Novel genomic resources for shelled pteropods: a draft genome and target capture probes for Limacina bulimoides, tested for cross-species relevance

Choo et al. BMC Genomics (2020) 21:11
https://doi.org/10.1186/s12864-019-6372-z

BMC Genomics

RESEARCH ARTICLE Open Access

Novel genomic resources for shelled pteropods: a draft genome and target capture probes for *Limacina bulimoides*, tested for cross-species relevance

Le Qin Choo[1,2*†], Thijs M. P. Bal[3†], Marvin Choquet[3], Irina Smolina[3], Paula Ramos-Silva[1], Ferdinand Marlétaz[4], Martina Kopp[3], Galice Hoarau[3] and Katja T. C. A. Peijnenburg[1,2*]

Abstract

Background: Pteropods are planktonic gastropods that are considered as bio-indicators to monitor impacts of ocean acidification on marine ecosystems. In order to gain insight into their adaptive potential to future environmental changes, it is critical to use adequate molecular tools to delimit species and population boundaries and to assess their genetic connectivity. We developed a set of target capture probes to investigate genetic variation across their large-sized genome using a population genomics approach. Target capture is less limited by DNA amount and quality than other genome-reduced representation protocols, and has the potential for application on closely related species based on probes designed from one species.

Results: We generated the first draft genome of a pteropod, *Limacina bulimoides*, resulting in a fragmented assembly of 2.9 Gbp. Using this assembly and a transcriptome as a reference, we designed a set of 2899 genome-wide target capture probes for *L. bulimoides*. The set of probes includes 2812 single copy nuclear targets, the 28S rDNA sequence, ten mitochondrial genes, 35 candidate biomineralisation genes, and 41 non-coding regions. The capture reaction performed with these probes was highly efficient with 97% of the targets recovered on the focal species. A total of 137,938 single nucleotide polymorphism markers were obtained from the captured sequences across a test panel of nine individuals. The probes set was also tested on four related species: *L. trochiformis*, *L. lesueurii*, *L. helicina*, and *Heliconoides inflatus*, showing an exponential decrease in capture efficiency with increased genetic distance from the focal species. Sixty-two targets were sufficiently conserved to be recovered consistently across all five species.

Conclusion: The target capture protocol used in this study was effective in capturing genome-wide variation in the focal species *L. bulimoides*, suitable for population genomic analyses, while providing insights into conserved genomic regions in related species. The present study provides new genomic resources for pteropods and supports the use of target capture-based protocols to efficiently characterise genomic variation in small non-model organisms with large genomes.

Keywords: Targeted sequencing, Exon capture, Genome, Non-model organism, Marine zooplankton

* Correspondence: leqin.choo@naturalis.nl; K.T.C.A.Peijnenburg@uva.nl
L.Q. CHOO and T.M.P. BAL are shared first authorship
†L. Q. Choo and T. M. P. Bal contributed equally to this work.
[1]Marine Biodiversity, Naturalis Biodiversity Center, Leiden, The Netherlands
Full list of author information is available at the end of the article

Background

Shelled pteropods are marine, holoplanktonic gastropods commonly known as 'sea butterflies', with body size ranging from a few millimetres (most species) to 1–2 cm [1]. They constitute an important part of the global marine zooplankton assemblage e.g. [2, 3] and are a dominant component of the zooplankton biomass in polar regions [4, 5]. Pteropods are also a key functional group in marine biogeochemical models because of their high abundance and dual role as planktonic consumers as well as calcifiers e.g. [6, 7]. Shelled pteropods are highly sensitive to dissolution under decreasing oceanic pH levels [2, 8, 9] because their shells are made of aragonite, an easily soluble form of calcium carbonate [10]. Hence, shelled pteropods may be the 'canaries in an oceanic coal mine', signalling the early effects of ocean acidification on marine organisms caused by anthropogenic releases of CO_2 [5, 11]. In spite of their vulnerability to ocean acidification and their important trophic and biogeochemical roles in the global marine ecosystem, little is known about their resilience towards changing conditions [5].

Given the large population sizes of marine zooplankton in general, including shelled pteropods, adaptive responses to even weak selective forces may be expected as the loss of variation due to genetic drift should be negligible [12]. Furthermore, the geographic scale over which gene flow occurs, between populations facing different environmental conditions, may influence their evolutionary potential [13] and consequently needs to be accounted for. It is thus crucial to use adequate molecular tools to delimit species and population boundaries in shelled pteropods.

So far, genetic connectivity studies in shelled pteropods have been limited to the use of single molecular markers. Analyses using the mitochondrial cytochrome oxidase subunit I (COI) and the nuclear 28S genes have revealed dispersal barriers at basin-wide scales in pteropod species belonging to the genera *Cuvierina* and *Diacavolinia* [14, 15]. For *Limacina helicina*, the Arctic and Antarctic populations were discovered to be separate species through differences in the COI gene [16, 17]. However, the use of a few molecular markers has often been insufficient to detect subtle patterns of population structure expected in high gene flow species such as marine fish and zooplankton [18–20]. In order to identify potential barriers to dispersal, we need to sample a large number of loci across the genome, which is possible due to recent developments in next-generation sequencing (NGS) technologies [21, 22].

Here, we chose a genome reduced-representation method to characterise genome-wide variation in pteropods because of their potentially large genome sizes and small amount of input DNA per individual. In species with large genomes, as reported for several zooplankton groups [20], whole genome sequencing may not be feasible for population-level studies. Reduced-representation methods can overcome the difficulty of sequencing numerous large genomes. Two common approaches are RADseq and target capture enrichment. RADseq [23], which involves the enzymatic fragmentation of genomic DNA followed by the selective sequencing of the regions flanking the restriction sites of the used enzyme(s), is attractive for non-model organisms as no prior knowledge of the genome is required. However, RADseq protocols require between 50 ng and 1 μg of high-quality DNA, with higher amounts being recommended for better performance [24], and has faced substantial challenges in other planktonic organisms e.g. [25, 26]. Furthermore, RADseq may not be cost efficient for species with large genomes [26]. Target capture enrichment [27–29] overcomes this limitation in DNA starting amount and quality, by using single-stranded DNA probes to selectively hybridise to specific genomic regions that are then recovered and sequenced [30]. It has been successfully tested on large genomes with just 10 ng of input DNA [31] as well as degraded DNA from museum specimens [32–35]. Additionally, the high sequencing coverage of targeted regions allows rare alleles to be detected [31].

Prior knowledge of the genome is required for probe design, however, this information is usually limited for non-model organisms. Currently, there is no pteropod genome available that can be used for the design of genome-wide target capture probes. The closest genome available is from the sister group of pteropods, Anaspidea (*Aplysia californica* (NCBI reference: PRJNA13635) [36]), but it is too distant to be a reference, as pteropods have diverged from other gastropods since at least the Late Cretaceous [37].

In this study, we designed target capture probes for the shelled pteropod *Limacina bulimoides* based on the method developed in Choquet et al. [26], to address population genomic questions using a genome-wide approach. We obtained the draft genome of *L. bulimoides* to develop a set of target capture probes, and tested the success of these probes through the number of single nucleotide polymorphisms (SNPs) recovered in the focal species. *L. bulimoides* was chosen as the probe-design species because it is an abundant species with a worldwide distribution across environmental gradients in subtropical and tropical oceans. The probes were also tested on four related species within the Limacinoidea superfamily (coiled-shell pteropods) to assess their cross-species effectiveness. Limacinoid pteropods have a high abundance and biomass in the world's oceans [2, 6, 37] and have been the focus of most ocean acidification research to date e.g. [2, 38, 39].

Results

Draft genome assembly

We obtained a draft genome of *L. bulimoides* (NCBI: SWLX00000000) from 108 Gb of Illumina data

sequenced as 357 million pairs of 150 base pair (bp) reads. As a first pass in assessing genomic data completeness, a k-mer spectrum analysis was done with JELLYFISH version 1.1.11 [40]. It did not show a clear coverage peak, making it difficult to estimate total genome size with the available sequencing data (Additional file 1: Appendix S1). Because distinguishing sequencing error from a coverage peak is difficult below 10-15x coverage, it is likely that the genome coverage is below 10-15x, suggesting a genome size of at least 6–7 Gb. The reads were assembled using the de novo assembler MaSuRCA [41] into 3.86 million contigs with a total assembly size of 2.9 Gbp (N50 = 851 bp, L50 = 1,059,429 contigs). The contigs were further assembled into 3.7 million scaffolds with a GC content of 34.08% (Table 1). Scaffolding resulted in a slight improvement, with an increase in the N50 to 893 bp and a decrease in the L50 to 994,289 contigs. Based on the hash of error corrected reads in MaSuRCA, the total haploid genome size was estimated at 4,801,432,459 bp (4.8 Gbp). Therefore, a predicted 60.4% of the complete genome was sequenced.

