina* is one of the rare plant species that is distributed throughout the entire northern hemisphere. It is, to the best of our knowledge, the only flowering plant that

has been studied thus far with a natural circumglobal distribution. Based on genome-wide genetic diversity, we confirm the origin of *Z. marina* to be Japan (Fig. 2), along with the earlier phylogenetic data on the radiation of the genus *Zostera* (Coyer et al., 2013). Using full-genome data, for the first time we are able to date several large-scale dispersal events via a time-calibrated, coalescence-based phylogeny. Considering the pattern of ocean currents, *Z. marina* in Japan has to first cross the Pacific Ocean to North America before the spread to other locations (Fig. 6a). There may have been three trans-Pacific dispersal events from the origin of the species (Fig. 6). San Diego (SD) was colonized by the first dispersal event, and was free from the influence of the subsequent dispersal events, probably owing to the Point Conception biogeographic boundary located between Bodega Bay and San Diego (Newman, 1976). Point Conception features a transition between cold northern and warm southern water masses, and is presented as a strong biogeographic break for many species (Doyle, 1985). The second dispersal led to the colonization of the Atlantic side through the Arctic Ocean. We found evidence on local extinction in Alaska after the second dispersal event, which erased the influence of the first and the second dispersal events. This probably happened during the glacial period MIS 6 (Fig. 5a). The Alaska was recolonized again during the subsequent dispersal event at 62 kya. The multiple trans-Pacific events formed admixture zone in Washington (WAS) and Bodega Bay (BB).

After the initial colonization at 222 kya, the Atlantic side was likely severely influenced by two glacial periods, resulting in much lower levels of genetic differentiation among locations (Fig. 3a) and genetic diversity (Fig. 2) compared with the Pacific side. During the Last Glacial Maximum (LGM), the Atlantic side had two separate refugia. One refugium located near northern Carolina (NC) served as source population for the entire Northwest and Northeast Atlantic, while the Mediterranean Sea had its own refugium throughout the LGM. In summary, the current *Z. marina* populations were derived from multiple trans-oceanic dispersal events, and many of the dispersal events were much more recent than anticipated. The recolonization of the Northeast Atlantic from a Northwest Atlantic source (after the LGM) highlights the propensity of *Z. marina* to swiftly cross the entire ocean basin on a time scale of 10,000s of years.

In particular for the Pacific side, our data show that among populations of a marine angiosperm, depicting the evolutionary history as a network may be a more appropriate representation than a bifurcating tree. Inherited as a haplotype, chloroplast genome is easily affected by genetic drift. For populations with input from two source populations, eventually the chloroplast genome will reach a fixed state, showing chloroplast haplotypes similar to the source population with larger proportion. This could be a reason for the frequently reported cytonuclear discordance (Morales-Briones et al., 2018; Zhang et al., 2019; Xu et al., 2021). Chloroplast genome may be misleading without nuclear data in population-level phylogenies, but combining nuclear and chloroplast genomes can provide better understanding of the evolutionary history, because chloroplast genome provides independent confirmation of the results based on nuclear genome.

The population is the basic unit of evolution, and population divergence is the precursor of any speciation events. Our finding will enhance the interpretation of the speciation of seagrasses. It is certainly no coincidence that the polyphyletic group of marine angiosperms (or seagrasses) only comprise 65 or so species (Larkum et al., 2018), although the recolonization of the marine environment happened at >70 mya (Larkum et al., 2018). Our population genomic data on the widespread model seagrass species *Z. marina* reveal frequent and swift colonization waves across entire ocean basins since 353.5 kya, with the most recent crossing from the Northwest to Northeast Atlantic only at 15.8 kya. This may explain a lack of speciation events across the seagrasses in general. As phylogeographic patterns and neutral genetic population structure are always the background for any assessments of selective processes, our data provide a baseline for further studies on adaptation in this important marine macrophyte.

Materials and Methods

Sampling

Zostera marina meadows consist of modular units with ability to live independently (= ramets). A genet/clone is the product of a single zygote, which is the collection of all the ramets derived from a single sexual event. We performed a range-wide sample collection of 190 *Z. marina* ramets from 16 geographic populations (Fig. 1; Supplementary Table 1). These sites represented a subset of the 50 Zostera Experimental Network (ZEN) sites (http://zenscience.org). Inter-ramet distance was maintained of minimally 2 m to reduce the possibility of collecting the same genet twice which was not always successful (Supplementary Table 3). Plant tissue (several 3-5 cm pieces) was selected from the basal meristematic part of the shoot within the leaf sheath to minimize epiphytes (bacteria and diatoms), frozen in liquid nitrogen and stored at -80 °C until DNA extraction.

DNA extraction and whole-genome resequencing

Genomic DNA was extracted using the Macherey-Nagel method (NucleoSpin plant II) following the manufacturer's instructions. Liquid nitrogen was used to keep low temperature while grinding the plant tissue. The method worked well starting from 100-200 mg fresh weight of basal leaf tissue, containing the meristematic region. DNA concentrations were in the range of 50-200 ng/uL. DNA QC was performed following JGI guidelines (https://jgi.doe.gov/wp-content/uploads/2013/11/Genomic-DNA-Sample-QC.pdf). Following genomic DNA QC, libraries were constructed from ~500 bp fragments. Whole-genome resequencing was performed by the US DOE-Joint Genome Institute (Berkeley, CA) on the Illumina HiSeq2000 or HiSeq2500 1T platform, aiming to produce 2x150 bp reads to a minimum of 40x coverage.

Quality check and filtering of raw Illumina reads

The quality of the raw Illumina reads was assessed by FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The FastQC output for all ramets were visualized by MultiQC (Ewels et al. 2016). BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/) was used to remove adapter contamination

and conduct a basic quality filtering: (1) reads with more than one Ns were discarded (maxns=1); (2) reads shorter than 50 bp after trimming were discarded (minlength=50); (3) reads with average quality below 10 after trimming were discarded (maq=10). FastQC and MultiQC were used for second round of quality check for the clean reads. Sequencing coverage was calculated for each ramet (Supplementary Data 1).

Identifying core genes that are shared by most of the collected ramets

Assuming different individuals within the same species have very similar genomes is inappropriate for a plant species with circumglobal distribution range. Genes are supposed to be more conservative, and are more likely to be shared by all individuals. Evidently, for our global SNP analysis, we were looking for genes present in all populations, thus core genes. Each of the 190 ramets were de novo assembled using HipMer (k=51) (Georganas et al., 2015). To categorize, extract, and compare core and variable (shell and cloud) genes, primary transcript sequences (n=21,483) from the *Z. marina* reference (V3.1) were aligned using BLAT (default parameters) (Kent, 2002) to each de novo assembly. Genes were considered present if the transcript aligned with either: 1) > 60% identity and > 60% coverage from a single alignment, or 2) > 85% identity and > 85% coverage split across 3 or fewer scaffolds. Individual PAV calls were combined into a matrix to classify genes into core, cloud, and shell categories based on their observation across the population. The total number of genes considered was 20,100. Because identical genotypes and fragmented, low-quality assemblies can bias and skew PAV analyses, to ensure only unique and high-quality assemblies were retained, only single representatives of clones and ramets with greater than 17,500 genes were kept (n = 141). Genes were classified using discriminant analysis of principal components (DAPC) (Jombart, 2008) into cloud, shell, and core gene clusters based on their observation frequency. Core genes were the largest category, with 18,717 genes that were on average observed in 97% of ramets.

SNP calling

Even although we are aware that there is substantial structural polymorphism among the 16 sampled populations, we here focus on single nucleotide polymorphisms to reconstruct recolonization history. The quality-filtered Illumina reads were mapped against the chromosome-level *Z. marina* reference genome V3.1 (Ma et al., 2021) using BWA MEM (Li & Durbin, 2009). The alignments were converted to BAM format and then sorted using Samtools (Li et al., 2009). The MarkDuplicates module in GATK4 (Van der Auwera & O'Connor, 2020) was used to identify and tag duplicate reads in the BAM files. Mapping rate for each ramet was calculated using Samtools (Supplementary Data 2). HaplotypeCaller (GATK4) was used to generate GVCF format file for each ramet, and all the GVCF files were combined by CombineGVCFs (GATK4). GenotypeGVCFs (GATK4) was used to call genetic variants.

SNP filtering

BCFtools (Li, 2011) was used to remove SNPs within 20 base pairs of an indel or other variant type (Supplementary Fig. 1). Then we extracted only SNPs (9,562,044 SNPs) and excluded the other variant types. VariantsToTable (GATK4) was used to extract INFO

annotations. SNPs meeting one or more than one of the following criteria were marked by VariantFiltration (GATK4): MQ < 40.0; FS > 60.0; QD < 10.0; MQRandSum > 2.5 or MQRandSum < -2.5; ReadPosRandSum < -2.5; ReadPosRandSum > 2.5; SOR > 3.0; DP > 10804.0 (2 * average DP). Those SNPs were excluded by SelectVariants (GATK4). A total of 3,975,407 SNPs were retained. VCFtools (Danecek et al., 2011) was used to convert individual genotypes to missing data when GQ < 30 or DP < 10. Individual homozygous reference calls with one or more than one reads supporting the variant allele, and individual homozygous variant calls with one or more than one reads supporting the reference allele, were also converted to missing data using a custom Python3 script. Only bi-allelic SNPs were kept (3,892,668 SNPs). The Pacific ramets and the Atlantic ramets were diverged to a large extent. To avoid the reference-related biases, we focused on the 18,717 core genes that were on average observed in 97% of ramets. Bedtools (Quinlan & Hall, 2010) was used to find overlap between the SNPs and the core genes, and only those SNPs were kept (ZM_HQ_SNPs, 763,580 SNPs).

Excluding duplicate genotypes and genotypes originating from selfing

Based on ZM_HQ_SNPs (763,580 SNPs), possible parent-descendant pairs under selfing (Supplementary Table 2) and clonemates (Supplementary Table 3) were detected based on "identity by heterozygosity" (Supplementary Fig. 3; Yu et al. 2022). For a parent-descendant pair under selfing, the descendant cannot be considered being formed by random mating. Accordingly, ten ramets possibly produced by selfing of the other collected genets were excluded. For each genet, one ramet was randomly selected and kept as representative, while the other ramets of the same genet were excluded (17 ramets).

Excluding ramets with high missing rate

Based on ZM_HQ_SNPs (763,580 SNPs), we calculated the missing rate for each ramet using a custom Python3 script. A histogram was plotted (Supplementary Fig. 4). Most of the ramets had missing rate < 15%, except 10 ramets (ALI01, ALI02, ALI03, ALI04, ALI05, ALI06, ALI10, ALI16, QU03, and SD08). Those 10 ramets were also excluded.

Demographic analysis

We first generated one mappability mask file for each of the six main chromosomes using SNPable (http://lh3lh3.users.sourceforge.net/snpable.shtml). Each file contained all regions on the chromosome, on which short sequencing reads can be uniquely mapped. We then generated one mask file for the core genes for each of the six main chromosomes.

The Multiple Sequentially Markovian Coalescent (Schiffels & Durbin, 2014) was run for each genotype per population. We generated one ramet-specific mask file for each of the six main chromosomes based on the bam format file using bamCaller.py (https://github.com/stschiff/msmc-tools). Each mask file contained the regions on the chromosome on which the genome of that ramet was covered sufficiently. The minDepth variable in bamCaller.py was set to 10, because in our SNP filtering step, we only kept genotype calls with DP >= 10. We also generated a ramet-specific vcf file for each of the six main chromosomes based on ZM_HQ_SNPs using a custom Python3 script.

Putatively neutral and non-linked SNPs (ZM_Core_SNPs, 11,705 SNPs)
Thirty-seven ramets were excluded from ZM_HQ_SNPs, including 10 ramets related to geitonogamous (within-clone) selfing, 17 ramets representing replicated genotypes, and 10 ramets with high missing rate (Supplementary Table 2 & 3, Supplementary Fig. 4). A total of 153 unique ramets were retained for analyses. After removing these ramets, some genomic sites became monomorphic, which were also excluded. SnpEff (http://pcingola.github.io/SnpEff/) was used to annotate each SNP. To obtain putatively neutral SNPs, we only kept SNPs annotated as "synonymous_variant" (ZM_Neutral_SNPs, 144,773 SNPs). For the SNPs in ZM_Neutral_SNPs (144,773 SNPs), only SNPs without any missing data were kept. To obtain putatively non-linked SNPs, we thinned sites using Vcftools to achieve a minimum pairwise distance (physical distance in the reference genome) of 3,000 bp (ZM_Core_SNPs, 11,705 SNPs).

Global genetic population structure
Principal component analysis (PCA) was used to study population structure. We used R packages to run PCA based on ZM_Core_SNPs (11,705 SNPs). Package vcfR (Knaus & Grunwald, 2017) was used to load the VCF format file, and function glPca in adegenet package was used to conduct PCA analysis. The visualization of the first three PCs was done by ggplot2 package (Wickham, 2016).

As an alternative method, STRUCTURE analysis was also used to study population structure (Pritchard et al., 2000). An individual may consist of multiple genetic components, which indicates admixture. STRUCTURE is time-consuming when large number of SNPs are used. To reduce running time, we randomly selected 2,353 SNPs from ZM_Core_SNPs to run STRUCTURE (Length of burn-in period, 300,000; Number of MCMC reps after burn-in, 2,000,000). K was set from 1 to 10. For each K, we repeated 10 times independently. StructureSelector (Li & Liu, 2018) was used to decide the optimal K based on Delta K method (Evanno et al., 2005), and to combine and visualize the STRUCTURE results of 10 independent runs for each K. We were aware that STRUCTURE fails to detect nested population structure, hence global run was complemented with analyses of populations from the Atlantic side and the Pacific side, respectively.

Split network
To check the reticulate evolutionary processes, we used SplitsTree4 to construct a split network, which is a combinatorial generalization of phylogenetic trees and is designed to represent incompatibilities (Huson & Bryant, 2006). A custom Python3 script was used to generate a fasta format file containing concatenated DNA sequences for all ramets based on ZM_Core_SNPs (11,705 SNPs). For a heterozygous genotype, one allele was randomly selected to represent the site. The fasta format file was converted to nexus format file using MEGAX (Kumar et al., 2018), which was fed to SplitsTree4. NeighborNet method was used to construct the split network.

Genetic diversity

Vcftools was used to calculate nucleotide diversity (π) for each population based on the first 29,371,071 bp (the minimum length for the six chromosomes) for each of the six chromosomes based on ZM_Core_SNPs (11,705 SNPs). Genomic heterozygosity for a given ramet = (number of heterozygous sites) / (total number of sites with available genotype calls). This was calculated by a custom Python3 script based on ZM_Core_SNPs (11,705 SNPs).

Pairwise Fst

We used R packages to calculate pairwise Fst based on ZM_Core_SNPs (11,705 SNPs). Package vcfR was used to load the VCF format file, and function stamppFst in StAMPP package (Pembleton et al., 2013) was used to do the calculation (Supplementary Table 4). P-values were generated by 1,000 bootstraps across loci.

Genetic population structure for the Pacific side

The ramets from the Pacific side were extracted from ZM_Neutral_SNPs (144,773 SNPs). Monomorphic sites were excluded. Then we only kept SNPs without any missing data. To obtain putatively independent SNPs, we thinned sites using Vcftools, so that no two sites are within the 3,000 bp distance (physical distance in the reference genome) from one another (ZM_Pacific_SNPs, 12,514 SNPs). Those 12,514 SNPs were used to run PCA. We then randomly selected 6,168 SNPs to reduce run times to run STRUCTURE (Length of burn-in period, 300,000; Number of MCMC reps after burn-in, 2,000,000). Possible K-values were set from 1 to 7. For each K, we repeated 10 times independently. StructureSelector was used to decide the optimal K based on Delta K method, and to combine and visualize the STRUCTURE results of 10 independent runs for each K.

Genetic population structure for the Atlantic side (Atlantic + Mediterranean Sea)

The ramets from the Atlantic and the Mediterranean Sea were extracted from ZM_Neutral_SNPs (144,773 SNPs). Monomorphic sites were excluded. Then we only kept SNPs without any missing data. To obtain putatively independent SNPs, we thinned sites using Vcftools, so that no two sites are within the 3,000 bp distance (physical distance in the reference genome) from one another (8,552 SNPs).

Those 8,552 SNPs were then used to run another separate PCA and STRUCTURE (Length of burn-in period, 300,000; Number of MCMC reps after burn-in, 2,000,000). For STRUCTURE analysis, K was set from 1 to 5. For each K, we repeated 10 times independently. StructureSelector was used to decide the optimal K based on Delta K method, and to combine and visualize the STRUCTURE results of 10 independent runs for each K.

D (ABBA-BABA) for each trio within the Pacific side and the Atlantic side, respectively

Patterson's D, also known as the ABBA-BABA statistic, provides a simple and powerful test for the deviation from a strict bifurcating evolutionary history. We used Dsuite (Malinsky et al., 2021) to calculate D for populations within the Pacific side and the Atlantic side, respectively. D was calculated for trios of *Z. marina* populations based on the dataset ZMZJ_D_SNPs (Supplementary Note), using the *Z. japonica* ramet as outgroup. Ruby script plot_d.rb

(https://github.com/mmatschiner/tutorials/blob/master/analysis_of_introgression_with_snp_data/src/plot_d.rb) was used to plot a heatmap. Assume there are three populations P1, P2, and P3, and P1 and P2 are sister populations. If P3 shares more derived alleles with P2 than with P1, Patterson' D will be positive, and a p value can be calculated by statistical test. For each trio, the three populations can be arranged in positions of P1, P2, and P3, to achieve a positive D value. A bi-color heatmap visualizes both D and the associated p value for each cell of the main diagram that depicts P2 and P3 on the horizontal and vertical axes, respectively. The color of the corresponding heatmap cell indicates the most significant D value across all possible populations in position P1. Red colors indicate higher D values, and more saturated colors indicate greater significance.

Phylogenetic tree with estimated divergence time

To estimate the divergence time among major groups, we used the method developed by Stange et al. (2018) to construct a phylogenetic tree with estimated divergence time. An accurate time estimate requires that the populations are not admixed, and thus some populations were removed due to admixture based on STRUCTURE (Fig. 3), splitstree and D statistics (Fig. 4). For the Pacific side, WAS, BB, and ALI were excluded, while for the Atlantic side, NC, SW, and CZ were selected to be representatives for the Northwest Atlantic, Northeast Atlantic, and the Mediterranean Sea, respectively. In the end, the analysis was based on four populations from the Pacific side (JS, JN, ASL, and SD) and three populations from the Atlantic side (NC, CZ, and SW). There was no strong admixture among the selected populations, which should make the estimation of the divergence time more accurate. We used the software SNAPP (Bryant et al., 2012) to run the analysis, and the input file was prepared by a Ruby script "snapp_prep.rb" (https://github.com/mmatschiner/snapp_prep).

