Robust phylogenomic inference through genome skimming as demonstrated by a systematic analysis of Nereididae

Dissertation
for the award of the degree
“Doctor rerum naturalium”

Division of Mathematics and Natural Sciences
at the Georg-August-Universität Göttingen
within the doctoral degree programme Genome Science
of the Georg-August University School of Science (GAUSS)

submitted by

Felix Thalén
from Norrstrand, Sweden

Göttingen, September 5, 2022

Thesis Committee

Prof. Dr. Christoph Bleidorn
Dept. for Animal Evolution and Biodiversity, Georg-August-Universität Göttingen

Prof. Dr. Burkhard Morgenstern
Dept. of Bioinformatics, Georg-August-Universität Göttingen

Dr. Nico Posnien
Dept. of Developmental Biology, Georg-August-Universität Göttingen

Members of the Examination Board

Reviewer: Prof. Dr. Christoph Bleidorn
Dept. for Animal Evolution and Biodiversity, Georg-August-Universität Göttingen

Second Reviewer: Prof. Dr. Burkhard Morgenstern
Dept. of Bioinformatics, Georg-August-Universität Göttingen

Further members of the Examination Board:

Dr. Nico Posnien
Dept. of Developmental Biology, Georg-August-Universität Göttingen

Prof. Dr. Elvira Hörandl
Dept. of Systematics, Biodiversity and Evolution of Plants, Georg-August-Universität Göttingen

Prof. Dr. Jan de Vries
Dept. of Applied Bioinformatics, Institute for Microbiology and Genetics, Georg-August-Universität Göttingen

Dr. Johannes Söding
Quantitative and Computational Biology, Max-Planck Institute for Biophysical Chemistry

Date of the oral examination: September 5, 2022

To Amanda, the love of my life

Abstract

The introduction of massively parallel sequencing in the mid-2000s has truly revolutionized the field of molecular phylogenetics and ultimately our understanding of the tree of life. Despite today's existence of long-read sequencing and hybrid strategies—with the ability to produce high-quality, near-chromosome-level assemblies—monetary costs, sampling restrictions, and other practical considerations still favors next-generation sequencing (NGS) in many instances. To this date, most large-scale phylogenetic or "phylogenomic" studies were conducted using a "genome reduction" strategy such as transcriptome sequencing or target enrichment. Until recently, whole-genome sequencing (WGS) was often dismissed or overlooked due higher sequencing costs and a lack of appropriate bioinformatic tools to process the data downstream. Despite this, WGS has many advantages over other techniques such as reduced laboratory workload, smaller DNA volume and quality requirements, and higher data re-usability, even outside phylogenetics. Now, advancements in short-read sequencing technology has reduced the costs of sequencing and the development of new, alignment-based bioinformatic software for working with WGS data should have phylogeneticists reconsider this sequencing approach. Still, we saw the need for a fast, scalable, and easy-to-use method for mining desired loci from raw reads or assembled contigs. Thus, we here present Patchwork, a new, alignment-based program, which mines phylogenetic markers from WGS data by "stitching" overlapping and or adjacent sequence regions. A novel sliding-window based algorithm trims non-coding regions from extracted markers. We ultimately demonstrate the utility of both Patchwork—and for using WGS in a phylogenomic context—by using this tool to reconstruct the phylogeny of the annelid family Nereididae. All previous attempts to infer the phylogeny of Nereididae have been limited to morphological data or by using one or a handful of mitochondrial genes. Most of these studies were also severely limited in taxonomic coverage. Here, we present trees inferred from a set of 777 near-universal single-copy orthologs and mitochondrial genomes, containing a total of 100 and 132 taxa respectively, to produce a well-supported and congruent phylogeny of the group.

Acknowledgements

First, I would like to thank my supervisor, Christoph Bleidorn, for all of the interesting discussions that we've had and for always being open to new ideas.

I also want to thank everyone involved in the sample collection and for sending samples our way: Christopher J. Glasby and Robin S. Wilson, Torkild Bakken, Torsten H. Struck, Teresa Darbyshire, Detlev Arendt, Kevin N. Mutemi, Tobias Gerber, Dinesh Kaippilly, Jithin Kothalil, Dazuo Yang, Maria C. Coralles, and Tulio F. Villalobos Guerrero.

Thanks to my committee members, Burkhard Morgenstern and Nico Posnien, for fruitfull discussions and sound advice on how to proceed with the project.

Thank you, Marta Tischer and Ludwik Gąsiorowski, for making this whole experience so much more fun and enjoyable. Thank you, Katharina Henze, for all of your hard work in the lab and for your kindness. I also want to thank the rest of the members of department and the members of the Genome Science program for their support.

Thanks to my parents for their continued support.

Lastly, thanks to all of my students: Clara G. Köhne, Frederick Hahner, and Thilo Schulze, for all of the effort that they have put into this project. I wish you all the best on your future endeavours.

This work has received funding from The German Research Foundation (DFG) and was partly supported by the Göttingen Graduate Center for Neurosciences, Biophysics, and Molecular Biosciencces of the Georg-August-Universität Göttingen.

This dissertation is dedicated to my fiancée, Amanda Sandberg, for the endless love, support, and encouragement that she gives me.

Table of Contents

List of Papers

This thesis is based on the following papers:

1. The untapped potential of genome skimming in the age of chromosome-scale genomics
2. Patchwork: reference- and alignment-based retrieval and concatenation of phylogenetic markers from whole-genome sequencing data
3. Phylogenomic analysis of Nereididae using a new genome-skimming approach

List of Contributions

Contributions by Felix Thalén:

1. Took the lead in manuscript preparation.
2. Conceived of the original idea together with the supervisor. Designed the program and implemented the program in collaboration with bachelor student Clara G. Köhne. Took the lead in manuscript preparation.
3. Contributed to the data collection. Performed the majority of the data analysis. Designed and implemented the phylogenomic pipeline. Took the lead in manuscript preparation.

List of Tables

List of Figures

List of Abbreviations

WGS	Whole-Genome Sequencing
LC-WGS	Low-Coverage Whole-Genome Sequencing
RNA-Seq	RNA Sequencing
NGS	Next-Generation Sequencing
TE	Target Enrichment
HGT	Horizontal Gene Transfer
SBS	Sequencing By Synthesis
UCE	Ultra-Conserved Element
USCO	Near-Universal Single-Copy Ortholog

1 Introduction

> *All living beings have much in common, in their chemical composition, their cellular structure, their laws of growth, and their liability to injurious influences... Therefore, on the principle of natural selection with divergence of character, it does not seem incredible that, from such low and intermediate form both animals and plants may have been developed; and, if we admit this, we must likewise admit that all the organic beings which have ever lived on this earth may have descended from someone primordial form*
>
> On the Origin of Species
> Charles Darwin

All life on earth—from bacteria to plants and humans—share a common ancestor. This idea, laid out by Darwin (Darwin, 1859; also see figure 1.1), has since been evidenced by the fossil record, shared morphological traits, and the existence of a shared universal genetic code. But how exactly did this event unfold, how *is* everything related? Phylogenetics is the field of study principally concerned with reconstructing the evolution of species (extinct or living), individuals, or genes (Baum & Smith, 2013). By modelling the evolutionary history of species—typically as phylogenetic trees—we can lay out a framework from which questions regarding relatedness can be answered. The utility of phylogenetics was clearly made evident in the ongoing COVID-19 pandemic, where the distribution and evolution of new variants could be tracked in real-time. Moreover, the degree of relatedness between these variants could be used to predict infection-rates as well as the origin of the virus itself (e.g., T. Li et al., 2020). Before molecular data was accessible, the phylogenetic history of a set of species was inferred through the comparison of shared morphological traits. With the advent of sequencing, pioneered by the works of Frederick Sanger, however, it became possible to "read" amino acid, and subsequently nucleotide, sequences of organisms (Heather & Chain, 2016). Various techniques and approach for comparing these genes were rapidly developed and the field of molecular phylogenetics was established (Delsuc et al., 2005). Researches did, however, often find

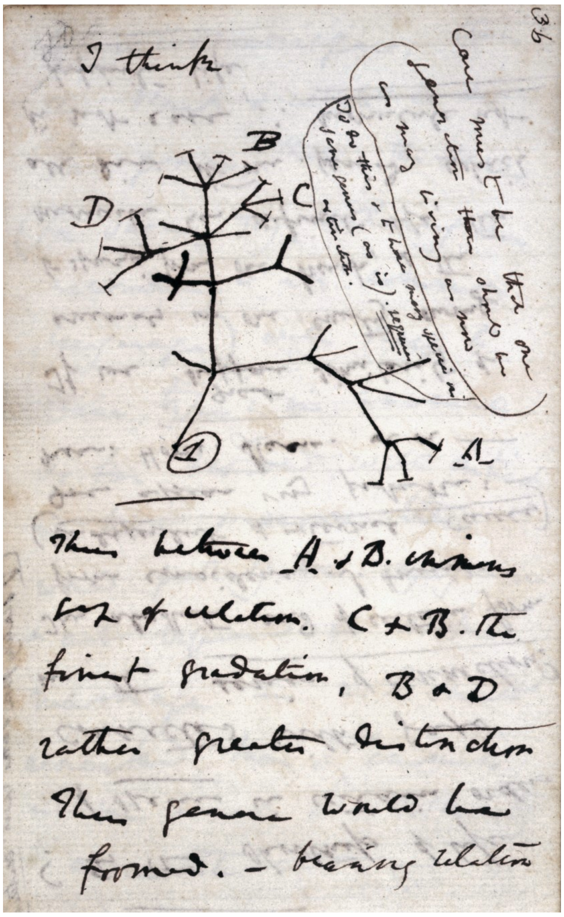

36
I think
B
C
D
1
A
Case must be that one generation then should be as many living as now
To do this & to have many species in same genus (as is). requires extinction.
Thus between A & B. immens gap of relation. C & B. the finest gradation, B & D rather greater distinction
Thus genera would be formed. — bearing relation

Figure 1.1: The first-known sketch of an evolutionary tree by Charles Darwin that describes the relationships among groups of organisms. Attribution: Charles Darwin/Public domain.

discrepancies when different genes and or models of molecular evolution were used to infer a phylogeny (Philippe et al., 2017). Advancements in sequencing technologies in the mid-2000s then transformed the field into phylogenomics, where large, multi-gene datasets paved way for phylogenetic trees that were less dependent on the individual gene, or the handful of genes, under study (Delsuc et al., 2005).

1.1 Phylogenomics: the end or the beginning of incongruence?

With the introduction of high-throughput sequencing technologies, it suddenly became possible to analyse genomic and or transcriptomic datasets with hundreds or even thousand of genes. Although initially introduced as mean to put an end to contradictory results (Gee, 2003; Rokas et al., 2003), researchers quickly realised that analyzing huge, concatenated gene sets came with its very own set of problems (Phillips et al., 2004; Soltis et al., 2004); different models of molecular evolution, tree reconstruction methods, choice of data type (i.e., nucleotides vs. amino acids), and or gene selection could potentially lead to different outcomes. Other sources of errors include violating the assumptions of orthology—i.e., that all genes under study are related via a speciation event as opposed to a gene duplication event or via Horizontal Gene Transfer (HGT)—and problems stemming from missing data, now further amplified by the these huge datasets (Jeffroy et al., 2006).

Nowadays, the availability of cost-efficient long-read technologies introduces yet another dimension to the field. With so many different sequencing approaches to choose from, what are actually the advantages and disadvantages of choosing between them?

1.2 Sequencing approaches

1.2.1 First generation (Sanger) sequencing

For nearly three decades, DNA sequencing—and thus also molecular phylogenetics—was dominated by Sanger sequencing, and it was this technique that was used to first sequence the human genome (Consortium et al., 2004). Although increased automation led Sanger sequencing to become easier and cheaper over time, this approach is still characterized by producing reads of around 500-1000 bp in length, with a high accuracy (around 99.999% with trimmed reads), in small quantities. As a consequence, most phylogenetic studies at the time were limited to a single, or a handful, of orthologous genes (Jeffroy et al., 2006).

1.2.2 Next-generation sequencing (NGS)

A new era of molecular phylogenetics began with the introduction of so called next-generation sequencing (NGS). Unlike its precursor, these new machines were capable of outputting genome-scale data in huge quantities and revolutionized the way in which molecular phylogenetic studies were conducted (Chan & Ragan, 2013). Now, entire genomes or transcriptomes of could be compared at a fraction of the cost. Not only did the depth at which the targets were sequenced but also the breadth—a broader selection of taxa could be targeted, thanks to the reduced costs of sequencing. Originally, the read lengths of these techniques were much shorter (36–72 bp), but modern platforms such as the Illumina NovaSeq 6000 can routinely produce reads of 250 bp in length. The accuracy of Illumina sequencing, which is today by far the most common platform for NGS, is around 99.25% (Quail et al., 2012) and today, the majority of transcriptome and resequencing projects are conducted using Illumina sequencing and NGS technology (Bleidorn, 2017). Unfortunately, the relatively short read lengths results in lower-quality assemblies, where many genes are missing, in fragments, and or are incorrectly assembled (Yin et al., 2019).

1.2.3 Long-read sequencing

Lately, the introduction of long-read sequencing technologies have made it easier to produce higher quality, near-chromosome-level genomes, thus overcoming some of the overarching challenges associated with short-read sequencing (Rhie et al., 2021). "Long-read sequencing", most often refer to one of the two currently leading platforms, namely Pacific Biosciences and Oxford Nanopore. These technologies can produce reads which are up to hundreds of thousands of bases long, about 30–300 times longer than those produced by short-read equivalents (ibid.). The accuracy of long-read technologies are lower, however, and the error rate of the PacBio platform is estimated to be around 0.5–15% (Hon et al., 2020; Weirather et al., 2017). Many phylogenomic projects today employ a "hybrid" strategy, where the high coverage of short-read sequencing technologies are combined with the long reads produced by long-read sequencer. This helps overcome the typically higher error rates and raw material requirements associated with long-read sequencing. However, low error rates of the latest PacBio sequencers has led to a situation where the PacBio HiFi sequencing method alone may yield sufficient quality (Hon et al., 2020). Despite all of these advantages, long-read or third-generation sequencing require high molecular weight DNA from freshly-collected and well-preserved material (Chakraborty et al., 2016), while sequence libraries for short-read sequencing can be constructed from low-quantity and more fragmented DNA. The higher associated costs has also limited the uses to larger consortia—such as Genome 10K (of Scientists, 2009) or The Darwin Tree of Life Project (of Life Project Consortium, 2022)—taking advantage of these approaches, rather than individuals labs, who may not be able to upfront the added cost. Given the choice of more depth (i.e., higher coverage) or more breadth (i.e, more taxonomic coverage), many researchers will likely opt for a broader taxon sampling. Collecting everything anew may not even be an option and in some cases and sub-optimal material from a museum collection may be all that is available and in those cases, long-read sequencing is not even an option anymore. The bottleneck of many phylogenomic projects today is not only producing high-quality genomes but to collect and correctly identify each species in the first place. Taxonomic expertise is rare and going out in the field collecting is time-consuming and posed with increasing amounts of paperwork. Having the flexibility to avoid some of these steps will likely be a welcome—if not necessary—option to phylogenomic researchers.

Taxon sampling and sequencing material from museum collections using NGS sequencing—something now known as museomics—will be discussed in more detail in the next section, followed by a section on various data collection approaches for undertaking these types of sequencing jobs.

1.3 Taxon sampling approaches

All phylogenetic projects begin by selecting representing taxa for analysis and an insufficient taxon sampling is often citepd as a major source of error in phylogenetic inference (Zwickl & Hillis, 2002). Exhaustive taxon sampling, however, is often impractical—if not impossible—to achieve due to monetary costs, lack of computational resources, time-constraints, need for sampling permits, or rarity of the species of interest. Hence, the taxon sampling for most reasonably sized taxonomic groups will be non-exhaustive and careful consideration should be taken into selecting appropriate taxa for analysis. This selection should ideally represent the phylogenetic diversity of the group, be based on community value (i.e., the community's interest in having sequences publicly available for that group), genome size (smaller genomes lead to lower costs of data generation), and taxonomic stability. Thus, the value added from a well-rounded sampling increases for less well-studied group where the taxonomic placement of some taxa may be less reliable. When possible, each taxon should also be sampled from or near the type locality to avoid taxonomic ambiguities or changes.

1.3.1 Museomics

One way to increase taxon sampling without having to collect everything anew is to include specimen from natural history museum collections. Although previously dismissed due to the damage and fragmentation caused by non-ideal preservation methods (Hofreiter, 2012), more and more researches are sequencing museum specimens (e.g., Breinholt et al., 2018; Cong et al., 2017; F. Zhang et al., 2019b. Researchers may be prohibited from collecting new material due to monetary reasons (e.g., funding reasons or high travel-related costs when covering a wide geographical range), increased bureaucracy (e.g., due to stricter regulations introduced with the Nagoya Protocol in 2014; explained in more detail in chapter 3), lack of taxonomic expertise, or rarity of the species of interest (Call et al., 2021b). Not only do natural history museum collections provide an alternative way to obtain DNA material from species that would be difficult to collect anew, it also a mean to study species throughout time and space. I.e., samples preserved in a museum collection are snapshots of a species, in a geographical location, at a specific point of time. This allow researchers to study genetic changes over time and also to study the genome of species in decline, facing the risk of extinction, or that are already extinct (e.g., Zedane et al., 2016).

1.4 Data collection approaches

In this section, I will outline the most popular data collection methods for phylogenomics, that are used when performing massively parallel sequencing (or NGS).

1.4.1 Reduced representation approaches

To this day, most large-scale phylogenetic (i.e., phylogenomic) studies were conducted using a "reduced-representation" strategy such as transcriptome sequencing (also known as RNA-Seq), or hybridization- or capture-based target enrichment (F. Zhang et al., 2019b). These approaches are called so because their aim is to sequence only a select portion of the genome (i.e., expressed portions of the genome in the case of transcriptomics and regions captured by designed probes in the case of targeted enrichment).

RNA-Seq

The transcriptome of an organsim is the expressed portion of the genome at a certain developmental stage or physiological condition (Wang et al., 2009). Various technologies have been developed to deduce and quantify transcriptomes, but when high-throughput sequencing, or NGS, is used, we call it RNA Sequencing (RNA-Seq). Notably, in RNA-Seq, the RNA itself is not sequenced and instead it (the RNA) is first converted to a library of cDNA fragments with adaptors attached to one or both ends, which are subsequently sequenced (Wang et al., 2009). RNA-Seq is a widely used approach to phylogenomics (e.g., Kocot et al., 2017), partly due to the reduced costs of sequencing, as only the expressed portion needs to be sequenced. Unfortunately, transcriptome sequence requires large quantities of high-quality RNA from freshly-collected or carefully-stored samples (Cronn et al., 2012b), meaning that smaller organisms have to be pooled together and older material, stored in museum collections, cannot be utilized when using this sequencing method.

Target enrichment

Target Enrichment (TE) is a way to capture a select portion of the genome through the use of so called probes, which can capture both genes and flanking regions. These regions can then be sequenced in greater depth, at reduced cost, as not all of the genome has to be sequenced (E. M. Lemmon & Lemmon, 2013). TE does not have the same materialistic restrictions as RNA-Seq and can be used for degraded DNA, e.g., from samples stored in sub-optimal conditions in a museum collection or similar entity. Designing "baits" does, however, require a reference, although this method can also be used on non-model organisms (Jones & Good, 2016). Even though this approach has a reduced computational costs, it comes with higher laboratory workloads and designing appropriate probes gets increasingly more difficult as the distance between the species of interest increases. Moreover, although target enrichment is a popular approach to conduct phylogenomic studies, the data that is generated has generally little use outside of phylogenetics and the whole study has to be conducted from scratch if other loci are to be included.

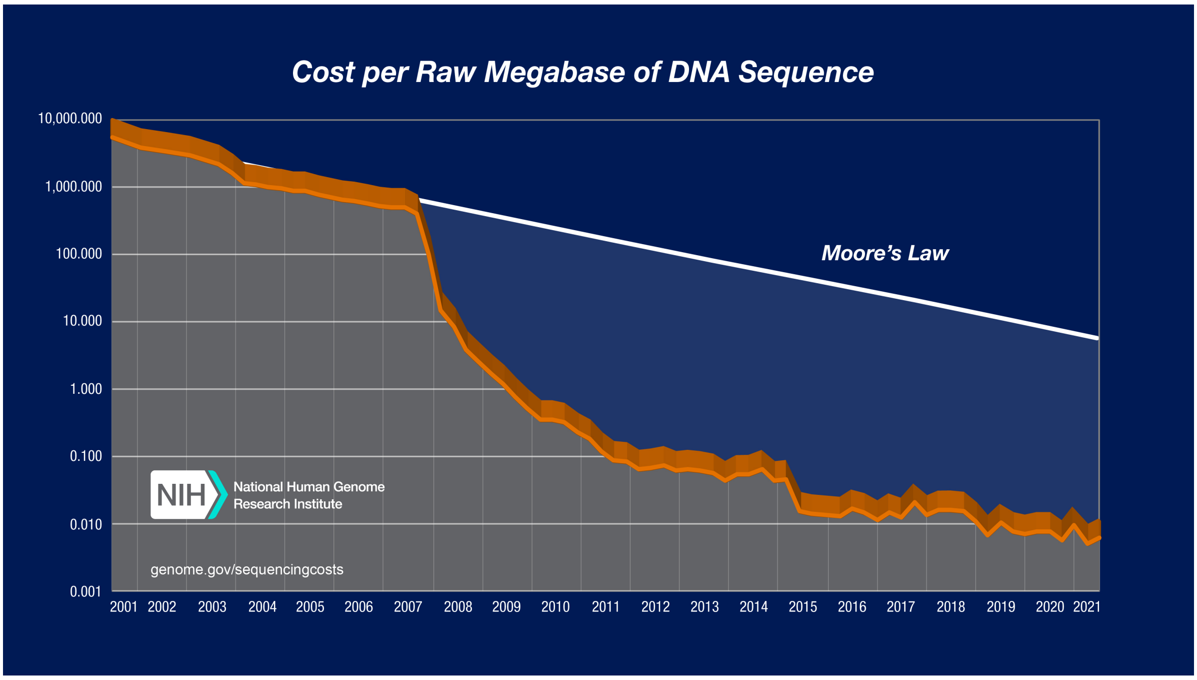


Figure 1.2: Sequencing cost per megabase in 2001–2021. Attribution: the National Human Genome Research Institute (NHGRI).

1.4.2 Whole-genome sequencing

An under-represented approach when it comes to phylogenomics is whole-genome sequencing (WGS). As the name suggests, the WGS approach means that an organisms genome—together with its organelle genomes (i.e., mitochondria or chloroplasts)—is sequenced in its entirety. The disuse of this method can most likely be attributed to higher monetary costs and a lack of appropriate bioinformatic tooling for handling the data it generates. Unlike "reduced-representation" strategies, WGS comes with increased sequencing costs because not only the regions of interest (i.e., the exome or an enriched portion of the genome) are sequenced. Now, decreasing sequencing costs (currently estimated to be around 0.01 US$ per one megabase [Mbp] of raw DNA sequences; also shown in figure 1.2; source: https://www.genome.gov/sequencingcostsdata) has made Whole-Genome Sequencing (WGS) more economically feasible and more widely applied in phylogenomics (e.g., W. Li et al., 2019).

There can be many advantages of using WGS over a reduced representation strategy. For instance, WGS uses a lower quantity of DNA than competing methods, typically somewhere between 50–200 ng for an Illumina library (ibid.). This can be advantageous in museomic studies, for example, or when sequencing organisms that are very small in size. In such cases, RNA-Seq may not even be an option because this typically requires freshly-collected material and pooling of smaller specimen, thus running the risk of mixing cryptic species together. Moreover, WGS has a much lower workload when compared with targeted enrichment, which requires that baits are designed and bait-design gets increasingly more difficult as the distance between the target species increases (Mamanova et al., 2010). Finally, one of the main benefits of using WGS over other approaches is that it doesn't restrict which data type is used and the data that is produced has future re-utility since it can be used in contexts other than phylogenetics. Greater depth of coverage is typically

On Low-Coverage Whole-Genome Sequencing

Because of a desire to maximize taxonomic coverage and because of the increased sequencing costs associated with WGS, most phylogenomic studies are relegated to using Low-Coverage Whole-Genome Sequencing (LC-WGS). By definition, Low-Coverage, "Shallow", or "Low-Depth" Whole-Genome Sequencing means that an organisms genome is sequenced at a lower coverage (<30x; G. Ribeiro et al., 2021) as a mean to reduce sequencing costs. G. Ribeiro et al., 2021 also found that the amount of sequencing error increases with a lower depth of coverage and that contamination has a potentially greater impact when coverage is low. The same study also found that for phylogenomic studies, depth of coverage of 5–10x is sufficient for inferring interspecies relationships. The ability to achieve this degree of coverage depends on an organism's genome size. The larger the genome, the higher the sequence depth required to achieve a certain amount of coverage. Although 5–10x coverage may provide sufficient for the fore-mentioned application, a higher average coverage (>70x sequencing depth; Faino and Thomma, 2014) may be required when assembling a high-quality draft genome of a eukaryote.

1.5 Mining phylogenetic markers

One of the biggest challenges with using LC-WGS in a phylogenetic context stems from the ability to accurately recover the loci of interest. Currently, bioinformatic programs exist for retrieving near-universal single-copy orthologs (USCOs; (Waterhouse et al., 2018)), ultra-conserved elements (Faircloth, 2016), and organellar genomes (Allio et al., 2020). Although LC-WGS genomes have routinely been used to mine high-copy number genomic regions such as mitochondrial loci and rDNA repeat regions (G. Ribeiro et al., 2021), targetting nuclear genomic regions is still less of a routine (F. Zhang et al., 2019b). Irregardless of the target, bioinformatic pipelines for processing WGS data are far from standardized and typically very computationally-intense due to the resource-heavy assembly process, which often has to be performed on dedicated clusters.

