Genomic variability and population structure in shelled pteropods

Choo, L.Q.

Publication date
2022

Citation for published version (APA):
3

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


*Shared first authorship
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 2,899 genome-wide target capture probes for L. bulimoides. The set of probes includes 2,812 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

This chapter is published as

Chapter 3

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 centimetres (Lalli and Gilmer, 1989). They constitute an important part of the global marine zooplankton assemblage (e.g., Bednaršek et al., 2012a; Burridge et al., 2017a) and are a dominant component of the zooplankton biomass in polar regions (Hunt et al., 2008; Manno et al., 2017). 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., Bé and Gilmer, 1977; Buitenhus et al., 2019). Shelled pteropods are highly sensitive to dissolution under decreasing oceanic pH levels (Bednaršek et al., 2012a; Comeau et al., 2012; Lischka et al., 2011) because their shells are made of aragonite, an easily soluble form of calcium carbonate (Mucci, 1983). 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 CO₂ (Bednaršek et al., 2017b; Manno et al., 2017). 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 (Manno et al., 2017).

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 (Peijnenburg and Goetze, 2013). Furthermore, the geographic scale over which gene flow occurs, between populations facing different environmental conditions, may influence their evolutionary potential (Sanford and Kelly, 2011) 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 (Burridge et al., 2015, 2019). For Limacina helicina, the Arctic and Antarctic populations were discovered to be separate species through differences in the COI gene (Hunt et al., 2010; Sromek et al., 2015). 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 (Bucklin et al., 2018; Gaggiotti et al., 2009; Waples, 1998). 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 (De Wit et al., 2015; McCormack et al., 2013).

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,
Novel genomic resources for shelled pteropods

as reported for several zooplankton groups (Bucklin et al., 2018), 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 (Baird et al., 2008), 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 (Andrews et al., 2016), and has faced substantial challenges in other planktonic organisms (e.g., Choquet et al., 2019; Deagle et al., 2015). Furthermore, RADseq may not be cost efficient for species with large genomes (Choquet et al., 2019). Target capture enrichment (Glenn and Faircloth, 2016; Jones and Good, 2016; Mamanova et al., 2010) 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 (Gnirke et al., 2009). It has been successfully tested on large genomes with just 10 ng of input DNA (Chung et al., 2016) as well as degraded DNA from museum specimens (Bi et al., 2013; Blaime et al., 2016; Kollias et al., 2015; McCormack et al., 2016). Additionally, the high sequencing coverage of targeted regions allows rare alleles to be detected (Chung et al., 2016).

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: PRUNA13635; Broad Institute, 2009), but it is too distant to be a reference, as pteropods have diverged from other gastropods since at least the Late Cretaceous (Burridge et al., 2017b).

In this study, we designed target capture probes for the shelled pteropod *Limacina bulimoides* based on the method developed in Choquet et al. (2019), 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. Limacinoide pteropods have a high abundance and biomass in the world’s oceans (Bé and Gilmer, 1977; Bednaršek et al., 2012a; Burridge et al., 2017b) and have been the focus of most ocean acidification research to date (e.g., Bednaršek et al., 2012a; Maas et al., 2018; Moya et al., 2016).
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 (Marçais and Kingsford, 2011). It did not show a clear coverage peak, making it difficult to estimate total genome size with the available sequencing data (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 (Zimin et al., 2013) 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 (Simão et al., 2015) 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 (Simão et al., 2015). 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>
<thead>
<tr>
<th>TABLE 1</th>
<th>Summary of draft genome statistics for <em>Limacina bulimoides</em>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly statistics</td>
<td>Value</td>
</tr>
<tr>
<td>Estimated total genome size</td>
<td>4,801,432,559 bp</td>
</tr>
<tr>
<td>Total assembly size</td>
<td>2,901,932,435 bp</td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>Value</td>
</tr>
<tr>
<td>&gt;= 0 bp</td>
<td>3,735,734</td>
</tr>
<tr>
<td>&gt;= 1000 bp</td>
<td>802,059</td>
</tr>
<tr>
<td>&gt;= 5000 bp</td>
<td>3,890</td>
</tr>
<tr>
<td>&gt;= 10,000 bp</td>
<td>116</td>
</tr>
<tr>
<td>&gt;= 25,000 bp</td>
<td>6</td>
</tr>
<tr>
<td>&gt;= 50,000 bp</td>
<td>3</td>
</tr>
<tr>
<td>N50</td>
<td>893 bp</td>
</tr>
<tr>
<td>L50</td>
<td>994,289</td>
</tr>
<tr>
<td>Smallest scaffold</td>
<td>200 bp</td>
</tr>
<tr>
<td>Largest scaffold</td>
<td>197,255 bp</td>
</tr>
<tr>
<td>Percentage of N’s</td>
<td>0.3307</td>
</tr>
<tr>
<td>GC content, %</td>
<td>34.08</td>
</tr>
</tbody>
</table>
Novel genomic resources for shelled pteropods

