Meta-Barcoding Accelerates Species Discovery and Unravels the High Biodiversity of Benthic Invertebrates in Marine Sediments of the Campos Basin, Brazil

Milena MDP **Schettini**, Raony GCCL **Cardenas**, Marcella AA **Detoni**, Mauro F **Rebelo**.

Instituto de Biofísica Carlos Chagas Filho. Universidade Federal do Rio de Janeiro. Rio de Janeiro, Rio de Janeiro. Brasil.

KEYWORDS: molecular taxonomy, 18S rRNA, 28 S rRNA, COI, metagenomics

ABSTRACT

Biodiversity is currently assessed for characterization and monitoring of the environment through morphological taxonomy, a laborious and time-consuming process. We used 18S rRNA, 28S rRNA, and cytochrome *c* oxidase I together with next-generation sequencing and bioinformatics to identify benthic invertebrate organisms from sediment samples collected at five stations in the Campos Basin in southeast Brazil, an important oil extraction area with one of the best-studied marine biota in Brazil. A total of 3.3 million sequences were clustered in operational taxonomic units, and more than 1.6 million sequences (about 50% of all reads) were assigned to 957 prokaryotes and 577 eukaryotes. BLAST identified 23 phyla, 60 classes, 62 orders, 70 families, 67 genera, and 46 species of eukaryotes. Using meta-barcoding, we identified phyla that are traditionally found in samples of marine benthos, including Annelida, Arthropoda, Mollusca, and Chordata, as well as rare phyla such as Entoprocta and Gastrotricha. Taxa identified through meta-barcoding were compared to data obtained through morphology from previous studies in the area (REVIZEE, Habitats Project) and geo-validated with the Global Biodiversity Information Facility database. This is the first report of a number of taxa in the Campos area, and the large number of operational taxonomic units that were identified reveal a high level of benthic biodiversity in the Campos Basin that has not been previously reported. Our study supports the application of meta-barcoding for environmental characterization and monitoring programs, which could greatly reduce the time currently required for species identification and biodiversity determination.

INTRODUCTION

Sediment fauna characterization and monitoring are mandatory requirements for obtaining oil and gas (O&G) environmental permits for exploration and production (E&P) activities. This requirement is expected to remain a key element for environmental management in the future, particularly in the frontiers of deep-sea offshore oil exploration areas such as the Equatorial Margin and the Santos Pre-salt Basin in Brazil.

Biodiversity identification, which is required for environmental characterization and monitoring, is commonly carried out by morphological taxonomy, a laborious and time-consuming process. As a general rule, taxonomic resolution at the species level is expected, but for some fauna groups, the expertise required for this task is so specialized that only a handful of individuals in the world are qualified to perform it. In addition, expert judgment is never 100% accurate, and a 50% rate of identification consistency among taxonomists has been reported (Culverhouse et al., 2003). The occurrence of pseudo-absence is frequent, especially for fragile organisms that require special fixation procedures (Costa-Paiva et al., 2007). As a result, invertebrate morphological identification efforts are often limited to few groups, including Mollusca, Crustacea, and Polychaeta (Scaramuzza, 2015), while some estimates suggest that more than 90% of all marine species have never been named (Scheffers et al., 2012).

The typical number of sediment samples in a monitoring campaign is in the range of tens, but in sedimentary basins as large as 300,000 km^2, this number can extends to tens of thousands of samples for baseline environmental characterization. The lack of expertise for morphological taxonomy of some groups is a major bottleneck in the process of identifying biodiversity (Hebert et al., 2003; Mora et al., 2013), and as a result, taxonomists are frequently unable to meet the demands for biodiversity assessment required for monitoring programs, delaying the development of economic activities and the discovery of new species.

According to the latest Report of the Convention on Biological Diversity (Diversity, 2016), Brazil is the most biologically-diverse country in the world, with more than 100,000 animal species described. However, only 184 marine invertebrates have had their conservation status assessed

(Scaramuzza, 2015). It is possible that current risk estimates of environmental impact are based on underestimated biodiversity inventories, representing a threat to species conservation (Wu, 1982). Developing new technologies and approaches that accelerate species discovery and reveal hidden biodiversity is crucial for setting conservation priorities and efforts.

Meta-barcoding uses genetic marker data generated through high-throughput next-generation sequencing (NGS) of environmental samples (Leray and Knowlton, 2015) to greatly accelerates species discovery and assess biodiversity. Since 2010, more than 600 papers have been published on the use of DNA-based identification methods for species conservation (Goldberg et al., 2015; Bergman et al., 2016), biodiversity inventory determination (Drummond et al., 2015), environmental monitoring (Bohmann et al., 2014; Chariton et al., 2015; Leray and Knowlton, 2015; Brown et al., 2015), and DNA extraction/detection (Pedersen et al., 2014; Eichmiller et al., 2014; Ficetola et al., 2016), and the technique has been considered a major tool for ocean sustainability in the 21st century (Aricò, 2015). This approach is particularly useful because of its sensitivity in identifying minute organisms and species in sediment (Wang et al., 2014). For eukaryotic organisms that have not yet had their genetic markers sequenced or have not yet been described morphologically, the concept of operational taxonomic unit (OTU) can be applied (Stackebrandt and Goebel, 1994; Pedersen et al., 2014).

In this study, we combined three different phylogenetic markers (18S rRNA, 28S rRNA, and cytochrome *c* oxidase subunit I [COI]) together with NGS and bioinformatics to identify benthic invertebrate organisms using metagenomes from sediment samples collected in the Campos Basin in southeast Brazil, an important oil extraction area with one of the best-studied marine biota in Brazil (Miloslavich et al., 2011).

MATERIAL AND METHODS

Sample collection and processing:

Samples were collected in a survey in 2009 as part of the Habitats Project – Campos Basin Environmental Heterogeneity coordinated by CENPES/PETROBRAS. Table 1 presents

information (collection date, geographic coordinates, and depth) on the five sampling stations (B3, B4, C2, G2, and F5) in the Campos Basin. Sediment samples were collected in triplicate by lowering a Van Veen grab at three different points around (150 m radius) each of the five stations, resulting in a total of 15 sediment samples. At the time of collection, there were no plans to have them genetically analyzed, and thus they were kept at –20°C for 4 years until our analysis was done in 2013.

A 200-g subsample of the 0–2 cm slice of sediment of each sample was manually homogenized, and DNA was extracted from 5 g of this subsample using the PowerMax Soil DNA Isolation kit (MoBio Inc) according to the manufacturer's instructions. DNA integrity was checked on a 1.2% agarose gel. Quantification was performed using a Qubit 2.0 Fluorometer (Life Technologies).

Biogeography data:

Biogeography data on the organisms identified in this study were extracted from previous studies. Data on Cnidaria, Crustacea, Echinodermata, Mollusca, Nematoda, Polychaeta, and Porifera groups were taken from the Brazilian REVIZEE (Living Resources in the Exclusive Economic Zone) program (Lavrado and Ignacio, 2006), whereas the data for organisms of the phyla Annelida, Arthropoda, Brachiopoda, Bryozoa, Cnidaria, Echinodermata, Echiura, Foraminifera, Haptophyte, Mollusca, Nematoda, Nemertea, Porifera, Priapulida, Protozoa, and Rodophyta were taken from the Habitats Project and provided by CENPES/PETROBRAS (unpublished data). We also used the Global Biodiversity Information Facility (GBIF) database (www.gbif.org) for organism geo-localization.

PCR and high-throughput sequencing:

Information on PCR of the COI, 18S rRNA, and 28S rRNA genes is presented in Figure 1 in Supplementary Material. We used the Ion Xpress™ Plus Fragment Library kit (Life Technologies) for preparing the libraries for sequencing according to the manufacturer's instructions for gDNA fragment library preparation. Template preparation and sequencing were done using the Ion

PGM™ Template OT2 400 kit. Sequencing was done using the Ion Personal Genome Machine (PGM™) System at the Life Technologies laboratories (São Paulo, SP), using Chip 318 v2.

Bioinformatics and Taxonomic Name Attribution:

Reads were prefiltered using the Torrent Suite software version 4.0.2 (Life Technologies) and assigned to samples based on a combination primer tail-Ion Xpress barcode. PRINSEQ version 0.20.4 (Schmieder and Edwards, 2011) was used to remove poly A/T tails longer than 5 bases, reads with unidentified (N) bases, reads shorter than 80 bp, and bad quality reads (Q<20). The remaining reads were clustered in operational taxonomic units (OTUs) using CD-HIT-EST version 4.6 (Li and Godzik, 2006) (up to 97% identity under 100% coverage within a bigger read, word size of 10, and 20 penalty points for gaps).

High quality and low redundancy sequences were compared to NCBI (http://www.ncbi.nlm.nih.gov) non-redundant nucleotide repositories (NR) using the Nucleotide Basic Local Alignment Search Tool (BLASTn) version 2.3.0+ (Zhang et al., 2000). Max e-value was of 10^{-5} and the number of events (referred to henceforth as 'hits') per query was limited to 100.

Taxonomic names were attributed to each read based on the read's group of BLAST hits using the Lowest Common Ancestor Assignment (LCA) algorithm in the MEGAN software (MEta Genome Analyzer v. 5.10.3; Huson et al., 2007) with parameter adjustment (Huson et al., 2011). Cladograms and rarefaction curves at the family taxonomic level were also built for each station using MEGAN.

The BLAST step was performed using the Elastic Compute Cloud (EC2) service of Amazon (aws.amazon.com). The BLAST for each of the 15 sets of reads corresponding to the 15 samples was run in a parallel scheme using eight threads on up to 96 AWS instances with 8 processors and 16 Gb of RAM each.

