

1 **Identification and qualification of 500 nuclear, single-copy, orthologous genes for the**
2 **Eupulmonata (Gastropoda) using transcriptome sequencing and exon-capture**

3 Luisa C. Teasdale^{1,2}, Frank Köhler³, Kevin D. Murray⁴, Tim O'Hara¹, and Adnan Moussalli¹

4

5 ¹ Sciences Department, Museum Victoria, 11 Nicholson St, Carlton, Vic, Australia 3053

6 ² School of BioSciences, The University of Melbourne, Parkville, Vic, Australia 3010

7 ³ Australian Museum, 6 College Street, Sydney, NSW, Australia 2010

8 ⁴ Division of Plant Sciences, Research School of Biology, Australian National University,

9 Australia 0200

10

11 Corresponding author: Luisa Teasdale, Mailing address: Sciences Department, Museum

12 Victoria, 11 Nicholson St, Carlton, Vic, Australia 3053, Fax Number: + 61 3 8341 7442,

13 Email: lteasnail@gmail.com

14

15 Key words: orthology, single-copy, phylogenomics, Pulmonata, transcriptome, targeted

16 enrichment

17 Running header: Orthologous genes for Eupulmonata

18

19

20 **ABSTRACT**

21 The qualification of orthology is a significant challenge when developing large, multi-
22 loci phylogenetic datasets from assembled transcripts. Transcriptome assemblies have various
23 attributes, such as fragmentation, frameshifts, and mis-indexing, which pose problems to
24 automated methods of orthology assessment. Here, we identify a set of orthologous single-
25 copy genes from transcriptome assemblies for the land snails and slugs (Eupulmonata) using
26 a thorough approach to orthology determination involving manual alignment curation, gene
27 tree assessment and sequencing from genomic DNA. We qualified the orthology of 500
28 nuclear, protein coding genes from the transcriptome assemblies of 21 eupulmonate species
29 to produce the most complete gene data matrix for a major molluscan lineage to date, both in
30 terms of taxon and character completeness. Exon-capture targeting 490 of the 500 genes
31 (those with at least one exon > 120 bp) from 22 species of Australian Camaenidae
32 successfully captured sequences of 2,825 exons (representing all targeted genes), with only a
33 3.7% reduction in the data matrix due to the presence of putative paralogs or pseudogenes.
34 The automated pipeline Agalma retrieved the majority of the manually qualified 500 single-
35 copy gene set and identified a further 375 putative single-copy genes, although it failed to
36 account for fragmented transcripts resulting in lower data matrix completeness. This could
37 potentially explain the minor inconsistencies we observed in the supported topologies for the
38 21 eupulmonate species between the manually curated and Agalma-equivalent dataset
39 (sharing 458 genes). Overall, our study confirms the utility of the 500 gene set to resolve
40 phylogenetic relationships at a broad range of evolutionary depths, and highlights the
41 importance of addressing fragmentation at the homolog alignment stage for probe design.

42

43 INTRODUCTION

44 Robust and well resolved phylogenies document the evolutionary history of
45 organisms and are essential for understanding spatio-temporal patterns of phylogenetic
46 diversification and phenotypic evolution. Despite the central role of phylogenies in
47 evolutionary biology, most phylogenetic studies in non-model systems have relied on a
48 limited number of readily sequenced genes due to cost restrictions and availability of
49 phylogenetic markers. However, both theoretical and empirical studies have shown that a
50 greater number of independently evolving loci are needed to resolve difficult phylogenetic
51 questions (Gontcharov *et al.* 2004; Wortley *et al.* 2005; Leaché & Rannala 2011). This need
52 has been addressed by rapid advances in phylogenomics, which capitalise on high-throughput
53 sequencing to acquire large multi-loci datasets. In particular, both transcriptome sequencing
54 and targeted-enrichment strategies are increasingly being employed to reconstruct
55 phylogenetic relationships across a wide range of taxonomic levels (e.g. Bi *et al.* 2012;
56 Lemmon *et al.* 2012; Faircloth *et al.* 2012; Zapata *et al.* 2014; O’Hara *et al.* 2014; Misof *et*
57 *al.* 2014). A common aim of these studies, especially targeted enrichment based studies, has
58 been to identify universal sets of orthologous loci that can readily be captured and sequenced
59 across a broad taxonomic spectrum (e.g. Lemmon *et al.* 2012; Faircloth *et al.* 2012; Hugall *et*
60 *al.* 2015). Obtaining such universal sets of orthologous genes allows for consistency and
61 comparison across studies, and ultimately contributes towards a more comprehensive Tree of
62 Life (ToL) meta-analysis.

63 One of the greatest challenges associated with developing large, multi-loci
64 phylogenomic datasets is the qualification of orthology. In the context of phylogenetic
65 analysis, genes need to be orthologous and single-copy across all taxa under study (Fitch
66 2000; Philippe *et al.* 2011; Struck 2013). To this end, a number of automated pipelines have
67 been developed to identify single-copy orthologous genes from assembled transcriptomes.
68 These methods generally involve two main steps. The first step is to identify and cluster

69 homologous sequences, either by direct reference to annotated genomes (e.g., O’Hara *et al.*
70 2014) or by reference to ortholog databases, which themselves are derived from genome
71 comparisons (e.g., Tatusov *et al.* 2003; Ranwez *et al.* 2007; Waterhouse *et al.* 2013;
72 Altenhoff *et al.* 2015). Alternatively, non-reference methods have been employed such as all-
73 by-all and reciprocal BLAST comparisons (Li *et al.* 2003; Dunn *et al.* 2013) followed by
74 clustering (Enright *et al.* 2002). In the second step, orthology is qualified using either
75 similarity based approaches, including best-hit reciprocal blasts (Ebersberger *et al.* 2009;
76 Waterhouse *et al.* 2013; Ward & Moreno-Hagelsieb 2014), and/or tree based methods, where
77 gene trees are used to identify sequences with purely orthologous relationships (e.g., Agalma,
78 Dunn *et al.* 2013; PhyloTreePruner, Kocot *et al.* 2013; TreSpEx, Struck 2014).

79 Despite rapid advances in automated approaches to homolog clustering and qualifying
80 orthology, there are many characteristics of transcriptome assemblies that challenge such
81 automated methods. These include frameshifts, mis-indexing, transcript fragmentation and
82 the presence of multiple isoforms. Not accounting for these issues can lead to erroneous
83 inclusion of paralogous sequences and/or the inadvertent removal of appropriate orthologous
84 sequences (Martin & Burg 2002; Pirie *et al.* 2007; Philippe *et al.* 2011). To address these
85 issues O’Hara *et al.* (2014) placed greater emphasis on careful manual curation and editing of
86 homolog alignments prior to orthology qualification. A key aspect of this approach was the
87 concatenation of transcript fragments into a single consensus sequence prior to tree-based
88 ortholog qualification, leading to a more complete final data matrix. This, in turn, allowed a
89 more robust probe design for subsequent exon-capture (Hugall *et al.* 2015). With the same
90 objective of deriving a gene set appropriate for exon-capture in future studies, here we
91 implement this approach to identify and qualify 500 single-copy orthologous genes for the
92 Eupulmonata, a major lineage of air breathing snails and slugs within the class Gastropoda.

93 Eupulmonata comprises over 20,000 species, with an evolutionary depth spanning
94 over 150 million years (Jörger *et al.* 2010; Lydeard *et al.* 2010). The evolutionary
95 relationships of the Eupulmonata, however, remain incompletely understood despite many
96 morphological and molecular phylogenetic studies over the last two decades (e.g., Ponder &
97 Lindberg 1997; Wade *et al.* 2001, 2006; Grande *et al.* 2004; Dinapoli & Klussmann-Kolb
98 2010; Holznagel *et al.* 2010; Dayrat *et al.* 2011). The lack of congruence between studies is
99 largely due to a combination of using insufficient genetic markers (Schrödl 2014), with many
100 studies relying on 28S rRNA or mitochondrial sequences, and widespread morphological
101 convergence (Dayrat & Tillier 2002). Therefore to resolve the ‘tree of life’ of the
102 eupulmonates, it is essential to identify more independently evolving markers, with a greater
103 range of substitution rates, to better estimate relationships across all evolutionary depths. To
104 achieve this, we sequenced and assembled transcriptomes for representatives of 15 families
105 across Eupulmonata. We used the owl limpet genome, *Lottia gigantea*, as a reference to
106 identify and cluster homologous sequences and visually assessed and manually edited
107 candidate homolog alignments accounting for transcript fragmentation, mis-indexing and
108 frameshifts. We then further qualified orthology by assessing individual gene trees and by
109 sequencing the orthologous gene set from genomic DNA using exon-capture as unexpressed
110 paralogs or pseudogenes will not be detected in transcriptome datasets. Lastly, as a
111 comparison and qualification of our approach we also analysed our transcriptome dataset
112 using the fully automated orthology determination pipeline Agalma (Dunn *et al.* 2013).

113 METHODS

114 *Transcriptome sequencing and assembly*

115 We sequenced transcriptomes for 21 species of terrestrial snails and slugs
116 representative of 15 families across Eupulmonata (Table 1). Total RNA was extracted from

117 foot or whole body tissue stored in RNAlater (Ambion Inc, USA) using the Qiagen RNeasy
118 extraction kit (Qiagen, Hilden, Germany). Library preparations were conducted using the
119 TruSeq RNA sample preparation kit v2 (Illumina Inc., San Diego, CA), and sequenced on the
120 Illumina HiSeq 2000 platform (100 bp paired end reads). We used the program Trimmomatic
121 v0.22 (Lohse *et al.* 2012) to remove and trim low quality reads and adaptor sequences, and
122 the program Trinity v2012-06-08 (Grabherr *et al.* 2011; Haas *et al.* 2013) with default
123 settings to assemble the transcriptomes.

124 ***Homolog clustering***

125 Our approach to homolog clustering and orthology qualification is largely consistent
126 with that detailed in O'Hara *et al.* (2014). A schematic representation of our pipeline is
127 provided in Figure 1. First, to generate clusters of putatively homologous sequences we
128 compared each assembly to the *Lottia gigantea* predicted gene dataset (hereon referred to as
129 the *L. gigantea* genes). The *L. gigantea* reference represents 23,851 filtered gene models
130 annotated in the most current draft genome (Grigoriev *et al.* 2012; Simakov *et al.* 2013). Each
131 transcriptome assembly was compared against the *L. gigantea* genes using blastx with an e-
132 value cut off of 1e-10. This is a relatively relaxed threshold given the small size of the *L.*
133 *gigantea* reference set. A relaxed e-value cutoff was used to ensure all closely related
134 homologs were assessed without allowing through too many spurious matches with non-
135 homologous sequences. We retained only the top hit for each assembled contig (i.e. the match
136 with the lowest e-value).

137 In addition to identifying homologous contigs from each transcriptome assembly, we
138 also identified putative paralogs within the *L. gigantea* genome itself, in order to aid the
139 identification of paralogous sequences within the eupulmonates. We ran an all-by-all BLAST
140 of the *L. gigantea* genes against themselves (blastp, cut off e-value of 1e-10), retaining all

141 hits to identify *L. gigantea* genes which had hits to *L. gigantea* genes other than themselves.
142 To qualify the all-by-all BLAST results, we also obtained orthology status for all *L. gigantea*
143 genes classified in the Orthologous MAtrix (OMA) ortholog database (Altenhoff *et al.* 2015).
144 A *L. gigantea* gene was considered to be single-copy if it was the only *L. gigantea* sequence
145 in its respective OMA group. While this information provided guidance, we were not reliant
146 on the *L. gigantea* orthology status when prioritising homolog clusters to assess (see below
147 for criteria used). We considered *L. gigantea* to be sufficiently divergent from the
148 eupulmonates (> 400 million years, Zapata *et al.* 2014) that single-copy status could differ.

