A genome resource for the marine annelid *Platynereis dumerilii*

Kevin Nzumbi Mutemi, Oleg Simakov*, Leslie Pan, Luca Santangeli, Ryan Null, Mette Handberg-Thorsager, Bruno Cossermelli Vellutini, Tomas Larsson, Emily Savage, Mireia Osuna Lopez, Rajna Hercog, Jan Provaznik, Diana Ordoñez-Rueda, Nayara Azevedo, Eve Gazave, Michel Vervoort †, Pavel Tomancak, Wenhua Tan, Sylke Winkler, Vladimir Benes, Jerome Hui, Conrad Helm, B. Duygu Özpolat*, Detlev Arendt*

AUTHORS

Kevin Nzumbi Mutemi European Molecular Biology Laboratory, Developmental Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
kevin.mutemi@embl.de ORCID: 0000-0002-6426-2363

Oleg Simakov Department for Neurosciences and Developmental Biology, University of Vienna, Austria oleg.simakov@univie.ac.at orcid: 0000-0002-3585-4511

Leslie Pan European Molecular Biology Laboratory, Developmental Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
leslie.pan@embl.de ORCID: 0000-0002-2408-4259

Luca Santangeli European Molecular Biology Laboratory, Developmental Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
luca.santangeli@embl.de ORCID: 0000-0001-9600-6164

Ryan W. Null Current: Department of Biology, Washington University in Saint Louis, MO, USA 63139 Previous: Eugene Bell Center for Regenerative Biology and Tissue Engineering, Marine Biological Laboratory, Woods Hole, MA. 02543. USA. nullr@wustl.edu
https://orcid.org/0000-0002-3830-4152

Mette Handberg-Thorsager Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, 01307 Dresden, Germany. metteht@gmail.com. ORCID 0000-0002-3908-7233

Bruno Cossermelli Vellutini - Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, 01307 Dresden, Germany. vellutini@mpi-cbg.de ORCID: 0000-0002-0000-9465

Tomas Larsson Tomas.Larsson@icm.uu.se Developmental Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. Current: Department of Cell and Molecular Biology, Uppsala University, Sweden

Emily L. Savage - emily.savage@cos.uni-heidelberg.de. European Molecular Biology Laboratory, Developmental Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany. Current: Centre for Organismal Studies, University of Heidelberg, Heidelberg. ORCID: 0000-0003-3391-7571

Mireia Osuna Lopez - European Molecular Biology Laboratory, Genomics Core Facility, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

mireia.osuna@embl.de ORCID: 0000-0001-7724-444X

Rajna Hercog - European Molecular Biology Laboratory, Genomics Core Facility, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
rajna.hercog@embl.de - not at EMBL anymore, do not know current address

Jan Provaznik - European Molecular Biology Laboratory, Genomics Core Facility, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
jan.provaznik@embl.de

Diana Ordoñez-Rueda European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. diana.ordonez.ics@gmail.com - ORCID: 0000–0002-8387-6707

Nayara Azevedo - European Molecular Biology Laboratory, Genomics Core Facility, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
nayara.azevedo@embl.de - not at EMBL anymore, do not know current address

Eve Gazave Université Paris Cité, CNRS, Institut Jacques Monod, F-75013, Paris, France. eve.gazave@ijm.fr ORCID 0000-0002-4647-6640

Michel Vervoort Université Paris Cité, CNRS, Institut Jacques Monod, F-75013, Paris, France. ORCID 0000-0002-5995-2140 - deceased

Pavel Tomancak - Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, 01307 Dresden, Germany. tomancak@mpi-cbg.de ORCID 0000-0002-2222-9370

Wenhua Tan - Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, 01307 Dresden, Germany. tan@mpi-cbg.de

Sylke Winkler - Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, 01307 Dresden, Germany. winkler@mpi-cbg.de ORCID: 0000-0002-0915-3316

Vladimir Benes - EMBL, Genomics Core Facility, Meyerhofstr. 1, 69117 Heidelberg, Germany benes@embl.de ORCID: 0000-0002-0352-2547

Jerome Hui - School of Life Sciences, Simon F.S. Li Marine Science Laboratory, State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong
jeromehui@cuhk.edu.hk 0000-0003-1355-8495

Conrad Helm - Animal Evolution and Biodiversity, University of Göttingen, Untere Karspüle 2, 37073 Göttingen, Germany. chelm@uni-goettingen.de
ORCID: 0000-0002-0308-8402

B. Duygu Özpolat Current: Department of Biology, Washington University in Saint Louis, MO, USA 63139 Previous: Eugene Bell Center for Regenerative Biology and Tissue Engineering, Marine Biological Laboratory, Woods Hole, MA. 02543. USA. bdozpolat@wustl.edu
https://orcid.org/0000-0002-1900-965X

Detlev Arendt European Molecular Biology Laboratory, Developmental Biology Unit, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
Centre for Organismal Studies (COS) University of Heidelberg, 69120 Heidelberg, Germany
arendt@embl.de ORCID: 0000-0001-7833-050X

Funding
BDÖ: NIH NIGMS MIRA 1R35GM138008-01; Hibbitt Startup Funds, DoveTail (Cantata Bio) for genome sequencing expenses
EG & MV: Labex "Who Am I" laboratory of excellence (No. ANR- 11-LABX-0071); ANR-19-CE27-0027-01
MHT, BCV, PT: DFG GZ: TO 563/7-1
BCV: EMBO Long Term Fellowship (ALTF 74-2018)
CH: German Research Foundation (DFG) (HE 7224/8-1)
KNM, LS, DA: Marie Skłodowska-Curie ITN "EvoCELL" #766053
LP, LS, DA: ERC Advanced grant NeuralCellTypeEvo #788921

Correspondence: arendt@embl.de, oleg.simakov@univie.ac.at, bdozpolat@wustl.edu

Abstract

The marine annelid *Platynereis dumerilii* is a model organism used in many research areas including evolution and development, neurobiology, ecology and regeneration. Here we present the genomes of *P. dumerilii* and of the closely related *P. massiliensis* and *P. megalops*, to facilitate comparative genomic approaches and help explore *Platynereis* biology. We used long-read sequencing technology and chromosomal-conformation capture along with extensive transcriptomic resources to obtain and annotate a draft genome assembly of ~1.47 Gbp for *P. dumerilii*, of which more than half represent repeat elements. We predict around 29,000 protein-coding genes, with relatively large intron sizes, over 38,000 non-coding genes, and 580 miRNA loci. We further explore the high genetic variation (~3% heterozygosity) within the *Platynereis* species complex. Gene ontology reveals the most variable loci to be associated with pigmentation, development and immunity. The current work sets the stage for further development of *Platynereis* genomic resources.

***Keywords*:** annelid, evo-devo, genome, model organism, *Platynereis*, *Platynereis dumerilii*, Spiralia

Introduction

Platynereis dumerilii (Audouin and Milne Edwards 1833) is found along European coastlines, including from the Azores, Mediterranean, North, Black and Red Seas, English Channel, as well as Atlantic and Pacific, Sea of Japan and Persian Gulf (Fauvel, 1914; Teixeira et al., 2022) (Fig. 1A). As a laboratory model species, it has been extensively studied in disciplines spanning ecology, behavior, physiology, development, evolution, regeneration and neurobiology (Fischer et al., 2010; Fischer & Dorresteijn, 2004; Özpolat et al., 2021) (Fig. 1B). However, an annotated genome has been challenging to achieve, due to high polymorphism, heterozygosity and repetitive content (Raible et al., 2005; Zantke et al., 2014). Here, we report the generation of a comprehensive genome resource for *P. dumerilii*, which also builds on previously published extensive transcriptomic data (Chou et al., 2018; Conzelmann et al., 2013; Raible et al., 2005; Zantke et al., 2014; Paré et al 2023). Utilizing long-read sequencing, dense transcriptomic sampling and high-throughput sequencing of chromosome conformation capture (Hi-C) data (Belton et al., 2012), we present a draft assembly and annotation of the *P. dumerilii* genome measuring ~1.47 giga-base pairs (Gbp) in size, which is larger than other annelid genomes sequenced and annotated to date (Liang et al., 2022; Martín-Durán et al., 2021; Simakov et al., 2013; Shao et al., 2020; Zakas et al., 2021) (Fig. 1C). We find that at least 51% of the genome consists of repetitive regions, and model around 29,000 protein-coding genes, with median intron sizes of ~1.3 kilo-base pairs (Kbp). This suggests that an increase in repeat content, intron sizes as well as modest gene duplication contributed to genome expansion. We explore evolutionary trends in genome organization, ancestral linkage groups, and gene content across Metazoa, focusing on comparisons across annelids and within the *Platynereis* genus. We also describe natural polymorphisms associated with habitats that *Platynereis* occupies globally. Finally, we present draft genomes of two sister species, *P. massiliensis* and *P. megalops,* which appear morphologically indistinguishable from *P. dumerilii* as juveniles, but present different reproductive, larval, and behavioral patterns.

Results and Discussion

Towards a chromosome-level assembly of the *P. dumerilii* genome

We sequenced high molecular weight genomic DNA (Supplementary Figs. 1-2), accessed from hundreds of progeny obtained from a single cross (one male with one female), from a culture bred in laboratory conditions since the 1960s (Fischer & Dorresteijn, 2004). The culture that was used for sequencing has been in-bred for several generations. We utilized PacBio Sequel II continuous long read (CLR) technology, sequencing at 200GB (corresponding roughly to 200x coverage, assuming ~1 Gbp genome size (Jha et al., 1995) (Materials and Methods; Supplementary Fig. 1). Read lengths ranged from 50 – 270,934 bp with a median of 48,761 bp (mean $\pm$ standard deviation [SD] = 49,814.5 $\pm$ 30,070.61 bp) (Supplementary Fig. 2). From these reads, we assembled the genome using CANU (Koren et al., 2017), yielding a ~3.41 Gbp assembly size with 9,431 contigs, and an N50 of 640 Kbp (Methods). At this state however, the genome size amounted to more than three times its previously predicted size (Jha et al., 1995). We re-evaluated the genome size, as well as ploidy, using GenomeScope2.0 and Smudgeplot (Ranallo-Benavidez et al., 2020) from high quality 150bp Illumina sequencing reads as well as DNA content quantified via flow cytometry. While these results estimated a genome size of ~940 Mbp (Supplementary Fig. 3), flow cytometry measurements suggested a genome size ranging between ~927 Mbp and ~1.2 Gbp (data not shown), when cross-referenced with the known *Drosophila melanogaster* genome size of ~220 Mbp (Bosco et al., 2007; Hjelmen et al., 2019), consistent with previous estimates of the *P. dumerilii* genome size (Jha et al., 1995). We thus reasoned that the large genome assembly size of 3.41 Gbp likely reflected a high number of recurrent contigs within the assembly.

To test this, we searched for universal single-copy gene orthologs in the assembly using BUSCO (Simão et al., 2015), and found that a large portion of the assembly contained recurrent and/or duplicate sequences i.e. BUSCO analysis with a metazoa database of 954 genes revealed 96.4% complete genes of which only 20.4% were single-copy and 76.0% duplicate, with 2.3% fragmented. We then identified and

separated out redundant contigs and probable haplotypes in the assembly as per Guan and colleagues (Guan et al., 2020) and Roach and colleagues (Roach et al., 2018) (Methods and Materials). Iterative 'purging' (i.e. removal of likely redundant sequences/contigs within the assembly) via the purge_dups and purge_haplotigs algorithms, significantly decreased duplicate contigs (and haplotypes) in the assembly, while retaining similar BUSCO gene completion scores with 95.4% complete (89.0% single-copy, 6.4% duplicate) and 3.1% fragmented gene models. To further eliminate likely redundant contigs in this assembly, we filtered out sequences less than 200 Kbp (Mutemi, 2023). This decreased BUSCO genome completion scores by 1.6% (93.8%), and the number of duplicate sequences by 3.5% (90.9% single-copy, 2.9% duplicate) and 3.1% fragmented. Notably, the initial assembly size decreased from ~3.41 to ~1.47 Gbp, with the number of contigs also decreasing from the original 9,431 to 964; with an increase in N50 from 640 Kbp to ~4.3 Mbp.

We then scaffolded the assembly using both long-reads and HiC data. Firstly, we used LINKS (Warren et al., 2015), an algorithm that relies on iterative k-mer pair matching over varying sequence lengths present in the raw long-read sequencing data and the assembly, to then build links between contigs. This approach scaffolded our assembly from 964 contigs to 647 scaffolds, increasing the N50 from 4.3 to 7.95 Mbp. We then polished the genome via POLCA (Zimin & Salzberg, 2020), using dense (~100X coverage assuming a ~1 Gbp genome size), high-quality accurate Illumina ~150bp paired-end sequenced reads (Methods and Materials), returning a consensus assembly quality value of 99.74%. Next, we made use of Hi-C mapping and scaffolding via the SALSA2 pipeline (Ghurye et al., 2019), resulting in 330 scaffolds with an N50 of 54.8 Mbp (Fig. 2). *P. dumerilii* is thought to have 14 pairs of chromosomes (2n=28), as measured by C-banding and silver staining methods (Jha et al., 1995). 50% of the total assembly is represented by 8 scaffolds (Fig. 2; Supplementary Fig. 4).

P. dumerilii genome annotation

Repeat elements

The *P. dumerilii* genome is among the largest assembled and annotated annelid genomes to date (Bhambri et al., 2018; Liang et al., 2022; Martín-Durán et al., 2021; Simakov et al., 2013; Zakas et al., 2021; Zwarycz et al., 2015) (Fig. 1C). Repeat elements (REs) comprise significant portions of animal genomes, and are thought to be major drivers of genome size evolution (López-Flores & Garrido-Ramos, 2012). Recent studies in the marine annelids *Dimorphilus gyrociliatus* (Martín-Durán et al., 2021) and *Streblospio benedictii* (Zakas et al., 2021) provided evidence supporting the hypothesis that the proportion of REs scale with genome size; *D. gyrociliatus* reported to have a genome size of ~71 Mbp, with REs accounting for ~8% of its genome content (Martín-Durán et al., 2021), in contrast to the ~700 Mbp genome of *S. benedictii* and the ~519 Mbp *Owenia fusiformis* genome revealing ~43% RE content (Liang et al., 2022; Zakas et al., 2021).

We identified and modeled REs using RepeatModeler (Flynn et al., 2020), and masked the genome using RepeatMasker (Tarailo-Graovac & Chen, 2009) (Materials and Methods). We then estimated ~51% of the *P. dumerilii* genome to be composed of REs (Fig. 3A; Table 1), listing as the annelid with the largest proportion of genomic RE-content sequenced and annotated thus far (Fig. 3B). Of the known REs, retroelements comprise the largest proportion, occupying 14.62% of the *P. dumerilii* genome, with low-complexity REs only making up 0.11% of the total genome (Fig. 3A-B). This was consistent also across the five annelid genomes studied. In *H. robusta* and *D. gyrociliatus* however, simple repeats make up an equally high fraction of REs (*H. robusta*; ~10.74% retroelements and ~9.67% simple repeats, and *D. gyrociliatus*; ~2.65% retroelements and 2.5% simple repeats; Fig. 3B). Across all annelids explored however, the 'unclassified' REs made up the bulk of repetitive sequences found in the genome. Despite comprising such a large fraction of the genome (particularly in the larger *O. fusiformis*, *S. benedictii* and *P. dumerilii* assemblies) very little is known about these Unclassified REs. We attempted to group

‘unclassified/unknown’ REs found in annelids and found very little similarity across all five species, with these elements clustering in a species-specific manner (Supplementary Fig. 5). This finding suggests that the bulk of REs found within annelid genomes are likely lineage- and/or species-specific.

P. dumerilii RE sizes varied significantly according to the type of REs (Table 2; Supplementary Fig. 6A) (Kruskal-Wallis test, df = 11, Benjamini and Hochberg adjusted *p-value* < *2.2e-16*). To explore whether the number of REs as well as their lengths partially explain genome size expansion or compaction, we also compared RE lengths across the five sequenced and annotated annelid genomes with *P. dumerilii* (Supplementary Fig. 6). We found that RE lengths were not always increased in larger genomes (Supplementary Fig. 6). For instance, *S. benedictii* unclassified REs (n = 1,155,954; median $\pm$ MAD = 126 $\pm$ 90.44 bp) were shorter than *O. fusiformis* (n = 772,141; median $\pm$ MAD = 147 $\pm$ 120.09 bp), despite the *S. benedictii* genome being larger (Supplementary Fig. 6B) (pairwise Wilcoxon rank sum test with continuity correction, *p-value* < *2e-16*). In *D. gyrociliatus,* DNA Retroelements (n = 4,001; median $\pm$ MAD = 201 $\pm$ 200) were longer than all other annelids except *C. teleta* (n = 23,545; median $\pm$ MAD = 207 $\pm$ 195.70; pairwise Wilcoxon rank sum test with continuity correction, *p-value = 0.024*) despite having the smallest genome (Supplementary Fig. 6B). While this may be impacted by the repeat identification and assembly quality, we conclude that it is the number of REs, rather than their lengths that best explain genome expansion in the annelid genomes analyzed thus far.

We also explored the distribution of REs across the genome and found that approximately 36% of the *P. dumerilii* REs have some overlap with protein-coding genes loci, mainly in the intronic regions, with the majority of other REs residing in intergenic regions (Fig. 3C). A similar analysis performed for the other annelid genomes showed consistently that REs predominantly occupy intergenic portions of the genome, with the major exception being *D. gyrociliatus* (Supplementary Fig. 5), showing ~91% of REs found within genic regions; likely due to increased compaction of its genome (Martín-Durán et al., 2021).

In *P. dumerilii* protein-coding genes (see below), we found that most genes had some overlap with REs, with only around 9% - 2,608 out 28,985 genes - having no overlap with annotated REs. The majority of the annotated *P. dumerilii* REs within gene loci reside in intronic regions (Fig. 3D), which partially explains the increased size of protein coding genes in *P. dumerilii* (see below). Overall, we observed that 'Unknown/Unclassified' families comprised the majority of REs across all intragenic regions, followed by long-terminal repeats (LTRs) (Fig. 3E).

We also assessed the expression of REs, via the mapping RNA-seq reads to the annotated *P. dumerilii* genome (Supplementary Fig. 7). We found that an estimated ~68% of all annotated REs were expressed (above a threshold cut-off of 10 unique mapping reads) (Fig. 3G). Of these, 66% of reads overlapped within annotated protein-coding gene loci, suggesting that the majority of RE expression likely originates from intragenic sites (Fig. 3H). We have yet to test if this is true RE specific expression, or if REs are "by-riders" of the host gene expression, as many REs sit within intronic regions of the gene loci. All RE types were expressed in both intragenic and intergenic regions, suggesting no overall differences in RE family expression (Supplementary Fig. 7B).

Gene content and evolution

Protein-coding genes

To annotate protein-coding genes in *P. dumerilii*, we aggregated good-quality publicly available Illumina paired-end short RNA-seq reads (Conzelmann et al., 2013) spanning multiple developmental stages in *P. dumerilii* - allowing us to access all major stages of the *P. dumerilii* life-cycle - and mapped the reads to the genome using STAR (Dobin et al., 2012). We also mapped PacBio (see Materials and Methods) and Nanopore long-read RNA sequencing datasets (Paré et al 2023) (spanning early developmental stages and regenerating samples; see Methods and Materials), using Minimap2 (Li, 2018), to capture isoform diversity and full-length genes. A genome-guided *de novo* transcriptome, combining all three

transcriptomic datasets, was then reconstructed using StringTie (Kovaka et al., 2019; M. Pertea et al., 2015). To survey the protein-coding landscape of the *P. dumerilii* genome, we searched the transcriptome for open reading frames (ORFs) using TransDecoder (Haas, 2018). In total, we modeled 28,985 protein-coding genes in *P. dumerilii*. Many genes contained at least three transcripts (~3.2 per locus), underscoring that alternative splicing likely also plays a key role in proteome diversity in *P. dumerilii*, as in other species.

Gene content

Previous analyses have proposed that although there is a correlation of protein-gene content to genome size, this correlation is relatively weak compared to the non-coding genome and REs (Hou & Lin, 2009). We explored this notion further for annelids, taking into account previous genome analyses (Liang et al., 2022; Martín-Durán et al., 2021; Simakov et al., 2013; Zakas et al., 2021) (Fig. 4 and Table 1). The total gene count in *P. dumerilii* was higher than in most annelids analyzed, with *C. teleta* being the only exception with ~400 more genes than *P. dumerilii*, despite measuring a smaller genome size (Table 1). Despite the small sample size (N = 6 species), a formal test for correlation between protein-coding gene counts and genome size within annelids revealed a weak association (*Spearman's rank correlation rho*, ρ: $r_s(4) = 0.43$, *p-value* = 0.419), consistent with previous reports in many other species (Hou & Lin, 2009). However, this may be impacted by spurious gene predictions in larger genomes.