In the last few years, a new suite of tools have sprung into existence for retrieving phylogenetic markers through alignment and so-called "hit stitching", where overlapping and or adjacent alignments are merged to form larger and less fragmented markers. These tools can operate on raw reads and or assembled contigs and includes aTram (Allen et al., 2018b), AliBaSeq (Knyshov et al., 2021b), and GeMoMa (Keilwagen et al., 2019).

1.6 Short introduction to Nereididae Blainville, 1818

Nereididae Blainville, 1818, commonly known as ragworms, is a diverse family of annelids with over 700 species described worldwide (Read & Fauchald, 2020). Nereidids predominantly inhabit marine waters—from the deep-sea to the intertidal—but many species are also found in brackish- or freshwater, and even semi-terrestrial environments (Bakken et al., 2018). Because many nereidids are abundant and obtainable from easy-to-access, intertidal habitats, some members of this group has come to be used as fishing baits, as a food source in aquacultures, and as a subject for laboratory studies (ibid.). For example, *Platynereis dumerilii* has emerged as a model organism for developmental, ecology and toxicology, and finally evolutionary and neurobiological research (A. H. Fischer et al., 2010). Annelids are members of Lophotrochozoa (which also includes molluscs, for example), one of the three major branches of bilaterians (with the other two being Deuterostomia and Ecdysozoa), thus making *P. dumerilii* one of the most well-studied lophotrochozoan organisms. Despite all this, the interspecies relationships within Nereididae remain poorly understood. Until recently, all phylogenetic studies conducted on this group only investigated morphological characters (Bakken et al., 2018). Today, phylogenetic studies of Nereididae exist (e.g., Alves et al., 2020b) but are limited to mitochondrial genes and in their taxonomical coverage.

1.7 Case study: the phylogeny of Nereididae Blainville, 1818

Given all of this, how did we ourselves select an adequate sequencing approach when tackling the phylogeny of Nereididae Blainville, 1818? First, because several larger genera within the group were hypothesized to be non-monophyletic, based on morphological analyses conducted by Bakken et al. 2005 (Bakken & Wilson, 2005), we wanted to maximize the amount of taxa and be sure to include multiple species from the affected genera.

Thus, our sequencing approach and our tool of choice all have the following advantages: (i) we can sequence already collected species, stored in various research collections in non-optimal ways, (ii) we can collect mitochondrial genomes as a bi-product, (iii) the data we produced can be reused; once a better way to analyse the data-set emerges, we can instead apply this technique to our data, (iv) we can further increase our taxon coverage by mixing transcriptomic- and genomic data, and (v) our approach is optimized for the old and distantly-related group of Nereididae, where designing optimal probes would have been difficult and time-consuming.

2 Aim of thesis

The introduction of high-throughput sequencing has turned molecular phylogenetics into phylogenomics, where phylogenies are now inferred at an almost industrialized scale. A plethora of sequencing techniques, data collection strategies, and bioinformatic software now exists for conducting such studies. However, despite the promises of phylogenomics—or large-scale phylogenetics—putting an end to incongruencies seen in single-gene studies, many newer studies still suffer from methodological artifacts, improper choice of evolutionary models, violation of assumptions of orthology, limitations of heuristics in tree inference and other bioinformatic methods, and low-quality genomic data (Philippe et al., 2017). We have reasons to believe, however, that some of these issues can be avoided given proper bioinformatic methods for constructing phylogenetic data matrices and also for interpreting them and analyzing them on a larger scale. Furthermore, we want to re-emphasize the importance of large and appropriate taxon coverage to avoid some of the forementioned problems. In today's age of chromosome-level phylogenomics, we strongly advocate for using WGS data and genome skimming approach. Unlike genome reduction techniques such as RNA-Seq or target enrichment, this approach can be done with less sample quality and volume, involves less laboratory workload, and does not require probe synthesis. Recent advancements in high-throughput sequencing technology have reduced the prohibitive costs associated with WGS and new, bioinformatic tools for handling such data has made this a much more viable approach.

With this thesis, I wish (i) to outline the current status and future outlooks of genome skimming for phylogenomics, (ii) to present a newly developed method for mining phylogenetic markers from WGS data and, finally, (iii) to showcase the utility of this approach by using our newly developed method when inferring the phylogeny of the annelid family Nereididae, in the largest phylogenetic study of this group to this date.

3 The untapped potential of genome skimming in the age of chromosome-scale genomics

Felix Thalén[1,2*] and Christoph Bleidorn[1]

[1]Dept. for Animal Evolution and Biodiversity, Georg-August-Universität Göttingen, Untere Karspüle 2, 37073 Göttingen, Germany

[2]Cardio-CARE AG, Medizincampus Davos, Herman Burchard Str. 1, 7265 Davos, Switzerland

[*]Corresponding author: felix.thalen@cardio-care.ch

3.1 Abstract

Although whole-genome sequencing (WGS) provide many benefits over the most prevalent data collection strategies for large-scale phylogenetics, this approach has largely gone unutilized, most likely due to the additional costs of sequencing and a lack of appropriate bioinformatic tooling. Lately, increased output from high-throughput sequencers, combined with new, alignment-based methods for genome skimming, have reduced costs and widened its utility. Despite this, we continue to observe an unrealised potential when it comes to using WGS data for phylogenomic studies. Until now, genome skimming has, in the context of phylogenomics, primarily been used to target high-copy sequences such as organellar genomes (mitochondria and plastids) as well as repetitive elements. Here, we examine other areas of applications such as using WGS to estimate genomic parameters and or recovering nuclear genes. We compare this to alternative approaches in said applications while arguing the overall benefits of this strategy.

Keywords: genome skimming, whole-genome sequencing, high-throughput sequencing, phylogenomics, genomic partitioning, target enrichment, transcriptomics

3.2 Introduction

Advancements in high-throughput sequencing methods revolutionized the field of molecular systematics and genomics (Levy & Myers, 2016). Reduced costs and the unprecedented amount of sequencing data also allowed using "omics"-approaches for non-model organisms (Ekblom & Galindo, 2011). Especially, Illumina-based short-read sequencing became prominent and was widely used in both transcriptomic and genomic studies. Due to its short read size (< 1000 bp; usually around 150-250 bp), however, its limitations for the reconstruction of highly continuous eukaryotic genomes became obvious. E.g., resulting assemblies were shorter than expected, repeat regions were not well-resolved and some coding exons were completely missing (Alkan et al., 2011). Also, gene content analyses of these highly-fragmented draft genomes resulted in erroneous numbers and the usefulness of such data in comparative genomic analyses has been doubted (Denton et al., 2014). Nowadays, single-molecule long-read sequencing is the gold standard to achieve chromosome-scale assemblies of complex genomes (Rhie et al., 2021). However, the latter techniques need high molecular weight DNA from freshly sampled or well-preserved material. In contrast, sequencing libraries for short-read sequencing can be successfully constructed from low amounts and highly-fragmented DNA, which is mirrored in the achievements in the field of ancient genomics (Der Sarkissian et al., 2015). The combination of minimum requirements of the input DNA coupled with low sequencing costs and high output of the latest sequencer generations (e.g,, the Illumina NovaSeq 6000 platform) makes the generation of discontinuous draft genomes a cost-effective and accessible alternative in the broad field of evolutionary genomics. While analysing of such draft genomes (often low coverage) has been relegated to the retrieval of high-copy number markers (e.g., organellar genomes), we will line out the potential of such data in the light of the development of new bioinformatic tools.

3.3 Estimating genomic parameters

Estimating genome size and repeat content is not only necessary when investigating the evolution of genome size, but also when planning genome sequencing projects. In the latter case, the estimated genome size aids to fine-tune the needed sequencing depth for the targeted genome (Sims et al., 2014). Genome size among eukaryotes show a huge variation often attributed to differences in the content of repetitive elements (Lynch, 2007). Eukaryotic genome size ranges from around 2 mbp (in parasitic Microsporidia) to up to 150 Gbp (in the plant *Paris japonica*; Elliott and Gregory, 2015. Interestingly, it has been shown that this variation does not scale with complexity (e.g.., as measured in the number of cell types or protein coding genes), a notion which is long known as the "C-value paradox" (Thomas Jr, 1971). The "C-value" refers to the amount of DNA in a haploid nucleus, a measure which has been traditionally used to estimate genome size. In the wet lab, the haploid DNA content can be analyzed using flow cytometry (Doležel & Greilhuber, 2010). However, for this method nuclei have to be isolated from fresh material, often a severe limitation when working with non-model organisms. Moreover, as it represents a relative measure, comparison with a reference sample with known genome size is necessary. Alternatively, genome size can be also estimated from sequence reads. The two most widely used strategies are based on investigating k-mer distribution or estimation of coverage of single-copy genes from read mapping data (Pflug et al., 2020). For most approaches, a decent coverage (10x or higher) is required for reasonable genome size estimated from sequence data. A read-mapping approach is represented by ModEst (Pfenninger et al., 2022), a method which relies on statistics of mapping sequence reads back to the resulting assembly to estimate the sequencing depth distribution and to infer genome size estimates. Recently, it has been demonstrated by Sarmashghi et al., 2021 that analyses of k-mers, by their software RESPECT, can yield reliable estimates of the length and repeat content of a genome from as low as 1x-coverage datasets.

3.4 Taxon sampling

The first step towards any phylogenetic study is to collect all of the species that one wish to include. Although an inadequate taxon sampling is though to produce poorly supported phylogenies and contradictionary hypotheses (e.g., Pick et al., 2010), an exhaustive taxon sampling is not always feasible due to time-constraints, monetary reasons, geographical location, inaccessibility of the species of interest, administrative burdens, or all of the above.

Until recently, genetic material stored in museum collections were often thought to be too degraded to use the specimen for high-throughput sequencing (Hofreiter, 2012) and consequently, most large-scale phylogenetic analyses were conducted using freshly-collected material. Nowadays, more and more studies successfully sequence material from natural history museum collections using either Whole-Genome Sequencing (WGS; e.g. Cong et al., 2017; F. Zhang et al., 2019b) or a TE approach (e.g., Breinholt et al., 2018; Call et al., 2021b). This recent development has opened up a new field known as museomics. Not only does this recent development allow scientists to study pre-labeled genetic material in a non-destructive way, this also extends to the study of extinct species or populations over time.

With the introduction of the Nagoya protocol, which came into effect in 2014, phylogenetic researchers who wish to sample new biological material face an increasing amount of bureaucratic work (Neumann et al., 2018). The Nagoya protocol was introduced to ensure that the benefits of genetic resources are shared equally among those who acquire and those who provide biological material, and that those resources are obtained in a fair and sustainable manner and with the conscent from the affected authorities (Buck & Hamilton, 2011). In practice, however, it has been critized to introduce beaurocratic barriers that hinder or slow down scientific research (Neumann et al., 2018). Moreover, discrepancies among the implementation and interpretation of the Nagoya Protocol across different countries further adds to this problem (Sherman & Henry, 2020). E.g., there were disagreements among countries whether or not the protocol should apply retroactively to genetic material collected before the date in which the protocol came into effect and consequently, thus leaving this decision to each individual country (Lassen et al., 2016). Irregardles, the Nagoya Protocol drastically changes the way in how the affected scientists operate; this simultaneously opens up an area where material collected before the Nagoya Protocol came into effect are suddenly increasingly attractive due to the reduced amount of bureaucracy.

3.5 Collecting high-throughput phylogenomic data

3.5.1 Reduced representation strategies

Phylogenomics, in the last decades, have largely come to be dominated by transcriptome sequencing (RNA-seq; e.g., Cannon et al., 2016; Kocot et al., 2017) and hybridization or capture-based target enrichment (e.g., Bi et al., 2012 or A. R. Lemmon et al., 2012b). These methods are also known as genomic partitioning or reduced representation strategies, since they target a subset of the genome. By only sequencing a select portion of the genome, one can reduce both computational- and sequencing costs (Turner et al., 2009).

Transcriptomics

One way to reduce sequencing and data processing costs is to only sequence the transcriptome of an organism. The transcriptome is the expressed portion of the genome, and its quantity, at a certain developmental stage and under certain physiological conditions (Wang et al., 2009). In the infancy of phylogenomics, transcriptomics was the most commonly used approach to large-scale phylogenetics and it continuous to be a popular approach to this date. One of the major drawbacks of RNA-Seq is that normally a relatively large amount of freshly collected (or RNAlater-preserved) material is needed, thus excluding ethanol-stored samples and making it more difficult to sequence smaller species as these needs to be pooled together to meet quantity requirements (Allen, Boyd, Nguyen, et al., 2017).

Target Enrichment

Another cost-efficient approach for generating phylogenomic data, that even works with degraded, ethanol-preserved material, is to capture a selected set of loci using target enrichment methods. The targeted loci can then be sequenced in greater depth, since not all of the genome has to be targeted (E. M. Lemmon & Lemmon, 2013) and the data that is generated is easier to process (Mamanova et al., 2010). In Target Enrichmennt (TE), a set of oligonucleotide probes (also known as baits) are designed from a reference and these probes are subsequently used to target genomic regions of high sequence similarity (e.g., exons of a set of pre-selected orthologs). One of the main difficulties lies in designing probes that not only capture the species from which the probes were derived, but that can also capture across the entirety of the taxon sampling (Bragg et al., 2016b; Hawkins et al., 2016).

3.5.2 Whole-Genome Sequencing

Like the name suggests, Whole-Genome Sequencing (WGS) means that the genome—both coding and non-coding portions—of an organism is sequenced in its entirety, together with mitochondrial or chloroplastic genomes. To date, phylogenomic studies has largely been dominated by the use of a genome reduction approach (with some exceptions such as the Bird 10K genome project [G. Zhang, 2015] or a phylogenomic study on yeast [X.-X. Shen et al., 2016]) but as the cost of sequencing has decreased and as the output of high-throughput sequences has increased, we are quickly approaching a point at which sequencing the entire portion of the genome is the most optimal data collection strategy (E. M. Lemmon & Lemmon, 2013). Using WGS for phylogenomics has many advantages over genomic partioning strategies such as reduced laboratory workload, lower requirements in terms of quantity and quality of the material used, and higher diversity of targeted loci (F. Zhang et al., 2019b). With WGS, there is no need for marker development and optimization, and applications beyond phylogenetics—e.g., genome organization, studies using non-targeted genes, and assembly of mitochondrial and microbial associates—can be pursued using the same, unbiased representation of the generated genome (Allen, Boyd, Nguyen, et al., 2017). Some of the downsides of using WGS for phylogenomics include an increased computational workload and a lack of appropriate tooling and or standardized bioinformatic pipelines for processing such data.

A short comparison of the of three different sequencing techniques—namely, RNA-seq, target enrichment, and genome skimming—is shown in the table 3.1.

3.6 Processing Whole-Genome Sequencing Data

One of the main challenges when it comes to using whole-genome sequencing data for phylogenomics lies in the added computational costs; since not only targeted markers are sequenced, there is additional complexity when it comes to handling such datasets (Allen, Boyd, Nguyen, et al., 2017). If a reference genome is available, the genome can be assembled using a reference-guided assembler, and if no such genome exists, the assembly may be performed *de novo* (e.g., Prjibelski et al., 2020, Peng et al., 2010, or

Table 3.1: Comparison of three different sequencing strategies: RNA-seq, targetted enrichment, and genome skimming; highlighting their differences, areas of applications, as well as potential advantages and or disadvantages. Table adapted, with alterations, from (Richter et al., 2015).

	RNA-Seq	**Target Enrichment**	**Genome Skimming**
Material	RNA	DNA	DNA
Prior genomic resources needed	No	Yes	No
Recommended taxon number	Flexible	High number recommended	Flexible
Genome size of species	Less important	Less important	Important
Workload	Time-intensive	Time-intensive	Fast and easy
Ability to identify single-copy genes	Yes	Yes	Maybe
Ability to distinguish different isoforms	Yes	No	No
Ability to analyze expession levels	Yes	No	No
Ability to analyze intron-exon structure	No	Yes (with prior information)	Yes

Simpson et al., 2009). Depending on the sequencing depth and consequently the size of the data, these types of analyses typically requires having access to a high-performance cluster or a cloud computing platform. One way to reduce the computational costs of assembling NGS data is through using targeted locus assembly (e.g., aTRAM [Allen et al., 2015; Allen et al., 2018b] or Kollector [Kucuk et al., 2017]). By only assembling a reduced set of targeted loci, the amount of time spent assembling a genome can be significantly reduced, but depends on the size loci. For many phylogenetic studies, only a specific set of—typically conserved—loci are needed and thus the whole assembly process can be sped up without much loss of information.

3.6.1 Alignment-based Marker Discovery Techniques

Post-assembly, there are now several are available for extracting a specific set of markers from WGS data, directly suitable for molecular phylogenetic studies. E.g., for extracting ultra-conserved elements (Phyluce; Faircloth, 2016) or near-universal single-copy orhthologs (BUSCO; Waterhouse et al., 2018). Unfortunately, many phylogenomic studies include low-coverage genomes (around 2–10x), as a way to increase the number of taxa. Consequently, these sequencing runs typically results in poorly-assembled, fragmented genomes. Thus, one of the primary challenges when working with these types of datasets stems from the ability to correctly infer conserved loci—not only from multi-exon genes but also—from fragmented and incorrectly assembled genomes. A new suite of tools, (e.g., ALiBaSeq [Knyshov et al., 2021b] and Patchwork [Thalen et al., 2022; also see chapter 4]) overcomes this challenge by including a "hit stitching" phase, in which overlapping alignments are pieced together into larger regions. This not only helps putting together multi-exon genes, residing on two or more contigs, but also for retrieving markers from sub-optimal assemblies resulting from a limited amount of raw data, or limits of short-read technologies in general. Thus, these new approaches are especially useful when working with low-coverage (i.e., 2-10x) genomes, which are notoriously difficult to assemble because of fragmented or missing genes. Nevertheless, it might be these types of genomes that researchers have to work with, when trying to cover the widest range of taxa possible or when sequencing sub-optimally-stored samples, e.g., from a museum collection.

3.6.2 Organelle genomes as a byproduct

The organelles of eukaryotic cells, such as the mitochondria and chloroplasts, have their own genomes that differ significantly from eukaryotic nuclear genomes in terms of size and in that they are circular (Van Bruggen et al., 1966). Although organelle genome may be targetted more directly, they are typically sequenced alongside the nuclear genome and can thus be retrieved as a biproduct from another sequencing approach such as high-throughput WGS (Al-Nakeeb et al., 2017). Moreover, these may be 10-100 times more frequent in a cell than the nucleues (Robin & Wong, 1988), which in practice means that organelle genomes will have a higher read depth and a higher coverage than the nuclear genome. Several tools exist to obtain mitochondrial genomes from WGS data (Al-Nakeeb et al., 2017; Dierckxsens et al., 2017; Hahn et al., 2013). These may get the desired reads through alignment to a reference or by exploiting the forementioned fact that organelle reads typically have a higher coverage.

Once extracted, mitochondrial DNA (mtDNA) can be used as markers for phylogenetic inference (Hassanin et al., 2012), COI barcoding for species identification, delimitation, or confirmation (Hebert et al., 2003), or for comparing mitochondrial gene orders (Weigert et al., 2016). Several software programs already exists for assembling (Dierckxsens et al., 2017; D. Li et al., 2016; Nurk et al., 2017) and or annotating mitochondrial genomes (Allio et al., 2020; Bernt et al., 2013). Thus, organelle genomes recovered as a biproduct of another sequencing approach can easily be used to complement another analysis by providing barcodes for species identification and delimitation. Furthermore, mitochondrial gene order information and or phylogenetic analyses of organelle genes, done on the nucleotide- or the amino acid-level, may be used provide additional support for one or more hypotheses generated by another, genome-scale phylogenetic (phylogenomic) analysis, done using nuclear genes.

3.6.3 Future outlook

As the costs of sequencing has continued to go down in the past, it would only be reasonable to assume that this trend will continue for the foreseeable future. Like previously discussed, one of the main arguments *against* WGS has been the increased costs of sequencing. If the costs do indeed continue to go down to the point where we can even sequence larger genomes (>1000 Mbp), at a modest coverage (10-30x) and for a reasonable cost, then we have truly reached the point where this approach is the most reasonable one for most large-scale phylogenetic studies. Although high-throughput sequencing has largely come to be dominated by Illumina (formerly known as Solexa), new companies, challenging their dominance, continue to emerge. E.g., Ultima Genomics recently described a new massively parallel sequencing by synthesis SBS approach in which samples are sequenced reads of 300bp in length are sequenced with a high accuracy (Q30 > 85%) at a cost of 1$/Gb (Almogy et al., 2022).

At the same time, we think that as more and more data continues to be produced at an ever-accelerating speed and scale, the same time of thought and investment should—or will have to—go into analyzing these datasets. Lack of bioinformatic tooling or expertise continues to be a major bottleneck for this approach, and the lack of standardization in bioinformatic tooling for analyzing WGS data in a phylogenomic context means that only institutions with access to big compute clusters and bioinformatic expertise will be able to carry out these types of analyses. If instead the focus is shift to the development of fast and easy-to-use software tools and streamlined pipelines for processing this data type, this approach could be carried out by a much larger variety of researchers. The emergence of workflow management systems aimed at bioinformatics (primarily Nextflow [Di Tommaso et al., 2017] and SnakeMake [Köster and Rahmann, 2012]), in combination with a plethora of cloud providers (e.g., Amazon Web Services (AWS), Google Cloud Platform, or Microsoft Azure), environment managers (e.g., Anaconda), and virtualization by containers (e.g., Docker), has simultaneously led to a point where implementing, re-purposing, maintaining, and using bioinformatic pipelines is now easier than ever. Although cloud computing does not remove the cost of computing, it does enable researchers whom do not have access to a local high-performance cluster to perform the same type of resource-intense analyses as those who do. And although we are still far from a point where phylogenomic analyses have been standardized, new workflow managers, combined with containerization, could quickly shift the current situation into one where phylogenomic pipelines are more accessible than ever.

3.6.4 Conclusions

Large-scale phylogenetics continues to be dominated by target enrichment TE and RNA-Seq. Despite the emergence of long-read sequencing for routinely producing high-quality, near-chromosome level assemblies, sequencing costs and material requirements prohibits that this sequencing technique is used at a scale which would be desirable for the majority of molecular phylogeneticists. Increased output from the latest generation of high-throughput sequencing platforms, however, means that whole-genome sequencing WGS—in which an organism's genome is sequenced in its entirity—may now be used to sequence moderately-sized genomes, at the desired coverage (at least low-coverage, i.e., 2-10x), for a reasonable price. Furthermore, new sequencing platforms, right around the corner, and increased competition within this space, promises further price decreases down the line. Thus, it is safe to say that we are approaching—if we are not already at—the point at which WGS becomes the most sensible approach for the majority of large-scale phylogenetic studies. Indeed, WGS offers reduced laboratory workload, reduced material requirements, and broader utility and re-utility of the markers that are recovered.

One of the biggest bottleneck in WGS sequencing for phylogenomics today is the lack of appropriate tooling, especially when it comes to working with non-model taxa. The emergence of new bioinformatic software within this field, just within the last few years, promises reduced computational workloads by specifically targetting—and sometimes only assembling—a pre-selected set of ultra-conserved elements UCEs or near-universal single-copy orthologs USCOs. The difficulty in using some of these tools, however, means that these types of analyses are restricted to workgroups with in-house bioinformatic expertise. This could, on the one hand, be seen as the result of an institutional failure to fund and support research software engineers to work side-by-side other researchers in a more traditional lab setting. A majority of bioinformatic software today continues to be written as one-off scripts or solutions to accompany a scientific paper where the idea of active software maintenance comes as an afterthought. Irregardless, increased ease-of-use and accessibility of software tools for phylogenomics, especially when handling WGS data, will benefit everyone. Moreover, recent development in the space of workflow managers for bioinformatics has made the development of bioinformatic pipelines to streamline and standardize these types of analyses much easier.

In summary, we encourage other researchers to consider and utilize a WGS approach for phylogenomic studies. We especially think that continued development of bioinformatics software for mining phylogenetic markers and subsampling WGS data will encourage others to do so. Finally, increased output and reduced costs in up-and-coming, short-read sequencing platforms continues to push the limits of this approach.

4 Patchwork: alignment-based retrieval and concatenation of phylogenetic markers from genomic data

Felix Thalén[1,2*], Clara G. Köhne[1], and Christoph Bleidorn[1]

[1]Dept. for Animal Evolution and Biodiversity, Georg-August-Universität Göttingen, Untere Karspüle 2, 37073 Göttingen, Germany

[2]Cardio-CARE AG, Medizincampus Davos, Herman Burchard Str. 1, 7265 Davos, Switzerland

[*]Corresponding author: felix.thalen@cardio-care.ch

4.1 Abstract

Motivation: Increased output from the latest short-read sequencers makes low-coverage whole-genome sequencing (LC-WGS) an increasingly affordable approach to large-scale phylogenetics. Despite offering several advantages over prevailing sequencing strategies, few tools exist to work with this data type within a phylogenomic context. Due to the fragmented nature of LC-WGS genomes, their use have mostly been restricted to easy-to-assemble, high-copy-number regions such as organelle genomes and or ribosomal genes.

Results: We here present a new method for mining phylogenetic markers directly from an assembled genome. Homologous regions are obtained via an alignment search, followed by a "hit-stitching" phase, in which adjacent or overlapping regions are concatenated together. Finally, a novel sliding window technique is used to trim non-coding regions from the alignments. We demonstrate the utility of Patchwork by recovering near-universal single-copy orthologs (USCOs) in the annelid *Dimorphilus gyrociliatus*.

Availability: Patchwork is available from Github (github.com/fethalen/Patchwork) under the GNU General Public license version 3.