149 The BLAST results for both the transcriptomes compared to *L. gigantea* and the *L.*
150 *gigantea* all-by-all BLAST were used to produce clusters of homologous sequences linked by
151 having a match to a specific *L. gigantea* gene. Hence, a homolog cluster represents 1) all
152 contigs from all species transcriptomes that had a BLAST match to a given reference *L.*
153 *gigantea* gene (there were often multiple contigs per taxon with hits to a given *L. gigantea*
154 gene), and 2) all contigs having a hit to any of the closely related *L. gigantea* genes identified
155 by the all-by-all BLAST.

156 ***Orthology assessment***

157 After constructing the homologous clusters, we first visually assessed the alignments
158 for evidence of paralogy. Sequences for each cluster were placed into the correct reading
159 frame using coordinates output from the Blastx comparison for each transcriptome against *L.*
160 *gigantea*, and were then translated and aligned in amino acids using ClustalW (Thompson *et*
161 *al.* 1994) within the program BioEdit (Hall 1999). We only considered the coding region (i.e.
162 untranslated regions (UTRs) were removed) which was identified manually by reference to
163 the *L. gigantea* protein sequence for the relevant gene, which was included in the alignments.
164 Many of the homolog clusters contained multiple fragmented transcripts for a given species

165 that were shorter than the coding region but which often overlapped. These fragmented
166 transcripts were synthesised into consensus sequences by manual manipulation within
167 BioEdit, if the overlapping regions did not differ by more than three nucleotides. Non-
168 overlapping fragments were also concatenated if there were no competing contigs covering
169 the same region of the alignment and both sequences displayed a high degree of similarity to
170 non-fragmented sequences in closely related taxa.

171 By visually assessing each homolog alignment in both amino acid and nucleotides (in
172 Bioedit it is straight forward to toggle between the two), we were able to identify and
173 manually correct frameshifts. These were clearly evident as a large proportion of a contig
174 would not align with the rest of the sequences and the site of the frameshift was usually
175 associated with runs of adenines. We also manually edited the alignments to remove clearly
176 erroneous sequences which could not be aligned, clear out-paralogs (i.e. sequences which are
177 paralogous but the duplication event took place before the common ancestor of
178 eupulmonates) and redundant sequences (identical transcripts within a species). Mis-indexing
179 was identified as cases where, within the one assembly, two contigs were present for the
180 same region but one (typically the shorter contig having low coverage) matched the sequence
181 for another taxon exactly. Taxa containing paralogs were clearly evident in the alignments as
182 they frequently had > 5% dissimilarity at the nucleotide level between overlapping contigs
183 within the one sample. To further qualify that these sequences were paralogs we inspected
184 genealogies constructed using the neighbour joining method in MEGA (see Figure S1). Any
185 homolog cluster containing paralogs for any species was excluded from further consideration.
186 In certain cases paralogous sequences were closely related (3-5% dissimilarity), representing
187 either in-paralogs (see Remm *et al.* 2001) or genes exhibiting elevated allelic diversity (see
188 O'Hara *et al.* 2014). These genes were also excluded from further consideration as such
189 genes are not optimal for exon-capture.

190 Approximately 1,500 homologous clusters were visually assessed in order to find 500
191 which were orthologous across all 21 taxa assessed. This dataset size was chosen to represent
192 a balance between phylogenetic power at varying time scales (Leaché & Rannala 2011;
193 Philippe *et al.* 2011; Lemmon & Lemmon 2013) and a suitable size for subsequent exon-
194 capture probe design. To maintain consistency across studies, we first assessed homolog
195 alignments corresponding to the 288 *L. gigantea* genes used in a phylogenomic study of the
196 Mollusca (Kocot *et al.* 2011). Although there are two other published molluscan
197 phylogenomic datasets (Smith *et al.* 2011; Zapata *et al.* 2014), we focussed on the final
198 dataset of Kocot *et al.* (2011) as the *L. gigantea* gene IDs were documented in the
199 supplementary they provided, which in turn allowed us to easily identify and assess these
200 genes given our pipeline was based on the same reference. We then proceeded to assess and
201 qualify additional homolog clusters until we obtained a final set of 500 single-copy
202 orthologous genes. Accordingly, we prioritised homolog clusters with high taxonomic
203 representation (≥ 18 taxa), as completeness of the data matrix is critical for designing probes
204 across multiple lineages (Lemmon *et al.* 2012; Hugall *et al.* 2015). Where possible we also
205 prioritised homolog alignments for which the corresponding *L. gigantea* gene had a coding
206 region (CDS) ≥ 300 bp or had at least one exon ≥ 200 bp.

207 As a proxy measure of substitution rate variation across the final 500 gene set, we
208 calculated uncorrected distances (p-distance) for species pairs within the families Rhytididae
209 (*Terrycarlessia turbinata* and *Victaphanta atramentaria*) and Camaenidae (*Sphaerospira*
210 *fraseri* and *Austrochloritis kosciuszkoensis*). We chose to limit this analysis to intrafamilial
211 comparisons to avoid underestimation due to saturation. For comparison, we also calculated
212 the p-distances for two commonly used phylogenetic markers, CO1 and 28S, for the same
213 taxa.

214 ***Qualification of orthology using gene tree assessments***

215 Although only a single copy of each gene per taxon was present in our final ortholog
216 alignments, they may nevertheless be paralogous across taxa (see Struck 2014). To
217 investigate ‘hidden paralogy’ we used the program TreSpEx (Struck 2014) to assess
218 genealogies for conflict with *a priori* taxonomic hypotheses. Gene trees for each of the 500
219 genes were constructed using the GTRGAMMA model, codon specific partitioning, and 100
220 fast bootstraps in RAxML (Stamatakis 2006). TreSpEx then identified well supported
221 conflicting phylogenetic signal relative to five distinct and taxonomically well-established
222 eupulmonate clades (Limacoidea, Orthurethra, Helicoidea, the Australian Rhytididae (Table
223 1: see Hausdorf 1998; Wade *et al.* 2006, 2007; Herbert *et al.* 2015), and the
224 Stylommatophora). All nodes with ≥ 75 bootstrap support were first assessed for conflict with
225 the monophyly of each of the five clades. Strongly supported sister relationships between
226 sequences from different clades can indicate the presence of ‘hidden’ paralogous sequences.
227 TreSpEx flags very short terminal branches (parameter blt set to 0.00001) as indicative of
228 potential cross-contamination and internal branches which are five times greater than the
229 average (parameter lowbl set to 5), which, in addition to strong nodal support, may indicate
230 paralogy.

231 ***Qualification of orthology using exon-capture***

232 To further qualify orthology and identify unexpressed paralogs and pseudogenes, we
233 designed an exon-capture probe set to enrich and sequence exons from our 500 gene dataset.
234 As the divergence across the eupulmonates is too large for a single probe design we designed
235 a probe set for the Australian Camaenidae as a test case. It would be feasible, however, to
236 design a probe set from our alignments for any of the taxa we have assessed in this study. We
237 designed the baits based on two species of Australian Camaenidae, *Sphaerospira fraseri* and
238 *Austrocholritis kosciuszkoensis*, which represent two divergent lineages of the Australian
239 camaenids (Hugall & Stanisic 2011). Specifically, we included sequences from both taxa for

240 each gene in the probe design. The divergence between these taxa ranges up to ~12% (Figure
241 3) which is about the level of divergence tolerated by the probes (Hugall *et al.* 2015).
242 Including both taxa in the design increases the likelihood that we will capture sequences from
243 more divergent lineages within the Camaenidae for which we don't yet have transcriptome
244 sequences. Exon boundaries were first delineated using the program Exonerate v2.2.0 (Slater
245 & Birney 2005) in reference to the *L. gigantea* genomic sequences and then manually
246 qualified using the boundaries detailed in the *L. gigantea* genome annotation (JGI, Grigoriev
247 *et al.* 2012). All exons shorter than 120bp (the probe length) were excluded. This resulted in a
248 target consisting of 1,646 exons from 490 of the 500 genes (ten genes contained only exons
249 shorter than 120bp and were excluded from the bait design). Probes for the target sequences
250 were designed and produced by MYcroarray (Ann Arbor, Michigan) using MYbaits custom
251 biotinylated 120bp RNA baits at 2X tiling.

252 We tested the probe set on 22 camaenid species spanning much of the phylogenetic
253 breadth of the Australasian camaenid radiation, representing up to 30 million years (My) of
254 evolution (Hugall & Stanisic 2011) (Table 2). DNA was extracted using the DNeasy blood
255 and tissue kit (Qiagen) and sheared using the Covaris S2 (targeting a fragment size of 275bp).
256 Libraries were then constructed using the Kapa DNA Library Preparation Kit (Kapa
257 Biosystems, USA), modified to accommodate dual-indexing using the i7 and i5 index sets
258 (see Hugall *et al.* 2015). Up to eight libraries (normalised to 100 ng each) were pooled per
259 capture, and hybridised to the baits (at one-quarter dilution) for 36 hours, following the
260 MYbait protocol v1. A second hybridisation was then carried out on the fragments retained
261 from the first hybridisation to further enrich the capture. Several captures were then
262 multiplexed and sequenced on the Illumina MiSeq platform (v2), obtaining 150bp paired-end
263 reads.

264 We used FastUniq v1.1 (Xu *et al.* 2012) to remove duplicates, and Trimmomatic
265 v0.22 (Bolger *et al.* 2014) to trim and remove low quality reads and adaptor sequences
266 (minimum average quality score threshold of 20 per 8nt window). Reads shorter than 40
267 bases after trimming were discarded. The trimmed reads were then mapped onto the
268 transcriptome sequences used for the probe design using BFAST v0.7.0a (Homer *et al.* 2009)
269 with a single index of 22nt without mismatch. After creating pileup files using Samtools
270 v0.1.19 (Li *et al.* 2009), VarScan v2.3.7 (Koboldt *et al.* 2012) was used to call variants and
271 produce a final consensus sequence for each taxon per exon. Viewing the initial BAM
272 alignments showed that exon boundaries were often not conserved between *L. gigantea* and
273 the Camaenidae. In these cases (Figure S5) the reference exons were split to reflect the actual
274 exon boundaries in the Camaenidae. The reads were then mapped to the revised exon
275 reference and consensus sequences made as outlined above. To flag potential pseudogenes
276 and paralogs we identified consensus sequences with an elevated proportion of variable sites
277 (> 3% heterozygote sites) and reviewed the corresponding read alignments (BAM files) using
278 the Integrative Genomics Viewer (IGV: Thorvaldsdóttir *et al.* 2013). All sequences with
279 greater than 3% ambiguous sites were removed from the final dataset. Exons where more
280 than 10% of the taxa contained greater than 3% ambiguous sites were discarded entirely.