We also compared the draft genome to a previously generated transcriptome of *L. bulimoides* (NCBI: SRR10527256) (Peijnenburg et al., 2020) 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 (Wu and Watanabe, 2005). 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, 8,374 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 2,899 genome-wide probes, ranging from 105 to 1,095 bp, was designed for *L. bulimoides*. This includes 2,812 single copy nuclear targets of which 643 targets were previously identified as conserved pteropod orthologs (Peijnenburg et al., 2020), the 28S rDNA sequence, 10 known mitochondrial genes, 35 candidate biomineralisation genes (Mann and Jackson, 2014; Ramos-Silva and Marin, 2016), and 41 randomly selected non-coding regions (see Methods). The set of probes worked very well on the focal species *L. bulimoides*. Of the targeted regions, 97% (2,822 of 2,899 targets) 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 ([Figure 3](#)) and 87% of the recovered targets (2,446 of the 2,822 targets) had a sequence depth of 15x or more across at least 90% of their bases ([Figure 1A](#)). Of the 2,822 targets, 643 targets accounted for 50% of the total aligned reads in *L. bulimoides* ([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...
Chapter 3

The diagrams in the figure depict the number of recovered targets plotted against the average proportion of bases in each target, with at least 15x sequencing coverage averaged across nine individuals, for each of the five shelled pteropod species (Limacina bulimoides, L. trochiformis, L. lesueurii, L. helicina, and 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 (APPENDIX S5).
Novel genomic resources for shelled pteropods

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).

<table>
<thead>
<tr>
<th>Species</th>
<th>Raw reads (x1,000)</th>
<th>Final mapped reads (x1,000)</th>
<th>% HQ reads</th>
<th>% targets covered</th>
<th>Average depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. bulimoides</td>
<td>10,529±4,997</td>
<td>3,531±1,548</td>
<td>32.2±9.10</td>
<td>97.3±0.42</td>
<td>250±111</td>
</tr>
<tr>
<td>L. trochiformis</td>
<td>15,508±4,865</td>
<td>1,765±521</td>
<td>11.6±2.59</td>
<td>20.3±1.65</td>
<td>468±144</td>
</tr>
<tr>
<td>L. lesueurii</td>
<td>7,060±2,043</td>
<td>807±196</td>
<td>11.9±2.77</td>
<td>13.2±1.96</td>
<td>431±76.9</td>
</tr>
<tr>
<td>L. helicina</td>
<td>10,346±6,260</td>
<td>337±180</td>
<td>3.4±0.56</td>
<td>12.5±2.71</td>
<td>63.7±26.7</td>
</tr>
<tr>
<td>H. inflatus</td>
<td>3,089±1,126</td>
<td>66±30</td>
<td>2.0±0.30</td>
<td>8.2±1.34</td>
<td>31.9±14.9</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Hard-filtering</th>
<th>80% individuals, 5x depth</th>
<th>80% individuals, 2x depth</th>
<th>50% individuals, 5x depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. bulimoides</td>
<td>154,864</td>
<td>137,938</td>
<td>137,953</td>
</tr>
<tr>
<td>L. trochiformis</td>
<td>44,014</td>
<td>11,948</td>
<td>12,165</td>
</tr>
<tr>
<td>L. lesueurii</td>
<td>23,379</td>
<td>5,359</td>
<td>5,847</td>
</tr>
<tr>
<td>L. helicina</td>
<td>18,298</td>
<td>2,432</td>
<td>2,771</td>
</tr>
<tr>
<td>H. inflatus</td>
<td>13,041</td>
<td>1,371</td>
<td>1,559</td>
</tr>
</tbody>
</table>

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 bimemeralisation genes (Biom.,) conserved pteropod orthologues, 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.