Keywords: computational biology and bioinformatics, genome informatics, next-generation sequencing, phylogenomics, phylogenetics, software

4.2 Introduction

Advancements in high-throughput sequencing techniques have revolutionized the field of phylogenetics and ultimately our understanding of the tree of life (A. R. Lemmon et al., 2012a). The availability of genomic and or transcriptomic data for basically all desired taxa and for a reasonable price has transformed the field to phylogenomics—genome-scale phylogenetic systematic analyses (McCormack et al., 2013). Some challenges remain, however, as many studies still show incongruent results, lack of branch-support, or resolution (Philippe et al., 2017). Even though complete genomes are available for more and more eukaryotes, most large-scale phylogenomic studies to date were conducted using either transcriptome sequencing (e.g., Andrade et al., 2015b; Weigert et al., 2014) or a genome subsampling methods such as targeted-enrichment (e.g., Andermann et al., 2020; Call et al., 2021a; Sann et al., 2018) which focuses on a set of pre-selected loci.

4.2.1 Reduced representation vs. WGS strategies

Transcriptome sequencing offers a way to sequence only the expressed portion of a genome without prior sequence knowledge. Unfortunately, this approach requires freshly collected material or material stored in a specific manner (e.g., deeply shock-frozen, RNAlater; Cronn et al., 2012a). Furthermore, smaller specimen may need to be pooled together to attain sufficient amounts of mRNA and such practice risks mixing up individuals with undetected genetic variation (Allen, Boyd, Nguyen, et al., 2017). Unfortunately, a large amount of collected specimen only exist in natural history museum collections and most of these are ethanol-preserved and thus not usable for transcriptomic studies (Call et al., 2021a). This is undesirable as taxon sampling is considered one of the most important factors for accurate phylogenetic tree reconstruction (Heath et al., 2008).

Target-enrichment approaches, on the other hand, require prior knowledge of target sequences (e.g., from well-annotated genomes) for the construction of oligonucleotide probes. Moreover, the number of enriched targets are limited by the amount of oligonucleotides included in the enrichment kit of choice and the efficiency of such approaches decreases as the distance bait-to-target distance increases (Bragg et al., 2016a). Another downside is that the data produced have few applications outside of phylogenomics and if one wants to add a taxon to the study, they need to use the exact same markers as previous studies (Allen, Boyd, Nguyen, et al., 2017).

A viable alternative is low-coverage whole genome sequencing (LC-WGS; also known as “shallow genome sequencing”, or “genome skimming”) using short-read techniques such as Illumina sequencing. Relying solely on this technique has been shown to be inadequate for the reconstruction of highly contiguous reference-quality genomes (Rhie et al., 2021). However, due to the introduction of newer sequencing platforms (e.g., Illumina's NovaSeq sequencing platform) WGS became relatively cheap (Schwarz et al., 2021) and even highly fragmented DNA can be used as input. Consequently, LC-WGS can be used to generate data from various sources of targeted organisms to retrieve marker loci on a genome scale. While this so called "genome-skimming" approach has frequently been used to reconstruct organellar genomes (e.g., Jin et al., 2020; Richter et al., 2015), it is currently underutilized in the field of phylogenomics.

Short-read assemblies of eukaryotic genomes tend to be highly discontinuous and automated annotation of such large, fragmented genomes remains difficult (Salzberg, 2019). Eukaryotic genomes are characterized by the presence of "genes in pieces", where introns interrupt coding sequences and exons (Rogozin et al., 2005). Depending on the coverage, short-read draft genomes are characterized by low N50s in the range of few kbp (if at all) (Salzberg et al., 2012) and consequently, exons of a single gene usually end up on several contigs in fragmented genomes.

4.2.2 Marker discovery techniques

The disuse of genome skimming in large-scale phylogenetics could potentially be ascribed to the lack of suitable data analysis methods (Philippe et al., 2011; F. Zhang et al., 2019a). Existing software tools for working with WGS data in a phylogenomic context, such as aTRAM (Allen et al., 2018a), ALiBaSeq (Knyshov et al., 2021a), and GeMoMa (Keilwagen et al., 2018; Keilwagen et al., 2016), are either difficult-to-use and or written in an interpreted language (e.g., Perl or Python) which does not allow the program to scale well with the large biological datasets that are commonplace today (Knyshov et al., 2021a).

To address the limitations typically associated with LC-WGS, we present Patchwork, an alignment-based tool for mining phylogenetic markers, directly from WGS data. Patchwork utilizes the sequence aligner DIAMOND (Buchfink et al., 2021), and is written in the programming language Julia (Bezanson et al., 2017), to achieve the best possible speed, thus allowing Patchwork to scale well with today's genome-scale datasets. In addition, our implementation focuses on ease-of-use; while pre-existing methods may require the user to perform the sequence alignments separately, our program handles each step in the analysis—from start to finish.

4.3 Implementation

Patchwork is a reference- and alignment-based method for mining phylogenetic markers from WGS data. One or more reference protein sequences guide the "stitching" process, where the best-scoring, translated query nucleotide sequences are merged into continuous stretches of amino acid sequences. Merged sequences go through a masking step, where unaligned residues, ambiguous amino acid characters, and stop codons are removed from query sequences. Finally, Patchwork implements a sliding-window based alignment trimming step to rid the resulting sequences from poorly aligned residues, e.g., due to putative non-coding regions. The aim of Patchwork is to capture multi-exon or fragmented genes, scattered across different contigs in an assembled genome. Moreover, this method allows sequences of different data types (i.e., genomic and transcriptomic data) to be combined into a single dataset.

The core of Patchwork is implemented in the relatively new scientific programming language Julia. Julia strives to be as performant as possible, while still retaining a high level of productivity. Existing libraries such as BioAlignments.jl (https://github.com/BioJulia/BioAlignments.jl) and BioSequences.jl (https://github.com/BioJulia/BioSequences.jl) further sped up the development itself. Patchwork is obtainable from GitHub (https://github.com/fethalen/Patchwork), it is released under the GPLv3 license and targets both Linux and macOS. To make the installation of Patchwork easier, we also provide a Docker image that contain Julia, Patchwork, and DIAMOND, one of the external dependency.

Table 4.1: Comparison of software used for evaluation, reproduced from Knyshov et al., 2021a.

Software	Input	Sequence type	Search engine	Reference
Patchwork	Reads/assemblies	DNA ⟶ AA	DIAMOND	This study
AliBaSeq	Assemblies	DNA ⟶ DNA/AA	Multiple	Knyshov et al., 2021a
GeMoMa	Assemblies/genomes	DNA ⟶ DNA	MMseqs/BLAST	Keilwagen et al., 2016
aTRAM	Reads	DNA ⟶ DNA	BLAST	Allen et al., 2018a

4.4 Algorithmic Overview of Patchwork

Patchwork's workflow can be divided into five different steps: (i) pooling of reference sequences, database construction, and initial alignment, (ii) hit stitching , (iii) alignment masking, (iv) alignment trimming, and (v) final alignment, filtering, statistical reports, and plots.

4.4.1 Initial alignment and database construction

First, all reference protein sequences—regardless of whether they are spread across multiple FASTA files or not—are pooled together into a single FASTA file, from which a DIAMOND database is created. There is also the option to use an existing DIAMOND-formatted database or a BLAST output file in a tabular format by using the `-database` or `-tabular` option respectively. These files are both provided in the output of Patchwork and can thus be re-utilized when trying out different parameters. In either case, DIAMOND's BLASTX algorithm is used to align translated nucleotide sequences to one or more reference protein sequences.

Like DIAMOND, Patchwork, by default, scores alignments using the substitution matrix BLOSUM62 (Henikoff & Henikoff, 1996), a gap open penalty of 11, and an extension penalty of 1. Other—built-in or custom—substitution matrices may be used in place of the default option. User-chosen gap open- and gap extension penalties may also be chosen, as long as they are within the limits set by the substitution matrix of choice. DIAMOND-specific options may be set using the `-diamond-flags` option. For the user's convenience, flags such as `-evalue` for changing the maximum expected value. Lastly, DIAMOND sensitivity modes (e.g., `-very-sensitive` and `-ultra-sensitive`), Buchfink et al., 2021, all have easy-to-access options as well.

Since the alignment search is likely to result in more than one hit, certain measures are taken to ensure that none of these hits are overlapping: They are, "hit stitching" (also known as contig- or exon stitching; i.e., merging of overlapping regions), removal of unaligned residues, and concatenation of non-overlapping regions. A graphical overview of these can also be seen in figure 4.1.

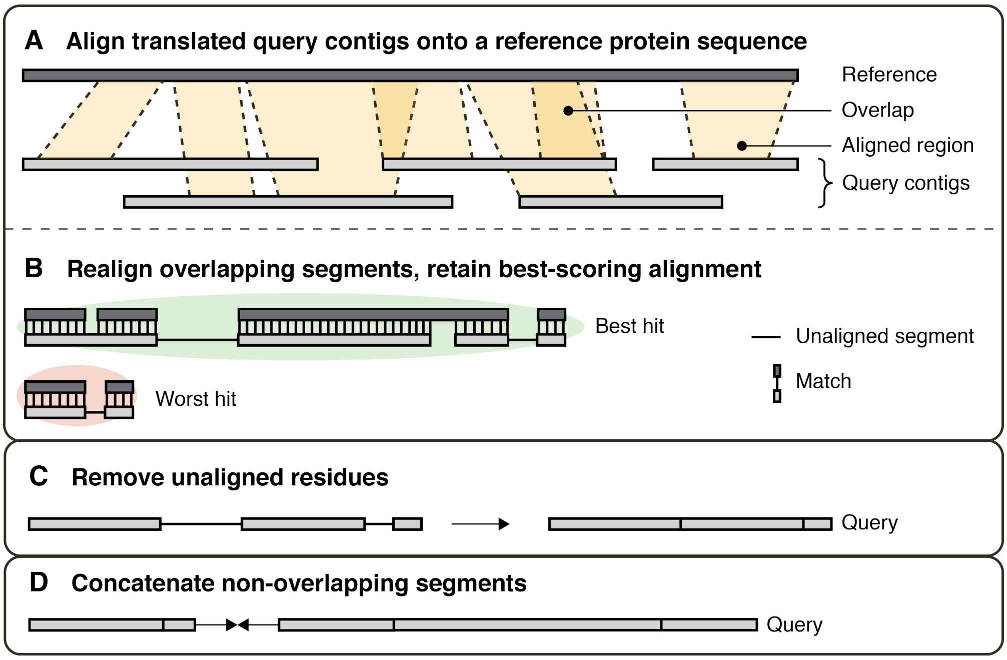

A Align translated query contigs onto a reference protein sequence
Reference
Overlap
Aligned region
Query contigs
B Realign overlapping segments, retain best-scoring alignment
Best hit
Unaligned segment
Match
Worst hit
C Remove unaligned residues
Query
D Concatenate non-overlapping segments
Query

Figure 4.1: Graphical overview of the Patchwork algorithm. First, (A) query sequences are aligned to the provided reference sequence. These alignments may or may not be overlapping. (B) Overlapping alignments are realigned but only in the area in which they overlap. The best-scoring alignment is retained while all others are discarded. (C) Non-aligned residues are then removed and (D) the remaining regions are concatenated into a single, continuous sequence.

4.4.2 Hit stitching

The merging algorithm works as follows: first, all regions are sorted by their first *and* last position at which they align to the reference sequence. The first region is added to the stack and then for each pair of regions, check if the two regions are overlapping. If they are not overlapping, add the rightmost region to the stack and continue. If they are overlapping, however, realign the overlapping region to identify the best-scoring sequence at that particular interval. Then, based on the realignment score, slice the sequences such that the best-scoring sequence stays at the overlapping region and so that non-overlapping, flanking regions, if existing, are retained as well.

Different aligned regions from the same contig are allowed to be stitched together. While "hit stitching" may result in the creation of chimeric sequences (i.e., two or more biological sequences incorrectly joined together), this procedure has the potential to increase coverage and to (correctly) join two or more regions that are located on separate contigs due to an erroneous assembly, a sequencing error, and or a multi-locus gene.

4.4.3 Alignment masking

At this step, unaligned residues, ambiguous amino acid characters, and stop codons (also known as "termination codons") are all removed from the resulting query sequence. Query sequences may contain residues which do not align to any particular region of the subject sequence. Such regions may, e.g., be non-coding regions or simply insertions. In either case, unaligned residues are removed on the basis that inserts are less likely to constitute phylogenetically informative sites and risks introducing untranslated regions and therefore biasing the downstream analysis. Similarly, ambiguous amino acids are most likely non-informative and stop codons are a clear indicator that non-coding characters have been included in the alignment. Although such regions are likely to be removed in the subsequent step (see section 4.4.4), the user may choose to keep stop codons and or ambiguous amino acid characters by providing the flags `-retain-stops` and or `-retain-ambiguous`.

4.4.4 Sliding window-based alignment trimming

One side effect of aligning translated nucleotide sequences to amino acid sequences is that one might recover noncoding portions of DNA, provided that the following two conditions are fulfilled: (i) the noncoding DNA is located in between two or more coding portions and (ii) there is a sequence region in the reference sequence that the noncoding region can align to. In the resulting alignment, noncoding portions are characterized by many indels, intercepted by occasional matches. The alignment of noncoding portions of DNA can already be observed in the alignments produced by DIAMOND and thus this side effect does not stem from Patchwork itself. In fact, the Patchwork algorithm will only include noncoding parts if nothing else aligns better to the affected region of the reference sequence.

To mitigate this effect, we have implemented a sliding window-based alignment trimming approach (see figure 4.2) to rid the alignments from these unwanted regions. This works by scanning the alignment from left to right, cutting all regions where the average distance is below the user-provided distance threshold. The window size and the distance threshold are both set by the user and this entire step can be skipped over in its entirety. This approach tries to avoid cases where a single bad, but correct, match would have otherwise been cut out.

4.4.5 Concatenation and realignment of remaining regions

Finally, the resulting set of ordered, non-overlapping sequence regions, are concatenated into one, continuous sequence. The concatenated sequence is then realigned to the reference to obtain the final output sequence and alignment score.

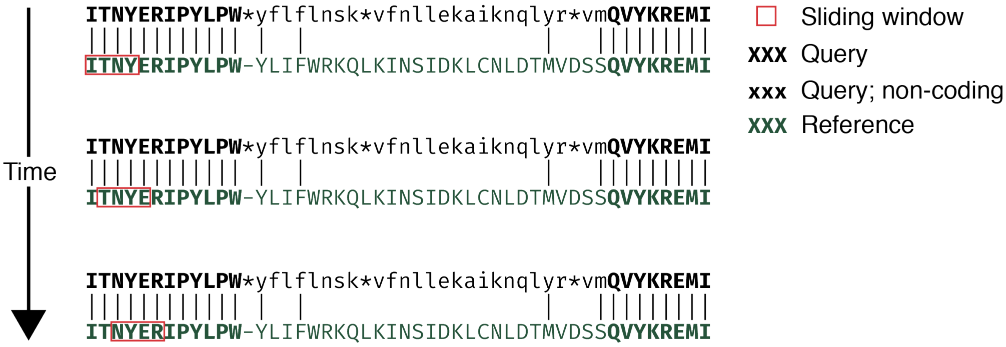

ITNYERIPYLPW*yflflnsk*vfnllekaiknqlyr*vmQVYKREMI
ITNYERIPYLPW-YLIFWRKQLKINSIDKLCNLDTMVDSSQVYKREMI
ITNYERIPYLPW*yflflnsk*vfnllekaiknqlyr*vmQVYKREMI
ITNYERIPYLPW-YLIFWRKQLKINSIDKLCNLDTMVDSSQVYKREMI
ITNYERIPYLPW*yflflnsk*vfnllekaiknqlyr*vmQVYKREMI
ITNYERIPYLPW-YLIFWRKQLKINSIDKLCNLDTMVDSSQVYKREMI
Time
□ Sliding window
XXX Query
xxx Query; non-coding
XXX Reference

Figure 4.2: Graphical depiction of the sliding window-based alignment trimming approach. The upper sequence represents the (translated) query sequence aligned to the lower reference sequence, here colored in green. The sliding window, shown in red, moves to the right and removes all residues within, when the average distance falls below the specified threshold. Here, retained residues *after* trimming are shown in a bold font. The putative, noncoding region is shown in lowercase.

4.4.6 Integrating Patchwork in a phylogenomic pipeline

Most phylogenomic studies include more than a handful taxa and manually concatenating these gets increasingly tedious as the dataset increases in size. Thus, to streamline and simultaneously speed up the downstream analysis, Patchwork includes a set of complementary tools for working with multiple datasets at once. First, `multi_patchwork.sh` can be used to (i) run Patchwork on multiple input files at once and to (ii) concatenate homologous sequences from different taxa into one and the same file(s). The resulting amino acid sequences are in FASTA format and thus the exact downstream analysis used is highly flexible. Nevertheless, a hypothetical bioinformatic pipeline for large-scale phylogenetics may look like the following: (i) BUSCO (Manni et al., 2021; Simão et al., 2015) is used to obtain an initial set of near-universal single-copy orthologs (USCOs) from a (preferably) high-quality genome or transcriptome, sequenced and assembled *de novo* or obtained from a publicly available database; (ii) `multi_patchwork.sh` is used to run Patchwork on a set of assembled genomes, using the USCO-set as a reference, (iii) all resulting markers are aligned using a sequence aligner such as MAFFT (Katoh et al., 2009) or an alignment and alignment trimming program such as GUIDANCE (Penn et al., 2010); finally, (iv) resulting alignments could be filtered for missing data and concatenated into a supermatrix using a tool such as PhyloPyPruner (github.com/fethalen/phylopypruner; Thalén et al. in prep). The resulting data matrix may then be analysed using a multitude of phylogenetic tree inference methods; e.g., maximum likelihood, Bayesian inference, or a coalescent-based methods.

4.5 Benchmarking

To assess the utility and the overall performance of Patchwork, we designed a small study set to recover near-universal single-copy orthologs (USCOs) from genomic Illumina short-read sequence data of the marine annelid *Dimorphilus gyrociliatus*, with the data originally generated as part of a study by Martın-Durán et al., 2021. Aside from having an annotated version of the genome publicly available, one of the main advantage of using *D. gyrociliatus* for this study is that the genome is relatively small in size (i.e., 73.82 Mb). As a consequence, assembling a read data for an annelid genome with such a high a high gene density (208.86 genes per Mb) is easier because the read depth and coverage is much higher as a result. However, as we only used short-reads, we created a highly discontinous assembly with low N50 as typical for low-coverage genomic datasets.

This benchmark consists of two phases: In phase 1, we align a short-read-only assembly of *Dimorphilus gyrociliatus* against near-universal single-copy genes found in the long-read, annotated assembly of itself, mentioned above. In phase 2, we instead align the same short-read assembly against USCOs from a chromosome-level assembly of the nereidid *Alitta virens* (GenBank assembly accession: GCA_932294295.1), published and processed by the Wellcome Sanger Institute.

Elapsed time was calculated from the `real` time as reported by the GNU `time` utility, rounded to the nearest second. All analyses were performed on the Ubuntu v20.04 operating system, with a Linux kernel version of 5.13.0, using an Intel® Xeon® Gold 5120 CPU with 28 threads, running at 2.20GHz.

4.5.1 Genome assembly and quality assessment

Raw sequence data was obtained from the Sequence Read Archive (SRA) using the fasterq-dump v2.10.0 tool and the integrity of the data was verified using md5sum v8.30.

Quality control (QC) of raw and trimmed reads was performed using FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/) and Trim Galore! v0.6.6 (https://www.bioinformatics.babraham.ac.uk/) was used for automated adapter trimming. The *de novo* assembly was performed using SPAdes v3.15.3 (Nurk et al., 2013), using a K-mer size of 55, and the quality of the assembly was assessed using QUAST v5.0.2 (Gurevich et al., 2013), the results of which are shown in Table 4.2. We further assessed the quality of the assembly by using BUSCO v5.0.0 (Simão et al., 2015) while utilizing the Metazoa Odb10 dataset, searching a total of 954 BUSCO groups. Out of the 954 BUSCOs, 844 (88.5%) were complete, 822 (86.2%) were complete single-copies, 22 (2.3%) of these were complete and duplicated, 56 (5.9%) were fragmented, and a total number of 54 (5.6%) were missing.

Table 4.2: Genome assembly quality assessment results of a short-read *de novo* assembly of *Dimorphilus gyrociliatus*, as reported by QUAST (v5.0.2).

K-mer	No. of contigs	Largest contig	Total length	N50	L50
55	55267	682753	123971200	19851	1348

4.5.2 *D. gyrociliatus* SPAdes assembly X *D. gyrociliatus* USCOs

We ran our contigs from our *de novo* assembly with sequences from the annelid *Dimorphilus gyrociliatus* through Patchwork v0.5.0, against a pre-annotated protein sequences from the same species (primary accession no. PRJEB37657). The following settings were used for Patchwork: DIAMOND was run using the flag `-iterate` and while discarding any hits with an E-value above $1 \cdot 10^{-3}$. Concatenated output sequences that were shorter than 30 AAs were discarded and alignment trimming was performed using a window size of 4 and with a mean minimum required distance of 2.

4.5.3 *D. gyrociliatus* SPAdes assembly X *A. virens* USCOs

Subsequently, we ran the same set of contigs from our *de novo* assembly, made with SPAdes, against a set of near-universal single-copy orthologs (USCOs) from *Alitta virens*. Although both organisms are annelids, they are estimated to have diverged more than 480 million years ago (Dos Reis et al., 2015). In today's age, one will typically be able to find publicly-available, qualitatively equivalent sequencing data from a more closely-related taxa, but we wanted to see how Patchwork performs on two such distantly-related groups. The chromosome-level assembly of *A. virens* was first obtained from GenBank (assembly accession: GCA_932294295.1). We ran BUSCO v.5.3.1 (Simão et al., 2015), utilizing the "Metazoa" lineage, to recover a total of 897 USCOs from the chromosome-level assembly of *Alitta virens*. For running Patchwork itself, we ran

DIAMOND using the flag `-iterate` and while discarding any hits with an E-value above $1 \cdot 10^{-3}$. Alignment trimming was performed using a window size of 5 and with a mean minimum distance of -11. No short-sequence filtering was performed in this instance. We subsequently used the `BenchmarkUscos.jl` module from Patchwork to realign the markers we obtained against the pre-annotated set of USCOs in *D. gyrociliatus* itself.

4.6 Results

4.6.1 *D. gyrociliatus* SPAdes assembly X *D. gyrociliatus* USCOs

From the initial 815 markers, we retrieved a set of 788 markers, which corresponds to 96.7% of the total. 27, or 3.3% of the recovered markers were discarded because they were shorter than the 30 AA threshold and 20, or 2.5% of the recovered markers were below 90% in gap-excluded similarity. A detailed summary of these results is shown in table 4.3. Visualizations of percent identity and query coverage from this run are shown in figure 4.3. Running Patchwork took a total of 7 minutes and 7 seconds, from start to finish.

Table 4.3: Results from Patchwork when using a *Dimorphilus gyrociliatus* SPAdes assembly as the query and USCOs from a long-read assembly of *Dimorphilus gyrociliatus* as a reference.

variable	mean	min	median	max
reference_len	449.643	77	351.5	2748
query_len	216.972	33	171.0	2174
regions	1.69543	1	1.0	38
contigs	2.86294	1	1.0	134
matches	215.61	33	167.0	2174
mismatches	1.36168	0	0.0	42
deletions	232.671	0	151.0	2204
query_coverage	55.2736	5.49	51.87	100.0
identity	98.9936	67.06	100.0	100.0

4.6.2 *D. gyrociliatus* SPAdes assembly X *A. virens* USCOs

Out of the 897 *A. virens* USCOs used as a reference, a total of 826 (92.1%) corresponding markers from *D. gyrociliatus* were obtained. 716 of these USCOs successfully aligned back to the pre-annotated set of USCOs found in *D. gyrociliatus*. A detailed summary of these results is shown in table 4.4. Visualizations of percent identity and query coverage from this run are shown in figure 4.4. Running *D. gyrociliatus* against the *A. virens* USCOs took a total of 6 minutes and 10 seconds.

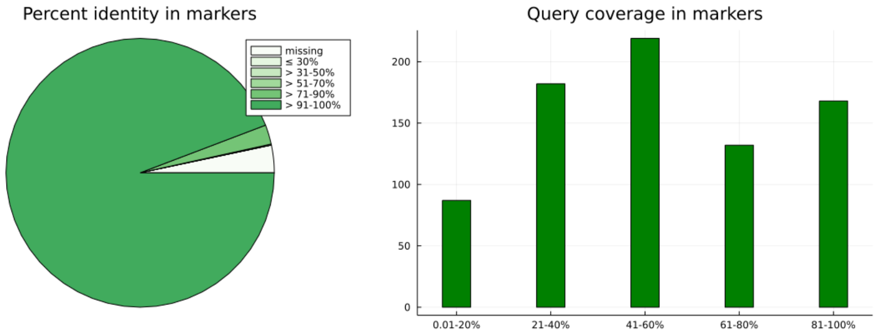


Figure 4.3: Percent identity- and query coverage in markers based on a Patchwork analysis of a SPAdes assembly of *Dimorphilus gyrociliatus*, targetting 815 single-copy orthologs from itself.