281 We again used TreSpEx to assess conflicting phylogenetic signal. We screened for
282 hidden paralogs based on five *a priori* phylogenetic hypotheses representing well supported
283 clades ($\geq 75\%$ bootstrap support) within the Australasian camaenid radiation as delineated by
284 Hugall and Stanisic (2011), namely the Hadroid group (clade 1 – 4 inclusive), the far-
285 northern (sister clades 5 and 6) and north-eastern (clade 7) Chloritid groups, a group
286 dominated by arid and monsoonal camaenids (clade 11) previously recognised as the
287 subfamily Sinumeloninae (e.g. Solem 1992), and a phenotypically and ecologically diverse
288 group dominated by eastern Australian wet forest taxa (sister clades 8 and 9). Gene trees for

289 each of the 490 genes (exons from the same gene were combined as one partition) were
290 constructed using the GTRGAMMA model and 100 fast bootstraps in RAxML (Stamatakis
291 2006). TreSpEx was run using the same settings as the analysis for the transcriptome dataset
292 (i.e. TreSpEx considered nodes for strong conflict, long branches, and short branches in that
293 order with parameters upbl and lowbl set to 5 and blt 0.00001).

294 ***Comparison to the Agalma pipeline***

295 As an independent qualification of the manually curated 500 gene set we ran the fully
296 automated orthology determination pipeline Agalma (Dunn *et al.* 2013) (Figure 1). We
297 commenced this pipeline from the ‘postassemble’ step which first identified open reading
298 frames and putative coding regions (Dunn *et al.* 2013). Homolog clusters were then identified
299 using an all-by-all tblastx, followed by clustering using the Markov Clustering algorithm
300 (MCL) (Figure 1). Homolog clusters were then translated and aligned using MAFFT (Katoh
301 & Standley 2013) and gene trees estimated using RAxML. To identify orthologous
302 sequences, the genealogies were then screened for ‘optimally inclusive subtrees’ which
303 contain only a single representative of each species. Multiple orthologous subtrees can be
304 delineated per homolog cluster, potentially allowing paralogs to be separated and retained.
305 The surviving subtrees were filtered based on the number of taxa (set to greater than four
306 taxa) and realigned for subsequent phylogenetic analysis. We then identified Agalma
307 homologous clusters that corresponded to the manually curated 500 gene set using BLAST
308 (blastp, e-value cut off of 1e-10).

309 ***Phylogenetic analysis***

310 After removal of paralogs or sequences with excessive polymorphism (>3%
311 dissimilarity), our phylogenomic datasets were refined by removing any regions of
312 ambiguous alignment through the use of Gblocks (Castresana 2000) (which is built into the

313 Agalma pipeline) and manual masking. We reconstructed maximum likelihood trees using
314 the program RAxML (Stamatakis 2006) for datasets resulting from both the manual curation
315 and the Agalma pipeline. PartitionFinder (Lanfear *et al.* 2012, 2014) was used to identify
316 suitable models and partitioning schemes, implemented with 1% heuristic r-cluster searches,
317 optimized weighting, RAxML likelihood calculations, and model selection based on BIC
318 scores. In all cases, nodal support was assessed by performing 100 full non-parametric
319 bootstraps.

320 We analysed two datasets resulting from the Agalma pipeline. The first dataset
321 comprised ortholog clusters that corresponded to the manually curated 500 gene set (here on
322 referred to as the ‘Agalma equivalent dataset’). The second dataset consisted of all ortholog
323 clusters which had high taxon coverage (≥ 18), and were derived from homolog clusters
324 containing only a single ortholog cluster (from here on referred to as the ‘Agalma best
325 dataset’); that is, Agalma homolog clusters containing multiple copies, albeit diagnosable,
326 were not considered further. Finally, we reconstructed a phylogeny for the camaenid dataset
327 obtained through exon-capture and included sequences from the five camaenid
328 transcriptomes presented herein, as well as sequences of *Cornu aspersum* as an outgroup.

329 RESULTS

330 *Transcriptome assembly and homolog clustering*

331 The number of paired reads obtained for each of the 21 eupulmonate species
332 sequenced ranged from 7.8M to 31.6M (Table 3). Trimming and de novo assembly statistics
333 are presented in Table 3. The number of *L. gigantea* reference genes with BLAST matches
334 ranged from 7,011 to 9,699 per assembly (Table 3), 5,490 of which had homologous
335 sequences in at least 18 of the 21 transcriptome assemblies.

336 Of the 288 genes used in a previous molluscan phylogenomic study (Kocot *et al.*
337 2011), 130 were single-copy for all eupulmonates considered here, while 146 contained
338 paralogs in at least one species (mean p-distance between paralogs within a sample was 0.28,
339 ranging from 0.16-0.46). We could not unambiguously qualify the remaining 12 genes from
340 the Kocot *et al.* study as they were poorly represented in our transcriptomes. Prioritising
341 genes with high taxon coverage and long exon length, we assessed additional alignments of
342 candidate homolog clusters until we reached a total of 500 single-copy genes. In addition to
343 the 146 Kocot genes shown to be paralogous within the eupulmonates, we identified and
344 qualified 62 multi-copy genes during the course of this work. The resulting manually curated
345 500 single-copy gene set is 98.5% taxa complete (i.e. sequence present for each gene and
346 taxon) and 93.1% character complete (Figure 4d), with an average gene length of 1,190nt,
347 ranging from 228nt to 6,261nt. In total, the final alignment of this gene set represents
348 512,958nt. Approximately 12% of the sequences in the final gene-by-species matrix were
349 derived by merging fragmented transcripts.

350 Based on the all-by-all BLAST comparison of the *L. gigantea* genes, 347 of our final
351 500 genes had a single hit at an e-value threshold of 1e-10 (i.e. single copy status was
352 consistent between the *L. gigantea* reference and the eupulmonates), while the remainder had
353 multiple hits, indicative of the presence of close paralogs in the reference. Conversely, of the
354 208 genes qualified as multiple-copy for the eupulmonates (146 from the Kocot gene set plus
355 62 from this study), 134 only had one hit within the *L. gigantea* gene set (i.e. just over half of
356 the multiple-copy gene set are potentially single copy for patellogastropods). These results
357 broadly correspond to the orthology designation in the OMA (Orthologous MAtrix) database.

358 Across the 500 single-copy genes, the p-distance between the two rhytidids,
359 *Terrycarlessia turbinata* and *Victaphanta atramentaria*, ranged from 0.02 to 0.13 (average of
360 0.06; Figure 3). This family is thought to have originated 120 Mya (Bruggen 1980; Upchurch

361 2008). However, the Australian rhytidids probably represent a more recent radiation (Herbert
362 et al. 2015, Moussalli and Herbert 2016). Similarly, p-distance between the two camaenids,
363 *Sphaerospira fraseri* and *Austrochloritis kosciuszkoensis*, ranged from 0.01 to 0.13 (average
364 of 0.04). This group is thought to have originated in the Oligo-Miocene approximately 30
365 Mya (Hugall & Stanisic 2011). All genes had a higher relative substitution rate than the
366 commonly used phylogenetic marker 28S, and were on average approximately four times
367 slower than COI (Figure 3).

368

369 ***Qualification of orthology using gene tree assessments***

370 TreSpEx analyses of all 500 genes found no well supported conflict with the *a priori*
371 phylogenetic hypotheses, suggesting that hidden paralogs (i.e., genes represented by a single
372 sequence per taxon yet paralogous across multiple taxa) were absent from our dataset.
373 Furthermore, this analysis also showed no evidence of cross sample contamination, nor any
374 evidence of suspect long internal branches within the Stylommatophora.

375 ***Qualification of orthology using exon-capture***

376 We enriched and sequenced all 1,646 targeted exons, from 490 genes, when
377 considering all 22 samples collectively. We first mapped reads to the original reference used
378 in the probe design with exon boundaries delineated based on the *L. gigantea* genome.
379 Examination of the resulting read alignments (BAM files) identified 437 exons which
380 contained multiple internal exon boundaries within the Camaenidae (Figure S4).
381 Accordingly, the mapping reference was modified to account for exon-splitting (including the
382 removal of 163 exons that were shorter than 40 bp after splitting), with the final revised
383 reference comprising 2,648 exons representing 417,846nt (Supplementary Table 1). We
384 targeted an average of five exons per gene.

385 We then remapped reads to the revised reference (coverage and specificity statistics
386 presented in Table 4) and flagged resulting consensus sequences which exhibited elevated
387 polymorphism (> 3% heterozygote sites). There were 508 exons where at least one taxon
388 exhibited elevated polymorphism. Of these, 105 exons had greater than 10% of the taxa
389 (typically two or more taxa, taking into account missing taxa) exhibiting elevated
390 polymorphism. Based on an examination of the corresponding read alignments, 95 exons
391 were classified as having lineage specific pseudogenes or paralogs, four contained evidence
392 of processed pseudogenes, and six where the alignment was complicated by the mapping of
393 unrelated reads containing small, highly similar domains (see Figure S4-S8 for examples of
394 each case). These 105 exons were removed prior to phylogenetic analyses. For the remaining
395 403 exons only the consensus sequences for the taxa with elevated polymorphism were
396 removed from the final alignment. In total, 3.7% of the sequences were removed from final
397 data matrix due to elevated polymorphism. The final exon capture data matrix was 98% taxa
398 complete and 95% character complete.

399 Based on the TreSpEx analyses, four genes did not support the monophyly of the ‘Far
400 North Chloritid’ group, but rather placed (*Nannochloritis layardi* and *Patrubella buxtoni*) as
401 sister to the ‘North-East Chloritid’ group (Figure 5). We concluded that this was not the result
402 of hidden paralogy, but rather due to insufficient lineage sorting of relatively conserved
403 genes. An additional five genes were in conflict with the *a priori* taxonomic hypotheses,
404 however, these represented cases where the genes were small and the proportion of
405 phylogenetically informative sites was low. Five genes were flagged as having at least one
406 internal branch which was greater than five times the average. Assessment of the alignments
407 and corresponding genealogies indicated that they represented deep basal divergence between
408 well supported major clades, and was not reflective of hidden paralogy.

409 Finally, we enriched another representative of *Sphaerospira fraseri*, one of the
410 reference species used in the probe design. Comparing the mapped consensus genomic
411 sequence to the transcriptome reference we found only minor mismatch, reflective of
412 intraspecific variation as the two samples came from different populations (the exons had a
413 median p-distance of 0.8%). Furthermore, for this species at least, all reference genes
414 constructed from multiple transcript fragments were consistent with those captured from
415 genomic DNA (i.e. chimeras of unrelated fragments were not created) and showed no
416 evidence of paralogy or elevated heterozygosity.

417

418 ***Comparison to Agalma pipeline***

419 Using the Agalma pipeline we identified 11,140 ortholog clusters. Of these ortholog
420 clusters 635 corresponded to 457 of our 500 single-copy gene set. We refer to this dataset as
421 the “Agalma equivalent” dataset, and is 61% taxa complete and 54% character complete.
422 Many of the genes were represented by multiple ortholog clusters in the Agalma analysis,
423 many of which contained fewer taxa relative to that obtained via manual curation (Figure 2).
424 Rather than paralogs, in all cases fragmentation in the transcriptome assemblies resulted in
425 the splitting of homolog clusters into multiple ortholog clusters, each representing the same
426 locus but containing a different subset of taxa (see example in Figure S3). Of the 43 single-
427 copy genes not picked up by Agalma, five were not annotated in the ‘postassemble’ step, 12
428 were annotated but not recovered by the all-by-all BLAST, 18 were recovered by the all-by-
429 all BLAST but dropped during the clustering step, and eight made it to the initial clusters but
430 failed the alignment and trimming step prior to the gene tree reconstruction. Failure to
431 recover these genes during the BLAST comparison, clustering and alignment steps is most
432 likely due to a combination of frameshift errors and transcript fragmentation, and in certain
433 cases, resulting in the taxon sampling threshold and cluster size criteria not being met.