<table>
<thead>
<tr>
<th>Species</th>
<th>Biom. (35)</th>
<th>Orthologues (643)</th>
<th>Mt genes (10)</th>
<th>28S (1)</th>
<th>Coding (2,169)</th>
<th>Non-coding (41)</th>
<th>Total (2,899)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. bulimoides</td>
<td>32</td>
<td>635</td>
<td>1</td>
<td>1</td>
<td>2,140</td>
<td>13</td>
<td>2,822</td>
</tr>
<tr>
<td>L. trochiformis</td>
<td>7</td>
<td>169</td>
<td>3</td>
<td>1</td>
<td>436</td>
<td>4</td>
<td>620</td>
</tr>
<tr>
<td>L. lesueurii</td>
<td>0</td>
<td>90</td>
<td>2</td>
<td>1</td>
<td>209</td>
<td>0</td>
<td>302</td>
</tr>
<tr>
<td>L. helicina</td>
<td>0</td>
<td>52</td>
<td>3</td>
<td>1</td>
<td>121</td>
<td>0</td>
<td>177</td>
</tr>
<tr>
<td>H. inflatus</td>
<td>0</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>61</td>
<td>0</td>
<td>83</td>
</tr>
</tbody>
</table>

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 2,899 targets) in H. inflatus to
20.32% (620 out of 2,899 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 (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* (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 1,371 (*H. inflatus*) and 12,165 SNPs (*L. trochiformis*) based on a genotyping 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* (Figure 3). With less stringent filtering, such as a 50% genotyping rate, the total number of SNPs obtained per species could be increased (Table 5).

![Figure 2](image)

**Figure 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.
Novel genomic resources for shelled pteropods

Figure 3: Density of single nucleotide polymorphisms (SNPs, present in 80% of individuals) plotted against coverage. 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 nine individuals.
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 (Figure 2). With an increase in genetic distance from *L. bulimoides*, the total number of SNPs obtained across the five shelled pteropod species decreased exponentially (Figure 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).

![Figure 4 Log-scaled number of SNPs against genetic divergence from *L. bulimoides* (substitutions/base)](image-url)

Figure 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.
Novel genomic resources for shelled pteropods

**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 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*) (Simakov et al., 2013) to 2.7 Gbp in the Californian two-spot octopus (*Octopus bimaculoides*) (Albertin et al., 2015). 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) (Broad Institute, 2009; Sayers et al., 2019).

Further, when considering marine gastropod genome size estimates in the Animal Genome Size Database (Gregory, 2019), 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 (Schell et al., 2017). As far as we know, high levels of repetitiveness within molluscan genomes are common (Takeuchi, 2017), and also makes *de novo* assembly using only short reads challenging (Treangen and Salzberg, 2012). In order to overcome this challenge, genome sequencing projects should combine both short and long reads to resolve repetitive regions that span across short reads (Koren et al., 2012; Rice and Green, 2018). 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 (Kigan et al., 2018). 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 (10X Genomics, 2019) that uses microfluidics to leverage barcoded subpopulations of genomic DNA or Hi-C (Belton et al., 2012), 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 (Suren et al., 2016). 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 (Figure 3). Coverage of reads across the recovered targets was somewhat variable (Figure S2A in Appendix S2), although the SNPs were obtained from the large proportion of sufficiently well-covered targets (>15x, Table 4; 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 (Peijnenburg and Goetze, 2013). 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 (Choquet et al., 2019).

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 (Kocot et al., 2016; McDougall and Degnan, 2018). 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 (Bi et al., 2012) 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 (Fourdrilis et al., 2016; Thomaz et al., 1996) and chaetognaths (Marléz et al., 2017). 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 (Figure 3). The number of targets whose reads accounted for 50% of all reads for each species was low (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 (Figure 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., Förster et al., 2018; Portik et al., 2016). Genetic divergence of 4-10% from the focal species resulted in an abrupt decline in coverage (e.g., Bi et al., 2012; Bragg et al., 2016). 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 (Lachance and Tishkoff, 2013).

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 (Figure 2). As a general rule, slowly evolving genomic regions are more likely to hybridise successfully to the probes (Bi et al., 2013; Pajjmans et al., 2016). 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 (Li et al., 2013). 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., Quattrini et al., 2018; Teasdale et al., 2016), these probes are unlikely to be suitable to study population structure and estimate 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.
Chapter 3

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 (Maisano Delser et al., 2016) 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 (Bucklin et al., 2018; Peijnenburg and Goetze, 2013), which is essential information for species that are proposed as indicators of ocean change.

Materials and 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 (Appendix S3 and Table 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 centrifugation times (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 (Bio 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 (https://www.bioin-
formatics.babraham.ac.uk/projects/fastqc/). Duplicated reads were removed using FastUniq version 0.11.5 (Xu et al., 2012). The remaining reads were then assembled by the MaSuRCA genome assembler version 3.2.1 (Zimin et al., 2013) 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 (Boetzer et al., 2011) requiring a minimum of three linkers and a minimum overlap of 12 bp to merge adjacent contigs (Boetzer et al., 2011). 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 (Marçais and Kingsford, 2011) 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 (Simão et al., 2015). BUSCO made use of AUGUSTUS version 3.3 (Stanke and Morgenstern, 2005) with the self-training mode utilised to predict gene models. Assembly quality was assessed with QUAST (Gurevich et al., 2013).