Two ramets were randomly selected from each of the 7 populations. The 14 *Z. marina* ramets plus the one *Z. japonica* ramet were extracted from ZMZJ_Neutral_SNPs (Supplementary Note). Monomorphic sites were excluded. Then we only kept SNPs without any missing data. To obtain putatively independent SNPs, we thinned sites using Vcftools, so that no two sites are within the 3,000 bp distance (physical distance in the reference genome) from one another (6,169 SNPs). A Ruby script was used to prepare the input file for SNAPP (https://github.com/mmatschiner/snapp_prep). The divergence time between *Z. japonica* and *Z. marina* was used as calibration point, which was represented by a lognormal distribution (Supplementary Note, mean = 11.1542 MYA, SD = 0.07).

A large proportion of the 6,169 SNPs above represented the genetic differences between *Z. japonica* and *Z. marina*, and were monomorphic among the *Z. marina* ramets, and thus the relationships within *Z. marina* were compressed. To obtain a better estimation among *Z. marina* populations, we performed a separate SNAPP analysis for the target species *Z. marina* only. The same 14 *Z. marina* ramets were extracted from ZM_Neutral_SNPs (144,773 SNPs). Monomorphic sites were excluded. Then we only kept SNPs without any missing data. To obtain putatively independent SNPs, we thinned sites using Vcftools, so that no two sites are within the 3,000 bp distance (physical distance in the reference genome)

from one another (13,732 SNPs). A Ruby script was used to prepare the input file for SNAPP (https://github.com/mmatschiner/snapp_prep). The crown divergence for all *Z. marina* populations was used as calibration point. This was extracted from the first SNAPP analysis, and was represented by a lognormal distribution (mean = 0.3564 MYA, SD = 0.1). The same analysis was repeated three times with different sample sets, giving consistent results.

Recolonization models after the LGM for the Atlantic

We used approximate Bayesian computations (ABC-modeling) to distinguish between competing alternative models of the recolonization history of *Z. marina* after the LGM. Based on our other results, the Atlantic side has been severely influenced by the LGM, and thus is the focus here. Considering that the Mediterranean Sea had its own glacial refugium, the ABC-modeling was conducted for only the Atlantic Ocean. We constructed three recolonization scenarios (Fig. 9): (i) NC and MA represent the only glacial refugia in the Atlantic. The glacial refugia recolonized QU, and then QU recolonized Northeast Atlantic. (ii) NC and MA represent the only glacial refugia in the Atlantic. Both QU and Northeast Atlantic were directly recolonized by the glacial refugia. (iii) NC and MA represent the southern glacial refugia in the Northwest Atlantic. There were also refugia in the Northeast Atlantic, which was not sampled in our study. QU was recolonized by the refugia in the Northwest Atlantic, and Northeast Atlantic was recolonized by the refugia in the Northeast Atlantic. DIYABC-RF (Collin et al., 2021) was used to run simulations under each scenario, and to choose the most likely scenario.

Chloroplast haplotypes

Chloroplast genome was de novo assembled by NOVOPlasty (Dierckxsens et al., 2017). Chloroplast DNA is represented by a circular molecule of 143,968 bp with a classic quadripartite structure: two identical inverted repeats (IRa and IRb) of 24,127 bp each, large single-copy region (LSC) of 83,312 bp, and small single-copy region (SSC) of 12,402 bp. All regions were equally taken into SNP calling analysis except for 9,818 bp encoding 23S and 16S RNAs due to supposed bacteria contamination in some samples.

The raw Illumina reads of each individual were aligned by BWA MEM (Li & Durbin, 2009) to the assembled chloroplast genome. The alignments were converted to BAM format and then sorted using Samtools (Li et al., 2009). Genomic sites were called as variable positions when frequency of variant reads >50% (Supplementary Fig. 8) and the total coverage of the position >30% of the median coverage (174 variable positions). Then 11 positions likely related to microsatellites and 12 positions reflecting minute inversions caused by hairpin structures (Petit & Vendramin, 2007) were removed from the final set of variable positions for the haplotype reconstruction (151 SNPs). Chloroplast haplotype network was constructed by Integer Neighbour-Joining Network method with the reticulation tolerance of 0.1 implemented by PopART (Leigh & Bryant, 2015).

References

Altmann, A., Weber, P., Bader, D., Preuss, M., Binder, E. B., & Muller-Myhsok, B. (2012). A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum. Genet., 131*(10), 1541-1554. doi:10.1007/s00439-012-1213-z

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol., 29*(8), 1917-1932. doi:10.1093/molbev/mss086

Collin, F. D., Durif, G., Raynal, L., Lombaert, E., Gautier, M., Vitalis, R., . . . Estoup, A. (2021). Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. *Mol. Ecol. Resour., 21*(8), 2598-2613. doi:10.1111/1755-0998.13413

Coyer, J. A., Hoarau, G., Kuo, J., Tronholm, A., Veldsink, J., & Olsen, J. L. (2013). Phylogeny and temporal divergence of the seagrass family Zosteraceae using one nuclear and three chloroplast loci. *System. Biodivers., 11*(3), 271-284. doi:10.1080/14772000.2013.821187

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res., 45*(4), e18. doi:10.1093/nar/gkw955

Doyle, R. F. (1985). *Biogeographical Studies of Rocky Shores Near Point Conception, California.* (Doctoral dissertation), University of California, Santa Barbara.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol., 14*(8), 2611-2620. doi:10.1111/j.1365-294X.2005.02553.x

Fang, B., Merilä, J., Matschiner, M., & Momigliano, P. (2020). Estimating uncertainty in divergence times among three-spined stickleback clades using the multispecies coalescent. *Mol. Phylogenet. Evol., 142*, 106646. doi:10.1016/j.ympev.2019.106646

Georganas, E., Buluç, A., Chapman, J., Hofmeyr, S., Aluru, C., Egan, R., . . . Yelick, K. (2015). *HipMer: an extreme-scale de novo genome assembler.* Paper presented at the SC'15: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis.

Grant, W. S., & Chenoweth, E. (2021). Phylogeography of sugar kelp: Northern ice-age refugia in the Gulf of Alaska. *Ecol. Evol., 11*(9), 4670-4687. doi:10.1002/ece3.7368

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Nature, 405*(6789), 907-913. doi:10.1038/35016000

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol., 23*(2), 254-267. doi:10.1093/molbev/msj030

Janes, J. K., Miller, J. M., Dupuis, J. R., Malenfant, R. M., Gorrell, J. C., Cullingham, C. I., & Andrew, R. L. (2017). The $K = 2$ conundrum. *Mol. Ecol., 26*(14), 3594-3602. doi:10.1111/mec.14187

Jenkins, T. L., Castilho, R. & Stevens, J. R. (2018). Meta-analysis of northeast Atlantic marine taxa shows contrasting phylogeographic patterns following post-LGM expansions. *PeerJ*, 6. doi:10.7717/peerj.5684

Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics, 24*(11), 1403-1405. doi:10.1093/bioinformatics/btn129

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res., 12*(4), 656-664. doi:10.1101/gr.229202

Knaus, B. J., & Grunwald, N. J. (2017). Vcfr: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour., 17*(1), 44-53. doi:10.1111/1755-0998.12549

Knoop, V. (2004). The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Curr. Genet., 46*(3), 123-139. doi:10.1007/s00294-004-0522-8

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol., 35*(6), 1547-1549. doi:10.1093/molbev/msy096

Larkum, A. W. D., Kendrick, G. A., & Ralph, P. J. (2018). *Seagrasses of Australia: Structure, Ecology and Conservation*: Springer.

Larkum, A. W. D., Orth, R. J., & Duarte, C. M. (2006). *Seagrasses: Biology, Ecology, and Conservation*: Springer.

Leigh, J. W., & Bryant, D. (2015). popart: full-feature software for haplotype network construction. *Methods Ecol. Evol., 6*(9), 1110-1116. doi:10.1111/2041-210X.12410

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987-2993. doi:10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, Y., & Liu, J. (2018). StructureSelector: A web-based software to select and visualize the optimal number of clusters using multiple methods. *Mol. Ecol. Resour., 18*(1), 176-177. doi:10.1111/1755-0998.12719

Linder, C. R., & Rieseberg, L. H. (2004). Reconstructing patterns of reticulate evolution in plants. *Am. J. Bot., 91*(10), 1700-1708. doi:10.3732/ajb.91.10.1700

Ma, X., Olsen, J. L., Reusch, T. B. H., Procaccini, G., Kudrna, D., Williams, M., . . . Van de Peer, Y. (2021). Improved chromosome-level genome assembly and annotation of the seagrass, *Zostera marina* (eelgrass). *F1000research, 10*. doi:10.12688/f1000research.38156.1

Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Mol. Ecol. Resour., 21*(2), 584-595. doi:10.1111/1755-0998.13265

Morales-Briones, D. F., Liston, A., & Tank, D. C. (2018). Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). *New Phytol., 218*(4), 1668-1684. doi:10.1600/036364418X696897

Newman, W. (1976). *California Transition Zone: significance of short range endemics.* Paper presented at the Historical Biogeography, Plate Tectonics and the Changing Environment, Proceedings of the 37th Annual Biology Colloquium and Selected Papers, 1976.

Noble, J. C., Bell, A. D., & Harper, J. L. (1979). The population biology of plants with clonal growth: I. The morphology and structural demography of *Carex arenaria*. *J. Ecol., 67*, 983-1008. doi:10.2307/2259224

Olsen, J. L., Rouzé, P., Verhelst, B., Lin, Y.-C., Bayer, T., Collen, J., . . . Maumus, F. (2016). The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. *Nature, 530*(7590), 331-335. doi:10.1038/nature16548

Olsen, J. L., Stam, W. T., Coyer, J. A., Reusch, T. B., Billingham, M., Bostrom, C., . . . Wyllie-Echeverria, S. (2004). North Atlantic phylogeography and large-scale population differentiation of the seagrass *Zostera marina* L. *Mol. Ecol., 13*(7), 1923-1941. doi:10.1111/j.1365-294X.2004.02205.x

Palumbi, S. R. (1992). Marine speciation on a small planet. *Trends Ecol. Evol., 7*(4), 114-118. doi:10.1016/0169-5347(92)90144-Z

Palumbi, S. R. (1994). Genetic divergence, reproductive isolation, and marine speciation. *Annu. Rev. Ecol. Syst., 25*(1), 547-572. doi:10.1146/annurev.es.25.110194.002555

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., . . . Reich, D. (2012). Ancient admixture in human history. *Genetics, 192*(3), 1065-1093. doi:10.1534/genetics.112.145037

Pembleton, L. W., Cogan, N. O., & Forster, J. W. (2013). StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Mol. Ecol. Resour., 13*(5), 946-952. doi:10.1111/1755-0998.12129

Petit, R. J., & Vendramin, G. G. (2007). Plant phylogeography based on organelle genes: an introduction. In S. Weiss & N. Ferrand (Eds.), *Phylogeography of southern European refugia* (pp. 23-97): Springer.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959. doi:10.1093/genetics/155.2.945

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat. Genet., 46*(8), 919-925. doi:10.1038/ng.3015

Schiffels, S., & Wang, K. (2020). MSMC and MSMC2: the multiple sequentially markovian coalescent. In J. Y. Dutheil (Ed.), *Statistical population genomics* (pp. 147-166): Springer Nature.

Stange, M., Sánchez-Villagra, M. R., Salzburger, W., & Matschiner, M. (2018). Bayesian divergence-time estimation with genome-wide single-nucleotide polymorphism data of sea catfishes (Ariidae) supports Miocene closure of the Panamanian Isthmus. *Syst. Biol., 67*(4), 681-699. doi:10.1093/sysbio/syy006

Sun, J., Wang, Y., Liu, Y., Xu, C., Yuan, Q., Guo, L., & Huang, L. (2020). Evolutionary and phylogenetic aspects of the chloroplast genome of Chaenomeles species. *Sci. Rep., 10*, 11466. doi:10.1038/s41598-020-67943-1

Talbot, S. L., Sage, G. K., Rearick, J. R., Fowler, M. C., Muniz-Salazar, R., Baibak, B., . . . Ward, D. H. (2016). The Structure of Genetic Diversity in Eelgrass (*Zostera marina* L.) along the North Pacific and Bering Sea Coasts of Alaska. *PLoS ONE, 11*(4). doi:10.1371/journal.pone.0152701

Van der Auwera, G. A., & O'Connor, B. D. (2020). *Genomics in the cloud: using Docker, GATK, and WDL in Terra*: O'Reilly Media.

White, C., Selkoe, K. A., Watson, J., Siegel, D. A., Zacherl, D. C., & Toonen, R. J. (2010). Ocean currents help explain population genetic structure. *Proc. Royal Soc. B, 277*(1688), 1685-1694. doi:10.1098/rspb.2009.2214

Whitlock, M. C. (2003). Fixation probability and time in subdivided populations. *Genetics, 164*(2), 767-779. doi:10.1093/genetics/164.2.767

Wicke, S., Schneeweiss, G. M., Depamphilis, C. W., Müller, K. F., & Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol., 76*(3), 273-297. doi:10.1007/s11103-011-9762-4

Wickham, H. (2016). *Elegant graphics for data analysis*: Spinger.

Xu, L., Yu, R., Lin, X., Zhang, B., Li, N., Lin, K., . . . Bai, W. (2021). Different rates of pollen and seed gene flow cause branch-length and geographic cytonuclear discordance within Asian butternuts. *New Phytol.* doi:10.1111/nph.17564

Yu, L., Stachowicz, J. J., DuBois, K. & Reusch, T. B. H. (2022). Using “identity by heterozygosity (IBH)” to detect clonemates under prevalent clonal reproduction in multicellular diploids. *bioRxiv*, 2022.2002.2016.480681. doi:10.1101/2022.02.16.480681

Zhang, B., Xu, L., Li, N., Yan, P., Jiang, X., Woeste, K. E., . . . Bai, W. (2019). Phylogenomics reveals an ancient hybrid origin of the Persian walnut. *Mol. Biol. Evol., 36*(11), 2451-2461. doi:10.1093/molbev/msz112

Acknowledgements
This study was supported by a 4-year PhD scholarship from the China Scholarship Council (CSC) to L.Y. (No. 201704910807), and by a grant to J.L.O. from the U.S. Department of Energy (DOE) Joint Genome Institute (JGI). We thank D. Gill for lab assistance. We thank T. Bayer for discussions on bioinformatic problems. We thank Y. Li for assistance with the ABC-RF analysis.

Author contributions
J.J.S., J.S., J.L.O. and T.B.H.R. conceived and designed the study, M.K. analyzed the chloroplast data, L.Y., M.M. and A.H. conducted the phylogenetic analyses, A.H. identified the core genes, L.Y. calculated D-statistic with assistance from M.M., L.Y. conducted all the other analyses, T.D. assisted with bioinformatic analyses, L.Y., M.K., M.M. A.H., J.L.O., T.D. and T.B.H.R. discussed and interpreted the results, L.Y. and T.B.H.R. wrote the paper, with input from all authors.

Competing interests
The authors declare no competing interests.

Supplementary Note

Estimating the divergence time between *Zostera japonica* and *Z. marina*

The fourfold degenerative third-codon transversion rate (4DTv) for each of the 1,072 syntenic paralog pairs in the *Z. marina* reference V3.1 (as determined by comparative genomics pipeline GENESPACE; Schmutz & Lovell, 2021) was calculated. The 4DTv rates showed a distinct peak at approximately 0.48, indicating the *Z. marina* whole genome duplication event (~67 MYA; Olsen et al., 2016). The 90% confidence interval of the peak (0.4640-0.5164) was estimated based on 10,000 bootstraps.

The divergence time between *Z. marina* and *Z. japonica* was estimated based on the divergence of homologies. A constrained protein homology search was performed using BLAT and GeMoMa (v1.7; Keilwagen et al., 2019). To generate a constrained sequence database to search, *Z. marina* transcripts were aligned to the *Z. japonica* genome assembly. Each transcript's best hit location (+500 bp sequence buffer) was extracted from the *Z. japonica* genome and was used for GeMoMa protein prediction with default parameters. The protein prediction pipeline found 8,687 peptide sequences. Gffread (Pertea & Pertea, 2020) was used to extract each peptide coding sequence, and 1:1 orthologs (n=7,154) between *Z. japonica* and *Z. marina* were discovered using best reciprocal BLAT hits. The 4DTv rate among orthologs was calculated and estimated using 10,000 bootstrap estimates (0.0795-0.0816; 90% CI). Applying the WGD age to the 4DTv neutral rate, the divergence time between *Z. japonica* and *Z. marina* was estimated to be between 9.86 - 12.67 MYA, which was consistent with the upper bound estimate of 11.9 MYA from Coyer et al. (2013).

Estimating mutation rate

The divergence time between *Z. japonica* and *Z. marina* was represented as a lognormal distribution (Mean: 11.1542 MYA; SD: 0.07). An estimate of a neutral mutation rate was calculated based on the synonymous mutation rate (Ks) among *Z. japonica* orthologs and four *Z. marina* populations (JS, BB, NC, and FR). The median Ks peak between *Z. japonica* and *Z. marina* was approximately 0.211, which was consistent across all 4 populations. The neutral mutation rate was calculated as r = Ks / (2 * divergence age) = 9.461883 e-9.

SNP calling and filtering including a *Z. japonica* ramet

One *Z. japonica* ramet was included for genetic variant calling. The GVCF format file for the *Z. japonica* ramet was generated in the same way with the *Z. marina* ramets. All the GVCF files (190 *Z. marina* ramets + 1 *Z. japonica* ramet) were combined by CombineGVCFs (GATK4). GenotypeGVCFs (GATK4) was used to call genetic variants. BCFtools was used to remove SNPs within 20 base pairs of an indel or other variant type. Then we extracted only SNPs (10,562,762 SNPs), and excluded the other variant types. VariantsToTable (GATK4) was used to extract INFO annotations. SNPs meeting one or more than one of the following criteria were marked by VariantFiltration (GATK4): MQ < 40.0; FS > 60.0; QD < 10.0; MQRandSum > 2.5 or MQRandSum < -2.5; ReadPosRandSum < -2.5; ReadPosRandSum > 2.5; SOR > 3.0; DP > 11185.0 (2 * average DP). Those SNPs were

excluded by SelectVariants (GATK4). A total of 4,873,274 SNPs were retained. VCFtools was used to convert individual genotypes to missing data when GQ < 30 or DP < 10. Individual homozygous reference calls with one or more than one reads supporting the variant allele, and individual homozygous variant calls with one or more than one reads supporting the reference allele, were also converted to missing data using a custom Python3 script. Only bi-allelic SNPs were kept (4,775,984 SNPs). To avoid the reference-related biases, we focused on the core genes. Bedtools was used to find overlap between the SNPs and the core genes in the six main chromosomes (18,717 genes), and only those SNPs were kept (1,483,603 SNPs). Thirty-seven ramets were excluded from ZM_HQ_SNPs, including 10 ramets related to geitonogamous (within-clone) selfing (Supplementary Table 2), 17 ramets representing replicated genotypes (Supplementary Table 3), and 10 ramets with high missing rate (Supplementary Fig. 1). After excluding these ramets, some SNPs became monomorphic, which were then excluded. SnpEff was used to annotate each SNP. To obtain putatively neutral SNPs, we only kept SNPs annotated as “synonymous_variant” (ZMZJ_Neutral_SNPs, 171,756 SNPs). Then only SNPs without any missing data were kept (ZMZJ_D_SNPs, 20,341 SNPs).