434 Of the 11,140 ortholog clusters there were 546 clusters that contained sequences of at
435 least 18 taxa and that had one ortholog cluster per homolog cluster. Of these, 171 were also
436 contained in our 500 single-copy gene set. Hence, the Agalma pipeline identified 375 genes
437 in addition to the 500 manually curated genes, which had optimum taxon sampling. The
438 majority of these genes also represented the full CDS with 89% representing at least 80% of
439 the length of the respective *L. gigantea* gene. We refer to this dataset as the “Agalma best”
440 dataset and is 92% taxa complete and 85% character complete.

441 ***Phylogenetic analysis***

442 We reconstructed phylogenies from three ortholog datasets for comparison: (1) the
443 manually curated 500 single-copy gene set (Figure 4a, d), (2) the Agalma equivalent dataset
444 consisting of 635 orthologous clusters which corresponded to 457 of the 500 single-copy
445 genes (Figure 4b, e), and (3) the Agalma best dataset consisting of 546 orthologous clusters
446 which had 18 or more taxa and were the only orthologous cluster from the respective
447 homolog cluster (Figure 4c, f). Of the manual curated dataset, 1.6% of the alignment was
448 removed by Gblocks prior to phylogenetic analysis. The phylogenies for the 500 single-copy
449 gene set and the Agalma best dataset had identical topologies, supporting all major clades
450 with very high bootstrap support, namely Helicoidea, Limacoidea, Orthurethra, the Australian
451 rhytidids and the Stylommatophora (Figure 4a, c). In terms of phylogenetic relationships, the
452 Rhytididae forms a sister relationship with the Limacoidea, and the Helicoidea occupies a
453 basal position within Stylommatophora. In contrast, while also supporting the monophyly of
454 all major clades, the phylogeny based on the ‘Agalma equivalent’ dataset places Orthurethra
455 in a basal position within Stylommatophora, (Figure 4b).

456 Of the Camaenidae exon capture dataset, 5% of the alignment was removed by
457 Gblocks prior to phylogenetic analysis. The resulting phylogeny supported all major groups

458 previously recognised by Hugall and Staniscic (2011). In terms of phylogenetic relationships,
459 the two Chloritid groups formed a clade with the Hadroid group, with the Far-northern
460 chloritids sister to the hadroids. There was poor resolution regarding the phylogenetic
461 positions of the two remaining groups, the Eastern rainforests and the arid and monsoonal
462 NW Australian clades (Figure 5).

463

464 DISCUSSION

465 The identification and qualification of orthology is a critical prerequisite for sound
466 phylogenetic inference. Our approach of orthology assessment involved an initial assessment
467 and manual editing of homolog clusters, allowing us to correct for multiple isoforms and
468 errors such as sequence fragmentation, frame-shifts and mis-indexing. Using this approach,
469 we qualified the orthology and single-copy status of 500 genes across the eupulmonates, 130
470 of which were used in a previous phylogenomic study of the Mollusca (Kocot *et al.* 2011).
471 The resulting 500 gene data matrix is the most complete produced for a major molluscan
472 lineage to date, both in terms of taxon and character completeness. We further qualified
473 orthology by capturing and sequencing 490 of the 500 genes from genomic DNA, revealing
474 the presence of paralogs and/or pseudogenes otherwise not evident from the transcriptome
475 data. Although the automated pipeline Agalma recovered the majority of the 500 genes as
476 single copy and identified 375 additional putatively orthologous genes for the eupulmonates,
477 it was hampered by transcript fragmentation within the assemblies. Furthermore, supported
478 topologies for the 21 eupulmonate species were not entirely consistent between the manually
479 curated and Agalma equivalent dataset, potentially a consequence of lower data matrix
480 completeness in the latter. We discuss approaches to ortholog determination and implications
481 for phylogenetic inference below.

482 ***Ortholog determination***

483 To date, most transcriptome based phylogenomic studies have focused on resolving
484 relatively deep evolutionary relationships (e.g. Kocot *et al.* 2011; Smith *et al.* 2011; Zapata *et*
485 *al.* 2014; O’Hara *et al.* 2014; Misof *et al.* 2014), and a number have relied on annotated
486 ortholog databases for the initial screening of suitable genes, such as OMA (Altenhoff *et al.*
487 2015), OrthoDB (Waterhouse *et al.* 2013), and the ortholog dataset associated with HaMStR
488 (Ebersberger *et al.* 2009). Such databases are typically limited in the number of
489 representatives per lineage (e.g., Tatusov *et al.* 2003; Ranwez *et al.* 2007; Waterhouse *et al.*
490 2013; Altenhoff *et al.* 2015). Nevertheless, it is a reasonable assumption that orthologous
491 genes qualified as single-copy across many highly divergent taxa are more likely to maintain
492 single-copy status with greater taxonomic sampling. We tested this idea at a preliminary stage
493 of our work by first assessing genes used in a phylogenomic study of the Mollusca (Kocot *et*
494 *al.* 2011). In that study, orthologous genes were identified using the program HaMStR, based
495 on a 1,032 ortholog set resulting from the Inparanoid orthology database (Ostlund *et al.*
496 2010). We found that just under half of the genes used in Kocot *et al.* (2011) were paralogous
497 within the eupulmonates. To some extent the high proportion of the Kocot *et al.* gene set
498 being paralogous is due to the limited representation of eupulmonates in that study, and for
499 these few taxa paralogs may have been absent. Alternatively, in such deep phylogenomic
500 studies lineage-specific duplication may have manifested as in-paralogs and were dealt with
501 by retaining one copy from the in-paralog set at random (Kocot *et al.* 2011; Dunn *et al.* 2013)
502 or based on sequence similarity (Ebersberger *et al.* 2009). However, with an increase in
503 taxonomic sampling, such paralogy may extend across multiple taxa and, unless conservation
504 of function can be established (i.e. isorthology, Fitch 2000), these genes would no longer be
505 suitable for phylogenetic analysis.

506 When the 500 gene set was compared to the OMA database (Altenhoff *et al.* 2015),
507 which at the time of this analysis only incorporated a single molluscan genome, namely *L.*
508 *gigantea*, we found a similarly high proportion of eupulmonate specific paralogy. A more
509 interesting result arising from this comparison, however, was that many genes classified as
510 having putative paralogs in *L. gigantea* were single-copy across the eupulmonates. We cannot
511 ascertain at this stage whether this is a consequence of duplication being derived within
512 Patellogastropoda, the lineage containing *L. gigantea*, or the consequence of duplicate loss in
513 the ancestral eupulmonate. Nevertheless, this result highlights that potentially suitable genes
514 may be overlooked when restricted to ortholog database designations, especially when such
515 databases have poor representation of the relevant lineage. Accordingly, although we used the
516 *L. gigantea* gene set as a reference with which to identify and cluster homologous sequences,
517 we did not rely on orthology database designations of the *L. gigantea* gene set to guide which
518 genes to consider when assessing orthology across the eupulmonates examined here.

519 *Automated vs manually curated aided pipelines*

520 Pipelines that fully automate homology searches and clustering, orthology
521 qualification, and final alignments are highly desirable for efficiency, consistency, and
522 repeatability. Moreover, reference free methods, like that implemented in Agalma, are also
523 highly desirable in cases where the study taxa are poorly represented in ortholog databases.
524 There are characteristics of assembled transcriptome sequences, however, that can challenge
525 fully automated methods, including transcript fragmentation, mis-indexing, frameshifts and
526 contamination, and these aspects necessitate careful manual appraisal and editing (Philippe *et*
527 *al.* 2011; O'Hara *et al.* 2014). Although recent phylogenomic studies have, to varying
528 degrees, incorporated manual appraisal, such checks are typically conducted at the final
529 proofing stage (e.g. Kocot *et al.* 2011; Simmons & Goloboff 2014). In this study, we
530 purposefully addressed the abovementioned issues at an early stage following the initial

531 alignment of homologous sequences. The most important aspect of our manual curation was
532 the creation of consensus sequences from fragmented transcripts (see also: O'Hara *et al.*
533 2014), which in turn ensured maximum retention of data (particularly for probe design) and
534 placed subsequent orthology assessment on a sounder footing. Consequently, our final data
535 matrix was highly complete (93% character complete whereas the 'Agalma best' dataset was
536 85% character complete).

537 The Agalma analysis confirmed the single-copy, orthology status for the majority of
538 the 500 manually curated gene set, but it was hampered by transcript fragmentation within
539 the transcriptome assemblies. In all cases where multiple ortholog clusters were derived using
540 Agalma for any one of our 500 genes, this was due to transcript fragmentation, not missed
541 paralogy. In essence, alignments of fragmented transcripts (whether or not they were partially
542 overlapping) resulted in poorly reconstructed gene trees, which in turn misled subsequent tree
543 pruning and ortholog clustering (e.g. Figure S3). Consequently, for the 'Agalma equivalent'
544 dataset, both taxon and character completeness was poor relative to the manually curated data
545 matrix. To our knowledge, no fully automated phylogenomics pipeline currently implements
546 the consensus of fragmented sequences, and studies that have made the effort to retain
547 multiple fragments, as in this study, have decided which sequences to retain and merge
548 manually (e.g., Rothfels *et al.* 2013; O'Hara *et al.* 2014). The issue of working with
549 fragmented assemblies can be addressed, however, by incorporating an automated consensus
550 making algorithm such as TGICL (Pertea *et al.* 2003) into the pipeline to address
551 fragmentation at the homolog alignment stage. Doing so is particularly desirable, given that
552 manual curation of homologous sequences requires considerable time investment.

553 A major strength of automated pipelines is that they enable a more comprehensive
554 screening of putative orthologous genes. Manual curation requires considerable effort, and
555 while more candidate genes were identified than were assessed, we ceased the manual

556 assessment once our target of 500 genes had been attained. The Agalma analyses had no
557 constraints, however, hence all possible orthologous clusters were considered. Consequently,
558 we identified an additional 375 ortholog clusters which met a strict taxa completeness
559 threshold (18 taxa or more) and represented the only ortholog cluster arising from original
560 homolog clusters. These genes (i.e. the ‘Agalma best’ dataset) reconstructed a phylogeny that
561 was very similar to the manually curated dataset. While beyond the scope of this study, there
562 is potential for these genes to be included in future probe designs and further qualification of
563 these additional genes using exon-capture (see below) would be highly desirable.

564 ***Phylogenetic inference***

565 The 500 gene set represents a significant contribution towards advancing molecular
566 phylogenetics of the eupulmonates, providing the capacity to resolve both evolutionary
567 relationships at shallow to moderate depths, and deep basal relationships. The phylogenetic
568 reconstructions presented here are well resolved and support the *a priori* taxonomic
569 hypotheses used as part of the orthology assessment. In terms of deeper relationships,
570 reconstructions based on the two most complete datasets are consistent, namely the
571 monophyly of Stylommatophora, within which Helicoidea is basal, and the sister relationship
572 between the Rhytidoidea and the Limacoidea. For the less complete Agalma equivalent
573 dataset, however, Orthurethra is basal within Stylommatophora, albeit with marginal support.
574 Without greater taxonomic sampling of all the major lineages within the eupulmonates,
575 however, a comprehensive phylogenetic assessment is beyond the scope of this study.
576 Nevertheless, these phylogenomic datasets do afford greater resolution of deeper
577 relationships than obtained in previous molecular studies (Wade *et al.* 2001, 2006). Secondly,
578 convergence in supported topology between the two most complete and largely independent
579 datasets (only 171 genes were in common), and the inconsistency between the manually

580 curated and Agalma equivalent dataset (sharing 458 genes), suggests the possible importance
581 of data matrix completeness in resolving short, basal internodes.