We then explored whether gene sizes may have played a significant role in genome expansion in *P. dumerilii*. We extracted overall gene size (including UTRs, exons and introns) and compared *P. dumerilii* to *S. benedictii*, *H. robusta*, *C. teleta*, *D. gyrociliatus* and *O. fusiformis* (Fig. 4, Table 3). *P. dumerilii* genes (n = 28,983) tended to be greater in lengths than the other annelid genes (median ± MAD = 8,693 ± 9,939 bp) (Fig. 4A and Table 3). This measures to a significantly larger gene size than that of the smallest annelid genome *D. gyrociliatus* (n_{genes} = 13,905; median ± MAD = 2,155 ± 1,524 bp), and also much larger than the ~700 Mbp *S. benedictii* genome (n_{genes} = 20,221; median ± MAD = 1,043 ± 1,290 bp). As this analysis includes all gene features including exons, introns, 5'UTR and 3'UTR and their respective

sizes, we wanted to further test if there is a specific gene feature that would explain the gene size expansion in *P. dumerilii* relative to other annelids (Supplementary Fig. 8; Table 3). Across all annelid species investigated except *P. dumerilii* we consistently observed that per gene exon and intron counts were highly correlated with gene size (Pearson's product-moment correlation $r > 0.6$; Supplementary Fig. 9). *P. dumerilii* exon counts were weakly correlated with transcript size (Pearson's product-moment correlation $r = 0.08$, *p-value* $< 2.2e^{-16}$; Supplementary Fig. 9A). The average exon size per gene was however poorly correlated with gene size in *D. gyrociliatus* (Pearson's product-moment correlation $r = 0.05$, *p-value* $< 2.2e^{-16}$) and negatively correlated with gene size in all other annelids (Pearson's product-moment correlation $r < -0.01$; Supplementary Fig. 9A-B). Unlike exon size however, intronic size was better correlated with gene size in all annelids analyzed (Pearson's product-moment correlation $r > 0.27$; Supplementary Fig. 9A-B, Table 3), consistent with previous hypotheses supporting the role of intron size in gene size expansion. We also calculated the median values of all exon, intron 5'- and 3'-UTR sizes, and counts across all five annelid species included in this study (Table 2). We found that of all the features that show any correlation with gene size, the 3'UTR and the intron sizes were larger in larger genomes. Taken together, we argue that 3'UTR and intron sizes differences best explain the variation in gene/transcript length sizes in annelids, relative to other gene features such as counts (Table 3, Supplementary Fig. 8 and 9). The increase in intron size in *P. dumerilii* genes is likely a result of REs occupying these regions (Fig. 3D-E).

We then analyzed orthology of protein-coding genes in annelids. We first extracted the longest isoform for each protein-coding gene in *P. dumerilii* and downloaded publicly available proteome datasets for the other annelid genomes (Methods and Materials). Using OrthoFinder (Emms & Kelly, 2015), we found 116,625 genes (~83%) of the total proteome (i.e. all six annelid proteomes used in the dataset) could be assigned to 17,752 orthogroups. The percentage of the protein-coding genes in orthogroups varied from ~72% (*H. robusta* - 15,505 genes) to ~87% (*D. gyrociliatus* - 12,155 genes), with around ~83-84% in *P. dumerilii* - 24,053 and *O. fusiformis* - 22,468 genes, with ~ 85-86% in *S. benedictii* - 17,241 and *C. teleta*

- 25,203 genes (Fig. 4B; Table 1). We interpret that those genes without orthogroups could represent novel single-copy genes (i.e. do not belong to a family) for that particular species, or its lineage. Alternatively, those genes may be modeling artifacts. Increasing the number of sequenced and annotated annelid genomes will help distinguish between these two possibilities, allowing the distinction between species-specific (or novel) single-copy and lineage-specific genes. Gene ontology (GO) term analysis showed minimal differences in GO-term enrichment, when comparing genes assigned to unassigned orthogroups (Supplementary Fig. 10); suggesting that these two different gene sets are likely similar in terms of cellular, molecular function and biological processes.

Further analysis highlighted 25.7% of *P. dumerilii* genes with an orthology assignment as *P. dumerilii*-specific (i.e. these genes form orthogroups with other *P. dumerilii* genes but no other annelid genes), with *D. gyrociliatus* possessing 10.9% species-specific orthogroups (Fig. 4B). We note that the total gene count, proportion of genes in orthogroups or species-specific orthogroups was not correlated with genome size (Fig. 4B; Pearson's product-moment correlation, $p > 0.1$). This result suggests that gene duplication is not a major factor in genome size expansion in the annelid species analyzed thus far.

A diverse and rich repertoire of RNA genes in P. dumerilii

We modeled the non-coding transcriptome of *P. dumerilii* using both RNA-seq reads (putative lncRNAs and miRNAs) and *de novo* genome searches for tRNAs and rRNAs.

Transfer RNAs

P. dumerilii tRNAs were modeled *de novo* from the genome sequence alone using tRNAscan-SE (Chan et al., 2019; Lowe & Eddy, 1997) (Materials and Methods). We identified approximately 4,719 high-confidence tRNA genes (excluding the mitochondrial genome) in *P. dumerilii*, ranking as one of the few animal genomes annotated thus far to have such a high number of tRNA copy genes (Table 1). Consistent with previous findings, we do not find an obvious association of tRNA gene number and other

genome and/or organismal traits (Bermudez-Santana et al., 2010). We did note however that high-confidence tRNA counts were much higher in *P. dumerilii* than in the other annelid genomes with *C. teleta* having 1,219 genes and *H. robusta* possessing the fewest tRNA genes (168) (Table 1). We did not see any association of tRNA gene count and genome size as *D. gyrociliatus* had an estimated 320 tRNA genes (nearly double that of *H. robusta*), despite having a genome size 6 times smaller than *H. robusta* (Table 1).

Long non-coding RNAs

We made use of the available RNA-seq datasets described before to model putative lncRNAs. When generating the transcriptome, we limited further analyses to transcripts with a minimum length of 200 nucleotides (Materials and Methods). Of the total 67,020 loci, 28,985 were protein coding with a remaining set of 38,035 non-coding loci. This non-coding gene set comprised ~287 Mbp (~19.5% of the total *P. dumerilii* genome). Incorporating these sequences with databases that contain lncRNAs, although promising, remains a challenge as many of these annotated lncRNA sequences are from model organisms that are phylogenetically distant from *Platynereis* and further experimental validation is necessary (Cao et al., 2018; Zhao et al., 2021). Another recent study in *P. dumerilii* has also identified putative lncRNAs with specific expression in germ cells (Ribeiro et. al, 2024).

microRNAs

miRNAs are small non-coding RNAs that post-transcriptionally regulate the expression of target mRNAs. Previous studies in *P. dumerilii* identified 34 miRNA genes common to protostomes and deuterostomes (Christodoulou et al., 2010). A more rigorous survey of *P. dumerilii* miRNAs however is lacking. To this end, we sequenced small RNAs from multiple early developmental stages for *de novo* prediction of miRNA genes using MiRDeep2 (Friedländer et al., 2012), identifying 587 miRNA unique gene loci (Supplementary File 1). miRNA genes were annotated according to their sequence similarity to aligned

conserved miRNA families present in the MIRGeneDB2.1 database. In the absence of additional experimental validation, evolutionary conservation was used for validation.

Following these criteria we found a total of 109 miRNA genes belonging to phylogenetically conserved miRNA families (Supplementary File 2 and 3) (Supplementary Fig. 11). Therefore, *P. dumerilii* shows nearly complete conservation of the miRNA genes present at each considered phylogenetic node: Bilateria (bl), Protostomia (pr), Lophotrochozoa (lp), Annelida (an)**,** with only 5 miRNA families missing from the total of 61. Notably, four miRNA genes, belonging to four different families, are missing from the current genome assembly, but present in a previous versions (data not published): mir-137 (bl), mir-2001 (bl), mir-33 (bl), mir-1994 (lp). Regarding the missing annelid-specific duplicate of the Lophotrochozoan-specific mir-1990 family, *P. dumerilii* shows two copies that cannot be conclusively aligned to the mir-1990 duplicates in other Lophotrochozoans. Alternatively, the P1 and P2 copies of mir-1990 might be specific for the Sedentaria (Supplementary File 3).

Furthermore, we see multiple independent duplication events for several miRNA families leading to a substantial expansion of the miRNA repertoire in *P. dumerilii*. Almost one third of the conserved miRNAome results from these duplications, which gave rise to a total of 37 duplicates. In humans, miRNA genes located within 50 kb from each other are generally co-expressed, often deriving from polycistronic transcripts (Baskerville et al, 2005). Sincen all above-mentioned lineages of miRNA gene clusters are conserved (Fig. 6), this might well extend to their coordinated transcription. Overall, the conserved structure of the miRNA gene clusters reflects the stability and conservation of the *P. dumerilii* genome (Supplementary File 4)**.**

Phylogenetic mapping of the *Pdu* miRNAs yields insight into miRNA cluster evolution. We find two sets of clusters (mir-1990-P1_mir-1986 and mir-22-P1_mir-22-P2) shared between the annelids, *P. dumerilii* and *C. teleta*, molluscs and the brachiopod *L. anatina*. This suggests that these clusters are at least

Lophotrochozoa-specific. Since mir-22-P1 and mir-22-P2 are conserved across Bilateria but only clustered in Lophotrochozoa, the mir-22 cluster might have predated Bilateria and clustering subsequently has been lost except in the Lophotrochozoan lineage. Within annelids, three clusters are shared between *C. teleta* and *P. dumerilii* composed of annelid-specific genes, which might therefore be annelid-specific. Finally, we also detect four miRNA clusters that are only found in *Pdu*. These are made of miRNAs belonging to the same (homo-seed) miRNA family, which means that they most likely share the same mRNA targets. This suggests that their duplication might relate to a dosage increase of mature miRNAs to achieve higher target silencing efficiency.

Exploring natural variation in the *Platynereis* populations

Platynereis represents a cosmopolitan species complex (Fauvel, 1914; Read & Fauchald, 2020) (Teixiera et al., 2021), with previous descriptions or citings of assumed *P. dumerilii* based on morphology likely partially attributable to *Platynereis* sister species, in particular the sympatric sibling species *P. massiliensis*. The laboratory culture sequenced here corresponds to *Platynereis dumerilii sensu stricto* as defined by Teixera et al. (2021). We collected RNA-seq data equaling ~20 million read pairs (for each location) from *Platynereis spp.* populations collected in six locations (Las Cruces, Chile; Algarve, Portugal; Kristineberg, Sweden; Ischia, Italy; Oban, Scotland; and Mayotte, France). These data were mapped against the current assembly and annotated genome (Fig. 5A). We found that samples from Ischia best matched our own lab culture RNA-seq data (~47% of reads from the Ischia samples mapped to the genome as opposed to ~56% of reads from lab culture samples), followed by samples from Faro (~32%), Oban (~18%), Kristineberg (~13%) and Mayotte (~1.2%), with Las Cruces samples showing very poor mapping statistics (~0.6%) (Fig. 5B). This likely suggests these samples are not from the same species; highlighting the utility of our genome resource in qualitatively testing how *P. dumerilii* species (including the *Platynereis spp.* complex [see below]) are related to each other (Fig. 5A).

To study phylogenetic relationships of these populations, we conducted transcriptome assemblies for each of the sites. Consistent with the raw-mapping data, Las Cruces was most distantly related to all sites, with Mayotte samples forming an outgroup to all remaining sites (Fig. 5C). Samples from Oban, Kristineberg, Ischia and Faro form a monophyletic group, including the lab cultures, suggesting that these are either all *P. dumerilii s.s.* or closely related (Valvassori et al., 2015; Wäge et al., 2017; Teixera et al. 2021)). According to the transcriptome data, we also noted that Faro samples were more closely related to lab cultures than samples from Ischia, where the culture originated in the 1960s (Fischer & Dorresteijn, 2004).

From the mapped data, we then called single nucleotide polymorphisms (SNPs) and insertions/deletions (InDels) using SnpEff & SnpSift (Cingolani et al., 2012). We found that Ischia and Faro samples shared a higher proportion of their indels (~7%) with our lab-cultures, followed by Kristineberg (~4.5%) and Oban (~0.8%) samples (Fig. 5D). Mayotte samples showed no overlap of their indels with our lab cultures. As expected of RNA-seq genome mapping data from the site-specific presumed *P. dumerilii* samples, majority of the SNPs and indels were found in exons (~47%), with ~19% of variants called 5 Kbp upstream of the genes (not including the 5' UTR) and 24% of variants called 5 Kbp downstream of genes (not including the 3' UTR) (Fig. 5E). This was consistent across all sites for which we had RNA-seq data (Supplementary Figs. 12-13; for example gene locus see Supplementary Fig. 14). We further grouped genes according to their variable SNP/In(Del) counts, identified on the different sites (Fig. 5D). We limited our analyses to a maximum of 5000 genes that showed the highest variation in SNP/In(Del) counts across sites (Fig. 5F). Hierarchical clustering of these SNP/In(Del) counts showed that for any given gene, the Ischia samples showed similar numbers of SNPs/In(Dels) to the lab cultures - acting as our reference samples - (Fig. 5F; Supplementary Fig. 14), with Mayotte samples displaying a more distinct pattern from all other samples. To explore the likely function of these genes, we analyzed their GO term enrichment using WEGO (Ye et al., 2018) (Fig. 5G, Supplementary Fig. 15). The statistically significant enriched terms encoded for terms associated with pigmentation (pigment metabolic processes

[GO:0042440], pigmentation [GO:0043473]), development (developmental maturation [GO:0021700], pattern specification process [GO:0007389]) and immune-responses (cytokine production [GO:0001816], process utilizing autophagic mechanism [GO:0061919], immune system development [GO:0002520], oxidoreductase activity [GO:0016491]).

Comparative genomic analyses of three *Platynereis* species: *P. dumerilii*, *P. massiliensis* and *P. megalops*

Currently, there are around 33 *Platynereis* species, widely distributed across the globe (Read and Fauchald, 2023), of which four species (i.e. *Platynereis bicanaliculata, Platynereis dumerilii, Platynereis massiliensis* and *Platynereis megalops*) have been well described with respect to their development, life-cycle and sexual behaviors (Fischer and Dorresteijn, 2004; Hauenschild, 1951; Just, 1914, 1915, 1922; Roe, 1975). *P. dumerilii* is sympatric - occurs in the same area - with *P. massiliensis* (Hauenschild, 1951; Valvassori et al., 2015), a sibling species that has a distinct early developmental program (Helm et al., 2014; Schneider et al., 1992). However, development then converges and individuals become morphologically indistinguishable during juvenile/adult stages. The two species can be again distinguished morphologically only after sexual maturation, based on pigmentation and mating behaviors (Fischer et al., 2010; Hauenschild, 1951; Helm et al., 2014; Lucey et al., 2015; Schneider et al., 1992; Valvassori et al., 2015). Intriguingly, *P. dumerilii* sex maturation is thought to transition directly to either male or female. On the contrary, *P. massiliensis* undergoes sequential hermaphroditism from males to females (Hauenschild, 1951; Helm et al., 2014). The mechanisms underlying these sex-determination modes are not known. *P. megalops* on the other hand, has only been described from Woods Hole (Just, 1914), or the Vineyard Sound region in Massachusetts (Verrill, 1873). Curiously, *P. megalops* appears to show a very similar development to *P. dumerilii*, however the mating behaviour is once more drastically distinct from that of *P. dumerilii* and *P. massiliensis* (Just, 1914, 1922). *P. dumerilii* and *P. massiliensis* display external fertilisation, whereas *P. megalops* fertilisation is internal. These observations highlight

the 'intra-genus' variation in this clade, and the dynamic evolution of these phenotypes. Little is known about the genotypic/genomic variation among these diverse *Platynereis* species. The mitochondrial genomes show little difference in gene number across *Platynereis* species, with only *P. massiliensis* possessing an additional tRNA (total of 23 as opposed to 22) (Alves et al., 2020). To better explore the patterns and dynamics through which genes have been gained/lost in annelids, we carried out genome sequencing, annotation and comparative analyses of *P. massiliensis* and *P. megalops* genomes.

Assembly statistics across the three studied **Platynereis** *genomes*

While genomes of *P. dumerilii* and *P. megalops* (version 1.0) could be assembled to 'pseudo-chromosome' resolution, the *P. massiliensis* (version 1.0) was only assembled to scaffold-resolution (Table 4). The assemblies were generated using the 'hifiasm' algorithm (Cheng et al., 2021), and further scaffolded using the Arima pipeline and SALSA2 (Materials and Methods). All three assemblies showed greater than 90% BUSCO completeness values (Table 4), implying that the majority of their sequence information was captured by the sequencing data. *P. massiliensis* possessed the most fragmented assembly at 1,478 scaffolds compared to 163 and 370 for *P. dumerilii* and *P. megalops* respectively (Table 4). The *P. megalops* assembly was the largest at ~1.88 Gbp, approximately 400 Mbp larger than both *P. dumerilii* and *P. massiliensis* at ~1.42 and ~1.43 Gbp respectively (Table 7). Significantly, for both the version 1.0 (submitted to Ensembl Metazoa) and version 2.0 (only HiFi assembly, no annotations as-of-yet) of the *P. dumerilii* assemblies, the genome size remained stable (see Table 1 and Table 4), suggesting that this assembly size may reflect the 'actual' genome size much more fittingly than the initial ~980 Mbp estimate (Jha et al., 1995), and the short-read DNA Illumina sequence data (see above).

The previous analyses on the first version of the *P. dumerilii* genome revealed RE content to be a major player in genome size differences. To explore if this is also the case between closely related species, REs were identified, revealing that their content scaled with genome size in the three *Platynereis* species

(Table 4 - 5). Of all the REs that scaled according to genome size, the Unclassified elements appeared to be the most dominant (Table 5), comprising 31% of the total genome length in *P. dumerilii* and *P. massiliensis* and 38% of *P. megalops* genome. All other REs were distributed similarly across the different species, suggesting that these elements do not change rapidly.

All three *Platynereis* species show overall conservation of scaffold homology, which likely extends to chromosomes. However, several breaks in collinearity within homologous regions could be observed indicative of their evolutionary distance (Fig. 7). Additionally, some translocations between scaffolds can be observed for *P. megalops*. While this suggests that overall no major rearrangements of karyotypes had happened in these three species, further sequencing efforts are required to test chromosomal conservation and prevalence of inter-chromosomal translocations in this clade.

Insights into chromosomal evolution in annelids

Conserved chromosomal-level conservation (macro-synteny) can be traced over very large phylogenetic distances spanning the animal tree of life (Putnam et al., 2007, 2008; Simakov et al., 2022). Previous studies showed that Bilateria linkage groups (BLGs, (Simakov et al., 2022)) can be used to represent karyotypes of many animal genomes. Such algebraic operations involve duplications of ancestral chromosomal elements (e.g., in vertebrates (Simakov et al., 2020)) or various fusion processes (Simakov et al., 2020, 2022) with and without mixing of genes (Fig. 8A). We sought to explore the chromosomal-level organization of the *P. dumerilii* genome in the context of annelids, and more generally other Spiralia assemblies currently available and their BLG representation. In total, we analyzed 51 spiralian genomes at near-chromosomal level resolution, including 25 annelids, nine mollusks, six flatworms, four bryozoans, two nemerteans, three rotifers, a brachiopod and a phoronid (Table 6). Of particular interest to us was the identification of chromosomal fusion-with-mixing events that have been

suggested to comprise strong irreversible synapomorphic characters (Simakov et al., 2020, 2022) (Fig. 8A), informative for validation of phylogenetic relationships (Schultz et al., 2023).

BLG configurations towards and within Errantia

Consistent with previous findings in other animal lineages, the majority of annelid genomes showed retention of the BLG complement with a high degree of chromosome conservation. Clitellata are an exception (n = 7) (Table 6; Supplementary Fig. 9) with BLG 'scrambling' (Fig. 8C; Supplementary Fig. 9), consistent with earlier sequencing results and recent studies (Simakov et al., 2013; Schultz et al 2024; Lewin et al 2024, Vargas-Chávez et al., 2024). For the non-clitellate species with clear BLG conservation we annotated additional BLG fusions-with-mixing events. For example, most annelids (except clitellates) possessed a J2⊗L⊗C2 fusion-with-mixing (Table 6, Fig. 8C), with only *Owenia* lacking this specific event (Fig. 8C), which supports *Owenia* (and Paleoannelida) representing the sister group to other annelids (note that J2⊗L alone represents a shared spiralian fusion (Simakov et al, 2022)). We also noted that *Dimorphilus* did not show statistically significant enrichment of the C2 BLG associated genes, likely due to faster evolutionary rates in genes and chromosomes in *Dimorphilus* (Table 6, Supplementary Fig. 9B).