References

Coyer, J. A., Hoarau, G., Kuo, J., Tronholm, A., Veldsink, J., & Olsen, J. L. (2013). Phylogeny and temporal divergence of the seagrass family Zosteraceae using one nuclear and three chloroplast loci. Systematics and Biodiversity, 11(3), 271-284.

Keilwagen, J., Hartung, F., & Grau, J. (2019). GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. Methods in molecular biology (Clifton, NJ), 1962, 161-177.

Olsen, J. L., Rouzé, P., Verhelst, B., Lin, Y. C., Bayer, T., Collen, J., ... & Van de Peer, Y. (2016). The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. Nature, 530(7590), 331-335.

Pertea, G., & Pertea, M. (2020). GFF utilities: GffRead and GffCompare. F1000Research, 9.

Schmutz, J., & Lovell, J. (2021). GENESPACE R Package (GENESPACE) v1. 0 (No. GENESPACE R Package). HudsonAlpha Institute for Biotechnology; Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States).

Supplementary Tables

Supplementary Table 1: Population metadata for *Zostera marina*.

	Region	Population code	Population location	Latitude	Longitude	*ZEN Cross ref	Sample Numbering
West Pacific							
	Japan North	JN	Akkeshi-ko Estuary, Hokkaido, Japan	43.021N	144.903E	JN-A	JN101-JN112
	Japan South	JS	Ikunoshima, Japan (Inland sea)	34.298N	132.916E	JS-A	JS201-JS 212
East Pacific							
	Bering Sea **LME	ASL	Safety Lagoon, Alaska, USA	64.485N	164.762W	Non-ZEN	ASL1-ASL12
	Gulf of Alaska LME	ALI	Izembek Lagoon, Alaska, USA	55.329N	162.821W	Non-ZEN	ALI301-ALI312
	California Current LME	WAS	Willapa Bay, Washington State, USA	46.474N	124.028W	WAS-A	WAS401-WAS412
	California Current LME	BB	Westside Park, Bodega Bay, California, USA	38.320N	123.055W	BB-A	BB501-BB512
	California Current LME	SD	Shelter Island, San Diego Bay, California, USA	32.714N	117.225W	SD-A	SD601-SD612
West Atlantic							
	Quebec	QU	Pointe-Lebel, Quebec, Canada	49.112N	68.176W	QU-A	QU701-QU711
	Massachusetts	MA	Dorothy Cove, Massachusetts, USA	42.420N	70.915W	MA-A	MA801-MA812
	North Carolina	NC	Middle Marsh, North Carolina, USA	34.692N	76.623W	NC-A	NC 1103-NC 1112
East Atlantic							
	Northern Norway	NN	Røvika, Norway (near Bødo)	67.268N	15.257E	NN-B	NN1201-NN1212
	Sweden	SW	Torserød, Sweden (west coast)	58.313N	11.549E	SW-A	SW1401-SW1412
	Wales	WN	Port Dinllaen, Wales, UK	52.991N	4.450W	WN-A	WN1501-WN1512
	Portugal	PO	Culatatra, Ria Formosa, S. Portugal	37.040N	7.910W	PO-A	PO1601-PO1611
Mediterranean Sea							
	France	FR	Bouzigues, Thau Lagoon, France	43.447N	3.662E	FR-A	FR1701-FR1712
	Croatia	CZ	Adriatic Sea, Posedarje, Croatia	44.212N	15.491E	CR-A	CZ1801-CZ1812

*The Zostera Experimental Network (ZEN) has produced extensive ecological and physical metadata available from Jay Stachowicz jjstachowicz@ucdavis.edu. **LME, Large marine ecosystem.

Supplementary Table 2: Possible parent-descendant pairs under selfing as detected by "identity by heterozygosity".

Selfing Pair	Parent Ramet	Descendant Ramet
SP_01	NN05	NN02
SP_02	NN05	NN06
SP_03	NN05	NN07
SP_04	NN05	NN09
SP_05	NN05	NN10
SP_06	NN08	NN02
SP_07	NN08	NN06
SP_08	NN08	NN07
SP_09	NN08	NN09
SP_10	NN08	NN10
SP_11	SD03	SD02
SP_12	SD03	SD12
SP_13	SD12	SD02
SP_14	WN02	WN03
SP_15	WN02	WN07
SP_16	WN12	WN05

Supplementary Table 3: Clonemates detected based on "identity by heterozygosity".

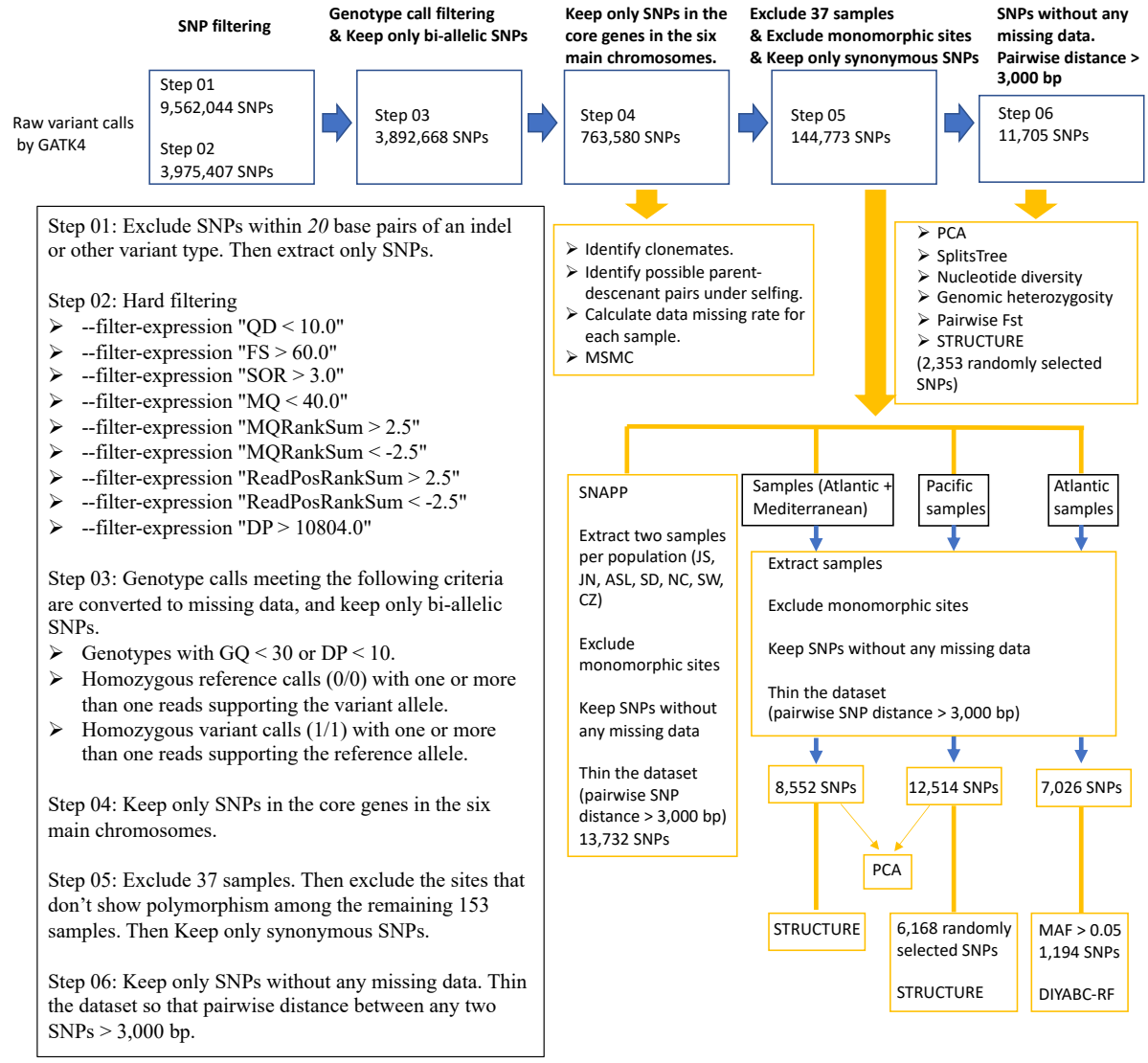
Genet	Clonemates						
Genet_01	BB04	BB05					
Genet_02	BB09	BB10					
Genet_03	JS03	JS04					
Genet_04	NN02	NN06	NN07	NN09	NN10		
Genet_05	PO02	PO05	PO07	PO08	PO10	PO11	PO12
Genet_06	PO03	PO04	PO06	PO09			
Genet_07	SD04	SD11					
Genet_08	SD06	SD09					
Genet_09	WN04	WN09					
Genet_10	WN06	WN10					
Genet_11	NN05	NN08					

Supplementary Table 4: Pairwise F_{ST}.

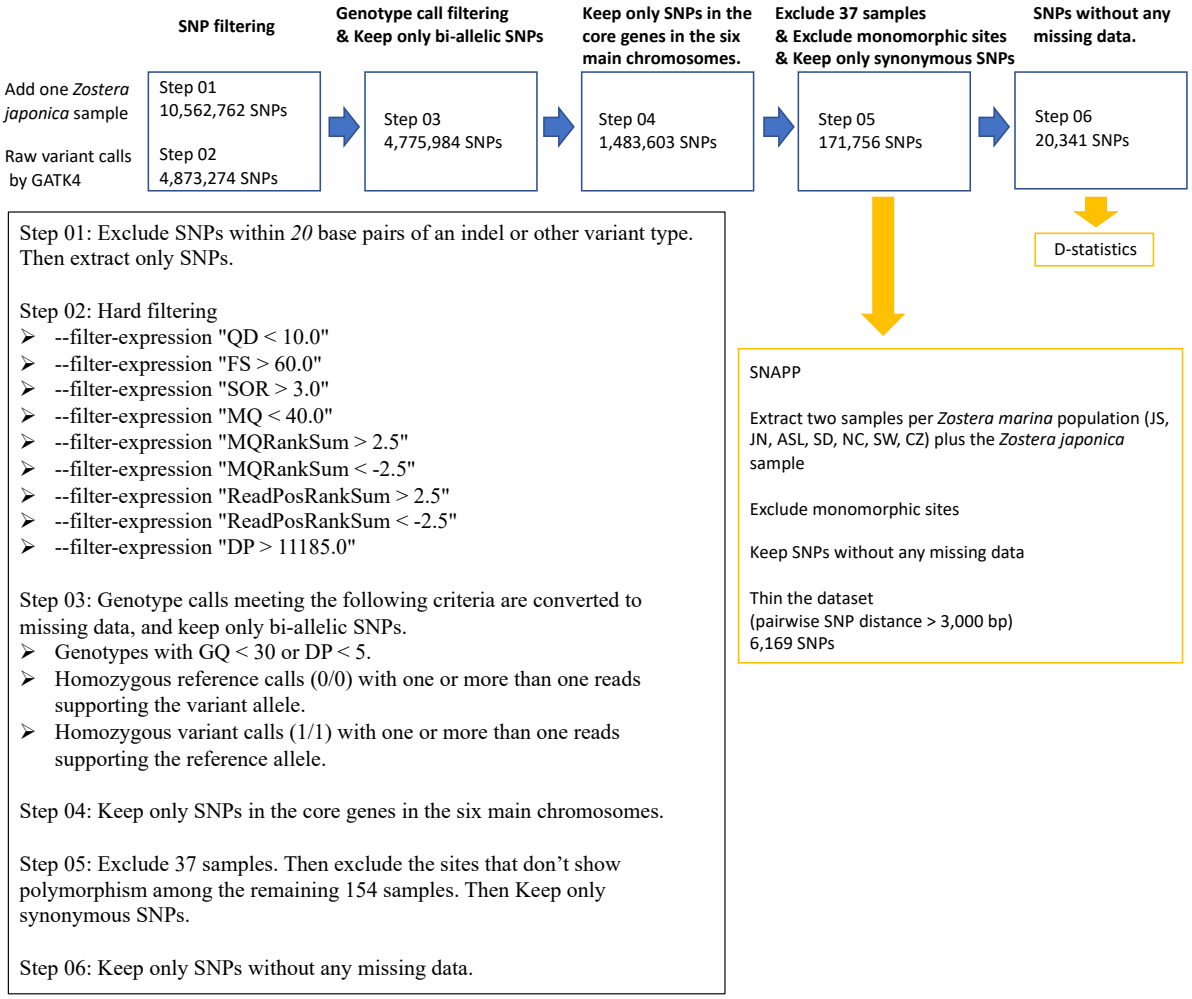
Pop\Pop	JN	JS	ASL	ALI	WAS	BB	SD	QU	MA	NC	NN	SW	WN	PO	FR	CZ
JN	0.0	0.523	0.399	0.31	0.474	0.525	0.692	0.629	0.635	0.616	0.58	0.634	0.587	0.518	0.64	0.633
JS	0.523	0.0	0.698	0.618	0.649	0.641	0.724	0.731	0.738	0.723	0.69	0.74	0.697	0.628	0.739	0.736
ASL	0.399	0.698	0.0	0.503	0.613	0.686	0.881	0.898	0.885	0.851	0.897	0.89	0.889	0.878	0.881	0.877
ALI	0.31	0.618	0.503	0.0	0.498	0.59	0.852	0.916	0.896	0.845	0.923	0.904	0.908	0.876	0.887	0.882
WAS	0.474	0.649	0.613	0.498	0.0	0.389	0.704	0.782	0.783	0.761	0.753	0.785	0.755	0.708	0.783	0.779
BB	0.525	0.641	0.686	0.59	0.389	0.0	0.556	0.782	0.784	0.764	0.748	0.787	0.752	0.69	0.784	0.78
SD	0.692	0.724	0.881	0.852	0.704	0.556	0.0	0.909	0.904	0.881	0.897	0.908	0.895	0.857	0.9	0.898
QU	0.629	0.731	0.898	0.916	0.782	0.782	0.909	0.0	0.391	0.358	0.594	0.445	0.473	0.738	0.609	0.57
MA	0.635	0.738	0.885	0.896	0.783	0.784	0.904	0.391	0.0	0.229	0.457	0.328	0.345	0.644	0.539	0.498
NC	0.616	0.723	0.851	0.845	0.761	0.764	0.881	0.358	0.229	0.0	0.345	0.277	0.28	0.502	0.464	0.427
NN	0.58	0.69	0.897	0.923	0.753	0.748	0.897	0.594	0.457	0.345	0.0	0.26	0.401	0.792	0.587	0.545
SW	0.634	0.74	0.89	0.904	0.785	0.787	0.908	0.445	0.328	0.277	0.26	0.0	0.175	0.658	0.517	0.455
WN	0.587	0.697	0.889	0.908	0.755	0.752	0.895	0.473	0.345	0.28	0.401	0.175	0.0	0.704	0.54	0.495
PO	0.518	0.628	0.878	0.876	0.708	0.69	0.857	0.738	0.644	0.502	0.792	0.658	0.704	0.0	0.601	0.609
FR	0.64	0.739	0.881	0.887	0.783	0.784	0.9	0.609	0.539	0.464	0.587	0.517	0.54	0.601	0.0	0.503
CZ	0.633	0.736	0.877	0.882	0.779	0.78	0.898	0.57	0.498	0.427	0.545	0.455	0.495	0.609	0.503	0.0

All pairwise F_{ST} have significant p values ($p < 0.001$), based on 1,000 bootstraps. JN shows the smallest F_{ST} with all populations from the Atlantic side (Atlantic + Mediterranean Sea) compared with the other Pacific populations.

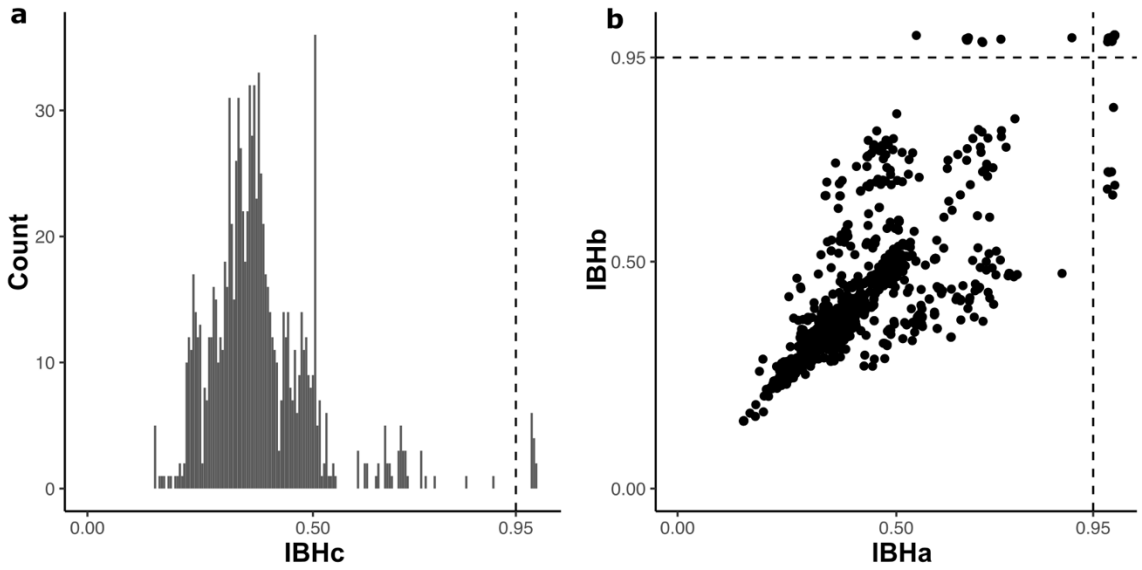
Supplementary Figures



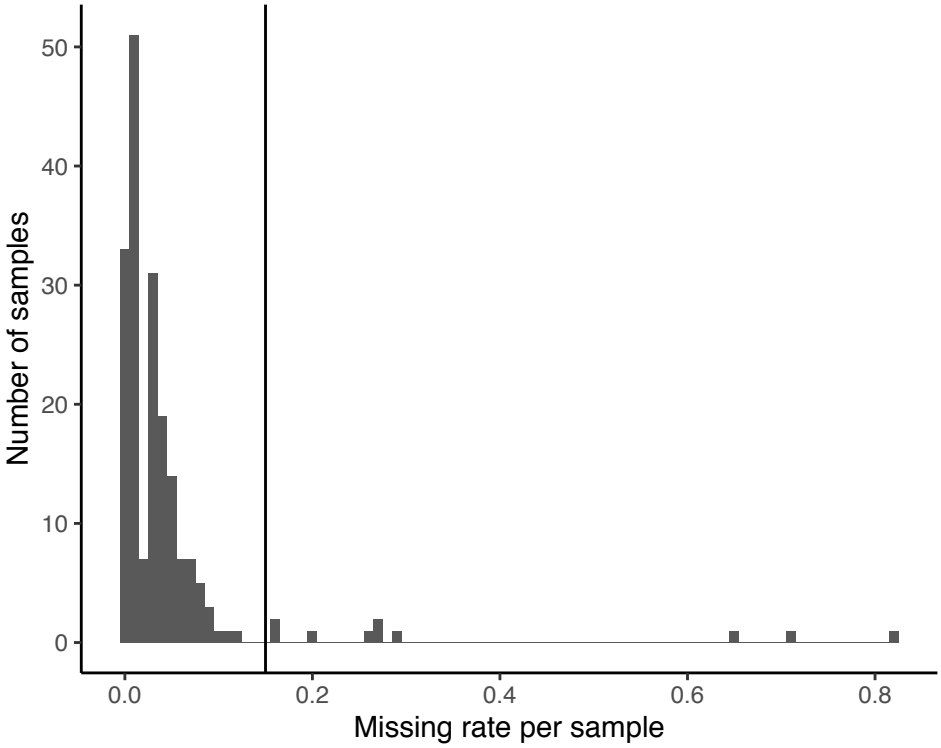
Supplementary Fig. 1 | Workflow for analyses based on *Zostera marina*.