Genome completeness based on the assembled draft genome was measured in BUSCO version 3.0.1 [42] and resulted in the detection of 60.2% of near universal orthologues that were either completely or partially present in the draft genome of *L. bulimoides* (Table 2). This suggests that around 40% of gene information is missing or may be too divergent from the BUSCO sets [42]. Although the use of BUSCO on a fragmented genome may not give reliable estimates as orthologues may be partially represented within scaffolds that are too short for a positive gene prediction, this percentage of near-universal orthologues coincides with the estimate of genome size by MaSuRCA.

Table 1 Summary of draft genome statistics for *Limacina bulimoides*

Assembly statistics	Value
Estimated total genome size	4,801,432,559 bp
Total assembly size	2,901,932,435 bp
Number of scaffolds	
> = 0 bp	3,735,734
> = 1000 bp	802,059
> = 5000 bp	3890
> = 10,000 bp	116
> = 25,000 bp	6
> = 50,000 bp	3
N50	893 bp
L50	994,289
Smallest scaffold	200 bp
Largest scaffold	197,255 bp
Percentage of N's	0.3307
GC content, %	34.08

Table 2 Summary of BUSCO analysis showing the number of metazoan near universal orthologues that could be detected in the draft genome of *Limacina bulimoides*

	Present in draft genome
Complete	296 (30.3%)
Complete and single-copy	262 (26.8%)
Complete and duplicated	34 (3.5%)
Fragmented	292 (29.9%)
Missing	390 (39.8%)
Total BUSCO groups searched	978

We also compared the draft genome to a previously generated transcriptome of *L. bulimoides* (NCBI: SRR10527256) [43] to assess the completeness of the coding sequences and aid in the design of capture probes. The transcriptome consisted of 116,995 transcripts, with an N50 of 555 bp. Even though only ~ 60% of the genome was assembled, 79.8% (93,306) of the transcripts could be mapped onto it using the splice-aware mapper GMAP version 2017-05-03 [44]. About half of the transcripts (46,701 transcripts) had single mapping paths and the other half (46,605 transcripts) had multiple mapping paths. These multiple mapping paths are most likely due to the fragmentation of genes over at least two different scaffolds, but may also indicate multi-copy genes or transcripts with multiple spliced isoforms. Of the singly mapped transcripts, 8374 mapped to a scaffold that contained two or more distinct exons separated by introns. Across all the mapped transcripts, 73,719 were highly reliable with an identity score of 95% or higher.

Target capture probes design and efficiency

A set of 2899 genome-wide probes, ranging from 105 to 1095 bp, was designed for *L. bulimoides*. This includes 2812 single copy nuclear targets of which 643 targets were previously identifed as conserved pteropod orthologs [43], the 28S rDNA sequence, 10 known mitochondrial genes, 35 candidate biomineralisation genes [45, 46], and 41 randomly selected non-coding regions (see Methods). The set of probes worked very well on the focal species *L. bulimoides*. 97% (2822 of 2899 targets) of the targeted regions were recovered across a test panel of nine individuals (Table 3) with 137,938 SNPs (Table 4) identified across these targeted regions. Each SNP was present in at least 80% of *L. bulimoides* individuals (also referred to as genotyping rate) with a minimum read depth of 5x. Coverage was sufficiently high for SNP calling (Fig. 3) and 87% of the recovered targets (2446 of the 2822 targets) had a sequence depth of 15x or more across at least 90% of their bases (Fig. 1a). Of the 2822 targets, 643 targets

Table 3 Target capture efficiency statistics, averaged ± standard deviation across nine individuals, for each of five pteropod species, including raw reads, final mapped reads, % High Quality reads (reads mapping uniquely to the targets with proper pairs), % targets covered (percentage of bases across all targets covered by at least one read), average depth (sequencing depth across all targets with reads mapped)

Species	Raw reads (× 1,000)	Final mapped reads (× 1,000)	% HQ reads	% targets covered	Average depth
L. bulimoides	10,529 ± 3997	3531 ± 1548	33.23 ± 9.10	97.36 ± 0.42	250 ± 111
L. trochiformis	15,508 ± 4865	1765 ± 521	11.61 ± 2.59	20.32 ± 1.65	468 ± 144
L. lesueurii	7060 ± 2043	807 ± 196	11.93 ± 2.77	13.28 ± 1.96	431 ± 76.9
L. helicina	10,346 ± 6260	337 ± 180	3.47 ± 0.56	12.57 ± 2.71	63.7 ± 26.7
H. inflatus	3089 ± 1126	66 ± 30	2.07 ± 0.30	8.21 ± 3.34	31.9 ± 14.9

accounted for 50% of the total aligned reads in *L. bulimoides* (Additional file 1: Figure S2A in Appendix S2). For *L. bulimoides*, SNPs were found in all categories of targets, including candidate biomineralisation genes, non-coding regions, conserved pteropod orthologues, nuclear 28S and other coding sequences (Table 5). Of the 10 mitochondrial genes included in the capture, surprisingly, only the COI target was recovered.

The hybridisation of the probes and targeted re-sequencing worked much less efficiently on the four related species. The percentage of targets covered by sequenced reads ranged from 8.21% (83 out of 2899 targets) in *H. inflatus* to 20.32% (620 out of 2899 targets) in *L. trochiformis* (Table 3). Of these, only five (*H. inflatus*) to 42 (*L. trochiformis*) targets were covered with a minimum of 15x depth across 90% of the bases (Additional file 1: Table S1). The number of targets that accounted for 50% of the total aligned reads varied across species, with 4 of 620 targets for *L. trochiformis* that accounted for 50% of reads, 2 of 302 targets for *L. lesueurii*, 14 of 177 targets for *L. helicina* and 5 of 83 targets for *H. inflatus* (Additional file 1: Figure S2B-E in Appendix S2). In these four species, targeted regions corresponding to the nuclear 28S gene, conserved pteropod orthologues, mitochondrial genes and other coding sequences were obtained (Table 4). The number of mitochondrial targets recovered ranged between one and three: ATP6, COB, 16S were obtained for *L. trochiformis*, ATP6, COI for *L. lesueurii*, ATP6, COII, 16S for *L. helicina*, and only 16S for *H. inflatus*. Additionally, for *L. trochiformis*, seven biomineralisation candidates and four non-coding targeted regions were recovered. The number of SNPs ranged between 1371 (*H. inflatus*) and 12,165 SNPs (*L. trochiformis*) based on a gentoyping rate of 80% and a minimum read depth 5x (Table 5). The maximum depth for SNPs ranged from ~150x in *H. inflatus*, *L. helicina* and *L. lesueurii* to ~375x in *L. trochiformis* (Fig. 3). With less stringent filtering, such as a 50% genotyping rate, the total number of SNPs obtained per species could be increased (Table 5).

