scientific data

Check for updates

OPEN

DATA DESCRIPTOR

The First Genome Assembly Of The Dogwhelk *Nucella lapillus*, a Bioindicator Species For The Marine Environment

Juned Kadiwala[1,3], Andrew Hesketh[1,3], Heleen De Weerd[2], Helen Ritch[2] & Rameen Shakur[1]✉

The dogwhelk (*Nucella lapillus*) is a predatory marine gastropod widely distributed across temperate intertidal zones. Renowned for its ecological role in controlling prey populations, *N. lapillus* is also an important bioindicator species for marine pollution through imposex. The molecular genetic basis of imposex, characterised by the abnormal development of male sex organs in females and reductions in fertility and lifespan, however remains poorly understood due to the absence of a reference genome sequence. Here we provide the first genome assembly comprising 2.41 Gb of sequence, predicted to encode 47,238 proteins. This inaugural assembly lays the foundations for implementing genomic approaches to better quantify and characterise imposex, in addition to elucidating adaptations to life within changeable intertidal ecosystems. To counter challenges of DNA fragmentation and contamination often associated with the sequencing of marine organisms, we found that a hybrid approach that integrates complementary long-read data from PacBio HiFi and Oxford Nanopore Technology (ONT) platforms helped maximise the final assembly. This innovative combination may be a useful approach for similar marine species.

Background & Summary

Intertidal organisms face extreme environmental variability, including fluctuating salinity, temperature, and oxygen levels. The ability of *Nucella lapillus* to thrive in such conditions makes it an ideal candidate for exploring the genetic mechanisms of environmental resilience in the context of climate change. Additionally, its predatory adaptations, including specialized enzymes and toxins for prey capture[1], provide a unique opportunity to investigate the molecular and genetic basis of its ecological success. The representation of predatory marine invertebrates from intertidal environments in genome sequence databases is however sparse. To date, only a handful of genome assemblies have been made available for the Muricidae family to which *N. lapillus* also belongs, with those for *Rapana venosa* and *Stramonita haemastoma* being amongst the most complete[2,3].

N. lapillus is an important and well-established bioindicator species for assessing endocrine disruption by pollutants within marine ecosystems[4–7]. Chronic exposure to TBT and other endocrine-disrupting chemicals can induce imposex in female dogwhelks, a phenomenon where male sexual characteristics are developed leading to sterility and shortened lifespan[6,8–10]. Increasing water temperature may also have a potentiating effect on the activity of endocrine disruptors[11]. The development and progression of imposex in dogwhelk is typically assessed by a visual phenotypic characterisation of the gonad tissues, and hence there is a need for more rigorous genomics-based approaches to standardise and improve its identification globally[6,8,12,13]. The provision of a reference genome for *N. lapillus* would facilitate a deeper understanding of the genetic basis of female masculinization due to pollution, and provide new opportunities to study the impact of climate stress on intertidal ecosystems. *N. lapillus* plays a critical role in the marine food web as both a predator and prey, impacting the health of other marine species, including those consumed by humans. Understanding how water temperature

[1]Brighton Integrative Genomics (BIG) Unit and the Centre for Precision Health and Translational Medicine, School of Applied Sciences, University of Brighton, Brighton, BN2 6DN, UK. [2]Edinburgh Genomics, The University of Edinburgh, Ashworth Laboratories, The King's Buildings, Charlotte Auerbach Road, Edinburgh, EH9 3FL, UK. [3]These authors contributed equally: Juned Kadiwala, Andrew Hesketh. ✉e-mail: r.shakur@brighton.ac.uk

Input reads type	Assembly	Total length (Gb)	Contigs (n)	Contigs N50 (Kb)	BUSCO (5925 orthologs, mollusca_odb10)		
					% complete	% single	% duplicated
ONT	flye	2.246	82,846	93	77.2	67.7	9.5
PacBio	hifiasm	2.122	17,114	225	79.8	63.6	16.2
ONT + PacBio	Hybrid hifiasm	2.116	13,248	257	79.0	60.1	18.9
ONT + PacBio	verkko	2.810	56,517	75	74.5	33.5	41.0
ONT + PacBio	Quickmerge of two assemblies	2.853	41,532	323	86.1	63.0	23.1
PacBio + ONT	Quickmerge of two assemblies	2.414	11,397	338	84.0	64.2	19.8

Table 1. Comparison of different genome assembly schemes.

coupled with pollutants affect the physiology and reproductive success of the dogwhelk and related species is important for assessing risks to seafood safety and implications to human health, as contaminated shellfish can pose a direct health hazard[14]. Furthermore, comparative genomics analysis between this species and humans can uncover molecular pathways related to environmental stress responses, endocrine disruption, and pollutant detoxification, with relevance in human diseases. For example, insights into how the dogwhelk copes with toxic organotin exposure will advance our knowledge of similar endocrine-disrupting chemicals affecting human health.