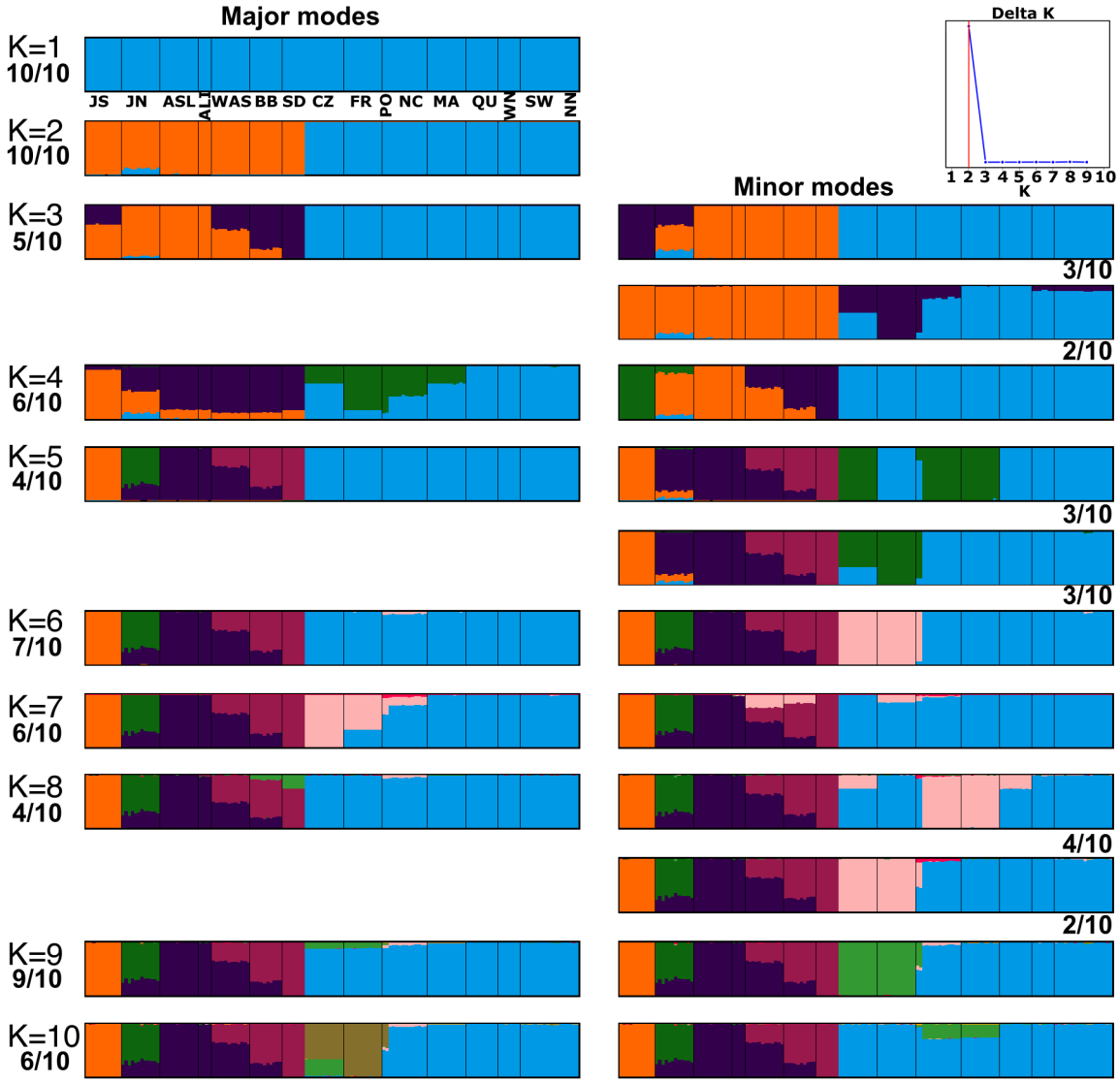
Supplementary Fig. 2 | Workflow for analyses that require an outgroup.



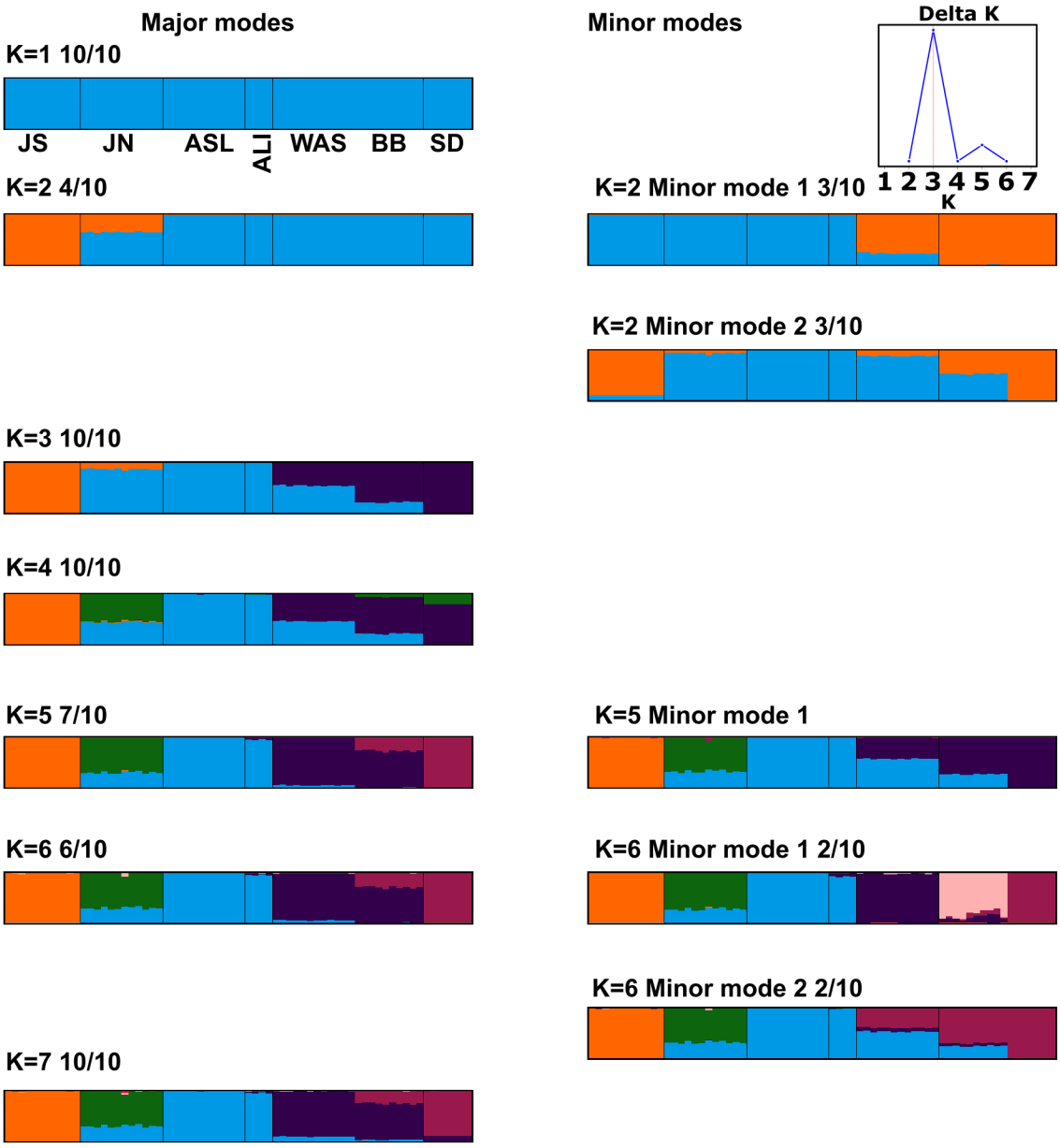
Supplementary Fig. 3: Detecting clonemates and possible parent-descendant pairs under selfing. For a pair of *Zostera marina* ramets (denoted as rametA and rametB, respectively), N_a denotes the number of the heterozygous loci in the genome of rametA, N_b denotes the number of the heterozygous loci in the genome of rametB, and N_{IBH} denotes number of the loci where A and B have identical heterozygous genotypes. $IBH_a = N_{IBH} / N_a$, $IBH_b = N_{IBH} / N_b$, and IBH similarity, $IBH_c = \min(IBH_a, IBH_b)$. **a,** When there is no somatic mutations and genotyping errors, IBH_c for a pair of clonemates will be equal to 1. To take into account somatic mutations and genotyping errors, we set the threshold to 0.95. Ramet pairs with $IBH_c > 0.95$ are considered clonemates. **b,** Under selfing, the descendants only inherit a subset of the heterozygous genotypes as in the parent. We used the same threshold as the one used in detecting clonemates, which was 0.95. When $IBH_a > 0.95$ & $IBH_b < 0.95$, we consider that rametA is produced by selfing of rametB. When $IBH_a < 0.95$ & $IBH_b > 0.95$, we consider that rametB is produced by selfing of rametA.



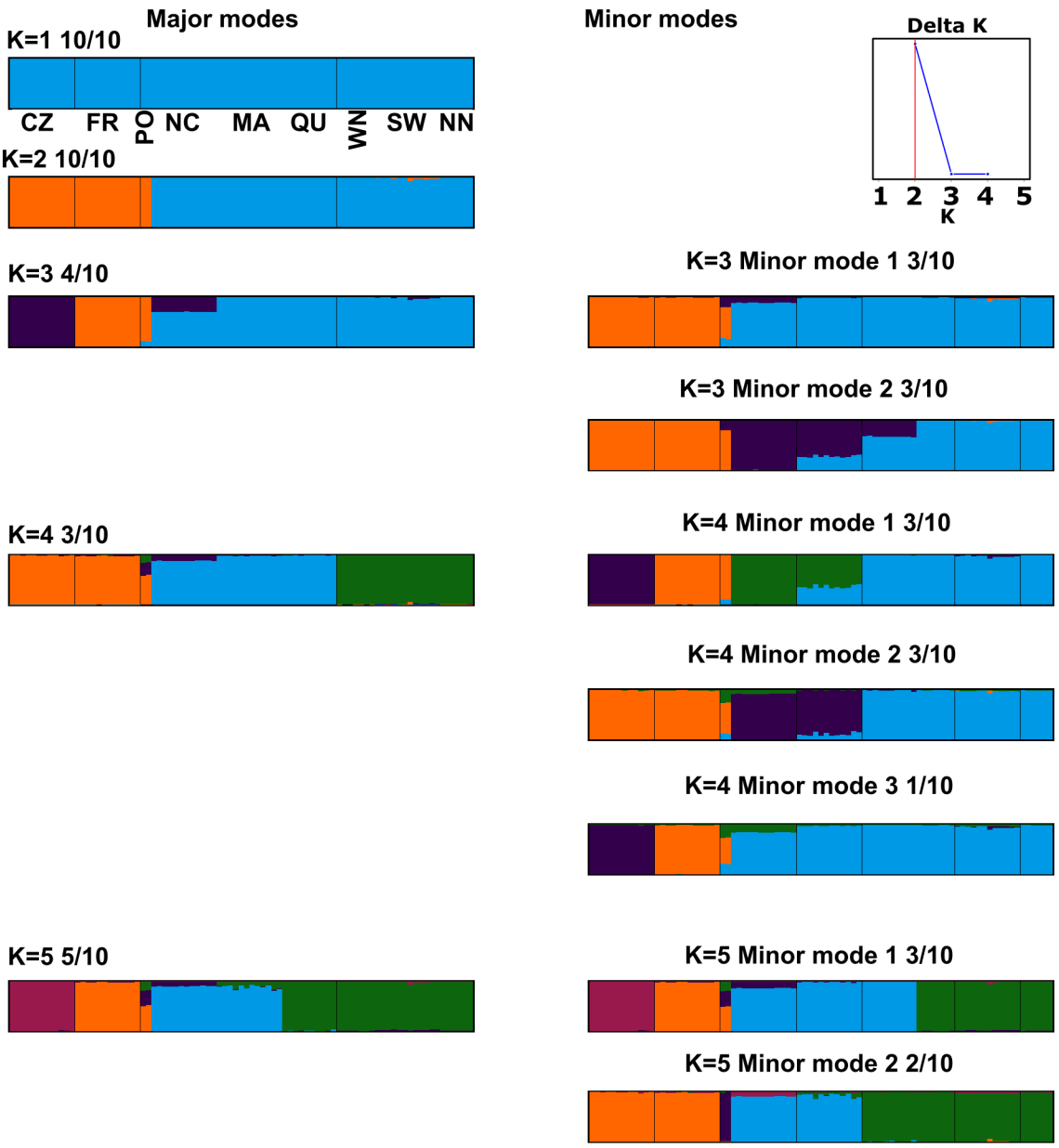
Supplementary Fig. 4 | Missing data rate for each ramet. The data missing rate for each ramet was calculated based on SNP data set ZM_HQ_SNPs (763,580 SNPs). Missing rate=number of loci with missing data/763,580. Samples with <15% missing data were retained. Ten of the 190 samples had a missing rate >15% (ALI01, ALI02, ALI03, ALI04, ALI05, ALI06, ALI10, ALI16, QU03 and SD08) and were excluded from the data set.



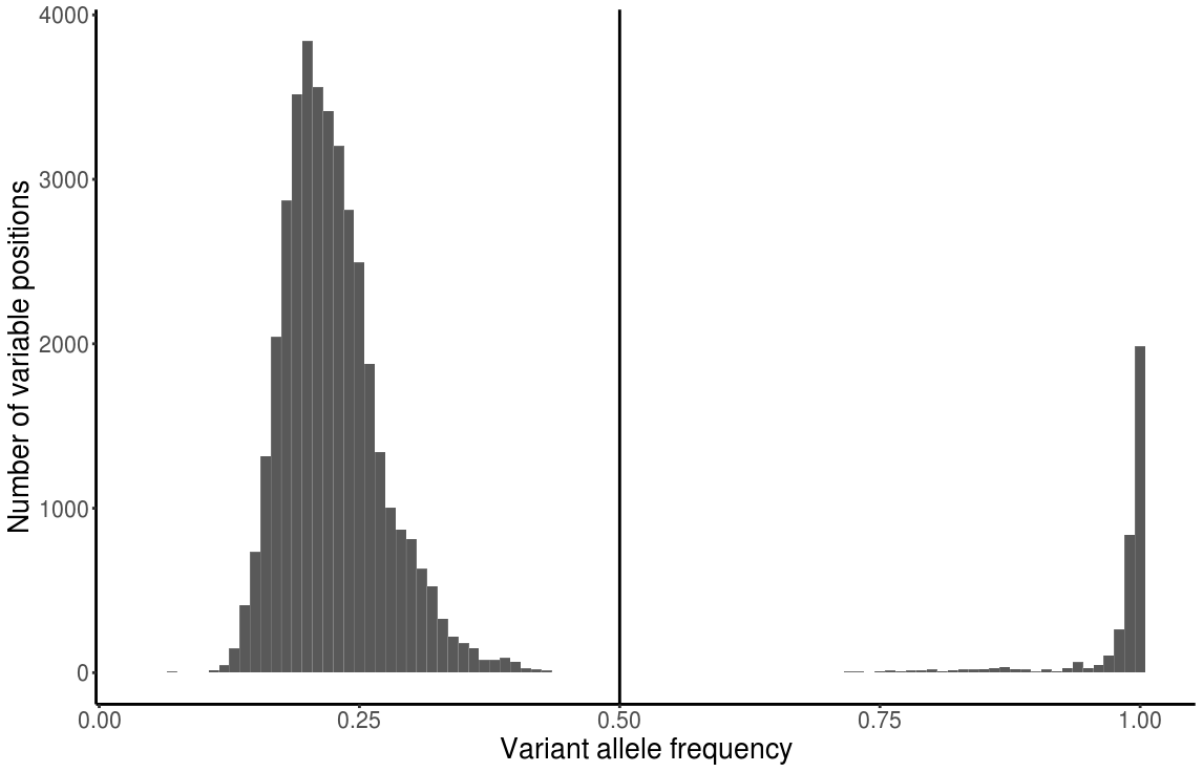
Supplementary Fig. 5 | Global STRUCTURE results for K from 1 to 10. The analysis was repeated 10 times with the same parameters. The number indicates how many times the mode occurred among the 10 runs.



Supplementary Fig. 6 | STRUCTURE results for only Pacific for K from 1 to 7. The analysis was repeated 10 times with the same parameters. The number indicates how many times the mode occurred among the 10 runs.



Supplementary Fig. 7 | STRUCTURE results for only Atlantic side (Atlantic + Mediterranean Sea) for K from 1 to 5. The analysis was repeated 10 times with the same parameters. The number indicates how many times the mode occurred among the 10 runs.



Supplementary Fig. 8 | Variant allele frequency based on chloroplast genomes. Variant allele frequency indicates the frequency of reads supporting the variant allele at a given locus. We only focused on the fixed variant alleles by setting a threshold of >0.5.