RESULTS

We obtained an average of 4.83 µg of DNA from each of the 15 samples. Sequencing generated approximately 4.8 million sequences with an average size of 155.1 bp. Over 3.6 million sequences (75.35%) passed quality control, and of these around 3.3 million were clustered in OTUs by CD-HIT. Table 2 shows (1) the total number of OTUs, (2) the number of OTUs with no hits in BLAST, (3) the number of OTUs with reads not attributed to any taxa by LCA, and (4) the number of OTUs with attributed reads. For the five sampling stations, more than 1.6 million sequences (about 50% of all reads) were assigned to 957 prokaryotes and 577 eukaryotes by the LCA algorithm in MEGAN using hits produced by the BLAST similarity algorithm with any of the three molecular markers (18S rRNA, 28S rRNA, or COI). LCA further identified 23 phyla, 60 classes, 62 orders, 70 families, 67 genera, and 46 species of eukaryotes. Figure 1A shows the distribution of the 13 invertebrate phyla OTU identified by meta-barcoding for each of the five stations, and Figure 1B shows the same for the 38 invertebrate families OTU identified. All other prokaryotes and eukaryotes identified in this study by any of the three molecular markers and classified to the taxonomic level of family are listed in the cladograms available in Supplementar Material 2 in for each of the five sampling stations.

Our analysis identified 38 families of invertebrates in the 15 samples from the five sampling stations in the Campos Basin. Figure 2 compares the spatial distribution of families identified by meta-barcoding from phyla with most abundant frequencies (Annelida, 9 families, Fig. 2A; Arthropoda, 10 families, Fig. 2B; and Mollusca, 7 families, Fig. 2C) in relation to previously published morphologic taxonomy results for stations B3, B4, C2, F5 and G2.

Initially, the LCA algorithm identified 46 species, of which 27 were invertebrates not previously described in the region. A text search of the list of BLAST hits allowed for identification of an additional 45 species of invertebrates that had previously been identified in the Campos Basin. The full list of species identified in this study is in Supplementary Material 3.

DISCUSSION

In this study we report the first meta-barcoding description of eukaryote biodiversity in the deep-sea Brazilian continental shelf. Of the 3.3 million sequences that were classified as OTUs, more than 1.6 million were assigned to 957 prokaryotes and 577 eukaryotes. Even though the association between a given OTU and a given species should be established carefully, the remaining 1.6 million OTUs that were not identified in the current study based on the sequences of genetic markers available in Genbank suggests that the benthic biodiversity of the Campos Basin could be orders of magnitude higher than that established by previous studies based on morphological taxonomy.

One of the differentiating characteristics of our study is the fact that we used samples collected from areas were E&P activities have been carried out and where several previous morphological taxonomic studies were performed, either by oil companies as part of the process to obtain environmental permits or those involved in conservation programs (such as the Habitats Project) or by the scientific community (specially the REVIZEE program).

The approximately 4.8 million sequences we found are within the expected range for the 318 v2 chip. Even though the average read size of 155.1 bp was below the expected number for the OT2 400 kit, it did not compromise our analysis.

When further analyzing the OTUs distributed in the 23 phyla, we found that a considerable number of reads were assigned to the families Hominidea and Bovidae, increasing the number of reads belonging to the Chordate phylum. However, these were the result of read alignments generated against the whole human and bovine genomes or chromosomes, as opposed to the three specific genetic markers. As the focus of this study was on benthic invertebrates because of their significance for legally mandated environmental characterization and monitoring in offshore areas, these artifact findings in the Chordata were ignored.

Our meta-barcoding analysis identified phyla that are traditionally found in samples of marine benthos, including Annelida, Arthropoda, Mollusca and Chordata, as well as more rarely found phyla such as Bryozoa, Cnidaria, Echinodermata, Nematoda, Nemertea, Platyhelminthes,

Porifera, and Priapulida, and even rarer phyla such as Entoprocta and Gastrotricha (Figure 1 and Figure 2 in Supplementary Material).

The large number of OTUs for Annelida, Arthropoda, and Mollusca found by meta-barcoding agrees with previous results for the Campos Basin (Lavrado and Ignacio, 2006) obtained by the REVIZEE project and also with those from the Habitats Project. A recent meta-barcoding study (Leray and Knowlton, 2015) also identified Annelida and Arthropoda as the phyla with the most OTUs among the 22 phyla identified in approximately 0.09 m^3 of sediment from coral reef regions in Virginia and Florida in the United States.

The Entoprocta (or Kamptozoa) phylum comprises about 170 aquatic and sessile species that are between 0.5 and 5.0 mm in size and are mostly marine (Zhang, 2011). Until 2011, only 18 species of Entoprocta were known on the Brazilian coast (Vieira and Migotto, 2011). In this study, all Entoprocta OTUs (6 at the C2 station and 24 at the G2 station) were attributed to the genus *Loxosomella* via the 28S rRNA marker and had over 86% similarity to sequences found in Genbank. This result expands the distribution of the genus, which was previously limited to six species collected off the coast of São Paulo (Vieira and Migotto, 2011). As for the cosmopolitan Gastrotricha phylum that comprises about 790 species of aquatic organisms up to 1 mm in length (Zhang, 2011), all 22 OTUs assigned to the phylum (C2 station) were in the *Tetranchyroderma* genus, with over 81% similarity to COI sequences found in Genbank. This finding also expands the distribution of this genus, which was previously limited to São Paulo beaches located approximately 1000 km away from the Campos Basin (Garraffoni and Araújo, 2010).

Our meta-barcoding results were compared to those from a recent comprehensive morphological taxonomy effort (the Habitats Project coordinated by CENPES/PETROBRAS) that analyzed the same samples used in our study. The Habitats Project generated a databank with data for almost 50,000 specimens and identified 17 phyla, 27 classes, 63 orders, 354 families, 768 genera, and 749 species. The comparison between the findings obtained with molecular and morphological taxonomies was restricted, however, since 1211 (68%) of the 1773

macroinvertebrate taxa identified by morphological taxonomy did not have an entry in Genbank for any of the three markers (18S rRNA, 28S rRNA, and COI) used in this study (indicating a major underrepresentation of Brazilian marine species in Genbank and a need to increase efforts to have sequences from more Brazilian species deposited in that database). Other factors that limited the comparison between the two taxonomic approaches included uncertainty about how much DNA was still available in the sediment samples that had been preserved at –20°C for 4 years as well as the limited amount of sample analyzed at each station (5 g of the 200 g of the 0–2 cm slice of sediment that was homogenized, compared to 4 L of the 0–10 cm slice for the morphological study). Finally, for many species, only partial sequences of the markers were available in Genbank, and thus it was not possible to ensure that they were sufficiently aligned with a given read to permit assignment to a taxon. These restrictions suggest that some families that were apparently absent may not actually have been so. The continuing effort to add more sequences to Genbank should clarify this issue.

Out of the 70 families identified by meta-barcoding, 21 were invertebrate. The families Amphinomidae, Enchytraeidae, Glyceridae, Orbiniidae, Serpulidae, and Spionidae, belonging to the phylum Annelida, were previously identified in the Campos Basin by the Habitats Project, which also identified 28 other Annelida families not found by meta-barcoding. Hormogastridae, which was found in our study, is most likely a false positive since it is not a marine family.

The families Solenoceridae, Cylindroleberididae, and Mysidae, belonging to the phylum Arthropoda, have previously been identified in the Campos Basin and in southeastern Brazil by other authors (Cardoso, 2007; Serejo et al., 2007; Tâmega et al., 2013), while 29 arthropod families previously reported by the Habitats Project were not identified by meta-barcoding. The families Miridae, Chalcididae, and Formicidae, all of which were found in our study, are most likely false positives since they are non-marine insects.

All Mollusca families identified by meta-barcoding in the Campos basin except for Mytilidae have previously been found in the region (Lavrado and Ignacio, 2006; Dornellas and Simone, 2011;

239 Tâmega et al., 2013), although not by the Habitats Project. The latter identified 15 Mollusca
240 families not identified by meta-barcoding.
241 Meta-barcoding was also able to find families at every sampling station that had not been
242 previously reported by the Habitats Project, such as Echiuridae and Pectinariidae in the
243 Annelida; Desmosomatidae and Hippolytidae in the Arthropoda; and Arcidae, Mactridae, and
244 Pectinidae in the Mollusca. This indicates that the distribution of these families may be broader
245 than that suggested by morphological taxonomy.
246 Of the 46 species identified by meta-barcoding, none of the 24 benthic invertebrates had been
247 previously described by the Habitats Project and may represent new occurrences in the region,
248 with the exception of the arthropod *Eurythenes gryllus*, previously identified in REVIZEE. We
249 must remember that even though species level resolution is desirable, taxonomic penetration to
250 the family level is accepted by environmental agencies, and most specimens in previous studies
251 have been identified only to this level. We found records of the families of all newly observed
252 species in the Habitats Project, REVIZEE and GBIF databases, which supports the argument that
253 these are new occurrences, even though we cannot rule out the possibility of false positive.
254 The comparison between our data and that of the Habitats Projects was limited by the
255 availability of sequences of the three genetic markers (18S rRNA, 28S rRNA and COI) deposited
256 in Genbank. Only 64 out of the 749 organisms identified to the species level by the Habitats
257 Project had at least one genetic marker sequence found in Genbank and thus were eligible for
258 molecular identification. However, none of those 64 species were identified by meta-barcoding.
259 We believe that these 64 missing species are pseudo-absence results, and that they were not
260 found because the samples had been preserved at –20°C for 4 years. However, another
261 explanation is possible. When analyzing our data, we noticed that even after calibration of the
262 parameters for the LCA algorithm (data not shown), some incongruence in the attribution of the
263 taxonomic name to a species could happen due to the selection of an unlikely BLAST hit to name
264 the query OTU. To overcome this problem, we text-searched the names of the organisms of all
265 BLAST hits that were associated to a given OTU and compared the names found for those of the

64 species found by the Habitat Project. We were thus able to identify 45 additional species that had previously been described by morphological taxonomy but were not picked by the LCA algorithm. The full list of species identified by molecular and morphological taxonomies, together with their genetic markers that are available in Genbank, are listed in Supplementary Material 3. Other pseudo-absence results could have been generated by the occurrence of synonymous names at the species level. For instance, according to recent estimates, more than 80% of the algae in some genera and 38% of Mollusca have synonymous names. For marine species, this percentage could reach 40% (Costello et al., 2013). An ongoing effort is dedicated to resolving synonymous names found in the GBIF database.