Table 4.4: Results from Patchwork, using *Dimorphilus gyrociliatus* SPAdes assembly as a query and *Alitta virens* USCOs as a reference, and then re-aligning the resulting query sequences against USCOs from a long-read assembly of *Dimorphilus gyrociliatus*.

variable	mean	min	median	max
reference_len	448.65	77	358.0	2748
query_len	232.62	25	172.0	2169
no_matches	197.82	25	150.5	2169
no_mismatches	28.90	0	2.0	601
no_insertions	2.15	0	0.0	162
no_deletions	15.63	0	4.0	355
distance	970.73	127	745.0	11612
query_cover	56.88	3.29	58.925	100.0
percent_identity	89.92	33.59	97.905	100.0

4.7 Conclusions

Patchwork is a new software for quickly mining phylogenetic markers from WGS data. Since Patchwork can retrieve homologous regions even in distantly related taxa, this program lends itself especially well for recovering phylogenetic markers for phylogenomic studies, where high-quality transcriptomes are not available for all species of interest. It is simultaneously an efficient way for increasing marker occupancy in poorly assembled genomes and or in the presence of multi-locus exons. Finally, Patchwork allows the user to combine two different data types—transcriptomic and genomic data—into a single dataset, thus further enabling an even larger taxon sampling and encouraging data reusability.

Special consideration should be taken to avoid the creation of chimeric sequences. One way in which such sequences may arise is when orthologous (i.e., genes related via a speciation event) and paralogous (i.e., genes related via a gene duplication event)

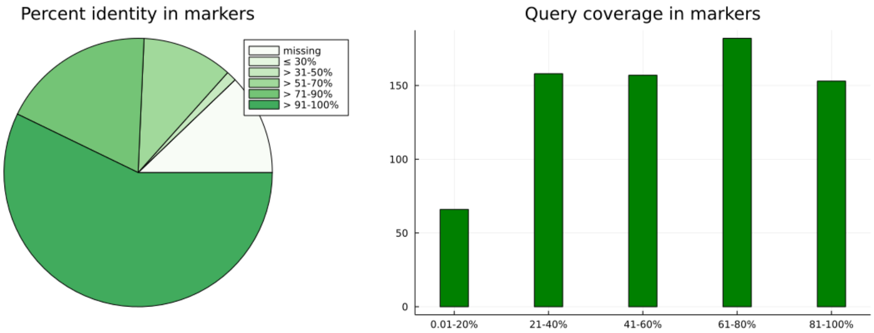


Figure 4.4: Percent identity- and query coverage in markers based on a Patchwork analysis of a SPAdes assembly of *Dimorphilus gyrociliatus*, targetting a set of 897 near-universal single-copy orthologs (USCOs) from the distantly-related *Alitta virens*. Results shown here are when realigning the recovered USCOs against a set of 826 single-copy orthologs from *D. gyrociliatus* itself.

sequences are merged together. To circumvent this issue, we recommend that the user limit the use of reference sequences to near-universal single-copy orthologs (USCOs). Many programs—e.g., the forementioned program BUSCO—exists for retrieving such sequences from an already assembled genome and these could be used as reference sequences.

The accuracy and the robustness of the results depends on how closely related the two species under study are. The difficulty stems from the ability to accurately predict non-coding regions in aligned contigs; because alignment-trimming relies on gap-excluded identity, choosing the correct cutoff threshold gets increasingly easier as the level of identity approaches 100% (the identity of non-coding regions is likely to stay the same, while the the identity to coding-regions increases). On the upside, high-quality genomes for practically all major lineages exists and are readily available online.

4.8 Back matter

Data availability

The supplementary data are available at github.com/Animal-Evolution-and-Biodiversity/benchmarking-patchwork.

Code availability

The source code of Patchwork is available at GitHub (https://github.com/fethalen/patchwork) under the GPLv3 license.

Acknowledgments

This work was supported by the German Research Foundation (DFG) BL787/8-1 “Phylogenomic analysis of Nereididae (Annelida)”.

Competing interests

The authors declare no competing interests.

Author's contributions

F.T. and C.B. conceived of the original design of the software and the study itself. C.K. and F.T. made the software implementation together. F.T. took the lead in writing the manuscript with contributions from C.B. All authors have read and approved of its final version.

5 Phylogenomic analysis of Nereididae using a new genome-skimming approach

Felix Thalén[1,2*], Thilo Schulze[1], Frederick Hahner[1], Clara G. Köhne[1], Katharina Henze[1], Christopher J. Glasby[7], Conrad Helm[1], Daniel Martin[8], Dazuo Yang[12], Detlev Arendt[3], Dinesh Kaippilly[11], Jithin Kothalil[11], Kevin N. Mutemi[3], M. Teresa Aguado[1], Maria Capa[5], Robin Wilson[10], Teresa Darbyshire[6], Tobias Gerber[3], Torkild Bakken[4], Tulio Villalobos[9], and Christoph Bleidorn[1]

[1]Dept. for Animal Evolution and Biodiversity, Georg-August-Universität Göttingen, Untere Karspüle 2, 37073 Göttingen, Germany

[2]Cardio-CARE AG, Medizincampus Davos, Herman Burchard Str. 1, 7265 Davos, Switzerland

[3]Developmental Biology Unit, EMBL Heidelberg, Meyerhofstraße 1, Heidelberg, 69117, Baden-Württemberg, Germany

[4]Norwegian University of Science and Technology, NTNU University Museum, Trondheim, 7012, Trøndelag, Norway

[5]Biology Department, University of the Balearic Islands, Carretera de Valldemossa, Palma, 07122, Balearic Islands, Spain

[6]Department of Biodiversity and Systematic Biology, National Museum of Wales, Cardiff, CF10 3NP, Wales, United Kingdom

[7]Museum Art Gallery of the Northern Territory, GPO Box 4646, Darwin, GPO Box 46460801, Northern Territory, Australia

[8]Centre d'Estudis Avançats de Blanes (CEAB-CSIC), carrer d'accés a la Cala Sant Francesc 14, 17300 Blanes (Girona), Catalunya, Spain

[9]Depto. Sistemática y Ecología Acuática, El Colegio de la Frontera Sur, Chetumal,

Quintana Roo, 77000, Mexico

[9]Depto. Sistemática y Ecología Acuática, El Colegio de la Frontera Sur, Chetumal, Quintana Roo, 77000, Mexico

[10]Museum Victoria, GPO Box 666 Melbourne, Vic. 3001, Australia

[11]Department of Aquaculture, Kerala University of Fisheries and Ocean Studies (KUFOS), Panangad Road, Kochi, 682506, Kerala, India

[12]Key laboratory of Marine Bio-resource Restoration and Habitat Reparation in Liaoning province, Dalian Ocean University, Dalian 116023, PR China

[*]Corresponding author: felix.thalen@cardio-care.ch

5.1 Abstract

Nereididae Blainville, 1818 constitute a family of marine and freshwater-living worms within Annelida with over 770 species of nereidids described worldwide. Their overall abundance, wide distrubtion, and ease of culturing, makes some nereidids suitable as laboratory species. Notably, *Platynereis dumerilii* has continuously been bred in laboratories for more than 70 years and is now one of the most well-studied species within Lophotrochozoa (also known as Spiralia). Despite this, interspecies relationships—and the relationships between currently accepted subfamilies—within this group are poorly resolved. Prior phylogenetic studies have limited taxonomic coverage and are based on morphological characters or a handful of mitochondrial markers. Here, we perform a phylogenomic study using low-coverage whole-genome sequencing data from 100 individuals from 78 species in 24 genera, covering 4 different subfamilies. Near-universal single-copy orthologs (USCOs) were obtained using a new, alignment-based method, making this the first phylogenomic study to apply this approach. We additionally analyzed mitochondrial genomes from 132 individuals, by combining our own data with publically available sequences. Our study implies that Gymnonereididae sensu Banse 1977 is the sister group to all other nereidids. Furthermore, larger genera such as *Alitta*, *Nereis*, and *Neanthes* are recovered as non-monophyletic. Analyses of mitochondrial genes and mitochondrial gene order in nereidids provide additional consistent support to our findings.

Supplementary information: Supplementary data are available from GitHub at github.com/Animal-Evolution-and-Biodiversity/nereididae-phylogenomic.

Keywords: Genome skimming, low-coverage whole-genome sequencing, Nereididae, ragworm, Annelida, phylogenomics, mitochondrial genes, gene order

5.2 Introduction

Nereididae Blainville, 1818, commonly known as ragworms, constitute a family of marine and freshwater-living species within Annelida, containing more than 770 valid species and 43 accepted genera (Read & Fauchald, 2021). Their worldwide distribution, overall abundance, and presence in easy-to-access intertidal zones has made them commercially and ecologically important (Bakken et al., 2018). Indeed, some members in this group have been used as fishing baits, as a food complement in aquacultures, and as a subject for laboratory studies (Simon et al., 2021). Notably, *Platynereis dumerilii* has continuously been bred in the laboratory since 1953 (A. Fischer & Dorresteijn, 2004) and has been established as a model for studying reproduction, regeneration, development, evolution, chronobiology, neurobiology, ecology, ecotoxicology, as well as and single-cell genomics (Özpolat et al., 2021). Nereidids, together with molluscs, brachiopods, bryozoans, acanthocephalans, as well as platyhelminths, are members of Lophotrochozoa (also called Spiralia), one of the three major branches of bilaterians (with the other two being Deuterostomia and Ecdysozoa; [Halanych, 2004]), thus making *P. dumerilii* one of the most prominent and widely studied model organisms within this group. Moreover, nereidids have often been used to represent Annelida and "Polychaeta" as a whole. Unfortunately, the evolutionary relationships among nereidids remain poorly understood. Pre-existing phylogenetic studies, based on morphological data (Bakken & Wilson, 2005; Fitzhugh, 1987; C. Glasby, 1991; Santos et al., 2006), are incongruent and suffer from poor taxonomic coverage. Similarly, prior molecular studies of the group (Alves et al., 2020a; Liu et al., 2013; Tosuji et al., 2019; Villalobos-Guerrero et al., 2022) have poor resolution due to limited taxon sampling and in that they only examine a handful of mitochondrial genes. Furthermore, the monophyly of species-rich genera—such as *Neanthes*, *Nereis*, and *Perinereis*—has been disputed (Bakken & Wilson, 2005). Indeed, two recent molecular analyses of *Neanthes* (Drennan et al., 2021; Villalobos-Guerrero et al., 2022) has further reinforced the idea of *Neanthes* as a convoluted genus, when members of this group were recovered as sister taxa or taxon to other genera such as *Alitta*, *Cheilonereis*, *Dendronereis* and *Nectoneanthes*. Within that same study, Villalobos-Guerrero et al., 2022 remarks on the issue with identification uncertainty and misidentification of many species and how only slightly above 10% of all nereidids have their COI sequence available in the Barcode of Life Database or GenBank. On top of this, cryptic diversity observed in some members of this group (Kara et al., 2020; Virgilio et al., 2009) further highlights the need for a joint effort from taxonomists and molecular biologists to publish COI sequences along with species identifications, in order to lay the groundwork of a more accurate species identification framework on top of which more precise phylogenetic analyses can be made. Although Nereididae as a whole is well-supported and considered monophyletic (C. J. Glasby, 1993), the relationships within Nereididae are still disputed. Five subfamilies have been proposed, based on morphological data. They are: Nereidinae Blainville 1818, Namanereidinae Hartman 1959, Dendronereidinae Pillai 1961, Notophycinae Knox and Cameron 1970, and Gymnonereidinae Banse, 1977. Notably, two subfamilies, Dendronereinae Pillai, 1961 and Notophycinae Knox and Cameron, 1970, are still under *incertae sedis* (Read et al., 2012).

Liu et al., 2013 conducted the first molecular phylogenetic study of Nereididae based on cytochrome c oxidase subunit I (COI) sequences from 21 species and eight genera. Despite a relatively small taxon sampling, their study recovered several larger genera (i.e., Platynereis, Perinereis, and Nereis) as paraphyletic. Subsequently, Alves et al., 2020a analysed nucleotide and amino acid sequences of whole mitochondrial genomes in 20 species. Interestingly, they found that—like with most other annelids (Jennings & Halanych, 2005; Weigert et al., 2016; Zhong et al., 2008)—protein-coding gene order is highly preserved within nereidis and and that tRNA rearrangements can be phylogenetically informative. Park et al., 2016 observed two arrangements and divided them into group 1 and 2 respectively; group one has the order tRNA-Tyr, ATP8, tRNA-Met and tRNA-Asp while the second group has tRNA-Met, tRNA-Asp, ATP8 and tRNA-Tyr. Because phylogenetic studies using one or a handful of genes often yield poor support and contradictory results, however, phylogenomic approaches—i.e., phylogenetic analyses done using hundreds or thousand of markers—are now the gold standard (Bleidorn, 2017; E. M. Lemmon & Lemmon, 2013) and performed on routine at an almost industrial scale. Phylogenomic analyses of annelids, focusing on the backbone of the tree, have been performed primarily using transcriptomic data (Andrade et al., 2015a; Laumer et al., 2015; Struck, 2011; Struck et al., 2015). Unfortunately, RNA sequencing (RNA-Seq) requires freshly-collected material, preserved in RNAlater (Allen, Boyd, Nguyen, et al., 2017). Another cost-effective approach to large-scale phylogenetics is targetted enrichment, in which only a select portion of the genome is sequenced (E. M. Lemmon & Lemmon, 2013). However, this approach requires that oligonucleotide probes (or baits) are designed and the difficulty lies in designing probes that capture across the entirety of the taxon sampling (Bragg et al., 2016b; Hawkins et al., 2016), thus rendering this strategy less than ideal for nereidids, which are estimated to have diverged from the rest of the annelids around 350 m.y.a (Helsem, 2021). As sequencing costs decreases (currently estimated to be 0.01 US$ per one Mbp of raw DNA sequences; https://www.genome.gov/sequencingcostsdata), however, we are approaching a threshold where whole-genome sequencing (WGS; i.e., sequencing an organism's genome and it's organelles in their entirety) is the most reasonable approach (E. M. Lemmon & Lemmon, 2013). G. Ribeiro et al., 2021 found that an average coverage of 5–10x can be sufficient for inferring interspecies relationships. Previously hampered by both the increased sequencing costs and additional computation processing (F. Zhang et al., 2019b), lower costs and new bioinformatic tools for handling such data (e.g., aTRAM Allen, Boyd, Nguyen, et al., 2017; ALiBaSeq Knyshov et al., 2021b; and Patchwork Thalen et al., 2022) has led to renewed interest in this space. Although phylogenomic approaches are now commonplace, no such study have previously been conducted on Nereididae.

In the present study, we carried out the largest phylogenetic study of Nereididae, based on molecular markers, to date. Our nuclear gene dataset includes a combination of transcriptomic and genomic data from a total of 100 individuals (96 ingroups and 4 outgroups) and 78 nereidid species across 24 genera, with representatives from four different subfamilies. We have included transcriptome data from eight ingroup taxa and five outgroup taxa from four different families and all members of the suborder Neridiformia: *Amphiduros fuscescens* (Hesionidae), *Chrysopetalum occidentale* (Chrysopetalidae), *Eulalia viridis* (Phyllodocidae), *Nephytis caeca* (Nephtyidae), and *Phyllodoce medipapillata* (Phyllodocidae). Our choice of outgroup taxa (i.e., Phyllodocida, Hesionidae, Chrysopetalidae) was informed by the phylogenetic hypothesis that nereidids are members of the order Phyllodocida and within this group, they are probably the sister-group to Chrysopetalidae and or Hesionidae (Dahlgren et al., 2000; C. J. Glasby, 1993; Pleijel & Dahlgren, 1998). Additionally, we retrieved and annotated mitochondrial genomes for all species within the fore-mentioned dataset and combined them with publically available sequences. Our mitochondrial gene dataset targets 15 mitochondrial genes from 132 individuals (7 outgroups and 125 ingroup [nereidid] individuals), from 79 nereidid species in 24 genera and 4 subfamilies.

We utilize our newly developed program, Patchwork (Thalen et al., 2022), to perform alignment-based phylogenetic marker retrieval from whole-genome sequencing WGS data. In detail, we used a set of near-universal single-copy orthologs (USCOs) in the newly published *Alitta virens* genome as inferred by BUSCO v5.0.0 (Simão et al., 2015). With this dataset, we aim to (i) investigate the validity and relationship between the currently established subfamilies, (ii) examine polyphyly within the most species-rich genera (i.e., *Neanthes*, *Nereis*, and *Perinereis*), and (iii) demonstrate the utility of genome skimming—and more specifically our newly developed method—as an effective approach to phylogenomic studies.

5.3 Methods

5.3.1 Sample collection

The taxonomic sampling for this study is the result of an international, multi-cohort assemblage of researchers whom collected the material anew and or provided pre-identified samples from various museums or other research collections. Notably, these samples vary in collection year, preservation method (most samples were stored in ethanol but at different concentration), and storage conditions. Some samples are as old as 1985, although most samples were collected at a much more recent data. Exact information for each sample is available in the supplementary material. Samples collected for this study specifically were obtained between 2019 and 2020 in Roscoff, France, and Ferrol, Spain, from sediment in-between intertidal rocks at low tide or from the benthic by using a dredge on a small research vessel. Specimens were examined under a stereo microscope and identified, based on morphological characters, according to the the 25th Volume of the *Fauna Iberica* (Viéitez, 2004), the Handbook of the marine fauna of North-West Europe (Hayward & Ryland, 2017), and individual species descriptions. Samples used for DNA sequencing were preserved in 99.9% ethanol, while those used for RNA-sequencing were stored in cryotubes with RNAlater at −80 °C.

5.3.2 Transcriptome sequencing and assembly

We generated novel RNA-seq data for seven taxa: *Alitta virens*, *Hediste diversicolor*, *Namalycastis abiuma*, *Perinereis cultrifera*, *Ceratonereis australis*, and *Neanthes nubila*. The rest of the eight transcriptomes used for this study were obtained from the NCBI Sequence Read Archive (SRA). Both novel samples and raw Illumina data taken from NCBI SRA are listed in Table 5.4.1.

Total RNA was isolated using the standard TRIzol™ LS Reagent and subsequently purified using the QIAGEN RNeasy MinElute Cleanup Kit. Purified samples were sent to Eurofins Genomics in Germany for library preparation and sequencing. Each library was sequenced using Illumina 2 X 150 bp paired-end chemistry.

At this point, novel samples and publicly available Illumina sequences were treated the same: Paired-end Illumina reads were assessed for their quality using FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/). After quality control, Trim Galore! v0.4.1 (https://www.bioinformatics.babraham.ac.uk/) was used for quality and adapter trimming with a Phred quality threshold of 20 and a minimum required length of 55 base pairs. After trimming, the quality of the resulting, trimmed sequences was checked again using FastQC v0.11.9. Finally, trimmed Illumina reads were assembled using the RNA-Seq *de novo* assembler Trinity v2.11.0 (Grabherr et al., 2011) under default settings.

For each assembly, we ran BUSCO v.5.3.1 (Simão et al., 2015) to search for universal single-copy orthologs (USCOs) and to assess the quality of the assemblies generated. BUSCO was ran using the "Metazoa" lineage, with AUGUSTUS v3.4.0 (Stanke et al., 2006; Stanke & Morgenstern, 2005) as a gene predictor under default settings, but with the self-training mode set to on. The number of amino acid (AA) characters in the resulting dataset was calculated using the `stats` module of the SeqKit toolkit for working with FASTA and FASTQ files (W. Shen et al., 2016).

5.3.3 Genome sequencing and assembly

We performed DNA extraction on a total of 130 samples, of which 100 ended up in the final tree. The species used were selected based on their assumed phylogenetic position (i.e., according to the literature), the quality of the DNA extraction, or the quality of the assembly. Individuals from species which had already been sequenced were down prioritized and *vice versa*.

Prior to DNA extraction, each sample was placed, with an open lid, inside a Eppendorf ThermoMixer at 45°C, in order to evaporate the EtOH. When still present, a lysis buffer containing Proteinase K was added to each tube and was left over night. The elution buffer was heated to 65°C and 35 or 50 µL of the buffer was used to elute the DNA, depending on the size of the specimen. DNA was extracted from single individuals, using the Quick-DNA Miniprep Plus Kit (Zymo Research), while following the manufacturer's protocol for extrating DNA from solid tissue. After DNA extraction, DNA concentration was checked using a Qubit Fluorometric (ThermoFisher), utilizing the Qubit dsDNA BR (Broad-Range) Assay Kit and while following the manufacturer's protocol. Samples where the amount of DNA extracted was less then 100 ng were not sequenced. However, for species of particular interest, DNA amplification was performed using the REPLI-g Mini Kit from Qiagen, following the manufacturer's instruction for 2,5 µL template DNA.

All sequencing was performed using an Illumina NovaSeq 2000 or 6000 system (150 bp paired-end) with at least 5 M read pairs per nereidid library. Known sizes of nereidids genomes ranges from 782–2,454 Mbp (Gregory, 2002). Samples collected in China and India were sequenced on-site, while the majority of the sequencing was carried out by Eurofins Genomics (Konstanz, Germany).

Paired-end reads were assessed for quality using FastQC v0.11.9 (https://www.bioinformatics.babraham.ac.uk/). and subsequently, adapter trimming was done using Trim Galore! v0.6.6 (https://www.bioinformatics.babraham.ac.uk/). For adapter trimming, we used a Phred quality score threshold of 20 and reads below 50 residues in length were discarded. *De novo* assembly was performed using SPAdes v3.15.3 (Nurk et al., 2013) and a K-mer size setting of 33, 55, or 91. The majority of assemblies were done using a K-mer size of 33, with the goal of generating as much data as possible rather than aiming at the best possible contig size. In our own tests, we found that Patchwork performs best on this type of data, since fragments are anyways stitched together. Assembly quality assessment was done using QUAST v5.0.2 (Gurevich et al., 2013). The assembly pipeline used (https://github.com/ThiloSchulze/eukaryotic-genome-assembly) was run on the local scientific compute cluster of the GWDG at the University of Göttingen, Germany.

5.3.4 Orthology assignment, marker discovery, and supermatrix construction

A chromosome-level assembly of the nereidid *Alitta virens*, published and processed by the Wellcome Sanger Institute, was obtained from GenBank (GenBank assembly accession: GCA_932294295.1) and used as a reference. We used BUSCO v.5.3.1 (Simão et al., 2015) to search the fore-mentioned assembly for universal single-copy orthologs (USCOs), by utilizing the "Metazoa" lineage. A total of 777 USCOs were were detected this way and we used these to obtain USCOs from the other assemblies. Depending on whether they were transcriptome assemblies or genomic assemblies, this was done in two different ways: (i) For our transcriptome assemblies, we re-ran BUSCO, again using the Metazoa dataset, and then kept all USCOs that were also found in our reference dataset. The BUSCO results for each transcriptome assembly is summarized in table 5.1. (ii) USCOs from the genome were obtained using Patchwork v1.0.2 (Thalen et al., 2022), using the *A. virens* USCOs as a reference, a sliding window size of 4 bps, an average distance cutoff of -7, an E-value threshold at 1^{-3}, and while running DIAMOND with the `-iterate` flag. Because the novelty of this software, we've included a short description of how it operates in section 5.3.4.

Overview of Patchwork

Patchwork is a new, alignment-based program for mining phylogenetic markers from genomic data. It uses translated nucleotide sequences to search against the provided set of reference sequences (the *A. virens* USCOs, in our case). Overlapping hits are merged by retaining the best-scoring sequence for each position. Finally, alignment masking and sliding window-based alignment trimming are both applied in order to rid the resulting sequence from putative, non-coding regions.

Alignment trimming

To rid the alignments of poorly aligned residues and to simultaneously speed up the downstream analysis, we used the alignment trimming software GUIDANCE v2.02 (Sela et al., 2015). GUIDANCE alternates between different guide trees to identify unreliable positions in the resulting alignments. For this, we used a sequence cutoff value of 0.6 and a column cutoff value of 0.93. This means that sequences with a GUIDANCE score below 60% were removed and columns below 93% were also removed. MAFFT v7.487 (Katoh & Standley, 2013) was the alignment program that was used within GUIDANCE and the alignment algorithm was automatically selected for each marker, using the `-auto` option.

Supermatrix construction

Next, we utilized the tree-based orthology program PhyloPyPruner v1.2.6 (github.com/fethalen/phylopypruner) to (i) get a statistical overview of the sequence composition and to (ii) construct the supermatrix and the partition file. Although PhyloPyPruner can perform tree-based orthology inference, this aspect of the program was not utilized. Nonetheless, gene trees for each input alignment were inferred, as this is required by the program. It's worth pointing out that these trees have no impact whatsoever on the supermatrix construction, since all of the input alignments were already 1:1 orthologs. Gene trees were constructed using FastTree v2.1.10 (Price et al., 2010), using the `-slow` and `-gamma` options. The resulting gene trees and their corresponding alignments were used as an input for PhyloPyPruner. Using PhyloPyPruner, we removed 5 taxa that were represented in less than 70% of the alignments.

5.3.5 Mitochondrial genome assembly and annotation

We developed a custom pipeline (github.com/ThiloSchulze/mitogenome-extraction), written in the workflow management system Nextflow (Di Tommaso et al., 2017), to extract mitochondrial genomes from our genome assemblies. Our pipeline uses BLASTN (v2.5.0; Altschul et al., 1990) to retrieve the best-matching contig to a user-provided reference. For our purposes, we selected a word-size from 11–25 and used the mitochondrial genome of *Platynereis dumerilii* (Genbank accession AF178678; Boore and Brown, 2000). This pipeline also uses an iterative coverage filter where, at the first stage, only contigs with 60x coverage and up are kept. If no matching contig is found, the coverage threshold is lowered to 10x and the BLASTing step is repeated. If multiple contigs were found, mitochondrial nucleotide data was reassembled using using NOVOPlasty (v4.3.1; Dierckxsens et al., 2017). If reassembly using NOVOPlasty failed, a reassembly using PRICE (v1.0.1 Ruby et al., 2013) was attempted. In case the best match was located on the opposite strand, a reverse complement was obtained using EMBOSS (v6.6.0; Rice et al., 2000). We obtained the 650 bp Cytochrome c oxidase I (COX1) subunit for barcoding by BLASTing against a reference from *P. dumerilii* (accession no. KR916915). Once a candidate contig near our specified target-size (i.e., 16500 bp, the average length of metazoan mitogenomes Bernt et al., 2013) had been found, we performed *de novo* mitogenome annotation using MITOS (v2.0.8; Bernt et al., 2013). Extracted barcode sequences were searched against The Barcode of Life Data System (BOLD; http://www.barcodinglife.org; Ratnasingham and Hebert, 2007) and NCBI GenBank Benson et al., 1993 for the best match, in order to re-verify species identification.

