Draft genomes of the three northern hemisphere blue mussel lineages: North and South European *Mytilus edulis* and Mediterranean *Mytilus galloprovincialis*.

Alexis Simon[1,*], Christine Arbiol[1], and Nicolas Bierne[1]

[1]CNRS, Univ Montpellier, …

June 17, 2022

Abstract

Using the 10X chromium long reads technology, we provide draft genomes for three closely related blue mussel species from the *Mytilus* species complex. The objective was to produce affordable genomic resources for population and evolutionary genomic studies. Genomes are fragmented but represent a large portion of the genome, with good sizes and BUSCO scores.

Keywords: *Mytilus edulis*, *Mytilus galloprovincialis*, Genome assembly, 10X chromium

* **Corresponding author**: alexis.simon@normalesup.org

1 Rationale and objectives

The *Mytilus* species complex has been a model system in population genetics, adaptation, hybridization and speciation since genetic variants could be identified (Ahmad et al., 1977; Bierne et al., 2003; Fraïsse et al., 2016; Koehn & Mitton, 1972; Milkman & Beaty, 1970; Quesada, Wenne, et al., 1995; Simon et al., 2021; Skibinski et al., 1978).

The *Mytilus* species complex is composed of three taxonomically recognized and partially reproductively isolated species in the northern hemisphere, *Mytilus edulis*, *M. galloprovincialis* and *M. trossulus*. Within each species, evolutionary relevant lineages can be identified. Two lineages of *M. galloprovincialis* can be identified, one Atlantic lineage and one Mediterranean lineage separated by a hybrid zone along the Almeria-Oran front (El Ayari et al., 2019; Fraïsse et al., 2016; Quesada, Zapata, et al., 1995). Three lineages of *M. edulis* can be identified: (i) an American lineage, (ii) a Southern European lineage, and (iii) a Northern European lineage (Fraïsse et al., 2016; Simon et al., 2020).

As an effort to diversify the genomic resources available for the *Mytilus* species complex, we assembled and annotate the genomes of three lineages using the 10X chromium technology. While our assemblies were initially fragmented due to the high level of heterozygosity, we leveraged the existence of a chromosome scale assembly of a sister species to scaffold our assemblies against it. We obtained assemblies equivalent to published ones in term of completeness for three new lineages of the species complex using a low sequencing budget and publicly available data. The resources produced and the assembly pipeline are freely available for use by the community.

2 Methods

General notes: The entire assembly was carried out using a Snakemake (Mölder et al., 2021) pipeline available on github at https://github.com/alxsimon/assembly_10x. Where deemed important, parameters are given in the text. For brevity and simplicity, not all information might be available in the text. However, all parameters, software versions and steps are retrievable from the repository.

Important caveat: The assembled genome of MeduEUN (*M. edulis* Northern lineage) was initially thought to be *M. trossulus*. Therefore some assembly and annotation steps wrongly used *M. trossulus* transcriptomes. While this is not ideal, we think results have not been strongly impacted by this issue.

2.1 Biological material and DNA extraction

One individual for each species of interest was collected and processed fresh.

Collection locations:

- *M. galloprovincialis* Mediterranean MgalMED; Sète, France

2

53   · *M. edulis* Southern European MeduEUS; ???

54   · *M. edulis* Northern Europe MeduEUN; ???

55   Whole mussels were placed in 50 mL falcon tubes containing 25 ml of TNES-Urea
56   solution and incubated for 4-6 weeks at room temperature (TNES-Urea: 10 mM Tris-
57   HCl pH 7.4; 120 mM NaCl; 10 mM EDTA pH 8.0; 0.5% SDS; 4 M urea).

58   After this period of pre-treatment at ambient temperature, proteinase K was added
59   at a final concentration of 150 μg/mL and the solution was incubated overnight at
60   56°C.

61   High Molecular weight genomic DNA was extracted following Nakayama et al.
62   (1994). We used a 15 mL Phase Lock Gel Heavy extraction with 3 steps of phenol/chloroform/isoamylalc
63   (25/24/1) Tris pH 8,1 followed by 2 chloroform extractions. After the last extraction, the
64   aqueous supernatant was precipitated with 2 volumes of 100% EtOH and the pellet
65   was hooked from the solution with a sterile glass Pasteur pipette. The pellet was
66   rinsed several times in 80% EtOH before being dried at room temperature.

67   DNA was resuspended with an appropriate volume of biomolecular water at 65°C
68   for several hours, the duration of this incubation depending on the size of the granules.
69   DNA was then stored at 4°C.

70   Prior to the construction of the DNA libraries, DNA was repaired with NEBNext
71   FFPE DNA Repair Mix according to the manufacturer's instructions.

2.2   Library preparation and sequencing

73   The 10X chromium library preparation and sequencing was subcontracted to the MGX
74   platform (Montpellier, France). The 10X linked reads libraries for each individual were
75   produced following the 10X Genomics Genome Reagen Kit (*Genome Solution*) proto-
76   col using a Chromium microfluidic chip. Libraries were subsequently sequenced on
77   an Illumina NovaSeq 6000 with an S4 flowcell.

2.3   Preprocessing 10X reads

79   We preprocessed 10X reads using the following pipeline for use in several algorithms
80   (section 2.4). We first removed duplicate reads using `Nubeam-dedup` (Dai and Guan,
81   2020; commit 25dd385). We then used `proc10xG` (https://github.com/ucdavis-bioinformatics/
82   proc10xG; commit 7afbfcf) to split the reads from their 10X barcodes for further pro-
83   cessing. We used a custom filtering step designed to remove under- and over-represented
84   barcodes and their associated reads from the data (`filter_barcodes.R` script, see
85   **Bary\&Gagnaire??**). [[Add supfigs of filter_barcodes]]. Reads were additionally fil-
86   tered with `fastp` (v0.20.1; Chen et al., 2018) with the main objective to remove poly-G
87   tails created by the Illumina Novaseq sequencing technology. We obtained at this
88   point reads equivalent to a short-read sequencing run, usable in some parts of the
89   assembly and quality control pipeline. Additionally, to obtain reads compatible to
90   10X genomics tools, we filtered and reassembled reads with their 10X barcodes in
91   `proc10xG` using the `filter_10xReads.py` and `regen_10xReads.py` scripts.

2.4 Initial genome assemblies

For each genome we used `Supernova` v2.1.1 (Weisenfeld et al., 2017) to assemble raw 10X reads. To avoid hard stops in `Supernova` due to both data quantity slightly under 10X genomics recommendations and an overestimation of genome size by Supernova, we used all available reads (`--maxreads='all'`) and accepted extreme coverage (`--accept-extreme-coverage`). We produced every style of Fasta output available in `Supernova` but only used the pseudo-haploid output in the following pipeline.

To remove duplicate haplotypes in the assemblies, we followed the `purge_dups` pipeline (Guan et al., 2020; commit e1934bb). We first used the `longranger` v2.2.2 align algorithm to map preprocessed reads (section 2.3) to the pseudo-haploid genomes to use in the `ngscstat` step. `Minimap2` (v2.17; Li, 2017) was used in the contig self-mapping step. We obtained the purged assemblies using the `get_seqs` steps without restricting the purging to the end of contigs (without the -e option).

We used `AGOUTI` (https://github.com/svm-zhang/AGOUTI; commit a7e65d6; Zhang et al., 2016) to improve scaffolding by using paired end RNA-seq reads. For each species, a different set of published transcriptomes were used (Supplementary Table [[...]] for accessions). `AGOUTI` require a gene prediction as input in addition to RNA-seq reads. We used `Augustus` (v3.3.3; Stanke et al., 2008) to produce an intermediate annotation for each assembly using the *Caenorhabditis* model. RNA-seq reads were first cleaned using `rcorrector` (v1.0.4; Song and Florea, 2015) and `trimgalore` (v0.6.6; https://www.bioinformatics.babraham.ac.uk/projects/trim_galore; `--quality 20 --stringency 1 -e 0.1 --length 70`). RNA-seq reads were mapped independently using `bwa-mem2` (v2.2.1; Vasimuddin Md et al., 2019), and then merged as a common bam file for each assembly with `samtools` (v1.12; Li et al., 2009). Finally, the `AGOUTI` scaffolding pipeline was run using the gene prediction and mapped RNA-seq reads (-minMQ 20 -maxFracMM 0.05; using `python` v2.7.15 and `samtools` v1.10 for compatibility).