The use of biogeographic databases (Habitats Project, REVIZEE and GBIF) to verify and adjust the meta-barcode observations has proven to be a good strategy. False positive results can be an artifact of the low representativeness of Brazilian species in Genbank. Due to similarities of genetic sequences shared among species belonging to the same genus, BLAST can associate, with very low error probability, a read from a species not present in the Genbank to a sequence from a phylogenetically similar species from a different habitat. By using metadata on the distribution of the species selected by BLAST, we managed to identify at least one instance of this occurring in our analysis. The small (25–85 mm) gastropod *Haliotis diversicolor*, which was identified in our study, is native to the Indo-Pacific Ocean, with geo-referenced records on the coast of Japan, Thailand, and Australia (GBIF, 2016). Thus this result might have represented a new occurrence of this species in a completely new environment or, alternatively, a false positive. *Haliotis aurantium*, a small gastropod belonging to the same genus as *H. diversicolor*, has previously been identified in the Campos Basin, and we believe that because Genbank contained no sequence of *H. aurantium* corresponding to any of the three genetic markers, the LCA algorithm may have erroneously assigned an OTU from *H. aurantium* to *H. diversicolor.* A system able to resolve these types of incongruences would greatly improve meta-barcoding analysis.

In an attempt to prevent false positive results, we tested whether redundant identification by a second or third genetic marker could confirm potential positive results of species that were

identified by a single marker. Unfortunately, that was not the case. Of the 46 species identified by meta-barcoding, 16 had sequences of all three genetic markers available in Genbank but were identified by just one of the three markers and not by the others. It was frequently the case that, even though the sequence for a genetic marker for a specific organism was available in Genbank, multiple names were attributed to the gene, only partial sequences were available, or sequences had not been validated experimentally. Genbank is the best repository for genetic sequences yet available, but it still does not offer a high level of confidence when it comes to the names attributed to genetic sequences. Our research team is currently working on developing new algorithms to help overcome this limitation.

The problems related to the presence of false positives and pseudo absences could be solved if biodiversity characterization was done in a taxonomic-free context, looking only at OTUs to compare biodiversity profiles among samples. The frequency and abundance of OTUs could then be related to environmental changes, either spatial or seasonal, and species discovery would be accelerated by identification of OTUs that vary according to environmental conditions. Were such a strategy to be adopted, not only would it allow us to work with the hidden biodiversity of the thousands of 'no hit' OTUs, but OTU profiles and their distribution could inform us of environmental changes. Species names are fundamental to ecology, and in spite of all the uncertainty that they entail, a lot of the accumulated knowledge in biology is associated with these units. Although we may never give up on the idea of naming species, the ease of gathering OTU data is unprecedented, making the prospect of a taxonomic-free ecology complementary to the traditional one more and more likely in the years to come.

CONCLUSION

This study confirms that meta-barcoding can be a reliable, fast, and low cost tool for environmental characterization and monitoring. It may be the only suitable method for producing information critical for decision making about extensive, little-explored areas within a short time period, thus safeguarding the environment without delaying economic activities. It

may also accelerate species discovery and contribute to ecological knowledge in ways that have not yet been fully explored.

The methodology could be improved by adding more sequences of native species to public and proprietary databases, but it is our opinion that meta-barcoding can already be considered to be the best available technique for generating biodiversity inventories in marine sediments, and it should be acknowledged as such by oil operators, environmental authorities, and the scientific community at large.

Brazil has one of the strictest set of environmental laws and regulations for the O&G sector in the world, and one that is constantly being improved. Recent changes made under resolution CONAMA 422/11 minimized bureaucracy in the application process, increased transparency by sharing information online, and reduced liability for the O&G operators. The Brazilian environmental authority IBAMA (Brazilian Institute of the Environment and Renewable Natural Resources) establishes guidelines and best practices for environmental licensing and monitoring by means of 'reference terms.' By becoming an early adopter of meta-barcoding, IBAMA could play a leading role in the implementation of this innovative methodology that can greatly contribute to the conservation of deep-sea environments worldwide.

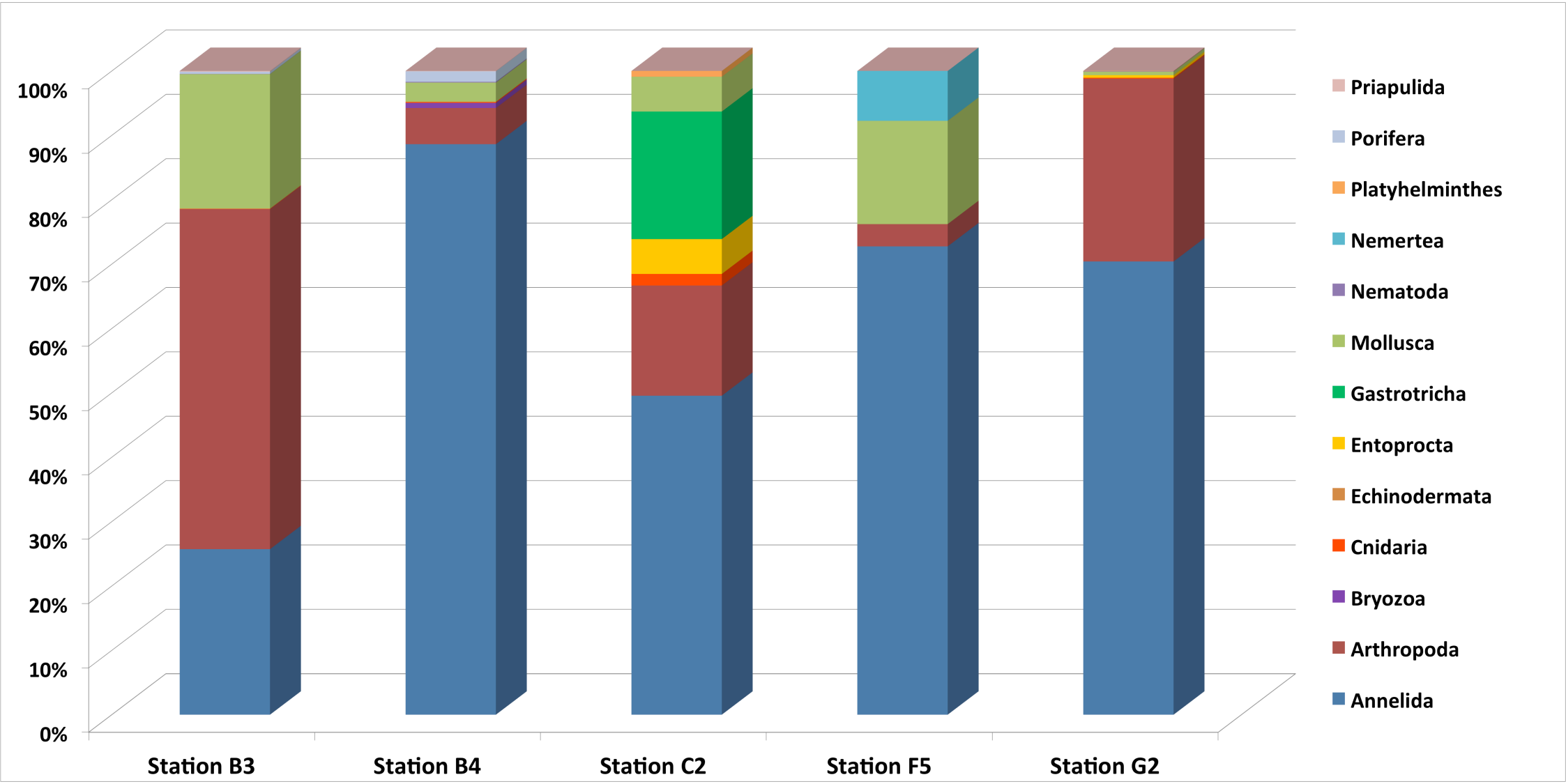
Table 1 – Sampling date, location and depth Location of sampling stations B3, B4, C2, F5 and G2 in Campos Basin, southeast Brazil.