We found a potential synapomorphic fusion-with-mixing event A1⊗E within errant annelids including Nereididae (n = 2) and Polynoidae (n = 5), belonging to the taxonomic order Phyllodocida. We note however that Hesionidae (n = 1) and Sigalionidae (n = 1), both thought to belong to the same taxonomic order of Phyllodocida (Weigert and Bleidorn, 2016), lacked this fusion-with-mixing event, suggesting placement of Sigalionidae outside the group comprising Nereididae and Polynoidae. Alternatively, A1⊗E could have fused independently in Nereididae and Polynoidae. Corroborating this, we did identify several potential independent fusion-with-mixing events on the annelid tree. For instance, the two lineages Oweniidae (n = 1) and Sipuncula (n = 1) both showed a I⊗E fusion-with-mixing event (Fig. 8C; Table 6). Between Hesionidae (n = 1) and Serpulidae (n = 1), the D⊗E fusion is shared, however this is only a single fusion character that unites two distinct lineages belonging to Errantia and Sedentaria (Weigert &

Bleidorn, 2016), thus it has higher chance of being a convergent fusion. Within Sedentaria, Dinophilidae (n = 1) and Terebelliformia (n = 1) shared the P⊗E fusion event, despite belonging to distinct lineages (Fig. 8C). Single convergent fusion-with-mixing events, while rare, are not unlikely to be observed at such vast evolutionary distances (Schultz et al, 2023).

Within Errantia, *Platynereis* and its relatively closely related genus *Alitta* showed several shared fusion-with-mixings, including C1⊗F, A1⊗E, and addition of several other fusions on-top of the pre-existing 'spiralian'-fusions (G to H⊗Q, and O2⊗K⊗C2⊗B1 to J2⊗L). This suggests that the reduced chromosomal complement within this taxon can be at least partially explained by these chromosomal fusions.

Conclusion

Despite its prominent research history, *Platynereis* has so far been lacking an adequate genome resource. Advances in sequencing technology and computational sequence analysis have been catalysts in achieving the goal of building the current resource. Together, the data and analyses presented provide the reference genome assembly and annotation of a highly sought after spiralian model species and its two sibling species. While in the draft state, the assembly will clearly benefit from further chromosomal-scale sequencing efforts of closely related populations and species to disentangle the high genetic variation and produce a usable linkage map. This will in turn allow for the study of its evolutionary resilience and change at the cell-type level.

Materials and Methods

The current version of the *P. dumerilii* genome assembly is available at NCBI under accession number GCA_026936325.1 (WGS Project: JAPTHN01).

DNA extraction and sequencing

Dovetail Genomics processed *P. dumerilii* samples for PacBio Sequel II CLR DNA sequencing. An additional sample was prepared for both Illumina sequencing. We used a protocol adapted from Oxford High Molecular Weight (HMW) extraction protocol from The Jackson Laboratory. Briefly, a single *P. dumerilii* sexually mature male (post sperm release) was incubated in 10mL lysis buffer (5M NaCl, 1M Tris pH8.0, 0.5M EDTA pH8.0, 10% SDS, 100mg/mL RNAse, Nuclease free H_2O) for 1 hour at 37°C. 50-500 µL of Proteinase K was added to the lysis buffer and after mixing, sample was incubated at 50°C overnight (~12-16 hours). DNA precipitation and elution were performed using the QIAGEN MaXtract High Density protocol kits (Cat. No. / ID: 129073). DNA was eluted in 300uL of TE pH8.0 buffer and stored at 4°C or -80°C, for long-term storage. All our DNA Illumina sequencing was performed in EMBL, Heidelberg at the Genomics Core Facility. For the genome size estimates and polishing Illumina data, DNA was extracted from a single 'wild-type' lab culture male (after sperm release) using the HMW QIAGEN Phenol-Chloroform kits. Approximately 500 nanograms (ng) – 1 microgram (µg) of DNA was used for library preparations and quality control. This sample was then sequenced using two HiSeq4000 lanes with 150 PE (thus a total of 300 bases). All protocols yielded good quality DNA at A260/230 and A260/230 ratios greater than 1.8.

Genome size measurements

We estimated genome size using k-mer distributions and DNA quantity. To calculate k-mer distributions from Illumina 150PE sequenced (see below) data, we first trimmed high-quality reads based on length

using Trimmomatic (Bolger et al., 2014) (command/parameters used '*java -jar trimmomatic-0.38.jar PE -threads 16 illumina_reads_R1.fastq illumina_reads_R2.fastq paired_R1.fq unpaired_R1.fq paired_R2.fq unpaired_R2.fq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:105*'). Although the data was of high quality prior to trimming (evaluated using FastQC (Andrews, 2010)), we nevertheless trimmed data to ensure that only the best quality reads were retained for analysis. Duplicate reads were then discarded using SuperDeduper (Petersen, 2015) – re-implemented as part of HTStream – command/parameters: '*hts_SuperDeduper -1 trimmed_paired_R1.fq -2 trimmed_paired_R2.fq -f deduped*'). Deduplicated reads were further normalized (k-mer coverage to 100 times) using BBNorm from the BBTools suite (version 37.68; command/parameters '*bbnorm.sh prefilter=t usejni=t in=deduped_R1.fastq in2=deduped_R2.fastq out=normalized_R1.fq out2=normalized_R2.fq target=100 min=5*'). Counting of k-mers, size 21bp, was performed with Jellyfish version 2.2.7 (Marçais & Kingsford, 2011) using the commands: '*jellyfish count -t 16 -C -m 21 -s 6G -o 21mer_out --min-qual-char=? normalized_R1.fq normalized_R2.fq*' followed by '*jellyfish histo -o 21mer_out.histo 21mer_out*'. We used Smudgeplot (version 0.2.1) and GenomeScope (version 2.0) (Ranallo-Benavidez et al., 2020) to estimate ploidy, heterozygosity and genome size respectively, from the normalized Illumina DNA-sequencing dataset. To estimate ploidy and heterozygosity, the commands '*smudgeplot.py cutoff 21mer_out.histo L*' and '*smudgeplot.py cutoff 21mer_out.histo U*' were executed so as to determine the lower and upper coverage cutoffs/thresholds. '*jellyfish dump -c -L $L -U $U 21mer_out | smudgeplot.py hetkmers -o kmer_pairs*' was the command used to extract heterozygous k-mer pairs. Plots were derived using '*smudgeplot.py plot kmer_pairs_coverages.tsv -o pdum*'. The GenomeScope command executed to estimate genome size: '*genomescope2 -i 21mer_out.histo -o output -k 21*'.

We also stained dissociated *Drosophila melanogaster* embryos and *P. dumerilii* male and female adults into nuclei and stained them with DAPI to estimate DNA content via FACS. Briefly, frozen *Drosophila* embryos (stored at -80ºC) were resuspended in 500µL ice-cold lysis buffer (10mM Tris-Cl pH 8.0, 10mM NaCl, 0.2% IGEPAL CA-630 and 1X cOmplete PI) and carefully ground using pre-chilled metal pestles

until embryos were completely lysed. After incubating on ice for 15 mins, the homogenized suspension was centrifuged at 2,000 x *g* at 4ºC for 10 mins, and the supernatant discarded. Remaining pellet was washed at least once with 500µL ice-cold lysis buffer, and later with 500µL ice-cold PBTriton 0.1%. Nuclei were then mechanically extracted using a series of 20G and 22G syringes, carefully pipetting the mixture. Solution was left in 4ºC for no more than a week for experiments. Nuclei were then counted using beads, and then an equivalent DAPI:cell number staining ratio was added so that all nuclei across different samples had equal amounts of DAPI stain. A similar approach was taken for the *P. dumerilii* samples with the exceptions that animals were homogenized using 250mM sucrose, 25mM KCl, 5mM $MgCl_2$ 10mM Tris-HCl pH 8.0 and 0.1% Nonidet P40/IGEPAL solution. After homogenizing the samples followed by low-speed centrifugation at 100 x *g* for 1 min at 4ºC, the supernatant was saved and washed at least once with ice-cold homogenization buffer and at least twice with ice-cold 1X PBS with centrifugation at 400-500 x *g* for 4-7 mins each. Nuclei suspensions in 1X PBS were then filtered using 40µM Flowmi strainers and 10µM filters.

Genome assembly

Three different long-read genome assembly algorithms were tested against our PacBio Sequel II CLR data; CANU (Koren et al., 2017), FLYE (Kolmogorov et al., 2019) and wtdbg2 (Ruan & Li, 2020) (Mutemi, 2023). We compared all assemblies, and found that CANU returned the more contiguous and complete genome assembly (Mutemi, 2023). The specific parameters used for CANU (version 2.1) were: '*canu -minReadLength=1000 -minOverlapLength=500 -genomeSize=1g -pacbio pacbio_sequel_II_CLR_subreads.fastq.gz -useGrid=false*'. Iterative purging was performed using purge_dups (Guan et al., 2020) and purge_haplotigs (Roach et al., 2018) (see Mutemi, 2023 for protocol details). Throughout the purging iterations, we ran BUSCO (Simão et al., 2015) (version 5.0.0) to evaluate likely genome completeness: '*busco -i canu_asm.fa --config config.ini -l metazoa_odb10 -m genome –long -c 24*'. The final assembly was subjected to LINKS scaffolding (Warren et al., 2015) (see Mutemi, 2023 and below for recipe details), followed by Illumina-read polishing via POLCA (Zimin &

Salzberg, 2020) with the command '*polca.sh -a links_canu_16.scaffolds.fa -r "trimmed_paired_R1.fq trimmed_paired_R2.fq" -t 32 -m 1G*'; approximately ~2 million substitution errors and ~1.8 million insertion/deletion errors were corrected, giving a final consensus quality of 99.74. We upload the assembly to NCBI under the GenBank assembly ID GCA_026936325.1 (whole genome sequence [WGS] project JAPTHN01) to facilitate its further exploration, annotation and quality maintenance, in hopes that it can be a useful resource for a broader audience, and not only the *Platynereis* community.

k-mer pairs and Hi-C scaffolding

After assembling and polishing the genome (contigs=1007, N50=2.4 and size=1.46 Gbp), we subjected the assembly through iterative scaffolding methods. Firstly we made use of LINKS (Warren et al., 2015), which searched for *k*-mers of length 25 bp in the genome assembly as well as the raw PacBio Sequel II CLR datasets. The k-mer pairs used for scaffolding the contigs from the raw reads were sampled at different intervals (ranging from 200 to 1 bp) and over several distances (spanning from 10 to 200 Kbp); '*LINKS -f asm.fa -s long_reads.txt -d 10000-100000 -t 200-1 -k 25 -b links_pdum -v 1*' (see Mutemi, 2023 and (Warren et al., 2015) for detailed iterative protocol). The LINKS scaffolded genome was then subjected to Hi-C scaffolding via SALSA2 (Ghurye et al., 2019), with 10 iterations. The Hi-C plot (Fig. 2) was generated using the 'convert.sh' script in the SALSA2 pipeline and visualized using Juicebox (Robinson et al., 2018).

A single *P. dumerilii* Hi-C library was prepared by Dovetail Genomics. For this dataset, the Hi-C library was prepared from hundreds of progenies from two parents (one male ♂and one female ♀). Briefly, chromatin was fixed in place with formaldehyde in the nucleus and then extracted. Fixed chromatin was digested with *DpnII*, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed, and the DNA purified from proteins. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and a sequencing library was generated using Illumina-compatible adapters.

Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of the library. The library was sequenced on three different Illumina lanes; a HiSeqX platform and two more HiSeq4000 lanes.

Repeat modeling and evolutionary analysis

We modeled and masked REs using RepeatModeler (version 2.02) (Flynn et al., 2020) and RepeatMasker (version) (Tarailo-Graovac & Chen, 2009), on the scaffolded genome. We first built a database of likely REs using Dfam release 3.3 (Storer et al., 2021) with the commands '*BuildDatabase -name pdum scaffolds.fasta*' and '*RepeatModeler -database pdum -pa 6 -LTRStruct > run.out*'. REs were then masked using RepeatMasker (version 4.1.1) (Tarailo-Graovac & Chen, 2009) with the commands: '*RepeatMasker -lib pdum-families.fa -pa 10 -dir pdum_scaffolds_masked -xsmall -gff -cutoff 250 -xm scaffolds.fasta*'. The same commands were executed when re-annotating REs for the other annelid genomes. The RMouttobed.pl script (Kapusta, 2017) was used to convert the *.out file from the RepeatMasker command into a *.bed file. This file was then used to extract length statistics and compare inter- and intra-genic occupancy of repeat elements using custom R scripts.

Mapping of quality-trimmed RNA-seq reads (see below) to an annotated 'repeat_element.gtf' file was performed via STAR (version 2.7.1a) (Dobin et al., 2012), using the commands: '*STAR --runThreadN 24 --runMode genomeGenerate --genomeDir Reindex/ --genomeFastaFiles pdum_scaffolds.masked.fa --sjdbGTFfile pdum_repeats.gtf --limitSjdbInsertNsj 2637930*' followed by '*STAR --runThreadN 24 --genomeDir Reindex/ --readFilesIn trimmedRNA_paired_R1.fq trimmedRNA_paired_R2.fq --quantMode TranscriptomeSAM GeneCounts*'. The 'ReadsPerGene.out.tab' output file was further analyzed for RE expression counts in different loci. We considered only those loci whose read counts amounted to greater than or equal to 15 (Chen et al., 2016).

Gene modeling and annotation

Transfer-RNA genes were predicted using tRNAscan-SE (version 2.0.7) (Chan et al., 2019; Lowe & Eddy, 1997) using the commands: '*tRNAscan-SE -EHQ -o# -f# -m# -s# -a# -l# --detail --thread 16 -p pdum_v2_tRNA pdum_scaffolds.fasta*' followed by '*EukHighConfidenceFilter --result pdum_v2_tRNA.out --ss pdum_v2_tRNA.ss -p pdum_v2_tRNAQC -o eukqualfilt_pdumv2_tRNAs*'. Ribosomal RNA genes were predicted *de novo* from the assembly using barrnap (version 0.9) (Seemann, 2018), via the commands: '*barrnap --kingdom euk --threads 6 --reject 0.1 scaffolds.fasta --outseq pdum_rRNA.fa*'. We further mapped partial *P. dumerilii* rRNA sequences previously identified (Hui et al., 2007), and found consistent ribosomal gene loci, with the exception of novel 5S rRNA genes found with the barrnap models. Short PE Illumina RNA-seq reads were quality trimmed using Trimmomatic (version 0.38) with the commands '*java -jar trimmomatic-0.38.jar PE -threads 16 pdum_totalRNA_R1.fq pdum_totalRNA_R2.fq trimmedRNA_paired_R1.fq trimmedRNA_unpaired_R1.fq trimmedRNA_paired_R2.fq trimmedRNA_unpaired_R2.fq LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:45*'. These reads were then mapped to the scaffolded and masked genome using STAR (version 2.7.1a) (Dobin et al., 2012) with the commands: *STAR --runThreadN 16 --runMode genomeGenerate --genomeDir index/ --genomeFastaFiles pdum_scaffolds.fasta,* followed by *STAR --runThreadN 16 --genomeDir index/ --readFilesIn trimmed_paired_R1.fq trimmed_paired_R2.fq*. Nanopore and PacBio transcriptomic long-reads were mapped to the same genome using Minimap2 (version 2.17) (Li, 2018); '*minimap2 -ax splice:hq -uf pdum_scaffolds.fasta pacbio_isoseq.fa > pb.sam*' and '*minimap2 -ax splice pdum_scaffolds.fasta nanopore.fq > ont.sam*'.

StringTie (version 2.1.7) (Kovaka et al., 2019; M. Pertea et al., 2015) was used to reconstruct transcripts from the sorted .BAM mapping files generated for both short- and long-reads; '*stringtie -o *.gtf *sorted.bam --conservative -p 16*', using the '-L' option for long-reads. Using GFF-Utilities (G. Pertea & Pertea, 2020), we grouped overlapping transcripts from the three datasets (i.e. Illumina PE, PacBio and Nanopore) into loci using '*gffcompare -i {gtf_list.txt}*' and '*gffcompare gffcmp.combined.gtf*' - repeated at

least three times to completely collapse overlapping exons/transcripts. Transcripts were extracted via '*gffread -w pdum_transcripts.fa -g scaffolds.fasta gffcmp.combined.gtf*' and BUSCO values were calculated (using the transcriptome mode) giving a BUSCO score of: C, 96.7% [S: 44.0%, D: 52.7%], F, 0.6% and M, 2.7% from a Metazoa list of 954 genes. To convert to the Ensembl Gene-Transfer-Format (GTF), the AGAT tool (version 0.5.1) command 'agat_sp_ensembl_output_style.pl' was used (Dainat, 2020). Transcripts were extracted from the genome *.gtf using TransDecoder (version 5.5.0) 'util' scripts: '*perl ~/util/gtf_genome_to_cdna_fasta.pl pdum_ensembl_genome.gtf scaffolds.fasta > pdum_transcripts.fasta* and *perl ~/util/gtf_to_alignment_gff3.pl pdum_ensembl_genome.gtf > pdum_transcripts.gff3*'. Protein-coding genes were then retrieved via TransDecoder (version 5.5.0) using '*TransDecoder.LongOrfs -t pdum_transcripts.fasta* and *TransDecoder.Predict -t pdum_transcripts.fasta*'. Finally a genome-based coding region annotated file was built using '*perl ~/util/cdna_alignment_orf_to_genome_orf.pl pdum_transcripts.transdecoder.gff3 pdum_transcripts.gff3 pdum_transcripts.fasta > pdum_transcripts.transdecoder.genome.gff3* '. In total, we modelled 166,199 mRNA transcripts likely originating from 69,573 loci. Of the 166,199 transcripts, 72,852 had multiple exons. 24,237 loci (of the 69,573) contained multiple transcripts with (~ 2.4 transcripts/loci), suggesting that approximately 35% of *P. dumerilii* genes contain more than one isoform. From the 166,199 transcripts (incl. isoforms), 93,240 had predicted ORFs; amounting to a total of 28,985 likely protein-coding genes. Selecting the longest peptide sequences for each transcript with an ORF, resulted in a total of 78,322 sequences. We subsequently annotated these protein sequences using orthology- (i.e. EGGNOGMAPPER version 2.1.5 (Cantalapiedra et al., 2021; Huerta-Cepas et al., 2019)), revealing only 37,664 with annotations, and 5,418 protein isoforms having a unique annotation.

For miRNA annotation and analysis, several developmental samples were pooled and subjected to single-end sequencing at ~80bp length. RNA was extracted from several *P. dumerilii* developmental stages (24, 36, 48, 72 and 144 hours post-fertilisation [hpf]), a minimum of three replicates for each developmental stage, using the Direct-zolTM RNA MiniPrep (Cat. No.: R2050). Sequencing libraries

were prepared with the NextFlex smRNA kit. The 3' adapter sequence 'TGGAATTCTCGGGTGCCAAGG' was trimmed using AdapterRemoval (version 2.3.2) (Schubert et al., 2016) using the commands: '*AdapterRemoval --file1 pdum_total_smallRNAs.fq --adapter1 TGGAATTCTCGGGTGCCAAGG --output1 pdum_trim_smallRNAs.fq*', with additional random bases appearing immediately 5' and 3' to the insert removed using '*seqtk trimfq -b 4 -e 4 pdum_trim_smallRNAs.fq > pdum_trim_clip_smallRNAs.fq*'. miRNAs were predicted using miRDeep2 (Friedländer et al., 2012) with the commands: '*bowtie-build pdum_scaffolds.fasta pdum*' to generate the index, '*mapper.pl pdum_trim_clip_smallRNAs.fq -e -h -I -j -m -l 18 -p pdum -s pdum_all_filt_collapsed.fa -t pdum_all_collapsed_genome.arf -v*' to process and map reads to the genome and '*miRDeep2.pl pdum_all_filt_collapsed.fa pdum_scaffolds.fasta pdum_all_collapsed_genome.arf none cte.fas none 2>report.log*'.

This identified a total of 587 miRNA 'unique' gene locations, of which many could be grouped into miRNA gene clusters. Conserved miRNA sequences were identified using the Blast tool from MiRGeneDB 2.1 database (Fromm et al, 2022). Each output sequence from the MirDeep2 prediction was searched in MiRGeneDB 2.1. miRNA gene names were thus assigned according to the best hits (i.e. lowest E value and/or highest Bits score) and by comparison to correspondent miRNA gene families of related species. Pending experimental validations, any predicted sequence that had no significant alignments with conserved miRNA families (thus putative *P. dumerilii*-specific genes) was not annotated and thus not considered in subsequent analysis. A compendium detailing the annotation of each conserved miRNA gene and gene families was built (Supplementary File 03).

Identification and annotation of clustered miRNA genes:

For each selected organism, miRNAs genome coordinates were downloaded from miRGeneDB 2.1 (//mirgenedb.org/browse). Genes were then sorted according to their genomic position and considered to be clustered when they were within 50KB (Baskerville et al., 2005). The same approach was used to identify the clusters in *P. dumerilii*, considering the genome location of each annotated

miRDeep2-predicted miRNA. A table summarizing the details of the conserved miRNA clusters in each specie was generated (Supplementary File 04).