Across the five species of Limacinoidea, we found an exponential decrease in the efficiency of the targeted re-sequencing congruent with the genetic distance from the focal species *L. bulimoides*. Only 62 targets were found in common across all five species, comprising 14 conserved pteropod orthologues, 47 coding regions, and a 700 bp portion of the 28S nuclear gene. Based on the differences in profiles of number of SNPs per target and total number of SNPs, the hybridisation worked differently between the focal and non-focal species. In *L. bulimoides*, the median number of SNPs per target was 45, whereas in the remaining four species, most of the targets had only one SNP and the median number of SNPs per target was much lower: 11 for *L. trochiformis*, 10 for *L. lesueurii*, six for *L. helicina*, and seven for *H. inflatus*. The number of SNPs per target varied between one and more than 200 across the targets (Fig. 2). With an increase in genetic distance from *L. bulimoides*, the total number of SNPs obtained across the five shelled pteropod species decreased

Table 4 Number of single nucleotide polymorphism (SNPs) recovered after various filtering stages for five species of shelled pteropods. Hard-filtering was implemented in GATK3.8 VariantFiltration using the following settings: QualByDepth <2.0, FisherStrand >60.0, RMSMappingQuality <5.0, MQRankSumTest <-5.0 and ReadPositionRankSum <-5.0. The hard-filtered SNPs were subsequently filtered to keep those with a minimum site coverage of 5x and present in at least 80% of the individuals. Other filtering options were less stringent, such as a minimum depth of 2x and site presence in at least 50% of individuals

	Hard-filtering	80% individuals, 5x depth	80% individuals, 2x depth	50% individuals,5x depth
L. bulimoides	154,864	137,938	137,953	147,763
L. trochiformis	44,014	11,948	12,165	20,518
L. lesueurii	23,379	5359	5847	8487
L. helicina	18,298	2432	2771	4613
H. inflatus	13,041	1371	1559	2092

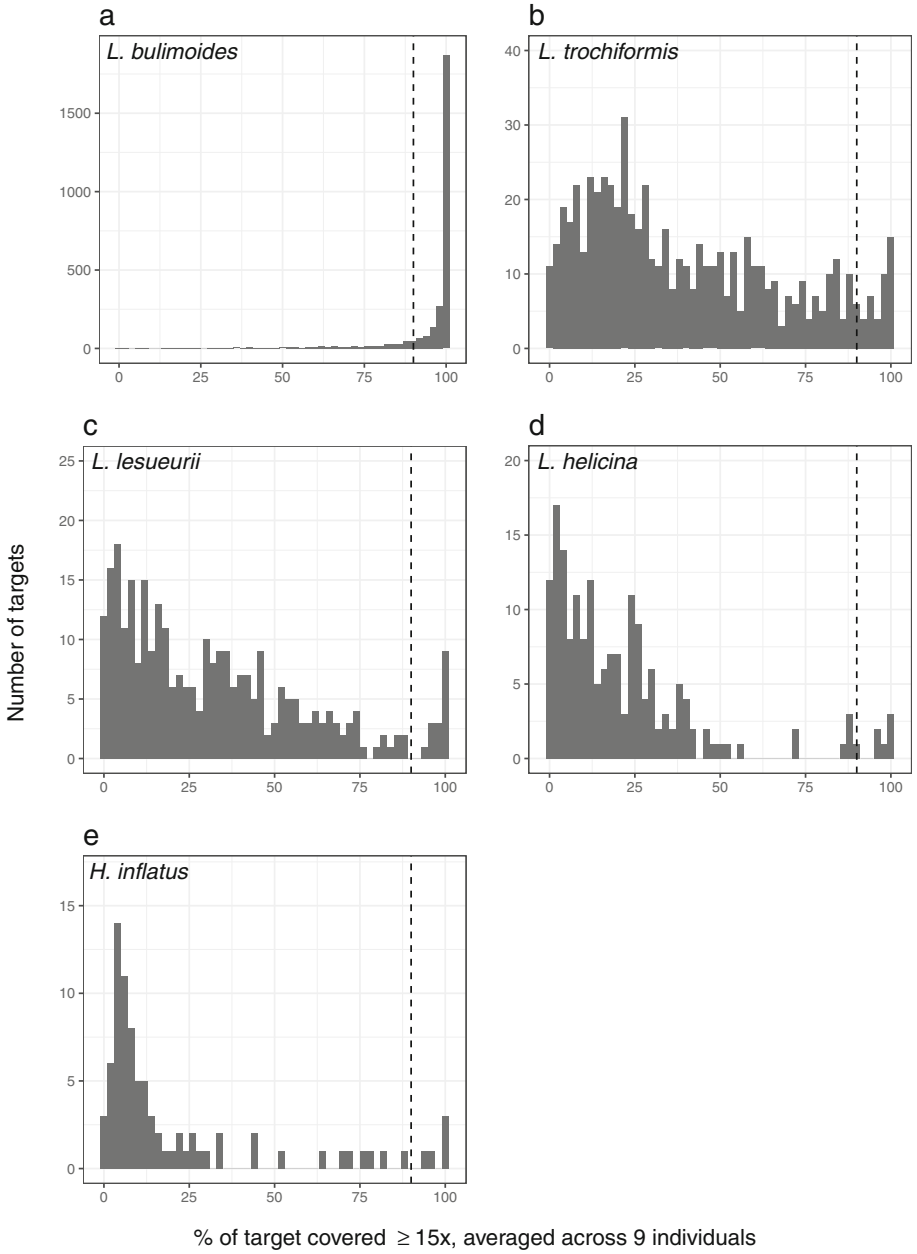

Fig. 1 Number of recovered targets plotted against average proportion of bases in each target, with at least 15x sequencing coverage averaged across nine individuals, for each for the five shelled pteropod species (**a**: *Limacina bulimoides*, **b**: *L. trochiformis*, **c**: *L. lesueurii*, **d**: *L. helicina*, and **e**: *Heliconoides inflatus*). Bars on the right of the dashed vertical line represent the number of targets where more than 90% of the bases in each target was sequenced with ≥15x depth. Note the differences in y-axes between the plots. There is no peak at one SNP for *L. bulimoides* (Additional file 1: Appendix S5)

exponentially (Fig. 4). There was an initial 10-fold decrease in number of SNPs between *L. bulimoides* and *L. trochiformis* with a maximum likelihood (ML) distance of 0.07 nucleotide substitutions per base between them. The subsequent decrease in number of SNPs was smaller in *L. lesueurii* (ML distance from *L. bulimoides*, subsequently ML dist = 0.11), *L. helicina* (ML dist = 0.18) and *H. inflatus* (ML dist = 0.29).

Discussion

First draft genome for pteropods

To assess the genetic variability and degree of population connectivity in coiled-shell pteropods, we designed a set of target capture probes based on partial genomic and transcriptomic resources. As a first step, we de novo assembled a draft genome for *L. bulimoides*, the first for a planktonic gastropod. We obtained an assembly size of

Table 5 Number of targets with at least one single nucleotide polymorphism (based on 80% genotyping rate, 5x depth) was calculated according to category: candidate biomineralisation genes (Biomin.), conserved pteropod orthologues (Ortholog.), mitochondrial (Mt genes), nuclear 28S, and other coding and non-coding regions for each of five pteropod species. Numbers in brackets represent the total number of targets in that category on the set of target probes designed for *Limacina bulimoides*

Species	Biomin. (35)	Ortholog. (643)	Mt genes (10)	28S (1)	Coding (2169)	Non-coding (41)	Total (2899)
L. bulimoides	32	635	1	1	2140	13	2822
L. trochiformis	7	169	3	1	436	4	620
L. lesueurii	0	90	2	1	209	0	302
L. helicina	0	52	3	1	121	0	177
H. inflatus	0	20	1	1	61	0	83

2.9 Gbp but the prediction of genome size together with the prediction of genome completeness suggest that only ~ 60% of the genome was sequenced. Therefore, we postulate that the genome size of *L. bulimoides* is indeed larger than the assembly size, and estimate it at 6–7 Gbp. In comparison, previously sequenced molluscan genomes have shown a wide variation in size across species, ranging from 412 Mbp in the giant owl limpet (*Lottia gigantea)* [47] to 2.7 Gbp in the Californian two-spot octopus (*Octopus bimaculoides)* [48]. The closest species to pteropods which has a sequenced genome is *Aplysia californica*, with a genome size of 927 Mbp (Genbank accession assembly: GCA_000002075.2) [36, 49]. Further, when considering marine gastropod genome size estimates in the Animal Genome Size Database [50], genome sizes range from 430 Mbp to 5.88 Gbp with an average size of 1.86 Gbp. Hence, it appears that *L. bulimoides* has a larger genome size than most other gastropods.