	Sampling date	Latitude (SIRGAS2000)	Longitude (SIRGAS2000)	Depth (m)
Station B3	02/20/2009	-22,99701	-41,352583	77
Station B4	02/21/2009	-23,16851	-41,052264	107
Station C2	07/16/2009	-22,62599	-41,365082	54
Station F5	02/24/2009	-22,29010	-40,110584	143
Station G2	02/25/2009	-21,98502	-40,419918	56

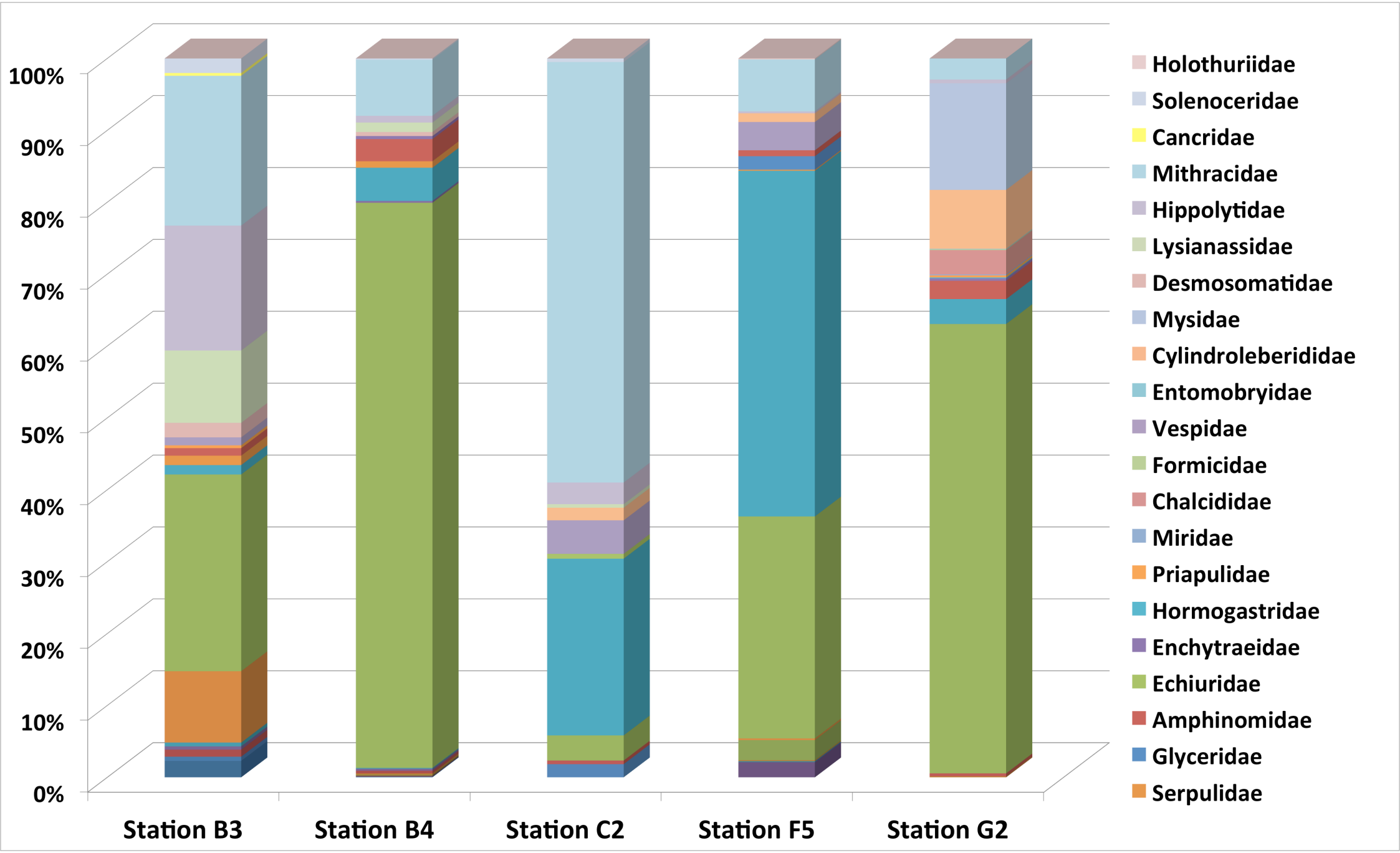
Table 2 – OTU per sample. OTU without a similar sequence on Genbank NR are under 'No fragments . OTU that did not comply with established LCA parameters (e.g. score bellov not add up to a node are under 'non attributed reads'. Also under 'non-attributed' are attributed by rRNA16S, taxa attributed by genes other than the 3 targets and taxa defir Genebank as 'undefined'. They were also disabled at the cladograms.

Sample	Total OTU	No Hits	Non attributed	Attributed
St. B3 rep. #1	101,966	20,505	73,653	7,808
St. B3 rep. #2	379,812	65,557	97,849	222,406
St. B3 rep. #3	84,180	12,167	57,290	14,723
St. B4 rep. #1	103,053	25,721	57,290	14,723
St. B4 rep. #2	332,953	35,384	64,066	236,503
St. B4 rep. #3	302,290	50,143	65,134	187,013
St. C2 rep. #1	245,233	34,452	40,687	170,094
St. C2 rep. #2	307,780	59,289	60,866	187,625
St. C2 rep. #3	249,969	56,247	81,114	112,608
St. F5 rep. #1	139,992	50,900	35,349	53,743
St. F5 rep. #2	105,435	32,435	47,684	25,316
St. F5 rep. #3	83,962	43,377	34,877	5,708
St. G2 rep. #1	173,740	71,230	60,632	41,780
St. G2 rep. #2	312,446	88,627	79,156	144,663
St. G2 rep. #3	347,494	32,832	120,519	194,143
TOTAL	3,270,206	678,866	959,986	1,631,453

Figure 1 – OTU occurence in each station. Percentage of OTU for phyla (A) and Family (B) in each station.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2103v3 | CC BY 4.0 Open Access | rec: 22 Jul 2016, publ: 22 Jul 2016

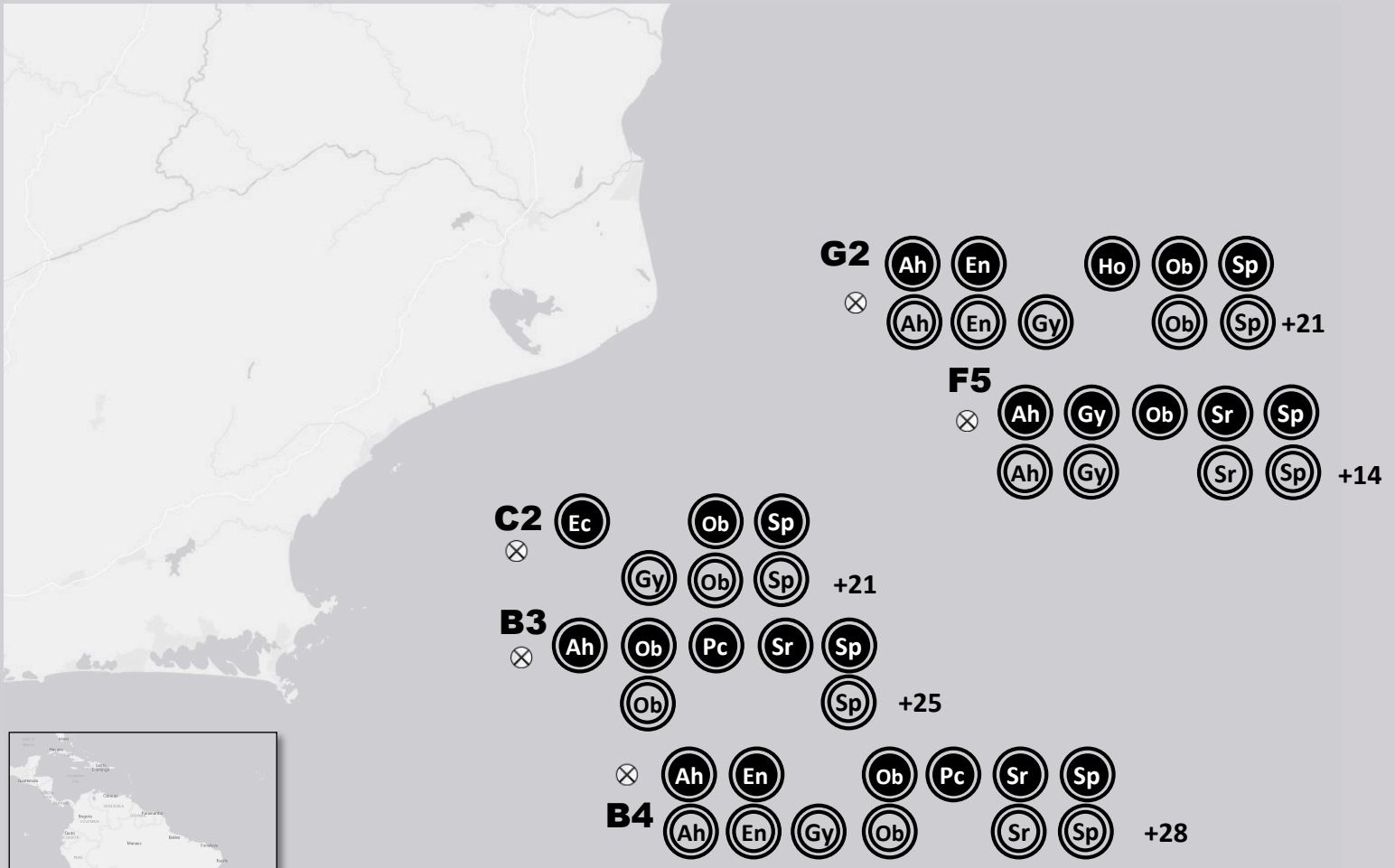


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2103v3 | CC BY 4.0 Open Access | rec: 22 Jul 2016, publ: 22 Jul 2016

Figure 2 – Distribution of the main invertebrate phylum identifyed by molecular and morphological taxonomy in Campos Basin. A) annelida distribution, b) arthropoda distribution, C) mollusca distribution.