582 *Exon-capture*

583 One of the overarching objectives of this study was to identify and qualify 500 genes
584 suitable for exon-capture work within the eupulmonates. Here we sequenced and analysed a
585 small dataset for the family Camaenidae principally as a means to further qualify orthology.
586 There are two principle outcomes from this exploration. First, for all reference sequences
587 based on the concatenation of fragmented transcripts, there was no evidence that erroneous
588 chimeric sequences were created. Second, as was the case with the increased sampling in the
589 transcriptome work, the pervasiveness of lineage-specific duplication was also evident from
590 the exon-capture experiment. Despite qualification of single-copy orthology of the
591 transcriptome dataset, increased taxonomic sampling within the family Camaenidae revealed
592 lineage-specific duplication for potentially as high as one fifth of the targeted exons. In the
593 great majority of cases, however, a very small proportion of taxa exhibited putative paralogy
594 or pseudogenes, and removal of the affected exon per taxon only reduced the completeness of
595 the final dataset by 3.7%. Similar results were achieved for the brittle stars with 1.5% of their
596 target discarded due to putative paralogs or pseudogenes (Hugall *et al.* 2015). It is possible
597 that these putative paralogs were only detected in the genomic sequencing because they were
598 not expressed in the transcriptomes.

599 Within the Australian Camaenidae, uncorrected distances for the majority of the genes
600 did not exceed 13%. This level of sequence variability is within the range of mismatch that is
601 tolerated by in-solution exon-capture protocols (Bi *et al.* 2012; Bragg *et al.* 2015; Hugall *et*
602 *al.* 2015). This was qualified here given the high proportion of target recovery (>95%) across
603 a broad representation of the camaenid diversity. As was the case for the Euplumonata

604 phylogeny presented above, our preliminary phylogenomic dataset for camaenids provides
605 considerable resolution, particularly among the chloritis and hadroid groups which to date
606 have been difficult to resolve (Hugall & Stanisc 2011).

607 Expanding the bait design to enrich across the Australasian camaenid radiation,
608 indeed the family Helicoidea, would require the incorporation of multiple divergent reference
609 taxa into the bait design. Recent “anchored enrichment” approaches to bait design (e.g.
610 Lemmon *et al.* 2012; Faircloth *et al.* 2012) target highly conserved regions to allow capture
611 across highly divergent taxa. By contrast, the approach taken here is to target both conserved
612 and highly variable regions, and where possible the full coding region (Bi *et al.* 2012; Bragg
613 *et al.* 2015; Hugall *et al.* 2015). Accordingly, this would require substantially greater
614 reference diversity to be incorporated into the bait design relative to the anchored approach to
615 capture across highly divergent lineages (e.g. across families). Recently, Hugall *et al.* (2015)
616 used a similar approach to the one in the present study, but designed baits based on ancestral
617 sequences, rather than representative tip taxa, to reduce the overall size of the reference set.
618 Using this approach, Hugall *et al.* successfully enriched and sequenced both conserved and
619 highly variable exons across the entire echinoderm class Ophiuroidea, spanning
620 approximately 260 million years. Here we have presented a simple bait design targeting a
621 specific family, but our transcriptome dataset could be used to produce a more diverse bait
622 design to facilitate a more comprehensive study of Eupulmonata phylogenetics and
623 systematics.

624

625 ACKNOWLEDGEMENTS

626 We thank Andrew Hugall for bioinformatics advice and Felipe Zapata for assistance with
627 running Agalma; Devi Stuart-Fox, Claire Mclean and Mark Phuong for critical feedback on

628 the manuscript; and Dai Herbert for providing images for Figure 4. We also thank the staff at
629 Australian Genome Research Facility (AGRF) and the Georgia Genomics Facility (GGF),
630 specifically Travis Glenn and Roger Neilsen, for guidance and assistance in transcriptome
631 and exon-capture sequencing. LT was supported by Victorian Life Science Computing
632 Initiative (VLSCI) scholarship. The work was supported by a Holsworth Wildlife Research
633 Endowment to LT, an Australian Biological Resources Study (ABRS) grant to FK and AM
634 (RF213-12). This research was undertaken with the assistance of resources from the National
635 Computational Infrastructure (NCI), which is supported by the Australian Government.

636

637 REFERENCES

- 638 Altenhoff AM, Skunca N, Glover N *et al.* (2015) The OMA orthology database in 2015: function
639 predictions, better plant support, synteny view and other improvements. *Nucleic Acids Research*,
640 **43**, D240–D249.
- 641 Bi K, Vanderpool D, Singhal S *et al.* (2012) Transcriptome-based exon capture enables highly cost-
642 effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*,
643 **13**, 403.
- 644 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data.
645 *Bioinformatics (Oxford, England)*, 1–7.
- 646 Bragg JG, Potter S, Bi K, Moritz C (2015) Exon capture phylogenomics: efficacy across scales of
647 divergence. *Molecular Ecology Resources*, DOI: 10.1111/1755-0998.12449.
- 648 Bruggen AC Van (1980) Gondwanaland Connections in the Terrestrial Molluscs of Africa and
649 Australia. *Journal of the Malacological Society of Australia*, **4**, 215–222.
- 650 Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in
651 phylogenetic analysis. *Molecular biology and evolution*, **17**, 540–552.
- 652 Dayrat B, Conrad M, Balayan S *et al.* (2011) Phylogenetic relationships and evolution of pulmonate
653 gastropods (Mollusca): New insights from increased taxon sampling. *Molecular Phylogenetics
654 and Evolution*, **59**, 425–437.
- 655 Dayrat B, Tillier S (2002) Evolutionary relationships of euthyneuran gastropods (Mollusca): a
656 cladistic re-evaluation of morphological characters. *Zoological Journal of the Linnean Society*,
657 **135**, 403–470.
- 658 Dinapoli A, Klussmann-Kolb A (2010) The long way to diversity - phylogeny and evolution of the
659 Heterobranchia (Mollusca: Gastropoda). *Molecular phylogenetics and evolution*, **55**, 60–76.
- 660 Dunn CW, Howison M, Zapata F (2013) Agalma: an automated phylogenomics workflow. *BMC
661 bioinformatics*, **14**, 330.
- 662 Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden markov model based search
663 for orthologs in ESTs. *BMC evolutionary biology*, **9**, 157.
- 664 Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of

- 665 protein families. *Nucleic acids research*, **30**, 1575–1584.
- 666 Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved Elements Anchor
667 Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Systematic biology*,
668 **61**, 717–726.
- 669 Fitch WM (2000) Homology: a personal view on some of the problems. *Trends in genetics*, **16**, 227–
670 231.
- 671 Gontcharov AA, Marin B, Melkonian M (2004) Are combined analyses better than single gene
672 phylogenies? A case study using SSU rDNA and rbcL sequence comparisons in the
673 Zygnematophyceae (Streptophyta). *Molecular biology and evolution*, **21**, 612–24.
- 674 Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq
675 data without a reference genome. *Nature Biotechnology*, **29**, 644–652.
- 676 Grande C, Templado J, Cervera JL, Zardoya R (2004) Molecular phylogeny of Euthyneura (Mollusca:
677 Gastropoda). *Molecular biology and evolution*, **21**, 303–313.
- 678 Grigoriev I V, Nordberg H, Shabalov I *et al.* (2012) The genome portal of the Department of Energy
679 Joint Genome Institute. *Nucleic acids research*, **40**, D26–D32.
- 680 Haas BJ, Papanicolaou A, Yassour M *et al.* (2013) De novo transcript sequence reconstruction from
681 RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, **8**,
682 1494–1512.
- 683 Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program
684 for Windows 95/98/NT. *Nucleic Acids Symposium Series*.
- 685 Hausdorf B (1998) Phylogeny of the Limacoidea Sensu Lato (Gastropoda: Stylommatophora).
686 *Journal of Molluscan Studies*, **64**, 35–66.
- 687 Herbert DG, Moussalli A, Griffiths OL (2015) Rhytididae (Eupulmonata) in Madagascar: reality or
688 conjecture? *Journal of Molluscan Studies*, **81**, 1–10.
- 689 Holznagel WE, Colgan DJ, Lydeard C (2010) Pulmonate phylogeny based on 28S rRNA gene
690 sequences: a framework for discussing habitat transitions and character transformation.
691 *Molecular phylogenetics and evolution*, **57**, 1017–25.
- 692 Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome
693 resequencing. *PloS one*, **4**, e7767.
- 694 Hugall AF, O’Hara TD, Hunjan S, Nilsen R, Moussalli A (2015) An Exon-Capture System for the
695 Entire Class Ophiuroidea. *Molecular Biology and Evolution*.
- 696 Hugall AF, Stanisic J (2011) Beyond the prolegomenon: a molecular phylogeny of the Australian
697 camaenid land snail radiation. *Zoological Journal of the Linnean Society*, **161**, 531–572.
- 698 Jörger KM, Stöger I, Kano Y *et al.* (2010) On the origin of Acochlidia and other enigmatic
699 euthyneuran gastropods, with implications for the systematics of Heterobranchia. *BMC*
700 *Evolutionary Biology*, **10**, 323.
- 701 Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7:
702 Improvements in Performance and Usability. *Molecular biology and evolution*, **30**, 772–780.
- 703 Koboldt DC, Zhang Q, Larson DE *et al.* (2012) VarScan 2: Somatic mutation and copy number
704 alteration discovery in cancer by exome sequencing. *Genome Research*, **22**, 568–576.
- 705 Kocot KM, Cannon JT, Todt C *et al.* (2011) Phylogenomics reveals deep molluscan relationships.
706 *Nature*, **477**, 452–456.
- 707 Kocot KM, Citarella MR, Moroz LL, Halanych KM (2013) PhyloTreePruner: A Phylogenetic Tree-
708 Based Approach for Selection of Orthologous Sequences for Phylogenomics. *Evolutionary*
709 *bioinformatics online*, **9**, 429–35.
- 710 Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: Combined Selection of
711 Partitioning Schemes and Substitution Models for Phylogenetic Analyses. *Molecular biology*