**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. (2019). 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* (Peijnenburg et al., 2020) and mapped the transcript sequences of *L. bulimoides* against themselves using the splice-aware mapper GMAP version 2017-05-03 (Wu and Watanabe, 2005) 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 2,169 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 (Peijnenburg et al., 2020) to give a set of 2,812 single copy coding nuclear targets. Of the 63 transcripts that showed homology to biomeineralisation proteins (Mann and Jackson, 2014; Ramos-Silva and Marin, 2016), we included 35 of these candidate biomeineralisation genes in the final probe set as they could be mapped to contigs in the draft genome (SUPPORTING FILE 1).

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 (Burridge et al., 2017b) was added. The other nine targets (COI, COIII, ATP6, ND2, ND3, ND6, CYB, 12S, 16S) were identified from the draft genome assembly as described hereafter. We identified a 9,039 bp contig from the fragmented assembly as a partially assembled mitochondrial genome using BLAST+ version 2.6.0 (Camacho et al., 2009) 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 (Bernt et al., 2013) 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 2,899 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.

**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. bullimoides*, 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. bullimoides*, *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 5 and Table S2). DNA was extracted from each individual separately using either E.Z.N.A. insect or mollusc kit (Omega BioTek) with modifications to the protocol (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. 27 libraries of *L. bullimoides* 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 three days and performed the whole protocol twice using 4 µl and 1.5 µl of probe mix, respectively (Appendix S3). The captured library of the species *L. bullimoides* 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.
Novel genomic resources for shelled pteropods

**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.

<table>
<thead>
<tr>
<th>Species</th>
<th>Location</th>
<th>Latitude</th>
<th>Longitude</th>
<th>n</th>
<th>Collection date</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>L. bulimoides</em></td>
<td>South Atlantic</td>
<td>-18.32</td>
<td>-25.08</td>
<td>3</td>
<td>18/10/2014</td>
</tr>
<tr>
<td><em>L. bulimoides</em></td>
<td>South Atlantic</td>
<td>-24.45</td>
<td>-25.05</td>
<td>3</td>
<td>21/10/2014</td>
</tr>
<tr>
<td><em>L. bulimoides</em></td>
<td>South Atlantic</td>
<td>-27.77</td>
<td>-25.02</td>
<td>3</td>
<td>22/10/2014</td>
</tr>
<tr>
<td><em>L. trochiformis</em></td>
<td>South Atlantic</td>
<td>-14.67</td>
<td>-25.07</td>
<td>3</td>
<td>17/10/2014</td>
</tr>
<tr>
<td><em>L. trochiformis</em></td>
<td>South Atlantic</td>
<td>-18.32</td>
<td>-25.08</td>
<td>3</td>
<td>18/10/2014</td>
</tr>
<tr>
<td><em>L. trochiformis</em></td>
<td>North Pacific</td>
<td>22.65</td>
<td>-157.69</td>
<td>3</td>
<td>03/07/2017</td>
</tr>
<tr>
<td><em>L. lesueurii</em></td>
<td>North Atlantic</td>
<td>20.40</td>
<td>-38.61</td>
<td>3</td>
<td>24/10/2012</td>
</tr>
<tr>
<td><em>L. lesueurii</em></td>
<td>South Atlantic</td>
<td>-15.30</td>
<td>-25.07</td>
<td>3</td>
<td>05/11/2012</td>
</tr>
<tr>
<td><em>L. lesueurii</em></td>
<td>South Atlantic</td>
<td>-24.13</td>
<td>-25.00</td>
<td>3</td>
<td>09/11/2012</td>
</tr>
<tr>
<td><em>L. helicina</em></td>
<td>South Atlantic</td>
<td>-40.12</td>
<td>-30.92</td>
<td>3</td>
<td>26/10/2014</td>
</tr>
<tr>
<td><em>L. helicina</em></td>
<td>South Atlantic</td>
<td>-41.48</td>
<td>-33.87</td>
<td>3</td>
<td>27/10/2014</td>
</tr>
<tr>
<td><em>L. helicina</em></td>
<td>North Pacific</td>
<td>48.36</td>
<td>-126.31</td>
<td>3</td>
<td>06/03/2016</td>
</tr>
<tr>
<td><em>H. inflatus</em></td>
<td>North Atlantic</td>
<td>25.48</td>
<td>-39.00</td>
<td>3</td>
<td>22/10/2012</td>
</tr>
<tr>
<td><em>H. inflatus</em></td>
<td>South Atlantic</td>
<td>-8.08</td>
<td>-25.04</td>
<td>3</td>
<td>03/11/2012</td>
</tr>
<tr>
<td><em>H. inflatus</em></td>
<td>South Atlantic</td>
<td>-38.08</td>
<td>-38.31</td>
<td>3</td>
<td>16/11/2012</td>
</tr>
</tbody>
</table>