Supplementary Data

Supplementary Data 1: Data coverage.

Sample	Number of bases (Raw fastq)	Coverage (Raw fastq)	Number of bases (Clean fastq)	Coverage (Clean fastq)
ALI01	1,1689E+10	44.88	1,1249E+10	43.18
ALI02	1,2377E+10	47.51	1,1742E+10	45.08
ALI03	1,2053E+10	46.27	1,1588E+10	44.48
ALI04	1,1049E+10	42.42	1,0383E+10	39.86
ALI05	1,136E+10	43.61	1,0778E+10	41.38
ALI06	1,2805E+10	49.15	1,2245E+10	47.01
ALI07	1,0477E+10	40.22	1,0086E+10	38.72
ALI09	1,2407E+10	47.63	1,1562E+10	44.39
ALI10	1,203E+10	46.18	1,1467E+10	44.03
ALI11	1,1384E+10	43.7	1,0897E+10	41.83
ALI12	1,265E+10	48.56	1,2095E+10	46.43
ALI16	1,2304E+10	47.23	1,1813E+10	45.35
ASL01	1,1955E+10	45.9	1,1318E+10	43.45
ASL02	1,2068E+10	46.33	1,1454E+10	43.97
ASL03	1,757E+10	67.45	1,6965E+10	65.13
ASL04	1,6107E+10	61.83	1,2668E+10	48.63
ASL05	1,6599E+10	63.72	1,5296E+10	58.72
ASL06	1,0937E+10	41.99	1,0506E+10	40.33
ASL07	1,3849E+10	53.16	1,2937E+10	49.66
ASL08	1,2928E+10	49.63	1,1854E+10	45.51
ASL09	7237391042	27.78	6746383742	25.9
ASL10	6343489162	24.35	5945697760	22.82
ASL11	1,6133E+10	61.93	1,5522E+10	59.59
ASL12	1,2004E+10	46.08	1,1576E+10	44.43
BB01	7476630610	28.7	6809351806	26.14
BB02	1,4176E+10	54.42	1,383E+10	53.09
BB03	1,2599E+10	48.37	1,1984E+10	46.0
BB04	1,1873E+10	45.58	1,1148E+10	42.79
BB05	1,5334E+10	58.87	1,4475E+10	55.57
BB06	1,2774E+10	49.04	1,1982E+10	46.0
BB07	1,264E+10	48.52	1,2046E+10	46.24
BB08	1,1599E+10	44.52	1,0879E+10	41.76
BB09	1,3036E+10	50.04	1,223E+10	46.95
BB10	1,3293E+10	51.03	1,2411E+10	47.65
BB11	1,3288E+10	51.02	1,2477E+10	47.9
BB12	1,1571E+10	44.42	1,064E+10	40.84

JN01	1,1961E+10	45.92	1,1542E+10	44.3
JN02	1,4443E+10	55.45	1,3817E+10	53.05
JN03	1,6379E+10	62.88	1,5332E+10	58.86
JN04	1,3677E+10	52.5	1,2384E+10	47.54
JN05	1,0602E+10	40.7	1,0168E+10	39.03
JN06	1,2215E+10	46.9	1,176E+10	45.14
JN07	1,6108E+10	61.84	1,5911E+10	61.08
JN08	1,1825E+10	45.4	1,1394E+10	43.74
JN09	1,095E+10	42.03	1,0502E+10	40.31
JN10	1,2323E+10	47.31	1,1782E+10	45.23
JN11	1,1347E+10	43.56	1,0695E+10	41.06
JN12	9902835190	38.01	8999354952	34.55
JS01	9115651050	34.99	8541668138	32.79
JS02	9788727510	37.57	9294997038	35.68
JS03	1,0133E+10	38.9	9741517214	37.4
JS04	9903528280	38.01	9335611228	35.84
JS05	1,2193E+10	46.81	1,1558E+10	44.37
JS06	1,0831E+10	41.58	1,0091E+10	38.73
JS07	1,3207E+10	50.7	1,2275E+10	47.12
JS08	1,1091E+10	42.58	1,0299E+10	39.53
JS09	1,2339E+10	47.37	1,1816E+10	45.36
JS10	1,0504E+10	40.32	9992228282	38.36
JS11	1,062E+10	40.77	1,0187E+10	39.11
JS12	1,1587E+10	44.48	1,0983E+10	42.16
MA01	6046660006	23.21	5544151672	21.28
MA02	6460523826	24.8	5923832850	22.74
MA03	7388921958	28.37	6692542900	25.69
MA04	6899697058	26.49	6321606412	24.27
MA05	1,2647E+10	48.55	1,1933E+10	45.81
MA06	7440037270	28.56	6782267034	26.04
MA07	1,3163E+10	50.53	1,2628E+10	48.47
MA08	7648182616	29.36	6988320586	26.83
MA09	1,2742E+10	48.91	1,2112E+10	46.5
MA10	6941489630	26.65	6354844284	24.4
MA11	1,2929E+10	49.63	1,2326E+10	47.31
MA12	1,1948E+10	45.87	1,1229E+10	43.11
NC03	1,1944E+10	45.85	1,107E+10	42.5
NC04	6511241706	25.0	5967539164	22.91
NC05	1,1906E+10	45.7	1,1233E+10	43.13
NC06	1,2865E+10	49.39	1,1972E+10	45.96

NC07	1,1662E+10	44.76	1,0241E+10	39.31
NC08	6079544786	23.34	5471137078	21.0
NC09	1,2578E+10	48.29	1,1439E+10	43.91
NC11	6305300960	24.21	5675497378	21.79
NC12	1,1918E+10	45.75	1,1087E+10	42.56
NC13	6682212664	25.65	6004206268	23.05
NC14	1,1909E+10	45.72	1,109E+10	42.57
NC15	1,4693E+10	56.4	1,3647E+10	52.39
QU01	1,3573E+10	52.11	1,3056E+10	50.11
QU02	1,132E+10	43.46	1,0662E+10	40.93
QU03	9420789132	36.16	8986175100	34.49
QU05	1,1237E+10	43.14	1,0754E+10	41.29
QU06	1,4626E+10	56.15	1,4E+10	53.75
QU07	8722329572	33.48	7733844056	29.69
QU08	7593190530	29.15	6832298932	26.23
QU09	1,4347E+10	55.08	1,3273E+10	50.95
QU10	1,6767E+10	64.37	1,6019E+10	61.5
QU11	9947742590	38.18	9540020832	36.62
QU12	1,1648E+10	44.72	1,1019E+10	42.3
SD01	1,2299E+10	47.22	1,1787E+10	45.25
SD02	1,1916E+10	45.74	1,1042E+10	42.39
SD03	1,291E+10	49.56	1,2416E+10	47.66
SD04	1,1174E+10	42.89	1,0432E+10	40.05
SD05	1,1644E+10	44.7	1,0997E+10	42.21
SD06	1,676E+10	64.34	1,6255E+10	62.4
SD07	1,197E+10	45.96	1,136E+10	43.61
SD08	1,3675E+10	52.5	1,2766E+10	49.01
SD09	1,42E+10	54.51	1,3515E+10	51.88
SD10	1,1956E+10	45.9	1,1225E+10	43.09
SD11	1,3179E+10	50.59	1,2633E+10	48.5
SD12	1,2159E+10	46.68	1,1661E+10	44.76
WAS01	1,3605E+10	52.22	1,3055E+10	50.12
WAS02	1,5273E+10	58.63	1,4766E+10	56.69
WAS03	7473724766	28.69	7021998980	26.96
WAS04	1,2697E+10	48.74	1,1998E+10	46.06
WAS05	6800059406	26.1	6260169372	24.03
WAS06	1,3877E+10	53.27	1,2475E+10	47.89
WAS07	6733428542	25.85	6192831414	23.77
WAS08	1,1917E+10	45.75	1,1019E+10	42.3
WAS09	6744765622	25.89	6298279098	24.18

WAS10	1,3329E+10	51.16	1,2735E+10	48.89
WAS11	6385394682	24.51	5860596420	22.5
WAS12	1,1241E+10	43.15	1,068E+10	40.99
NN01	2,4098E+10	92.51	2,2297E+10	85.6
NN02	2,1913E+10	84.12	2,0146E+10	77.34
NN03	2,2066E+10	84.71	2,0357E+10	78.15
NN04	1,7674E+10	67.85	1,6746E+10	64.29
NN05	2,1148E+10	81.19	1,9517E+10	74.92
NN06	2,5075E+10	96.26	2,3151E+10	88.87
NN07	2,4548E+10	94.24	2,2705E+10	87.16
NN08	1,9313E+10	74.14	1,7795E+10	68.31
NN09	2,4162E+10	92.75	2,284E+10	87.68
NN10	2,2423E+10	86.08	1,9142E+10	73.48
NN11	2,5987E+10	99.76	2,4758E+10	95.04
NN12	2,505E+10	96.17	2,3179E+10	88.98
SW01	1,9571E+10	75.13	1,8286E+10	70.2
SW02	2,2076E+10	84.75	1,9613E+10	75.29
SW03	2,6233E+10	100.71	2,4589E+10	94.39
SW04	2,1655E+10	83.13	1,8687E+10	71.74
SW05	1,9877E+10	76.3	1,7219E+10	66.1
SW06	2,3457E+10	90.05	1,8653E+10	71.61
SW07	1,4376E+10	55.19	1,2244E+10	47.0
SW08	1,6486E+10	63.29	1,6296E+10	62.56
SW09	1,9767E+10	75.89	1,7E+10	65.26
SW10	1,6155E+10	62.02	1,5742E+10	60.43
SW11	1,4996E+10	57.57	1,3706E+10	52.62
SW12	1,1419E+10	43.84	1,1312E+10	43.43
WN01	2,4276E+10	93.19	2,221E+10	85.26
WN02	2,3121E+10	88.76	2,0671E+10	79.35
WN03	2,4551E+10	94.25	2,1978E+10	84.37
WN04	2,2169E+10	85.11	1,7544E+10	67.35
WN05	2,0224E+10	77.64	1,6698E+10	64.1
WN06	2,0573E+10	78.98	1,7043E+10	65.43
WN07	2,3762E+10	91.22	2,1919E+10	84.14
WN08	2,316E+10	88.91	2,1511E+10	82.58
WN09	1,8445E+10	70.81	1,8164E+10	69.73
WN10	2,5974E+10	99.71	2,5868E+10	99.31
WN11	2,4747E+10	95.0	2,2923E+10	88.0
WN12	1,948E+10	74.78	1,6196E+10	62.17
PO02	1,3687E+10	52.54	1,3562E+10	52.06

PO03	2,783E+10	106.84	2,549E+10	97.85
PO04	2,1427E+10	82.26	1,8965E+10	72.8
PO05	2,5289E+10	97.08	2,3178E+10	88.98
PO06	2,7643E+10	106.12	2,4744E+10	94.99
PO07	2,2902E+10	87.92	2,0909E+10	80.27
PO08	2,0943E+10	80.4	1,9451E+10	74.67
PO09	2,0131E+10	77.28	1,8268E+10	70.13
PO10	1,7224E+10	66.12	1,5956E+10	61.25
PO11	1,9753E+10	75.83	1,7739E+10	68.1
PO12	2,2093E+10	84.81	1,9957E+10	76.61
FR01	2,1228E+10	81.49	1,901E+10	72.98
FR02	1,4053E+10	53.95	1,3926E+10	53.46
FR03	2,1178E+10	81.3	1,8924E+10	72.65
FR04	1,9854E+10	76.22	1,8405E+10	70.65
FR05	1,7938E+10	68.86	1,5869E+10	60.92
FR06	2,0532E+10	78.82	1,8032E+10	69.22
FR07	1,9522E+10	74.94	1,7301E+10	66.42
FR08	1,4117E+10	54.19	1,2241E+10	46.99
FR09	2,1066E+10	80.87	1,9247E+10	73.89
FR10	2,3174E+10	88.96	2,0557E+10	78.91
FR11	1,6324E+10	62.67	1,4898E+10	57.19
FR12	1,5646E+10	60.06	1,4296E+10	54.88
CZ01	2,0968E+10	80.5	1,9194E+10	73.68
CZ02	2,5798E+10	99.04	2,3313E+10	89.5
CZ03	2,3528E+10	90.32	2,1302E+10	81.77
CZ04	2,3028E+10	88.4	2,1164E+10	81.25
CZ05	2,1934E+10	84.2	1,9074E+10	73.22
CZ06	2,2415E+10	86.05	1,8674E+10	71.69
CZ07	2,357E+10	90.48	2,0899E+10	80.23
CZ08	2,2505E+10	86.39	2,0085E+10	77.1
CZ09	1,8607E+10	71.43	1,6617E+10	63.79
CZ10	2,7159E+10	104.26	2,4036E+10	92.27
CZ11	2,2345E+10	85.78	1,9665E+10	75.49
CZ12	2,5286E+10	97.07	2,1629E+10	83.03

Supplementary Data 2: Mapping rate.

Sample	Mapped(%)	Properly_paired(%)
ALI01	34.78	31.91
ALI02	33.04	30.28
ALI03	21.55	19.77
ALI04	22.89	21.01
ALI05	47.87	44.07
ALI06	36.83	33.88
ALI07	55.08	50.54
ALI09	88.00	81.48
ALI10	34.50	31.60
ALI11	64.02	58.94
ALI12	57.19	52.93
ALI16	16.86	15.40
ASL01	99.45	93.04
ASL02	95.04	88.91
ASL03	83.56	77.64
ASL04	87.78	81.53
ASL05	78.60	73.14
ASL06	90.85	84.35
ASL07	90.98	84.64
ASL08	73.55	67.78
ASL09	94.12	86.97
ASL10	94.01	87.33
ASL11	95.96	89.77
ASL12	80.89	75.04
BB01	99.51	91.84
BB02	98.95	91.63
BB03	96.05	88.50
BB04	99.17	92.10
BB05	98.98	91.18
BB06	98.50	91.37
BB07	98.39	90.71
BB08	99.15	91.32
BB09	99.05	92.07
BB10	99.15	91.86
BB11	99.47	91.65
BB12	99.11	91.79
JN01	99.58	92.18
JN02	99.18	91.63

JN03	98.66	91.74
JN04	99.49	91.60
JN05	98.62	91.13
JN06	99.72	92.27
JN07	99.58	93.11
JN08	98.73	91.25
JN09	99.21	92.02
JN10	99.20	91.79
JN11	98.75	92.04
JN12	98.81	91.42
JS01	99.00	90.59
JS02	98.42	89.43
JS03	99.35	89.62
JS04	97.98	88.33
JS05	98.34	89.66
JS06	98.72	89.53
JS07	98.45	89.19
JS08	99.02	89.95
JS09	98.84	90.00
JS10	99.22	89.69
JS11	98.50	89.00
JS12	98.48	89.21
MA01	99.90	97.13
MA02	99.24	96.75
MA03	99.28	96.83
MA04	99.91	97.69
MA05	98.61	96.24
MA06	99.87	97.12
MA07	99.27	96.83
MA08	99.83	97.22
MA09	98.88	96.37
MA10	99.51	96.83
MA11	99.73	97.33
MA12	99.43	96.64
NC03	98.10	95.50
NC04	97.96	94.50
NC05	99.37	96.44
NC06	97.52	94.65
NC07	98.42	95.48
NC08	97.72	94.92

NC09	97.45	94.75
NC11	98.02	95.27
NC12	97.00	94.27
NC13	99.62	96.76
NC14	98.04	95.42
NC15	98.35	95.47
QU01	99.49	97.24
QU02	67.98	66.28
QU03	48.87	47.43
QU05	56.42	54.84
QU06	99.31	97.00
QU07	83.74	81.79
QU08	87.20	84.97
QU09	47.80	46.46
QU10	99.32	96.88
QU11	66.92	65.05
QU12	66.07	64.34
SD01	83.97	76.03
SD02	96.36	87.89
SD03	99.04	90.30
SD04	97.30	88.61
SD05	93.91	85.45
SD06	96.45	89.02
SD07	97.94	89.13
SD08	48.71	44.10
SD09	98.11	89.87
SD10	96.34	87.92
SD11	95.94	87.92
SD12	97.88	89.62
WAS01	88.91	82.10
WAS02	98.95	91.28
WAS03	98.51	91.14
WAS04	98.80	91.25
WAS05	97.66	90.16
WAS06	98.40	91.61
WAS07	94.41	87.13
WAS08	97.86	90.18
WAS09	99.16	91.44
WAS10	99.17	91.64
WAS11	97.03	89.06

WAS12	97.71	89.95
NN01	99.93	98.69
NN02	99.44	98.32
NN03	99.95	98.81
NN04	99.50	98.10
NN05	99.91	98.61
NN06	99.68	98.35
NN07	99.96	98.60
NN08	99.42	98.00
NN09	99.92	98.39
NN10	98.01	96.64
NN11	99.77	98.34
NN12	99.64	98.22
SW01	99.74	98.30
SW02	98.41	97.00
SW03	99.17	97.77
SW04	99.89	98.41
SW05	97.26	96.02
SW06	99.63	98.34
SW07	96.78	95.52
SW08	97.58	96.30
SW09	97.97	96.61
SW10	98.26	96.57
SW11	99.58	98.00
SW12	98.92	97.15
WN01	91.13	89.77
WN02	91.88	90.56
WN03	91.78	90.48
WN04	93.58	92.27
WN05	97.58	96.22
WN06	91.35	89.98
WN07	89.44	88.02
WN08	88.53	87.18
WN09	89.19	87.47
WN10	95.41	93.77
WN11	89.52	88.11
WN12	97.69	96.28
PO02	99.56	97.03
PO03	99.74	97.07
PO04	98.59	95.78

PO05	99.56	96.77
PO06	99.64	96.83
PO07	99.57	96.88
PO08	99.28	96.50
PO09	99.80	97.11
PO10	99.40	96.73
PO11	98.29	95.65
PO12	99.37	96.72
FR01	99.28	96.84
FR02	97.69	94.98
FR03	98.32	96.27
FR04	99.77	97.49
FR05	98.77	96.70
FR06	99.30	97.08
FR07	99.83	97.40
FR08	98.39	96.10
FR09	99.38	97.08
FR10	98.71	96.34
FR11	99.78	97.59
FR12	99.43	97.07
CZ01	98.11	95.98
CZ02	98.48	96.29
CZ03	99.10	96.82
CZ04	98.62	96.35
CZ05	98.96	96.75
CZ06	99.87	97.57
CZ07	98.78	96.60
CZ08	99.20	96.97
CZ09	99.13	96.98
CZ10	99.29	97.07
CZ11	99.51	97.37
CZ12	98.66	96.55

Chapter 3: Using "identity by heterozygosity (IBH)" to detect clonemates under prevalent clonal reproduction in multicellular diploids

Lei Yu[1*], John J. Stachowicz[2, 3], Katie DuBois[2, 3], Thorsten B. H. Reusch[1]

[1]Marine Evolutionary Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany
[2]Center for Population Biology, University of California, Davis, CA, USA
[3]Department of Evolution and Ecology, University of California, Davis, CA, USA
[*]Corresponding author, lyu@geomar.de

Abstract

Clonal reproduction, the production of asexual modular units with the ability to live independently under multicellularity, is a very common life history. Genetic markers have been widely used to detect clonemates tracing back to a single sexual event. However, the currently available methods cannot work in cases where all collected samples belong to a few or even one single clone. Here we propose the concept of identity by heterozygosity (IBH), meaning that two diploid genomes are identically heterozygous at some genetic markers such as SNPs (single-nucleotide polymorphisms). Based on IBH, we develop IBH similarity (IBH_s), a novel measure of genetic similarity between two diploid genomes. The core of our IBH method for detecting clonemates is to use IBH_s in combination with the cutoff of 0.95. Two samples with IBH_s >0.95 are considered members of the same clone, while two samples with IBH_s <=0.95 are considered to represent different clones. The method requires a relatively large number of marker loci where the two target samples are identically heterozygous (i.e., N_{IBH}, on the order of >=3,000). Another potential application of IBH is to detect possible parent-descendant pairs under selfing because this reproductive mode leads to a predictable loss of heterozygosity. Our IBH method extends the detection of clonemates to any pair of samples, and will allow the detection of clonemates in extreme cases where all collected samples belong to one single clone.