5.3.6 Phylogenetic analysis

Maximum likelihood analysis of the 777 single-copy orthologs in the data matrix was performed using IQ-TREE v2.1.2 (Minh et al., 2020) under the best-fitting substitution model for each partition, as determined by ModelFinder Plus (Kalyaanamoorthy et al., 2017) and by using 1000 ultrafast bootstrap replicates (Hoang et al., 2018).

We additionally used single gene trees, generated from the same data matrix to perform a coalescent-based species tree estimation using ASTRAL-III v.5.7.1 (C. Zhang et al., 2018) and 1000 bootstrap replicates.

The final trees were visualized using the Interactive Tree Of Life (iTOL; Letunic and Bork, 2019) and subsequently annotated using Adobe Illustrator CS6.

A workflow of the bioinformatic pipeline used in this phylogenetic analysis is shown in figure 5.1.

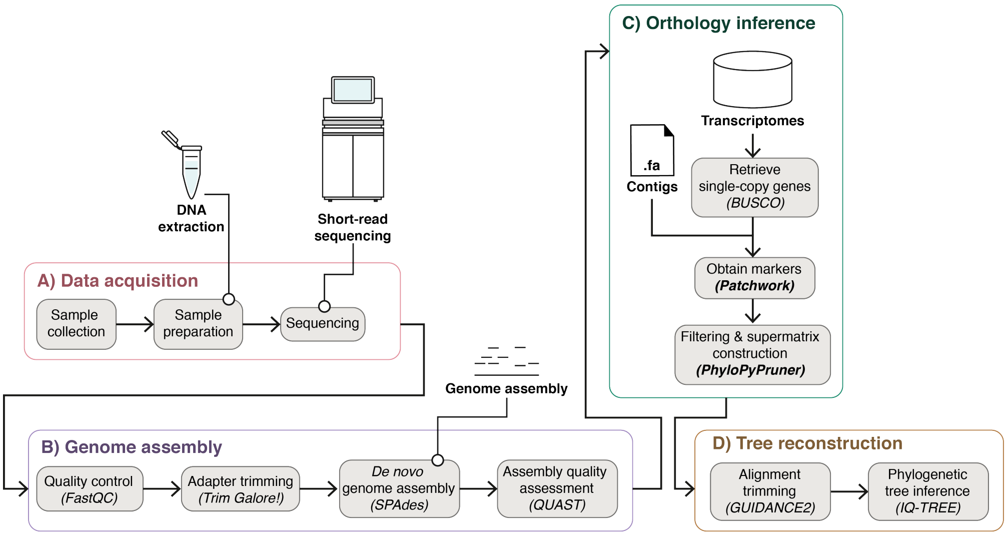

C) Orthology inference
Transcriptomes
.fa
Contigs
Retrieve single-copy genes (BUSCO)
Obtain markers (Patchwork)
Filtering & supermatrix construction (PhyloPyPruner)
DNA extraction
Short-read sequencing
A) Data acquisition
Sample collection
Sample preparation
Sequencing
Genome assembly
B) Genome assembly
Quality control (FastQC)
Adapter trimming (Trim Galore!)
De novo genome assembly (SPAdes)
Assembly quality assessment (QUAST)
D) Tree reconstruction
Alignment trimming (GUIDANCE2)
Phylogenetic tree inference (IQ-TREE)

Figure 5.1: Schematic workflow of the phylogenomic analyses carried out in this study.

5.4 Results

5.4.1 Transcriptome sequencing and assembly

Our transcriptome analysis resulted in a total of 13 assemblies, whereas two of these were subsequently removed due to the amount of missing data exceeding our 30% threshold. Each species, together with the number of USCOs, AAs, and the accession number (where applicable), is shown in table 5.1. Removed taxa have been highlighted in red.

5.4.2 Genome sequencing and assembly

We assembled paired-ended Illumina whole-genome sequencing data from a total of 105 nereidid species, with represents from four different subfamilies: Nereidinae, Dendronereidinae, Gymnonereidinae, and Namanereidinae. Assembly statistics, showing

Table 5.1: List of universal single-copy orthologs (USCOs) obtained from the *transcriptome* sequence used for this study. The number of orthologs corresponds to the total number of USCOs that were also found in the reference used (*Alitta virens*).

Ingroup			
Taxon	**No. orthologs (%)**	**AA positions**	**Accession no.**
Alitta virens	897 (100)	302,598	GCA_932294295.1
Alitta succinea	316 (35)	92,468	SRS954182
Ceratonereis australis	334 (37)	223,105	This study
Hediste diversicolor	474 (52)	128,130	This study
Namalycastis abiuma	533 (59)	169,923	This study
Neanthes nubila	310 (35)	71,787	This study
Perinereis aibuhitensis	527 (59)	189,356	In prep.
Platynereis dumerilii	549 (61)	205,095	In prep.
Outgroup			
Taxon	**No. orthologs (%)**	**AA positions**	**Accession no.**
Amphiduros fuscescens	315 (35)	114,121	SRR15277965
Chrysopetalum occidentale	526 (59)	189,952	SRR15277964
Eulalia viridis	393 (44)	136,286	SRS6018155
Nephytis caeca	487 (54)	161,758	SRS591169
Phyllodoce medipapillata	258 (29)	87,087	SRS933586

the K-mer size, number of assembled contigs, the mean contig length, and the N50 of each assembly is shown in the tables 5.3 and 5.5. In summary, our genome assemblies had an average assembled contig length of 956,470, an average total length of 563,323,652 amino acid positions, and an average N50 of 1,355.

5.4.3 Orthology assignment, marker discovery, and supermatrix construction

Our data matrix consisted of 777 different universal single-copy orthologs (USCOs) and a total of 104 operational taxonomic units (OTUs), with 99 OTUs per alignment on average and a mean sequence length of 99 amino acids (AAs). 39.1% of the positions were missing and the total concatenated supermatrix length was 312,912 AAs. A heatmap depicting the completeness of each marker is shown in figure 5.2.

5.4.4 Gene order arrangements

We observe, similar to Alves et al., 2020a, that—like with most annelids—protein-coding and rRNA gene order information within nereidids remain highly conserved and that tRNA arrangements can provide phylogenetic information. In addition to the two distinct tRNA gene order arrangements described in Park et al., 2016, we observe four

Figure 5.2: Two occupancy matrix partitions (representing two halves of the data matrix, split into chunks of 500 genes per chunk), produced by PhyloPyPruner v2.1.6 (Thalén et al. in prep.), depicting the completeness (number of positions in an alignment, gaps or complete absence are treated as missing data) of each gene on the X-axis and each taxon on the Y-axis. 100% completeness is displayed in dark green, while fully missing genes from that taxa are displayed in white. (A) Occupancy matrix of the first 500 markers and (B) occupancy matrix of the 395 others. Markers marked in red were discarded as their occupancy fell below the 30% occupancy threshold.

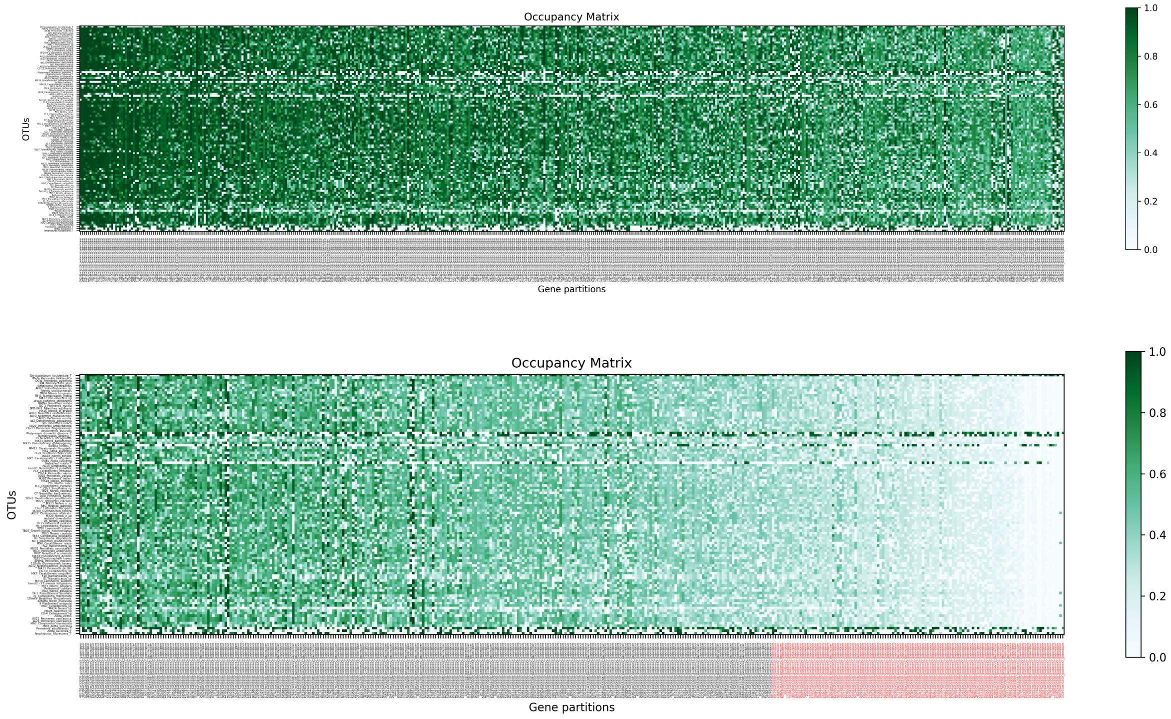
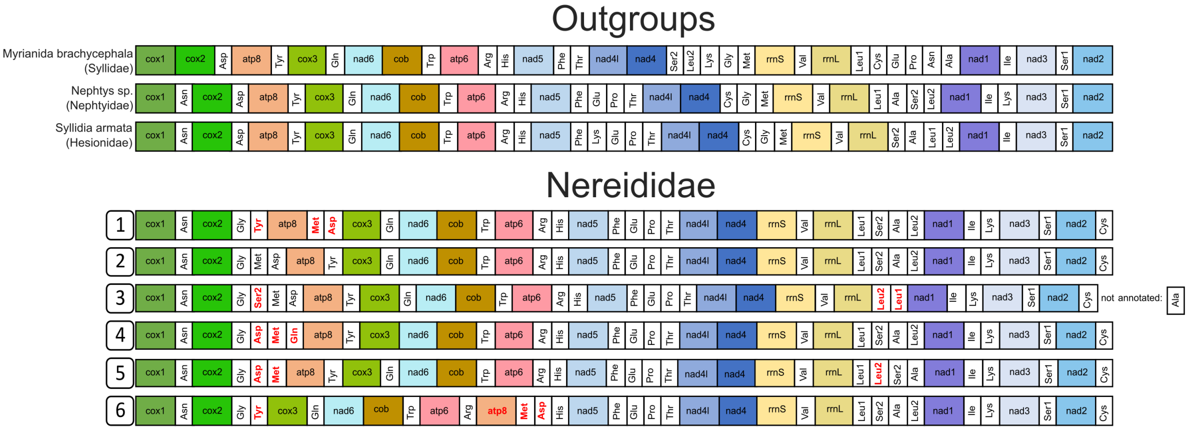


Figure 5.3: Six observed mitochondrial gene orders for Nereididae and three arrangements found in the outgroups. Their occurence have been annotated in the mitochondrial gene tree (see figure 5.5). With the assumption that the tRNA arrangements in group 2 are ancestral, deviations from this arrangement are highlighted in red.


Outgroups
Myrianida brachycephala (Syllidae) cox1 cox2 Asp atp8 Tyr cox3 Gln nad6 cob Trp atp6 Arg His nad5 Phe Thr nad4l nad4 Ser2 Leu2 Lys Gly Met rrnS Val rrnL Leu1 Cys Glu Pro Asn Ala nad1 Ile nad3 Ser1 nad2
Nephtys sp. (Nephtyidae) cox1 Asn cox2 Asp atp8 Tyr cox3 Gln nad6 cob Trp atp6 Arg His nad5 Phe Glu Pro Thr nad4l nad4 Cys Gly Met rrnS Val rrnL Leu1 Ala Ser2 Leu2 nad1 Ile Lys nad3 Ser1 nad2
Syllidia armata (Hesionidae) cox1 Asn cox2 Asp atp8 Tyr cox3 Gln nad6 cob Trp atp6 Arg His nad5 Phe Lys Glu Pro Thr nad4l nad4 Cys Gly Met rrnS Val rrnL Ser2 Ala Leu1 Leu2 nad1 Ile nad3 Ser1 nad2
Nereididae
1 cox1 Asn cox2 Gly Tyr atp8 Met Asp cox3 Gln nad6 cob Trp atp6 Arg His nad5 Phe Glu Pro Thr nad4l nad4 rrnS Val rrnL Leu1 Ser2 Ala Leu2 nad1 Ile Lys nad3 Ser1 nad2 Cys
2 cox1 Asn cox2 Gly Met Asp atp8 Tyr cox3 Gln nad6 cob Trp atp6 Arg His nad5 Phe Glu Pro Thr nad4l nad4 rrnS Val rrnL Leu1 Ser2 Ala Leu2 nad1 Ile Lys nad3 Ser1 nad2 Cys
3 cox1 Asn cox2 Gly Ser2 Met Asp atp8 Tyr cox3 Gln nad6 cob Trp atp6 Arg His nad5 Phe Glu Pro Thr nad4l nad4 rrnS Val rrnL Leu2 Leu1 nad1 Ile Lys nad3 Ser1 nad2 Cys not annotated: Ala
4 cox1 Asn cox2 Gly Asp Met Gln atp8 Tyr cox3 nad6 cob Trp atp6 Arg His nad5 Phe Glu Pro Thr nad4l nad4 rrnS Val rrnL Leu1 Ser2 Ala Leu2 nad1 Ile Lys nad3 Ser1 nad2 Cys
5 cox1 Asn cox2 Gly Asp Met atp8 Tyr cox3 Gln nad6 cob Trp atp6 Arg His nad5 Phe Glu Pro Thr nad4l nad4 rrnS Val rrnL Leu1 Leu2 Ser2 Ala nad1 Ile Lys nad3 Ser1 nad2 Cys
6 cox1 Asn cox2 Gly Tyr cox3 Gln nad6 cob Trp atp6 Arg atp8 Met Asp His nad5 Phe Glu Pro Thr nad4l nad4 rrnS Val rrnL Leu1 Ser2 Ala Leu2 nad1 Ile Lys nad3 Ser1 nad2 Cys

additional, previously undescribed, arrangements, including one protein-coding gene translocation in one species Three out of four newly observed groups belong to a new, hypothetical clade containing the genera *Paraleonnates*, *Solomonereis*, and *Ceratonereis*. Notably, these arrangements only occur in single species and are not synapomorphic for this group, and the other members of this clade have the "ancestral" state (group 2). The last arrangement can be found in *Nereis panamensis*, which, in our mitochondrial gene tree, is the sister group to all other members of the subfamily Nereidinae.

5.4.5 Phylogenetic analysis

In all, we generated three different phylogenetic trees: one nuclear-gene tree (figure 5.4), one mitochondrial-gene tree (figure 5.5), and one coalescent-based species tree (figure 5.6). Our analyses consistently supports the monophyly of Nereididae and while we observe discrepancies—especially in the coalescent-based species tree—among the three, the overarching pattern remains the same. E.g., Gymnonereididae, together with Namanereidinae and Dendronereidinae as well as the genus *Tylorrhynchus* are consistently recovered as the sister group to the rest of the nereidids within our study. Furthermore, we find in all trees support for grouping the genera *Paraleonnates*, *Solomonereis*, and *Ceratonereis* together into a new, hypothetical group. Finally, we consistently observe non-monophyly within larger genera such as *Alitta*, *Nereis*, and *Neanthes*.

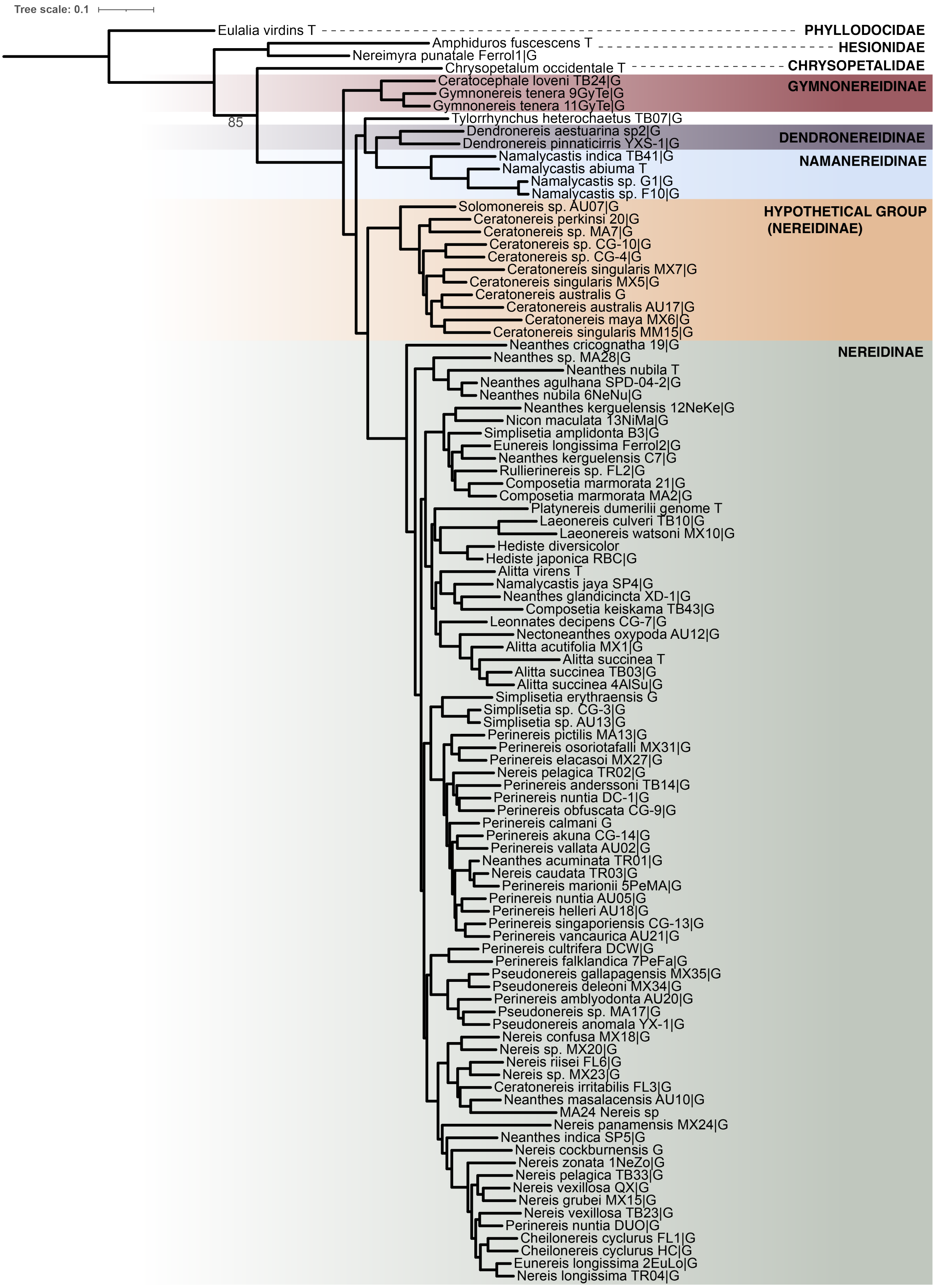

Tree scale: 0.1
Eulalia virdins T
PHYLLODOCIDAE
Amphiduros fuscescens T
HESIONIDAE
Nereimyra punatale Ferrol1|G
Chrysopetalum occidentale T
CHRYSOPETALIDAE
Ceratocephale loveni TB24|G
GYMNONEREIDINAE
Gymnonereis tenera 9GyTe|G
Gymnonereis tenera 11GyTe|G
85
Tylorrhynchus heterochaetus TB07|G
Dendronereis aestuarina sp2|G
DENDRONEREIDINAE
Dendronereis pinnaticirris YXS-1|G
Namalycastis indica TB41|G
NAMANEREIDINAE
Namalycastis abiuma T
Namalycastis sp. G1|G
Namalycastis sp. F10|G
Solomonereis sp. AU07|G
HYPOTHETICAL GROUP
(NEREIDINAE)
Ceratonereis perkinsi 20|G
Ceratonereis sp. MA7|G
Ceratonereis sp. CG-10|G
Ceratonereis sp. CG-4|G
Ceratonereis singularis MX7|G
Ceratonereis singularis MX5|G
Ceratonereis australis G
Ceratonereis australis AU17|G
Ceratonereis maya MX6|G
Ceratonereis singularis MM15|G
Neanthes cricognatha 19|G
NEREIDINAE
Neanthes sp. MA28|G
Neanthes nubila T
Neanthes agulhana SPD-04-2|G
Neanthes nubila 6NeNu|G
Neanthes kerguelensis 12NeKe|G
Nicon maculata 13NiMa|G
Simplisetia amplidonta B3|G
Eunereis longissima Ferrol2|G
Neanthes kerguelensis C7|G
Rullierinereis sp. FL2|G
Composetia marmorata 21|G
Composetia marmorata MA2|G
Platynereis dumerilii genome T
Laeonereis culveri TB10|G
Laeonereis watsoni MX10|G
Hediste diversicolor
Hediste japonica RBC|G
Alitta virens T
Namalycastis jaya SP4|G
Neanthes glandicincta XD-1|G
Composetia keiskama TB43|G
Leonnates decipens CG-7|G
Nectoneanthes oxypoda AU12|G
Alitta acutifolia MX1|G
Alitta succinea T
Alitta succinea TB03|G
Alitta succinea 4AlSu|G
Simplisetia erythraensis G
Simplisetia sp. CG-3|G
Simplisetia sp. AU13|G
Perinereis pictilis MA13|G
Perinereis osoriotafalli MX31|G
Perinereis elacasoi MX27|G
Nereis pelagica TR02|G
Perinereis anderssoni TB14|G
Perinereis nuntia DC-1|G
Perinereis obfuscata CG-9|G
Perinereis calmani G
Perinereis akuna CG-14|G
Perinereis vallata AU02|G
Neanthes acuminata TR01|G
Nereis caudata TR03|G
Perinereis marionii 5PeMA|G
Perinereis nuntia AU05|G
Perinereis helleri AU18|G
Perinereis singaporiensis CG-13|G
Perinereis vancaurica AU21|G
Perinereis cultrifera DCW|G
Perinereis falklandica 7PeFa|G
Pseudonereis gallapagensis MX35|G
Pseudonereis deleoni MX34|G
Perinereis amblyodonta AU20|G
Pseudonereis sp. MA17|G
Pseudonereis anomala YX-1|G
Nereis confusa MX18|G
Nereis sp. MX20|G
Nereis riisei FL6|G
Nereis sp. MX23|G
Ceratonereis irritabilis FL3|G
Neanthes masalacensis AU10|G
MA24 Nereis sp
Nereis panamensis MX24|G
Neanthes indica SP5|G
Nereis cockburnensis G
Nereis zonata 1NeZo|G
Nereis pelagica TB33|G
Nereis vexillosa QX|G
Nereis grubei MX15|G
Nereis vexillosa TB23|G
Perinereis nuntia DUO|G
Cheilonereis cyclurus FL1|G
Cheilonereis cyclurus HC|G
Eunereis longissima 2EuLo|G
Nereis longissima TR04|G

Figure 5.4: Maximum likelihood tree showing Nereididae evolutionary relationships. Topology and branch lengths are based on the concatenation of a data matrix consisting of 777 near-universal single-copy orthologs (USCOs), from 100 individuals, 312,912 amino acid (AA) positions, with 60.9% occupancy or 39.1% missing data, as inferred using IQ-TREE v2.1.2 under the best-fitting model for each locus. 1000 bootstrap replicates when <100% are also shown. G = genome data, T = transcriptome.