Gene-content evolution analyses

For protein-coding gene count and statistics (i.e. exon, intron etc.), *.gff3 files for *Helobdella*, *Capitella* and *Dimorphilus* were downloaded from the NCBI Assembly database. AGAT (version 0.5.1) was used to quickly test for .gff3 file compatibility and basic statistics (Dainat, 2020). The longest isoforms per gene locus were extracted via the AGAT and GFF-Utilities tools using '*agat_sp_keep_longest_isoform.pl -gff transcripts.fasta.transdecoder.genome.gff3 -o single_isoform.transdecoder.genome.gff3* followed by *gffread -w pdum_single_isoform.transdecoder.fa -g scaffolds.fasta single_isoform.transdecoder.genome.gff3*'. The calculated BUSCO completion for this single isoform/gene fasta file stood at C, 92.1% [S: 88.4%, D: 3.7%], F, 0.6% and M, 7.3% from a Metazoa list of 954 genes; missing ~4.6% of genes from the original proteome file (see above). For the comparative analyses, the *Capitella*, *Helobdella* and *Dimorphilus* CDS *.fasta files were downloaded from the GenBank Assembly database. For the *Streblospio* genome, we first extracted the longest protein-coding isoform/gene from the *.gff3 file, and extracted the transcripts (similar to the workflow for *Platynereis*) and filtered the non-coding tRNAs using pyfaidx (commands: '*faidx streblospio_single_isoform.cds.fa -v -g "_nc_"*'. To identify orthologous groups, we used OrthoFinder (version 2.5.4) (Emms & Kelly, 2015) using the default settings commands: '*orthofinder -f metazoa_proteomes_folder*'. We annotated these orthology groups using eggNOG-mapper (version 2.1.5) (Cantalapiedra et al., 2021) for each species to further explore if these species-specific orthogroups fall into specific functional categories.

Variation analyses

P. dumerilii animals were sampled from the different sites with guided experts from the local marine stations. RNA was collected and sequenced using Illumina PE, and only quality-trimmed-paired reads were used for further analyses. We sampled the datasets to 20 million read pairs using '*seqtk sample -s10 RNA_trimmed.paired.R.fq 20000000 > RNA_trimmed.paired.sampled.R.fq*', and mapped the data using STAR (Dobin et al., 2012) (version 2.7.9a). Variants were identified using bcftools (version 1.13) (Li, 2011) using the commands '*bcftools mpileup -Ou -f scaffolds.fasta rna_aligned.sortedByCoord.out.bam | bcftools call -mv -Ob -o calls.bcf*'. Preliminary descriptions of variants were performed using the SnpEff & SnpSift tool (version 5.0) (Cingolani et al., 2012). A *P. dumerilii* specific database was built using the commands: '*java -jar snpEff.jar build -v pdumv2.0*'. SNP and In/Del quantification was summarized using the commands '*java -jar snpEff.jar eff pdumv2.0 <vcf_file>*'. To count number of SNPs/In(Del)s per gene, we prepared genome annotation and variant call *.BED files (from *.GFF3 and *.VCF respectively) using the BEDOPS (version 2.4.39) (Neph et al., 2012) '*convert2bed*' and '*vcf2bed*' commands. We then summarized the counts of SNPs and In(Del)s using the BEDTOOLS (version 2.30.0) '*annotate*' command.

The various *P. dumerilii* samples' transcriptomes were built using RNA-Bloom (version 1.3.1) (Nip et al., 2020) with the commands: '*rnabloom --left <site_platynereis_sample>.R1.fq --right <site_platynereis_sample>.R2.fq -rcr -ntcard -outdir <site_sample>*'. Transcriptomes were subsequently filtered for sequences greater than 1000 bp, and Orthofinder (version 2.5.4) (Emms & Kelly, 2015) was used to identify orthology groups and generate a 'species' tree using: '*orthofinder -f variants_transcriptome_folder -M msa -A mafft -T fasttree -d -S blast_nucl*'. To test which genes showed site-specific variants, the eggNOG-mapper (version 2.1.5) (Cantalapiedra et al., 2021; Huerta-Cepas et al., 2019) annotated genes - commands: '*emapper.py -m diamond --itype CDS -i pdum.fasta -o emapper_output --cpu 50*' - were intersected with the positions at which variants were detected.

Synteny and gene-content evolution analyses

A set of *Branchiostoma floridae* highly conserved protein-coding genes were searched for using BLAST+ (version 2.12.0) against the hard-masked genome versions (i.e. all REs converted to 'N' nucleotides). Nucleotide databases for the hard-masked annelid genomes were built using the commands '*makeblastdb -in <hard-masked.fasta> -parse_seqids -dbtype nucl -out <database_name>*'. The conserved *B. floridae* proteins were then searched using '*tblastn -query <B.floridae_proteins.pep> -db <database_name> -out <blast_output_name> -max_target_seqs 1 -outfmt 6 -evalue 1e-2 -num_threads 24*'. The blast output file was then parsed for synteny analysis using custom scripts; the BLAST+ output result was processed using the following

'*perl makeMap_Blast.pl <blast.outfmt6.output> . > blast.outfmt6.output.chrom*'

'*perl prepMsynt2.pl blast.outfmt6.output.chrom . threeway_final.allmbh.clus > blast.outfmt6.output.3waymbh_final.msynt*'.

We accessed publicly available assembled annelid genomes (n = 22, plus our *P. dumerilii* assembly) (at either chromosome or near-chromosome resolution) available via the NCBI Assembly database (see Materials and Methods). This included an annelid genome from the Oweniidae (*Owenia fusiformis*), which together with Magelonidae comprises the Paleoannelida – earliest major branch of Annelida – another from the Sipuncula lineage (*Sipunculus nudus*) which together with Amphinomidae form a sister group to the Pleistoannelida (Errantia and Sedentaria) (Weigert & Bleidorn, 2016). From the Errantia lineage, two Nereididae (*Alitta virens* and *Platynereis dumerilii*), five Polynoidae (*Acholoe squamosa*, *Alentia gelatinosa*, *Branchiopolynoe longqiensis*, *Harmothoe impar* and *Lepidonotus clava*), a Hesionidae species (*Amphiduros pacificus*) and a Sigalionidae species *Sthenelais limicola*, and a total of 12 Sedentaria with six from Clitellata (*Bimastos eiseni*, *Branchellion lobata*, *Helobdella robusta*, *Lumbricus rubellus*, *Metaphire vulgaris* and *Piscicola geometra*) and a single representative from each of Terebelliformia (*Terebella lapidaria*), Capitellidae (*Capitella teleta*), Spionidae (*Streblospio benedictii*), Serpulidae (*Protula*), Siboglinidae (*Paraescarpia*) and Dinophilidae (*Dimorphilus gyrociliatus*).

References

Andrade, S. C. S., Novo, M., Kawauchi, G. Y., Worsaae, K., Pleijel, F., Giribet, G., & Rouse, G. W. (2015). Articulating "Archiannelids": Phylogenomics and Annelid Relationships, with Emphasis on Meiofaunal Taxa. *Molecular Biology and Evolution*, *32*(11), 2860–2875. https://doi.org/10.1093/molbev/msv157

Andrews, S. (2010). *A Quality Control Tool for High Throughput Sequence Data.* [Computer software]. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Baskerville S, Bartel DP. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA. 2005 Mar;11(3):241-7. doi: 10.1261/rna.7240905. PMID: 15701730; PMCID: PMC137071.

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*, *58*(3), 268–276. https://doi.org/10.1016/j.ymeth.2012.05.001

Bermudez-Santana, C., Attolini, C. S.-O., Kirsten, T., Engelhardt, J., Prohaska, S. J., Steigele, S., & Stadler, P. F. (2010). Genomic organization of eukaryotic tRNAs. *BMC Genomics*, *11*(1), 270. https://doi.org/10.1186/1471-2164-11-270

Bhambri, A., Dhaunta, N., Patel, S. S., Hardikar, M., Bhatt, A., Srikakulam, N., Shridhar, S., Vellarikkal, S., Pandey, R., Jayarajan, R., Verma, A., Kumar, V., Gautam, P., Khanna, Y., Khan, J. A., Fromm, B., Peterson, K. J., Scaria, V., Sivasubbu, S., & Pillai, B. (2018). Large scale changes in the transcriptome of Eisenia fetida during regeneration. *PLOS ONE*, *13*(9), e0204234. https://doi.org/10.1371/journal.pone.0204234

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bosco, G., Campbell, P., Leiva-Neto, J. T., & Markow, T. A. (2007). Analysis of Drosophila Species Genome Size and Satellite DNA Content Reveals Significant Differences Among Strains as Well as Between Species. *Genetics*, *177*(3), 1277. https://doi.org/10.1534/genetics.107.075069

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *bioRxiv*, 2021.06.03.446934. https://doi.org/10.1101/2021.06.03.446934

Cao, H., Wahlestedt, C., & Kapranov, P. (2018). Strategies to Annotate and Characterize Long Noncoding RNAs: Advantages and Pitfalls. *Trends in Genetics*, *34*(9), 704–721. https://doi.org/10.1016/j.tig.2018.06.002

Chan, P. P., Lin, B. Y., Mak, A. J., & Lowe, T. M. (2019). tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes. *bioRxiv*, 614032. https://doi.org/10.1101/614032

Chen, Y., Lun, A., & Smyth, G. (2016). From reads to genes to pathways: Differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]. *F1000Research*, *5*(1438). https://doi.org/10.12688/f1000research.8987.2

Chou, H.-C., Acevedo-Luna, N., Kuhlman, J. A., & Schneider, S. Q. (2018). PdumBase: A transcriptome database and research tool for Platynereis dumerilii and early development of other metazoans. *BMC Genomics*, *19*(1), 618. https://doi.org/10.1186/s12864-018-4987-0

Christodoulou, F., Raible, F., Tomer, R., Simakov, O., Trachana, K., Klaus, S., Snyman, H., Hannon, G. J., Bork, P., & Arendt, D. (2010). Ancient animal microRNAs and the evolution of tissue identity. *Nature*, *463*(7284), 1084–1088. https://doi.org/10.1038/nature08744

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Conzelmann, M., Williams, E. A., Krug, K., Franz-Wachtel, M., Macek, B., & Jékely, G. (2013). The neuropeptide complement of the marine annelid Platynereis dumerilii. *BMC Genomics*, *14*(1), 906. https://doi.org/10.1186/1471-2164-14-906

Dainat, J. (2020). *AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format.* (0.5.1) [Computer software]. https://github.com/NBISweden/AGAT

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

dos Reis, M., Thawornwattana, Y., Angelis, K., Telford, M. J., Donoghue, P. C. J., & Yang, Z. (2015). Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. *Current Biology*, *25*(22), 2939–2950. https://doi.org/10.1016/j.cub.2015.09.066

Eigen, M., & Winkler-Oswatitsch, R. (1981). Transfer-RNA, an early gene? *Naturwissenschaften*, *68*(6), 282–292. https://doi.org/10.1007/BF01047470

Emms, D. M., & Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 157. https://doi.org/10.1186/s13059-015-0721-2

Fauvel, P. (1914). Annélides polychètes non-pélagiques provenant des campagnes de l'Hirondelle et de la Princesse-Alice (1885-1910). *Résultats Des Campagnes Scientifiques Accompliés Par Le Prince Albert I*, *46*, 1–432.

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*, *26*(3), 314–325.

Fischer, A., Henrich, T., & Arendt, D. (2010). The normal development of Platynereis dumerilii (Nereididae, Annelida). *Frontiers in Zoology*, *7*(31). https://doi.org/10.1186/1742-9994-7-31

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451. https://doi.org/10.1073/pnas.1921046117

Fromm B, Høye E, Domanska D, Zhong X, Aparicio-Puerta E, Ovchinnikov V, Umu SU, Chabot PJ, Kang W, Aslanzadeh M, Tarbier M, Mármol-Sánchez E, Urgese G, Johansen M, Hovig E, Hackenberg M, Friedländer MR, Peterson KJ. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. Nucleic Acids Res. 2022 Jan 7;50(D1):D204-D210. doi: 10.1093/nar/gkab1101. Erratum in: Nucleic Acids Res. 2022 Nov 11;50(20):11991. PMID: 34850127; PMCID: PMC8728216

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A. M., & Koren, S. (2019). Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLOS Computational Biology*, *15*(8), e1007273. https://doi.org/10.1371/journal.pcbi.1007273

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

Haas, B. (2018). *TransDecoder* (5.5.0) [Computer software]. https://github.com/TransDecoder/TransDecoder/wiki

Hjelmen, C. E., Blackmon, H., Holmes, V. R., Burrus, C. G., & Johnston, J. S. (2019). Genome Size Evolution Differs Between Drosophila Subgenera with Striking Differences in Male and Female Genome Size in Sophophora. *G3 Genes|Genomes|Genetics*, *9*(10), 3167–3179. https://doi.org/10.1534/g3.119.400560

Hou, Y., & Lin, S. (2009). Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes. *PLOS ONE*, *4*(9), e6978. https://doi.org/10.1371/journal.pone.0006978

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Hui, J. H. L., Kortchagina, N., Arendt, D., Balavoine, G., & Ferrier, D. (2007). Duplication of the ribosomal gene cluster in the marine polychaete Platynereis dumerilii correlates with

ITS polymorphism. *Journal of the Marine Biological Association of the United Kingdom*, *87*(2), 443–449. https://doi.org/10.1017/S002531540705566X

Jha, A. N., Hutchinson, T. H., Mackay, J. M., Elliott, B. M., Pascoe, P. L., & Dixon, D. R. (1995). The chromosomes of Platynereis dumerilii (Polychaeta: Nereidae). *Journal of the Marine Biological Association of the United Kingdom*, *75*(3), 551–562.

Jyothi Kara, Cinthya S. G. Santos, Angus H. H. Macdonald, & Carol A. Simon. (2020). Resolving the taxonomic identities and genetic structure of two cryptic Platynereis Kinberg species from South Africa. *Invertebrate Systematics*, *34*(6), 618–636. https://doi.org/10.1071/IS19072

Kapusta, A. (2017). *Parsing-RepeatMasker-Outputs* [Computer software]. https://github.com/4ureliek/Parsing-RepeatMasker-Outputs

Knapp, J., Nakao, M., Yanagida, T., Okamoto, M., Saarma, U., Lavikainen, A., & Ito, A. (2011). Phylogenetic relationships within Echinococcus and Taenia tapeworms (Cestoda: Taeniidae): An inference from nuclear protein-coding genes. *Molecular Phylogenetics and Evolution*, *61*(3), 628–638. https://doi.org/10.1016/j.ympev.2011.07.022

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546. https://doi.org/10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278. https://doi.org/10.1186/s13059-019-1910-1

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., & Worsaae, K. (2015). Spiralian Phylogeny Informs the Evolution of Microscopic Lineages. *Current Biology*, *25*(15), 2000–2006. https://doi.org/10.1016/j.cub.2015.06.068

Lewin, Thomas D., Isabel Jiah-Yih Liao, and Yi-Jyun Luo. 2024. "Annelid Comparative Genomics and the Evolution of Massive Lineage-Specific Genome Rearrangement in Bilaterians." bioRxiv. https://doi.org/10.1101/2024.05.15.594353.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Liang, Y., Martín-Zamora, F. M., Guynes, K., Carrillo-Baltodano, A. M., Tan, Y., Moggioli, G., Seudre, O., Tran, M., Mortimer, K., Luscombe, N. M., Hejnol, A., Marlétaz, F., & Martín-Durán, J. M. (2022). Annelid functional genomics reveal the origins of bilaterian life cycles. *bioRxiv*, 2022.02.05.479245. https://doi.org/10.1101/2022.02.05.479245

López-Flores, I., & Garrido-Ramos, M. A. (2012). The Repetitive DNA Content of Eukaryotic Genomes. In *Genome Dynamics* (Vol. 7, pp. 1–28). https://doi.org/10.1159/000337118

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research*, *25*(5), 955–964. https://doi.org/10.1093/nar/25.5.955

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770. https://doi.org/10.1093/bioinformatics/btr011

Marlétaz, F., Peijnenburg, K. T. C. A., Goto, T., Satoh, N., & Rokhsar, D. S. (2019). A New Spiralian Phylogeny Places the Enigmatic Arrow Worms among Gnathiferans. *Current Biology*, *29*(2), 312-318.e3. https://doi.org/10.1016/j.cub.2018.11.042

Martín-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., Kerbl, A., Marquez, Y., Bekkouche, N., Chourrout, D., Gómez-Skarmeta, J. L., Irimia, M., Lenhard, B., Worsaae, K., & Hejnol, A. (2021). Conservative route to genome compaction in a miniature annelid. *Nature Ecology & Evolution*, *5*(2), 231–242. https://doi.org/10.1038/s41559-020-01327-6

Mutemi, K.N. 2023. PhD Dissertation.

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., & Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics*, *28*(14), 1919–1920. https://doi.org/10.1093/bioinformatics/bts277

Nip, K. M., Chiu, R., Yang, C., Chu, J., Mohamadi, H., Warren, R. L., & Birol, I. (2020). RNA-Bloom enables reference-free and reference-guided sequence assembly for

single-cell transcriptomes. *Genome Research*, *30*(8), 1191–1200. https://doi.org/10.1101/gr.260174.119

Özpolat, B. D., Randel, N., Williams, E. A., Bezares-Calderón, L. A., Andreatta, G., Balavoine, G., Bertucci, P. Y., Ferrier, D. E. K., Gambi, M. C., Gazave, E., Handberg-Thorsager, M., Hardege, J., Hird, C., Hsieh, Y.-W., Hui, J., Mutemi, K. N., Schneider, S. Q., Simakov, O., Vergara, H. M., … Arendt, D. (2021). The Nereid on the rise: Platynereis as a model system. *EvoDevo*, *12*(1), 10. https://doi.org/10.1186/s13227-021-00180-3

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000Research*, *9*(304). https://doi.org/10.12688/f1000research.23297.2

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Petersen. (2015). *SuperDeduper* [Computer software]. https://github.com/s4hts/HTStream

Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.-K., Benito-Gutiérrez, E., Dubchak, I., Garcia-Fernàndez, J., Gibson-Brown, J. J., Grigoriev, I. V., Horton, A. C., de Jong, P. J., Jurka, J., Kapitonov, V. V., … Rokhsar, D. S. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, *453*(7198), 1064–1071. https://doi.org/10.1038/nature06967

Putnam, N. H., Srivastava Mansi, Hellsten Uffe, Dirks Bill, Chapman Jarrod, Salamov Asaf, Terry Astrid, Shapiro Harris, Lindquist Erika, Kapitonov Vladimir V., Jurka Jerzy, Genikhovich Grigory, Grigoriev Igor V., Lucas Susan M., Steele Robert E., Finnerty John

R., Technau Ulrich, Martindale Mark Q., & Rokhsar Daniel S. (2007). Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science*, *317*(5834), 86–94. https://doi.org/10.1126/science.1139158

Raible, F., Tessmar-Raible, K., Osoegawa, K., Wincker, P., Jubin, C., Balavoine, G., Ferrier, D., Benes, V., de Jong, P., Weissenbach, J., Bork, P., & Arendt, D. (2005). Vertebrate-type intron-rich genes in the marine annelid Platynereis dumerilii. *Science*, *310*(5752), 1325–1326. https://doi.org/10.1126/science.1119089

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1432. https://doi.org/10.1038/s41467-020-14998-3

Read, G., & Fauchald, K. (2020). *World Polychaeta database. Platynereis Kinberg, 1865*. http://www.marinespecies.org/aphia.php?p=taxdetails&id=129381

Ribeiro, Rannyele P., Null, R.W., Ozpolat, B.D. 2024. "Sex-Biased Gene Expression Precedes Sexual Dimorphism in the Agonadal Annelid Platynereis Dumerilii." bioRxiv. https://doi.org/10.1101/2024.06.12.598746.