Despite moderate sequencing efforts, our genome is highly fragmented. Increasing the sequencing depth could result in some improvements, although other sequencing methods will be required to obtain a better genome. Roughly 350 million paired-end (PE) reads were used for the de novo assembly, but 50% of the assembly is still largely unresolved with fragments smaller than 893 bp. The absence of peaks in the k-mer distribution histogram and low mean coverage of the draft genome may indicate insufficient sequencing depth caused by a large total genome size, and/or high heterozygosity which complicates the assembly. In the 1.6 Gbp genome of another gastropod, the big-ear radix, *Radix auricularia*, approximately 70% of the content consisted of repeats [51]. As far as we know, high levels of repetitiveness within molluscan genomes are common [52], and also makes de novo assembly using only short reads challenging [53]. In order to overcome this challenge, genome sequencing projects should combine both short

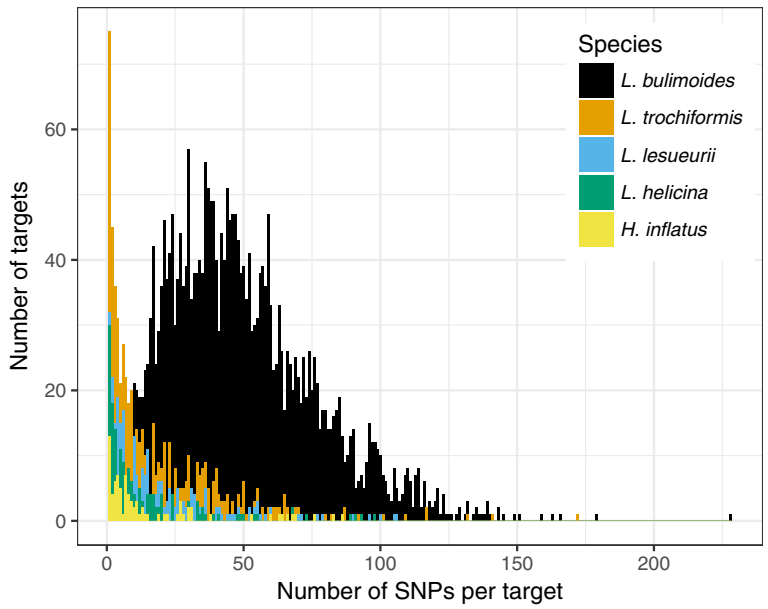

Fig. 2 Number of single nucleotide polymorphisms (SNPs) per recovered target for the five pteropod species of the superfamily Limacinoidea (see legend), based on filtering settings of minimum presence in 80% of individuals with at least 5x read depth

and long reads to resolve repetitive regions that span across short reads [54, 55]. Single molecule real time (SMRT) sequencing techniques which produce long reads recommend substantial DNA input, although some recent developments in library preparation techniques have lowered the required amount of DNA [56]. These SMRT techniques also tend to be high in cost, which may be a limiting factor when choosing between sequencing methods. Constant new developments in sequencing-related technologies may soon bring the tools needed to achieve proper genome assembly even for small-sized organisms with large genomes. Potential methods to improve current shotgun assemblies include 10x Genomics linked-reads [57] that uses microfluidics to leverage barcoded subpopulations of genomic DNA or Hi-C [58], which allow sequences in close physical proximity to be identified as linkage groups and enable less fragmented assemblies.

Target capture probes for *Limacina bulimoides*

Our results show that generating a draft genome and transcriptome to serve as a reference in the design of target capture probes is a promising and cost-effective approach to allow population genomics studies in non-model species of small sizes. Despite the relatively low N50 of the assembled genome, we were able to map 79.8% of the transcript sequences onto it. The combined use of the transcriptome and fragmented genome allowed us to identify the expressed genomic regions reliably and include intronic regions, which may have contributed to the probe hybridisation success [59]. In addition, the draft genome was useful in obtaining single-copy regions. This allowed us to filter out multi-copy regions at the probe design step, and hence reducing the number of non-target matches during the capture procedure.

The target capture was highly successful in the focal species *L. bulimoides*, with more than 130,000 SNPs recovered across nine individuals (Fig. 3). Coverage of reads across the recovered targets was somewhat variable (Additional file 1: Figure S2A in Appendix S2), although the SNPs were obtained from the large proportion of sufficiently well-covered targets (>15x, Table 4; Additional file 1: Table S1) and thus, can provide reliable genomic information for downstream analyses, such as delimiting population structure. The high number of SNPs may be indicative of high levels of genetic variation, congruent with predictions for marine zooplankton with large population sizes [12]. The number of SNPs recovered (Table 4) and percentage of properly paired reads mapping uniquely to the targets (Table 3) are comparable to the results from a similar protocol on copepods [26].

Targets corresponding to candidate biomineralisation genes and mitochondrial genes were less successfully recovered compared to conserved pteropod orthologues and other coding sequences (Table 4). This could be because biomineralisation-related gene families in molluscs are known to evolve rapidly, with modular proteins composed of repetitive, low complexity domains that are more likely to accumulate mutations due to unequal cross-over and replication slippage [60, 61]. Surprisingly, only the COI gene was recovered out of the 10 mitochondrial genes included in the set of probes. This is despite the theoretically higher per cell copy number of mitochondrial than nuclear genomes [62] and thus a higher expected coverage for mitochondrial targets compared to nuclear targets. High levels of mitochondrial polymorphism among individuals of *L. bulimoides* could have further complicated the capture, resulting in low capture success of mitochondrial targets. Hyperdiversity in mitochondrial genes, with more than 5% nucleotide diversity in synonymous sites has been reported for several animal clades, including gastropods [63, 64] and chaetognaths [65]. Only 13 of the 41 non-coding targeted regions were recovered, which may indicate that these regions were also too divergent to be captured by the probes.

Cross-species relevance of target capture probes

The success of targeted re-sequencing of the four related pteropod species (*L. trochiformis*, *L. lesueuri*, *L. helicina* and *Heliconoides inflatus*) decreased exponentially with increasing genetic distance from the focal species *L. bulimoides*. Even within the same genus, divergence was sufficiently high to show an abrupt decrease in coverage (Fig. 3). The number of targets whose reads accounted for 50% of all reads for each species was low (Additional file 1: Figure S2B-E in Appendix S2), indicating that representation across the targets could be highly uneven. The number of SNPs recovered also decreased rapidly with genetic distance (Fig. 4), leading to less informative sites across the genome that can be used in downstream analyses for these non-focal species. While direct comparisons are not possible due to differences in the probe design protocol and measurements used, we also see a decreasing trend in success of target capture applied with increasing levels of genetic divergence in other studies e.g. [66, 67]. Genetic divergence of 4–10% from the focal species resulted in an abrupt decline in coverage e.g. [62, 68]. Another possible reason for the decrease in capture success is different genome sizes across the species. While we used the same amount of DNA per individual in a capture reaction, pooling different species of unknown genome sizes into the same capture reaction may have resulted in different genome copy numbers sequenced per species. Our results may thus be attributed to high levels of polymorphism and/or possible differences in genome size, both leading to ascertainment bias [69].