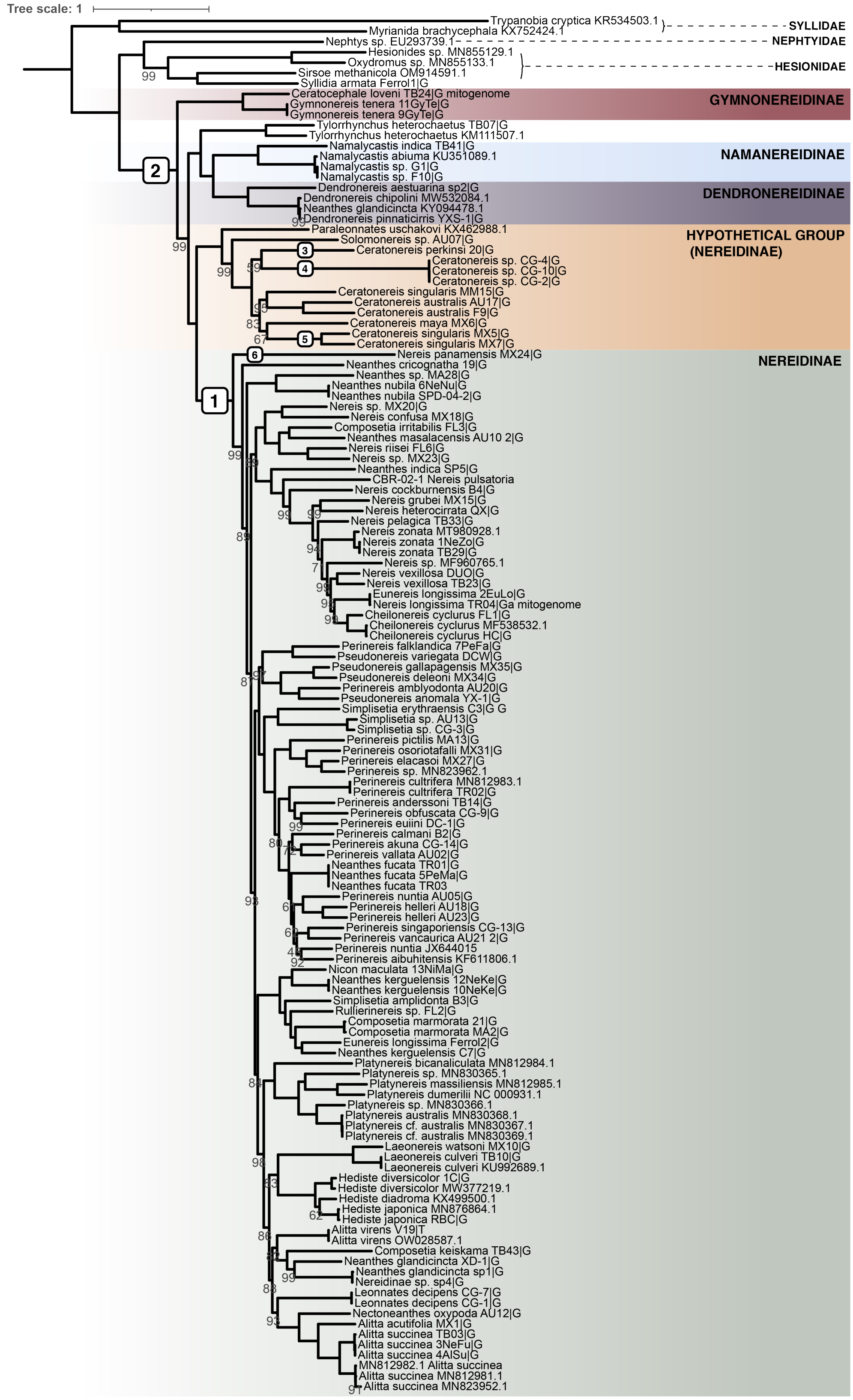

Tree scale: 1
Trypanobia cryptica KR534503.1
Myrianida brachycephala KX752424.1
SYLLIDAE
Nephtys sp. EU293739.1
NEPHTYIDAE
Hesionides sp. MN855129.1
Oxydromus sp. MN855133.1
HESIONIDAE
99
Sirsoe methanicola OM914591.1
Syllidia armata Ferrol1|G
Ceratocephale loveni TB24|G mitogenome
Gymnonereis tenera 11GyTe|G
Gymnonereis tenera 9GyTe|G
GYMNONEREIDINAE
Tylorrhynchus heterochaetus TB07|G
Tylorrhynchus heterochaetus KM111507.1
Namalycastis indica TB41|G
Namalycastis abiuma KU351089.1
Namalycastis sp. G1|G
Namalycastis sp. F10|G
NAMANEREIDINAE
2
Dendronereis aestuarina sp2|G
Dendronereis chipolini MW532084.1
Neanthes glandicincta KY094478.1
99
Dendronereis pinnaticirris YXS-1|G
DENDRONEREIDINAE
Paraleonnates uschakovi KX462988.1
99
Solomonereis sp. AU07|G
3
Ceratonereis perkinsi 20|G
HYPOTHETICAL GROUP
(NEREIDINAE)
99
59
4
Ceratonereis sp. CG-4|G
Ceratonereis sp. CG-10|G
Ceratonereis sp. CG-2|G
Ceratonereis singularis MM15|G
95
Ceratonereis australis AU17|G
Ceratonereis australis F9|G
83
Ceratonereis maya MX6|G
67
5
Ceratonereis singularis MX5|G
Ceratonereis singularis MX7|G
6
Nereis panamensis MX24|G
NEREIDINAE
Neanthes cricognatha 19|G
Neanthes sp. MA28|G
Neanthes nubila 6NeNu|G
Neanthes nubila SPD-04-2|G
1
Nereis sp. MX20|G
Nereis confusa MX18|G
Composetia irritabilis FL3|G
Neanthes masalacensis AU10 2|G
Nereis riisei FL6|G
99
Nereis sp. MX23|G
89
Neanthes indica SP5|G
CBR-02-1 Nereis pulsatoria
Nereis cockburnensis B4|G
Nereis grubei MX15|G
99
99
Nereis heterocirrata QX|G
Nereis pelagica TB33|G
89
Nereis zonata MT980928.1
Nereis zonata 1NeZo|G
94
Nereis zonata TB29|G
71
Nereis sp. MF960765.1
Nereis vexillosa DUO|G
99
Nereis vexillosa TB23|G
Eunereis longissima 2EuLo|G
95
Nereis longissima TR04|Ga mitogenome
99
Cheilonereis cyclurus FL1|G
Cheilonereis cyclurus MF538532.1
Cheilonereis cyclurus HC|G
Perinereis falklandica 7PeFa|G
Pseudonereis variegata DCW|G
Pseudonereis gallapagensis MX35|G
87
97
Pseudonereis deleoni MX34|G
Perinereis amblyodonta AU20|G
Pseudonereis anomala YX-1|G
Simplisetia erythraensis C3|G G
Simplisetia sp. AU13|G
Simplisetia sp. CG-3|G
Perinereis pictilis MA13|G
Perinereis osoriotafalli MX31|G
Perinereis elacasoi MX27|G
Perinereis sp. MN823962.1
Perinereis cultrifera MN812983.1
Perinereis cultrifera TR02|G
Perinereis anderssoni TB14|G
Perinereis obfuscata CG-9|G
99
Perinereis euiini DC-1|G
Perinereis calmani B2|G
80
72
Perinereis akuna CG-14|G
Perinereis vallata AU02|G
Neanthes fucata TR01|G
Neanthes fucata 5PeMa|G
Neanthes fucata TR03
93
Perinereis nuntia AU05|G
61
Perinereis helleri AU18|G
Perinereis helleri AU23|G
62
Perinereis singaporiensis CG-13|G
Perinereis vancaurica AU21 2|G
48
Perinereis nuntia JX644015
92
Perinereis aibuhitensis KF611806.1
Nicon maculata 13NiMa|G
Neanthes kerguelensis 12NeKe|G
Neanthes kerguelensis 10NeKe|G
Simplisetia amplidonta B3|G
Rullierinereis sp. FL2|G
Composetia marmorata 21|G
Composetia marmorata MA2|G
Eunereis longissima Ferrol2|G
Neanthes kerguelensis C7|G
Platynereis bicanaliculata MN812984.1
Platynereis sp. MN830365.1
84
Platynereis massiliensis MN812985.1
Platynereis dumerilii NC 000931.1
Platynereis sp. MN830366.1
Platynereis australis MN830368.1
Platynereis cf. australis MN830367.1
Platynereis cf. australis MN830369.1
Laeonereis watsoni MX10|G
Laeonereis culveri TB10|G
98
Laeonereis culveri KU992689.1
Hediste diversicolor 1C|G
83
Hediste diversicolor MW377219.1
Hediste diadroma KX499500.1
62
Hediste japonica MN876864.1
Hediste japonica RBC|G
86
Alitta virens V19|T
Alitta virens OW028587.1
82
Composetia keiskama TB43|G
Neanthes glandicincta XD-1|G
99
Neanthes glandicincta sp1|G
Nereidinae sp. sp4|G
88
Leonnates decipens CG-7|G
Leonnates decipens CG-1|G
93
Nectoneanthes oxypoda AU12|G
Alitta acutifolia MX1|G
Alitta succinea TB03|G
Alitta succinea 3NeFu|G
Alitta succinea 4AlSu|G
MN812982.1 Alitta succinea
Alitta succinea MN812981.1
91
Alitta succinea MN823952.1

Figure 5.5: Maximum likelihood tree, based on nucleotide data from 15 mitochondrial genes, showing evolutionary relationships within Nereididae. This phylogeny is based on the concatenation of mitochondrial data from 132 taxa, and was inferred using IQ-TREE v2.1.2 under the best-fitting model for each locus. Bootstrap support values, based on 1000 bootstrap replicates, and <100% are displayed

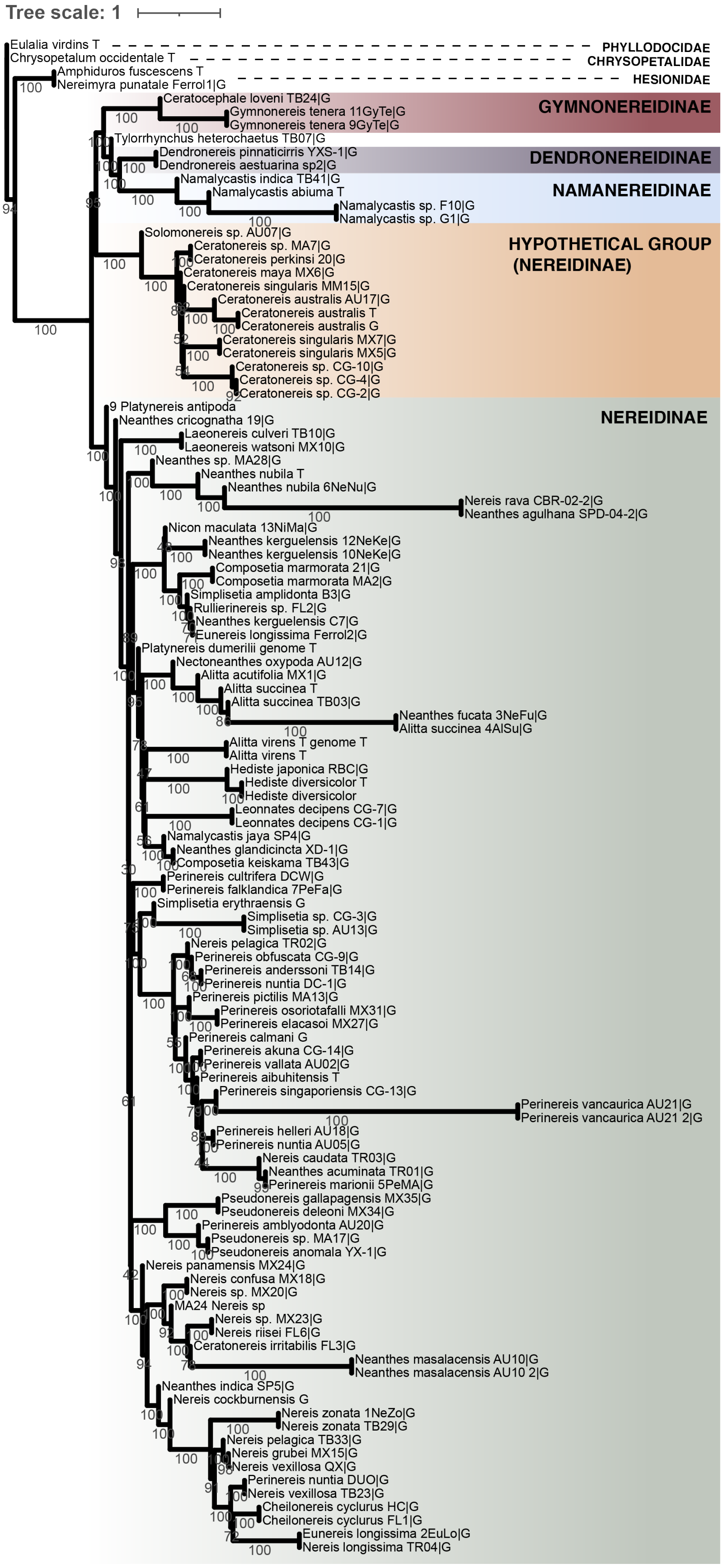

Tree scale: 1
Eulalia virdins T
PHYLLODOCIDAE
Chrysopetalum occidentale T
CHRYSOPETALIDAE
Amphiduros fuscescens T
Nereimyra punatale Ferrol1|G
HESIONIDAE
Ceratocephale loveni TB24|G
Gymnonereis tenera 11GyTe|G
Gymnonereis tenera 9GyTe|G
GYMNONEREIDINAE
Tylorrhynchus heterochaetus TB07|G
Dendronereis pinnaticirris YXS-1|G
Dendronereis aestuarina sp2|G
DENDRONEREIDINAE
Namalycastis indica TB41|G
Namalycastis abiuma T
Namalycastis sp. F10|G
Namalycastis sp. G1|G
NAMANEREIDINAE
Solomonereis sp. AU07|G
Ceratonereis sp. MA7|G
Ceratonereis perkinsi 20|G
Ceratonereis maya MX6|G
Ceratonereis singularis MM15|G
Ceratonereis australis AU17|G
Ceratonereis australis T
Ceratonereis australis G
Ceratonereis singularis MX7|G
Ceratonereis singularis MX5|G
Ceratonereis sp. CG-10|G
Ceratonereis sp. CG-4|G
Ceratonereis sp. CG-2|G
HYPOTHETICAL GROUP
(NEREIDINAE)
9 Platynereis antipoda
Neanthes cricognatha 19|G
Laeonereis culveri TB10|G
Laeonereis watsoni MX10|G
Neanthes sp. MA28|G
Neanthes nubila T
Neanthes nubila 6NeNu|G
Nereis rava CBR-02-2|G
Neanthes agulhana SPD-04-2|G
Nicon maculata 13NiMa|G
Neanthes kerguelensis 12NeKe|G
Neanthes kerguelensis 10NeKe|G
Composetia marmorata 21|G
Composetia marmorata MA2|G
Simplisetia amplidonta B3|G
Rullierinereis sp. FL2|G
Neanthes kerguelensis C7|G
Eunereis longissima Ferrol2|G
Platynereis dumerilii genome T
Nectoneanthes oxypoda AU12|G
Alitta acutifolia MX1|G
Alitta succinea T
Alitta succinea TB03|G
Neanthes fucata 3NeFu|G
Alitta succinea 4AlSu|G
Alitta virens T genome T
Alitta virens T
Hediste japonica RBC|G
Hediste diversicolor T
Hediste diversicolor
Leonnates decipens CG-7|G
Leonnates decipens CG-1|G
Namalycastis jaya SP4|G
Neanthes glandicincta XD-1|G
Composetia keiskama TB43|G
Perinereis cultrifera DCW|G
Perinereis falklandica 7PeFa|G
Simplisetia erythraensis G
Simplisetia sp. CG-3|G
Simplisetia sp. AU13|G
Nereis pelagica TR02|G
Perinereis obfuscata CG-9|G
Perinereis anderssoni TB14|G
Perinereis nuntia DC-1|G
Perinereis pictilis MA13|G
Perinereis osoriotafalli MX31|G
Perinereis elacasoi MX27|G
Perinereis calmani G
Perinereis akuna CG-14|G
Perinereis vallata AU02|G
Perinereis aibuhitensis T
Perinereis singaporiensis CG-13|G
Perinereis vancaurica AU21|G
Perinereis vancaurica AU21 2|G
Perinereis helleri AU18|G
Perinereis nuntia AU05|G
Nereis caudata TR03|G
Neanthes acuminata TR01|G
Perinereis marionii 5PeMA|G
Pseudonereis gallapagensis MX35|G
Pseudonereis deleoni MX34|G
Perinereis amblyodonta AU20|G
Pseudonereis sp. MA17|G
Pseudonereis anomala YX-1|G
Nereis panamensis MX24|G
Nereis confusa MX18|G
Nereis sp. MX20|G
MA24 Nereis sp
Nereis sp. MX23|G
Nereis riisei FL6|G
Ceratonereis irritabilis FL3|G
Neanthes masalacensis AU10|G
Neanthes masalacensis AU10 2|G
Neanthes indica SP5|G
Nereis cockburnensis G
Nereis zonata 1NeZo|G
Nereis zonata TB29|G
Nereis pelagica TB33|G
Nereis grubei MX15|G
Nereis vexillosa QX|G
Perinereis nuntia DUO|G
Nereis vexillosa TB23|G
Cheilonereis cyclurus HC|G
Cheilonereis cyclurus FL1|G
Eunereis longissima 2EuLo|G
Nereis longissima TR04|G
NEREIDINAE

Figure 5.6: Coalescent-based species tree estimation using ASTRAL v5.7.1 based on 777 single-copy gene trees, generated using FastTree v2.1.10. "local posterior probability" support values have also been added to the figure.

5.5 Discussion

5.5.1 Nereidid phylogeny

Our phylogenetic of Nereididae is based on nuclear genomic-, transcriptomic-, and mitochondrial data. Three trees were presented: two maximum likelihood trees based on the best-fitting model for each partition (figure 5.4 and 5.5) and one coalescent-based species tree (figure 5.6).

Both our mitochondrial and nuclear trees recover Gymnonereidinae, restricted here to *Ceratocephale loveni*, and *Gymnonereis tenera*, as a sister-group to all other nereidids in our analysis. This restricted version of Gymnonereidinae—sensu Banse, 1977—excludes both *Tylorrhynchus* and *Laeonnareis*. Gymnonereidinae sensu Banse 1977 is characterized by bifid ventral and subacicular notopodial chaetae and was well-supported in a morphological study by Santos et al., 2006 and a mtDNA analysis by Alves et al., 2020a.

Although prior phylogenetic analyses (e.g., Bakken and Wilson, 2005; Fitzhugh, 1987; C. Glasby, 1991; Pleijel and Dahlgren, 1998; Santos et al., 2006) places Namanereidinae as the sister-group to the remaining subfamilies and genera, our analysis indicates that Namanereidinae, here consisting of *Namalycastis indica*, *Namalycastis abiuma*, and *Namalycastis sp.*, is nested within a larger clade which also contain the genera *Tylorrhynchus* and *Dendronereis*. Namanereidinae contain two genera, are adapted to low-salinity conditions, and are characterized by a bare pharynx and reduced notopodia. While we find consistent support this group being monophyletic, our analyses suggests that these morphological characters are an adaptation to their unique living-conditions rather than being the ancestral state of Nereididae.

Dendronereidinae, characterized by the presence of branchiae, are recovered as monophyletic, but again within in a larger clade which also includes *Tylorrhynchus* and *Namanereidinae*. The subfamily Dendronereidinae was originally erected by Pillai, 1961, and included three taxa: Tambalagamia Pillai, 1961, Dendronereis Peters, 1854 and Dendronereides Southern, 1921. Both *Tambalagamia* and *Dendronereides* are absent in our study and the monophyly of this grouping can thus not be tested. Dendronereidinae's subfamily status has been criticized by Banse, 1977 and Santos et al., 2006 rejected the homology of the morphological character upon which this grouping is based on. Dendronereidinae is currently considered *nomen dubium* and our restricted analysis of this group does not support a subfamily status of this group.

Like the parsimonious study, based on morphological data, by Bakken and Wilson, 2005, we find consistent evidence for that some of the largest genera, *Nereis*, *Neanthes*, and *Perinereis* are polyphyletic. Furthermore, *Alitta*, is recovered here as a paraphyletic group.

Similar to Alves et al., 2020a, we find the genus *Laeonereis*, here represented by *L. culveri* and *L. watsoni*, to be nested within Nereidinae instead of Gymnonereidinae, to which they are currently ascribed. Our mitochondrial analysis further indicates that the species *Neanthes glandicincta*, used in the fore-mentioned study, is a mis-identification and that the actual species is actually *Dendronereis chipolini*. Thus, the monophyly of Nereidinae as currently described is still rejected due to the placement of the genus *Laeonereis* therewithin. However, the reason for why one member of Nereidinae was recovered *outside* of this clade was because of a misidentified taxon.

We also find consistent support for grouping the genera *Solomonereis*, *Ceratonereis*, and possibly also *Paraleonnates uschakovi* into a new, hypothetical clade, which then forms a sister-group to the subfamily Nereidinae. Bakken and Wilson, 2005 already observed morphological similarities in these two genera and the phylogenetic analysis by Santos et al., 2006 gave further support to this clade. Bakken and Wilson, 2005 also placed *Unanereis*, which is very similar to *Ceratonereis* except for the presence of a single antenna, together with this group and it is possible that this genus belongs here as well.

Within Nereidinae, some of the larger genera are recovered as polyphyletic. Notably, *Neanthes cricognatha* is the sister group to all other members of Nereidinae but then other species within *Neanthes* are placed within their own, separate clade or as a sister species to, e.g., *Nicon*, *Composetia*, or *Nereis*. *Nereis* is another large genus that is also grouped together with several different groups of genera such as *Perinereis*, *Eunereis*, and *Perinereis*. Furthermore, although *Alitta* is only observed in one single clade, it is placed both as a sister group to a clade consisting of *Namalycastis*, *Neanthes*, and *Composetia*, but also within its own clade and as a sister group to *Nectoneanthes*.

When comparing our nuclear-gene tree to our mitochondrial gene tree, our main findings—e.g., Gymnonereidinae is the sister group to all other nereidids, the genera *Solomonereis* and *Ceratonereis* forms a previously unrecognized hypothetical group—remain the same throughout. Within the subfamily Nereidinae, however, we do observe some incongruencies: the clade consisting of *Perinereis* and *Pseudonereis* is in the nuclear-gene tree placed as a sister group to two larger clade which include *Cheilonereis cyclurus* and more, while in the mitochondrial gene tree, it forms a sister group to the only clade that includes *Simplisetia sp.*.

The discerning reader will also observe some obvious mislabeled species. E.g., *Namalycastis jaya* in Nereidinae, which is a member of the subfamily Namanereidinae. To account for these cases, we have performed barcoding of all of the nereidids included in this study (see supplementary data). Although we were not able to get the mislabeled *Namalycastis jaya* to the species level, our barcode comparison says that its sequence is closest to a *Nereis sp.*, making us confidence that this is not a member of Namanereidinae and instead Nereidinae. A list of misidentified species, together with a corrected identification, based on a similarity search against the BOLD database can be seen in table 5.2. Our nuclear-gene and coalescent-based tree display IDs in their original, misidentified form, while the mitochondrial gene tree have their labels corrected.

Table 5.2: A list of misidentified species and their corrected species identification label, based on a similarity search against the BOLD database.

Original label	**ID**	**Corrected label**	**Similarity**
Namalycastis jaya	SP4	Nereis sp.	82.17%
Composetia keiksama	TB43	Leonnates decipens	80.46%
Neanthes fucata	3NeFU	Alitta succinea	100%

5.5.2 Mitochondrial gene order evolution

Like prior studies of mitochondrial gene order evolution within nereidids (Alves et al., 2020a; Park et al., 2016), we can, with our expanded dataset, confirm that protein-coding and rRNA gene order is highly preserved throughout Nereididae as a whole. We do observe the translocation of the protein-coding gene ATP8 in group 6, but this is only observed in a single species. We have presented four additional tRNA arrangement, in addition to the ones first described in Park et al., 2016. Given the position within the tree of group 6 (*Nereis panamensis*), there are three possibilities for how this arrangement happened: (i) group 1 is the ancestral state within Nereidinae and group 6 is derived from there, (ii) group 6 is the ancestral state and and group 1 is derived from that state, or (iii) both changed happened independently and evolved from something similar to that in group 2. Given that the only difference is the location of tRNA-Met, tRNA-Asp, and ATP8, the most parsimonous explanation is (i) that group 1 is the ancestral state and there was one translocation event where tRNA-Met, tRNA-Asp, and ATP8 all moved to a different location. The second explanation (ii) requires that two different translocation events whereas the second event resulted in something akin to the ancestral state by pure chance and the third explanation (iii) is very unlikely given the similarities that we observe in group 1 and 6.

5.5.3 Evaluating Patchwork

Using our new genome skimming approach, we were able to recover a well-supported phylogeny, based on a large taxon sampling and nearly 800 markers. By basing our analysis on WGS data, we were able to successfully include many older, ethanol-preserved samples and, since we did not use target enrichment, our strategy had very little work up-front since we did not have to design any oligonucleotide probes. Processing one sample, in Patchwork, takes around 7–15 minutes (depending on the number of reads) and supporting scripts already support multi-sample processing such that the user does not have to perform each run manually. Furthermore, Patchwork supports that the ability to work in an incremental fashion—samples can be added or removed, without having to redo the same analysis, as long as the markers remain the same. As suggested by Alves et al., 2020a, mitochondrial gene order information is well-preserved within nereidids, as in other annelids, and phylogenetically informative. As a result, we were able to evaluate our method by comparing it to our mitochondrial gene tree. Small discrepancies between the two suggests that this method works well for our purposes of inferring a phylogeny, despite the old age of the group. I.e., we do not observe any strange placements due methodological bias as the distance to the reference used (*Alitta virens*) increases. We do, however, notice that the interspecies distances in the nuclear phylogeny are sometimes exaggerated when compared to the mitochondrial phylogeny. OTUs which were recovered as identical in the mitochondrial gene tree have a small, yet observable, pairwise distance in the nuclear gene tree. We suspect that this is a result of markers having unequal coverage across taxa. Mitochondrial genes are high-copy markers and have a very high coverage, while the nuclear genes are more variable and have a much lower coverage in general. This is further exaggerated by the fact that we are using low-coverage genomes and this effect could potentially be mitigated by increasing the sequencing depth.

5.6 Conclusion

We have presented a phylogeny of Nereididae, based for the first time on genomic and transcriptomic sequences from 100 ingroup species and 4 outgroup species from 3 different families. A coalescent-based species tree, a maximum likelihood tree based on 15 mitochondrial genes, both from our own and publically available sequences, and an analysis of mitochondrial gene order preservation in Nereididae complements and provides further support and insight into our phylogenomic analysis.