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1), 460. https://doi.org/10.1186/s12859-018-2485-7

Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdóttir, H., Mesirov, J. P., & Aiden, E. L. (2018). Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Systems*, *6*(2), 256-258.e1. https://doi.org/10.1016/j.cels.2018.01.001

Rota-Stabelli, O., Daley, A. C., & Pisani, D. (2013). Molecular Timetrees Reveal a Cambrian Colonization of Land and a New Scenario for Ecdysozoan Evolution. *Current Biology*, *23*(5), 392–398. https://doi.org/10.1016/j.cub.2013.01.026

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, *17*(2), 155–158. https://doi.org/10.1038/s41592-019-0669-3

Schaeffer, S. W. (2018). Muller "Elements" in Drosophila: How the Search for the Genetic Basis for Speciation Led to the Birth of Comparative Genomics. *Genetics*, *210*(1), 3–13. https://doi.org/10.1534/genetics.118.301084

Schultz, D. T., Haddock, S. H. D., Bredeson, J. V., Green, R. E., Simakov, O., & Rokhsar, D. S. (2023). Ancient gene linkages support ctenophores as sister to other animals. *Nature*, *618*(7963), 110–117. https://doi.org/10.1038/s41586-023-05936-6

Schultz, Darrin T., Elizabeth A. C. Heath-Heckman, Christopher J. Winchell, Dian-Han Kuo, Yun-Sang Yu, Fabian Oberauer, Kevin M. Kocot, Sung-Jin Cho, Oleg Simakov, and David A. Weisblat. 2024. "Acceleration of Genome Rearrangement in Clitellate Annelids." bioRxiv : The Preprint Server for Biology, May. https://doi.org/10.1101/2024.05.12.593736

Seemann, T. (2018). *BAsic Rapid Ribosomal RNA Predictor.* (0.9) [Computer software]. https://github.com/tseemann/barrnap

Simakov, O., Bredeson Jessen, Berkoff Kodiak, Marletaz Ferdinand, Mitros Therese, Schultz Darrin T., O'Connell Brendan L., Dear Paul, Martinez Daniel E., Steele Robert E., Green Richard E., David Charles N., & Rokhsar Daniel S. (2022). Deeply conserved synteny and the evolution of metazoan chromosomes. *Science Advances*, *8*(5), eabi5884. https://doi.org/10.1126/sciadv.abi5884

Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., Savage, R., Osoegawa, K., de Jong, P., Grimwood, J., Chapman, J. A., Shapiro, H., Aerts, A., Otillar, R. P., Terry, A. Y., … Rokhsar, D. S. (2013). Insights into bilaterian evolution from three spiralian genomes. *Nature*, *493*(7433), 526–531. https://doi.org/10.1038/nature11696

Simakov, O., Marlétaz, F., Yue, J.-X., O'Connell, B., Jenkins, J., Brandt, A., Calef, R., Tung, C.-H., Huang, T.-K., Schmutz, J., Satoh, N., Yu, J.-K., Putnam, N. H., Green, R. E., & Rokhsar, D. S. (2020). Deeply conserved synteny resolves early events in vertebrate evolution. *Nature Ecology & Evolution*, *4*(6), 820–830. https://doi.org/10.1038/s41559-020-1156-z

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*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Storer, J., Hubley, R., Rosen, J., Wheeler, T. J., & Smit, A. F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, *12*(1), 2. https://doi.org/10.1186/s13100-020-00230-y

Struck, T. H., & Purschke, G. (2005). The sister group relationship of Aeolosomatidae and Potamodrilidae (Annelida: "Polychaeta")—A molecular phylogenetic approach based on 18S rDNA and cytochrome oxidase I. *Zoologischer Anzeiger - A Journal of Comparative Zoology*, *243*(4), 281–293. https://doi.org/10.1016/j.jcz.2005.01.001

Struck, T. H., Schult, N., Kusen, T., Hickman, E., Bleidorn, C., McHugh, D., & Halanych, K. M. (2007). Annelid phylogeny and the status of Sipuncula and Echiura. *BMC Evolutionary Biology*, *7*(1), 57. https://doi.org/10.1186/1471-2148-7-57

Tandonnet, S., Koutsovoulos, G. D., Adams, S., Cloarec, D., Parihar, M., Blaxter, M. L., & Pires-daSilva, A. (2019). Chromosome-Wide Evolution and Sex Determination in the Three-Sexed Nematode Auanema rhodensis. *G3 Genes|Genomes|Genetics*, *9*(4), 1211–1230. https://doi.org/10.1534/g3.119.0011

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics*, *25*(1), 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25

Teixeira, M. A. L., Ngyren, A., Ravara, A., Vieira, P. E., Hernández, J. C., & Costa, F. O. (2021). The small polychaete Platynereis dumerilii revealed as a large species complex with fourteen MOTUs in European marine habitats. *ARPHA Conference Abstracts*, *4*, e64937. https://doi.org/10.3897/aca.4.e64937

Valvassori, G., Massa-Gallucci, A., & Gambi, M. C. (2015). Reappraisal Of Platynereis Massiliensis (Moquin-Tandon) (Annelida, Nereididae), A Neglected Sibling Species Of Platynereis Dumerilii (Audouin & Milne Edwards)/Platynereis Massiliensis (Moquin-Tandon) (Annelida, Nereididae), Una Specie Criptica Dimenticata Di Platynereis Dumerilii (Audouin & Milne Edwards). *Biologia Marina Mediterranea*, *22*(1), 113–116.

Vinther, J., Sperling, E. A., Briggs, D. E. G., & Peterson, K. J. (2012). A molecular palaeobiological hypothesis for the origin of aplacophoran molluscs and their derivation

from chiton-like ancestors. *Proceedings of the Royal Society B (Biological Sciences)*, *279*(1732). https://doi.org/10.1098/rspb.2011.1773

Wäge, J., Valvassori, G., Hardege, J. D., Schulze, A., & Gambi, M. C. (2017). The sibling polychaetes Platynereis dumerilii and Platynereis massiliensis in the Mediterranean Sea: Are phylogeographic patterns related to exposure to ocean acidification? *Marine Biology*, *164*(10), 199. https://doi.org/10.1007/s00227-017-3222-x

Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J. M., & Birol, I. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, *4*(1), 35. https://doi.org/10.1186/s13742-015-0076-3

Weigert, A., & Bleidorn, C. (2016). Current status of annelid phylogeny. *Organisms Diversity & Evolution*, *16*(2), 345–362. https://doi.org/10.1007/s13127-016-0265-7

Wu, L., Hiebert, L. S., Klann, M., Passamaneck, Y., Bastin, B. R., Schneider, S. Q., Martindale, M. Q., Seaver, E. C., Maslakova, S. A., & Lambert, J. D. (2020). Genes with spiralian-specific protein motifs are expressed in spiralian ciliary bands. *Nature Communications*, *11*(1), 4171. https://doi.org/10.1038/s41467-020-17780-7

Ye, J., Zhang, Y., Cui, H., Liu, J., Wu, Y., Cheng, Y., Xu, H., Huang, X., Li, S., Zhou, A., Zhang, X., Bolund, L., Chen, Q., Wang, J., Yang, H., Fang, L., & Shi, C. (2018). WEGO 2.0: A web tool for analyzing and plotting GO annotations, 2018 update. *Nucleic Acids Research*, *46*(W1), W71–W75. https://doi.org/10.1093/nar/gky400

Zakas, C., Harry, N. D., Scholl, E. H., & Rockman, M. V. (2021). The genome of the poecilogonous annelid Streblospio benedicti. *bioRxiv*, 2021.04.15.440069. https://doi.org/10.1101/2021.04.15.440069

Zakas, C., Harry, N. D., Scholl, E. H., & Rockman, M. V. (2022). The Genome of the Poecilogonous Annelid Streblospio benedicti. *Genome Biology and Evolution*, *14*(2), evac008. https://doi.org/10.1093/gbe/evac008

Zantke, J., Bannister, S., Rajan, V. B. V., Raible, F., & Tessmar-Raible, K. (2014). Genetic and Genomic Tools for the Marine Annelid Platynereis dumerilii. *Genetics*, *197*(1), 19–31. https://doi.org/10.1534/genetic.112.148254

Zhao, L., Wang, J., Li, Y., Song, T., Wu, Y., Fang, S., Bu, D., Li, H., Sun, L., Pei, D., Zheng, Y., Huang, J., Xu, M., Chen, R., Zhao, Y., & He, S. (2021). NONCODEV6: An updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Research*, *49*(D1), D165–D171. https://doi.org/10.1093/nar/gkaa1046

Zimin, A. V., & Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLOS Computational Biology*, *16*(6), e1007981. https://doi.org/10.1371/journal.pcbi.1007981

Zwarycz, A. S., Nossa, C. W., Putnam, N. H., & Ryan, J. F. (2015). Timing and Scope of Genomic Expansion within Annelida: Evidence from Homeoboxes in the Genome of the Earthworm Eisenia fetida. *Genome Biology and Evolution*, *8*(1), 271–281. https://doi.org/10.1093/gbe/evv243.

Tables

Table 1

Comparative genome assembly attributes in annelids

	Platynereis	***Streblospio***	***Helobdella***	***Capitella***	***Dimorphilus***	***Owenia***
Genome size	1.47 Gbp	701 Mbp	228 Mbp	324 Mbp	74 Mbp	519 Mbp
Karyotype	2n=28	2n=22	2n=18	2n=20	2n=24	2n=24
GC content	38%	38%	33%	40%	32%	35%
Protein-coding genes	28,985	20,221	23,426	31, 977	14,031	26,957
o Within OGs	23,598	16,783	16,866	26,880	11,786	
o No OGs	4,834	3,438	6,650	5,097	2,245	
RE content	51%	42%	33%	31%	11.44%	47.32%
tRNAs	4,719	595	168	1219	320	(undetermined)
rRNAs	91	37	184	138	90	(undetermined)
miRNAs	-	-	-		-	-

Table 2

Repeat element sizes in P. dumerilii

RE type	No. elements	mean ± SD (bp)	median (MAD) (bp)	min - max (bp)
DNA	96,971	362.21 ± 517.76	200 (173.46)	9 - 11,327
LINE	150,127	497.39 ± 644.24	258 (272.80)	9 - 13,478
Low complexity	22,053	70.45 ± 118.06	43 (16.31)	9 - 4,301
LTR	394,173	374.79 ± 729.65	192 (173.46)	9 - 21,967
Rolling-circles	6,196	359.00 ± 900.17	162 (80.60)	9 - 9,295
Satellite	8,663	644.00 ± 1,295.79	226 (263.90)	9 - 75,744
Simple repeats	225,376	62.48 ± 76.84	42 (23.72)	4 - 7,786
SINE	24,734	153.45 ± 83.79	154 (74.13)	9 - 1,100
SINE?	71	102.30 ± 7.54	104 (0)	44 - 106
snRNA	73	158.95 ± 76.40	212 (23.72)	37 - 243
*tRNA	16,985	196.19 ± 242.55	150 (115.64)	9 - 8,727
Unclassified	1,797,443	272.49 ± 348.39	170 (140.85)	9 - 106,830

Table 3

Protein-coding gene feature size and counts in annelids

Species	***Platynereis***	***Streblospio***	***Helobdella***	***Capitella***	***Dimorphilus***	***Owenia***
gene size	8,693 (9,939)	1,043 (1,290)	2,700 (2,322)	1,984 (1,739)	2,155 (1,524)	6,174 (6,933)
exon No.	4 (4)	3.69 ± 4.65**	4 (3)	4 (3)	5 (4)	5 (6)
exon size	604 (592)	270 (210)	183 (96)	213 (128)	345 (240)	273 (187)
intron No.	3 (4)	2.69 ± 4.65**	3 (3)	3 (3)	4 (4)	4 (6)
intron size	1,353 (1,316)	654 (522)	357 (249)	273 (283)	106 (63)	1,091 (882)
5’ UTR No.	1.36 ± 1.29	0.32 ± 0.82	0.2 ± 0.6	0.21 ± 0.56	1.31 ± 1.05	1.1 ± 0.9
5’ UTR size	143 (209)	76 (63)	89 (79)	46 (64)	85 (125)	94 (105)
3’ UTR No.	1.25 ± 1.11	0.26 ± 0.68	0.3 ± 0.7	0.32 ± 0.55	0.63 ± 1.07	0.96 ± 0.92
3’ UTR size	704 (833)	528 (428)	292 (224)	189 (208)	352 (302)	263 (286)

Note: To avoid the influence of outliers, data in the table - except the UTR counts which are presented as mean ± SD - are presented as median (median absolute deviation). The trends observed are also consistent when using the mean ± SD. The sizes are given in base pairs (bp). All values were rounded to the nearest small whole number. ** The median values gave 1 (0) and thus the mean ± SD was thought to be appropriate for estimating these values for *S. benedictii*.

Table 4

Assembly features of *Platynereis* genomes

	***P. dumerilii* (HiFi assembly) #**	***P. massiliensis***	***P. megalops***
'Predicted' genome size	~940 Mbp	~730 Mbp	~440 Mbp**
Assembly size	1.42 Gbp	1.43 Gbp	1.88 Gbp
Scaffold No.	163	1,478	370
N50	71 Mbp	4 Mbp	49 Mbp
L50	7	59	11
RE content	51%	58%	66%
Heterozygosity	3.2%	1.2%	10%**
BUSCO	96.0% complete (1.3% dupl.)	94.2% complete (2.5% dupl.)	94.8% complete (0.4% dupl.)

Note: The 'predicted' genome size and the heterozygosity are calculated based on short-read Illumina DNA sequencing using the same algorithm (see Materials and Methods). The polymorphism rate can greatly impact the estimated values and they are highly correlated (**Ranallo-Benavidez et al., 2020**). BUSCO version 5 was used to approximate genome completion for all three species. *dupl is the shortened text for duplication, a measure of how many conserved single-copies appear more than once in the assembly. # The *P. dumerilii* Hi-Fi assembly is not yet annotated, and was not the assembly used for the prior analyses. This genome will be annotated in unison with the other *Platynereis* genome assembly, for more in-depth comparative analyses.

Table 5

RE distribution in *Platynereis* species

	P. dumerilii	***P. massiliensis***	***P. megalops***
Total genome RE content	**50.38%**	**58.40%**	**65.96%**
Retroelements	14.83%	20.92%	20.56%
SINEs	*0.16%*	*0.63%*	*0.28%*
LINEs	*5.00%*	*7.22%*	*7.99%*
LTR elements	*9.67%*	*13.07%*	*12.28%*
DNA transposons	2.79%	5.06%	5.66%
Rolling-circles	0.12%	0.36%	0.26%
Unclassified	30.93%	30.56%	38.19%
Total interspersed repeats**	48.56%	56.54%	64.41%
Small RNA	0.36%	0.59%	0.11%
Satellites	0.23%	0.03%	0.001%
Simple-repeats	1.00	1.03%	1.08%
Low-complexity	0.11%	0.10%	0.09%

Note: The total interspersed repeats include all REs that are not found in tandem, which include all Retroelements, DNA transposons, Rolling-circles and Unclassified repeats.

Table 6 Analyses of Spiralian genomes at near-chromosomal level resolution

Animal	Species	Karyotype	ALGs (total of 24 elements)	ALG fusions
Annelids	*Acholoe squamosa*	-	all	A1⊗E, + 'spiralian'-fusions (J2⊗L⊗C2)
	Alentia gelatinosa	-	all	A1⊗E, I⊗N, + 'spiralian'-fusions (H⊗Q⊗J2⊗L⊗C2, O2⊗K⊗M)
	Alitta virens	-	all	C1⊗F, A1⊗E, + 'spiralian'-fusions (H⊗Q⊗G and J2⊗L⊗O2⊗K⊗C2⊗B1)
	Amphiduros pacificus	-	all	A2⊗J1, M⊗B1, D⊗G, + 'spiralian'-fusions (J2⊗L⊗C2⊗I, H⊗Q⊗B2, O1⊗R⊗N)
	Aporrectodea icterica	-	none	scrambled
	Bimastos eiseni	-	none	scrambled
	Branchellion lobata	-	none	scrambled
	Branchiopolynoe longqiensis	-	all	M⊗I, C1⊗H, A1⊗E, + 'spiralian'-fusions (J2⊗L⊗C2)
	Capitella teleta	10	all	I⊗G, + 'spiralian'-fusions (J2⊗L⊗C2)
	Dimorphilus gyrociliatus	12	missing: K, A2, M, R, O2, C2	NxI, A1xH, B2xB3, J1⊗F, B1⊗D, O1⊗G, C1⊗Q, P⊗E 'spiralian'-fusions (J2⊗L only)
	Harmothoe impar		all	A1⊗E, + 'spiralian'-fusions (J2⊗L⊗C2)
	Helobdella robusta	9	none	scrambled
	Lamellibrachia satsuma		all	A2xM, A1xD, B2⊗J1, D⊗P, + 'spiralian'-fusions (O1⊗R⊗B3 and J2⊗L⊗C2)
	Lepidonotus clava	-	all	A1⊗E, + 'spiralian'-fusions (J2⊗L⊗C2)
	Lumbricus rubellus	-	missing: H, L, C2, J1, B2, C1, P, I, O2, R, F, M, G, A1, A2, Q, N, B3, B1,	scrambled

			E, O1, D, J2, K	
	Metaphire vulgaris	-	none	scrambled
	Owenia fusiformis	12	all	C1⊗A2, B1⊗B3⊗C2, I⊗E, + 'spiralian'-fusions (O1⊗R⊗M and J2⊗L⊗P⊗B2⊗J1)
	Paraescarpia ochinospica	-	all	ExB1, A2xM, B2⊗J1, D⊗P, + 'spiralian'-fusions (O1⊗R⊗B3 and J2⊗L⊗C2)
	Piscicola geometra	16	none	scrambled
	Platynereis dumerilii	14	all	C1⊗F, A1⊗E, + 'spiralian'-fusions (H⊗Q⊗G and J2⊗L⊗O2⊗K⊗C2⊗B1)
	Protula sp. h YS-2021	-	all	C1⊗A1, B1⊗B2, I⊗A2, B3⊗J1⊗G, D⊗E, + 'spiralian'-fusions (J2⊗L⊗C2⊗P, O2⊗K⊗F)
	Sipunculus nudus	-	all	J1⊗M, I⊗E, + 'spiralian'-fusions (J2⊗L⊗C2)
	Sthenelais limicola	-	all	C1⊗M⊗B1, D⊗E, A1⊗G, J1⊗B3⊗N, + 'spiralian'-fusions (J2⊗L⊗I⊗C2, H⊗Q⊗F⊗B2, O2⊗K⊗P)
	Streblospio benedictii	11	all	M⊗G, C1⊗B2, P⊗N, A1⊗B3, A2⊗D, + 'spiralian'-fusions (O1⊗R⊗E, O2⊗K⊗J1, J2⊗L⊗C2⊗F)
	Terebella lapidaria	-	all	G⊗F + 'spiralian'-fusions (J2⊗L⊗C2, H⊗Q-P-E* *'mixing-without-fusion'*)
Brachiopods	*Lingula anatina*	10	all	Only the J2⊗L and H⊗Q spiralian fusions
Bryozoans	*Bugulina stolonifera*	-	missing: A2, B1, R, B2	C2⊗A1, C1⊗D⊗G⊗E⊗B3, N⊗F, I⊗L⊗O1⊗Q (only J2⊗L⊗J1⊗I⊗O1⊗Q and O2⊗K⊗P 'spiralian'-fusions)
	Membranipora membranicia	-	missing: J1, B2, C2, R, A2, B1, B3	C1⊗D⊗G⊗E, N⊗F, I⊗L⊗H⊗Q (only the O2⊗K, H⊗Q⊗P 'spiralian'-fusions)
	Cryptosula palasiana		missing: R, B2, O1, B1, J2, B3, A2	D⊗G⊗E⊗C1, C2⊗A1, I⊗L, N⊗F (only the O2⊗K⊗P , H⊗Q, 'spiralian'-fusions)
	Watersipora subatra	-	missing: A2, B1, B2, J1, R	C1⊗D⊗G⊗E⊗B3, N⊗F, C2⊗A1, O1⊗H, I⊗L⊗Q (only the L⊗J2⊗I⊗Q⊗O1, O2⊗K⊗P 'spiralian'-fusions)
Mollusks	*Acanthopleura granulata*	-	all	D⊗P, C1⊗F, + 'spiralian'-fusions (O1⊗R⊗I)
	Anadara kagoshimensis		all	P⊗B3 + 'spiralian'-fusions
	Crassostrea virginica	10	all	B1⊗C2, M⊗A2⊗B3⊗F, C1⊗A1, I⊗N⊗D⊗G⊗P + 'spiralian'-fusions (O1⊗R⊗B2⊗J1, H⊗Q⊗J2⊗L)

	Doryteuthis pealeii		missing: J2, B3, O2, R	scrambled (J1⊗L, C2⊗I, D⊗G, D⊗A2, K⊗M, G⊗A1, G⊗F, E⊗H, L⊗F)
	Gari tellinella		all	A2⊗B2, + 'spiralian'-fusions
	Mytilus edulis		all	D⊗A2, N⊗A1, I⊗C1, M⊗B1⊗B2, + 'spiralian'-fusions (J2⊗L⊗E)
	Octopus sinensis		C1, D, G, I, M, P, N, B1, B2, A1 and F	no 'spiralian'-fusion
	Pomacea canaliculata	14	all	G⊗J1, P⊗B2⊗C2, I⊗B1, + 'spiralian'-fusions (H⊗Q⊗B3, O1⊗R⊗F)
	Ruditapes philippinarum	19	all	A2⊗B2, + 'spiralian'-fusions
Nemerteans	*Lineus longissimus*		all	C1⊗G, + 'spiralian'-fusions
	Notospermus geniculatus	19	all	C1⊗G, + 'spiralian'-fusions (J2⊗L)
Phoronids	*Phoronis australis*		missing: C2, J2, K, B2, F, Q, J1, R and O2	fragmented assembly
Platyhelminthes	*Clonorchis sinensis*	28	missing: H, B2, C1, C2, R, O2, I, G, M, F, Q, B1, K, J2, O1, E	scrambled (A1⊗N, B3⊗A2⊗J1, L⊗D)
	Echinococcus granulosus	9	missing: B1, B3, N, E, O1, J2, Q, A2, P, I, R, O2, G, M, H, L, J1, B2, C2	scrambled (K⊗D)
	Hymenolepis microstoma	6	missing: G, I, P, R, O2, J1, B2, C2, H, L, K, O1, E, J2, D, B1, N, B3, Q,	scrambled + (A1⊗M)