Annelida distribution

Molecular | Morphological



- Ah Amphinomidae
- En Enchytraeidae
- Ec Echiuridae*
- Gy Glyceridae
- Ho Hormogastridae**
- Ob Orbiniidae
- Pc Pectinariidae*
- Sr Serpulidae
- Sp Spionidae

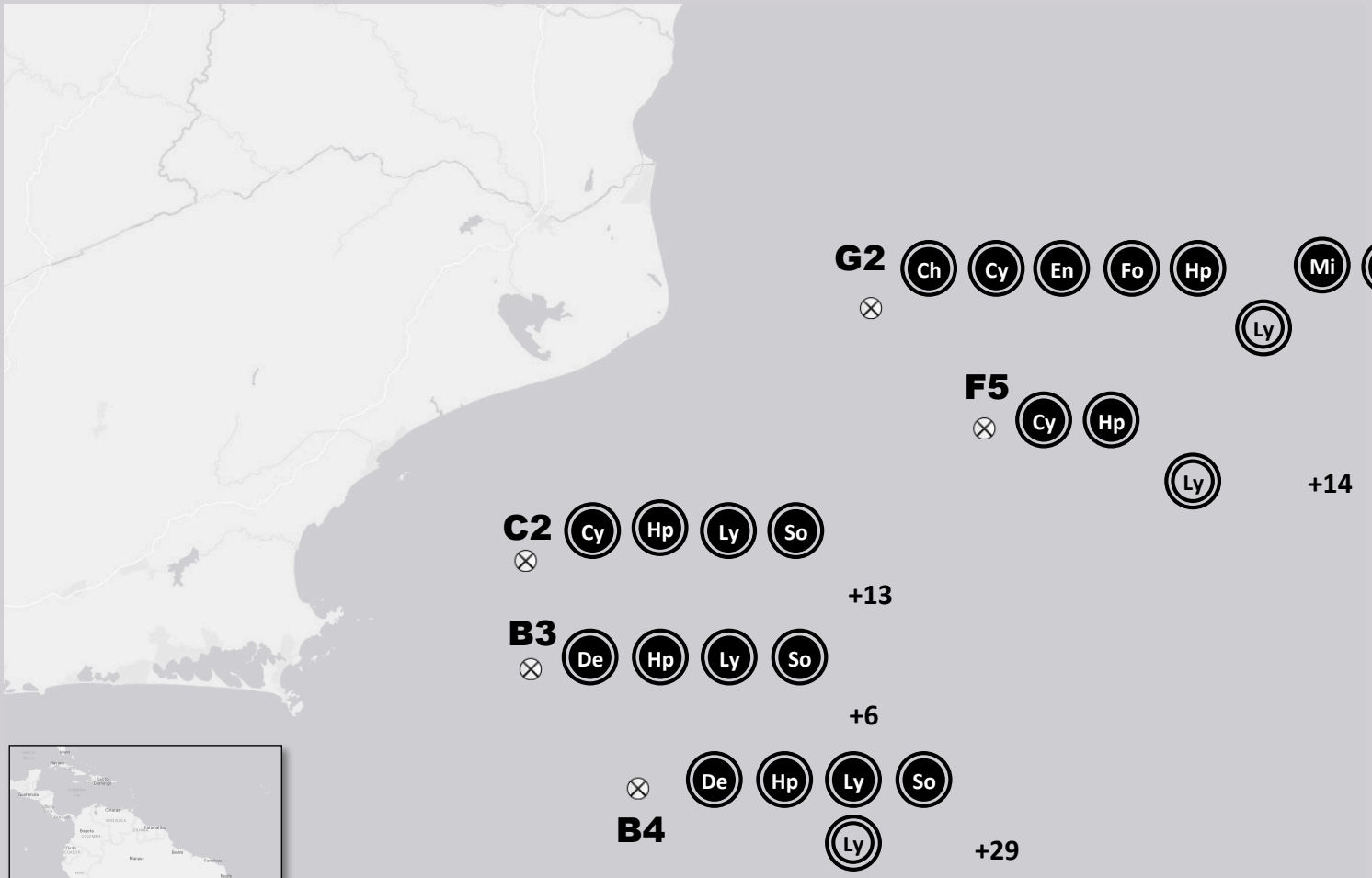
* Present in other stations of Habitats

** Non-marine family

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2103v3 | CC BY 4.0 Open Access | rec: 22 Jul 2016, publ: 22 Jul 2016

Arthropoda distribution

Molecular | Morphological



0 15 30km

- Ch Chalcididae***
- Cy Cylindroleberididae**
- De Desmosomatidae*
- En Entomobryidae
- Fo Formicidae***
- Hp Hippolytidae*
- Ly Lysianassidae
- Mi Miridae***
- My Mysidae**
- So Solenoceridae**

* Present in other stations of Habitats Project

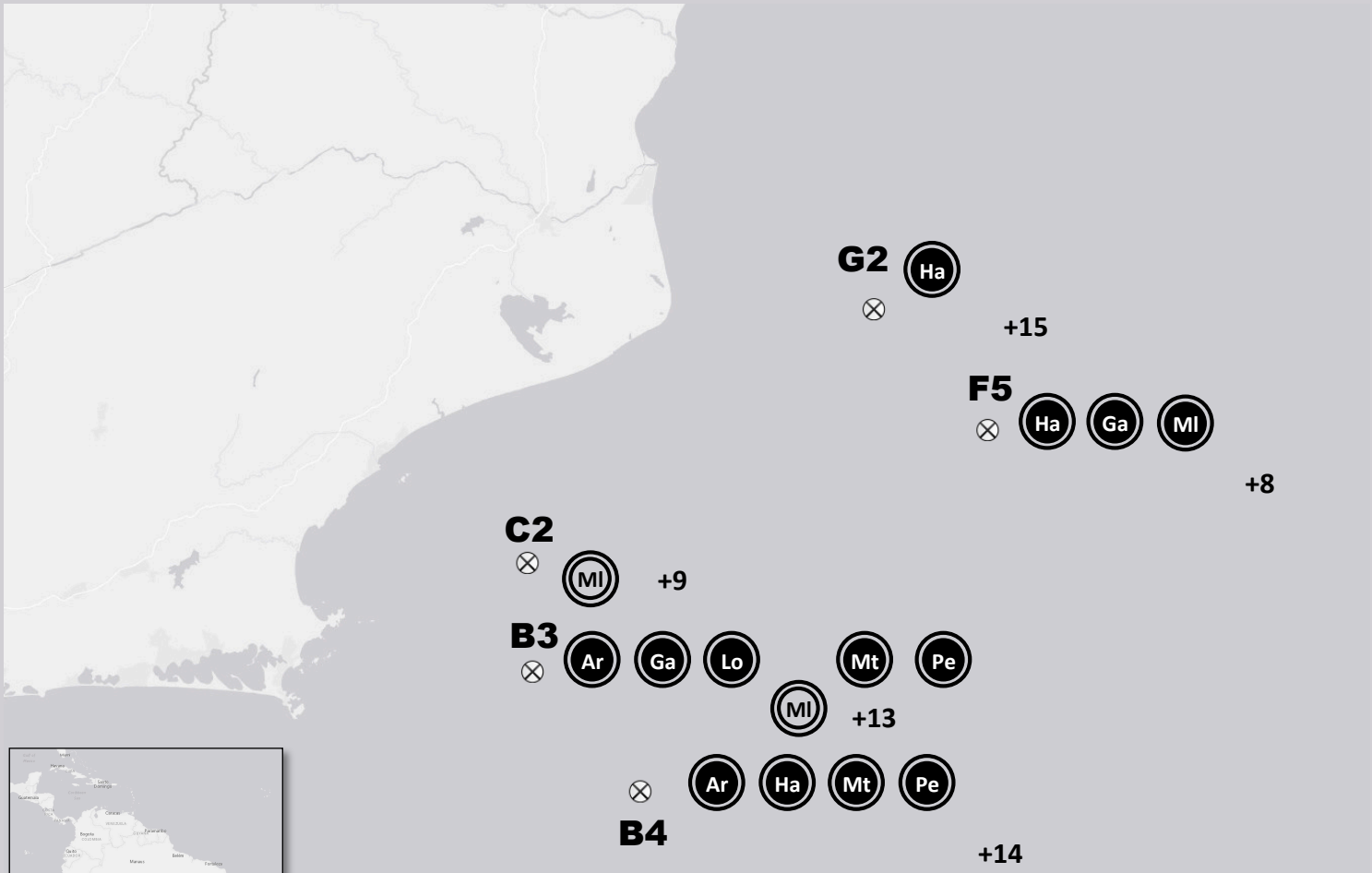
** Previous studies in Campos basin

***Non-marine family

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2103v3 | CC BY 4.0 Open Access | rec: 22 Jul 2016, publ: 22 Jul 2016