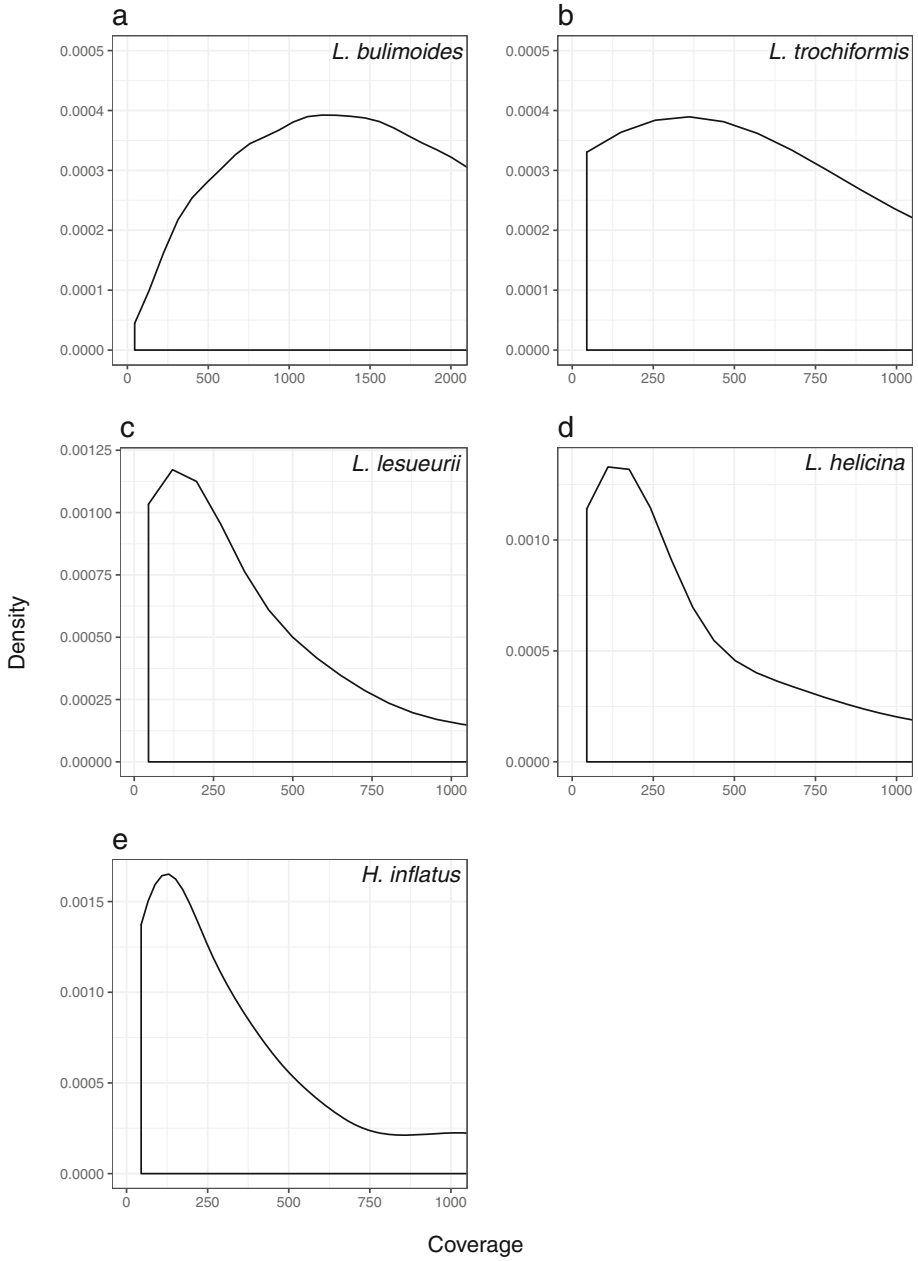

Fig. 3 Density of single nucleotide polymorphisms (SNPs, present in 80% of individuals) plotted against coverage for each of the five pteropod species (**a**: *Limacina bulimoides*, **b**: *L. trochiformis*, **c**: *L. lesueurii*, **d**: *L. helicina*, and **e**: *Heliconoides inflatus*). The plots were truncated at coverage = 2000x for *L. bulimoides* and coverage = 1000x for the other four species. Note that minimum coverage is 45x due to filtering settings of a minimum 5x depth for 9 individuals

The targets that hybridised successfully and were sequenced across species were conserved genes with low levels of genetic variation. This probably indicates that high levels of genetic diversity and divergence from the focal species resulted in the targeted regions not being able to hybridise to the probes. Indeed, from the four non-focal pteropod species, most of the recovered targets had low diversity, containing only a single SNP (Fig. 2). As a general rule, slowly evolving genomic regions are more likely to hybridise successfully to the probes [33, 70]. This may vary across targeted regions, as a mismatch tolerance of 40% between the baits and targeted region can still result in successful enrichment in specific cases [71]. While it is possible to design probes to be relevant across broader phylogenetic scales, by including conserved orthologues across the various target species e.g. [72, 73], these probes are unlikely to be suitable to study population structure and estimate

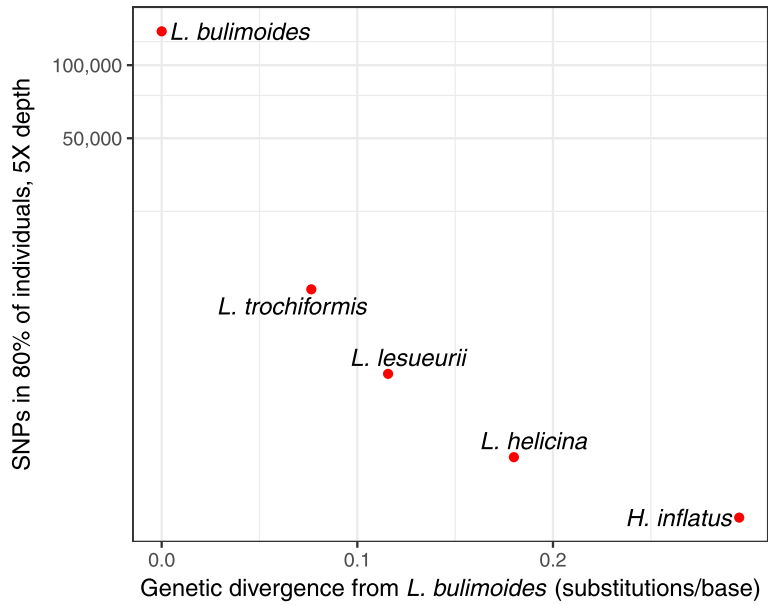

Fig. 4 Log-scaled number of SNPs against genetic divergence from the focal species *Limacina bulimoides* shows that there is a sharp reduction in the SNPs recovered with genetic distance

levels of gene flow in the focal species. Nonetheless, the low diversity targets that were recovered can be useful in resolving relationships at a deeper phylogenetic scale.

Conclusion

We show that using a combination of a draft genome and transcriptome is an efficient way to develop a database for capture probes design in species without prior genomic resources. These probes can be useful for analyses in closely related species, though cross-species hybridisation was limited to conserved targets and capture success decreased exponentially with increasing genetic distance from the focal species. Since the target capture approach can be successfully applied with low DNA input and even with poor quality or degraded DNA, this technique opens the door to population genomics of zooplankton, from recent as well as historical collections.

With more than 130,000 SNPs recovered in *L. bulimoides* and > 10,000 SNPs in *L. trochiformis*, our set of probes is suitable for genome-wide genotyping in these two globally distributed pteropod species. The high and consistent coverage across targeted genomic regions increases the range of analyses that can be applied to these organisms, such as identifying dispersal barriers, inferring ancestry and demographic history, and detecting signatures of selection across the genome. The statistical strength from analysing many genomic loci overcomes the limitation of an incomplete sampling of the metapopulation [74] and increases the capacity to detect even subtle patterns in population structure. This is especially relevant in widespread marine zooplankton where there is likely to be cryptic diversity and undiscovered species [12, 20], which is essential information for species that are proposed as indicators of ocean change.

Methods

Draft genome sequencing and assembly

A single adult *L. bulimoides* (1.27 mm total shell length) was used to generate a draft genome (NCBI: SWLX00000000). This individual was collected from the southern Atlantic subtropical gyre (25°44′S, 25°0′W) during the Atlantic Meridional Transect (AMT) cruise 22 in November 2012 (Additional file 1: Appendix S3 and Figure S3) and directly preserved in 95% ethanol at − 20 °C. Back in the lab, 147.2 ng of genomic DNA was extracted from the whole specimen using the E.Z.N.A. Insect DNA Kit (Omega Bio-Tek) with modifications to the manufacturer's protocol regarding reagents volumes and centrifuge times (Additional file 1: Appendix S3). The extracted DNA was randomly fragmented via sonication on a S220 Focused-ultrasonicator (Covaris) targeting a peak length of approximately 350 bp. A genomic DNA library was prepared using the NEXTflex Rapid Pre-Capture Combo Kit (Bioo Scientific) following the manufacturer's protocol. Subsequently, the library was sequenced in two runs of NextSeq500 (Illumina) using mid-output v2 chips producing 150 bp PE reads.

The resulting forward and reverse sequencing reads were concatenated in two separate files and quality-checked using FastQC version 0.11.4 [75]. Duplicated reads were removed using FastUniq version 0.11.5 [76]. The remaining reads were then assembled by the MaSuRCA genome assembler version 3.2.1 [41] using a k-

mer length of 105 as this produced the least fragmented assembly compared to other assemblers (Platanus, SOAPdenovo2). Further contig extension and scaffolding were carried out by running SSPACE-Basic version 2 [77] requiring a minimum of three linkers and a minimum overlap of 12 bp to merge adjacent contigs [77]. The total genome size was roughly estimated using MaSuRCA (as a by-product of calculating optimal assembly parameters), based on the size of the hash table containing all error corrected reads. A second estimate of the genome size was made by searching for k-mer peaks in sequencing reads using JELLYFISH version 1.1.11 [40] with various k-mer lengths between 15 and 101. To assess the completeness of the generated draft genome, the in-built BUSCO metazoan dataset containing 978 near-universal orthologues of 65 species was used to search for key orthologous genes with BUSCO version 3.0.1 [42]. BUSCO made use of AUGUSTUS version 3.3 [78] with the self-training mode utilised to predict gene models. Assembly quality was assessed with QUAST [79].