As a last step, we ran *Blobtoolkit* (v2.4.0; Challis et al., 2020) on the three assemblies to evaluate quality and potential contamination levels. We used a custom script (`btk_conta_extraction.py`) to filter the assembly contigs based on the taxids found by the *Blobtoolkit* pipeline to remove the most obvious contaminations. Contigs matching taxids associated with viruses, bacteria and non-mollusca eukaryotes were removed. More specifically contigs for were removed for virus contamination if they presented a hit percentage of more than 10 % of their length. Contigs were removed for eukaryote contamination only when presenting only hits to taxa outside Mollusca on more than 10 % of their length. The list of retained contigs was used to filter the fasta assembly file using `seqkit` (v0.13.2; Shen et al., 2016).

2.5 Scaffolding on the *Mytilus coruscus* genome

At the time of assembly, the closest relative of the *Mytilus* species of interest with a chromosome scale assembly was *Mytilus coruscus* (GCA_017311375.1; Yang et al., 2021). Given the conserved number of 14 chromosomes in the *Mytilus* genus, we decided to scaffold our contigs on this high quality reference. Additionally, we had Oxford Nanopore reads for the MeduEUN individual. We used `minimap2` (v2.17; Li, 2017) and `LRScaf` (v1.1.10; https://github.com/shingocat/lrscaf) to first improve the MeduEUN assembly with this small amount of long reads.

136 Then for each assembly, we ran the scaffolder `RagTag` (v1.1.1; Alonge et al., 2021) to
137 position contigs on the *M. coruscus* chromosomal assembly.

138 A final polishing step was performed using `Pilon` (v1.24; Walker et al., 2014). `Pilon`
139 attempts to improve the assembly based on mapped read information (gap filling
140 and error corrections). We used `bwa-mem2` to map the debarcoded and filtered 10X
141 reads (section 2.3). In addition, Oxford Nanopore reads for MeduEUN were also used
142 in `Pilon` for the given assembly.

2.6 Repeats

144 We masked repeats for the purpose of annotation using `RepeatModeler` (v2.0.1; Flynn
145 et al., 2020) and `RepeatMasker` (v4.1.2-p1; Smit et al., 2013–2015) through the TETools
146 DFAM container (v1.3.1; https://hub.docker.com/r/dfam/tetools). We first built re-
147 peat databases for each of five assemblies MgalMED, MeduEUS, MeduEUN, *M. cor-*
148 *uscus* (GCA_017311375.1; Yang et al., 2021) and *M. galloprovincialis* from the Atlantic
149 (GCA_900618805.1; Gerdol et al., 2020). Then, we built a common *Mytilus* database
150 of repeats by merging those five databases using `cd-hit` (v4.8.1; Fu et al., 2012). We
151 used the same options as used by default in `RepeatModeler`: `-aS 0.8 -c 0.8 -g 1`
152 `-G 0 -A 80 -M 10000`. Finally, soft repeat masking was performed on the assemblies
153 with `RepeatMasker` using the merged database.

2.7 Annotation

155 To obtain an annotation for each assembly, we used both database and RNA-seq in-
156 formation. We used `Braker2` (v2.1.6; Brůna et al., 2021) as an initial step, using both a
157 protein database and RNA-seq reads (preprocessed in section 2.4). To build the pro-
158 tein database, we used all *Mollusca* proteins (taxid 6447) from OrthoDB (v10.1; Krivent-
159 seva et al., 2019). To provide gene presence hints, we mapped all RNA-seq reads for
160 each species using `HISAT2` (v2.2.1; Kim et al., 2019).

161 We used the `Mantis` pipeline (https://github.com/PedroMTQ/mantis; commit c6cb597;
162 Queirós et al., 2021) to obtain consensus annotations of genes based on multiple
163 databases. Protein sequences for each assembly were built from the `Braker2` an-
164 notation using the `python` module `gff3tool` (v2).

2.8 NCBI submission

166 The NCBI submission process identified a few errors that needed correcting. A small
167 number of adaptor sequences and duplicates were removed to comply with NCBI re-
168 quirements (see the rule `ncbi_submission_changes.smk` in the pipeline for more de-
169 tails). Assemblies are available under the following accessions: JAKGDF000000000
170 for MgalMED, JAKGDG000000000 for MeduEUS, and JAKGDH000000000 for MeduEUN.

2.9   Quality assessments and comparisons

Preprocessed 10X reads (section 2.3) were used to first estimate estimate genome size and heterozygosities of the three individuals. We used the reference free k-mer based method `genomescope` (https://github.com/tbenavi1/genomescope2.0; commit 5034ed4; Ranallo-Benavidez et al., 2020) and the fork of the `KMC` k-mer counting program (https://github.com/tbenavi1/KMC; commit 1df71f6).

Assembly statistics were computed with the python module `assembly_stats` (v0.1.4; Trizna, 2020).

To assess the remaining level of duplication in the assemblies, we used the program `KAT` (v2.4.2; Mapleson et al., 2017) to compare k-mer spectra from reads (preprocessed 10X) and from the assembly.

Finally, gene completion analyses were carried out using `BUSCO` (v5.1.1; Manni et al., 2021). We used both a Metazoan (`metazoa_odb10.2021-02-24`) and Molluscan (`mollusca_odb10.2020-08` database to assess and compare new and published assemblies. We compared our assemblies to the following published ones:

- *M. coruscus*, GCA_017311375.1, Yang et al. (2021);

- *M. galloprovincialis* from the Altantic lineage, GCA_900618805.1, Gerdol et al. (2020);

- *M. edulis* from the Southern European lineage, GCA_905397895.1, Corrochano-Fraile et al. (2021)

- *M. edulis* from the American lineage, GCA_019925275.1.

2.10   Phylogenetic species tree

We compiled protein sequences from published *Mytilus* genomes and transcriptomes, in addition to the current three genomes and annotations, to build a species tree. We used transcriptomes produced in Popovic and Riginos (2020) for American *M. edulis*, Mediterranean *M. galloprovincialis*, *M. trossulus*, *M. californianus*. Transcriptomes were translated using the `seqkit` program (v2.2.0; Shen et al., 2016) We used genomes and associated annotations of *M. coruscus* (GCA_017311375.1; Yang et al. (2021)), Southern European *M. edulis* (GCA_905397895.1; Corrochano-Fraile et al. (2021)), Atlantic *M. galloprovincialis* (GCA_900618805.1; Gerdol et al. (2020)), and American *M. edulis* (GCA_019925275.1; annotation as personal communication of Tiago Hori, PEIMSO). For GCA_019925275.1, protein sequences where retrieve from the fasta and gff files using the python module gff3tool (v2.1.0).

We used `Orthofinder` (v2.5.2; Emms and Kelly, 2015, 2019) to find orthogroups and orthologue genes. Species tree was inferred using the MSA method of `Orthofinder` (Emms & Kelly, 2018) and `IQTREE` (v2.2.0_beta; Minh et al., 2020) tree inference.

3   Results and discussion

The pre-assembly k-mer analysis carried out using Genomescope (using 21-mers) showed that, as expected, the genomes were highly heterozygous with values going from 3.49 to 8.77 % (fig. 1). Genomescope provide a bad fit and estimation in the case of the MeduEUS data due to a reduced coverage of the genome in that case, compared to the two other assemblies.