- 712 *and evolution*, **29**, 1695–1701.
- 713 Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A (2014) Selecting optimal partitioning
714 schemes for phylogenomic datasets. *BMC evolutionary biology*, **14**, 82.
- 715 Leaché AD, Rannala B (2011) The accuracy of species tree estimation under simulation: a
716 comparison of methods. *Systematic biology*, **60**, 126–37.
- 717 Lemmon AR, Emme S, Lemmon EM (2012) Anchored hybrid enrichment for massively high-
718 throughput phylogenomics. *Systematic biology*, **61**, 727–744.
- 719 Lemmon EM, Lemmon AR (2013) High-Throughput Genomic Data in Systematics and
720 Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 99–121.
- 721 Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic
722 genomes. *Genome research*, **13**, 2178–2189.
- 723 Lohse M, Bolger AM, Nagel A *et al.* (2012) RobiNA: a user-friendly, integrated software solution for
724 RNA-Seq-based transcriptomics. *Nucleic Acids Research*, **40**, W622–W627.
- 725 Lydeard C, Cowie RH, Ponder WF *et al.* (2010) The Global Decline of Nonmarine Mollusks.
726 *BioScience*, **54**, 321–330.
- 727 Martin AP, Burg TM (2002) Perils of paralogy: using HSP70 genes for inferring organismal
728 phylogenies. *Systematic biology*, **51**, 570–587.
- 729 Misof B, Liu S, Meusemann K *et al.* (2014) Phylogenomics resolves the timing and pattern of insect
730 evolution. *Science*, **346**, 763–767.
- 731 O’Hara TD, Hugall AF, Thuy B, Moussalli A (2014) Phylogenomic resolution of the class
732 Ophiuroidea unlocks a global microfossil record. *Current biology*: CB, **24**, 1874–1879.
- 733 Ostlund G, Schmitt T, Forslund K *et al.* (2010) InParanoid 7: new algorithms and tools for eukaryotic
734 orthology analysis. *Nucleic acids research*, **38**, D196–203.
- 735 Perteua G, Huang X, Liang F *et al.* (2003) TIGR gene indices clustering tools (TGICL): A software
736 system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.
- 737 Philippe H, Brinkmann H, Lavrov D V *et al.* (2011) Resolving difficult phylogenetic questions: why
738 more sequences are not enough. *PLoS biology*, **9**, e1000602.
- 739 Pirie MD, Vargas MPB, Botermans M, Bakker FT, Chatrou LW (2007) Ancient paralogy in the
740 cpDNA trnL-F region in Annonaceae: implications for plant molecular systematics. *American
741 Journal of Botany*, **94**, 1003–1016.
- 742 Ponder WF, Lindberg DR (1997) Towards a phylogeny of gastropod molluscs: an analysis using
743 morphological characters. *Zoological Journal of the Linnean Society*, **119**, 83–265.
- 744 Ranwez V, Delsuc F, Ranwez S *et al.* (2007) OrthoMaM: a database of orthologous genomic markers
745 for placental mammal phylogenetics. *BMC evolutionary biology*, **7**, 241.
- 746 Remm M, Storm CE V, Sonnhammer ELL (2001) Automatic Clustering of Orthologs and In-paralogs
747 from Pairwise Species Comparisons. *Journal of molecular biology*, **314**, 1041–1052.
- 748 Rothfels CJ, Larsson A, Li F-W *et al.* (2013) Transcriptome-mining for single-copy nuclear markers
749 in ferns. *PloS one*, **8**, e76957.
- 750 Schrödl M (2014) Time to say “Bye-bye Pulmonata”? *Spixiana*, **37**, 161–164.
- 751 Simakov O, Marletaz F, Cho S-J *et al.* (2013) Insights into bilaterian evolution from three spiralian
752 genomes. *Nature*, **493**, 526–531.
- 753 Simmons MP, Goloboff PA (2014) Dubious resolution and support from published sparse
754 supermatrices: The importance of thorough tree searches. *Molecular phylogenetics and
755 evolution*, **78**, 334–348.
- 756 Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison.
757 *BMC bioinformatics*, **6**, 31.

- 758 Smith SA, Wilson NG, Goetz FE *et al.* (2011) Resolving the evolutionary relationships of molluscs
759 with phylogenomic tools. *Nature*, **480**, 364–367.
- 760 Solem A (1992) Camaenid land snails from southern and eastern Australia, excluding Kangaroo
761 Island. *Records of the South Australian Museum, Monograph Series*, **2**, 1–425.
- 762 Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
763 thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.
- 764 Struck TH (2013) The impact of paralogy on phylogenomic studies - a case study on annelid
765 relationships. *PloS one*, **8**, e62892.
- 766 Struck TH (2014) TreSpEx—Detection of Misleading Signal in Phylogenetic Reconstructions Based
767 on Tree Information. *Evolutionary Bioinformatics*, 51–67.
- 768 Tatusov RL, Fedorova ND, Jackson JD *et al.* (2003) The COG database: an updated version includes
769 eukaryotes. *BMC bioinformatics*, **4**, 41.
- 770 Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive
771 multiple sequence alignment through sequence weighting, position-specific gap penalties and
772 weight matrix choice. *Nucleic acids research*, **22**, 4673–4680.
- 773 Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): High-
774 performance genomics data visualization and exploration. *Briefings in Bioinformatics*, **14**, 178–
775 192.
- 776 Upchurch P (2008) Gondwanan break-up: legacies of a lost world? *Trends in ecology & evolution*, **23**,
777 229–236.
- 778 Wade CM, Hudelot C, Davison A, Naggs F, Mordan PB (2007) Molecular phylogeny of the helicoid
779 land snails (Pulmonata: Stylommatophora: Helicoidea), with special emphasis on the
780 Camaenidae. *Journal of Molluscan Studies*, **73**, 411–415.
- 781 Wade CM, Mordan PB, Clarke B (2001) A phylogeny of the land snails (Gastropoda: Pulmonata).
782 *Proceedings. Biological sciences / The Royal Society*, **268**, 413–422.
- 783 Wade CM, Mordan PB, Naggs F (2006) Evolutionary relationships among the Pulmonate land snails
784 and slugs (Pulmonata, Stylommatophora). *Biological Journal of the Linnean Society*, **87**, 593–
785 610.
- 786 Ward N, Moreno-Hagelsieb G (2014) Quickly Finding Orthologs as Reciprocal Best Hits with BLAT,
787 LAST, and UBLAST: How Much Do We Miss? *PloS one*, **9**, e101850.
- 788 Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva E V. (2013) OrthoDB: a hierarchical
789 catalog of animal, fungal and bacterial orthologs. *Nucleic acids research*, **41**, D358–D365.
- 790 Wortley AH, Rudall PJ, Harris DJ, Scotland RW (2005) How much data are needed to resolve a
791 difficult phylogeny?: case study in Lamiales. *Systematic biology*, **54**, 697–709.
- 792 Xu H, Luo X, Qian J *et al.* (2012) FastUniq: A Fast De Novo Duplicates Removal Tool for Paired
793 Short Reads. *PLoS ONE*, **7**, e52249.
- 794 Zapata F, Wilson NG, Howison M (2014) Phylogenomic analyses of deep gastropod relationships
795 reject Orthogastropoda. *Proceedings of the Royal Society B: Biological Sciences*, **281**,
796 20141739.

797

798

799 DATA ACCESSIBILITY

800 Raw high-throughput sequence reads: NCBI Bioproject PRJNA304185

801 Transcriptome assemblies, gene and exon alignments for the transcriptome analyses, the

802 Camaenidae exon-capture probe set and the data sets used for phylogenetic inference: Dryad

803 (doi:10.5061/dryad.fn627)

804

805 AUTHOR CONTRIBUTIONS

806 LCT and AM designed the study. LCT lead the analysis with contribution from AM, TOH,
807 and KDM. LCT, AM and FK collected samples. LCT and AM wrote the manuscript. All
808 authors reviewed and edited the manuscript prior to submission.

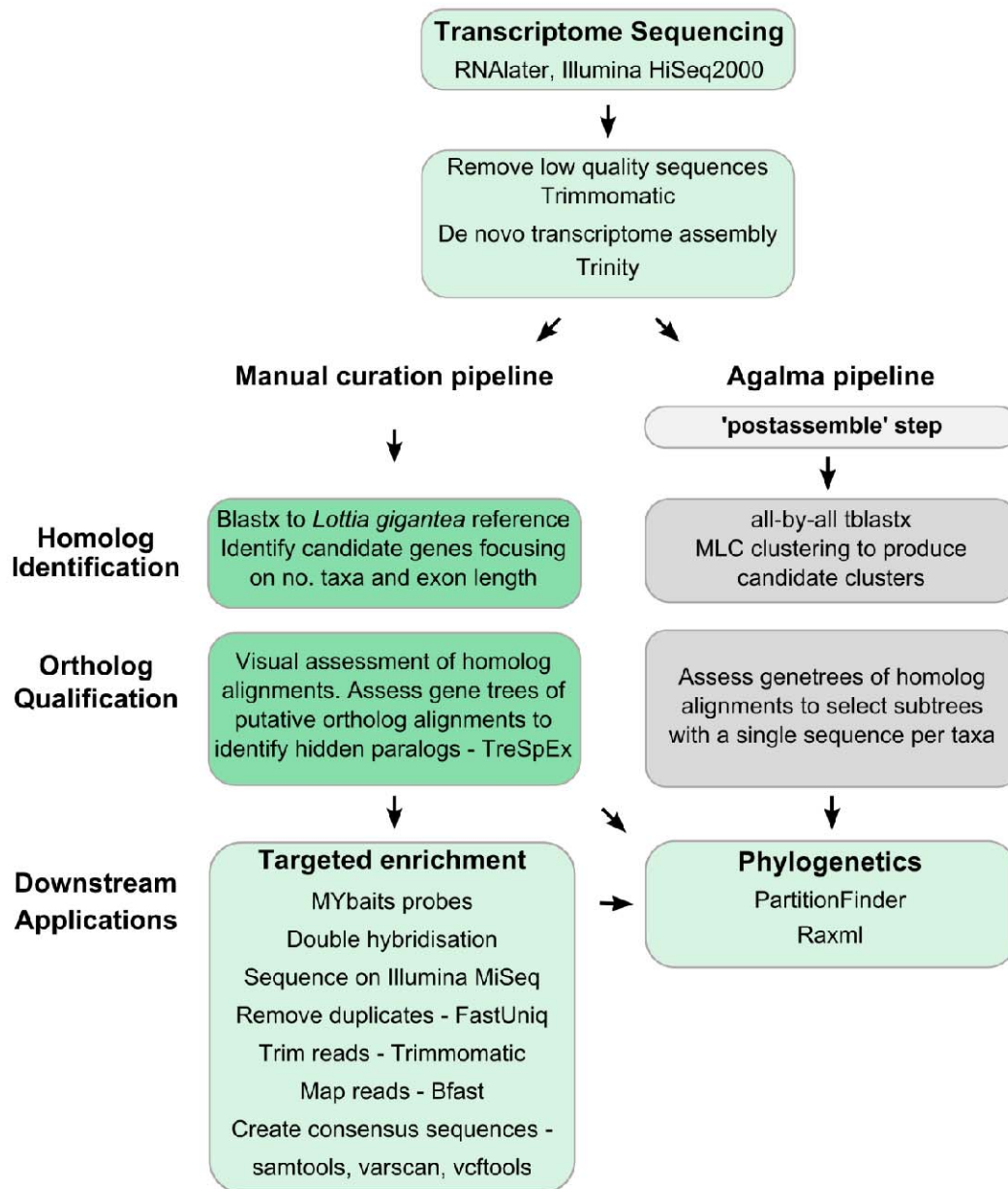
809

810

811

812

813



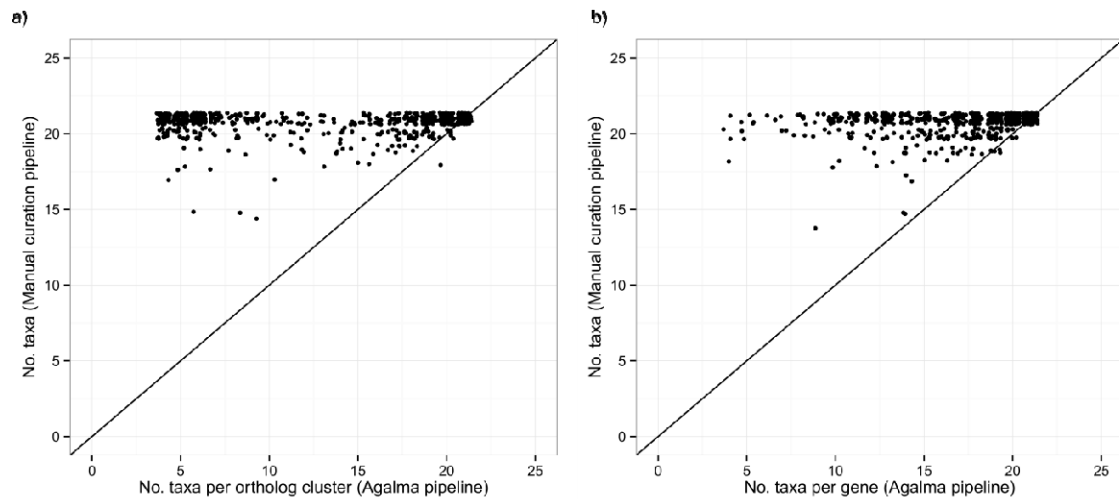
814

815 Fig. 1. *Analysis Pipelines*. Outline of the two pipelines used to detect single-copy,
816 orthologous genes from 21 eupulmonate transcriptomes.