Target capture probes design

We designed the target capture probe set by using the draft genome and transcriptome as a reference, following the workflow recommended by Choquet et al. [26]. Firstly, we aimed to select only single-copy coding DNA sequences (CDS) in order to achieve a high specificity of the target capture probes and to reduce false-positive SNPs from multi-copy genes. We used the previously generated transcriptome of *L. bulimoides* [43] and mapped the transcript sequences of *L. bulimoides* against themselves using the splice-aware mapper GMAP version 2017-05-03 [44] with a k-mer length of 15 bp and no splicing allowed. Only unique transcripts with one mapping path were selected as potential target sequences. We then mapped these selected transcript sequences (with splicing allowed) directly to the contigs of the genomic assembly to identify expressed regions and their respective exon-intron boundaries. We selected only the subset of genomic sequences that mapped to unique transcripts with minimum pairwise identity scores of 90%. Using this approach, we selected 2169 coding target sequences. Additionally, 643 transcripts that mapped to unique contigs in the draft genome were selected from a set of conserved orthologues from a phylogenomic analysis of pteropods [43] to give a set of 2812 single copy coding nuclear targets. Of the 63 transcripts that showed homology to biomineralisation proteins [45, 46], we included 35 of these candidate biomineralisation genes in the final probe set as they could be mapped to contigs in the draft genome (Additional file 2).

Secondly, sequences of mitochondrial genes, 28S and non-coding targets were added to the baits design. A fragment of the COI gene (NCBI: MK642914), obtained by sanger sequencing as in [37] was added. The other nine targets (COII, COIII, ATP6, ND2, ND3, ND6, CYB, 12S, 16S) were identified from the draft genome assembly as described hereafter. We identified a 9039 bp contig from the fragmented assembly as a partially assembled mitochondrial genome using BLAST+ version 2.6.0 [80] and comparing the mitochondrial genes of three related mollusc species (NCBI Bioprojects: PRJNA10682, PRJNA11892, PRJNA12057) to the draft genome. Gene annotation was then carried out on this contig using the MITOS webserver [81] with the invertebrate genetic code and the parameters 'cut-off', 'fragment quality factor' and 'start/stop range' set to 30, 12 and 10, respectively. From this, we identified the seven protein-coding genes and the two rRNA genes as separate target sequences which we added to the probe design. Finally, we added the commonly-used nuclear 28S Sanger-sequenced fragment (NCBI: MK635470) and randomly chose 41 unique non-coding genomic regions. The final design comprised of 2899 target sequences with a total size of 1,866,005 bp. Probe manufacturing was performed by Arbor Biosciences (MI, USA) using myBaits custom biotinylated probes of 82-mer with 2x tiling density (Additional file 3).

Targeted sequencing of five pteropod species

We selected five shelled pteropod species from the genera *Limacina* and *Heliconoides* (superfamily Limacinoidea), including the focal species *L. bulimoides*, to evaluate the efficiency of the target capture probes on species of varying genetic relatedness. For each species, we aimed to test the capture efficiency across three sampling locations with three individuals per location (Table 6). Specimens from each species (*L. bulimoides*, *L. trochiformis*, *L. lesueurii*, *L. helicina*, *H. inflatus*) were collected across various sites during the AMT22 and AMT24 cruises in the Atlantic and from two sites in the Pacific Ocean (Table 6 and Additional file 1: Table S2). DNA was extracted from each individual separately using either E.Z.N.A. insect or mollusc kit (Omega Bio-Tek) with modifications to the protocol (Additional file 1: Appendix S3). The DNA was then sheared by sonication, using a Covaris S220 ultrasonicator with the peak length set to 300 bp. This fragmented DNA was used to prepare individual libraries indexed using the NEXTflex Rapid Pre-Capture Combo Kit (Bioo Scientific). Libraries were subsequently pooled into equimolar concentrations for the capture reaction using the myBaits Custom Target Capture kit (Arbor Biosciences). Hybridisation was carried out using the myBaits protocol with the following modifications. Twenty-seven libraries of *L. bulimoides* were pooled together for one capture reaction, of which nine individuals were analysed in this study. The other four species were pooled in groups of 22–23 specimens per capture. We extended the hybridisation time to 3 days and performed the whole protocol twice using 4 μL and 1.5 μL of probe mix, respectively (Additional file 1:

Table 6 Collection details of specimens from five shelled pteropod species: *Limacina bulimoides*, *L. trochiformis*, *L. lesueurii*, *L. helicina* and *Heliconoides inflatus*. Three individuals per site were included from localities in the Atlantic and Pacific Oceans. Latitude and longitude are presented in the decimal system, with positive values indicating North and East and negative values, South and West, respectively

Species	Location	Latitude	Longitude	n	Collection Date
L. bulimoides	South Atlantic	−18.32	−25.08	3	18/10/2014
L. bulimoides	South Atlantic	− 24.45	−25.05	3	21/10/2014
L. bulimoides	South Atlantic	−27.77	−25.02	3	22/10/2014
L. trochiformis	South Atlantic	−14.67	−25.07	3	17/10/2014
L. trochiformis	South Atlantic	−18.32	−25.08	3	18/10/2014
L. trochiformis	North Pacific	22.65	− 157.69	3	03/07/2017
L. lesueurii	North Atlantic	20.40	−38.61	3	24/10/2012
L. lesueurii	South Atlantic	−15.30	−25.07	3	05/11/2012
L. lesueurii	South Atlantic	−24.13	−25.00	3	09/11/2012
L. helicina	South Atlantic	−40.12	−30.92	3	26/10/2014
L. helicina	South Atlantic	−41.48	−33.87	3	27/10/2014
L. helicina	North Pacific	48.36	−126.31	3	06/03/2016
H. inflatus	North Atlantic	25.48	−39.00	3	22/10/2012
H. inflatus	South Atlantic	−8.08	−25.04	3	03/11/2012
H. inflatus	South Atlantic	−38.08	−39.31	3	16/11/2012

Appendix S3). The captured library of the species *L. bulimoides* was sequenced on the NextSeq500 (Illumina) using a high-output v2 chip producing 150 bp PE reads. The captured libraries of the other species were sequenced together on the same NextSeq500 mid-output v2 chip.

Assessment of target capture probes efficiency

The following pipeline of bioinformatic analyses was largely adapted from Choquet et al. [26]. Raw sequencing reads were de-multiplexed and mapped using BWA version 0.7.12 [82] with default settings to targets concatenated with the perl script concatFasta.pl [83]. The resulting BAM files were then cleaned and sorted using SAMtools version 1.4.1 [84] to retain only the reads paired and uniquely mapped in proper pairs. With Picard version 2.18.5 [85], duplicates were marked and removed. Coverage of targeted regions was assessed with the GATK version 3.8 [86] DepthOfCoverage tool. Next, SNP calling was performed using GATK version 3.8 with GNU Parallel [87] following the recommended Variant Discovery pipeline [88, 89] as a first trial for SNP calling in pteropods. Variants were called per individual using HaplotypeCaller with emitRefConfidence output, and the resulting gVCF files were combined according to their species with CombineGVCFs. The combined gVCF files for each species, with nine individuals each, were then genotyped in GenotypeGVCFs. SNPs were extracted from the raw variants with SelectVariants (–SelectType SNP). Given the lack of a calibration set of SNPs, the hard filters were first evaluated by plotting the density of annotation values and checking them against the planned filtering parameters. The SNPs were then hard-filtered with VariantFiltration using QualByDepth (QD) < 2.0, FisherStrand (FS) > 60.0, RMSMappingQuality < 5.0, MQRankSumTest (MQRankSum) < – 5.0, ReadPositionRankSum (ReadPosRankSum) < – 5.0 to retain reliable SNPs. The processed SNPs were further filtered using VCFtools version 0.1.13 [90] to keep those with a minimum coverage of 5x and represented in at least 80% of the individuals.

In order to investigate the relative effect of the different SNP filters, other less conservative VCFtools filtering settings such as a reduced genotyping rate of 50% or reduced depth requirement of 2x were used, and the relative increase in number of SNPs recovered for each species was recorded. For each species, the resulting VCF files were then annotated with the names and coordinates of the original targets using retabvcf.pl [83]. The targets represented in each species and the number of SNPs per target were then extracted from the annotated VCF files (Additional file 1: Appendix S4).