Highly heterozygous genomes brings the risk of having a large number of dupli-cate contigs due to the separate assembly of the two alleles from a same locus. For this reason, we used the `purge_dups` pipeline to try removing a maximum of such bias. This procedure reduced the number of complete duplicated genes in all assemblies (fig. S1, v1 to v4).
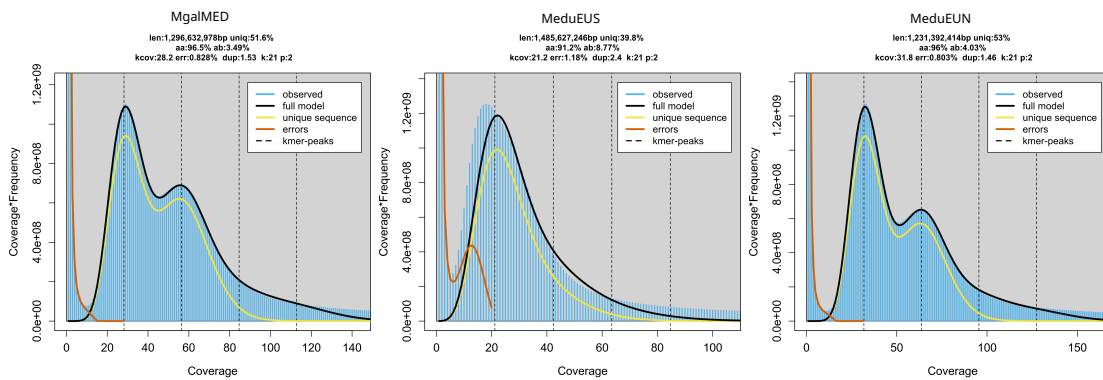
Figure 1: Genomescope

Table 1: Assembly statistics comparisons. C for contigs and S for scaffolds.

assembly	C.L50	C.N50	C.median	C.sequence_count	S.L50	S.N50	S.gc_content	S.median	S.sequence_count	S.total_bps
Mcor_GCA017311375	261	1481111	42077	6449	6	99542347	32.4	21293	4434	1566529938
MgalATL_GCA900618805	4922	77035	39489	22922	1903	207642	32.1	77568	10577	1282208009
MeduEUS_GCA905397895	977	511485	184638	5966	464	1097279	32.2	292104	3339	1827085763
MeduAM_GCA019925415	835	490737	62522	9686	6	116503180	32.3	28931	1119	1651313236
MgalMED_v7	16906	26323	7475	115913	9	71952646	32.1	4531	41122	1658656017
MeduEUS_v7	24467	23624	6478	169324	415	228263	34.6	4999	73616	2076685641
MeduEUN_v7	15862	29157	8961	105892	9	77102752	32.1	6164	28220	1764246486

To assess the completeness of our assemblies, we compared them to four *Mytilus* published assemblies on the basis of a Metazoan and a Molluscan set of single copy orthologous genes using BUSCO (fig. 2). Overall, we show that our assemblies are equivalent to the published ones in terms of completeness.

BUSCO scores Metazoa_odb10 (954 genes)

Assembly	Complete single copy	Complete duplicated	Fragmented	Missing
Mcor_GCA017311375	832	17	34	71
MgalATL_GCA900618805	681	74	88	111
MeduEUS_GCA905397895	662	224	39	29
MeduAM_GCA019925415	754	112	43	45
MgalMED_v7	727	112	70	45
MeduEUS_v7	726	166	34	28
MeduEUN_v7	701	158	62	33

BUSCO scores Mollusca_odb10 (5295 genes)

Assembly	Complete single copy	Complete duplicated	Fragmented	Missing
Mcor_GCA017311375	4232	56	180	827
MgalATL_GCA900618805	3447	288	253	1307
MeduEUS_GCA905397895	3644	730	230	691
MeduAM_GCA019925415	3782	409	212	892
MgalMED_v7	3604	316	346	1029
MeduEUS_v7	3824	307	280	884
MeduEUN_v7	3511	528	291	965

categories: Complete single copy, Complete duplicated, Fragmented, Missing

Figure 2: Busco scores

Acknowledgements

References

Ahmad, M., Skibinski, D. O. F., & Beardmore, J. A. (1977). An estimate of the amount of genetic variation in the common mussel Mytilus edulis. *Biochemical Genetics*, *15*(9-10), 833–846. https://doi.org/10.1007/BF00483980

Alonge, M., Lebeigle, L., Kirsche, M., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2021). Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. *bioRxiv*, 2021.11.18.469135. https://doi.org/10.1101/2021.11.18.469135

Bierne, N., Borsa, P., Daguin, C., Jollivet, D., Viard, F., Bonhomme, F., & David, P. (2003). Introgression patterns in the mosaic hybrid zone between *Mytilus edulis* and *M. galloprovincialis*. *Molecular Ecology*, *12*(2), 447–461. https://doi.org/10.1046/j.1365-294X.2003.01730.x

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUS-TUS supported by a protein database. *NAR Genomics and Bioinformatics*, *3*(1), lqaa108. https://doi.org/10.1093/nargab/lqaa108

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 Genes|Genomes|Genetics*, *10*(4), 1361–1374. https://doi.org/10.1534/g3.119.400908

242   Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ prepro-
243   cessor. *Bioinformatics*, *34*, i884–i890. https://doi.org/10.1093/bioinformatics/
244   bty560

245   Corrochano-Fraile, A., Davie, A., & Carboni, S. (2021). Evidence of multiple genome
246   duplication events in Mytilus evolution. *bioRxiv*. https://doi.org/10.1101/2021.
247   08.17.456601

248   Dai, H., & Guan, Y. (2020). Nubeam-dedup: A fast and RAM-efficient tool to de-duplicate
249   sequencing reads without mapping (I. Birol, Ed.). *Bioinformatics*, *36*(10), 3254–
250   3256. https://doi.org/10.1093/bioinformatics/btaa112

251   El Ayari, T., Trigui El Menif, N., Hamer, B., Cahill, A. E., & Bierne, N. (2019). The hidden
252   side of a major marine biogeographic boundary: A wide mosaic hybrid zone at
253   the Atlantic–Mediterranean divide reveals the complex interaction between
254   natural and genetic barriers in mussels. *Heredity*, *122*, 770–784. https://doi.
255   org/10.1038/s41437-018-0174-y

256   Emms, D. M., & Kelly, S. (2018). STAG: Species Tree Inference from All Genes. *bioRxiv*,
257   267914. https://doi.org/10.1101/267914

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

261   Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for com-
262   parative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/
263   s13059-019-1832-y

264   Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F.
265   (2020). RepeatModeler2 for automated genomic discovery of transposable el-
266   ement families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–
267   9457. https://doi.org/10.1073/pnas.1921046117

268   Fraïsse, C., Belkhir, K., Welch, J. J., & Bierne, N. (2016). Local interspecies introgression
269   is the main cause of extreme levels of intraspecific differentiation in mussels.
270   *Molecular Ecology*, *25*(1), 269–286. https://doi.org/10.1111/mec.13299

271   Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the
272   next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://
273   doi.org/10.1093/bioinformatics/bts565

274   Gerdol, M., Moreira, R., Cruz, F., Gómez-Garrido, J., Vlasova, A., Rosani, U., Venier, P.,
275   Naranjo-Ortiz, M. A., Murgarella, M., Greco, S., Balseiro, P., Corvelo, A., Frias,
276   L., Gut, M., Gabaldón, T., Pallavicini, A., Canchaya, C., Novoa, B., Alioto, T. S., …
277   Figueras, A. (2020). Massive gene presence-absence variation shapes an open
278   pan-genome in the Mediterranean mussel. *Genome Biology*, *21*(1), 275. https:
279   //doi.org/10.1186/s13059-020-02180-3

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

283   Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome
284   alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotech-
285   nology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

286   Koehn, R. K., & Mitton, J. B. (1972). Population Genetics of Marine Pelecypods. I. Ecolog-
287   ical Heterogeneity and Evolutionary Strategy at an Enzyme Locus. *The Ameri-
288   can Naturalist*, *106*(947), 47–56. https://doi.org/10.1086/282750

289   Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdob-
290   nov, E. M. (2019). OrthoDB v10: Sampling the diversity of animal, plant, fungal,

protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, *47*(D1), D807–D811. https://doi.org/10.1093/nar/gky1053

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H. (2017). *Minimap2: Pairwise alignment for nucleotide sequences*.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., & Clavijo, B. J. (2017). KAT: A K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, *33*(4), 574–576. https://doi.org/10.1093/bioinformatics/btw663

Milkman, R., & Beaty, L. (1970). Large-Scale Electrophoretic Studies of Allelic Variation in *Mytilus edulis*. *Biological Bulletin*, *139*, 430.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mölder, F., Jablonski, K., Letcher, B., Hall, M., Tomkins-Tinch, C., Sochat, V., Forster, J., Lee, S., Twardziok, S., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake [version 2; peer review: 2 approved]. *F1000Research*, *10*(33). https://doi.org/10.12688/f1000research.29032.2

Nakayama, I., Foresti, F., Tewari, R., Schartl, M., & Chourrout, D. (1994). Sex chromosome polymorphism and heterogametic males revealed by two cloned DNA probes in the ZW/ZZ fish Leporinus elongatus. *Chromosoma*, *103*(1), 31–39. https://doi.org/10.1007/BF00364723

Popovic, I., & Riginos, C. (2020). Comparative genomics reveals divergent thermal selection in warm- and cold-tolerant marine mussels. *Molecular Ecology*, *29*(3), 519–535. https://doi.org/10.1111/mec.15339

Queirós, P., Delogu, F., Hickl, O., May, P., & Wilmes, P. (2021). Mantis: Flexible and consensus-driven genome annotation. *GigaScience*, *10*(6), giab042. https://doi.org/10.1093/gigascience/giab042

Quesada, H., Wenne, R., & Skibinski, D. O. F. (1995). Differential introgression of mitochondrial DNA across species boundaries within the marine mussel genus *Mytilus*. *Proceedings of the Royal Society B: Biological Sciences*, *262*(1363), 51–56. https://doi.org/10.1098/rspb.1995.0175

Quesada, H., Zapata, C., & Alvarez, G. (1995). A multilocus allozyme discontinuity in the mussel *Mytilus galloprovincialis*: The interaction of ecological and life-history factors. *Marine Ecology Progress Series*, *116*, 99–115.