These phylogenies demonstrates that the subfamily Namanereidinae is not the basal-most node within the group as previously thought. Instead, we find consistent support for that a subset of Gymnonereidinae—as currently described—is most likely the sister group to all other nereidids. Furthermore, the genera *Tylorrhynchus* and *Laeonereis* are recovered outside of Gymnonereidinae to which they currently belong. We further found support for grouping the genera *Ceratonereis*, *Solomononereis*, and most likely *Paraleonnates* into a previously unrecognised clade. Finally, an analysis of mitochondrial gene orders across the major groups provided further supported for the major clades we recovered.

This is also the first study in which Patchwork is used for mining phylogenetic markers from whole-genome sequencing data. The completeness of our data matrix as well as the well-supported phylogeny, which to a large extent matches our mitochondrial genome-based tree, clearly demonstrates the effectiveness of this low-workload approach to phylogenomics. Unlike with a pure transcriptome analysis, we were able to include more markers and ethanol-preserved specimen from various museums and research collections, and unlike with targetted- or hybrid enrichment, we forewent the error-prone and computationally intensive step of probe-design.

The mislabling of *Neanthes glandicincta* in the phylogenetic analysis by Alves et al., 2020a highlights the importance of publishing DNA barcode sequences, which can be used for species identification, delimiation, and re-verification. Because the mitochondrial genome in question had been published on GenBank (accession no. NC_035893), we were able to reject the previous species identification.

5.7 Back matter

Data availability

Individuals of all species (or populations) used in the phylogenomic analyses will be deposited as vouchers in the Zoological Museum at the Georg-August-University Göttingen. All raw sequence raw data will be submitted to the sequence-read archive (SRA) and transcriptome assemblies to the assembly database of NCBI GenBank (http://www.ncbi.nlm.nih.gov/genbank/). All published alignments, data matrices and phylogenetic trees will be uploaded to the Dryad repository (http://datadryad.org/). It is intended to publish open access, thereby allowing unrestricted access to all results. This will be supported by the open access fund at the University of Göttingen (https://www.sub.uni-goettingen.de/en/electronic- publishing/open-access/open-access-publication-funding/).

Code availability

Our workflow for performing quality control, trimming, assembly, and assembly quality assessment on WGS data is published on GitHub at https://github.com/ThiloSchulze/eukaryotic-genome-assembly The pipeline used to extract and annotate mitochondrial genomes from WGS data can be found on GitHub: https://github.com/ThiloSchulze/mitogenome-extraction. The implementation of Patchwork is described in its manuscript (Thalen et al., 2022) and instructions on obtaining and installing Patchwork can be found on GitHub: https://github.com/fethalen/Patchwork. PhyloPyPruner (Thalén et al. in prep.), a tree-based orthology inference program which in our case was used to filter low-occupancy genes and for constructing the supermatrices is found at https://github.com/fethalen/phylopypruner. Other supplementary data for this article can be downloaded from https://github.com/Animal-Evolution-and-Biodiversity/nereididae-phylogenomic.

Acknowledgments

This work was supported by the German Research Foundation (DFG) and was partly supported by the Göttingen Graduate Center for Neurosciences, Biophysics, and Molecular Biosciencces of the Georg-August-Universität Göttingen.

Competing interests

The authors declare no competing interests.

Author's contributions

C.B.: conceptualization, acquisition of funding, supervision, and review and editing. **K.H.:** DNA- and RNA extraction, sequence amplification, and purification. **F.H.:** RNA extraction, transcriptome assembly and bioinformatic analysis. **T.S.:** co-implemented the assembly pipeline together with **F.T**, performed genome assemblies, implemented a pipeline for extracting and annotating mitochondrial genomes, performed the gene order analysis and inferred the mitochondrial-gene tree. **C.G.K.:** co-implemented the "Patchwork" program for phylogenetic marker discovery together with **F.T.** and performed genome assemblies. **F.T.:** designed the phylogenomic pipeline, performed bioinformatic analyses, implementation of scripts for automation, supervision, writing and editing. Authors not previously mentioned here were part of the worldwide effort to collect, sequence, and share nereidid samples, without whom this study wouldn't have been possible.

5.8 Supplementary material

The following are the Supplementary data to this article:

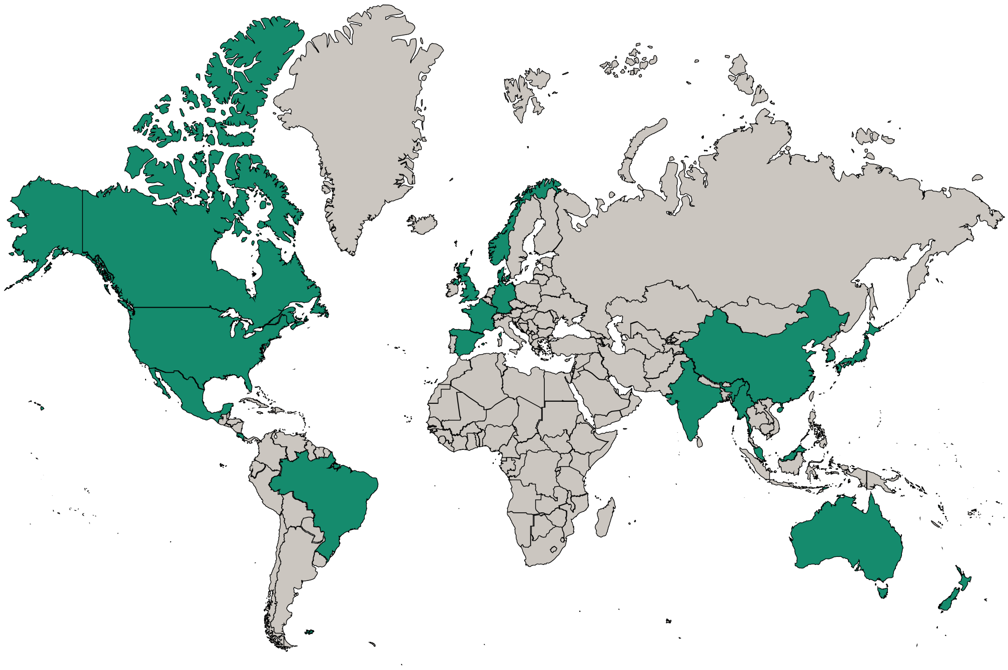


Figure 5.7: Sampling locations for from which specimen were taken. Countries from which one or more samples were taken are highlighted in green.

Table 5.3: Assembly statistics for the first 50 species used in the nuclear-gene analysis, showing the voucher, K-mer size, number of assembled contigs, mean contig length, and the N50.

Species	**Voucher**	**K-mer**	**Contigs**	**Avg. contig len.**	**N50**
Alitta acutifolia	-	33	788193	749,46	944
Alitta succinea	4AlSu	33	1462068	482,53	556
Alitta succinea	TB03	33	1506301	457,71	517
Alitta virens	V19-2	91	434064	976,95	1450
Ceratocephale loveni	TB25	33	249611	1304,59	2091
Ceratonereis australis	AU17	33	793589	579,56	1232
Ceratonereis australis	F9	55	5332124	203,25	220
Ceratonereis singularis	-	33	975135	495,74	827
Ceratonereis irritabilis	FL3	33	529025	997,77	1445
Ceratonereis maya	-	33	600323	883,24	1568
Ceratonereis perkinsi	20	55	1725482	501,52	922
Ceratonereis singularis	-	33	413492	911,39	2160
Ceratonereis singularis	MM15	33	510542	815,67	1852
Ceratonereis sp.	CG-10	33	637007	778,42	1143
Ceratonereis sp.	CG-2	33	564843	882,89	1332
Ceratonereis sp.	CG-4	33	561679	878,72	1308
Ceratonereis sp.	MA7	33	828790	668,09	871
Eunereis longissima	Ferrol2	33	499581	962,22	1343
Neanthes sp.	MA28	33	1100390	467,11	532
Nereis sp.	MA24	33	1353413	387,37	421
Pseudonereis sp.	MA17	33	1756523	421,75	597
Cheilonereis cyclurus	FL1	33	715627	618,93	1576
Cheilonereis cyclurus	HC	91	1111743	622,24	2100
Composetia keiskama	TB43	33	751017	586,19	1172
Composetia marmorata	21	55	718392	1003,14	1477
Composetia marmorata	MA2	33	813784	598,62	742
Dendronereis aestuarina	SP2	55	654940	543,78	1169
Dendronereis pinnaticirris	YXS-1	91	1239990	516,02	1465
Eunereis longissima	2EuLo	33	252716	1349,95	2024
Gymnonereis tenera	11GyTe	33	231708	1446,5	2348
Gymnonereis tenera	9GyTe	33	689839	579,86	1251
Hediste diversicolor	1C	55	4070369	214,46	227
Hediste japonica	RBC	91	3071209	423,12	766
Laeonereis culveri	TB10	33	294599	1093,24	3063
Laeonereis watsoni	-	33	508699	684,83	1706
Leonnates decipiens	CG-1	33	352070	1036,06	1383
Leonnates decipiens	CG-7	33	361012	993,26	1313
Namalycastis abiuma	F10	55	176374	1514,85	6302
Namalycastis abiuma	G1	55	182023	1420,92	8827
Namalycastis indica	TB41	33	479861	605,72	1427
Namalycastis jaya	SP4	55	1006169	728,32	1927
Neanthes acuminata	TR01	33	482209	893,39	1132
Neanthes agulhana	SPD-04-2	33	517137	1042,32	1390
Neanthes cricognatha	19	55	1199799	834,83	1140
Neanthes fucata	3NeFu	33	1386545	488,19	563
Neanthes glandicincta	SP1	55	1017058	717,8	1893
Neanthes glandicincta	XD-1	91	983702	701,06	3522
Neanthes indica	SP5	55	963575	479,39	763
Neanthes kerguelensis	10NeKe	33	500796	519,46	619
Neanthes kerguelensis?	12NeKe	33	492914	499,64	581

Table 5.4: Assembly statistics for species 51–100 used in the nuclear-gene analysis, showing the voucher, K-mer size, number of assembled contigs, mean contig length, and the N50.

Species	**Voucher**	**K-mer**	**Contigs**	**Avg. contig len.**	**N50**
Neanthes kerguelensis	C7	55	1379626	610,83	1295
Neanthes masalacensis	AU10	33	1056097	597,34	720
Neanthes nubila	6NeNu	33	627193	775,14	972
Neanthes trifasciata	26	55	13191	584,33	2177
Nectoneanthes oxypoda	AU12	33	714669	564,66	645
Nereimyra cf punatale	Ferrol1	33	479950	1068,94	1583
Nereis caudata	TR03	33	523243	726,89	901
Nereis sp.	-	33	317121	1461,63	2383
Nereis cockburnensis	B4	71	377206	1154,86	2129
Nereis confusa	-	33	468334	1020,16	1424
Nereis longissima	TR04	33	235025	1492,53	2379
Nereis maxillodentata	7	125	24910	447,58	1088
Nereis sp.	-	33	486919	983,53	1306
Nereis sp.	-	33	627662	924,94	1201
Nereis panamensis	-	33	786853	382,44	416
Nereis pelagica	TB33	33	433845	894,97	1324
Nereis pelagica	TR02	33	592291	649,08	801
Nereis pulsatoria	CBR-02-1	55	1843858	533,91	811
Nereis rava	CBR-02-2	33	517137	1042,32	1390
Nereis riisei	FL6	33	567219	1141,4	1604
Nereis vexillosa	QX	91	2666085	468,36	1038
Nereis vexillosa	TB23	33	627274	660,69	1468
Nereis zonata	1NeZo	33	265467	1304,75	2067
Nereis zonata	TB29	33	257526	1381,02	2133
Nicon maculata	13NiMa	33	505692	716,01	917
Perinereis akuna	CG-14	33	997714	530,07	1070
Perinereis amblyodonta	AU20	33	630996	832,52	1087
Perinereis anderssoni	TB14	33	418546	1176,39	1746
Perinereis calmani	B2	91	818166	546,17	619
Perinereis cultrifera	DCW	91	1712718	538,04	1176
Perinereis elenacasoi	-	33	1201383	438,8	785
Perinereis falklandica	7PeFa	33	1011621	490,3	959
Perinereis helleri	AU18	33	527928	892,46	1195
Perinereis helleri	AU23	33	365620	285,38	264
Perinereis marionii	5PeMa	33	1137072	494,75	848
Perinereis nuntia	AU05	33	826771	596,91	776
Perinereis nuntia	DC-1	91	3196755	511,86	1564
Perinereis nuntia	DUO	91	2513136	417,84	707
Perinereis obfuscata	CG-9	33	832599	549,83	669
Perinereis osoriotafalli	-	33	837595	561,72	1169
Perinereis pictilis	MA13	33	1149879	577,31	690
Perinereis singaporiensis	CG-13	33	470169	1070,26	1480
Perinereis vallata	AU02	33	573173	639,59	784
Perinereis vancaurica	AU21	33	660596	745,56	956
Platynereis antipoda	9	55	5147224	544,5	1040
Pseudonereis anomala	YX-1	91	3650444	479,07	1194
Pseudonereis deleoni	-	33	394883	1129,05	1627
Pseudonereis gallapagensis	-	33	438283	1007,24	1406
Rullierinereis sp.	FL2	33	526726	1097,9	1781
Simplisetia amplidonta	B3	55	1201776	700,93	1203
Simplisetia erythraeensis	C3	91	473511	780,53	1078

Table 5.5: Assembly statistics for species 101–104 used in the nuclear-gene analysis, showing the voucher, K-mer size, number of assembled contigs, mean contig length, and the N50.

Species	**Voucher**	**K-mer**	**Contigs**	**Avg. contig len.**	**N50**
Simplisetia sp.	AU13	33	1251600	359,66	547
Simplisetia sp.	CG-3	33	1487795	330,89	477
Solomononereis sp.	AU07	33	1300356	404,99	650
Tylorrhynchus heterochaetus	TB07	33	431888	968,4	1315

Table 5.6: Sampling locations of all specimen which genomes were sequenced for study. **n** is the number of species from that were collected from each country.

Code	**Country**	**n**
AU	Australia	37
BR	Brazil	2
CA	Canada	1
CH	China	9
CR	Costa Rica	2
DE	Denmark	1
TL	East Timor	4
FK	Falkland Islands	6
FR	France	4
DE	Germany	2
IN	India	4
JP	Japan	1
MY	Malaysia	8
MX	Mexico	13
MM	Myanmar	1
NZ	New Zealand	1
NO	Norway	3
ZA	South Africa	1
KR	South Korea	1
ES	Spain	9
GB	United Kingdom	6
US	United States	4
VU	Vanuatu	1

Table 5.7: List of additional, publicly available species used for the mitochondrial gene tree analysis and corresponding Genbank numbers.

Taxa	**Data type**	**Genbank no.**	**Collection site**	**Reference**
Nephtys sp.	mtDNA	EU293739	N/A	Vallès et al. 2008
Perinereis nuntia	mtDNA	JX644015	Korea	Won et al. 2013
Perinereis aibuhitensis	mtDNA	KF611806	Korea	Kim et al. 2015
Tylorrhynchus heterochaetus	mtDNA	KM111507	China	Direct submission
Trypanobia cryptica	mtDNA	KR534503	Australia	Aguado et al. 2015
Namalycastis abiuma	mtDNA	KU351089	China	Lin et al. 2016
Laeonereis culveri	mtDNA	KU992689	Brazil	Seixas et al. 2016
Paraleonnates uschakovi	mtDNA	KX462988	Korea	Park et al. 2016
Hediste diadroma	mtDNA	KX499500	Korea	Kim et al. 2016
Myrianida brachycephala	mtDNA	KX752424	Germany	Aguado et al. 2016
Neanthes glandicincta	mtDNA	KY094478	China	Lin et al. 2017
Cheilonereis cyclurus	mtDNA	MF538532	South Korea	Park et al. 2017
Nereis sp.	mtDNA	MF960765	South Korea	Kim et al. 2017
Alitta succinea	mtDNA	MN812981	USA	Alves et al. 2020
Alitta succinea	mtDNA	MN812982	USA	Alves et al. 2020
Perinereis cultrifera	mtDNA	MN812983	France	Alves et al. 2020
Platynereis bicanaliculata	mtDNA	MN812984	USA	Alves et al. 2020
Platynereis massiliensis	mtDNA	MN812985	Wales	Alves et al. 2020
Alitta succinea	cDNA	multiple	USA	Alves et al. 2020
Perinereis sp.	cDNA	multiple	Panama	Alves et al. 2020
Platynereis sp.1	mtDNA	MN830365	Brazil	Alves et al. 2020
Platynereis sp.2	mtDNA	MN830366	Brazil	Alves et al. 2020
Platynereis cf. australis	mtDNA	MN830367	Chile	Alves et al. 2020
Platynereis cf. australis	mtDNA	MN830368	Chile	Alves et al. 2020
Platynereis cf. australis	mtDNA	MN830369	Chile	Alves et al. 2020
Hesionides sp.	cDNA	multiple	Panama	Alves et al. 2020
Oxydromus sp.	cDNA	multiple	Panama	Alves et al. 2020
Hediste japonica	mtDNA	MN876864	South Korea	Park et al. 2020
Nereis zonata	mtDNA	MT980928	N/A	Direct Submission
Hediste diversicolor	mtDNA	MW377219	Norway	Gomes-dos-Santos et al. 2021
Dendronereis chipolini	mtDNA	MW532084	China	Zhen et al. 2022
Platynereis dumerilii	mtDNA	AF178678	N/A	Boore and Brown 2000
Sirsoe methanicola	mtDNA	OM914591	N/A	Lim et al. 2022
Alitta virens	mtDNA	OW028587	United Kingdom	Direct Submission

Bibliography

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. Nature methods, vol. 8no. 1, 61–65.

Allen, J. M., Boyd, B., Nguyen, N.-P., Vachaspati, P., Warnow, T., Huang, D. I., Grady, P. G., Bell, K. C., Cronk, Q. C., Mugisha, L., et al. (2017). Phylogenomics from whole genome sequences using atram. Systematic Biology, vol. 66no. 5, 786–798.

Allen, J. M., Boyd, B., Nguyen, N.-p., Vachaspati, P., Warnow, T., Huang, D. I., Grady, P. G., Bell, K. C., Cronk, Q. C., Mugisha, L., Pittendrigh, B. R., Soledad Leonardi, M., Reed, D. L., & Johnson, K. P. (2017). Phylogenomics from Whole Genome Sequences Using aTRAM. Systematic Biology, syw105.

Allen, J. M., Huang, D. I., Cronk, Q. C., & Johnson, K. P. (2015). Atram-automated target restricted assembly method: A fast method for assembling loci across divergent taxa from next-generation sequencing data. BMC bioinformatics, vol. 16no. 1, 1–7.

Allen, J. M., LaFrance, R., Folk, R. A., Johnson, K. P., & Guralnick, R. P. (2018a). aTRAM 2.0: An Improved, Flexible Locus Assembler for NGS Data. Evolutionary Bioinformatics Online, vol. 14, 1176934318774546.

Allen, J. M., LaFrance, R., Folk, R. A., Johnson, K. P., & Guralnick, R. P. (2018b). Atram 2.0: An improved, flexible locus assembler for ngs data. Evolutionary Bioinformatics, vol. 14, 1176934318774546.

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). Mitofinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Molecular Ecology Resources, vol. 20no. 4, 892–905.

Almogy, G., Pratt, M., Oberstrass, F., Lee, L., Mazur, D., Beckett, N., Barad, O., Soifer, I., Perelman, E., Etzioni, Y., et al. (2022). Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. bioRxiv.

Al-Nakeeb, K., Petersen, T. N., & Sicheritz-Pontén, T. (2017). Norgal: Extraction and de novo assembly of mitochondrial dna from whole-genome sequencing data. BMC bioinformatics, vol. 18no. 1, 1–7.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, vol. 215no. 3, 403–410.

Alves, P. R., Halanych, K. M., & Santos, C. S. G. (2020a). The phylogeny of Nereididae (Annelida) based on mitochondrial genomes. Zoologica Scripta, vol. 49no. 3, 366–378 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/zsc.12413.

Alves, P. R., Halanych, K. M., & Santos, C. S. G. (2020b). The phylogeny of nereididae (annelida) based on mitochondrial genomes. Zoologica Scripta, vol. 49no. 3, 366–378.

Andermann, T., Torres Jiménez, M. F., Matos-Maraví, P., Batista, R., Blanco-Pastor, J. L., Gustafsson, A. L. S., Kistler, L., Liberal, I. M., Oxelman, B., Bacon, C. D., & Antonelli, A. (2020). A guide to carrying out a phylogenomic target sequence capture project. Frontiers in Genetics, vol. 10, 1407.

Andrade, S. C., Novo, M., Kawauchi, G. Y., Worsaae, K., Pleijel, F., Giribet, G., & Rouse, G. W. (2015a). Articulating “archiannelids”: Phylogenomics and annelid relationships, with emphasis on meiofaunal taxa. Molecular Biology and Evolution, vol. 32no. 11, 2860–2875.

Andrade, S. C., Novo, M., Kawauchi, G. Y., Worsaae, K., Pleijel, F., Giribet, G., & Rouse, G. W. (2015b). Articulating “Archiannelids”: Phylogenomics and Annelid Relationships, with Emphasis on Meiofaunal Taxa. Molecular Biology and Evolution, vol. 32no. 11, 2860–2875.

Bakken, T., Glasby, C. J., Santos, C. S., & Wilson, R. S. (2018). Nereididae blainville, 1818.

Bakken, T., & Wilson, R. S. (2005). Phylogeny of nereidids (polychaeta, nereididae) with paragnaths. Zoologica Scripta, vol. 34no. 5, 507–547.

Banse, K. (1977). Gymnonereidinae new subfamily: The nereididae (polychaeta) with bifid parapodial neurocirri. Journal of Natural History, vol. 11no. 6, 609–628.

Baum, D. A., & Smith, S. (2013). Tree thinking. An Introduction to Phylogenetic Biology. Roberts and Company Publishers.

Benson, D., Lipman, D. J., & Ostell, J. (1993). Genbank. Nucleic Acids Research, vol. 21no. 13, 2963–2965.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., & Stadler, P. F. (2013). Mitos: Improved de novo metazoan mitochondrial genome annotation. Molecular phylogenetics and evolution, vol. 69no. 2, 313–319.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A Fresh Approach to Numerical Computing. SIAM Review, vol. 59no. 1, 65–98.

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC genomics, vol. 13no. 1, 1–14.

Bleidorn, C. (2017). Phylogenomics: An introduction. Springer.

Boore, J. L., & Brown, W. M. (2000). Mitochondrial genomes of galathealinum, helobdella, and platynereis: Sequence and gene arrangement comparisons indicate that pogonophora is not a phylum and annelida and arthropoda are not sister taxa. Molecular Biology and Evolution, vol. 17no. 1, 87–106.

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016a). Exon capture phylogenomics: Efficacy across scales of divergence. Molecular Ecology Resources, vol. 16no. 5, 1059–1068
_eprint:
https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12449.

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016b). Exon capture phylogenomics: Efficacy across scales of divergence. Molecular ecology resources, vol. 16no. 5, 1059–1068.

Breinholt, J. W., Earl, C., Lemmon, A. R., Lemmon, E. M., Xiao, L., & Kawahara, A. Y. (2018). Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. Systematic Biology, vol. 67no. 1, 78–93.

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods, vol. 18no. 4, 366–368.

Buck, M., & Hamilton, C. (2011). The nagoya protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the convention on biological diversity. Review of European Community & International Environmental Law, vol. 20no. 1, 47–61.

Call, E., Mayer, C., Twort, V., Dietz, L., Wahlberg, N., & Espeland, M. (2021a). Museomics: Phylogenomics of the Moth Family Epicopeiidae (Lepidoptera) Using Target Enrichment (M. Mutanen, Ed.). Insect Systematics and Diversity, vol. 5no. 2, 6.

Call, E., Mayer, C., Twort, V., Dietz, L., Wahlberg, N., & Espeland, M. (2021b). Museomics: Phylogenomics of the moth family epicopeiidae (lepidoptera) using target enrichment. Insect Systematics and Diversity, vol. 5no. 2, 6.

Cannon, J. T., Vellutini, B. C., Smith, J., Ronquist, F., Jondelius, U., & Hejnol, A. (2016). Xenacoelomorpha is the sister group to nephrozoa. Nature, vol. 530no. 7588, 89–93.

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic acids research, vol. 44no. 19, e147–e147.

Chan, C. X., & Ragan, M. A. (2013). Next-generation phylogenomics. Biology direct, vol. 8no. 1, 1–6.

Cong, Q., Shen, J., Borek, D., Robbins, R. K., Opler, P. A., Otwinowski, Z., & Grishin, N. V. (2017). When coi barcodes deceive: Complete genomes reveal introgression in hairstreaks. Proceedings of the Royal Society B: Biological Sciences, vol. 284no. 1848, 20161735.

Consortium, H. G. S. et al. (2004). Finishing the euchromatic sequence of the human genome. Nature, vol. 431no. 7011, 931–945.

Cronn, R., Knaus, B. J., Liston, A., Maughan, P. J., Parks, M., Syring, J. V., & Udall, J. (2012a). Targeted enrichment strategies for next-generation plant biology. American Journal of Botany, vol. 99no. 2, 291–311.

Cronn, R., Knaus, B. J., Liston, A., Maughan, P. J., Parks, M., Syring, J. V., & Udall, J. (2012b). Targeted enrichment strategies for next-generation plant biology. American journal of botany, vol. 99no. 2, 291–311.

Dahlgren, T., Lundberg, J., Pleijel, F., & Sundberg, P. (2000). Morphological and molecular evidence of the phylogeny of nereidiform polychaetes (annelida).

Journal of Zoological Systematics and Evolutionary Research, vol. 38no. 4, 249–253.

Darwin, C. (1859). The origin of species by means of natural selection (Vol. 247). EA Weeks.

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics, vol. 6no. 5, 361–375.