Key words: identity by heterozygosity, IBH similarity, clonal reproduction, clonemate

Introduction

Clonal reproduction, the production of asexual modular units with the ability to live independently, can happen by means of asexual budding, fragmentation or branching. This life history is widespread in multicellular animals, plants and fungi, shared by 66.5% of flora (Honnay & Bossuyt, 2005), and 40% of animal phyla (Buss, 1983). A ramet (Harper & White, 1974) is the basic modular unit with an ability to live independently. A genet (synonymous with clone) is the product of a single zygote from gamete fusion (Noble et al., 1979), and comprises all ramets that emerge during its lifetime (i.e., clonemates).

Ramets belonging to the same clone vs. different clones are often physically indistinguishable. Once they obtain autonomy and break off the parental module, genetic markers such as microsatellites or AFLP have been widely used for detecting clonemates (Arnaud-Haond et al., 2007). A ramet can be genotyped with multiple independent markers, and the genotypes at all these loci collectively form a multilocus genotype (MLG). In case of co-dominant markers such as microsatellites, identical or nearly identical MLGs are then interpreted as being members of the same clone or genet. However, repeated MLGs are not definitive evidence for clonal reproduction (Halkett et al., 2005), if for example there is low (or no) allelic diversity at particular loci. In order to calculate the probability of finding identical MLGs resulting from distinct zygotes, the population allelic frequencies are required, and the population is assumed to be under Hardy-Weinberg equilibrium (Parks & Werth, 1993). Evidently, estimates of such population-level frequencies are inaccurate in populations dominated by a few clones (Arnaud-Haond et al., 2007), and meaningless if a single site is dominated by one clone (Smith et al., 1992; Reusch et al., 1999; Arnaud-Haond et al., 2012; Japaud et al., 2015), thereby rendering distinction of clonemates from non-clonemates difficult.

Moreover, members of the same clone could also have slightly different MLGs, due to a combination of somatic mutations and genotyping errors (Douhovnikoff & Dodd, 2003; Yu et al., 2020). To overcome this issue and also to work with dominant markers that do not allow for the application of MLGs, an alternative suite of approaches is the analysis of pairwise genetic distances. Ramet pairs from the same clone will show much lower levels of genetic distances than those from different clones, and thus the frequency distribution of distances will be bimodal rather unimodal. Then, a threshold can be defined to separate the intra-clone vs. inter-clone distances (Douhovnikoff & Dodd, 2003; Bailleul et al., 2016; Meirmans, 2020). However, it is impossible to define the threshold without inter-clone distances, and again, such methods cannot work on cases where all sampled ramets belong to one single clone.

In the genomic era when whole-genome data can be easily obtained, we propose to detect clonemates in diploid species by directly assessing the genetic similarity between two target samples based on the thousands of segregating genetic markers throughout their genomes. Heterozygous sites are the only informative ones because they show the potential for segregation

in sexual reproduction, while remaining identically heterozygous in clonal reproduction. Here we propose the concept of identity by heterozygosity (IBH), meaning that two diploid genomes are identically heterozygous at some genetic markers such as SNPs (single-nucleotide polymorphisms). Based on IBH, we developed IBH similarity (IBH_s), a novel measure of genetic similarity between two diploid genomes.

IBH method is most suitable for large number of SNPs obtained from whole-genome resequencing, restriction-site associated DNA (RAD) sequencing, or SNP arrays. The goal of this study is to test the performance of our IBH method using seagrass *Zostera marina* clones of known clonal history and/ or known age.

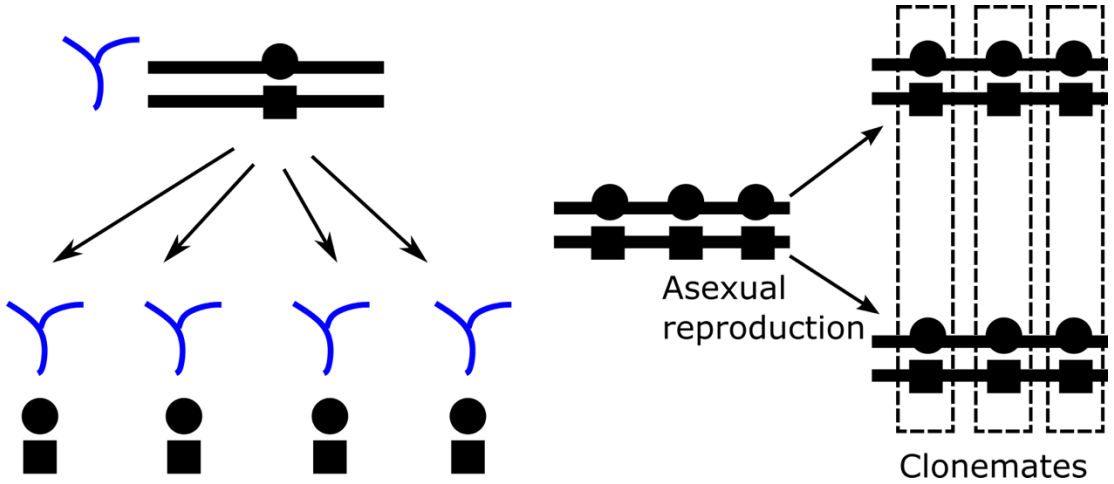


Figure 1: The maintenance of heterozygous genotypes under asexual reproduction. Under asexual reproduction, the heterozygous loci in the parent will remain identically heterozygous in all asexual descendants, except for rare somatic mutations.

Materials and methods

Detecting clonemates based on identity by heterozygosity (IBH)

For a pair of samples in multicellular diploids, N_{IBH} denotes the number of the genetic markers where the two samples are identically heterozygous. We define $IBH_X = N_{IBH}$ / (number of the heterozygous loci in the sample named "X"). IBH similarity (IBH_s) is defined as the smaller value of IBH_X for the two target samples. If assuming no somatic mutations, a pair of clonemates would share all the heterozygous genotypes (Fig. 1), which leads to $IBH_s = 1$. Somatic mutations and genotyping errors will cause the IBH_s to be slightly smaller.

Threshold (N_{IBH} and IBH_s) decided based on a *Z. marina* clone as old as 750 to 1,500 years old

To set the threshold for N_{IBH} and IBH_s, respectively, we applied the method to the SNP data set in Yu et al. (2020). Yu et al. (2020) conducted whole-genome resequencing of 24 ramets of a seagrass *Z. marina* clone located in Finland. The clone was estimated to be between 750 and

1,500 years old. Here the clone was named “FIN”, and the 24 ramets were named from FIN01 to FIN24, respectively. The SNP data set in Yu et al. (2020) contained 38,831 genomic sites before excluding the ones where all 24 ramets are identically heterozygous (see Extended Data Fig. 2 in Yu et al., 2020).

To obtain a proper threshold for N_{IBH} and IBH_s, respectively, we randomly selected 1/10, 1/100 and 1/1000 from the 38,831 SNPs, based on which the pairwise IBH_s was calculated for each pair of ramets. Each selection was repeated 100 times, and the IBH_s values from the 100 repeats were pooled together to plot a histogram (Fig. 2).

Seagrass *Zostera marina* clones with known ancestry

Two *Z. marina* clones were selected from the eight clones cultured in adjacent 300-l outdoor flow-through mesocosms at Bodega Marine Laboratory (BML) under ambient light and temperature conditions (Hughes et al., 2009). Each mesocosm was initiated with a single terminal shoot collected from one of eight different locations in Bodega Harbor, California, in July 2004. The genetic distinctiveness of each of the original shoots was verified before planting. Six shoots (= ramets) were collected from each clone for genomic analysis in 2021, after the clones had grown in mesocosm for 17 years. One clone was named “GREEN”, and the six collected shoots were named G01, G02, G06, G07, G10 and G16, respectively. The other clone was named “RED”, and the six collected shoots were named R02, R03, R05, R08, R09 and R10, respectively.

DNA extraction, whole-genome resequencing, quality check and filtering of raw Illumina reads

Genomic DNA was extracted using NucleoSpin plant II kit from Macherey-Nagel following the instructions. The plant tissue (160-200 mg fresh weight) was ground with lN_2. Library construction and Illumina sequencing (PE150, paired end 150 bp) were conducted by BGI (Beijing Genomics Institute, Hong Kong). The quality of the raw Illumina reads was assessed by FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/) was used to remove adapter contamination and conduct a basic quality filtering: (1) reads with more than one Ns were discarded (maxns=1); (2) reads shorter than 50 bp after trimming were discarded (minlength=50); (3) reads with average quality below 10 after trimming were discarded (maq=10). FastQC was used for second round of quality check for the filtered reads.

SNP calling and filtering

The quality-filtered Illumina reads were mapped against the chromosome-level *Z. marina* reference genome V3.1 (Ma et al., 2021) using BWA MEM (Li & Durbin, 2009). The alignments were converted to BAM format and then sorted using Samtools (Li et al., 2009). The MarkDuplicates module in GATK4 (Van der Auwera & O'Connor, 2020) was used to identify

and tag duplicate reads in the BAM files. Sequencing depth and percentage of covered genome region for each ramet were calculated using Samtools and Bedtools (Quinlan & Hall, 2010), respectively. HaplotypeCaller (GATK4) was used to generate GVCF format file for each ramet, and all the GVCF files were combined by CombineGVCFs (GATK4). GenotypeGVCFs (GATK4) was used to call genetic variants.

BCFtools (Li, 2011) was used to remove SNPs within 20 base pairs of an indel or other variant type. Then we extracted only bi-allelic SNPs located in the six main chromosomes (2,969,820 SNPs). INFO annotations were extracted using VariantsToTable (GATK4). SNPs meeting one or more than one of the following criteria were marked by VariantFiltration (GATK4): MQ < 40.0; FS > 60.0; QD < 13.0; MQRandSum > 2.5 or MQRandSum < -2.5; ReadPosRandSum < -2.5; ReadPosRandSum > 2.5; SOR > 3.0; DP > 1192.45 (2 * average DP). Those SNPs were excluded by SelectVariants (GATK4). A total of 1,105,628 SNPs were retained. VCFtools (Danecek et al., 2011) was used to convert individual genotypes to missing data when GQ < 30 or DP < 20. Individual homozygous reference calls with one or more than one reads supporting the variant allele, and individual homozygous variant calls with one or more than one reads supporting the reference allele, were also converted to missing data using a custom Python3 script. Finally, we kept only genomic sites showing polymorphism (604,066 SNPs).

Calculation of IBH measures

IBH measures were calculated by a custom Phython3 script for each pair of *Z. marina* ramets. We only focused on the nucleotide positions in the reference genome where both samples had available genotype calls (not missing data).

Results

For the *Z. marina* clone in Yu et al. (2020), the pairwise N_{IBH} ranged from 29,338 to 32,756. All ramet pairs displayed IBH_s greater than 0.95 (Fig. 2a). Based on the subsets of the SNP data set (Fig. 2b-d), we found that the minimum IBH_s increased when N_{IBH} increased. When N_{IBH} ranged from 2,808 to 3,407, the minimum IBH_s was 0.96. Hence, a threshold of 0.95 would be proper. We suggest to use our method only when N_{IBH} is on the order of >=3000, and the threshold for IBH_s can be set to 0.95.

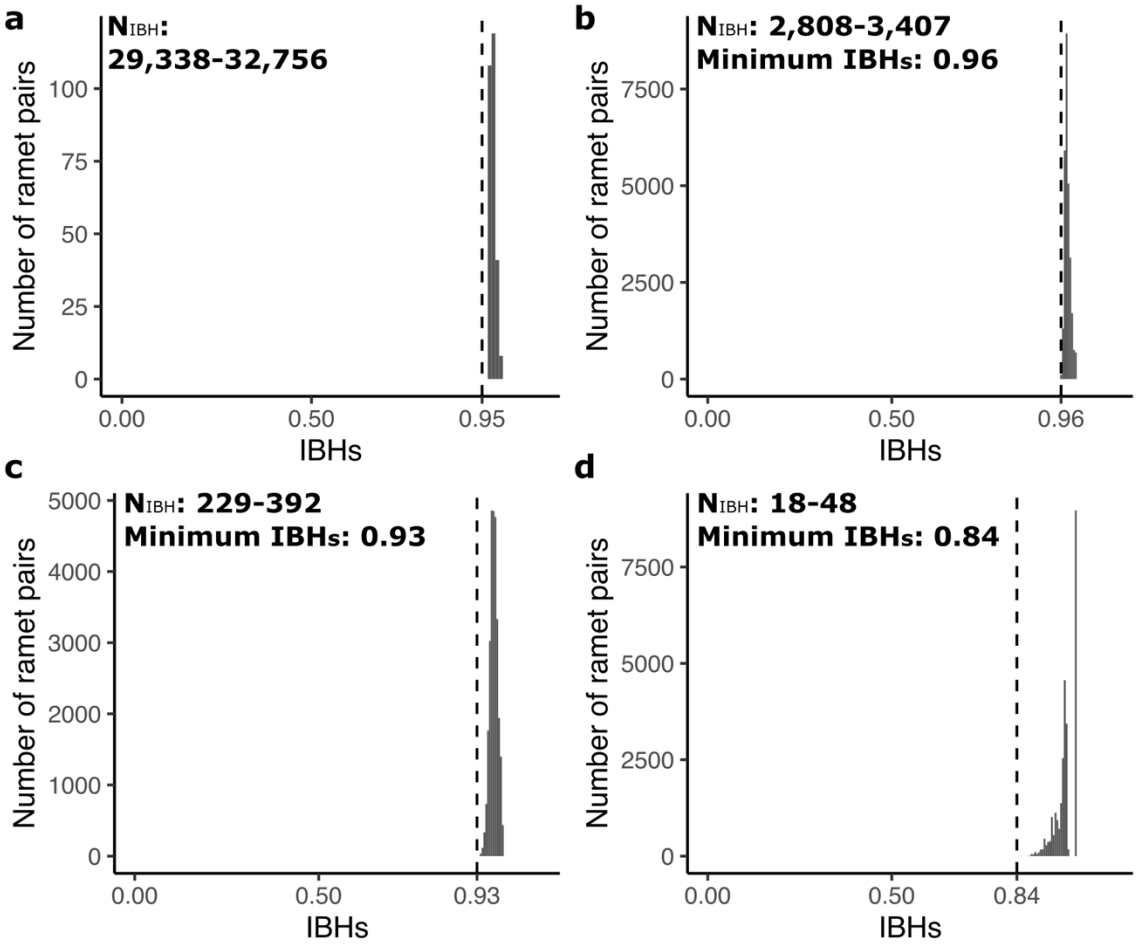


Figure 2: Frequency distribution of IBH similarity (IBH_S) for *Zostera marina* ramet pairs among the 24 ramets beloging to a clone of 750 to 1,500 years old. a, Frequency distribution based on all 38,831 SNPs. b, Frequency distribution based on 1/10 of the SNPs. c, Frequency distribution based on 1/100 of the SNPs. d, Frequency distribution based on 1/1000 of the SNPs.

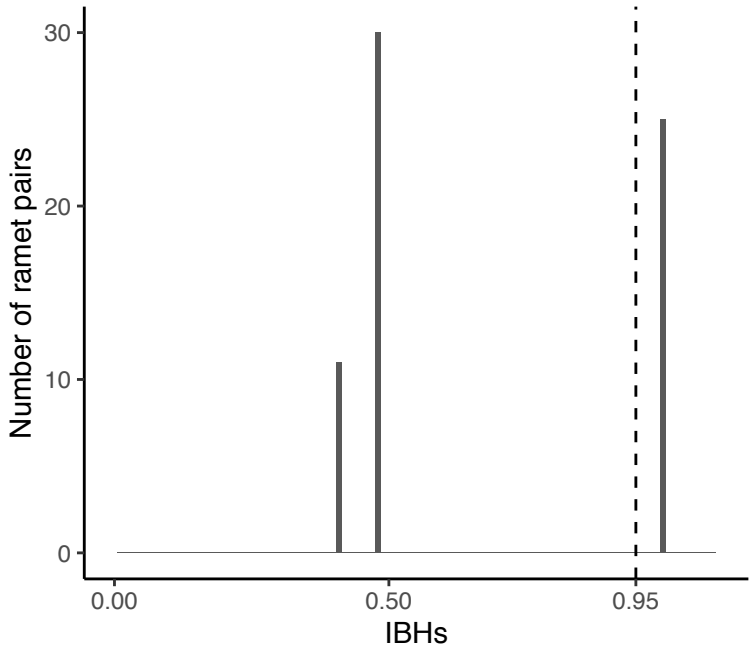


Figure 3: Frequency distribution of IBH similarity (IBH_s) for ramet pairs among the 12 cultured *Zostera marina* ramets.

Table 1: Coverage of the whole-genome resequencing data for the two cultured *Zostera marina* clones.