Mollusca distribution

Molecular | Morphological



Ar Ar **Arcidae* ****

Ga Ga **Galeommatidae****

Ha Ha **Haliotidae****

Lo Lo **Lottiidae****

Mt Mt **Mactridae* ****

Ml Ml **Mytilidae**

Pe Pe **Pectinidae* ****

* Present in other stations of Habitats Project

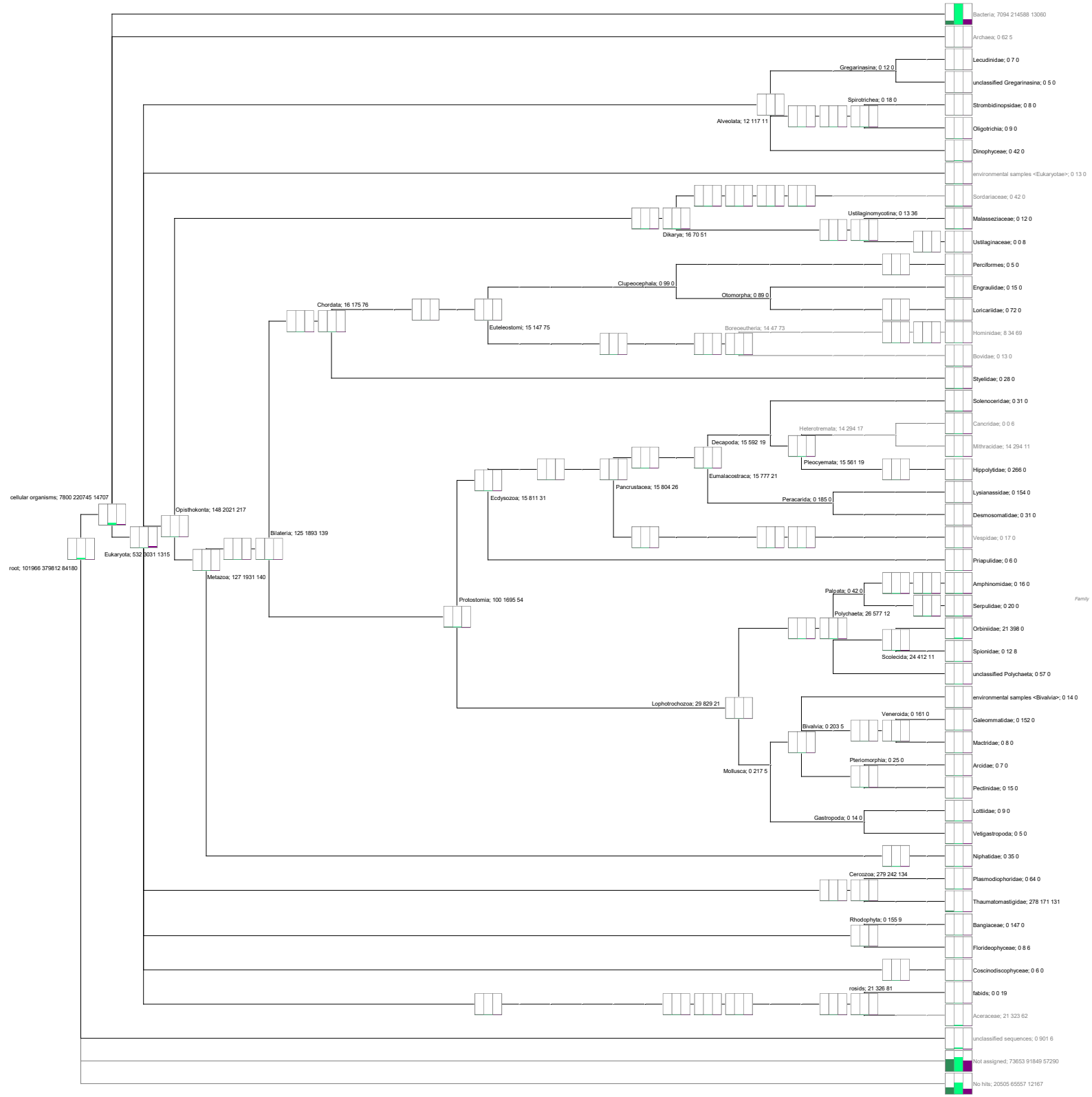
** Previous studies in Campos basin

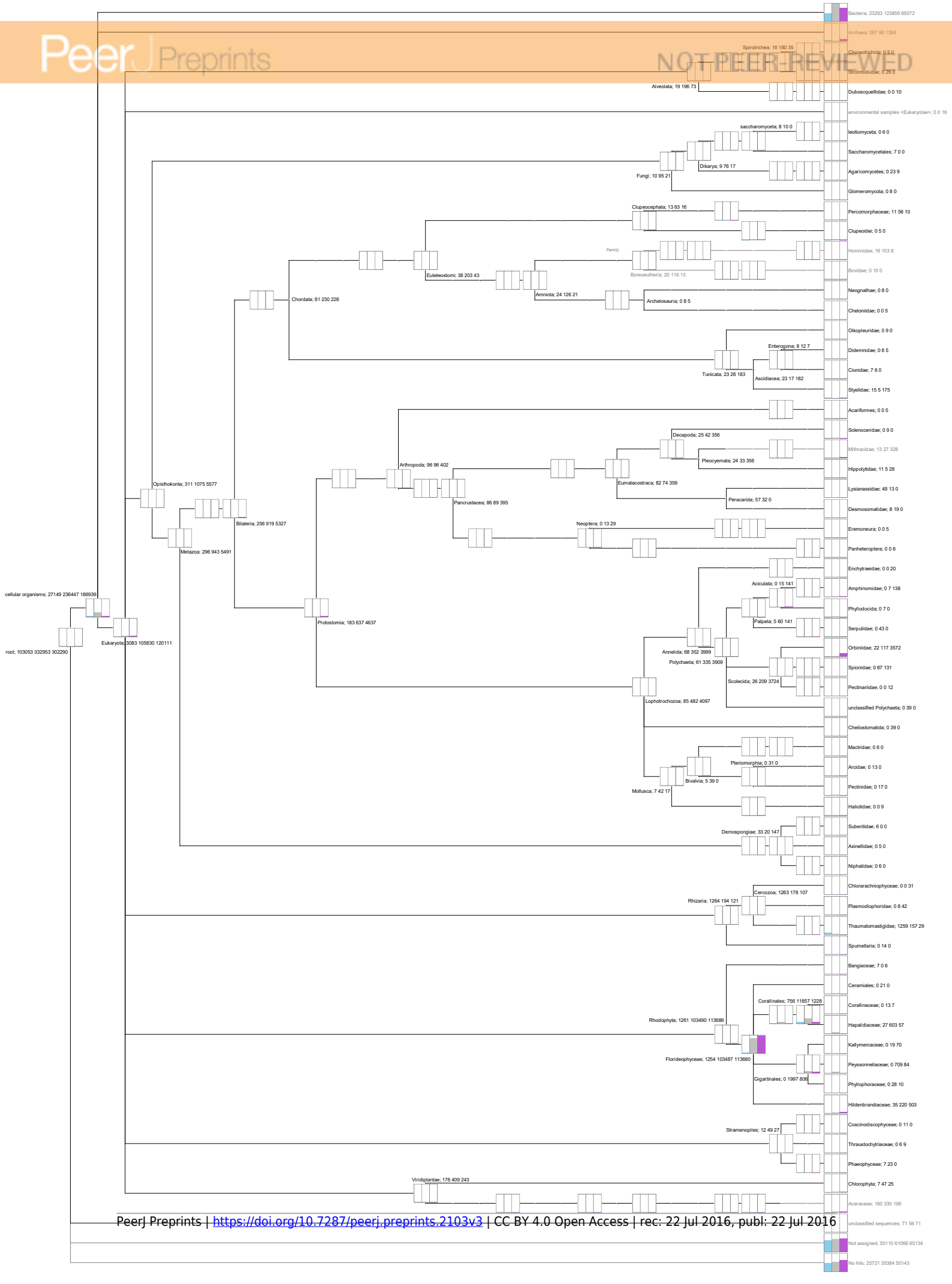
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2103v3 | CC BY 4.0 Open Access | rec: 22 Jul 2016, publ: 22 Jul 2016

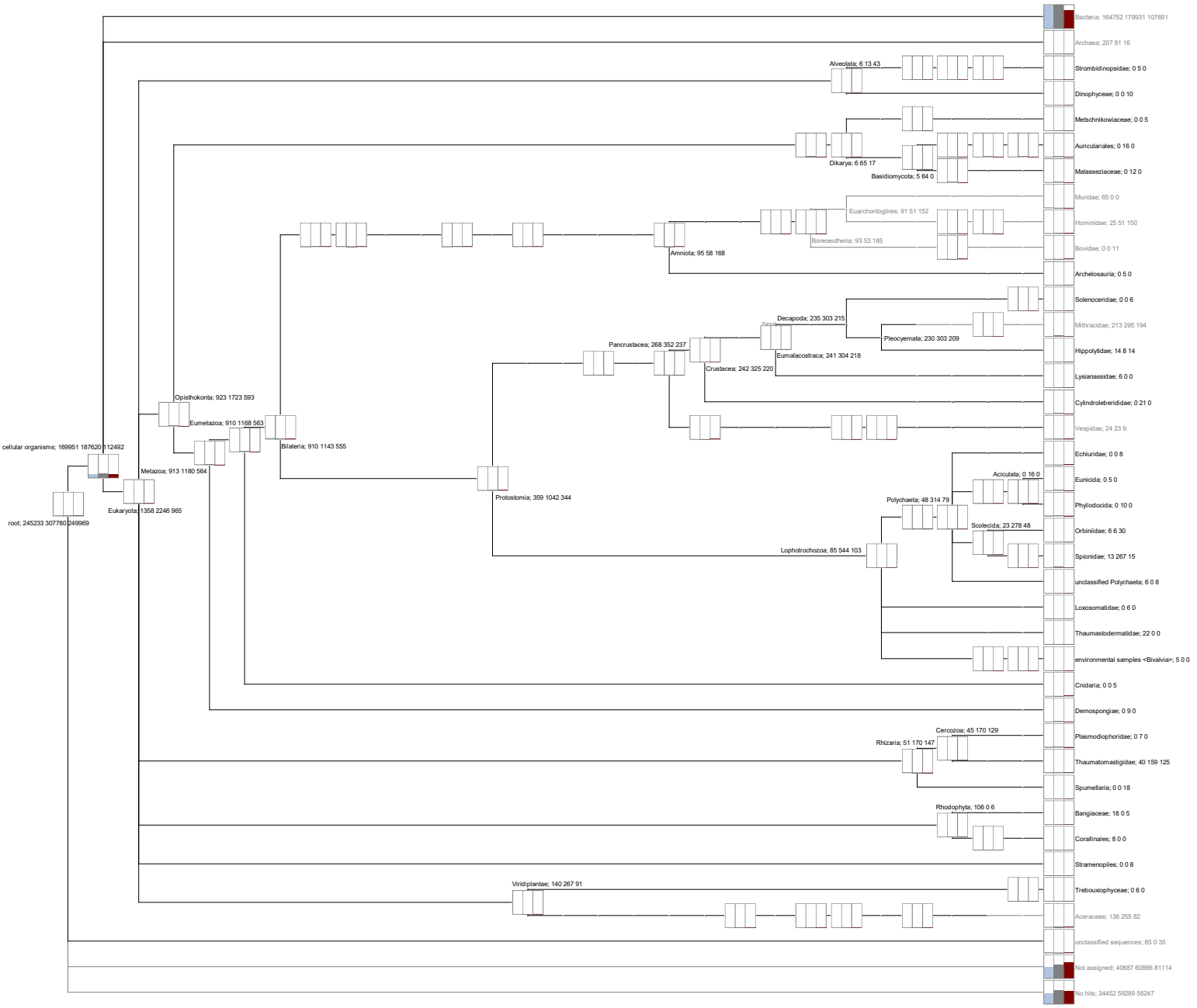
Suppelementar Material 1 – PCR primers and conditions. 1-5 µL of DNA template, 1 µL (5µM) of primers Forward and reverse), 5 µl of 10X buffer, 2 µl of $MgCl_2$ (25 mM), 1 µl of dNTP 10 µM (Fermentas), 0.2 µl de Platinum® Taq DNA Polymerase High Fidelity 5 U.µL-1 (Thermo Scientific) and ultra pure destilaed water (Invitrogen) to complete 50 µl final reaction volume.