817

818

819



820

821

822 Fig. 2. A comparison between two orthology detection pipelines. (a) shows the relationship
823 between the number of taxa per ortholog cluster for the ortholog clusters in common between
824 the manual curation and Agalma pipelines. The manually curated alignments resulted in more
825 taxa complete alignments than the corresponding Agalma alignments. (b) shows the same
826 relationship, however, the number of taxa per gene for the Agalma pipeline were calculated
827 across all ortholog clusters which matched the same *L. gigantea* gene. A comparison of the
828 two plots demonstrates that Agalma tended to produce multiple independent alignments per
829 *L. gigantea* gene, whereas a single alignment was produced through manual curation. Even
830 when the number of taxa recovered across all Agalma alignments associated with a given
831 gene are summed, taxa completeness of the Agalma dataset remained lower than that
832 obtained through manual curation (see also Figure 4e). These graphs are plotted using
833 `geom_jitter` in `ggplot2` to help visualise the large number of data points.

834

835

836

837

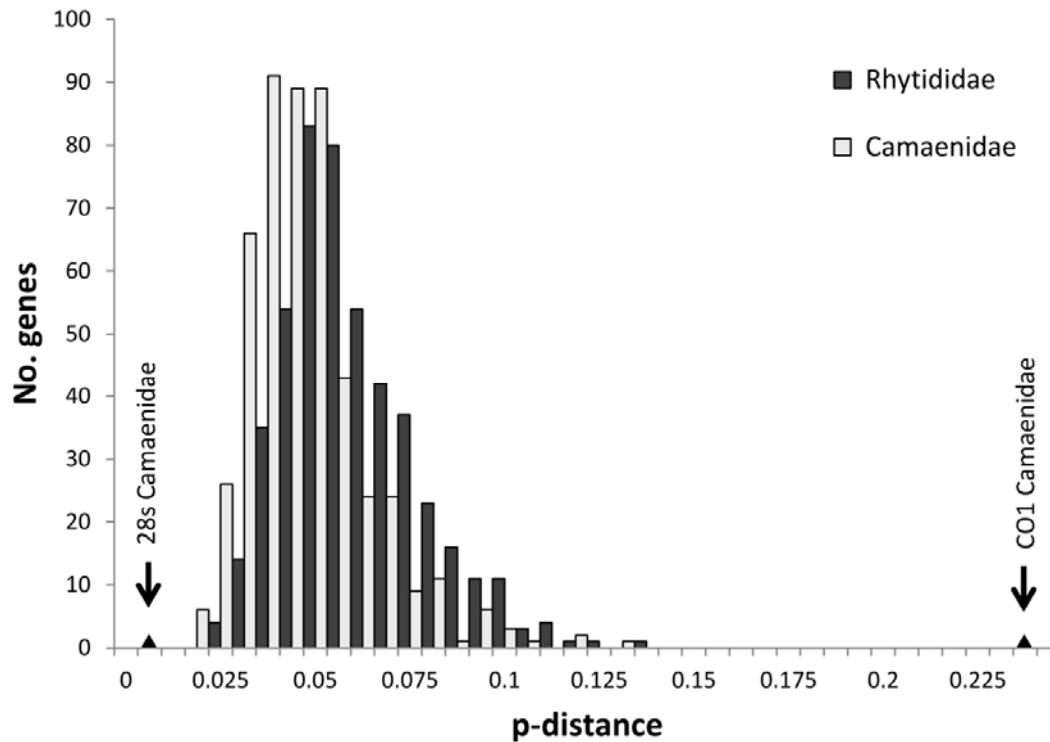
838

839

840

841

842



843

844 Fig. 3. Distribution of the p-distance for 500 single-copy orthologous genes across two
845 families. Uncorrected distances for both groups were calculated using alignments of
846 *Terrycarlessia turbinata* and *Victaphanta atramentaria* (Rhytididae), and *Austrochloritis*
847 *kosciuszkoensis* and *Sphaerospira fraseri* (Camaenidae). Triangles on the x-axis notate p-
848 distances of two commonly used phylogenetic markers, CO1 and 28S, for the Camaenidae.

849

850

851

852

853

854

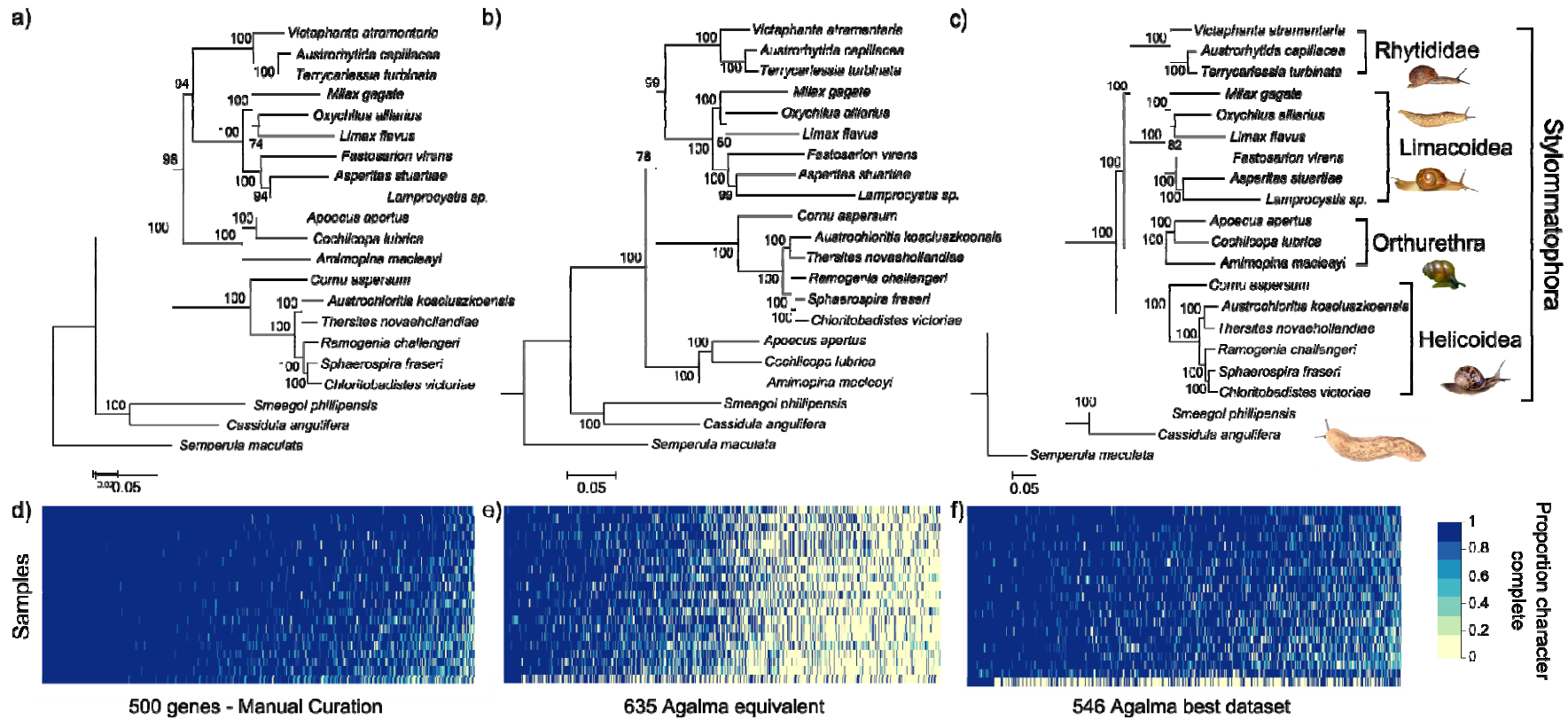


Fig. 4. *Maximum Likelihood phylogenies for 21 eupulmonates based on three datasets.* These datasets were (a) 500 nuclear single-copy, orthologous genes identified by manual curation, (b) 635 orthologous clusters identified by the automated pipeline Agalma, which correspond to the same 500 genes, and (c) 546 orthologous clusters identified by Agalma, where each orthologous cluster was the only one produced from the respective homolog cluster and had sequences for at least 18 taxa. Phylogenies are each based on analyses of amino acid sequences. Numbers on branches indicate bootstrap nodal support. Heat maps (d, e, f) indicate proportions of sequence obtained for each gene per sample for each dataset (sorted left to right by total proportion of data present per gene, top to bottom by total proportion of data present per sample).

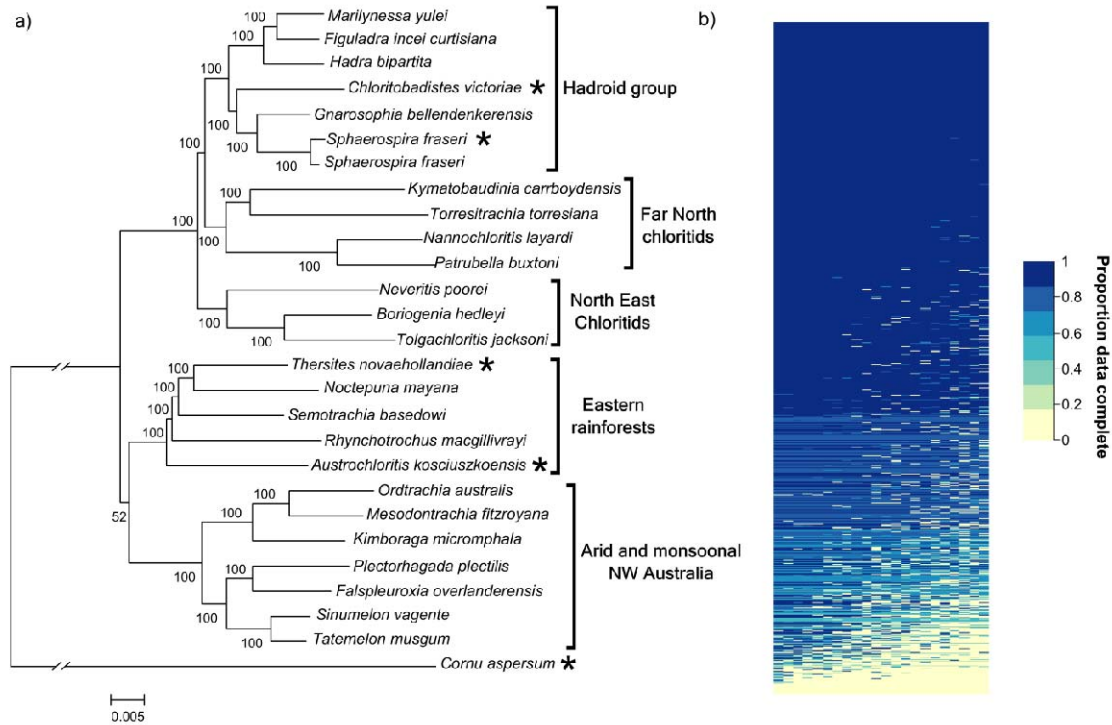


Fig. 5. *Maximum likelihood phylogeny of 26 Australian camaenid land snails.* (a) Phylogenetic reconstruction based on nucleotides sequences from 2,648 exons obtained through exon-capture. Sequences for the taxa marked with asterisks were derived from transcriptome datasets. Numbers on branches indicate bootstrap nodal support. (b) Heat map showing the proportion of available sequences for each sample per gene (sorted left to right by proportion of data present per sample; top to bottom by proportion of data present per exon).