			A2	
	Schistosoma japonica	8	missing: J2, E, B1, B3, Q, A2, G, M, R, O2, I, P, C1, B2, J1, C2, H	scrambled
	Schmidtea mediterranea	4	none	scrambled
	Taenia multiceps	-	K, O1, A1 and F	scrambled
Rotifers	*Adineta vaga*	6	none	scrambled
	Brachionus rubens	-	I, M, A2, N, B2, B3, A1 and F	A2⊗N⊗B2⊗B3, I⊗M
	Proales similies	-	missing: A2, B3, N, B1, O1, D, J2, K, H, L, C2, J1, P, O2, R, G	scrambled

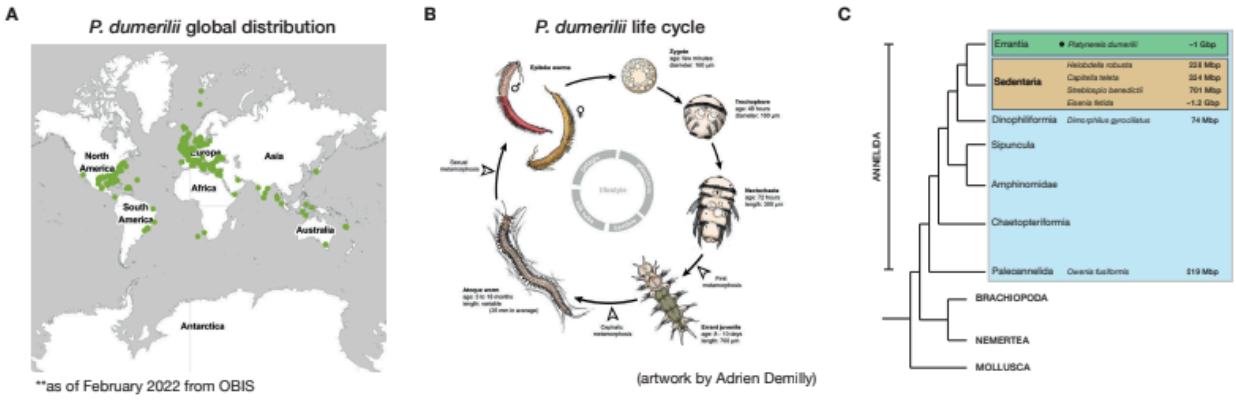


***Figure 1. Platynereis dumerilii* as a genetic model organism. A**, Global distribution of *P. dumerilii* animals, as of February 2022, according to observations by investigators worldwide, with data submitted to the open-access platform Ocean Biodiversity Information System (OBIS). Each green dot represents a specific location for which there has been a reported *P. dumerilii* observation or citing. **B**, a representative schematic summarizing the *P. dumerilii* life-cycle from Demilly et al 2013. **C**, a phylogenetic tree of major Spiralia/Lophotrochozoa groups, with the sequenced and 'annotated' annelid genome size estimates highlighted. The genome sizes are based off genome assemblies or DNA nuclei staining methods. The *Platynereis dumerilii* genome size numbers from (Jha et al., 1995), *Helobdella robusta* and *Capitella teleta* values were taken from (Simakov et al., 2013), the *Eisenia fetida* genome size estimates were taken from (Bhambri et al., 2018; Zwarycz et al., 2015), the *Streblospio benedictii* measurements were taken from (Zakas et al., 2022), the *Dimorphilus gyrociliatus* genome size from (Martín-Durán et al., 2021) and the *Owenia fusiformis*' from (Liang et al., 2022).

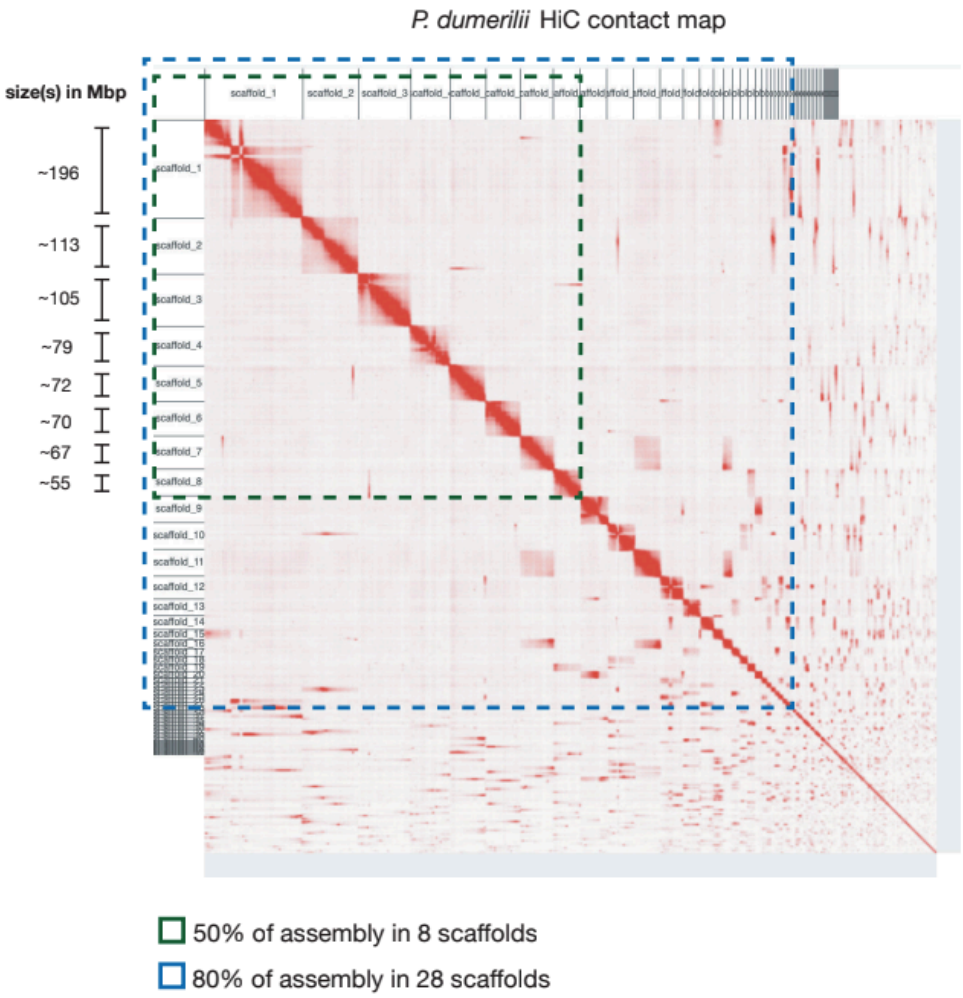


Figure 2. **A chromosomal scale *P. dumerilii* genome assembly.** A Hi-C contact map of all 330 *P. dumerilii* scaffolds. Highlighted in green are the 8 scaffolds that make up 50% of the assembly and in blue are the 28 scaffolds amounting to 80% of the assembly.

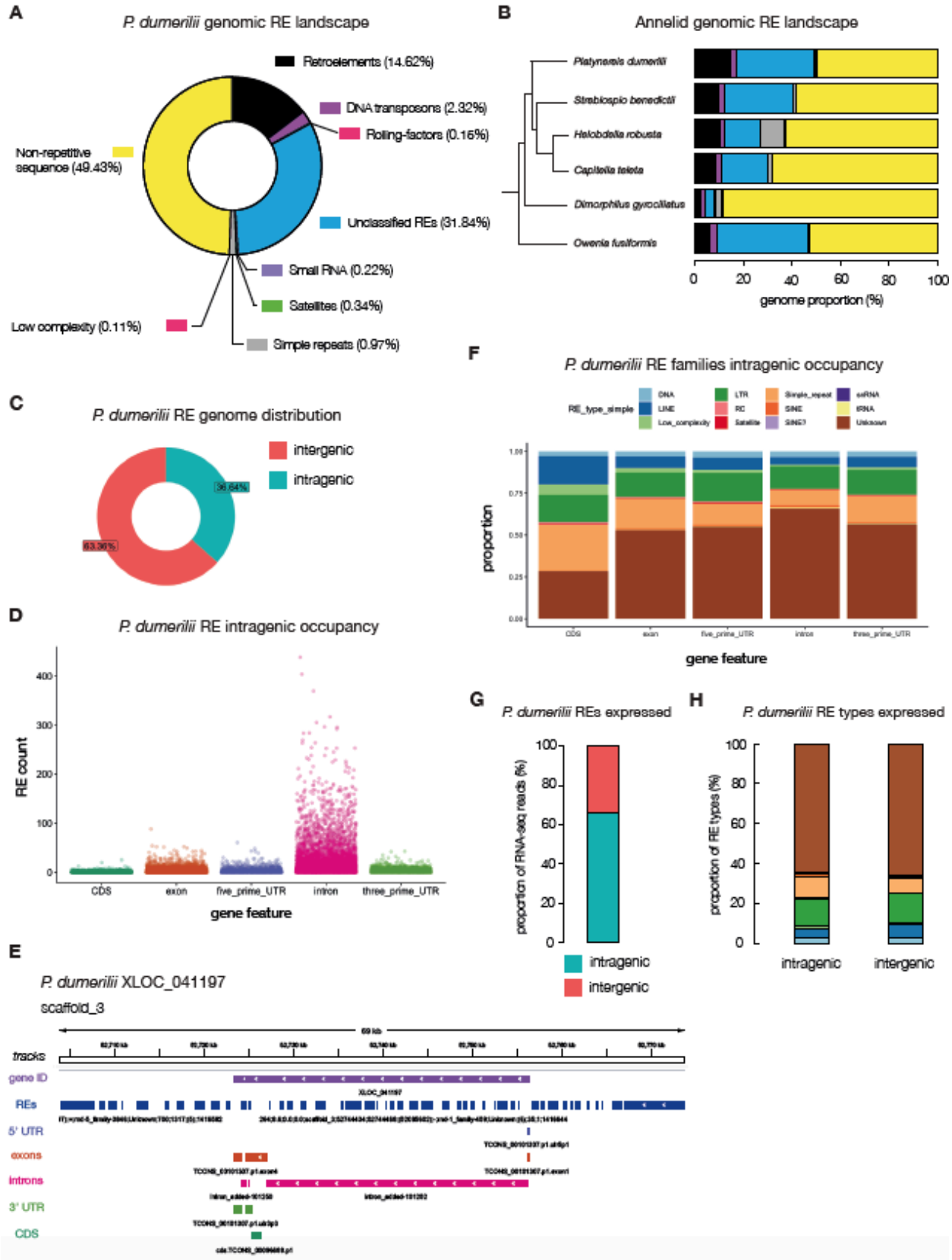

A
P. dumerilii genomic RE landscape
Retroelements (14.62%)
DNA transposons (2.32%)
Rolling-factors (0.16%)
Non-repetitive sequence (49.43%)
Unclassified REs (31.84%)
Small RNA (0.22%)
Satellites (0.34%)
Low complexity (0.11%)
Simple repeats (0.97%)
B
Annelid genomic RE landscape
Platynereis dumerilii
Streblospio benedicti
Helobdella robusta
Capitella teleta
Dimorphilus gyrociliatus
Owenia fusiformis
0
20
40
60
80
100
genome proportion (%)
C
P. dumerilii RE genome distribution
intergenic
intragenic
36.64%
63.36%
D
P. dumerilii RE intragenic occupancy
RE count
400
300
200
100
0
CDS
exon
five_prime_UTR
intron
three_prime_UTR
gene feature
E
P. dumerilii XLOC_041197
scaffold_3
tracks
gene ID
REs
5' UTR
exons
introns
3' UTR
CDS
F
P. dumerilii RE families intragenic occupancy
RE_type_simple
DNA
LINE
Low_complexity
LTR
RC
Satellite
Simple_repeat
SINE
SINE?
snRNA
tRNA
Unknown
proportion
1.00
0.75
0.50
0.25
0.00
CDS
exon
five_prime_UTR
intron
three_prime_UTR
gene feature
G
P. dumerilii REs expressed
proportion of RNA-seq reads (%)
100
80
60
40
20
0
intragenic
intergenic
H
P. dumerilii RE types expressed
proportion of RE types (%)
100
80
60
40
20
0
intragenic
intergenic

Figure 3. **The repeat-element landscape in *P. dumerilii*.** **A**, a doughnut plot illustrating the percentages of repeat and non-repeat elements found in the *P. dumerilii* genome. Percentages are of the total assembly (i.e. 49.43% of the entire genome is annotated as non-repetitive; yellow). **B**, annelids – whose relationships are shown in a phylogenetic tree – genomic repeat-element landscape. **C**, the distribution of intra – vs – inter-genic *P. dumerilii* repeat elements. **D**, counts of repeat elements represented as scatterplots within annotated intragenic regions of the *P. dumerilii* genome. **E**, an example gene locus (XLOC_041197) and its flanking regions on scaffold_3 highlighting repeat-element tracks (dark-blue) with the 5' and 3' UTRs (light-purple and green tracks respectively), exons (dark-orange track), introns (pink tracks) and the CDS (green tracks). **F**, proportion of repeat-element families and their occupancy at different intragenic regions. **G**, proportion of repeat-element specific RNA-seq reads mapping to intra – vs – inter-genic sites in *P. dumerilii*. **H**, proportion of RNA-seq reads mapping to intra – vs – inter-genic sites within specific RE types, colored according to the same legend in panel **F**.

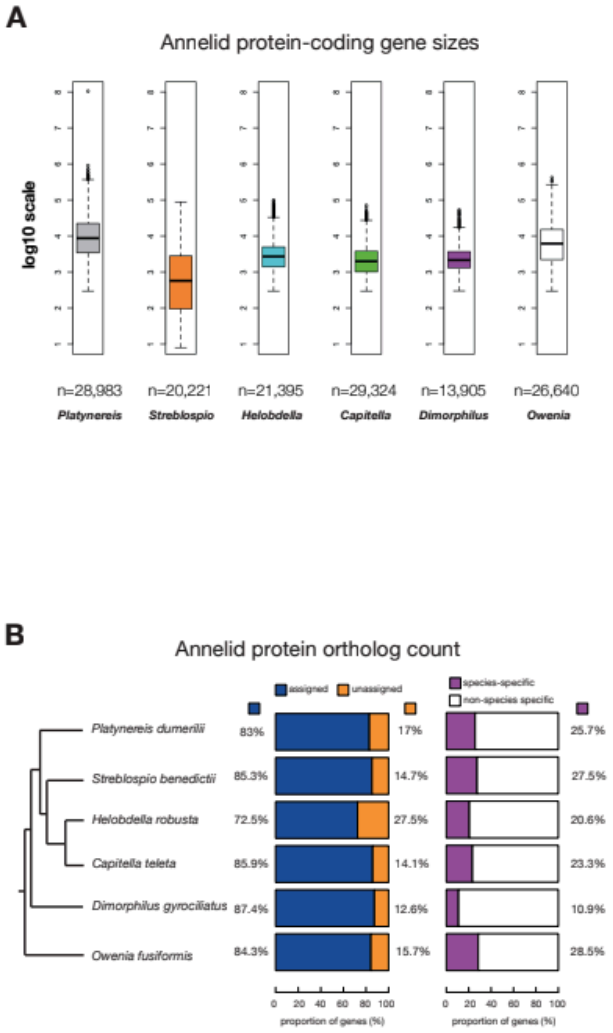


Figure 4. **The protein coding repertoire in annelids. A,** Annelid protein coding gene sizes plotted in $\log_{10}$ scale. The n values represent the total number of protein-coding genes that were measured for gene size, spanning the actual gene locus i.e. exons, introns and UTRs. The longest isoforms per gene were selected for the analysis. **B,** Proportion of annelid protein-coding genes in orthogroups.

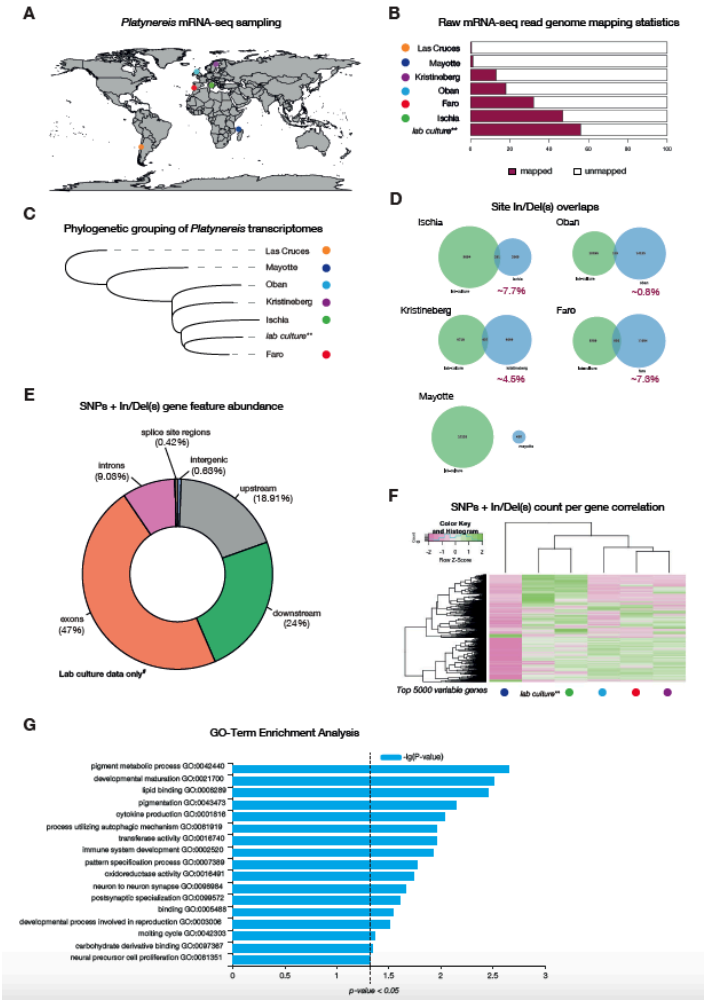


Figure 5. **Genomic and transcriptomic variation analyses on wild sampled *P. dumerilii*. A**, global map of sites of *P. dumerilii* mRNA sampling. **B,** histogram of raw mRNA-seq genome mapping percentages. **C**, phylogenetic grouping/sorting of wild sampled *Platynereis* transcriptomes via OrthoFinder. **D**, proportion of In/Del overlaps identified from the different *Platynereis* samples. **E**, gene feature abundance/occupancy of SNPs and In/Dels from mRNA-seq reads accessed from *P. dumerilii* lab cultures. **F**, SNP and In/Del counts from the same position on the genome correlation for the top 5,000 most variable genes (i.e. genes that showed the most variation in SNP and In/Del counts across the different sites). **G**, GO-term enrichment analysis of the top 5,000 variable genes.

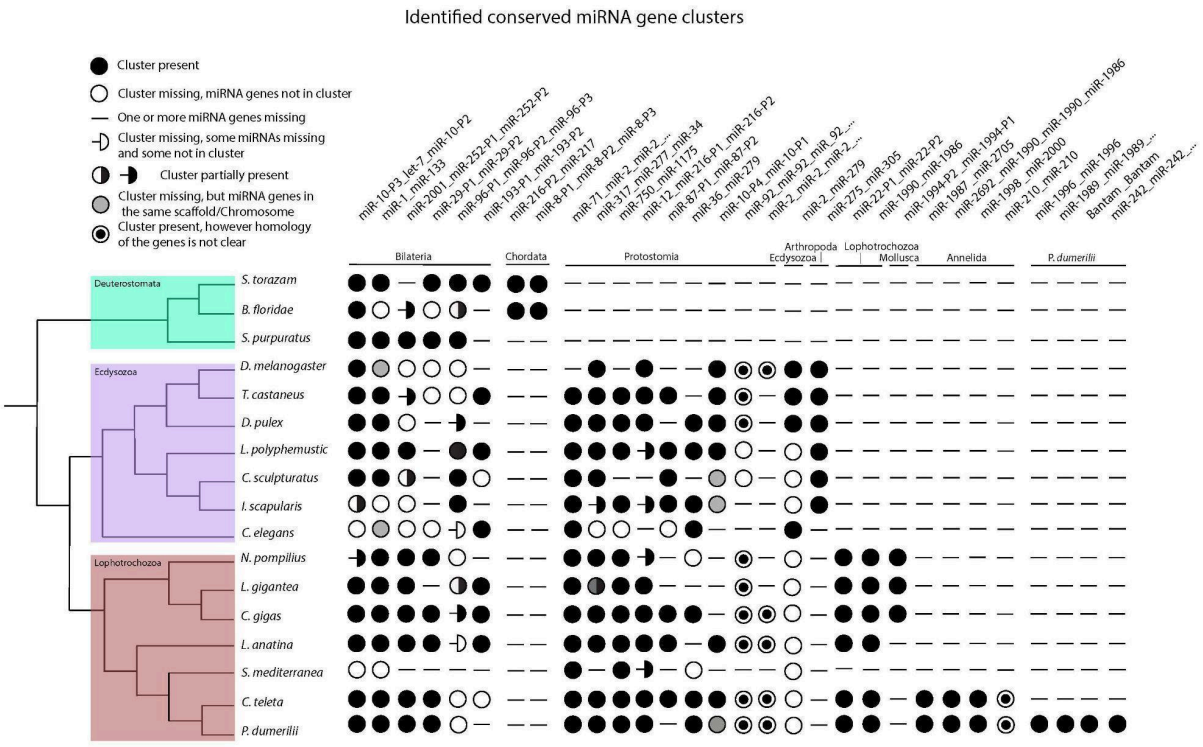


Figure 6. This scheme shows the distribution of phylogenetically conserved miRNA gene clusters in selected bilaterian phyla from the MirGeneDB database. Species are listed on the left, with *P. dumerilii* shown at the bottom. The tree, next to the species names, reflects state of the art of the lophotrochozoan clade phylogeny, with the branching taken from Marlétaz et al. (2019). The names of the miRNA clusters are defined by the comprised miRNA genes, separated by underscores (_), and are listed at the top of the figure. Since the order of miRNA genes in genomic clusters can vary between different species, the nomenclature follows three hierarchical criteria: 1. The gene order in the *P. dumerilii* genome; 2. The most common arrangement in the analysed species 3. Alphabetical order, when the first two criteria cannot be fulfilled. When the same miRNA gene name is repeated in the cluster, it indicates the presence of multiple copies, with uncertain homology, of the correspondent gene family. If the cluster name ends with three dots (...) more copies of the last listed gene are present. The number of copies can vary between

species. miRNA clusters are grouped and ordered according to their phylogenetic conservation, with the respective clades indicated just below the clusters. Here follows the description of the symbols. Full circle: the cluster listed above is present in the corresponding specie, with all genes included in the cluster. Empty circle: miRNA genes are found in the genome but not clustered together. Hyphen (-): corresponding miRNA genes are not found in the genome. Including when only one member of a two-gene cluster is missing. For clusters composed of three or more genes, there are three additional scenarios with respective symbols: 1. Hyphen followed by an empty half-circle: some miRNA genes are absent in the genome, while others are present but not clustered; 2. Circle half full: some miRNA genes are clustered, while others are not. 3. Hyphen followed by a full half-circle: some genes are clustered, while others are not present in the genome. Grey-filled circle: genes are not clustered but are found in the same genomic scaffold or chromosome. Circle with a filled center: indicates the presence of a cluster composed of gene copies with uncertain homology.