**Assessment of target capture probes efficiency**

The following pipeline of bioinformatic analyses was largely adapted from Choquet et al. (2019). Raw sequencing reads were de-multiplexed and mapped using BWA version 0.7.12 (Li, 2013) with default settings to targets concatenated with the perl script concatFasta.pl (Matz, 2019). The resulting BAM files were then cleaned and sorted using SAMTools version 1.4.1 (Li et al., 2009) to retain only the reads paired and uniquely mapped in proper pairs. With Picard version 2.18.5 (Broad Institute, 2019), duplicates were marked and removed. Coverage of targeted regions was assessed with the GATK version 3.8 (Mckenna et al., 2010) DepthOfCoverage tool. Next, SNP calling was performed using GATK version 3.8 with GNU Parallel (Tange, 2011) following the recommended Variant Discovery pipeline (Auwera et al., 2013; Depristo et al., 2011) 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, Fisher Strand (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 (Danecek et al., 2011) 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 VCF tools 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 (Matz, 2019). The targets represented in each species and the number of SNPs per target were then extracted from the annotated VCF files (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 (Peijnenburg et al., 2020). 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 (R Core Team, 2017).

**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 Supporting File 1 and 2 (https://doi.org/10.1186/s12864-019-6372-z). The additional information supporting the conclusions of this article are included as appendices within the Supplementary Information File.

**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.

**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.
Novel genomic resources for shelled pteropods

SUPPLEMENTARY MATERIAL

APPENDIX S1 DRAFT GENOME STATISTICS

Figure S1. Histogram of k-mer frequency distribution in the assembled Limacina bulmoides draft genome for k = 31. The x-axis represents the number of times a k-mer occurred and the y-axis represents the number of distinct k-mers for the given multiplicity.
APPENDIX S2 Density of coverage per target for each species

Figure S2 Density plot of coverage for each target, averaged across nine individuals, for each of the five shelled pteropod species (Limacina bulimoides, L. trochiformis, L. lesueurii, L. trochiformis, L. helicina and H. inflatus).
Novel genomic resources for shelled pteropods

APPENDIX S3 SPECIMEN COLLECTION AND MOLECULAR ANALYSES

Draft genome
The sequenced L. bulimoides individual, Lbul_AMT22_57_08 was collected during the Atlantic Meridional Transect 22 (AMT22) cruise in November 2012 in the southern gyre of the Atlantic Ocean (25°44'S, 25°0'W). The specimen was photographed in a standardised orientation using a Zeiss V20 stacking microscope (FIGURE S1).

Modified DNA extraction protocol with E.Z.N.A. Insect Kit (Omega Bio-tek) for extracting DNA from shelled pteropods
1. Soak the complete individual in nuclease-free water for 30 minutes.
2. Transfer individual into 2 ml screw cap grinding tube containing:
   - bashing beads
   - 400 μl CTL buffer
   - 28.5 μl proteinase K solution
3. Grind individual for 20 seconds at a frequency of 30 per second.
4. Incubate sample at 60 °C for 20 minutes.
5. Repeat step 3 and incubate sample at 60°C for 20 minutes.
6. Add 400 μl chloroform:isoamyl alcohol, vortex for 10 seconds.
7. Centrifuge at 13,500 g for 18 minutes at room temperature.
8. Prepare the HiBind DNA Mini Column:

Figure S3 Stacking microscopy photograph of the sequenced Limacina bulimoides specimen.

90
Chapter 3

9. Transfer upper aqueous phase (~320 μl) from step 7 to a new 1.5 ml eppendorf tube. Avoid the white milky surface.

10. Add one volume CBL buffer.

11. Add one volume 100% EtOH and vortex for 10 seconds.

12. Transfer 750 μl of the mixture to the equilibrated HiBind DNA Mini Column and centrifuge at 10,000 g for 1 minute. Discard the flow-through.