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., & Hahn, M. W. (2014). Extensive error in the number of genes inferred from draft genome assemblies. PLoS computational biology, vol. 10no. 12, e1003998.

Der Sarkissian, C., Allentoft, M. E., Ávila-Arcos, M. C., Barnett, R., Campos, P. F., Cappellini, E., Ermini, L., Fernández, R., Da Fonseca, R., Ginolhac, A., et al. (2015). Ancient genomics. Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 370no. 1660, 20130387.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature biotechnology, vol. 35no. 4, 316–319.

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). Novoplasty: De novo assembly of organelle genomes from whole genome data. Nucleic acids research, vol. 45no. 4, e18–e18.

Doležel, J., & Greilhuber, J. (2010). Nuclear genome size: Are we getting closer? Cytometry Part A, vol. 77no. 7, 635–642.

Dos Reis, M., Thawornwattana, Y., Angelis, K., Telford, M. J., Donoghue, P. C., & Yang, Z. (2015). Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. Current biology, vol. 25no. 22, 2939–2950.

Drennan, R., Wiklund, H., Rabone, M., Georgieva, M. N., Dahlgren, T. G., & Glover, A. G. (2021). Neanthes goodayi sp. nov.(annelida, nereididae), a remarkable new annelid species living inside deep-sea polymetallic nodules. European Journal of Taxonomy, vol. 760, 160–185.

Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, vol. 107no. 1, 1–15.

Elliott, T. A., & Gregory, T. R. (2015). What's in a genome? the c-value enigma and the evolution of eukaryotic genome content. Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 370no. 1678, 20140331.

Faino, L., & Thomma, B. P. (2014). Get your high-quality low-cost genome sequence. Trends in Plant Science, vol. 19no. 5, 288–291.

Faircloth, B. C. (2016). Phyluce is a software package for the analysis of conserved genomic loci. Bioinformatics, vol. 32no. 5, 786–788.

Fischer, A., & Dorresteijn, A. (2004). The polychaete platynereis dumerilii (annelida): A laboratory animal with spiralian cleavage, lifelong segment proliferation and a mixed benthic/pelagic life cycle. BioEssays, vol. 26no. 3, 314–325.

Fischer, A. H., Henrich, T., & Arendt, D. (2010). The normal development of platynereis dumerilii (nereididae, annelida). Frontiers in zoology, vol. 7no. 1, 1–39.

Fitzhugh, K. (1987). Phylogenetic relationships within the nereididae (polychaeta): Implications at the subfamily level. Bulletin of the Biological Society of Washington, no. 7, 174–183.

G. Ribeiro, P., Torres Jiménez, M. F., Andermann, T., Antonelli, A., Bacon, C. D., & Matos-Maravı, P. (2021). A bioinformatic platform to integrate target capture and whole genome sequences of various read depths for phylogenomics. Molecular Ecology, vol. 30no. 23, 6021–6035.

Gee, H. (2003). Ending incongruence. Nature, vol. 425no. 6960, 782–782.

Glasby, C. J. (1993). Family revision and cladistic analysis of the nereidoidea (polychaeta: Phyllodocida). Invertebrate Systematics, vol. 7no. 6, 1551–1573.

Glasby, C. (1991). Phylogenetic relationships in the nereididae (annelida: Polychaeta), chiefly in the subfamily gymnonereidinae, and the monophyly of the namanereidinae. Bulletin of Marine Science, vol. 48no. 2, 559–573.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Trinity: Reconstructing a full-length transcriptome without a genome from rna-seq data. Nature biotechnology, vol. 29no. 7, 644.

Gregory, T. R. (2002). Animal genome size database. http://www. genomesize. com.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). Quast: Quality assessment tool for genome assemblies. Bioinformatics, vol. 29no. 8, 1072–1075.

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic acids research, vol. 41no. 13, e129–e129.

Halanych, K. M. (2004). The new view of animal phylogeny. Annual Review of Ecology, Evolution, and Systematics, 229–256.

Hassanin, A., Delsuc, F., Ropiquet, A., Hammer, C., Van Vuuren, B. J., Matthee, C., Ruiz-Garcia, M., Catzeflis, F., Areskoug, V., Nguyen, T. T., et al. (2012). Pattern and timing of diversification of cetartiodactyla (mammalia, laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. Comptes rendus biologies, vol. 335no. 1, 32–50.

Hawkins, M. T., Hofman, C. A., Callicrate, T., McDonough, M. M., Tsuchiya, M. T., Gutiérrez, E. E., Helgen, K. M., & Maldonado, J. E. (2016). In-solution hybridization for mammalian mitogenome enrichment: Pros, cons and challenges associated with multiplexing degraded dna. Molecular Ecology Resources, vol. 16no. 5, 1173–1188.

Hayward, P. J., & Ryland, J. S. (2017). Handbook of the marine fauna of north-west europe. Oxford University Press.

Heath, T. A., Hedtke, S. M., & Hillis, D. M. (2008). Taxon sampling and the accuracy of phylogenetic analyses. Journal of Systematics and Evolution, 19.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing dna. Genomics, vol. 107no. 1, 1–8.

Hebert, P. D., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through dna barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences, vol. 270no. 1512, 313–321.

Helsem, S. A. (2021). Divergence time estimates for several phylogenies of annelida (lophotrochozoa) (Master's thesis).

Henikoff, J. G., & Henikoff, S. (1996). [6] Blocks database and its applications. Methods in Enzymology (pp. 88–105). Elsevier.

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). Ufboot2: Improving the ultrafast bootstrap approximation. Molecular biology and evolution, vol. 35no. 2, 518–522.

Hofreiter, M. (2012). Nondestructive dna extraction from museum specimens. Ancient dna (pp. 93–100). Springer.

Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J. W., Landolin, J. M., Maurer, N., Kudrna, D., Hardigan, M. A., Steiner, C. C., et al. (2020). Highly accurate long-read hifi sequencing data for five complex genomes. Scientific data, vol. 7no. 1, 1–11.

Jeffroy, O., Brinkmann, H., Delsuc, F., & Philippe, H. (2006). Phylogenomics: The beginning of incongruence? TRENDS in Genetics, vol. 22no. 4, 225–231.

Jennings, R. M., & Halanych, K. M. (2005). Mitochondrial genomes of clymenella torquata (maldanidae) and riftia pachyptila (siboglinidae): Evidence for conserved gene order in annelida. Molecular Biology and Evolution, vol. 22no. 2, 210–222.

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., DePamphilis, C. W., Yi, T.-S., & Li, D.-Z. (2020). Getorganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome biology, vol. 21no. 1, 1–31.

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. Molecular ecology, vol. 25no. 1, 185–202.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., & Jermiin, L. S. (2017). Modelfinder: Fast model selection for accurate phylogenetic estimates. Nature methods, vol. 14no. 6, 587–589.

Kara, J., Santos, C. S., Macdonald, A. H., & Simon, C. A. (2020). Resolving the taxonomic identities and genetic structure of two cryptic platynereis kinberg species from south africa. Invertebrate Systematics, vol. 34no. 6, 618–636.

Katoh, K., Asimenos, G., & Toh, H. (2009). Multiple alignment of dna sequences with mafft. Bioinformatics for dna sequence analysis (pp. 39–64). Springer.

Katoh, K., & Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: Improvements in performance and usability. Molecular biology and evolution, vol. 30no. 4, 772–780.

Keilwagen, J., Hartung, F., & Grau, J. (2019). Gemoma: Homology-based gene prediction utilizing intron position conservation and rna-seq data. Gene prediction (pp. 161–177). Springer.

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O., & Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinformatics, vol. 19no. 1, 189.

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. Nucleic Acids Research, vol. 44no. 9, e89–e89.

Knyshov, A., Gordon, E. R., & Weirauch, C. (2021a). New alignment-based sequence extraction software (ALiBaSeq) and its utility for deep level phylogenetics. PeerJ, vol. 9, e11019.

Knyshov, A., Gordon, E. R., & Weirauch, C. (2021b). New alignment-based sequence extraction software (alibaseq) and its utility for deep level phylogenetics. PeerJ, vol. 9, e11019.

Kocot, K. M., Struck, T. H., Merkel, J., Waits, D. S., Todt, C., Brannock, P. M., Weese, D. A., Cannon, J. T., Moroz, L. L., Lieb, B., et al. (2017). Phylogenomics of lophotrochozoa with consideration of systematic error. Systematic biology, vol. 66no. 2, 256–282.

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics, vol. 28no. 19, 2520–2522.

Kucuk, E., Chu, J., Vandervalk, B. P., Hammond, S. A., Warren, R. L., & Birol, I. (2017). Kollector: Transcript-informed, targeted de novo assembly of gene loci. Bioinformatics, vol. 33no. 12, 1782–1788.

Lassen, B. et al. (2016). The two worlds of nagoya: Abs legislation in the eu and provider countries, discrepancies and how to deal with them. Cape Town: Public Eye, Zurich and Natural Justice.

Laumer, C. E., Bekkouche, N., Kerbl, A., Goetz, F., Neves, R. C., Sørensen, M. V., Kristensen, R. M., Hejnol, A., Dunn, C. W., Giribet, G., et al. (2015). Spiralian phylogeny informs the evolution of microscopic lineages. Current Biology, vol. 25no. 15, 2000–2006.

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012a). Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Systematic Biology, vol. 61no. 5, 727–744.

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012b). Anchored hybrid enrichment for massively high-throughput phylogenomics. Systematic biology, vol. 61no. 5, 727–744.

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. Annual Review of Ecology, Evolution, and Systematics, vol. 44no. 1, 99–121.

Letunic, I., & Bork, P. (2019). Interactive tree of life (itol) v4: Recent updates and new developments. Nucleic acids research, vol. 47no. W1, W256–W259.

Levy, S. E., & Myers, R. M. (2016). Advancements in next-generation sequencing. Annu Rev Genomics Hum Genet, vol. 17no. 1, 95–115.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., & Lam, T.-W. (2016). Megahit v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods, vol. 102, 3–11.

Li, T., Liu, D., Yang, Y., Guo, J., Feng, Y., Zhang, X., Cheng, S., & Feng, J. (2020). Phylogenetic supertree reveals detailed evolution of sars-cov-2. Scientific reports, vol. 10no. 1, 1–9.

Li, W., Cong, Q., Shen, J., Zhang, J., Hallwachs, W., Janzen, D. H., & Grishin, N. V. (2019). Genomes of skipper butterflies reveal extensive convergence of wing patterns. Proceedings of the National Academy of Sciences, vol. 116no. 13, 6232–6237.

Liu, M., Liu, H., Wang, Q., Guan, S., & Ge, S. (2013). Phylogenetic relationships of twenty-one nereids species inferring two different evolutionary origins. Aquatic Science and Technology, vol. 1, 167–180.

Lynch, M. (2007). The origins of genome architecture (sinauer, sunderland, ma), p 494.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., & Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. Nature methods, vol. 7no. 2, 111–118.

Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A., & Zdobnov, E. M. (2021). Busco update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. arXiv preprint arXiv:2106.11799.

Martın-Durán, J. M., Vellutini, B. C., Marlétaz, F., Cetrangolo, V., Cvetesic, N., Thiel, D., Henriet, S., Grau-Bové, X., Carrillo-Baltodano, A. M., Gu, W., et al. (2021). Conservative route to genome compaction in a miniature annelid. Nature ecology & evolution, vol. 5no. 2, 231–242.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., & Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics [Morris Goodman Memorial Symposium]. Molecular Phylogenetics and Evolution, vol. 66no. 2, 526–538.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution, vol. 37no. 5, 1530–1534.

Neumann, D., Borisenko, A., Coddington, J., Häuser, C., Butler, C., Casino, A., Vogel, J., Haszprunar, G., & Giere, P. (2018). Global biodiversity research tied up by juridical interpretations of access and benefit sharing. Organisms Diversity & Evolution, vol. 18no. 1, 1–12.

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., Prjibelsky, A., Pyshkin, A., Sirotkin, A., Sirotkin, Y., et al. (2013). Assembling genomes and mini-metagenomes from highly chimeric reads. Annual International Conference on Research in Computational Molecular Biology, 158–170.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). Metaspades: A new versatile metagenomic assembler. Genome research, vol. 27no. 5, 824–834.

of Life Project Consortium, D. T. (2022). Sequence locally, think globally: The darwin tree of life project. Proceedings of the National Academy of Sciences, vol. 119no. 4, e2115642118.

of Scientists, G. 1. C. (2009). Genome 10k: A proposal to obtain whole-genome sequence for 10 000 vertebrate species. Journal of Heredity, vol. 100no. 6, 659–674.

Özpolat, B. D., Randel, N., Williams, E. A., Bezares-Calderón, L. A., Andreatta, G., Balavoine, G., Bertucci, P. Y., Ferrier, D. E., Gambi, M. C., Gazave, E., et al. (2021). The nereid on the rise: Platynereis as a model system. EvoDevo, vol. 12no. 1, 1–22.

Park, T., Lee, S.-H., & Kim, W. (2016). Complete mitochondrial genome sequence of the giant mud worm paraleonnates uschakovi khlebovich & wu, 1962 (polychaeta: Nereididae). Mitochondrial DNA Part B, vol. 1no. 1, 640–642.

Peng, Y., Leung, H., Yiu, S.-M., & Chin, F. Y. (2010). Idba–a practical iterative de bruijn graph de novo assembler. Annual international conference on research in computational molecular biology, 426–440.

Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., & Pupko, T. (2010). Guidance: A web server for assessing alignment confidence scores. Nucleic acids research, vol. 38no. suppl_2, W23–W28.

Pfenninger, M., Schönnenbeck, P., & Schell, T. (2022). Modest: Accurate estimation of genome size from next generation sequencing data. Molecular ecology resources, vol. 22no. 4, 1454–1464.

Pflug, J. M., Holmes, V. R., Burrus, C., Johnston, J. S., & Maddison, D. R. (2020). Measuring genome sizes using read-depth, k-mers, and flow cytometry: Methodological comparisons in beetles (coleoptera). G3: Genes, Genomes, Genetics, vol. 10no. 9, 3047–3060.

Philippe, H., de Vienne, D. M., Ranwez, V., Roure, B., Baurain, D., & Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. European Journal of Taxonomy, no. 283.

Philippe, H., Brinkmann, H., Copley, R. R., Moroz, L. L., Nakano, H., Poustka, A. J., Wallberg, A., Peterson, K. J., & Telford, M. J. (2011). Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature, vol. 470no. 7333, 255–258.

Phillips, M. J., Delsuc, F., & Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. Molecular biology and evolution, vol. 21no. 7, 1455–1458.

Pick, K., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D., Wrede, P., Wiens, M., Alié, A., Morgenstern, B., Manuel, M., et al. (2010). Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Molecular biology and evolution, vol. 27no. 9, 1983–1987.

Pillai, T. G. (1961). Annelida polychaeta of tambalagam lake, ceylon. Ceylon Journal of Science (Biological Sciences), vol. 4no. 1, 1–40.

Pleijel, F., & Dahlgren, T. (1998). Position and delineation of chrysopetalidae and hesionidae (annelida, polychaeta, phyllodocida). Cladistics, vol. 14no. 2, 129–150.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). Fasttree 2–approximately maximum-likelihood trees for large alignments. PloS one, vol. 5no. 3, e9490.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., & Korobeynikov, A. (2020). Using spades de novo assembler. Current protocols in bioinformatics, vol. 70no. 1, e102.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms: Comparison of ion torrent, pacific biosciences and illumina miseq sequencers. BMC genomics, vol. 13no. 1, 1–13.

Ratnasingham, S., & Hebert, P. D. (2007). Bold: The barcode of life data system (http://www. barcodinglife. org). Molecular ecology notes, vol. 7no. 3, 355–364.

Read, G., & Fauchald, K. (2020). World polychaeta database. Pseudopotamilla laciniosa.

Read, G., & Fauchald, K. (2021). World polychaeta database. nereididae blainville, 1818.

Read, G., Fauchald, K., & Glasby, C. (2012). Nereididae. Read G, Fauchald, K.(2012). World Polychaeta database. Accessed through: World Registe

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature, vol. 592no. 7856, 737–746.

Rice, P., Longden, I., & Bleasby, A. (2000). Emboss: The european molecular biology open software suite. Trends in genetics, vol. 16no. 6, 276–277.

Richter, S., Schwarz, F., Hering, L., Böggemann, M., & Bleidorn, C. (2015). The utility of genome skimming for phylogenomic analyses as demonstrated for glycerid relationships (annelida, glyceridae). Genome Biology and Evolution, vol. 7no. 12, 3443–3462.

Robin, E. D., & Wong, R. (1988). Mitochondrial dna molecules and virtual number of mitochondria per cell in mammalian cells. Journal of cellular physiology, vol. 136no. 3, 507–513.

Rogozin, I. B., Sverdlov, A. V., Babenko, V. N., & Koonin, E. V. (2005). Analysis of evolution of exon-intron structure of eukaryotic genes. Briefings in Bioinformatics, vol. 6no. 2, 118–134.

Rokas, A., Williams, B. L., King, N., & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature, vol. 425no. 6960, 798–804.

Ruby, J. G., Bellare, P., & DeRisi, J. L. (2013). Price: Software for the targeted assembly of components of (meta) genomic sequence data. G3: Genes, Genomes, Genetics, vol. 3no. 5, 865–880.

Salzberg, S. L. (2019). Next-generation genome annotation: We still struggle to get it right. Genome Biology, vol. 20no. 1, 92.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., Marçais, G., Pop, M., & Yorke, J. A. (2012). Gage: A critical evaluation of genome assemblies and assembly algorithms. Genome Research, vol. 22no. 3, 557–567.

Sann, M., Niehuis, O., Peters, R. S., Mayer, C., Kozlov, A., Podsiadlowski, L., Bank, S., Meusemann, K., Misof, B., Bleidorn, C., & Ohl, M. (2018). Phylogenomic analysis of Apoidea sheds new light on the sister group of bees. BMC Evolutionary Biology, vol. 18no. 1, 71.

Santos, C. S., Pleijel, F., Lana, P., & Rouse, G. W. (2006). Phylogenetic relationships within nereididae (annelida: Phyllodocida). Invertebrate Systematics, vol. 19no. 6, 557–576.

Sarmashghi, S., Balaban, M., Rachtman, E., Touri, B., Mirarab, S., & Bafna, V. (2021). Estimating repeat spectra and genome length from low-coverage genome skims with respect. PLoS computational biology, vol. 17no. 11, e1009449.

Schwarz, J. M., Lüpken, R., Seelow, D., & Kehr, B. (2021). Novel sequencing technologies and bioinformatic tools for deciphering the non-coding genome. Medizinische Genetik, vol. 33no. 2, 133–145.

Sela, I., Ashkenazy, H., Katoh, K., & Pupko, T. (2015). Guidance2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic acids research, vol. 43no. W1, W7–W14.

Shen, W., Le, S., Li, Y., & Hu, F. (2016). Seqkit: A cross-platform and ultrafast toolkit for fasta/q file manipulation. PloS one, vol. 11no. 10, e0163962.

Shen, X.-X., Zhou, X., Kominek, J., Kurtzman, C. P., Hittinger, C. T., & Rokas, A. (2016). Reconstructing the backbone of the saccharomycotina yeast phylogeny using genome-scale data. G3: Genes, Genomes, Genetics, vol. 6no. 12, 3927–3939.

Sherman, B., & Henry, R. J. (2020). The nagoya protocol and historical collections of plants. Nature Plants, vol. 6no. 5, 430–432.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). Busco: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, vol. 31no. 19, 3210–3212.

Simon, C. A., Muthumbi, A. W., Kihia, C. M., Smith, K. M., Cedras, R. B., Mahatante, P. T., Wangondu, V. W., & Katikiro, R. (2021). A review of marine invertebrates used as fishing baits and the implications for national and regional management in the western indian ocean. African Zoology, vol. 56no. 4, 237–263.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). Abyss: A parallel assembler for short read sequence data. Genome research, vol. 19no. 6, 1117–1123.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. Nature Reviews Genetics, vol. 15no. 2, 121–132.

Soltis, D. E., Albert, V. A., Savolainen, V., Hilu, K., Qiu, Y.-L., Chase, M. W., Farris, J. S., Stefanović, S., Rice, D. W., Palmer, J. D., et al. (2004). Genome-scale data, angiosperm relationships, and 'ending incongruence': A cautionary tale in phylogenetics. Trends in plant science, vol. 9no. 10, 477–483.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). Augustus: Ab initio prediction of alternative transcripts. Nucleic acids research, vol. 34no. suppl_2, W435–W439.

Stanke, M., & Morgenstern, B. (2005). Augustus: A web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic acids research, vol. 33no. suppl_2, W465–W467.

Struck, T. H. (2011). Direction of evolution within annelida and the definition of pleistoannelida. Journal of Zoological Systematics and Evolutionary Research, vol. 49no. 4, 340–345.

Struck, T. H., Golombek, A., Weigert, A., Franke, F. A., Westheide, W., Purschke, G., Bleidorn, C., & Halanych, K. M. (2015). The evolution of annelids reveals two adaptive routes to the interstitial realm. Current Biology, vol. 25no. 15, 1993–1999.

Thalen, F., Koehne, C. G., & Bleidorn, C. (2022). Patchwork: Alignment-based retrieval and concatenation of phylogenetic markers from genomic data. bioRxiv.

Thomas Jr, C. A. (1971). The genetic organization of chromosomes. Annual review of genetics, vol. 5no. 1, 237–256.

Tosuji, H., Bastrop, R., Götting, M., Park, T., Hong, J.-S., & Sato, M. (2019). Worldwide molecular phylogeny of common estuarine polychaetes of the genus hediste (annelida: Nereididae), with special reference to interspecific common haplotypes found in southern japan. Marine Biodiversity, vol. 49no. 3, 1385–1402.

Turner, E. H., Ng, S. B., Nickerson, D. A., & Shendure, J. (2009). Methods for genomic partitioning. Annual review of genomics and human genetics, vol. 10no. 1, 263–284.

Van Bruggen, E., Borst, P., Ruttenberg, G., Gruber, M., & Kroon, A. (1966). Circular mitochondrial dna. Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis, vol. 119no. 2, 437–439.

Viéitez, J. M. (2004). Fauna ibérica. vol. 25. annelida: Polychaeta i (Vol. 25). Editorial CSIC-CSIC Press.

Villalobos-Guerrero, T. F., Kara, J., Simon, C., & Idris, I. (2022). Systematic review of neanthes kinberg, 1865 (annelida: Errantia: Nereididae) from southern africa, including a preliminary molecular phylogeny of the genus. Marine Biodiversity, vol. 52no. 2, 1–30.

Virgilio, M., Fauvelot, C., Costantini, F., Abbiati, M., & Backeljau, T. (2009). Phylogeography of the common ragworm Hediste diversicolor (Polychaeta: Nereididae) reveals cryptic diversity and multiple colonization events across its distribution. Molecular Ecology, vol. 18no. 9, 1980–1994.

Wang, Z., Gerstein, M., & Snyder, M. (2009). Rna-seq: A revolutionary tool for transcriptomics. Nature reviews genetics, vol. 10no. 1, 57–63.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). Busco applications from quality assessments to gene prediction and phylogenomics. Molecular biology and evolution, vol. 35no. 3, 543–548.

Weigert, A., Golombek, A., Gerth, M., Schwarz, F., Struck, T. H., & Bleidorn, C. (2016). Evolution of mitochondrial gene order in annelida. Molecular Phylogenetics and Evolution, vol. 94, 196–206.

Weigert, A., Helm, C., Meyer, M., Nickel, B., Arendt, D., Hausdorf, B., Santos, S. R., Halanych, K. M., Purschke, G., Bleidorn, C., & Struck, T. H. (2014). Illuminating the Base of the Annelid Tree Using Transcriptomics. Molecular Biology and Evolution, vol. 31no. 6, 1391–1401.

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., & Au, K. F. (2017). Comprehensive comparison of pacific biosciences and oxford nanopore technologies and their applications to transcriptome analysis. F1000Research, vol. 6.

Yin, Z.-T., Zhu, F., Lin, F.-B., Jia, T., Wang, Z., Sun, D.-T., Li, G.-S., Zhang, C.-L., Smith, J., Yang, N., et al. (2019). Revisiting avian 'missing'genes from de novo assembled transcripts. Bmc Genomics, vol. 20no. 1, 1–10.

Zedane, L., Hong-Wa, C., Murienne, J., Jeziorski, C., Baldwin, B. G., & Besnard, G. (2016). Museomics illuminate the history of an extinct, paleoendemic plant lineage (hesperelaea, oleaceae) known from an 1875 collection from guadalupe island, mexico. Biological Journal of the Linnean Society, vol. 117no. 1, 44–57.

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). Astral-iii: Polynomial time species tree reconstruction from partially resolved gene trees. BMC bioinformatics, vol. 19no. 6, 15–30.

Zhang, F., Ding, Y., Zhu, C.-D., Zhou, X., Orr, M. C., Scheu, S., & Luan, Y.-X. (2019a). Phylogenomics from low-coverage whole-genome sequencing

(M. Matschiner, Ed.). Methods in Ecology and Evolution, vol. 10no. 4, 507–517.

Zhang, F., Ding, Y., Zhu, C.-D., Zhou, X., Orr, M. C., Scheu, S., & Luan, Y.-X. (2019b). Phylogenomics from low-coverage whole-genome sequencing. Methods in Ecology and Evolution, vol. 10no. 4, 507–517.

Zhang, G. (2015). Bird sequencing project takes off. Nature, vol. 522no. 7554, 34–34.

Zhong, M., Struck, T. H., & Halanych, K. M. (2008). Phylogenetic information from three mitochondrial genomes of terebelliformia (annelida) worms and duplication of the methionine trna. Gene, vol. 416no. 1-2, 11–21.

Zwickl, D. J., & Hillis, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. Systematic biology, vol. 51no. 4, 588–598.