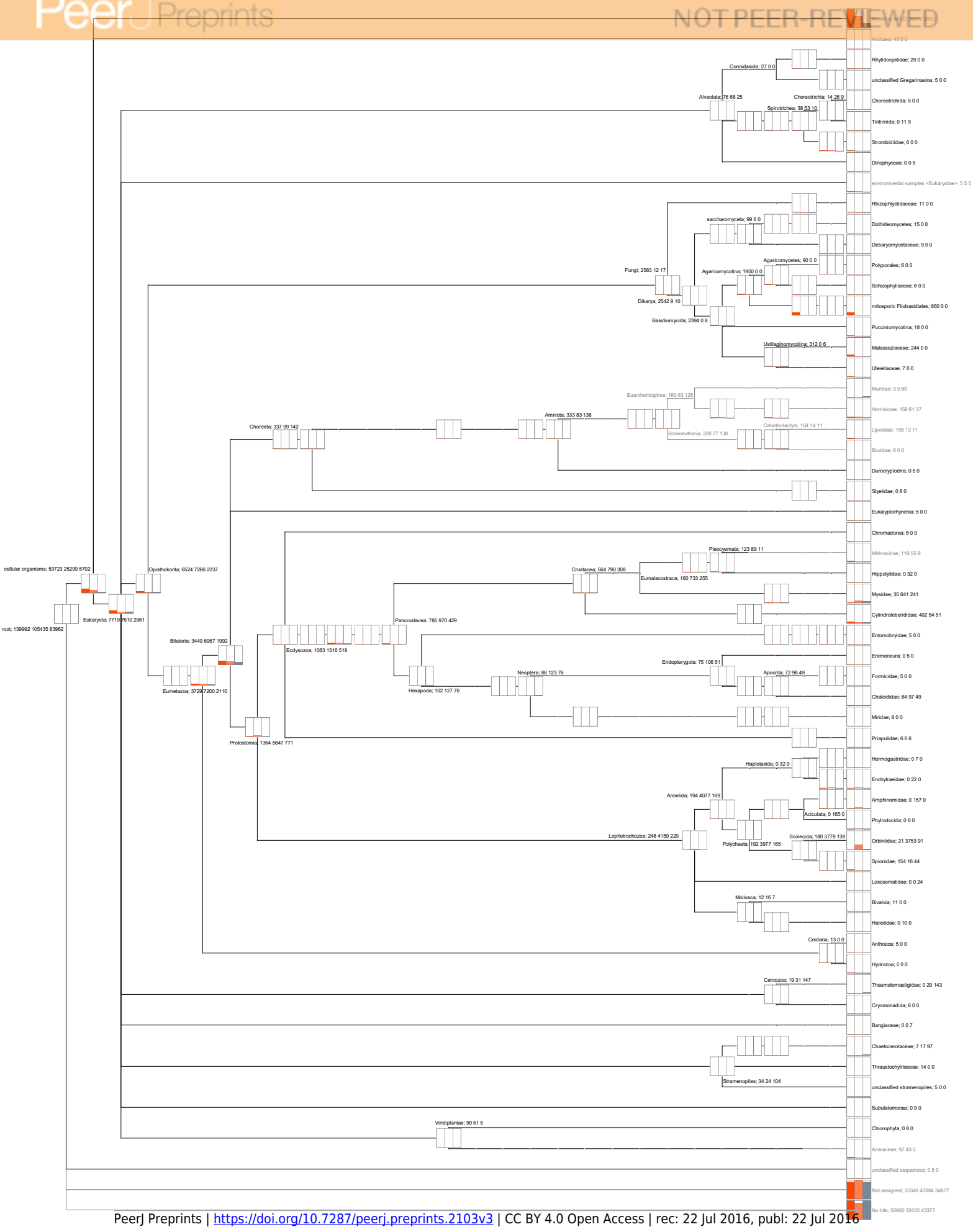
Target	Primer (F – Forward; R – reverse)	Denaturation	cycles	denaturation	anealing	Extension	Final extension	References
COI	TITCIAAYCAYAARGAYATTGG (F – jLCO1490); TAIACYTCIGGRTGICCRAARAAYCA (R – jHCO2198)	1' @94oC	10+30	30"@94oC	1'30"@61-52oC (-1oC per cycle) + 1'30"@61-52oC	1'@72oC	5'@72oC	Geller et al., 2013
rRNA 18S	ATGGTTGCAAAGCTGAAC (F – a2.0); GATCCTTCCGCAGGTTCACCTAC (R- 9R)	2' @94oC	40	30"@94oC	30'@55oC	1'@72oC	5'@72oC	Whiting et al., 1997; Whiting, 2002
rRNA 28S	ACCCGCTGAATTTAAGCAT (F – C1'); TGAACTCTCTCTTCAAAGTTCTTTTC (R- C2)	2' @94oC	40	30"@94oC	30'@55oC	1'@72oC	5'@72oC	Van Le et al., 1993; Chen et al., 2003

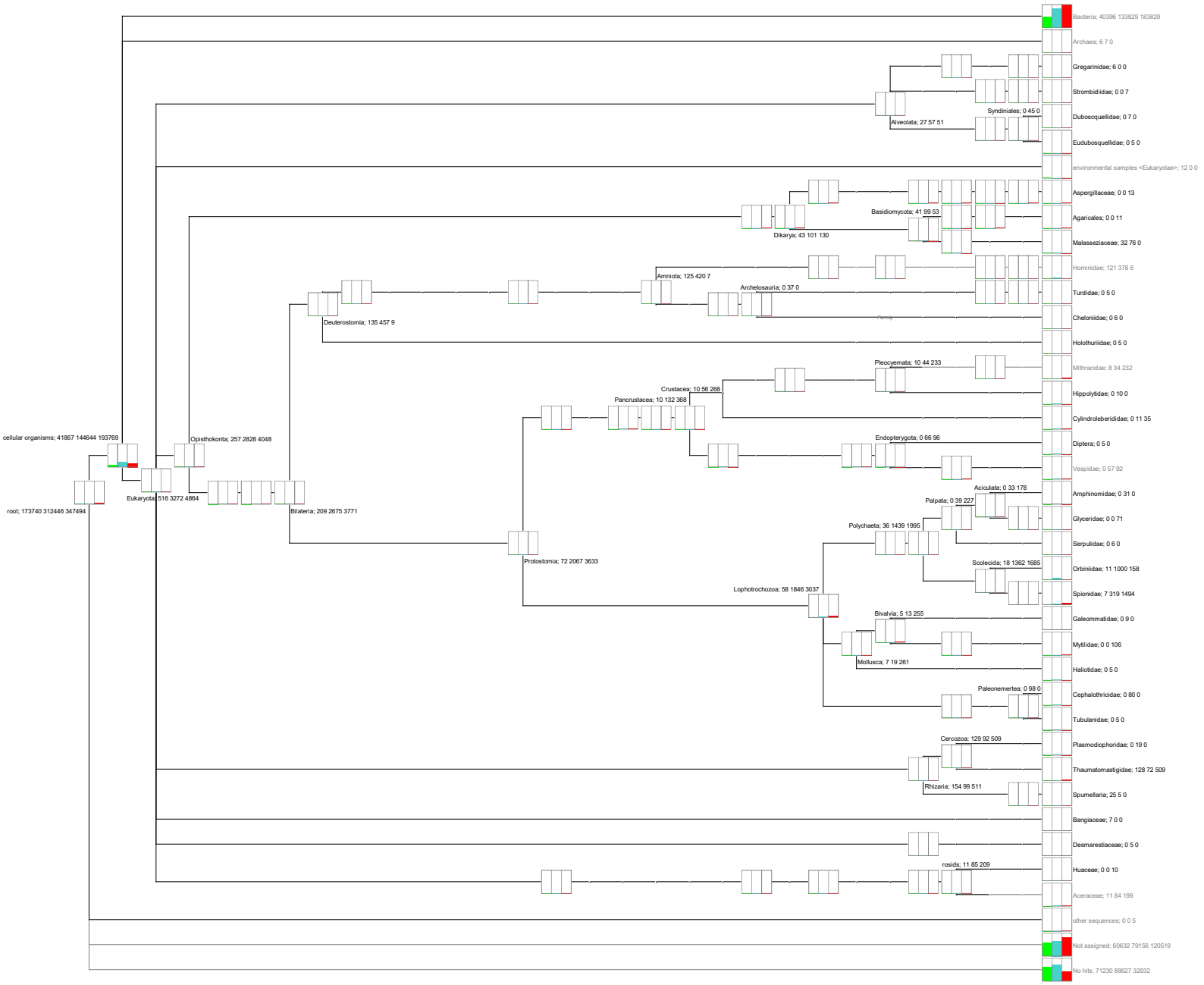
Suplementar material 2 – Family level Cladograms of the 5 sampling stations. Cladograms were built using speciments identified with any of the 3 target genes. Bar inside the squares represent the number of reads from each gene used to create the node. A) Family cladogram for station B3; b) Family cladogram for station B4; C) Family cladogram for station C2; D) Family cladogram for station G2; E) Family cladogram for station F5.