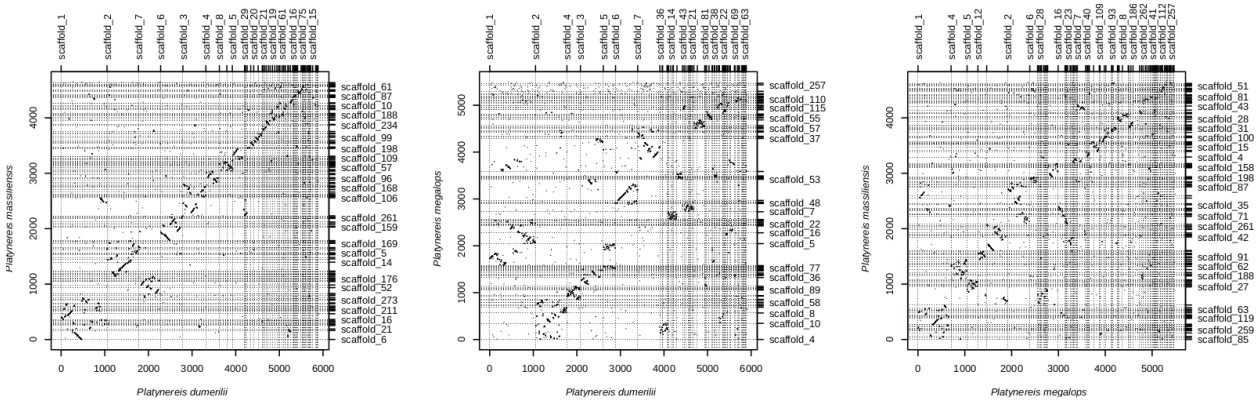


Figure 7. **Comparison of three *Platynereis* species.** Oxford dotplot comparison of the three *Platynereis* species genome assemblies. White homologies of most of the scaffolds can be identified, within scaffold inversions are common and present in all species.

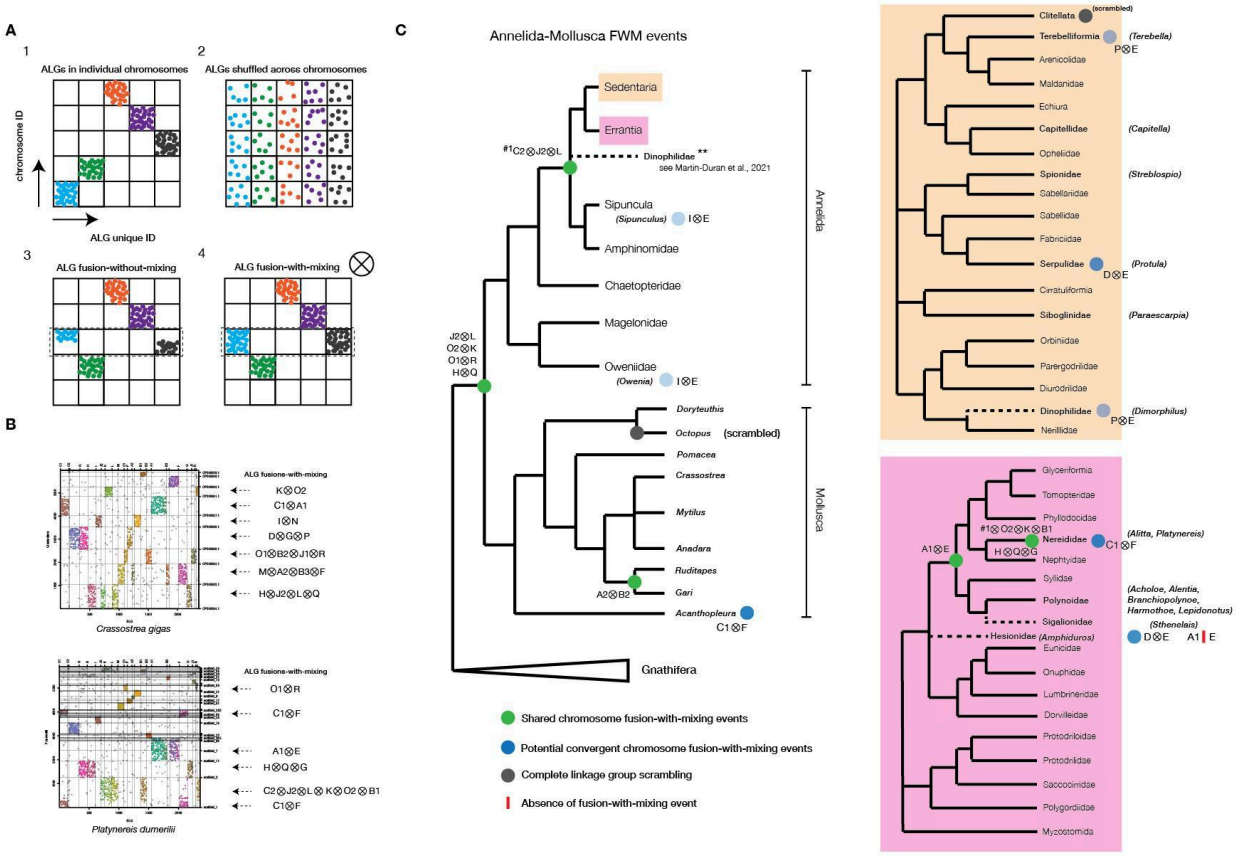
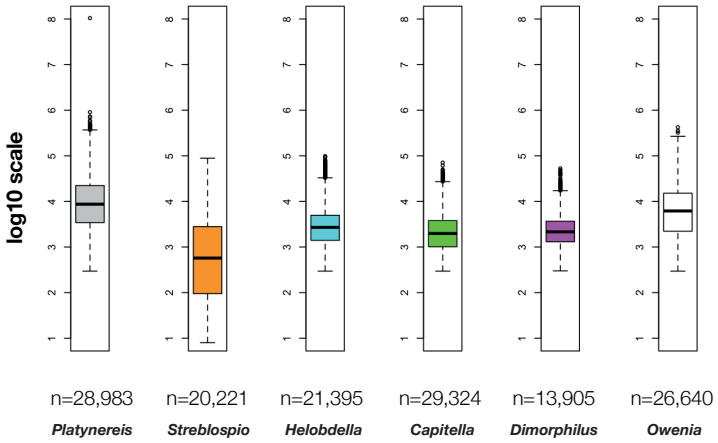


Figure 8. **Bilaterian Ancestral Linkage Group (bALG) fusion-with-mixing events towards and within Errantia. A,** potential linkage group evolutionary patterns described in animals. **B**, example annotation of bALGs and fusion-with-mixing (FWM) events detected in scaffolds or annotated chromosomes of *C.gigas* and *P. dumerilii* (this study). **C**, mapping of FWM events detected in this study onto the most up-to-date annelid-mollusc tree. Highlighted are the FWM events detected in annelid species belonging to specific groups within Sedentaria (orange) and Errantia (pink).