Table 1. Taxon sampling: Transcriptome sequencing

Superfamilies or higher unranked classification	Family	Species	Voucher specimen	Collection locality*
Helicoidea	Camaenidae	<i>Austrochloritis kosciuszkoensis</i> Shea & Griffiths, 2010	NMV F193285	Sylvia Creek, VIC
Helicoidea	Camaenidae	<i>Chloritobadistes victoriae</i> (Cox, 1868)	NMV F193288	Crawford River, VIC
Helicoidea	Camaenidae	<i>Ramogenia challengerii</i> (Gude, 1906)	NMV F193287	Noosa, QLD
Helicoidea	Camaenidae	<i>Sphaerospira fraseri</i> (Griffith & Pidgeon, 1833)	NMV F193284	Noosa, QLD
Helicoidea	Camaenidae	<i>Thersites novaehollandiae</i> (Gray, 1834)	NMV F193248	Comboyne, NSW
Helicoidea	Helicidae	<i>Helix aspersa</i> Müller, 1774	NMV F193280	Melbourne, VIC
Limacoidea	Dyakiidae	<i>Asperitas stuartiae</i> (Pfeiffer, 1845)	NMV F193286	North of Dili, Timor-Leste
Limacoidea	Helicarionidae	<i>Fastosarion cf virens</i> (Pfeiffer, 1849)	NMV F193282	Noosa, QLD
Limacoidea	Limacidae	<i>Limax flavus</i> Linnaeus, 1758	NMV F193283	Melbourne, VIC
Limacoidea	Microcystidae	<i>Lamprocystis</i> sp.	AM C.476947	Ramelau Mountains, Timor-Leste
Limacoidea	Milacidae	<i>Milax gagates</i> (Draparnaud, 1801)	NMV F226625	Melbourne, VIC
Limacoidea	Oxychilidae	<i>Oxychilus alliarius</i> (Miller, 1822)	NMV F226626	Melbourne, VIC
Orthurethra	Cerastidae	<i>Amimopina macleayi</i> (Brazier, 1876)	NMV F193290	Darwin, NT
Orthurethra	Cochlicopidae	<i>Cochlicopa lubrica</i> (Müller, 1774)	MV614	Blue Mountains, NSW
Orthurethra	Enidae	<i>Apoecus apertus</i> (Martens, 1863)	AM C.488753	Ramelau Mountains, Timor-Leste
Rhytidoidea	Rhytididae	<i>Austrorhytida capillacea</i> (Férussac, 1832)	NMV F193291	Blue Mountains, NSW
Rhytidoidea	Rhytididae	<i>Terrycarlessia turbinata</i> Stanisc, 2010	NMV F193292	Comboyne, NSW
Rhytidoidea	Rhytididae	<i>Victaphanta atramentaria</i> (Shuttleworth, 1852)	NMV F226627	Toolangi, VIC
Ellobioidea	Ellobiidae	<i>Cassidula angulifera</i> (Petit, 1841)	NMV F193289	Manatuto, Timor-Leste
Otinoidea	Smeagolidae	<i>Smeagol phillipensis</i> Tillier & Ponder, 1992	MVR13_138	Phillip Is., VIC
Veronicelloidea	Veronicellidae	<i>Semperula maculata</i> (Templeton, 1858)	AM C.476934	Manatuto, Timor-Leste

*All localities within Australia unless otherwise indicated

Table 2. Taxon sampling: Transcriptome sequencing

Species	Voucher specimen	Collection locality*
<i>Boriogenia hedleyi</i> (Fulton, 1907)	MV1082	Cairns, QLD
<i>Falspleuroxia overlanderensis</i> Solem, 1997	WAM S70235	Shark Bay, WA
<i>Figuladra incei curtisiana</i> (Pfeiffer, 1864)	NMV F219323	Mt Archer, QLD
<i>Gnarosophia bellendenkerensis</i> (Brazier, 1875)	NMV F226513	Alligator creek, QLD
<i>Hadra bipartita</i> (Férussac, 1823)	AM C.476663	Green Island, QLD
<i>Kimboraga micromphala</i> (Gude, 1907)	AM C.463554	Windjana Gorge, WA
<i>Kymatobaudinia carrboydensis</i> Criscione & Köhler, 2013	WAM 49172	Carr Boyd Ranges, WA
<i>Marilynessa yulei</i> (Forbes, 1851)	MV1265	Brandy Creek, QLD
<i>Mesodontachia fitzroyana</i> Solem, 1985	AM C.476985	Victoria River District, NT
<i>Nannochloritis layardi</i> (Gude, 1906)	AM C.477826	Somerset, QLD
<i>Neveritis poorei</i> (Gude, 1907)	MV1054	Mt Elliot, QLD
<i>Noctepuna mayana</i> (Hedley, 1899)	AM C.478270	Diwan, QLD
<i>Ordtrachia australis</i> Solem, 1984	AM C.462736	Victoria River District, NT
<i>Patrubella buxtoni</i> (Brazier, 1880)	AM C.478884	Moa Is., Torres Strait
<i>Plectorhagada plectilis</i> (Benson, 1853)	WAM S70240	Shark Bay, WA
<i>Rhynchotrochus macgillivrayi</i> (Forbes, 1851)	AM C.478271	Diwan, QLD
<i>Semotrachia basedowi</i> (Hedley, 1905)	AM C.476884	Musgrave Ranges, WA
<i>Sinumelon vagente</i> Iredale, 1939	WA 61253	Mt Gibson, WA
<i>Sphaerospira fraseri</i> (Griffith & Pidgeon, 1833)	MV1104	Benarkin State Forest, QLD
<i>Tatemelon musgum</i> (Iredale, 1937)	AM C.476881	Musgrave Ranges, WA
<i>Tolgachloritis jacksoni</i> (Hedley, 1912)	NMV F226521	Mt Garnet, QLD
<i>Torresitrachia torresiana</i> (Hombron & Jacquinot, 1841)	AM C.477860	Weipa, Cape York Peninsula, QLD

*All localities within Australia unless otherwise indicated

Table 3. Summary statistics for sequencing and *de novo* assembly of 21 eupulmonate transcriptomes

Species	Pairs of raw reads	Proportion of reads after trimming	Trinity contigs	BLAST hits 1e-10 (<i>L. gigantea</i>)	<i>L. gigantea</i> genes with hits	No. of the 500 single copy genes
<i>Ramogenia challengerii</i>	11,726,377	0.84	103,471	14,665	7,011	488
<i>Austrochloritis kosciuszkoensis</i>	11,357,080	0.85	107,810	16,238	7,522	495
<i>Sphaerospira fraseri</i>	31,594,841	0.85	179,695	23,910	9,433	500
<i>Thersites novaehollandiae</i>	15,620,892	0.85	118,298	17,330	7,869	492
<i>Chloritobadistes victoriae</i>	26,433,009	0.85	148,817	20,453	8,792	498
<i>Amimopina macleayi</i>	7,874,195	0.97	93,250	17,258	8,091	494
<i>Cochlicopa lubrica</i>	8,074,560	0.97	111,396	21,675	9,086	497
<i>Asperitas stuartiae</i>	9,322,853	0.97	104,942	15,491	7,460	491
<i>Cassidula angulifera</i>	14,281,906	0.97	105,803	16,981	8,083	489
<i>Apoecus cf apertus</i>	9,362,182	0.97	119,711	21,275	9,095	497
<i>Fastosarion cf virens</i>	14,904,669	0.84	127,454	18,306	7,987	494
<i>Cornu aspersum</i>	21,273,910	0.86	160,490	23,114	9,254	498
<i>Limax flavus</i>	14,907,395	0.84	116,088	19,071	8,349	497
<i>Lamprocystis</i> sp.	22,539,699	0.97	128,611	23,797	9,679	499
<i>Milax gagates</i>	11,263,950	0.97	92,337	16,541	7,041	490
<i>Oxychilus alliarius</i>	12,925,111	0.97	136,044	21,183	8,940	499
<i>Terrycarlessia turbinata</i>	16,985,068	0.84	141,421	17,073	7,778	489
<i>Victaphanta atramentaria</i>	11,312,274	0.86	101,127	16,584	7,466	490
<i>Austrorhytida capillacea</i>	10,154,817	0.96	88,525	15,352	7,118	477
<i>Smeagol phillipensis</i>	6,393,571	0.96	95,429	23,067	9,699	497
<i>Semperula maculata</i>	12,461,924	0.97	76,847	21,851	9,276	492

Table 4. Sequencing and mapping summary statistics for the exon capture experiment.

Species	No. raw paired end reads	Proportion of pairs of reads retained after duplicate removal	Proportion retained after Trimmomatic	Proportion of reads mapped to the final reference	Average coverage per exon	Proportion of exons captured (total 2648 exons)
<i>Boriogenia hedleyi</i>	836,437	0.60	0.97	0.64	145	0.96
<i>Falspleuroxia overlanderensis</i>	170,769	0.69	0.98	0.74	41	0.88
<i>Figuladra incei curtisiana</i>	1,117,954	0.57	0.96	0.6	167	0.97
<i>Gnarosophia bellendenkerensis</i>	1,490,686	0.57	0.98	0.63	235	0.98
<i>Hadra bipartita</i>	659,509	0.6	0.98	0.7	131	0.96
<i>Kimboraga micromphala</i>	186,942	0.86	0.99	0.73	55	0.90
<i>Kymatobaudinia carrboydensis</i>	666,965	0.78	0.98	0.63	145	0.94
<i>Marilynessa yulei</i>	865,712	0.56	0.97	0.62	139	0.97
<i>Mesodontrachia fitzroyana</i>	429,572	0.85	0.98	0.61	102	0.91
<i>Nannochloritis layardi</i>	179,432	0.86	0.97	0.72	50	0.90
<i>Neveritis poorei</i>	1,313,049	0.57	0.96	0.62	205	0.95
<i>Noctepuna mayana</i>	297,503	0.77	0.98	0.73	81	0.93
<i>Ordtrachia australis</i>	670,743	0.65	0.94	0.86	222	0.92
<i>Patrubella buxtoni</i>	492,474	0.82	0.97	0.7	125	0.92
<i>Plectorhagada plectilis</i>	220,636	0.81	0.98	0.76	65	0.90
<i>Rhynchotrochus macgillivrayi</i>	340,338	0.85	0.98	0.7	96	0.92
<i>Semotrachia basedowi</i>	290,966	0.92	0.88	0.83	119	0.92
<i>Sinumelon vagente</i>	282,838	0.86	0.97	0.75	86	0.92
<i>Sphaerospira fraseri</i>	796,591	0.56	0.98	0.66	130	0.98
<i>Tatemelon musgum</i>	242,614	0.87	0.99	0.7	66	0.91
<i>Tolgachloritis jacksoni</i>	1,207,039	0.38	0.97	0.65	139	0.95
<i>Torresitrachia torresiana</i>	192,031	0.87	0.98	0.74	61	0.90