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

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation (Q. Zou, Ed.). *PLOS ONE*, *11*(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Simon, A., Arbiol, C., Nielsen, E. E., Couteau, J., Sussarellu, R., Burgeot, T., Bernard, I., Coolen, J. W. P., Lamy, J.-B., Robert, S., Skazina, M., Strelkov, P., Queiroga, H., Cancio, I., Welch, J. J., Viard, F., & Bierne, N. (2020). Replicated anthropogenic hybridisations reveal parallel patterns of admixture in marine mussels. *Evolutionary Applications*, *13*(3), 575–599. https://doi.org/10.1111/eva.12879

Simon, A., Fraïsse, C., El Ayari, T., Liautard-Haag, C., Strelkov, P., Welch, J. J., & Bierne, N. (2021). How do species barriers decay? Concordance and local introgression in mosaic hybrid zones of mussels. *Journal of Evolutionary Biology*, *34*, 208–223. https://doi.org/10.1111/jeb.13709

Skibinski, D. O. F., Ahmad, M., & Beardmore, J. A. (1978). Genetic evidence of naturally occurring hybrids between *Mytilus edulis* and *M. galloprovincialis*. *32*(2), 354–364. https://doi.org/10.1038/nature01240

Smit, A., Hubley, R., & Green, P. (2013–2015). *RepeatMasker Open-4.0* (Version 4.1.2-p1).

Song, L., & Florea, L. (2015). Rcorrector: Efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience*, *4*(1), 48. https://doi.org/10.1186/s13742-015-0089-y

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013

Trizna, M. (2020). *Assembly_stats 0.1.4* (Version 0.1.4). https://doi.org/10.5281/zenodo.3968775

Vasimuddin Md, S. M., Misra, S., Li, H., & Aluru, S. (2019). *Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems*.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, *9*(11), 14.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, *27*(5), 757–767. https://doi.org/10.1101/gr.214874.116

Yang, J.-L., Feng, D.-D., Liu, J., Xu, J.-K., Chen, K., Li, Y.-F., Zhu, Y.-T., Liang, X., & Lu, Y. (2021). Chromosome-level genome assembly of the hard-shelled mussel Mytilus coruscus, a widely distributed species from the temperate areas of East Asia. *GigaScience*, *10*. https://doi.org/10.1093/gigascience/giab024

Zhang, S. V., Zhuo, L., & Hahn, M. W. (2016). AGOUTI: Improving genome assembly and annotation using transcriptome data. *GigaScience*, *5*(1), 31. https://doi.org/10.1186/s13742-016-0136-3

Supplementary Information
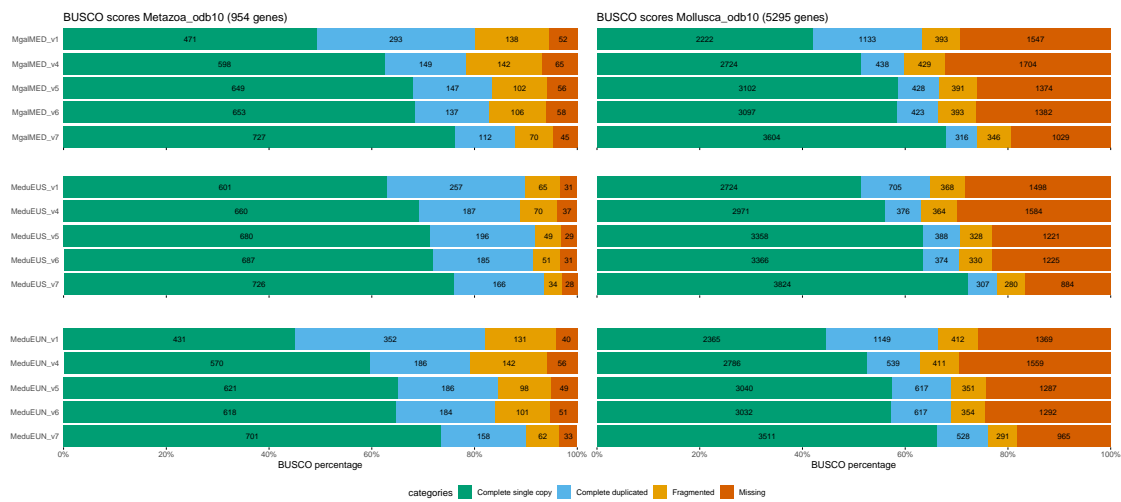
Figure S1: Busco scores