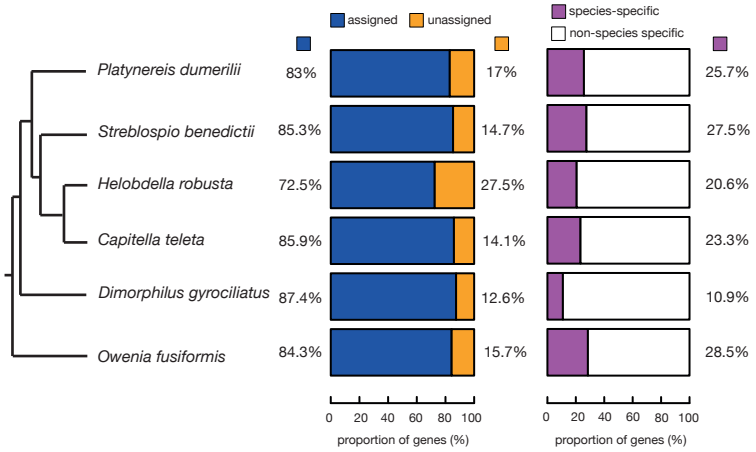
A

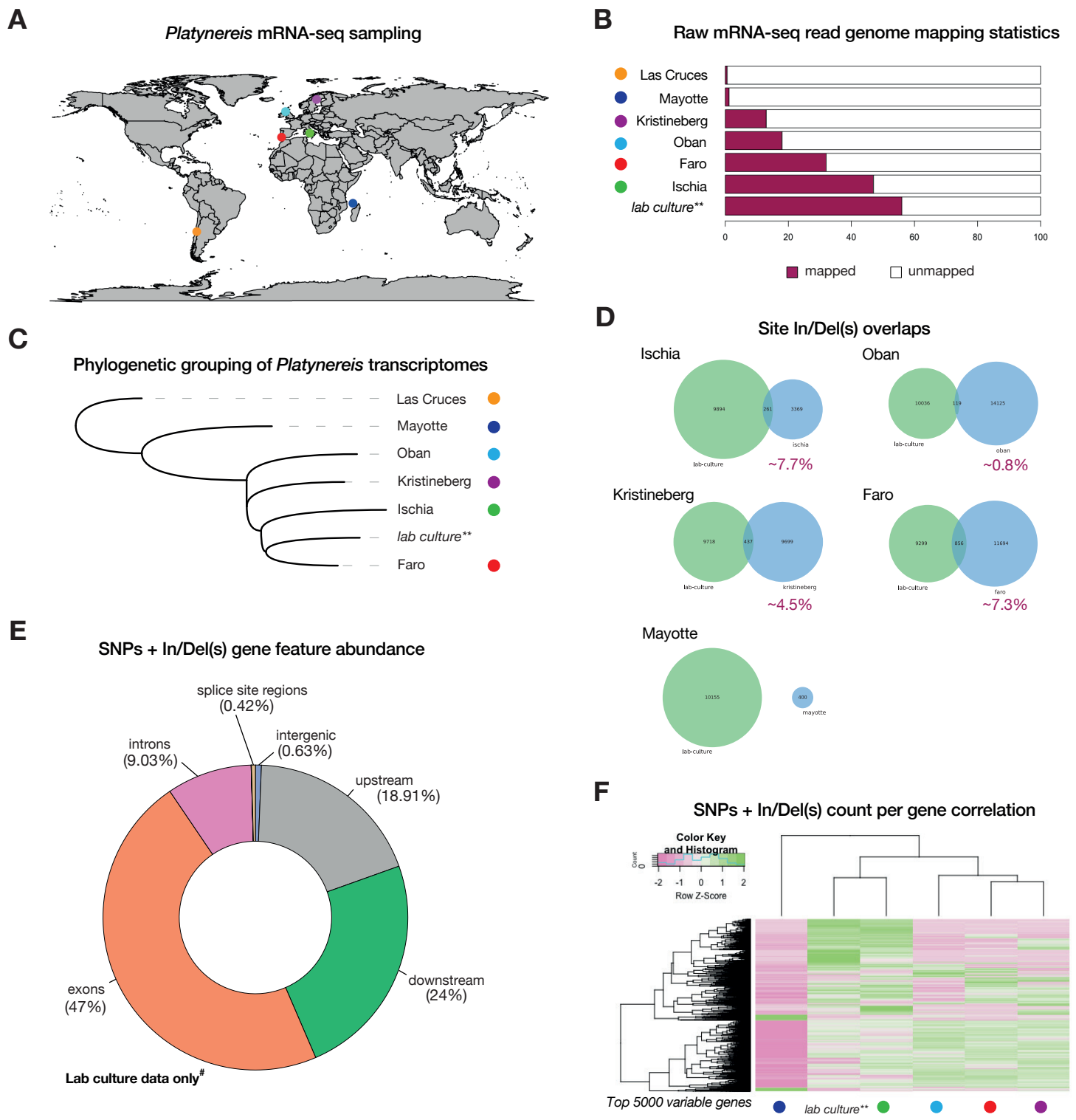
Annelid protein-coding gene sizes



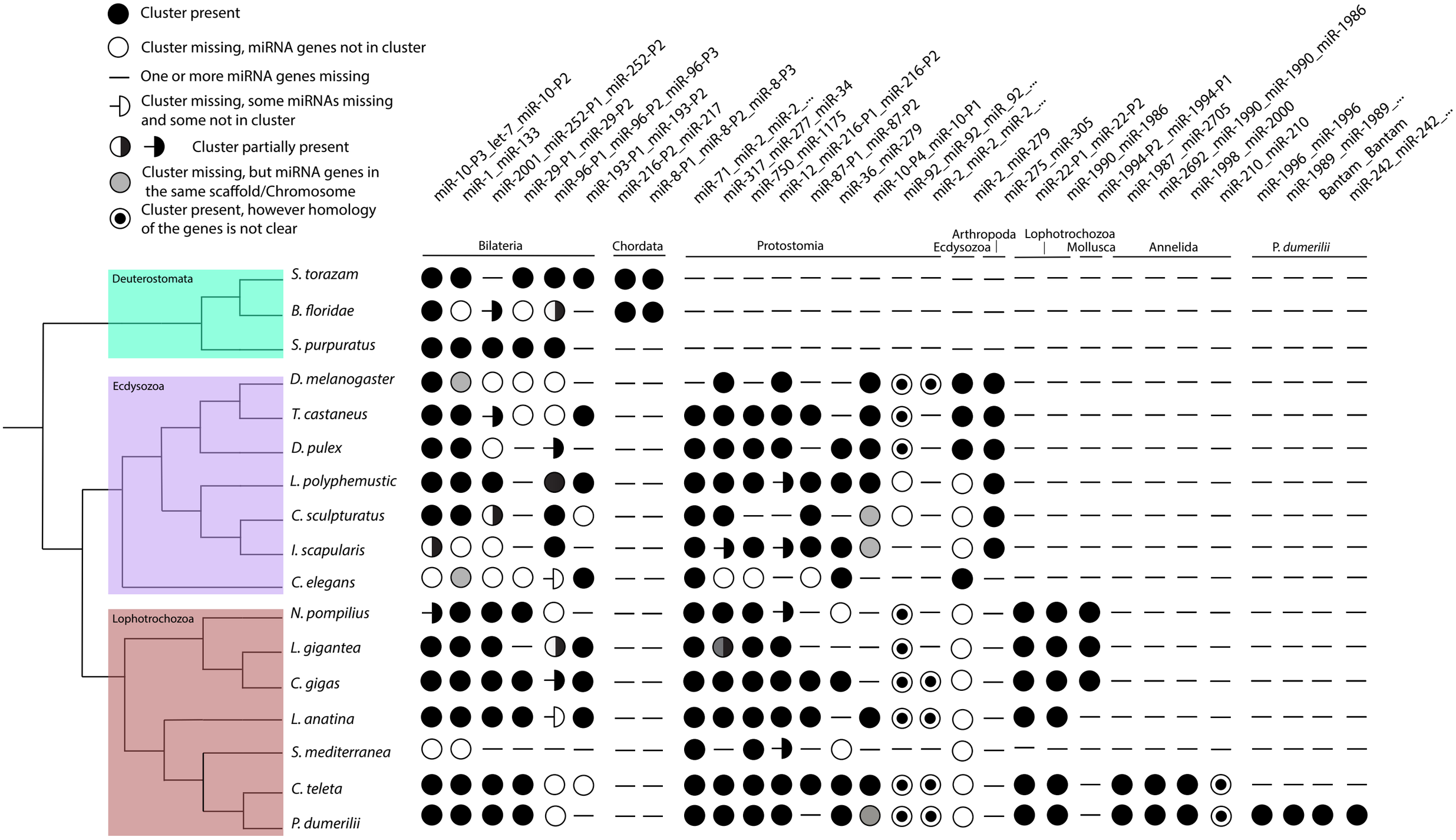
B

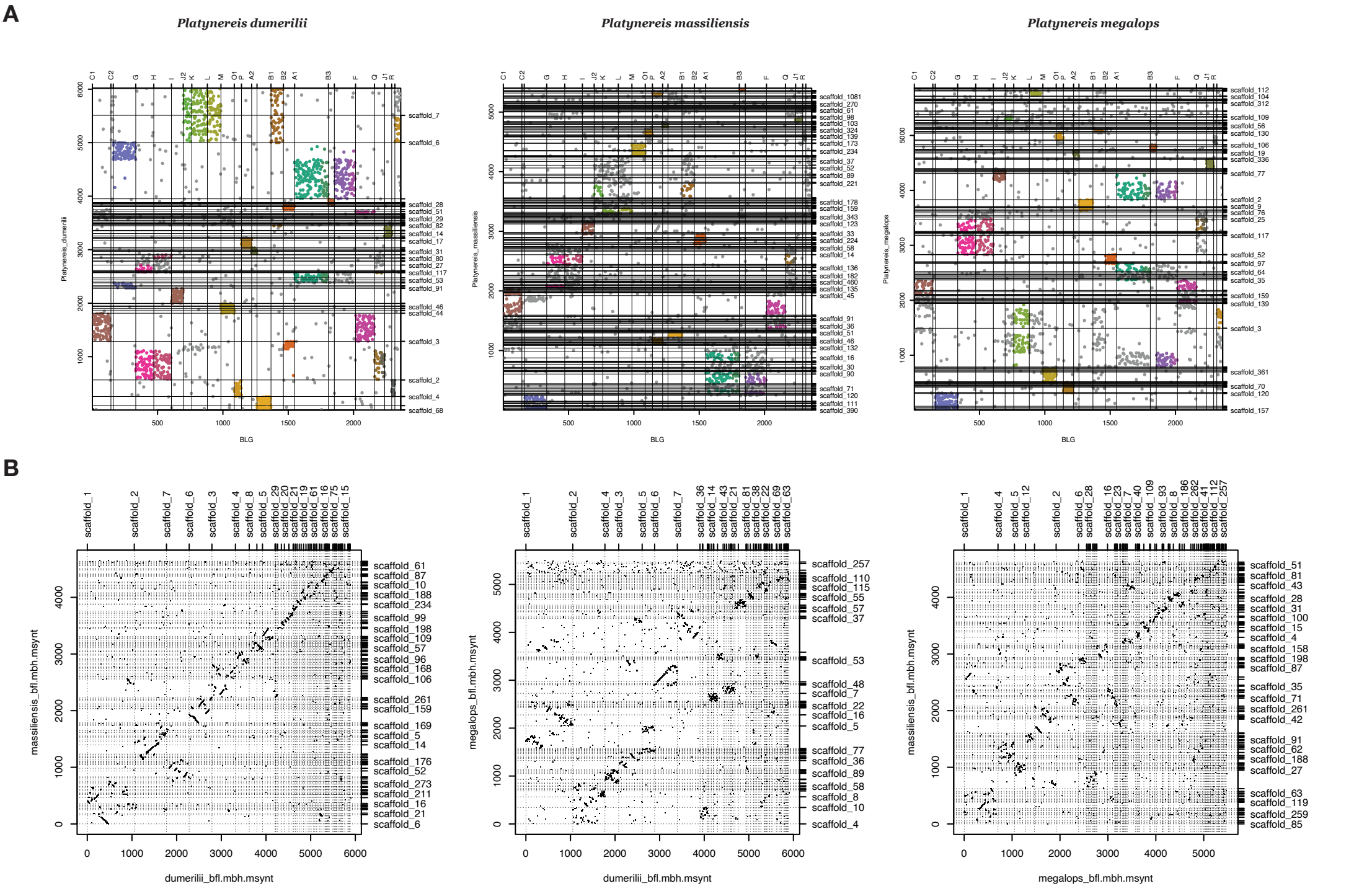
Annelid protein ortholog count

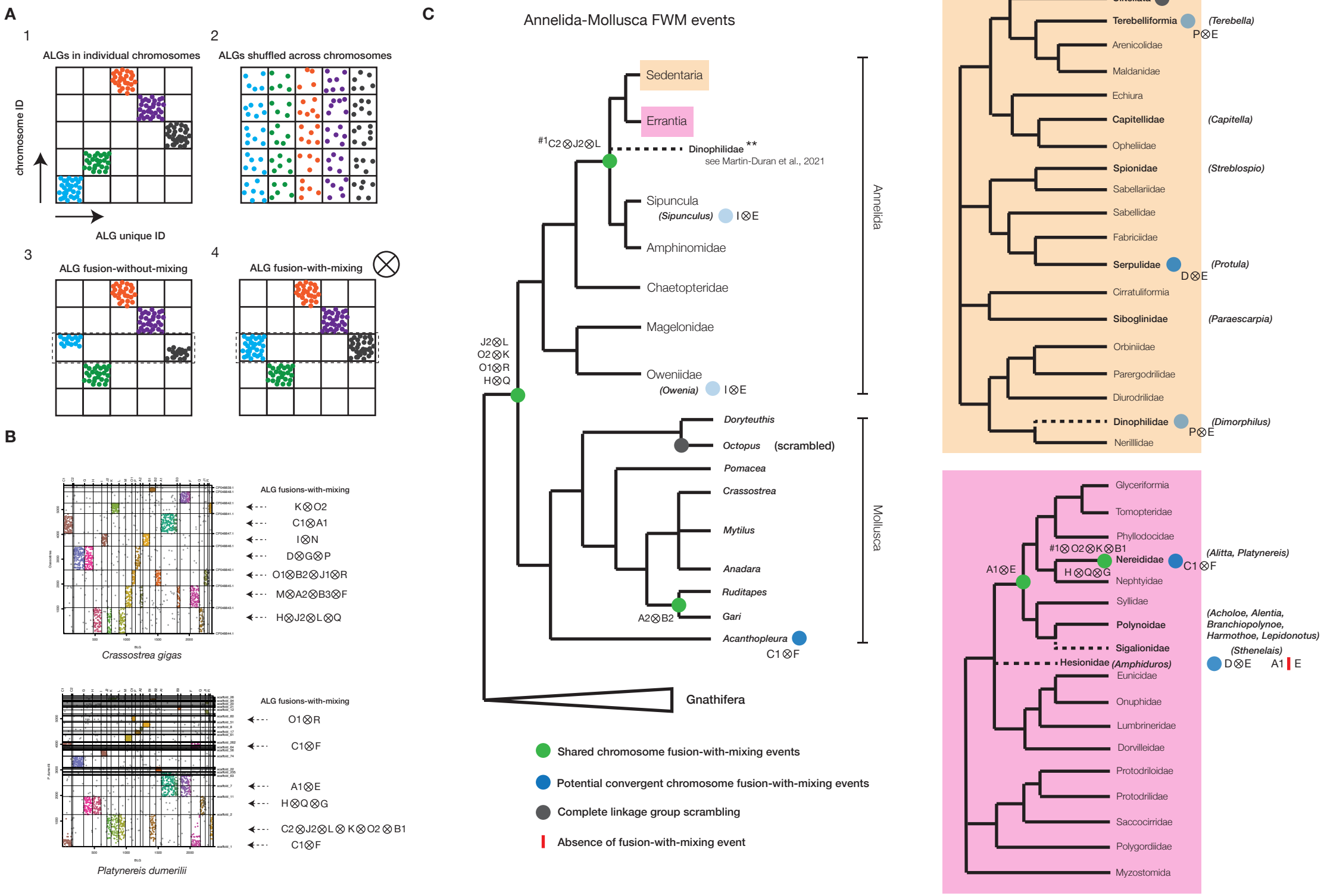



A
Platynereis mRNA-seq sampling
B
Raw mRNA-seq read genome mapping statistics
Las Cruces
Mayotte
Kristineberg
Oban
Faro
Ischia
lab culture**
0
20
40
60
80
100
mapped
unmapped
C
Phylogenetic grouping of Platynereis transcriptomes
Las Cruces
Mayotte
Oban
Kristineberg
Ischia
lab culture**
Faro
D
Site In/Del(s) overlaps
Ischia
Oban
~7.7%
~0.8%
Kristineberg
Faro
~4.5%
~7.3%
Mayotte
E
SNPs + In/Del(s) gene feature abundance
splice site regions (0.42%)
intergenic (0.63%)
introns (9.03%)
upstream (18.91%)
downstream (24%)
exons (47%)
Lab culture data only#
F
SNPs + In/Del(s) count per gene correlation
Color Key and Histogram
Count
-2
-1
0
1
2
Row Z-Score
Top 5000 variable genes
lab culture**
G
GO-Term Enrichment Analysis
-lg(P-value)
pigment metabolic process GO:0042440
developmental maturation GO:0021700
lipid binding GO:0008289
pigmentation GO:0043473
cytokine production GO:0001816
process utilizing autophagic mechanism GO:0061919
transferase activity GO:0016740
immune system development GO:0002520
pattern specification process GO:0007389
oxidoreductase activity GO:0016491
neuron to neuron synapse GO:0098984
postsynaptic specialization GO:0099572
binding GO:0005488
developmental process involved in reproduction GO:0003006
molting cycle GO:0042303
carbohydrate derivative binding GO:0097367
neural precursor cell proliferation GO:0061351
0
0.5
1
1.5
2
2.5
3
p-value < 0.05

Identified conserved miRNA gene clusters




A
Platynereis dumerilii
Platynereis massiliensis
Platynereis megalops
Platynereis_dumerilii
Platynereis_massiliensis
Platynereis_megalops
BLG
B
massiliensis_bfl.mbh.msynt
dumerilii_bfl.mbh.msynt
megalops_bfl.mbh.msynt


A
1
ALGs in individual chromosomes
2
ALGs shuffled across chromosomes
chromosome ID
ALG unique ID
3
ALG fusion-without-mixing
4
ALG fusion-with-mixing
B
ALG fusions-with-mixing
K⊗O2
C1⊗A1
I⊗N
D⊗G⊗P
O1⊗B2⊗J1⊗R
M⊗A2⊗B3⊗F
H⊗J2⊗L⊗Q
Crassostrea gigas
ALG fusions-with-mixing
O1⊗R
C1⊗F
A1⊗E
H⊗Q⊗G
C2⊗J2⊗L⊗K⊗O2⊗B1
C1⊗F
Platynereis dumerilii
C
Annelida-Mollusca FWM events
Sedentaria
Errantia
#1C2⊗J2⊗L
Dinophilidae **
see Martin-Duran et al., 2021
Sipuncula
(Sipunculus) I⊗E
Amphinomidae
Chaetopteridae
Magelonidae
Oweniidae
(Owenia) I⊗E
J2⊗L
O2⊗K
O1⊗R
H⊗Q
Annelida
Doryteuthis
Octopus (scrambled)
Pomacea
Crassostrea
Mytilus
Anadara
Ruditapes
Gari
A2⊗B2
Acanthopleura
C1⊗F
Mollusca
Gnathifera
Shared chromosome fusion-with-mixing events
Potential convergent chromosome fusion-with-mixing events
Complete linkage group scrambling
Absence of fusion-with-mixing event
Clitellata (scrambled)
Terebelliformia (Terebella)
P⊗E
Arenicolidae
Maldanidae
Echiura
Capitellidae (Capitella)
Opheliidae
Spionidae (Streblospio)
Sabellariidae
Sabellidae
Fabriciidae
Serpulidae (Protula)
D⊗E
Cirratuliformia
Siboglinidae (Paraescarpia)
Orbiniidae
Parergodrilidae
Diurodrilidae
Dinophilidae (Dimorphilus)
P⊗E
Nerilliidae
Glyceriformia
Tomopteridae
Phyllodocidae
#1⊗O2⊗K⊗B1
Nereididae (Alitta, Platynereis)
C1⊗F
A1⊗E
H⊗Q⊗G
Nephtyidae
Syllidae
Polynoidae (Acholoe, Alentia, Branchiopolynoe, Harmothoe, Lepidonotus)
Sigalionidae (Sthenelais)
Hesionidae (Amphiduros)
D⊗E
A1 E
Eunicidae
Onuphidae
Lumbrineridae
Dorvilleidae
Protodriloidae
Protodrilidae
Saccocirridae
Polygordiidae
Myzostomida

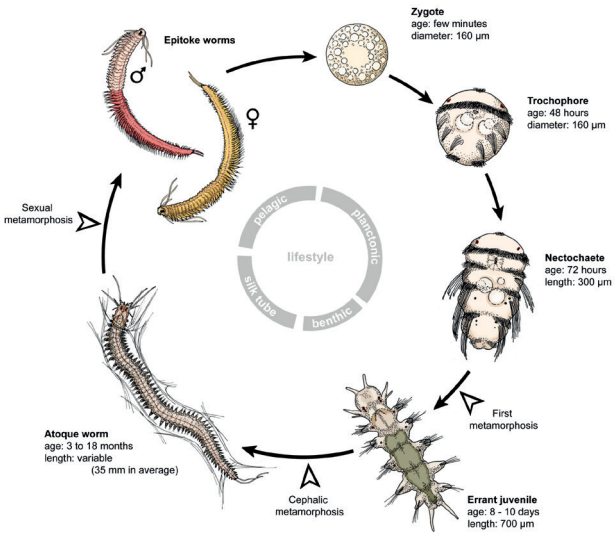
A

P. dumerilii global distribution

**as of February 2022 from OBIS

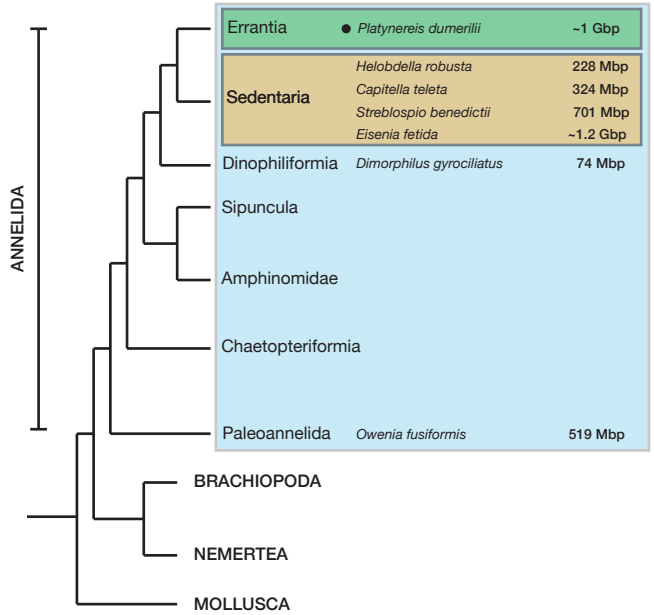
B

P. dumerilii life cycle

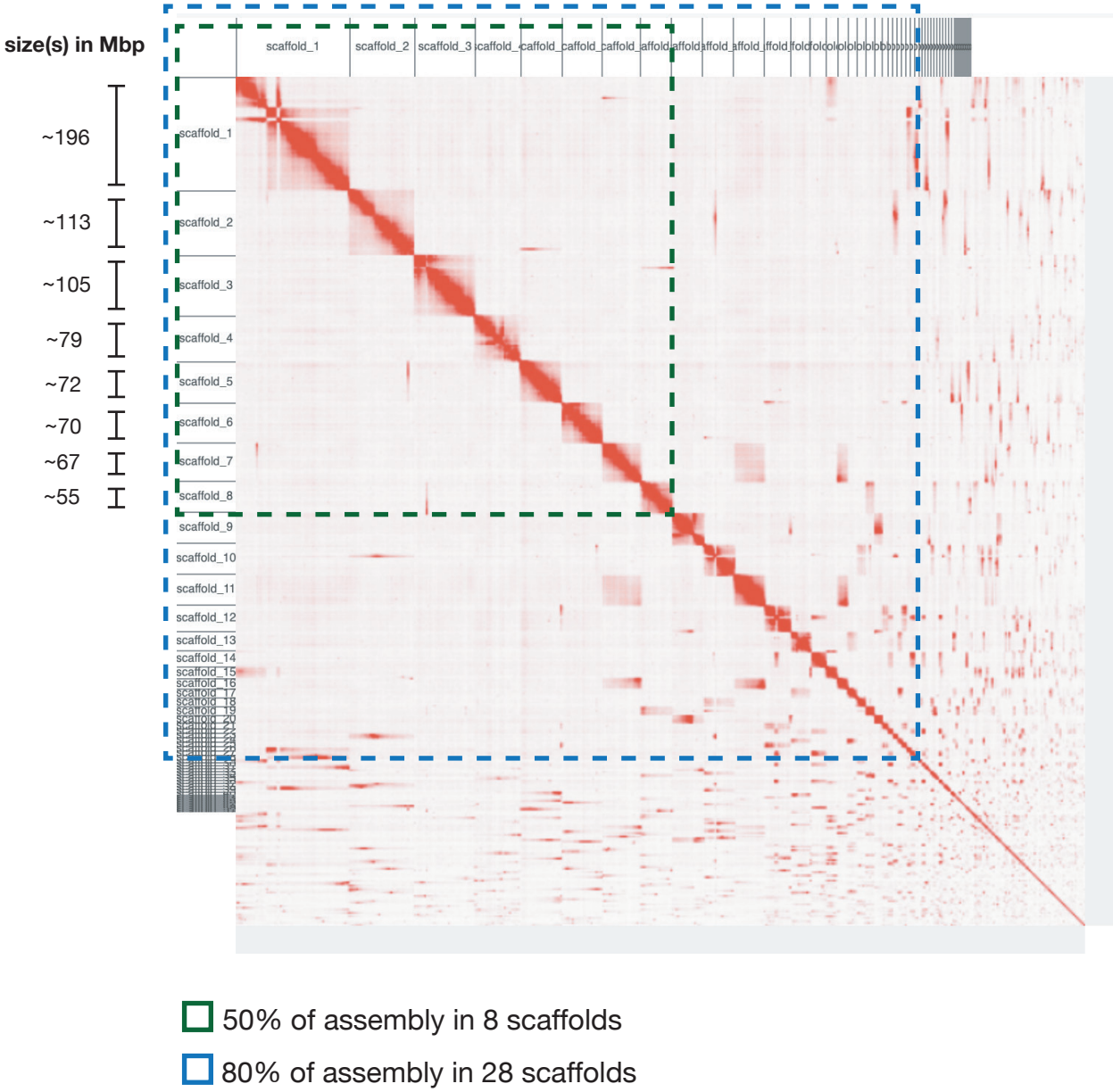


(artwork by Adrien Demilly)

C

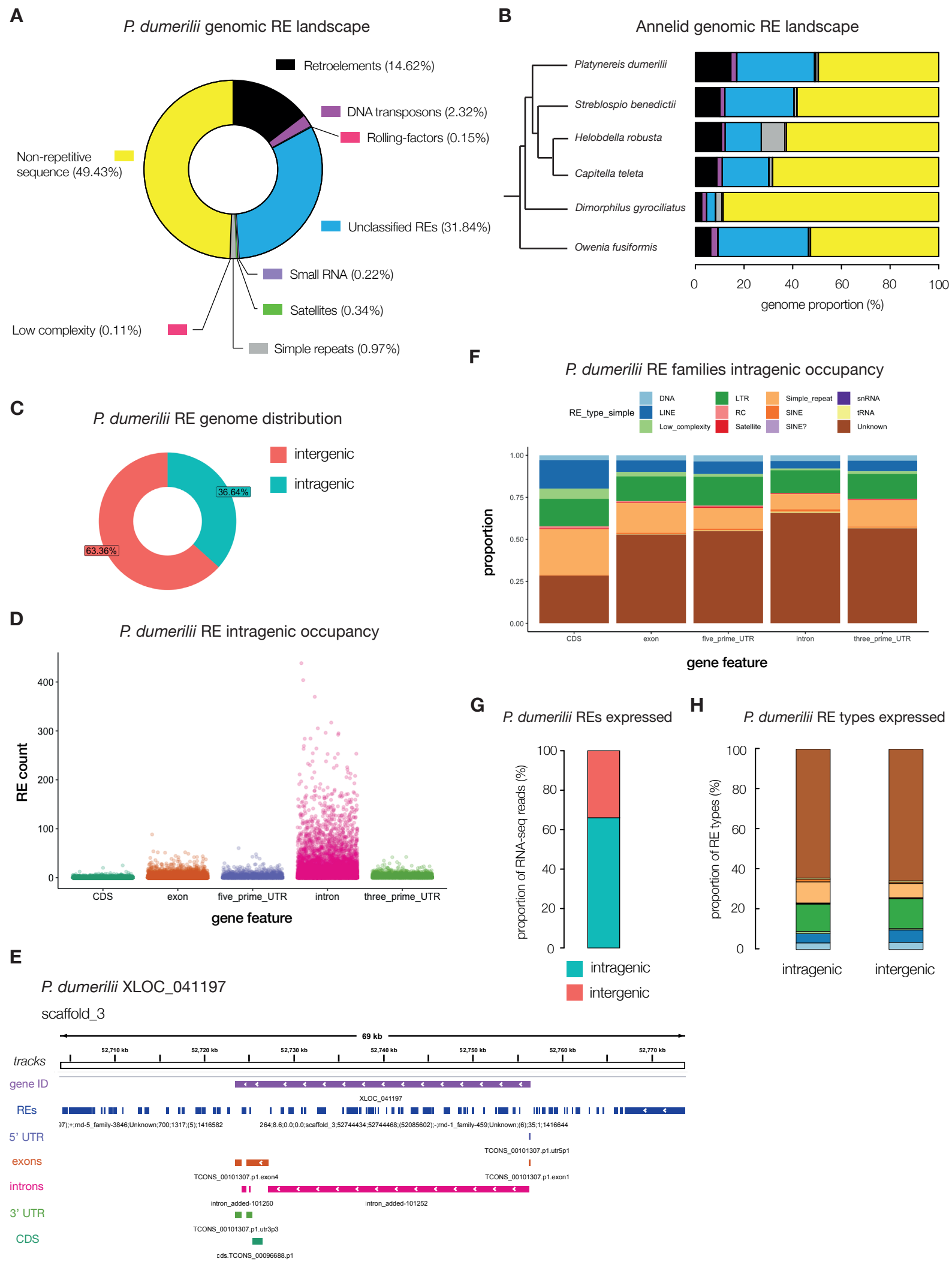


P. dumerilii HiC contact map



50% of assembly in 8 scaffolds

80% of assembly in 28 scaffolds


A
P. dumerilii genomic RE landscape
Retroelements (14.62%)
DNA transposons (2.32%)
Rolling-factors (0.15%)
Non-repetitive sequence (49.43%)
Unclassified REs (31.84%)
Small RNA (0.22%)
Satellites (0.34%)
Low complexity (0.11%)
Simple repeats (0.97%)
B
Annelid genomic RE landscape
Platynereis dumerilii
Streblospio benedictii
Helobdella robusta
Capitella teleta
Dimorphilus gyrociliatus
Owenia fusiformis
0
20
40
60
80
100
genome proportion (%)
C
P. dumerilii RE genome distribution
intergenic
intragenic
36.64%
63.36%
D
P. dumerilii RE intragenic occupancy
RE count
400
300
200
100
0
CDS
exon
five_prime_UTR
intron
three_prime_UTR
gene feature
F
P. dumerilii RE families intragenic occupancy
RE_type_simple
DNA
LINE
Low_complexity
LTR
RC
Satellite
Simple_repeat
SINE
SINE?
snRNA
tRNA
Unknown
proportion
1.00
0.75
0.50
0.25
0.00
CDS
exon
five_prime_UTR
intron
three_prime_UTR
gene feature
G
P. dumerilii REs expressed
proportion of RNA-seq reads (%)
100
80
60
40
20
0
intragenic
intergenic
H
P. dumerilii RE types expressed
proportion of RE types (%)
100
80
60
40
20
0
intragenic
intergenic
E
P. dumerilii XLOC_041197
scaffold_3
69 kb
52,710 kb
52,720 kb
52,730 kb
52,740 kb
52,750 kb
52,760 kb
52,770 kb
tracks
gene ID
XLOC_041197
REs
5' UTR
TCONS_00101307.p1.utr5p1
exons
TCONS_00101307.p1.exon4
TCONS_00101307.p1.exon1
introns
intron_added-101250
intron_added-101252
3' UTR
TCONS_00101307.p1.utr3p3
CDS
cds.TCONS_00096688.p1