To assess the applicability of probes designed from *L. bulimoides* and other related pteropod species, the relationship between sequence divergence and number of SNPs recovered was investigated. The genetic divergence between *L. bulimoides* and each of the four other species was calculated from the branch lengths of a maximum likelihood (ML) phylogeny of pteropods based on transcriptome data [43]. The number of SNPs recovered per species using the most conservative filtering settings (80% genotyping rate and 5x depth) was plotted against sequence divergence from *L. bulimoides* in R [91].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-019-6372-z.

Additional file 1: Supporting information containing Appendices S1-5 and Tables S1-2.

Additional file 2: Transcripts of *L. bulimoides* with homology to biomineralisation proteins.

Additional file 3: 82-mer probe sequences for *L. bulimoides*.

Abbreviations

AMT: Atlantic Meridional Transect; CDS: Coding DNA Sequence; COI: Cytochrome Oxidase subunit I; ML: Maximum Likelihood; NGS: Next Generation Sequencing; PE: Paired End; SMRT: Single Molecule Real Time; SNP: Single Nucleotide Polymorphism

Acknowledgements

We thank Erica Goetze, Nina Bednaršek, Alice Burridge, and Lisette Mekkes as well as the captains and crews of the research cruises for support and assistance with collecting zooplankton samples.

Authors' contributions

LQC and TMPB contributed equally to the study design, molecular work, bioinformatic analyses and manuscript writing. KTCAP and GH contributed to analyses, writing, designed the study and supervised the project. KTCAP also

collected samples and acquired funding. MC contributed substantially to the study design, analyses and manuscript writing. MK and IS contributed to the study design and molecular work. FM and PR-S analysed sequence data and contributed to the capture design. All authors provided feedback and approved of the final manuscript.

Funding
This research was supported by the Netherlands Organisation for Scientific Research (NWO) Vidi grant 016.161.351 to K.T.C.A.P. Fieldwork was also supported by NSF grants OCE-1029478 and OCE-1338959 (to Erica Goetze). The Atlantic Meridional Transect program is supported by the UK NERC National Capability funding.

Availability of data and materials
The genomic assembly (NCBI accession: SWLX00000000, BioSample ID: SAMN11131519), and raw sequencing data of the target capture are available in NCBI Genbank, under BioProject PRJNA527191. The transcriptome is available in NCBI Genbank under the NCBI accession SRR10527256 (BioSample ID: SAMN13352221, BioProject: PRJNA591100). The list of *L. bulimoides* contigs with homology to biomineralisation proteins and set of 82-mer probes developed for *L. bulimoides* are included as Additional file 2 and Additional file 3. The additional information supporting the conclusions of this article are included as appendices within the Additional file 1.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Marine Biodiversity, Naturalis Biodiversity Center, Leiden, The Netherlands. [2]Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Amsterdam, The Netherlands. [3]Faculty of Biosciences and Aquaculture, Nord University, Bodø, Norway. [4]Molecular Genetics Unit, Okinawa Institute of Science and Technology, Onna-son, Japan.

Received: 24 September 2019 Accepted: 5 December 2019
Published online: 03 January 2020

References

1. Lalli CM, Gilmer RW. Pelagic snails: the biology of holoplanktonic gastropod molluscs. California: Stanford University Press; 1989.
2. Bednaršek N, Možina J, Vogt M, O'Brien C, Tarling GA. The global distribution of pteropods and their contribution to carbonate and carbon biomass in the modern ocean. Earth Syst Sci Data. 2012;4(1):167–86.
3. Burridge AK, Goetze E, Wall-Palmer D, Le Double SL, Huisman J, Peijnenburg KTCA. Diversity and abundance of pteropods and heteropods along a latitudinal gradient across the Atlantic Ocean. Prog Oceanogr. 2017;158: 213–23.
4. Hunt BPV, Pakhomov EA, Hosie GW, Siegel V, Ward P, Bernard K. Pteropods in Southern Ocean ecosystems. Prog Oceanogr. 2008;78(3):193–221.
5. Manno C, Bednaršek N, Tarling GA, Peck VL, Comeau S, Adhikari D, et al. Shelled pteropods in peril: assessing vulnerability in a high CO2 ocean. Earth Sci Rev. 2017;169:132–45.
6. Bé AWH, Gilmer R. A zoogeographic and taxonomic review of Euthecosomatous Pteropoda. Ocean Micropaleontol. 1977;1(6):733–808.
7. Buitenhuis ET, Le Quéré C, Bednaršek N, Schiebel R. Large contribution of Pteropods to shallow CaCO3 export. Glob Biogeochem Cycles. 2019;33(3): 458–68.
8. Lischka S, Büdenbender J, Boxhammer T, Riebesell U. Impact of ocean acidification and elevated temperatures on early juveniles of the polar shelled pteropod Limacina helicina: mortality, shell degradation, and shell growth. Biogeosciences. 2011;8(4):919–32.
9. Comeau S, Gattuso JP, Nisumaa AM, Orr J. Impact of aragonite saturation state changes on migratory pteropods. Proc R Soc B Biol Sci. 2012; 279(1729):732–8.
10. Mucci A. The solubility of calcite and aragonite in seawater at various salinities, temperatures and one atmosphere total pressure. Am J Sci. 1983; 283:780–99.
11. Bednaršek N, Klinger T, Harvey CJ, Weisberg S, McCabe RM, Feely RA, et al. New Ocean, new needs: application of pteropod shell dissolution as a biological indicator for marine resource management. Ecol Indic. 2017;76:240–4.
12. Peijnenburg KTCA, Goetze E. High evolutionary potential of marine zooplankton. Ecol Evol. 2013;3(8):2765–83.
13. Sanford E, Kelly MW. Local adaptation in marine invertebrates. Annu Rev Mar Sci. 2011;3:509–35.
14. Burridge AK, Goetze E, Raes N, Huisman J, Peijnenburg KTCA. Global biogeography and evolution of *Cuvierina* pteropods. BMC Evol Biol. 2015; 15(1):1–16.
15. Burridge AK, Van der Hulst R, Goetze E, Peijnenburg KTCA. Assessing species boundaries in the open sea: an integrative taxonomic approach to the pteropod genus *Diacavolinia*. Zool J Linnean Soc. 2019:1–25.
16. Hunt B, Strugnell J, Bednarsek N, Linse K, Nelson RJ, Pakhomov E, et al. Poles apart: the "bipolar" pteropod species *Limacina helicina* is genetically distinct between the Arctic and Antarctic oceans. PLoS One. 2010;5(3):4–7.
17. Sromek L, Lasota R, Wolowicz M. Impact of glaciations on genetic diversity of pelagic mollusks: Antarctic *Limacina antarctica* and Arctic *Limacina helicina*. Mar Ecol Prog Ser. 2015;525:143–52.
18. Gaggiotti OE, Bekkevold D, Jørgensen HBH, Foll M, Carvalho GR, Andre C, et al. Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. Evol Int J Org Evol. 2009;63(11):2939–51.
19. Waples RS. Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. J Hered. 1998;89(5):438–50.
20. Bucklin A, DiVito KR, Smolina I, Choquet M, Questel JM, Hoarau G, et al. Population genomics of marine zooplankton. In: Population Genomics. Cham: Springer; 2018.
21. De Wit P, Pespeni MH, Palumbi SR. SNP genotyping and population genomics from expressed sequences - current advances and future possibilities. Mol Ecol. 2015;24(10):2310–23.
22. McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol. 2013;66(2):526–38.
23. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008;3(10):1–7.
24. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17(2):81–92.
25. Deagle BE, Faux C, Kawaguchi S, Meyer B, Jarman SN. Antarctic krill population genomics: apparent panmixia, but genome complexity and large population size muddy the water. Mol Ecol. 2015;24(19):4943–59.
26. Choquet M, Smolina I, Dhanasiri AKS, Kopp M, Jueterbock A, Sundaram AYM, et al. Towards population genomics in non-model species with large genomes; a case study of the marine zooplankton *Calanus finmarchicus*. R Soc Open Sci. 2019:1–36.
27. Jones MR, Good JM. Targeted capture in evolutionary and ecological genomics. Mol Ecol. 2016;25(1):185–202.
28. Glenn TC, Faircloth BC. Capturing Darwin's dream. Mol Ecol Resour. 2016; 16(5):1051–8.
29. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7(2):111–8.
30. Gnirke A, Melnikov A, Maguire J, Rogov P, Leproust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009;27(2):182–9.
31. Chung J, Son D-S, Jeon H-J, Kim K-M, Park G, Ryu GH, et al. The minimal amount of starting DNA for Agilent's hybrid capture-based targeted massively parallel sequencing. Sci Rep. 2016;6(1):26732.
32. Kollias S, Poortvliet M, Smolina I, Hoarau G. Low cost sequencing of mitogenomes from museum samples using baits capture and ion torrent. Conserv Genet Resour. 2015;7(2):345–8.
33. Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C. Unlocking the vault: next-generation museum population genomics. Mol Ecol. 2013; 22(24):6018–32.
34. McCormack JE, Tsai WLE, Faircloth BC. Sequence capture of ultraconserved elements from bird museum specimens. Mol Ecol Resour. 2016;16(5):1189–203.