13. Repeat step 12 until all of the mixture has been used.

14. Warm-up the Elution Buffer to 70 °C.

15. Add 500 μl HBC buffer and centrifuge at 10,000 g for 30 seconds. Discard the flow-through.

16. Add 500 μl DNA Wash Buffer and centrifuge at 10,000 g for 1 minute. Discard the flow-through.

17. Repeat step 15.

18. Re-insert HiBind DNA Mini Column into new collection tube, centrifuge at 15,000 g for 2 minutes.


20. Add 40 μl of pre-heated Elution Buffer directly to the center of the Mini Column membrane without touching it.

21. Wait for 2 minutes and then centrifuge at 10,000 g for 1 minute.

22. Repeat steps 20 and 21 for a final elution volume of 80 μl.

Target capture protocol

All DNA libraries were fragmented to an average size of 300 bp by sonication and were prepared using the NEXTflex™ Rapid Pre-Capture Combo Kit (Bioo Scientific, Austin, TX, USA), including a step of single adapter indexing of each library. Libraries were clean-up, amplified separately for 8 and pooled in the following way: the nine L. bulimoides were part of a pooled capture that included 27 specimens of L. bulimoides in total. The four other species were amplified separately for another 8 cycles to increase the amount of DNA, and then combined into two pools that contained all four species with a total of 22-23 specimens per pool.

We increased the efficiency of the hybridisation and aimed to maximise the number of on-target captured sequences by doing the capture reaction twice, splitting the total amount of baits required for one reaction in two. The first round of hybridisation was performed using 4 μl of baits for each reaction. The reaction was performed over three days at a temperature of 60 °C in order to maximise the specificity. Capture was performed consecutively using DYNAbeads MyOne Streptavidin C1 beads (Invitrogen) to bind the hybridised targets during 30 min at 65 °C. The captured DNA was amplified by PCR for 8 cycles using KAPA HiFi HotStart
Novel genomic resources for shelled pteropods

ReadyMix (Kapa Biosystems). A second round of hybridisation was conducted using 1.5 μL of baits for each of the two pools, followed by a second capture and 6 more cycles of post-capture PCR. Finally, the two mixed-species pools were mixed together in equal proportions and sequenced on a NextSeq 550 (Illumina) with a 2x150 bp mid-output kit v.2. The *L. bulimoides* pool was mixed equally with two other *L. bulimoides* pools containing 27 individuals and sequenced on a NextSeq 550 (Illumina) with a 2x150 bp high-output kit v.2.
APPENDIX S4 COMMANDS USED FOR SNP CALLING FROM FASTQ FILES

For manipulating targets fasta file: Used concatFasta.pl https://github.com/z0on/2bRAD_denovo/blob/master/concatFasta.pl (concatenate fasta file into user-specified number of “chromosomes”) and retabvcf.pl from https://github.com/z0on/2bRAD_denovo/blob/master/retabvcf.pl (re-annotate .vcf file) to increase speed and reduce memory requirements. This is done before mapping reads to the targets.

(1) Use mapping.sh on demultiplexed, raw fastq.gz files

Contents of mapping.sh:

```bash
## mappin
## collect file names and place in file called names
ls *_R1.fastq.gz >names
sed -e s/_R1.fastq.gz//g -i names

## use parallel to run 24 threads of bwa mem, estimate two threads per specimen, for computer with 48 threads, producing aligned bam file *_aln.bam per specimen
cat names | parallel --verbose -j 24 "
bwa mem -M Genomic_targets_Limacina_bulimoides_cc.fasta {}_R1.fastq.gz {}_R2.fastq.gz | samtools view -Sbh -o {}_aln.bam"

## cleaning initial bam files
cat names | parallel --verbose -j 10 "
samtools view {}_aln.bam | fgrepXA | cut -f 1 > bad_names_{}.txt
samtools view -h {}_aln.bam | fgrep -vf bad_names_{}.txt | samtools view -Sb - > {}_aln2.bam
samtools view {}_aln2.bam | fgrepSA | cut -f 1 > bad_names_{}.txt
samtools view -h {}_aln2.bam | fgrep -vf bad_names_{}.txt | samtools view -Sb - > {}_aln3.bam
samtools view -b -f 3332 -f 3 {}_aln3.bam > {}_aln3_cleaned.bam"

## sort reads (3min)
chmod 755 {}_aln3_cleaned.bam
cat names | parallel --verbose -j 10 "
samtools sort {}_aln3_cleaned.bam -o {}_sorted.bam"

## mark duplicates
cat names | parallel --verbose -j 10 "
java -jar picard.jar MarkDuplicates I={}_sorted.bam O={}dedup.bam M={}dedup_metricsfile ASSUME_SORT_ORDER=coordinate VALIDATION_STRINGENCY=SILENT REMOVE_DUPLICATES=true &>Log_{}_dedup.txt"
```
Novel genomic resources for shelled pteropods