Blobtoolkit

13
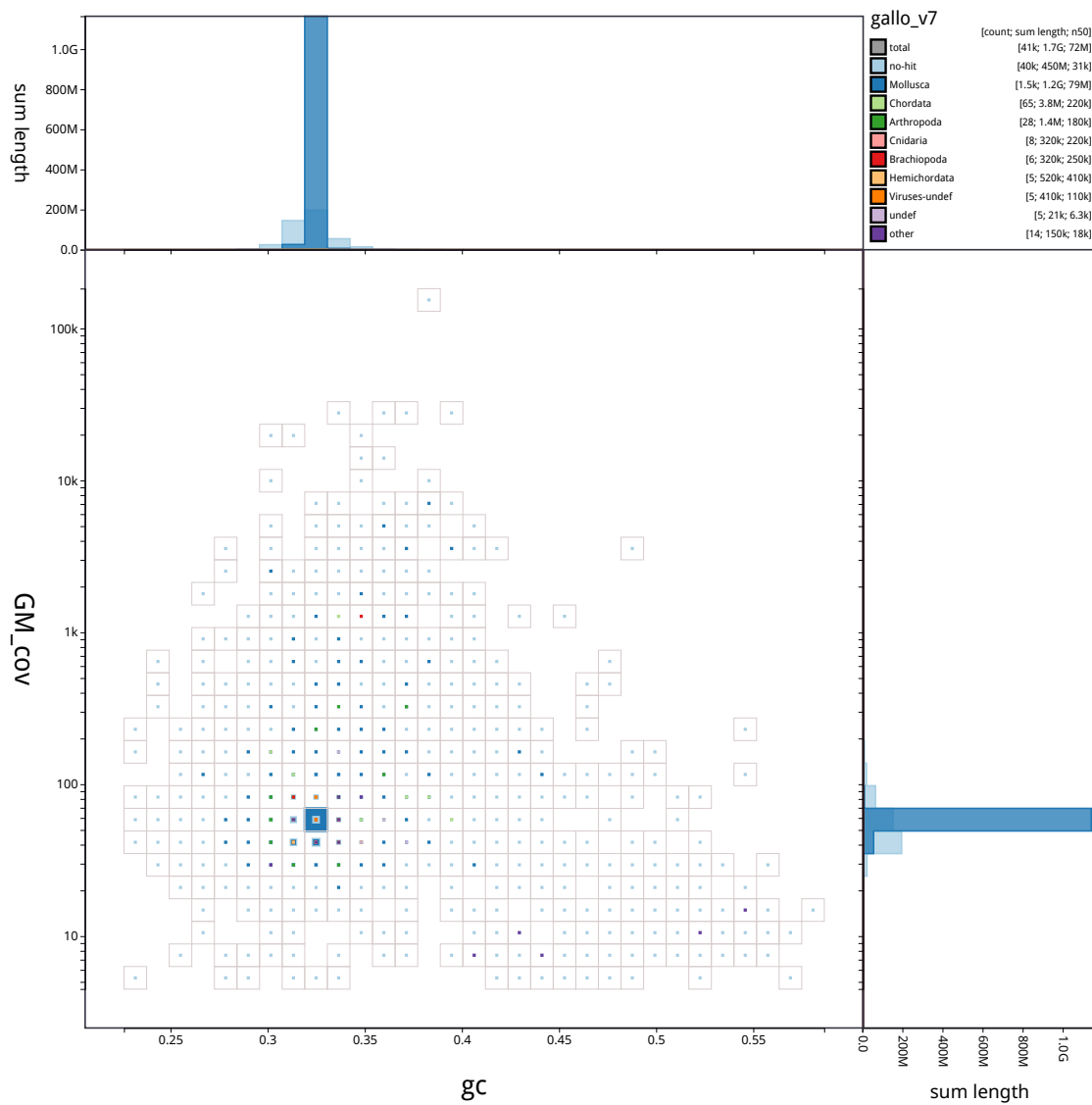
Figure S2: None-binned blob plot of base coverage in GM against GC proportion for scaffolds in assembly MgalMED_v7. Scaffolds are coloured by phylum and binned at a resolution of 30 divisions on each axis. Coloured nones within each bin are sized in proportion to the sum of individual scaffold lengths on a square-root scale, ranging from 1,005 to 1,127,105,406. Histograms show the distribution of scaffold length sum along each axis.
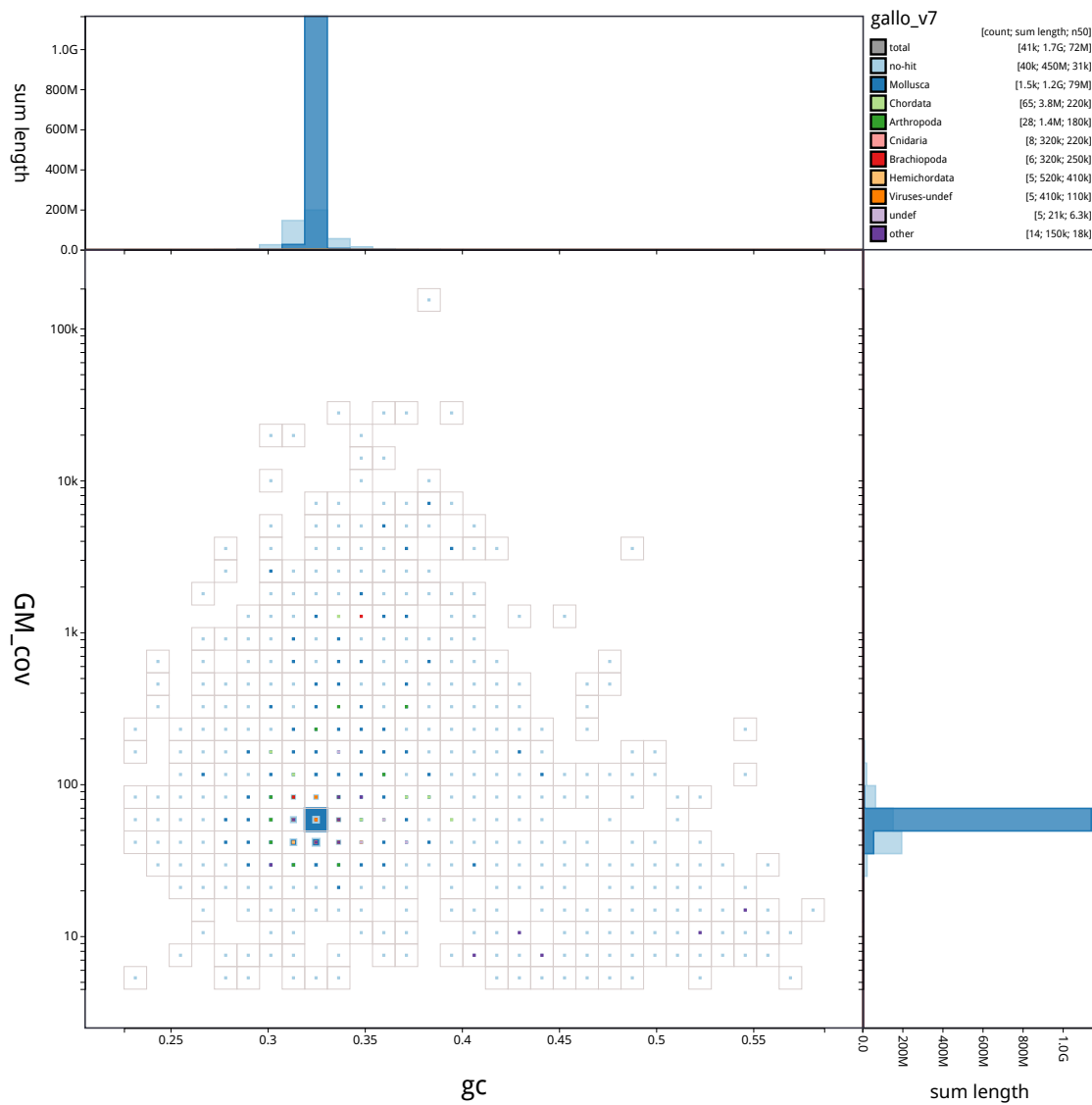
Figure S3: None-binned blob plot of base coverage in GM against GC proportion for scaffolds in assembly MeduEUS_v7. Scaffolds are coloured by phylum and binned at a resolution of 30 divisions on each axis. Coloured nones within each bin are sized in proportion to the sum of individual scaffold lengths on a square-root scale, ranging from 987 to 487,517,891. Histograms show the distribution of scaffold length sum along each axis.
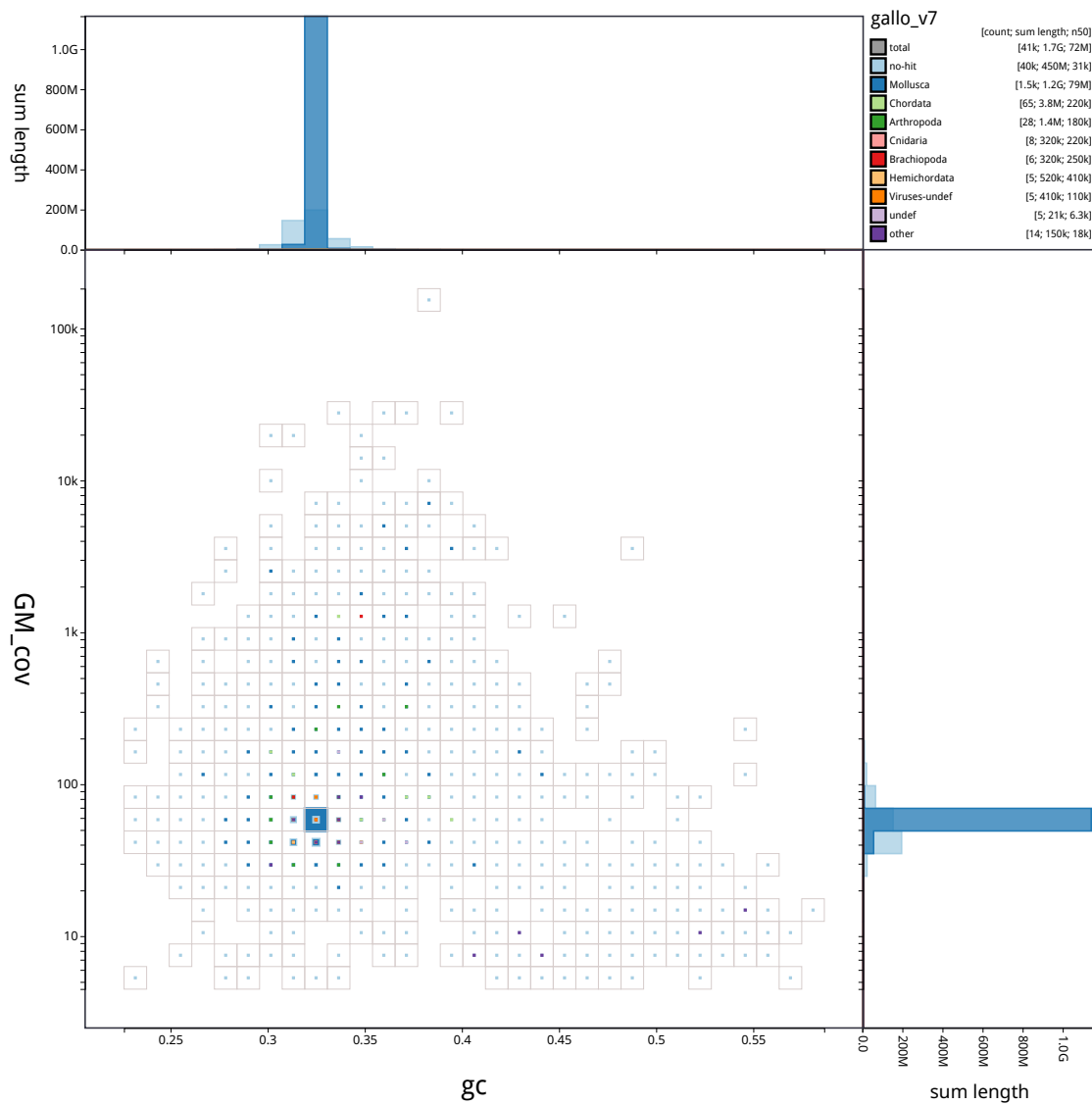
Figure S4: None-binned blob plot of base coverage in GM against GC proportion for scaffolds in assembly MeduEUN_v7. Scaffolds are coloured by phylum and binned at a resolution of 30 divisions on each axis. Coloured nones within each bin are sized in proportion to the sum of individual scaffold lengths on a square-root scale, ranging from 1,011 to 1,136,301,196. Histograms show the distribution of scaffold length sum along each axis.