35. Blaimer BB, Lloyd MW, Guillory WX, Brady SG. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. PLoS One. 2016;11(8):1–20.
36. Broad Institute. *Aplysia* genome project. 2009. Available from: https://www.broadinstitute.org/aplysia/aplysia-genome-project. Accessed 28 Nov 2019.
37. Burridge AK, Hörnlein C, Janssen AW, Hughes M, Bush SL, Marlétaz F, et al. Time-calibrated molecular phylogeny of pteropods. PLoS One. 2017;12(6):1–22.
38. Maas AE, Lawson GL, Bergan AJ, Tarrant AM. Exposure to CO2 influences metabolism, calcification, and gene expression of the thecosome pteropod *Limacina retroversa*. J Exp Biol. 2018;221:jeb.164400.
39. Moya A, Howes EL, Lacoue-Labarthe T, Forêt S, Hanna B, Medina M, et al. Near-future pH conditions severely impact calcification, metabolism and the nervous system in the pteropod Heliconoides inflatus. Glob Chang Biol. 2016;22(12):3888–900.
40. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70.
41. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013;29(21):2669–77.
42. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
43. Peijnenburg KTCA, Janssen AW, Wall-Palmer D, Goetze E, Maas A, Todd JA, et al. The origin and diversification of pteropods predate past perturbations in the Earth's carbon cycle. bioRxiv preprint. 2019. https://doi.org/10.1101/813386.
44. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859–75.
45. Mann K, Jackson D. Characterization of the pigmented shell-forming proteome of the common grove snail *Cepaea nemoralis*. BMC Genomics. 2014;15(1):249.
46. Ramos-Silva P, Marin F. Proteins as functional units of biocalcification – an overview. Key Eng Mater. 2016;672:183–90.
47. Simakov O, Marletaz F, Cho S, Edsinger-Gonzales E, Havlak P, Hellsten U, et al. Insights into bilaterian evolution from three spiralian genomes. Nature. 2013;493(7433):526–31.
48. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, et al. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature. 2015;524(7564):220–4.
49. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2019;47(D1):D94–9.
50. Gregory TR. Animal Genome Size Database. 2019. Available from: http://www.genomesize.com. Accessed 28 Nov 2019.
51. Schell T, Feldmeyer B, Schmidt H, Greshake B, Tills O, Truebano M, et al. An annotated draft genome for Radix auricularia (Gastropoda, Mollusca). Genome Biol Evol. 2017;9(3):585–92.
52. Takeuchi T. Molluscan genomics: implications for biology and aquaculture. Curr Mol Biol Reports. 2017:1–9.
53. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012;13(1):36–46.
54. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012;30(7):693–700.
55. Rice ES, Green RE. New approaches for genome assembly and scaffolding. Annu Rev Anim Biosci. 2018;7(1):17–40.
56. Kingan SB, Heaton H, Cudini J, Lambert CC, Baybayan P, Galvin BD, et al. A High-Quality De Novo Genome Assembly from a Single Mosquito using PacBio Sequencing. Genes (Basel). 2019;10(1):62.
57. 10x Genomics. 2019. Available from: https://www.10xgenomics.com/. Accessed 28 Nov 2019.
58. Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi–C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58(3):268–76.
59. Suren H, Hodgins K, Yeaman S. NurkowskiKA, Smets P, Rieseberg L, et al. exome capture from the spruce and pine giga-genomes. Mol Ecol Resour. 2016;16:1136–46.
60. McDougall C, Degnan BM. The evolution of mollusc shells. Wiley Interdiscip Rev Dev Biol. 2018;7(3):1–13.
61. Kocot KM, Aguilera F, McDougall C, Jackson DJ, Degnan BM. Sea shell diversity and rapidly evolving secretomes: insights into the evolution of biomineralization. Front Zool. 2016;13(1):23.
62. Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics. 2012;13(1):403.
63. Thomaz D, Guiller A, Clarke B. Extreme divergence of mitochondrial DNA within species of pulmonate land snails. Proc R Soc B Biol Sci. 1996;263(1368):363–8.
64. Fourdrilis S, Mardulyn P, Hardy OJ, Jordaens K, de Frias Martins AM, Backeljau T. Mitochondrial DNA hyperdiversity and its potential causes in the marine periwinkle *Melarhaphe neritoides* (Mollusca: Gastropoda). PeerJ. 2016;4:e2549.
65. Marlétaz F, Le Parco Y, Liu S, Peijnenburg KTCA. Extreme Mitogenomic variation in natural populations of Chaetognaths. Genome Biol Evol. 2017;9(6):1374–84.
66. Förster DW, Bull JK, Lenz D, Autenrieth M, Paijmans JLA, Kraus RHS, et al. Targeted resequencing of coding DNA sequences for SNP discovery in nonmodel species. Mol Ecol Resour. 2018;18(6):1356–73.
67. Portik DM, Smith LL, Bi K. An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). Mol Ecol Resour. 2016;16(5):1069–83.
68. Bragg JG, Potter S, Bi K, Moritz C. Exon capture phylogenomics: efficacy across scales of divergence. Mol Ecol Resour. 2016;16(5):1059–68.
69. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. BioEssays. 2013;35(9):780–6.
70. Paijmans JLA, Fickel J, Courtiol A, Hofreiter M, Förster DW. Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. Mol Ecol Resour. 2016;16(1):42–55.
71. Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJP. Capturing protein-coding genes across highly divergent species. Biotechniques. 2013;54(6):321–6.
72. Quattrini AM, Faircloth BC, Dueñas LF, Bridge TCL, Brugler MR, Calixto-Botía IF, et al. Universal target-enrichment baits for anthozoan (Cnidaria) phylogenomics: new approaches to long-standing problems. Mol Ecol Resour. 2018;18(2):281–95.
73. Teasdale LC, Köhler F, Murray KD, O'Hara T, Moussalli A. Identification and qualification of 500 nuclear, single-copy, orthologous genes for the Eupulmonata (Gastropoda) using transcriptome sequencing and exon capture. Mol Ecol Resour. 2016;16(5):1107–23.
74. Maisano Delser P, Corrigan S, Hale M, Li C, Veuille M, Planes S, et al. Population genomics of *C. melanopterus* using target gene capture data: demographic inferences and conservation perspectives. Sci Rep. 2016;6:1–12.
75. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
76. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, et al. FastUniq: a fast de novo duplicates removal tool for paired short reads. PLoS One. 2012;7(12):1–6.
77. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27(4):578–9.
78. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33:465–7.
79. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5.
80. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421.
81. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. MITOS: Improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 2013;69(2):313–9.
82. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint. 2013;00(00):1–3.
83. Matz M. Genome-wide de novo genotyping with 2bRAD. 2019. Available from: https://github.com/z0on/2bRAD_denovo. Accessed 28 Nov 2019.
84. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
85. Broad Institute. Picard. 2019. Available from: http://broadinstitute.github.io/picard/. Accessed 28 Nov 2019.
86. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.
87. Tange O. GNU Parallel: the command-line power tool. The USENIX Magazine. 2011;36(1):42-7.
88. Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–501.

89. van der Auwera G, Carneiro MO, Hartl C, Poplin R, Angel G, Levy-Moonshine A, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:1–33.
90. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, Depristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
91. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for statistical. Computing. 2017; Available from: https://www.r-project.org/.

Publisher's Note