##Add or replace read groups, most important is RGSM which allows GATK to call
genotypes by individuals
cat names | parallel —verbose -j 15 “
java -jar picard.jar AddOrReplaceReadGroups I={}_dedup.bam
O={} _dedup_RG.bam SORT_ORDER=coordinate RGLB=lib1 RGPL=illumina
RGPU=unit54 RGSM={}”

##index .bam file
for file in * _RG.bam; do
samtools index $file ”${file}.bai”;
done
#end of mapping.sh

(2) The steps below follow Best Practices for GATK3.8., for
genotyping SNPs from bam files
##HaplotypeCaller, use gnu-parallel to use multiple cores
cat names | parallel —verbose -j 20 “java -jar GenomeAnalysisTK.jar
-R Genomic_targets_Limacina_bulimoides_cc.fasta
-T HaplotypeCaller 
-l {}_dedup_RG.bam 
—emitRefConfidence GVCF 
-o $src/gvcf/{}.g.vcf”

##CombineGVCFs- files are combined per population at this step, so parallel is no
longer needed.
java -jar GenomeAnalysisTK.jar
-R ../Genomic_targets_Limacina_bulimoides_cc.fasta
-T CombineGVCFs 
—variant Ltro.list 
-o Ltro.g.vcf

##GenotypeGVCFs
java -jar GenomeAnalysisTK.jar
-R ../Genomic_targets_Limacina_bulimoides_cc.fasta
-T GenotypeGVCFs 
—variant Ltro.g.vcf 
—max_alternate_alleles 18 
#depends on organism (diploid) and number of indi-
viduals (n=9)
-o ../combined/Ltro_raw_snps.vcf

##SelectVariants
for file in *.vcf; do

94
Chapter 3

java -jar GenomeAnalysisTK.jar \
   -T SelectVariants \ 
   -R ../Genomic_targets_Limacina_bulimoides_cc.fasta \
   -V $file\ 
   -selectType SNP \ 
   -o $(file%variants*).snps.vcf; done

##VariantFiltration
for file in *_raw_snps.vcf; do
   java -jar GenomeAnalysisTK.jar \
   -T VariantFiltration \ 
   -R ../Genomic_targets_Limacina_bulimoides_cc.fasta \ 
   -V $file \ 
   --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 50.0 || MQRankSum < -5.0 || ReadPosRankSum < -5.0" \ 
   --filterName "choquet2018" \ 
   -o $(file%raw*)filtered_snps.vcf &> $(file%raw*)filtered_snps.log ; done

(3) The steps below use vcftools to filter the snps within the vcf file and retabvcf.pl to rename coordinates of concatenated contigs with the names of the targets

##Filter SNPs
for file in *filtered_snps.vcf; do
   vcftools --vcf $file --max-missing 0.8 --min-meanDP 5.0 --recode --recode-INFO-all --out $(file%filtered*)cleaned_80_minDP5; done

##Rename “chromosomes” in vcf file with name of the targets
for file in *cleaned_80_minDP5.recode.vcf; do
   ~/2bRAD_denovo/retabvcf.pl vcf=$file tab=../Genomic_targets_Limacina_bulimoides_cc.tab >retab_${file}; done

##To count number of targets present in the vcf file
for file in retab_*; do
   awk ‘/CHROM/,0’ $file |cut -f1 |uniq -c > $(file%cleaned*)target_uniqc.list; done
Novel genomic resources for shelled pteropods

**APPENDIX S5 NUMBER OF SNPS PER TARGET (INDIVIDUAL PLOTS)**

![Graphs showing number of SNPs per target for individual pteropod species.](image)

**Figure S4** Number of SNPs per recovered target for each of the five shelled pteropod species (*Limacina bulimoides*, *L. trochiformis*, *L. lesueurii*, *L. helicina* and *Heliconoides inflatus*), based on filtering settings of minimum presence in 80% of individuals with at least 5x reads.

**Table S1** Number of targets with a minimum coverage of 15x in at least one base, across the five shelled pteropod species *Limacina bulimoides*, *L. trochiformis*, *L. lesueurii*, *L. helicina* and *Heliconoides inflatus*. The number of targets with more than 90% or 50%, and less than 10% of bases with 15x coverage is also displayed for each species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Total targets recovered at 15x depth or more</th>
<th>≥90% bases recovered with 15x depth or more</th>
<th>≥50% bases recovered with 15x depth or more</th>
<th>≤10% bases recovered with 15x depth or more</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>L. bulimoides</em></td>
<td>2822</td>
<td>2446</td>
<td>2768</td>
<td>5</td>
</tr>
<tr>
<td><em>L. trochiformis</em></td>
<td>620</td>
<td>42</td>
<td>206</td>
<td>88</td>
</tr>
<tr>
<td><em>L. lesueurii</em></td>
<td>302</td>
<td>16</td>
<td>72</td>
<td>77</td>
</tr>
<tr>
<td><em>L. helicina</em></td>
<td>177</td>
<td>6</td>
<td>15</td>
<td>64</td>
</tr>
<tr>
<td><em>H. inflatus</em></td>
<td>83</td>
<td>5</td>
<td>13</td>
<td>44</td>
</tr>
</tbody>
</table>
Table S2 Sampling localities and raw sequencing results per specimen.