gallo_v7

	[count; sum length; n50]
total	[41k; 1.7G; 72M]
no-hit	[40k; 450M; 31k]
Mollusca	[1.5k; 1.2G; 79M]
Chordata	[65; 3.8M; 220k]
Arthropoda	[28; 1.4M; 180k]
Cnidaria	[8; 320k; 220k]
Brachiopoda	[6; 320k; 250k]
Hemichordata	[5; 520k; 410k]
Viruses-undef	[5; 410k; 110k]
undef	[5; 21k; 6.3k]
other	[14; 150k; 18k]
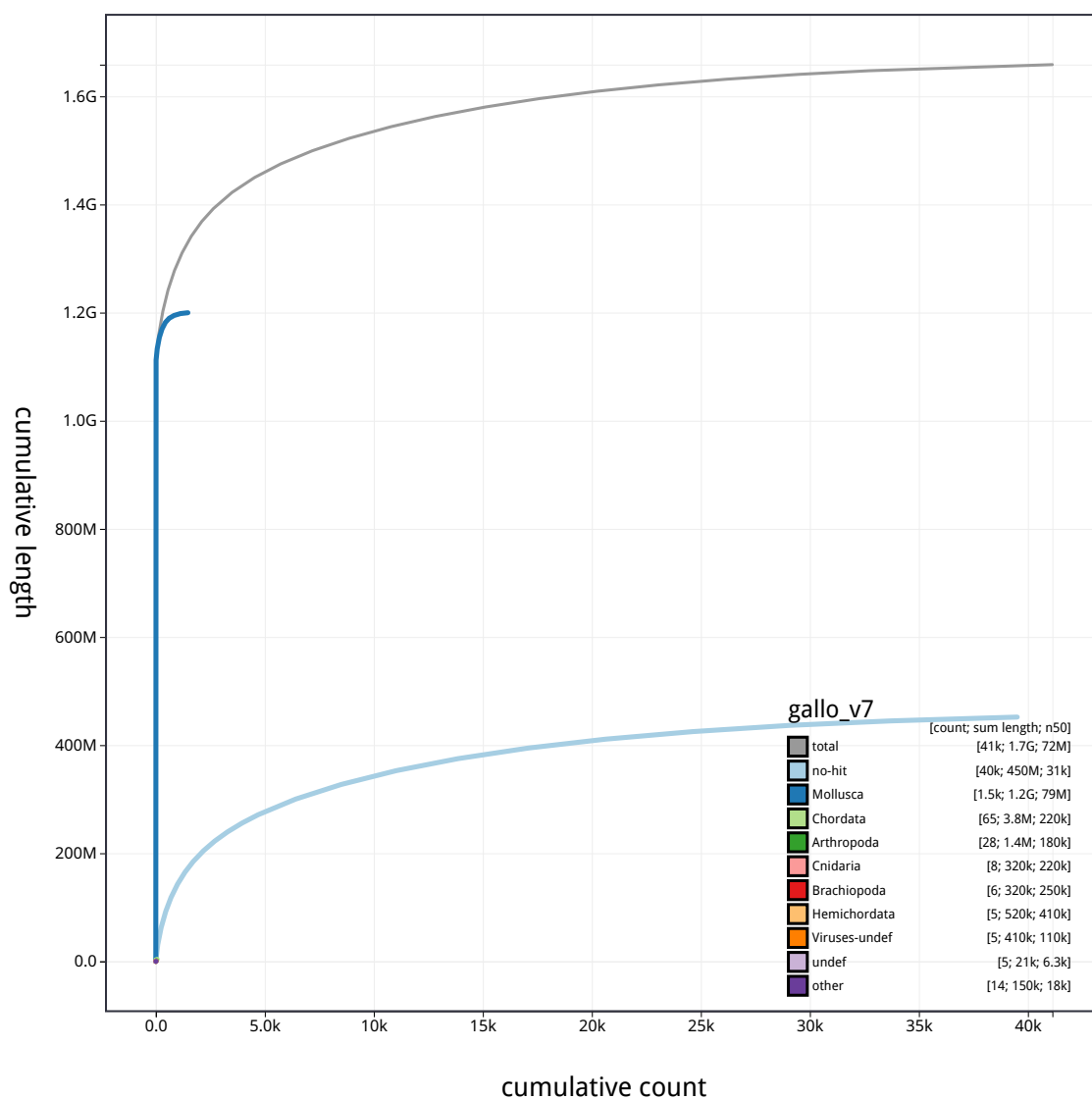
Figure S5: Cumulative scaffold length for assembly MgalMED_v7. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the bestsumorder taxrule.
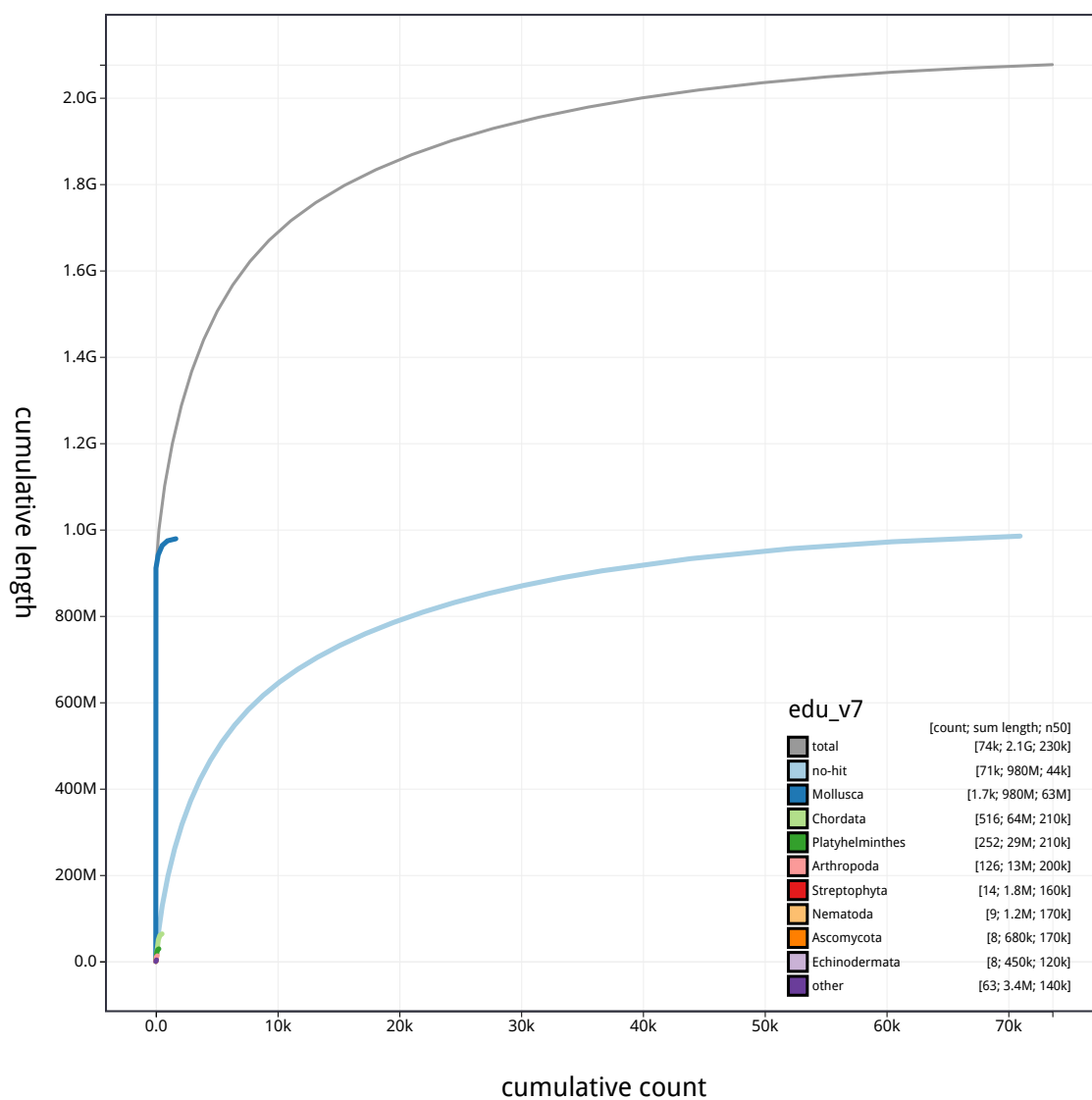
Figure S6: Cumulative scaffold length for assembly MeduEUS_v7. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the bestsumorder taxrule.
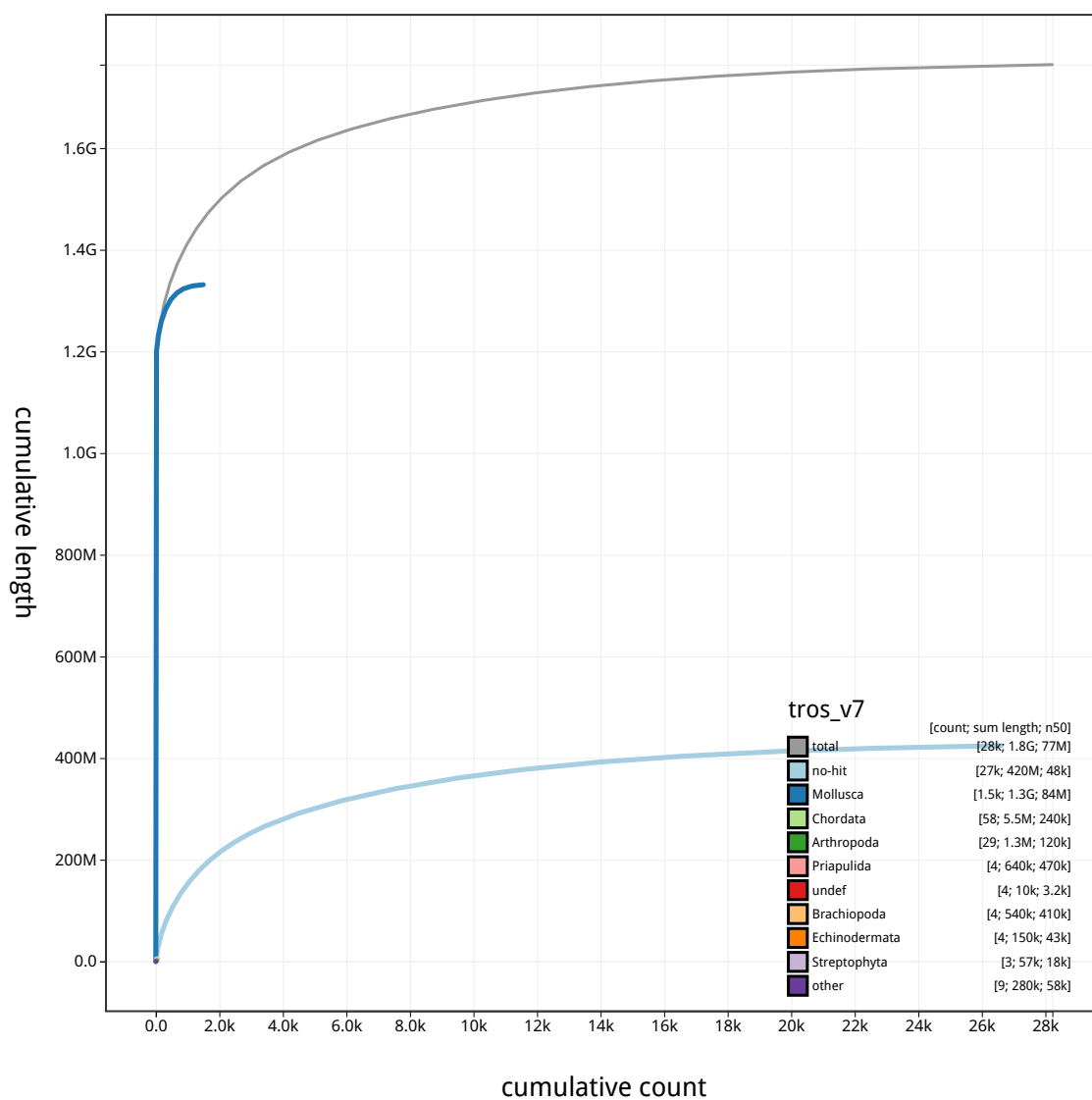