Sample	Average depth for the covered region	Genome region covered (%)
G01	68.64	93.33
G02	80.42	93.66
G06	79.05	93.72
G07	69.94	93.36
G10	77.49	93.52
G16	73.31	93.26
R02	81.55	93.86
R03	85.13	94.06
R05	68.14	93.57
R08	66.44	93.51
R09	71.73	93.59
R10	67.97	93.69

The whole-genome resequencing data of the 12 cultured *Z. marina* ramets covered >93% of the reference genome, and reached >66x coverage (Table 1). The pairwise N_{IBH} ranged from 108,585 to 288,606. All ramet pairs within the clone “GREEN” displayed IBH_s greater than 0.95 (Fig. 3). As for the clone “RED”, we detected one "contaminant" ramet (i.e., R10), obviously not

belonging to that clone. The five ramet pairs composed of R10 and one of the other five “RED” ramets showed IBH_s ranging from 0.4103 to 0.4114, indicating that R10 does not belong to this clone. Apart from R10, the ramet pairs among the other five “RED” ramets displayed IBH_s greater than 0.95 (Fig. 3). The ramet pairs composed of ramets from different clones showed IBH_s much smaller than 0.95, ranging from 0.4790 to 0.4803.

Discussion

Many species can propagate clonally to such an extent that the whole population is dominated by one or a few clones, such as fungi (Smith et al., 1992), seagrass species *Zostera marina* (Reusch et al., 1999) and *Posidonia oceanica* (Arnaud-Haond et al., 2012) and corals (Japaud et al., 2015). The situation of having a collection of samples belonging to one single clone (Yu et al., 2020) is thus not as rare as previously thought. The currently available methods cannot work in this extreme case (Arnaud-Haond et al., 2007). We have proposed the concept of identity by heterozygosity (IBH), meaning that two diploid genomes are identically heterozygous at some genetic markers. Based on IBH, we have developed IBH similarity (IBH_s), a novel measure of pairwise genetic similarity. The core of our IBH method for detecting clonemates is to use IBH_s in combination with the cutoff of 0.95 when N_{IBH} is on the order of >=3000. Two samples with IBH_s >0.95 are considered members of the same clone, while two samples with IBH_s <=0.95 are considered to represent different clones. Importantly, this method can be applied to any pair of ramets even if there is only one genet occurring within a "population"-level sampling. To the best of our knowledge, this is the only method that can work in cases where all collected ramets belong to one single clone.

One possible reason for the "contaminant" ramet (i.e., R10) in the clone “RED” is that R10 was produced by sexual event between members of the clone “RED” (i.e., selfing). In the first 15 years of growth, this would be the only way for contamination to occur, as the clones were cultured in isolated containers. Under selfing, each of the heterozygous loci in the clone “RED” can either remain identically heterozygous or change to a homozygous state in R10 with equal probabilities. As a consequence, the IBH_{R10} for the five ramet pairs composed of R10 and one of the other five “RED” ramets should be close to 1. However, that is not the case, as IBH_{R10} ranges from 0.4320 to 0.4331. Hence, R10 is not likely to be derived from selfing. In late 2019 or early 2020, these clones were transferred and planted in separate containers within in a single larger tank (~4 m diameter). All clones were in the same tank for at most one year. It is possible that either (a) outcrossing occurred and seeds set and germinated or (b) shoots dislodged and re-attached or were able to grow between pots.

The application of IBH_s requires a sufficient number of genetic markers where the two target samples are identically heterozygous (i.e., N_{IBH}). With the development of sequencing technology, it is now easy to obtain thousands of genetic markers using whole-genome resequencing, restriction-site associated DNA (RAD) sequencing, or SNP arrays. One should be

careful when N_{IBH} is small. For example, with N_{IBH} = 30, even 2 heterozygous genotypes caused by somatic mutations or genotyping errors will lead to IBH_s of 0.94, and thus the two samples would be wrongly categorized as different genets. We suggest this method will be reliable only when N_{IBH} is on the order of >=3,000, based on comparing our results from different subsets of SNPs (Fig. 2).

IBH can potentially be used to detect possible parent-descendant pairs under selfing, because this reproductive mode leads to a predictable loss of heterozygosity. Under selfing, on average half of the heterozygous loci in the parent will change to a homozygous state in the descendants. If not considering the ramet-specific heterozygous loci derived from mutations and genotyping errors, all the heterozygous loci in the descendant will be identically heterozygous in the parent, while there will be a number of loci that are only heterozygous in the parent. The predicted loss in heterozygosity can potentially be used to detect possible parent-descendant pairs under selfing.

Our IBH method takes advantage of the availability of the larger number of genetic markers in the genomic era, and extends the detection of clonemates to any pair of samples. It has filled the gap in detecting clonemates in extreme cases where all collected samples belong to one single clone.

Samping permit

The original sampling of *Zostera marina* clones was conducted under the permit no. SC-4874.

Data availability

DNA sequence data are available in the NCBI short read archive, BioProject no. PRJNA806459, SRA accession no. SRR18000159- SRR18000170.

Code availability

Custom-made computer code is available at Github https://github.com/leiyu37/Detecting-clonemates.

Acknowledgements

This study was supported by a 4-year PhD scholarship from the China Scholarship Council (CSC) to L.Y. (No. 201704910807). We thank D. Gill for lab assistance. We thank Y. Li for the comments on the early version of the manuscript.

Authors' contributions

L.Y. and T.B.H.R. conceived and designed the study, J.J.S and K.D. cultured and provided the samples of the two 17-year-old seagrass *Zostera marina* clones, L.Y. extracted DNA and analyzed the sequencing data, L.Y. performed the bioinformatic analyses, L.Y. and T.B.H.R. wrote the paper, with input from all authors.

Competing interests

The authors declare that they have no competing interests.

References

Arnaud-Haond, S., Duarte, C. M., Diaz-Almela, E., Marbà, N., Sintes, T. & Serrão, E. A. (2012). Implications of extreme life span in clonal organisms: millenary clones in meadows of the threatened seagrass *Posidonia oceanica*. *PLoS ONE, 7*(2), e30454. doi:10.1371/journal.pone.0030454

Arnaud-Haond, S., Duarte, C. M., Alberto, F. & Serrao, E. A. (2007). Standardizing methods to address clonality in population studies. *Mol. Ecol., 16*(24), 5115-5139. doi:10.1111/j.1365-294X.2007.03535.x

Bailleul, D., Stoeckel, S. & Arnaud-Haond, S. (2016). RClone: a package to identify MultiLocus Clonal Lineages and handle clonal data sets in r. *Methods Ecol. Evol., 7*(8), 966-970. doi:10.1111/2041-210X.12550

Buss, L. W. (1983). Evolution, development, and the units of selection. *Proc. Natl. Acad. Sci. U.S.A, 80*(5), 1387-1391. doi:10.1073/pnas.80.5.1387

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Douhovnikoff, V. & Dodd, R. S. (2003). Intra-clonal variation and a similarity threshold for identification of clones: application to Salix exigua using AFLP molecular markers. *Theor. Appl. Genet., 106*(7), 1307-1315. doi:10.1007/s00122-003-1200-9

Halkett, F., Simon, J. & Balloux, F. (2005). Tackling the population genetics of clonal and partially clonal organisms. *Trends Ecol. Evol. , 20*(4), 194-201. doi:10.1016/j.tree.2005.01.001

Harper, J. L. & White, J. (1974). The demography of plants. *Annu. Rev. Ecol. Syst., 5*(1), 419-463. doi:10.1146/annurev.es.05.110174.002223

Honnay, O. & Bossuyt, B. (2005). Prolonged clonal growth: escape route or route to extinction? *Oikos, 108*(2), 427-432. doi:10.1111/j.0030-1299.2005.13569.x

Hughes, R. A., Stachowicz, J. J. & Williams, S. L. (2009). Morphological and physiological variation among seagrass (*Zostera marina*) genotypes. *Oecologia, 159*(4), 725-733. doi:10.1007/s00442-008-1251-3

Japaud, A., Bouchon, C., Manceau, J. & Fauvelot, C. (2015). High clonality in *Acropora palmata* and *Acropora cervicornis* populations of Guadeloupe, French Lesser Antilles. *Mar. Freshw. Res., 66*(9), 847-851. doi:10.1071/MF14181

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987-2993. doi:10.1093/bioinformatics/btr509

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Ma, X., Olsen, J. L., Reusch, T. B. H., Procaccini, G., Kudrna, D., Williams, M., . . . Van de Peer, Y. (2021). Improved chromosome-level genome assembly and annotation of the seagrass, *Zostera marina* (eelgrass). *F1000research, 10*. doi:10.12688/f1000research.38156.1

Meirmans, P. G. (2020). genodive version 3.0: Easy-to-use software for the analysis of genetic data of diploids and polyploids. *Mol. Ecol. Resour., 20*(4), 1126-1131. doi:10.1111/1755-0998.13145

Noble, J. C., Bell, A. D. & Harper, J. L. (1979). The population biology of plants with clonal growth: I. The morphology and structural demography of *Carex arenaria*. *J. Ecol., 67*, 983-1008. doi:10.2307/2259224

Parks, J. C. & Werth, C. R. (1993). A study of spatial features of clones in a population of bracken fern, *Pteridium aquilinum* (Dennstaedtiaceae). *Am. J. Bot., 80*(5), 537-544. doi:10.1002/j.1537-2197.1993.tb13837.x

Quinlan, A. R. & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033

Reusch, T. B. H., Boström, C., Stam, W. T. & Olsen, J. L. (1999). An ancient eelgrass clone in the Baltic. *Mar. Ecol. Prog. Ser., 183*, 301-304. doi:10.3354/meps183301

Smith, M. L., Bruhn, J. N. & Anderson, J. B. (1992). The fungus *Armillaria bulbosa* is among the largest and oldest living organisms. *Nature, 356*(6368), 428-431. doi:10.1038/356428a0

Van der Auwera, G. A. & O'Connor, B. D. (2020). *Genomics in the cloud: using Docker, GATK, and WDL in Terra*: O'Reilly Media.

Yu, L., Boström, C., Franzenburg, S., Bayer, T., Dagan, T. & Reusch, T. B. H. (2020). Somatic genetic drift and multilevel selection in a clonal seagrass. *Nat. Ecol. Evol., 4*(7), 952-962. doi:10.1038/s41559-020-1196-4

Synthesis and perspective

Synthesis

Somatic genetic variation (SoGV) in clonal species

Clonal species, such as seagrass *Zostera marina*, have two levels of "population", the asexual population of ramets belonging to a single clone/genet, and the population of clones/genets, the "classical", sexually reproducing population. My thesis extended the knowledge on the multi-level genetic variation in the model seagrass species *Z. marina* using whole-genome resequencing. With manuscript I, I studied the somatic genetic variation (SoGV) in a *Z. marina* meadow tracing back to one single sexual event, and showed two levels of SoGV.

When a genotype is shared by all the cells within a ramet (i.e., fixed), the intra-ramet allele frequency should be 0.5 for heterozygotes, and 1 for homozygotes. The mosaic SoGV with different genotypes among cells are expected to show intermediate allele frequencies. Based on ultra-deep sequencing (1370x) of 3 ramets, I found high levels of mosaic SoGV. The only well-studied system which is comparable to my case is tumor. Neutral tumor evolution results in a power-law distribution in the histogram of variant read frequency (VRF) (Williams et al., 2016), and subclones are shown as additional peaks (Williams et al., 2018). In my case of *Z. marina*, the histogram of VRF showed approximate power-law distribution in the low-frequency part, together with multiple additional peaks. Similar to the case of tumor, this could indicate intra-ramet positive selection, which may favor a particular cell lineage possessing a driver mutation with which an entire cohort of neutral background mutations would "hitchhike". Another possible reason is the stratified architecture of the shoot apical meristem (SAM). In plants, the shoot apical meristem (SAM) contains the cells which can divide via mitosis to increase their number. The meristematic cells then differentiate to form different plant tissues. All SAMs examined thus far in flowering plants are organized into different cell populations or layers (Poethig, 1989; Pineda-Krch & Lehtilä, 2002; Klekowski, 2003). The layers are relatively isolated, and are responsible for the formation of different tissue types. The multiple "subclones" shown in the distribution of VRF may correspond to different layers in the SAM. This indicates that the stratified architecture of SAM is an important parameter, and should be taken into account in modeling work.

I also found 7,054 SoGV with different fixed genotypes among the 24 collected ramets. Previous studies have studied fixed SoGV among leaves of a single tree (Schmid-Siegert et al., 2017; Wang et al., 2019). This thesis was the first attempt to study fixed SoGV in clonal species. It has been overlooked that a mechanism is needed to turn mosaic SoGV into fixed SoGV. My thesis proposed that a stochastic process during the formation of new ramets/modules, *somatic genetic drift*, can lead to fixation of SoGV in the newly formed ramets. *Z. marina* can increase the number of ramets via the branching of the rhizome. The SAM on the tip of the rhizome divides into two SAMs during the branching process, with each new branch having its own SAM. Only a

small number of cells act as precursors to the SAM in the new branch. This bottleneck can change the relative proportion of different genotypes, eventually leading to a fixed genotype. Somatic genetic drift can explain the substantial amount of the fixed SoGV found here. In contrast to solitary trees, *Z. marina* can branch as long as space for expansion is available. Each branching, in turn, presents an opportunity for somatic genetic drift, eventually resulting in strong genetic heterogeneity among different ramets. Once a SoGV is fixed, it is equally important as the germline genetic variation, e.g., joining sexual cycle. This is an important source of genetic variation in clonal species, which has been ignored. Hence, efforts are needed to quantify the abundance of SoGV in clonal species, and how frequently they get fixed. To understand the fate of SoGV, modeling work on the evolutionary processes within and among ramets are needed. My thesis opened up avenues for the study of SoGV in clonal species which represent 66.5% of flora (Honnay & Bossuyt, 2005), and 40% of animal phyla (Buss, 1983).

Methodological issues for detecting low-frequency variants

One challenge in studying SoGV is to accurately quantify the intra-ramet allele frequency. Currently the frequency of the reads from next-generation sequencing is used as proxy for allele frequency. However, this is only solid when the sequencing coverage is high enough. Assuming that the library constructed for sequencing can accurately reflect the allele frequency, and the allele frequency of an allele in the sequencing library is f, the read frequency f_{read} which we can calculate based on sequencing data will be a sample from a binomial distribution with mean $\mu = f$ and s.d. $\sigma = \sqrt{f(1-f)/n}$ (Noorbakhsh & Chuang, 2017), given read depth n. When sequencing depth is low, for example $n = 50$, the s.d. σ of an allele with $f = 0.5$ will be 0.071, which is very large compared with the allele frequency f. If we increase the sequencing depth to, for example $n = 1000$, the s.d. σ of an allele with $f = 0.5$ will be 0.016, and the read frequency f_{read} can be a good proxy for allele frequency f. However, conducting whole-genome resequencing at >1000x coverage is very expensive. New technologies are needed to decrease the cost and increase the accuracy of obtaining intra-ramet allele frequency.

Complex worldwide evolutionary pathways detected by whole-genome resequencing

With manuscript II, I studied population genomics of *Z. marina*, based on the worldwide distribution of the third-level of genetic variation, i.e., germline genetic variation. This is the "classical" level of genetic variation considered in evolution and population genetics. My thesis extended the knowledge on seagrass evolution by upscaling to the worldwide population level. *Z. marina* is one of the rare plant species that is distributed throughout the entire Northern Hemisphere. Hence, I was particularly interested as to how *Z. marina* has spread from its origin to eventually colonize the entire North Pacific and North Atlantic. Olsen et al. (2004) conducted the most comprehensive population genetic study of *Z. marina*, based on nine microsatellite markers, one nuclear sequence and one chloroplast sequence. However, owing to the limitation of the markers used, the authors were not able to date the major evolutionary events.

A bifurcating tree is the traditional way to represent the relationships of species or populations. However, as it has recently become apparent, the evolutionary processes on population level are more complicated, and the evolutionary history may not be like a tree (Beddows et al., 2017; Marques et al., 2019; Marková et al., 2020). My results based on the whole-genome resequencing data indicated that *Z. marina* originated in Japan, and there have been swift and repeated trans-oceanic dispersal events. I was also able to estimate when these events happened. The three-cluster pattern of the chloroplast haplotype network seemed odd at first look, but it was actually consistent with the three trans-Pacific dispersal events. The multiple trans-Pacific dispersal events might be responsible for the admixture zone represented by Washington (WAS) and Bodega Bay (BB), but WAS and BB were fixed with different chloroplast haplotype states. My thesis suggested that one should never use chloroplast genome to reconstruct population-level phylogeny alone, but chloroplast haplotype network can complement nuclear phylogeny to interpret the evolutionary history, because chloroplast genome shows stepwise accumulation of genetic variation.

Interestingly, in the Pacific Ocean, the population located to the south of the Point Conception biogeographical boundary (San Diego, SD) formed a unique clade (Clade I) which was different from the other Pacific populations (Clade II), and the Atlantic side (Atlantic + Mediterranean Sea) was more closely related to the Clade II. This topology is consistent with the previously reported phylogeny based on ITS and *mat*K-intron sequences (see Fig. 1c in Olsen et al., 2004). Another interesting question is when *Z. marina* colonized the Atlantic side, and my estimate was 222 kya (95% HPD: 266.6 – 181.9 kya), consistent with the estimate based on ITS molecular clock. Using a clock of 0.8-2%/ Ma (Bakker et al., 1995), the ITS data in Olsen et al. (2004) supports the colonization of the Atlantic side to be 250-100 kya.

It is certainly no coincidence that the polyphyletic group of marine angiosperms (or seagrasses) only comprise 65 or so species (Larkum et al., 2018), although the recolonization of the marine environment happened at >70 mya (Larkum et al., 2018). My thesis proposed that frequent and swift colonization waves across entire ocean basins may explain a lack of speciation events across the seagrasses in general. My thesis also demonstrated the influence of historical glacial periods on *Z. marina* populations, and identified the refugia during the Last Glacial Maximum (LGM). All these findings are completely novel for marine plants. With the new set of seagrass genomes being completed right now, it would be interesting to see how other species have dispersed throughout their distributional area. My thesis provided a null hypothesis for the colonization history of seagrasses. Moreover, such a study on demographic history/phylogeography is also a prerequisite for making inferences on selection. Hence, my thesis served as foundation and opened up avenues for further research on adaptation to environmental gradients. For such questions, the model seagrass species *Z. marina* is particularly suitable, as populations grow in very different climate zones, water temperatures, light levels and pathogen pressure.

Detecting clonemates based on "identity by heterozygosity"

For the study of clonal species, it is necessary to accurately distinguish ramet vs. genet. The currently available methods cannot work on the case in manuscript I where all collected ramets belong to one single clone. The situation of having a collection of samples belonging to one single clone is not so rare as previously thought, because many species can propagate clonally to such an extent that the whole population is dominated by one or a few clones, such as fungi (Smith et al., 1992), seagrass species *Z. marina* (Reusch et al., 1999) and *Posidonia oceanica* (Arnaud-Haond et al., 2012) and corals (Japaud et al., 2015). With manuscript III, I provided a method that can distinguish whether any two samples are clonemates based on a relatively large number of maker loci. I expect that my method will be increasingly applied in situations where entire sites are comprised of a few or only one clonal lineage, in species ranging from fungi, colonial invertebrates such as corals, to plants.