Bacteria; 7094 214588 13060
Archaea; 0 62 5
Lecudinidae; 0 7 0
Gregarinasina; 0 12 0
unclassified Gregarinasina; 0 5 0
Spirotrichea; 0 18 0
Strombidinopsidae; 0 8 0
Alveolata; 12 117 11
Oligotrichia; 0 9 0
Dinophyceae; 0 42 0
environmental samples <Eukaryotae>; 0 13 0
Sordariaceae; 0 42 0
Ustilaginomycotina; 0 13 36
Malasseziaceae; 0 12 0
Dikarya; 16 70 51
Ustilaginaceae; 0 0 8
Perciformes; 0 5 0
Clupeocephala; 0 99 0
Engraulidae; 0 15 0
Otomorpha; 0 89 0
Chordata; 16 175 76
Loricariidae; 0 72 0
Euteleostomi; 15 147 75
Boreoeutheria; 14 47 73
Hominidae; 8 34 69
Bovidae; 0 13 0
Styelidae; 0 28 0
Solenoceridae; 0 31 0
Cancridae; 0 0 6
Heterotremata; 14 294 17
Decapoda; 15 592 19
Mithracidae; 14 294 11
Pleocyemata; 15 561 19
Hippolytidae; 0 266 0
Eumalacostraca; 15 777 21
Pancrustacea; 15 804 26
Lysianassidae; 0 154 0
Ecdysozoa; 15 811 31
Peracarida; 0 185 0
Desmosomatidae; 0 31 0
Vespidae; 0 17 0
cellular organisms; 7800 220745 14707
Opisthokonta; 148 2021 217
Bilateria; 125 1893 139
Priapulidae; 0 6 0
Eukaryota; 532 3031 1315
root; 101966 379812 84180
Metazoa; 127 1931 140
Amphinomidae; 0 16 0
Palpata; 0 42 0
Family
Protostomia; 100 1695 54
Serpulidae; 0 20 0
Polychaeta; 26 577 12
Orbiniidae; 21 398 0
Scolecida; 24 412 11
Spionidae; 0 12 8
unclassified Polychaeta; 0 57 0
environmental samples <Bivalvia>; 0 14 0
Lophotrochozoa; 29 829 21
Venerocida; 0 161 0
Galeommatidae; 0 152 0
Bivalvia; 0 203 5
Mactridae; 0 8 0
Pteriomorphia; 0 25 0
Arcidae; 0 7 0
Mollusca; 0 217 5
Pectinidae; 0 15 0
Lottiidae; 0 9 0
Gastropoda; 0 14 0
Veligastropoda; 0 5 0
Niphatidae; 0 35 0
Cercozoa; 279 242 134
Plasmodiophoridae; 0 64 0
Thaumatomastigidae; 278 171 131
Rhodophyta; 0 155 9
Bangiaceae; 0 147 0
Florideophyceae; 0 8 6
Coscinodiscophyceae; 0 6 0
rosids; 21 326 81
fabids; 0 0 19
Aceraceae; 21 323 62
unclassified sequences; 0 901 6
Not assigned; 73653 91849 57290
No hits; 20505 65557 12167









Suplementar material 3 – List of species identified by molecular and morhological taxonomy.

Specie	18S	20S	COI	Study	
Cnemidocarpa verrucosa	+	+	+	Schettini	
Desmarestia dudresnayi	+	+	+	Schettini	
Erythrophyllum delesserioides	+	+	+	Schettini	
Eurythenes gryllus	+	+	+	Schettini	
Galeomma turtoni	+	+	+	Schettini	
Grifola frondosa	+	+	+	Schettini	
Haliotis diversicolor	+	+	+	Schettini	
Hormogaster redii	+	+	+	Schettini	
Lysmata seticaudata	+	+	+	Schettini	
Malassezia globosa	+	+	+	Schettini	
Marenzelleria arctia	+	+	+	Schettini	
Mimachlamys varia	+	+	+	Schettini	
Mysidium columbiae	+	+	+	Schettini	
Parotocinclus maculicauda	+	+	+	Schettini	
Pinctada imbricata	+	+	+	Habitats	and Hits
Platynereis dumerilii	+	+	+	Habitats	and Hits
Pontocaris lacazei	+	+	+	Habitats	
Praxillella affinis	+	+	+	Habitats	
Progoniada regularis	+	+	+	Habitats	and Hits
Protodorvillea kefersteini	+	+	+	Habitats	
Pteria colymbus	+	+	+	Habitats	
Scalibregma inflatum	+	+	+	Habitats	and Hits
Scapharca broughtonii	+	+	+	Schettini	
Serpula vermicularis	+	+	+	Schettini	
Syllis gracilis	+	+	+	Habitats	and Hits
Syllis variegata	+	+	+	Habitats	and Hits
Travisia brevis	+	+	+	Habitats	and Hits
Travisia forbesii	+	+	+	Habitats	and Hits
Travisia pupa	+	+	+	Habitats	and Hits
Aglaophamus circinata		+	+	Habitats	and Hits
Alpheus formosus		+	+	Habitats	
Amphipholis squamata		+	+	Habitats	
Aricidea wassi		+	+	Habitats	and Hits
Chelonia mydas		+	+	Schettini	
Praxillella pacifica		+	+	Habitats	and Hits
Priapulus caudatus		+	+	Schettini	
Scolelepis bonnieri		+	+	Schettini	
Scolelepis foliosa		+	+	Schettini	
Amphimedon queenslandica	+		+	Schettini	
Axiothella rubrocincta	+		+	Habitats	and Hits
Bathyarca pectunculoides	+		+	Habitats	
Bathyglycinde profunda	+		+	Habitats	
Bathyglycinde sibogana	+		+	Habitats	
Caprella equilibra	+		+	Habitats	and Hits
Ceratocephale abyssorum	+		+	Habitats	and Hits
Ciona intestinalis	+		+	Schettini	
Clymenella torquata	+		+	Habitats	and Hits
Pectinaria granulata	+		+	Schettini	
Perna viridis	+		+	Schettini	
Protaspis grandis	+		+	Schettini	
Syllis hyalina	+		+	Habitats	and Hits
Didemnum candidum			+	Schettini	
Leodamas rubra			+	Habitats	and Hits
Leodia sexiesperforata			+	Habitats	
Leptochelia dubia			+	Habitats	
Leucothoe urospinosa			+	Habitats	and Hits
Lumbrineris latreilli			+	Habitats	and Hits
Lysidice ninetta			+	Habitats	and Hits
Lysmata anchisteus			+	Schettini	
Macrochaeta clavicornis			+	Habitats	
Marphysa bellii			+	Habitats	and Hits
Mendicula ferruginosa			+	Habitats	and Hits
Mooreonuphis pallidula			+	Habitats	and Hits
Neanthes acuminata			+	Habitats	and Hits
Nereimyra punctata			+	Habitats	and Hits

Specie	18S	20S	COI	Study	
Notomastus latericeus			+	Habitats	and Hits
Ophelina acuminata			+	Habitats	and Hits
Pyropia haitanensis			+	Schettini	
Scapharca kagoshimensis			+	Schettini	
Scoloplos armiger			+	Schettini	
Isolda pulchella			+	Habitats	and Hits
Apophlaea lyallii	+	+		Schettini	
Chaetoceros curvisetus	+	+		Schettini	
Coelomactra antiquata	+	+		Schettini	
Crassinella lunulata	+	+		Habitats	and Hits
Cryptococcus friedmannii	+	+		Schettini	
Cyclaspis alba	+	+		Habitats	
Cylichna alba	+	+		Habitats	and Hits
Engraulis japonicus	+	+		Schettini	
Euclymene oerstedi	+	+		Habitats	and Hits
Eulalia viridis	+	+		Habitats	and Hits
Eumida sanguinea	+	+		Habitats	and Hits
Exogone dispar	+	+		Habitats	and Hits
Galathowenia oculata	+	+		Habitats	
Glycera americana	+	+		Habitats	and Hits
Glycera southeastatlantica	+	+		Habitats	and Hits
Goniada emerita	+	+		Habitats	
Hesiospina aurantiaca	+	+		Habitats	and Hits
Patelloida striata	+	+		Schettini	
Scopelocheirus schellenbergi	+	+		Schettini	
Subulatomonas tetraspora	+	+		Schettini	
Ophelina cylindricaudata		+		Habitats	and Hits
Ophiactis lymani		+		Habitats	
Trypanosyllis zebra		+		Habitats	and Hits
Ahnfeltiopsis leptophylla	+			Schettini	
Crucigera zygophora	+			Schettini	
Leitoscoloplos pugettensis	+			Schettini	
Malassezia nana	+			Schettini	
Ophiura ljungmani	+			Habitats	
Owenia fusiformis	+			Habitats	and Hits
Panthalis oerstedi	+			Habitats	and Hits
Paralacydonia paradoxa	+			Habitats	and Hits
Paramphinome jeffreysii	+			Habitats	and Hits
Pholoe minuta	+			Habitats	
Phtisica marina	+			Habitats	
Phyllodoce longipes	+			Habitats	and Hits
Solenocera crassicornis	+			Schettini	
Strombidium paracalkinsi	+			Schettini	
Phagomyxa odontellae	+			Schettini	