Figure S7: Cumulative scaffold length for assembly MeduEUN_v7. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the bestsumorder taxrule.
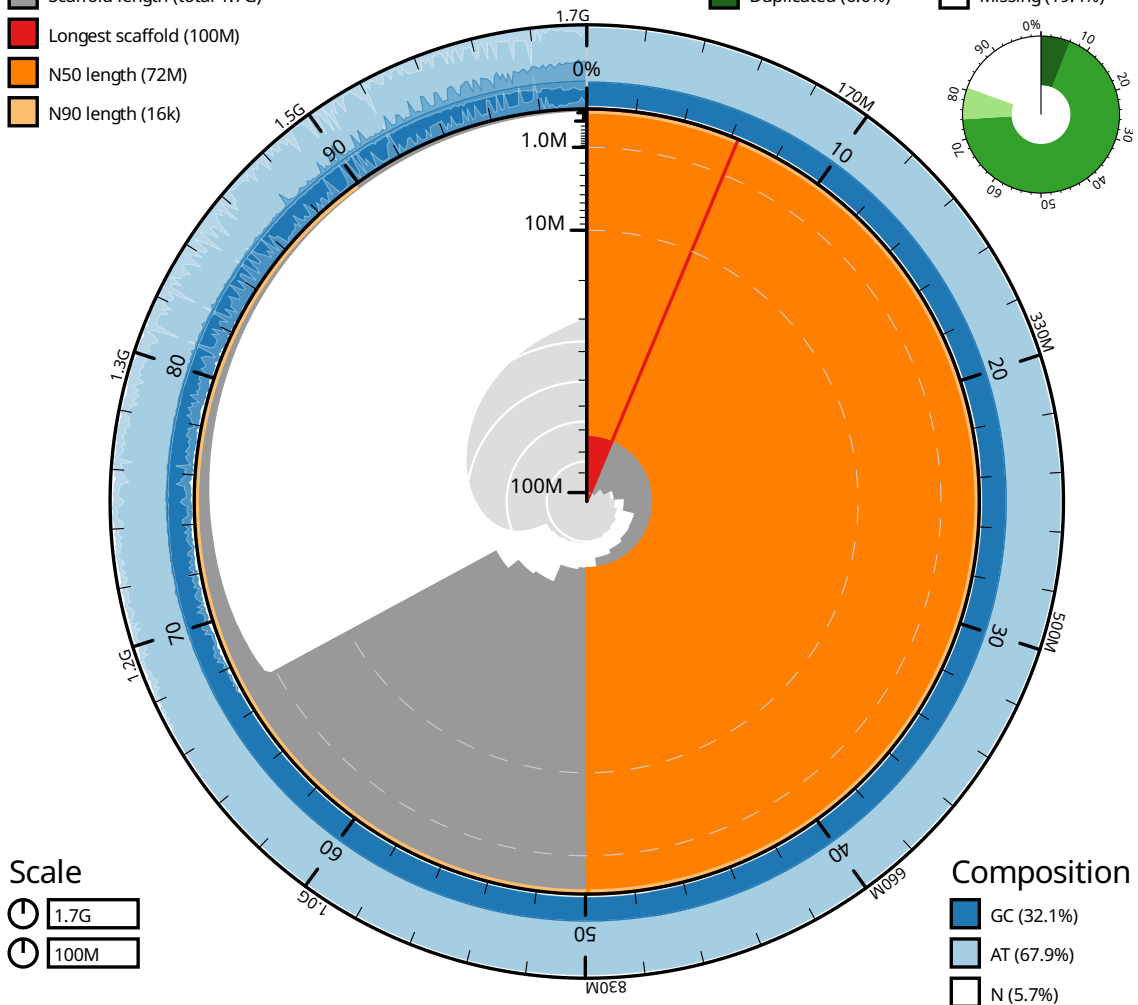
Figure S8: Snail plot summary of assembly statistics for assembly MgalMED_v7. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 1,658,656,017 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (104,729,878 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (71,952,646 and 16,011 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the mollusca_odb10 set is shown in the top right.
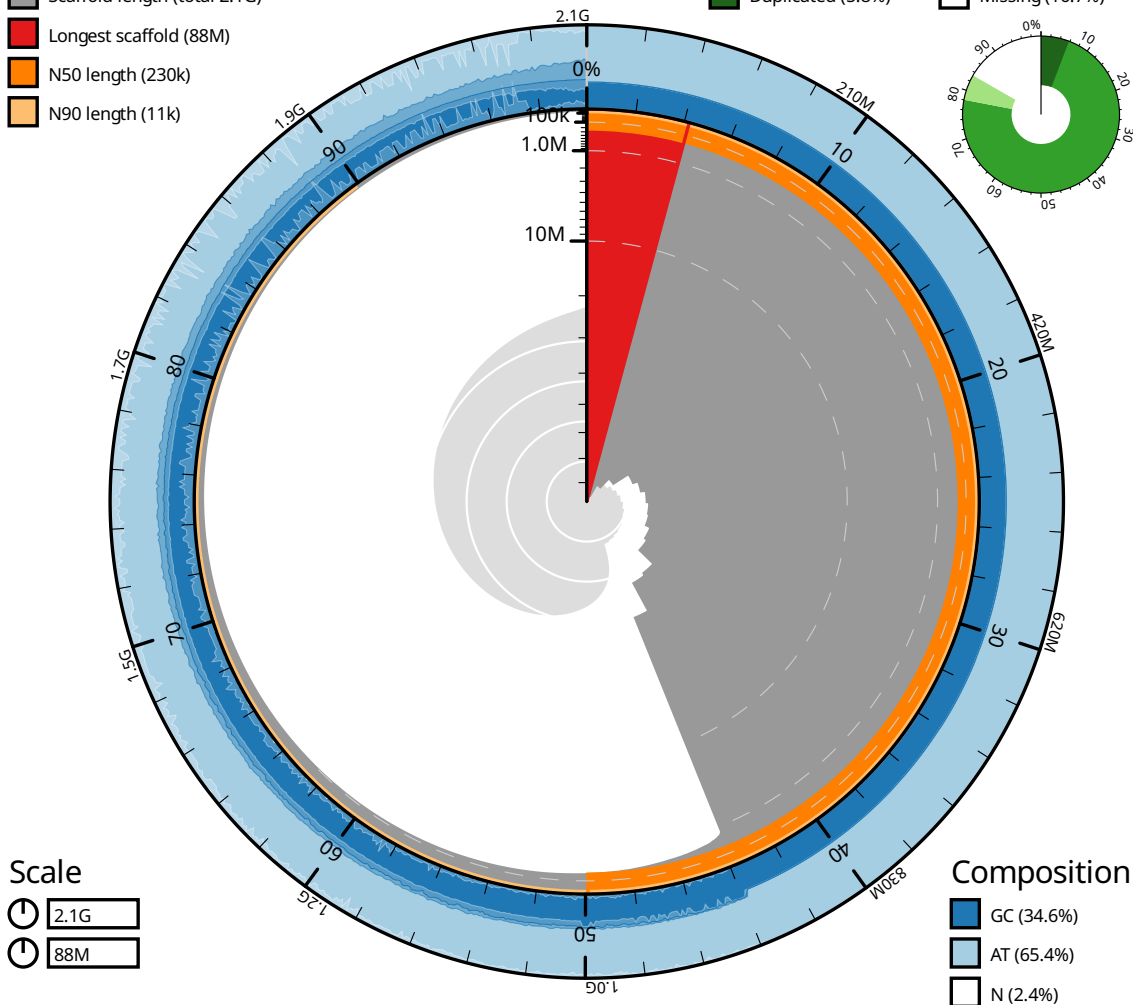