<table>
<thead>
<tr>
<th>Individual</th>
<th>RMNH number</th>
<th>NCBI number</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Raw reads</th>
<th>Final mapped reads</th>
<th>% HQ reads</th>
<th>% targets covered</th>
<th>Depth</th>
</tr>
</thead>
</table>

// De magenta omlijning geeft de netto maat aan en zal niet zichtbaar zijn in het eindproduct //

Chapter 3
Novel genomic resources for shelled pteropods

<table>
<thead>
<tr>
<th>Individual</th>
<th>RMNH number</th>
<th>NCBI</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Raw reads</th>
<th>Final mapped reads</th>
<th>% HQ reads</th>
<th>% targets covered</th>
<th>Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lhe_CCE_123_06</td>
<td>RMNH.MOL.340273</td>
<td>SAMN11131500</td>
<td>-48.36</td>
<td>-126.31</td>
<td>1721637</td>
<td>70278</td>
<td>4.08</td>
<td>8.23</td>
<td>23.13</td>
</tr>
<tr>
<td>Hinf_AMT22_21_02</td>
<td>RMNH.MOL.340283</td>
<td>SAMN11131510</td>
<td>25.48</td>
<td>-39.00</td>
<td>3890761</td>
<td>92610</td>
<td>2.38</td>
<td>9.99</td>
<td>40.60</td>
</tr>
<tr>
<td>Hinf_AMT22_21_03</td>
<td>RMNH.MOL.340284</td>
<td>SAMN11131511</td>
<td>25.48</td>
<td>-39.00</td>
<td>3469925</td>
<td>55336</td>
<td>1.59</td>
<td>6.16</td>
<td>22.73</td>
</tr>
<tr>
<td>Hinf_AMT22_21_06</td>
<td>RMNH.MOL.340285</td>
<td>SAMN11131512</td>
<td>25.48</td>
<td>-39.00</td>
<td>1609269</td>
<td>31112</td>
<td>1.93</td>
<td>5.13</td>
<td>23.76</td>
</tr>
<tr>
<td>Hinf_AMT22_45_01</td>
<td>RMNH.MOL.340286</td>
<td>SAMN11131513</td>
<td>-8.08</td>
<td>-25.04</td>
<td>3726617</td>
<td>77294</td>
<td>2.07</td>
<td>12.80</td>
<td>34.87</td>
</tr>
<tr>
<td>Hinf_AMT22_45_02</td>
<td>RMNH.MOL.340287</td>
<td>SAMN11131514</td>
<td>-8.08</td>
<td>-25.04</td>
<td>3920742</td>
<td>91946</td>
<td>2.35</td>
<td>12.58</td>
<td>55.57</td>
</tr>
<tr>
<td>Hinf_AMT22_45_03</td>
<td>RMNH.MOL.340288</td>
<td>SAMN11131515</td>
<td>-8.08</td>
<td>-25.04</td>
<td>1256559</td>
<td>24332</td>
<td>1.94</td>
<td>4.16</td>
<td>13.01</td>
</tr>
<tr>
<td>Hinf_AMT22_66_01</td>
<td>RMNH.MOL.340289</td>
<td>SAMN11131516</td>
<td>-38.08</td>
<td>-39.31</td>
<td>4229048</td>
<td>107418</td>
<td>2.54</td>
<td>10.85</td>
<td>52.23</td>
</tr>
<tr>
<td>Hinf_AMT22_66_02</td>
<td>RMNH.MOL.340290</td>
<td>SAMN11131517</td>
<td>-38.08</td>
<td>-39.31</td>
<td>3653019</td>
<td>71554</td>
<td>1.96</td>
<td>6.28</td>
<td>22.93</td>
</tr>
<tr>
<td>Hinf_AMT22_66_04</td>
<td>RMNH.MOL.340291</td>
<td>SAMN11131518</td>
<td>-38.08</td>
<td>-39.31</td>
<td>2042221</td>
<td>38566</td>
<td>1.89</td>
<td>5.89</td>
<td>20.99</td>
</tr>
</tbody>
</table>