Perspective

Within-clone evolutionary processes

Table 1: Similarities between genetic drift and somatic genetic drift.

	Genetic drift	Somatic genetic drift
Definition of allele frequency	Frequency of alleles among individuals within the population	Frequency of alleles among cells within the individual
Direct consequence	Fluctuation of allele frequency from one generation to the next generation of the population	Fluctuation of allele frequency from the parent to the clonally reproduced descendants
Ultimate consequence	Alleles fixed or lost within the population	Genotypes fixed or lost within the individual
Genome region affected	Whole linked genome as one single haplotype	Linked genome fragments
Underlying mechanism	Population size is not infinite	Clonally reproduced descendants are founded by a small number of precursor cells
Important parameters	Population size	Number of precursor cells, Number of proliferating cells within the individual

One of the key findings of chapter I using deep resequencing within a large seagrass clone was the discovery of somatic genetic drift. Somatic genetic drift is a process that stochastically changes the allele frequency among the population of cells from an individual to its clonally reproduced descendants, which is analogous to the concept of genetic drift in population genetics that stochastically changes the allele frequency among the population of sexually reproduced individuals from one generation to the next generation. Both somatic genetic drift and genetic drift are stochastic processes (Table 1). The key idea underlying genetic drift is that population

size is not infinite. Each individual is assumed to be a random combination of two alleles from the gene pool of the population. When the population size is infinite, the allele frequency remains stable across generations. Otherwise, the allele frequency will stochastically change. However, the key idea underlying somatic genetic drift is that a clonally reproduced individual is founded by a small number of precursor cells from the parent. The allele frequency among the precursor cells may be different from that among all the cells of the parent, which leads to the fluctuation of the allele frequency in the clonally reproduced descendants. Thus, the number of the precursor cells, and the number of the proliferating cells within the individual, are two important parameters for somatic genetic drift. One difference is that somatic genetic drift leads to the fixation of genotypes, while genetic drift leads to the fixation of alleles under recombination. Somatic mutation always generates one cell with a different genotype. Fixation via somatic genetic drift means that all the cells with a ramet/module share this novel genotype. Hence, for diploid species, a SoGV is usually fixed as heterozygote. On the contrary, fixation via genetic drift means that the population becomes monomorphic.

Future avenues to detect selection during somatic evolution

To understand the fate of SoGV in *Z. marina*, it is necessary to model the evolutionary processes within and among ramets. Some breakthroughs have been made for the system of cancer tumor (Williams et al., 2016; Williams et al., 2018). However, there are some fundamental differences between tumor and *Z. marina*. In *Z. marina*, shoot apical meristem (SAM) has the ability to divide to increase the number of cells, but SAM has a relatively fixed number of meristematic cells. In contrast, the tumor is a growing system, and the number of cells within the tumor keeps increasing. The stratified architecture of the SAM in *Z. marina* leads to another layer of complexity, as this may result in a pattern that is very similar to “subclones” derived from positive selection on cell lineages in tumor. Moreover, *Z. marina* has another level of genetic diversity among ramets tracing back to one single sexual event, owing to somatic genetic drift. Modeling work will be the foundation for further studying selection on intra-ramet and inter-ramet levels. To this end, more critical information on various process parameters is needed. For example, it is unknown to this date of how many dividing cells the meristem consists for most plants. It is also unknown as to how many cells are the foundation of a new ramet/module, which determines the extent of somatic genetic drift. These parameters need to be studied in the future.

The increasing importance of structural variation

In my dissertation, I have focused on SNPs as one major source of genetic variation, but it has become clear in recent years that structural variation (SV) might play an important role in evolution (Dorant et al., 2020; Mérot et al., 2020). SV can range from small insertion/deletion polymorphisms (indels) to entire duplicated genes or large inversions (Mérot et al., 2020). As an example of germline SV, Lamichhaney et al. (2016) showed that a 4.5-Mb inversion is responsible for alternative reproductive strategies in the ruff (*Philomachus pugnax*). As for somatic SV, Zhou et al. (2019) studied the evolutionary genomics of SVs in clonally propagated

grapevine cultivars and their outcrossing wild progenitors, and found that strong purifying selection acts against SVs but particularly against inversion and translocation events. My thesis was focused on SNPs, as the detection of SNPs is relatively easy and accurate. However, SVs would be a good next step to be included in population genomic studies, especially for the study of adaptation.

Conclusion

Population genetics has been focused on the evolution of germline genetic variation. This thesis demonstrated two levels of somatic genetic variation (SoGV) nested within "classical" germline genetic variation in clonal species, owing to somatic mutation and somatic genetic drift. The two levels of SoGV have long been predicted (Antolin & Strobeck, 1985; Gill et al., 1995; Otto & Orive, 1995; Orive, 2001; Pineda-Krch & Lehtilä, 2002; Klekowski, 2003), but this thesis was one of the first quantifying those. Somatic mutation leads to genetic mosaic within a single ramet/module. During the formation of new ramets/modules, the process of somatic genetic drift segregates mosaic SoGV into different fixed genotypes among ramets/modules (i.e., fixed SoGV). Some breakthroughs have been made to model the evolution of mosaic SoGV in the system of cancer tumor; however, the processes in clonal species are quite different. Moreover, the fixed SoGV seems to be restricted in tree species (Burian et al., 2016; Schmid-Siegert et al., 2017; Wang et al., 2019), but this thesis showed that clonal species may be able to accumulate fixed SoGV without limit, as long as the formation of new ramets/modules is not restricted. Hence, this thesis opened up new avenues of studying multi-level genetic variation by using clonal species as model. Further modeling and theoretical work based on clonal species will broaden the current population genetics theory.

References

Antolin, M. F. & Strobeck, C. (1985). The population genetics of somatic mutation in plants. *Am. Nat., 126*(1), 52-62. doi:10.1086/284395

Arnaud-Haond, S., Duarte, C. M., Diaz-Almela, E., Marbà, N., Sintes, T. & Serrão, E. A. (2012). Implications of extreme life span in clonal organisms: millenary clones in meadows of the threatened seagrass *Posidonia oceanica*. *PLoS ONE, 7*(2), e30454. doi:10.1371/journal.pone.0030454

Bakker, F. T., Olsen, J. L. & Stam, W. T. (1995). Evolution of nuclear rDNA ITS sequences in the *Cladophora albida/sericea* clade (Chlorophyta). *J. Mol. Evol., 40*(6), 640-651. doi:10.1007/BF00160512

Beddows, I., Reddy, A., Kloesges, T. & Rose, L. E. (2017). Population genomics in wild tomatoes—the interplay of divergence and admixture. *Genome Biol. Evol., 9*(11), 3023-3038. doi:10.1093/gbe/evx224

Burian, A., De Reuille, P. B. & Kuhlemeier, C. (2016). Patterns of stem cell divisions contribute to plant longevity. *Curr. Biol., 26*(11), 1385-1394. doi:10.1016/j.cub.2016.03.067

Buss, L. W. (1983). Evolution, development, and the units of selection. *Proc. Natl. Acad. Sci. U.S.A., 80*(5), 1387-1391. doi:10.1073/pnas.80.5.1387

Dorant, Y., Cayuela, H., Wellband, K., Laporte, M., Rougemont, Q., Mérot, C., . . . Bernatchez, L. (2020). Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species. *Mol. Ecol., 29*(24), 4765-4782. doi:10.1111/mec.15565

Gill, D. E., Chao, L., Perkins, S. L. & Wolf, J. B. (1995). Genetic mosaicism in plants and clonal animals. *Annu. Rev. Ecol. Syst., 26*(1), 423-444. doi:10.1146/annurev.es.26.110195.002231

Honnay, O. & Bossuyt, B. (2005). Prolonged clonal growth: escape route or route to extinction? *Oikos, 108*(2), 427-432. doi:10.1111/j.0030-1299.2005.13569.x

Japaud, A., Bouchon, C., Manceau, J. & Fauvelot, C. (2015). High clonality in *Acropora palmata* and *Acropora cervicornis* populations of Guadeloupe, French Lesser Antilles. *Mar. Freshw. Res., 66*(9), 847-851. doi:10.1071/MF14181

Klekowski, E. J. (2003). Plant clonality, mutation, diplontic selection and mutational meltdown. *Biol. J. Linn. Soc., 79*(1), 61-67. doi:10.1046/j.1095-8312.2003.00183.x

Lamichhaney, S., Fan, G., Widemo, F., Gunnarsson, U., Thalmann, D. S., Hoeppner, M. P., . . . Zhang, H. (2016). Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nat. Genet., 48*(1), 84-88. doi:10.1038/ng.3430

Larkum, A. W. D., Kendrick, G. A. & Ralph, P. J. (2018). *Seagrasses of Australia: Structure, Ecology and Conservation*: Springer.

Marková, S., Horníková, M., Lanier, H. C., Henttonen, H., Searle, J. B., Weider, L. J. & Kotlík, P. (2020). High genomic diversity in the bank vole at the northern apex of a range expansion: The role of multiple colonizations and end-glacial refugia. *Mol. Ecol., 29*(9), 1730-1744. doi:10.1111/mec.15427

Marques, D. A., Lucek, K., Sousa, V. C., Excoffier, L. & Seehausen, O. (2019). Admixture between old lineages facilitated contemporary ecological speciation in Lake Constance stickleback. *Nat. Commun., 10*(1), 1-14. doi:10.1038/s41467-019-12182-w

Mérot, C., Oomen, R. A., Tigano, A. & Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol. Evol., 35*(7), 561-572. doi:10.1016/j.tree.2020.03.002

Noorbakhsh, J. & Chuang, J. H. (2017). Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. *Nat. Genet., 49*(9), 1288-1289. doi:10.1038/ng.3876

Olsen, J. L., Stam, W. T., Coyer, J. A., Reusch, T. B., Billingham, M., Boström, C., . . . Lumiere, R. L. (2004). North Atlantic phylogeography and large-scale population differentiation of the seagrass *Zostera marina* L. *Mol. Ecol., 13*(7), 1923-1941. doi:10.1111/j.1365-294X.2004.02205.x

Orive, M. E. (2001). Somatic mutations in organisms with complex life histories. *Theor. Popul. Biol., 59*(3), 235-249. doi:10.1006/tpbi.2001.1515

Otto, S. P. & Orive, M. E. (1995). Evolutionary consequences of mutation and selection within an individual. *Genetics, 141*(3), 1173-1187. doi:10.1093/genetics/141.3.1173

Pineda-Krch, M. & Lehtilä, K. (2002). Cell lineage dynamics in stratified shoot apical meristems. *J. Theoret. Biol., 219*(4), 495-505. doi:10.1006/jtbi.2002.3139

Poethig, S. (1989). Genetic mosaics and cell lineage analysis in plants. *Trends Genet., 5*, 273-277. doi:10.1016/0168-9525(89)90101-7

Reusch, T. B. H., Boström, C., Stam, W. T. & Olsen, J. L. (1999). An ancient eelgrass clone in the Baltic. *Mar. Ecol. Prog. Ser., 183*, 301-304. doi:10.3354/meps183301

Schmid-Siegert, E., Sarkar, N., Iseli, C., Calderon, S., Gouhier-Darimont, C., Chrast, J., . . . Pagni, M. (2017). Low number of fixed somatic mutations in a long-lived oak tree. *Nat. Plants, 3*(12), 926-929. doi:10.1038/s41477-017-0066-9

Smith, M. L., Bruhn, J. N. & Anderson, J. B. (1992). The fungus *Armillaria bulbosa* is among the largest and oldest living organisms. *Nature, 356*(6368), 428-431. doi:10.1038/356428a0

Wang, L., Ji, Y., Hu, Y., Hu, H., Jia, X., Jiang, M., . . . Jia, Y. (2019). The architecture of intra-organism mutation rate variation in plants. *PLOS Biol., 17*(4), e3000191. doi:10.1371/journal.pbio.3000191

Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. *Nat. Genet., 48*(3), 238-244. doi:10.1038/ng.3489

Williams, M. J., Werner, B., Heide, T., Curtis, C., Barnes, C. P., Sottoriva, A. & Graham, T. A. (2018). Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet., 50*(6), 895-903. doi:10.1038/s41588-018-0128-6

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., . . . Gaut, B. S. (2019). The population genetics of structural variants in grapevine domestication. *Nat. Plants, 5*(9), 965-979. doi:10.1038/s41477-019-0507-8

Declaration of contribution

Chapter I: Yu, L., Boström, C., Franzenburg, S., Bayer, T., Dagan, T., & Reusch, T. B. (2020). Somatic genetic drift and multilevel selection in a clonal seagrass. Nature ecology & evolution, 4(7), 952-962.
Author contributions: T.B.H.R. and L.Y. designed the study, C.B. provided access to the biological samples and conducted the sampling, T.B. and T.D. assisted with the bioinformatic analyses, S.F. prepared the libraries and performed the re-sequencing, L.Y., T.D. and T.B.H.R. analyzed, interpreted and discussed the results, T.B.H.R. and L.Y. wrote the paper, with input from all authors.

Chapter II: Yu, L. et al. Frequent and swift trans-ocean dispersal events of a widespread marine angiosperm, the seagrass *Zostera marina* L. (manuscript).
Author contributions: J.J.S., J.S., J.L.O. and T.B.H.R. conceived and designed the study, M.K. analyzed the chloroplast data, L.Y., M.M. and A.H. conducted the phylogenetic analyses, A.H. identified the core genes, L.Y. calculated D-statistic with assistance from M.M., L.Y. conducted all the other analyses, T.D. assisted with bioinformatic analyses, L.Y., M.K., M.M. A.H., J.L.O., T.D. and T.B.H.R. discussed and interpreted the results, L.Y. and T.B.H.R. wrote the paper, with input from all authors.

Chapter III: Yu, L. et al. Using "identity by heterozygosity (IBH)" to detect clonemates under prevalent clonal reproduction in multicellular diploids (manuscript).
Author contributions: L.Y. and T.B.H.R. conceived and designed the study, J.J.S and K.D. cultured and provided the samples of the two 17-year-old seagrass *Zostera marina* clones, L.Y. extracted DNA and analyzed the sequencing data, L.Y. performed the bioinformatic analyses, L.Y. and T.B.H.R. wrote the paper, with input from all authors.

Curriculum Vitae

Personal Information
Name: Lei Yu
Date of birth: March 13th 1992
Place of birth: Qingdao, China
Nationality: Chinese

Education
October 2017 – present, PhD candidate in Marine Evolutionary Ecology, GEOMAR Helmholtz Center for Ocean Research Kiel (Kiel, Germany)
Thesis: "Multi-level genetic variation and somatic genetic drift in the model seagrass species, eelgrass (*Zostera marina* L.)"
Supervisor: Prof. Dr. Thorsten Reusch

September 2014 – July 2017, Master student
MSc in Marine Ecology
Institute of Oceanology of the Chinese Academy of Sciences, Qingdao, China
Supervisor: Prof. Dr. Jin-Xian Liu

September 2010 – June 2014, Bachelor student (BSc in Marine Biological Resources and Environment)
Ocean University of China, Qingdao, China

Oral Presentations
Yu L, Boström C, Franzenburg S, Bayer T, Dagan T, & Reusch TBH (2020). Somatic genetic drift and multilevel selection in a clonal seagrass. Talk at KPC mini-symposium, online.
Yu L, Boström C, Franzenburg S, Bayer T, Dagan T, Olsen JL & Reusch TBH (2020). Evolutionary biology of seagrass *Zostera marina*. Talk at IMPRS retreat, online.

Posters
Yu L, Boström C, Franzenburg S, Bayer T, Dagan T, & Reusch TBH (2019). Evolution by somatic genetic drift in species with modular organization. Poster at IMPRS retreat.
Yu L, Franzenburg S, Boström C, Bayer T, & Reusch TBH (2018). Somatic mutations are neglected in modular species. Poster at IMPRS retreat.

Workshops
Eukaryote genome annotation workshop 2019, Kiel, Germany.
Introduction to genome-wide association studies (GWAS) 2019, Kiel, Germany.

Acknowledgments

First and foremost, I would like to thank **Prof. Dr. Thorsten Reusch** for being my PhD supervisor. Under your supervision, I got useful suggestions whenever I had problems, and my independence and opinions were respected. You not only helped me to finish my PhD, but also taught me how to conceive my own ideas, and eventually to become an independent researcher. With your help, I really enjoyed my PhD research. Thanks a lot!

Thank my thesis advisory committee (TAC) members, Tal, Jeanine and Tobias, for the useful discussions and encouraging conversions.

Thank co-authors, Marina, Till, Adam and Micha, for the discussions and the insightful suggestions. Especially, thank Till for teaching me bioinformatics, which made my PhD much easier.

I would like to thank the whole Marine Evolutionary Ecology group. I learnt a lot from the lunch seminars. I would also like to thank Svend, Conny and Diana, for solving all the day-to-day problems.

Janina and Britta, thanks for your help during my first days in Kiel. Starting a new life in Kiel was not easy for me, and your help meant a lot to me.

Kechen, Yanan, and Guanlin, thanks for your excellent company. It was fantastic to drink beer with Kechen, and to play basketball with Yanan and Guanlin.

Henry, thanks for everything, e.g., postcards, relaxing conversations, cakes, chocolates and the plant that is still on my desk. I had a great time working in the same office with you.

Jamie, Kosmas and Melanie, thanks for the useful suggestions. Kwi, thanks for organizing the bioinformatic meetings, which I found very useful.

I would like to thank my parents for always encouraging me to pursue my scientific career. I thought about giving up, but your encouragements changed my mind.

Finally, **Xiaomei**, thanks for marrying me and giving birth to our son; **Niannian**, Thanks for joining our family.

Eidesstattliche Erklärung

Hiermit bestätige ich, Lei Yu, dass die folgende Dissertation

Multi-level genetic variation and somatic genetic drift in the model seagrass species, eelgrass (*Zostera marina* L.)

von mir, unter Beratung meiner Betreuer, selbstständig verfasst wurde, nach Inhalt und Form meine eigene Arbeit ist und keine weiteren Quellen und Hilfsmittel als die angegebenen verwendet wurden.

Die vorliegende Arbeit ist unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden und wurde weder im Rahmen eines Prüfungsverfahrens an anderer Stelle vorgelegt noch veröffentlicht. Veröffentlichte oder zur Veröffentlichung eingereichte Manuskripte wurden kenntlich gemacht. Mir wurde kein akadamischer Grad entzogen.

Ich erkläre mich einverstanden, dass diese Dissertation an die Bibliothek des GEOMAR Helmholtz Zentrum für Ozeanforschung Kiel und die Universitätsbibliothek der Christian-Albrechts-Universität zu Kiel weitergeleitet wird.

Kiel, March 2022

Lei Yu