Figure S9: Snail plot summary of assembly statistics for assembly MeduEUS_v7. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 2,076,685,641 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (88,305,666 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (228,263 and 10,773 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the mollusca_odb10 set is shown in the top right.
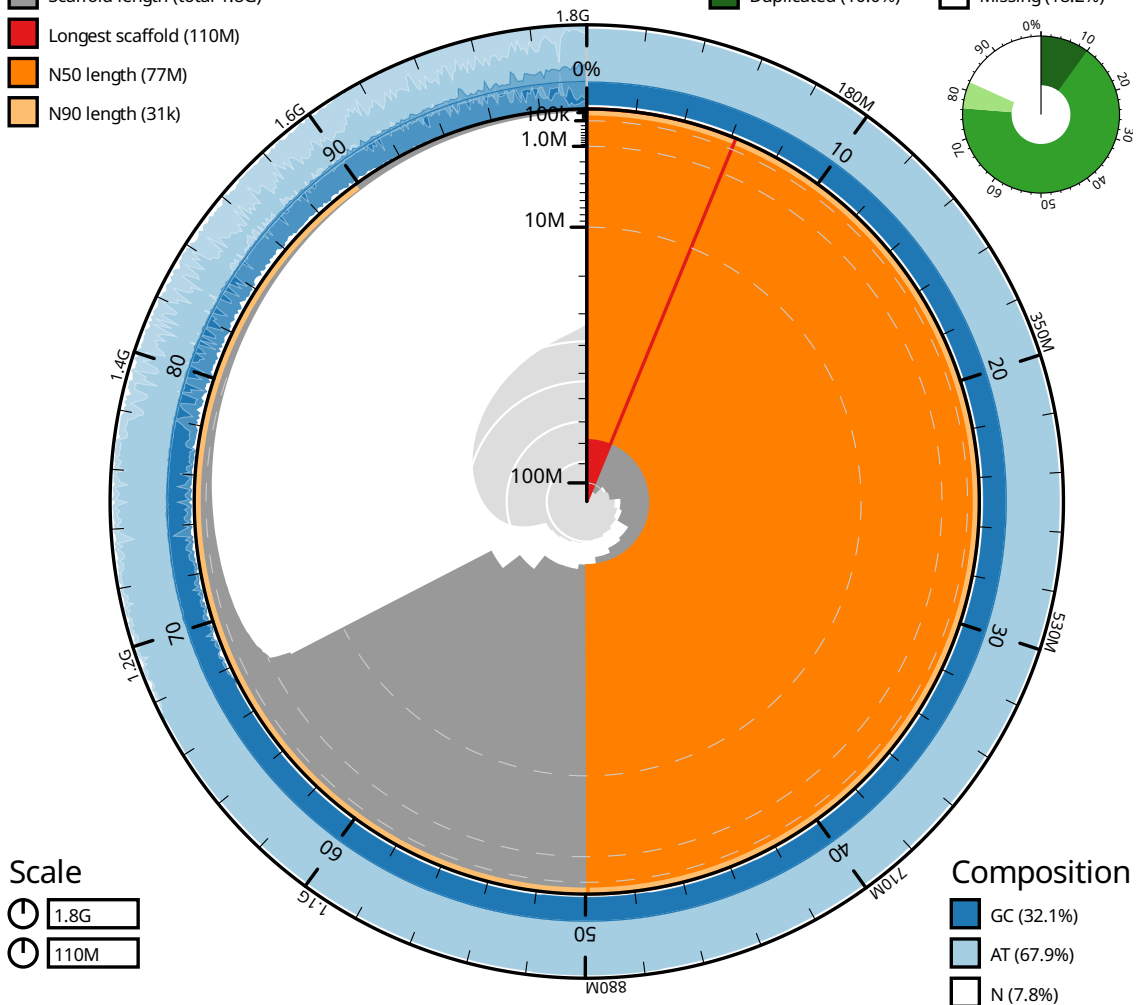
Figure S10: Snail plot summary of assembly statistics for assembly MeduEUN_v7. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 1,764,246,486 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (110,194,685 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (77,102,752 and 30,649 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the mollusca_odb10 set is shown in the top right.