In this study, we present a high-quality draft genome of *N. lapillus*, leveraging a hybrid sequencing approach that integrates complementary long-read data from PacBio HiFi and Oxford Nanopore Technology (ONT) platforms. This innovative combination capitalized on the high accuracy of PacBio HiFi reads and the extended read lengths of ONT, resulting in an assembly of superior completeness and contiguity. Our findings underscore the power of hybrid sequencing for overcoming challenges inherent to assembling complex, repetitive marine genomes. An advantage of using a combination of both long-read platforms was observed for this species during the assembly stage, revealing that their combined use resulted in the best assembly (Table 1). The final assembly consists of 11,397 contigs (N50 = 338 Kb), with a total size of 2.41 Gb and 84.0% assembly completeness as determined by BUSCO analysis[15]. It has 42.6% GC base composition and a large proportion of repeat sequences (~66%) (Fig. 1). Repeat sequences play an important role in regulating gene expression, inducing variation, and driving evolution. Unclassified repeat sequences and simple repeats are the most common type of repeats present in the *N. lapillus* genome in terms of number, but long interspersed nuclear elements (LINE) comprise the largest proportion of the genome (462 Mb, 19.2% of the assembled genome) (Fig. 1). Simple sequence repeats comprise an unusually high proportion of the genome (284 Mb and 11.8%) and are reported to promote genome plasticity in shrimp species[16].

The genome assembly was derived from 40.6 Gb of PacBio HiFi reads (read N50, 11291; N90, 9246), and 61.1 Gb of ONT data (read N50, 3643; N90, 1546). Annotation of the genome predicted 47,238 protein encoding genes, with an average of 5.6 exons per gene. BUSCO analysis of the predicted proteome indicates 76% completeness, which is comparable to those for the other Muricidae species available, *R. venosa* and *S. haemastoma* (Table 2), and also more broadly to those from other marine gastropods (Table 3). The statistics for the size, contig number and length of the genome assemblies are also comparable (Tables 2, 4). A comparative genomics analysis based on the identification of orthologous protein groups in predicted proteomes of different species[17], places *N. lapillus* within the Caenogastropoda subclass of Gastropoda, and more specifically alongside relatives from the Muricidae family, supporting the quality of the annotation of the protein sequences in the *N. lapillus* genome (Fig. 2a). About 90% of the annotated proteins in *N. lapillus* belong to orthologous groups, and the species-specificity or commonality of these groups is comparable to other gastropods (Fig. 2b,c). This indicates a general high level of evolutionary conservation among the proteins across the species being analysed. Alongside *C. gigas*, *N. lapillus* shows the highest number of species-specific orthologs, and these could represent the more frequent acquisition of specialised functions or adaptations compared to the other species analysed.

Excessive fragmentation of genomic DNA from marine species is often observed due to vigorous or lengthy isolation protocols aimed at overcoming problems associated with the tough consistency of the tissues available for extraction, and the need to remove co-purifying metabolites known to inhibit nanopore sequencing or enzymes involved in the preparation of sequencing libraries from samples of marine origin[18,19]. This can hinder the assembly of reads into contigs of sufficient length during genome construction, especially in marine gastropods that can also possess highly repetitive genomic DNA. In this study, the use of long sequencing reads generated by the different but complementary sequencing platforms, helped to maximise the assembly of the reference genome for *N. lapillus*, and will be a useful approach for other similar marine species.

Methods

Sample collection. Specimens of wild *Nucella lapillus*, measuring approximately 1.5–3 cm in length, were collected from a rocky shore (mid-upper shore) at low tide from near the quay at Portnahaven, Isle of Islay, Argyll and Bute, Scotland (National Grid Reference NR 16614 51966) on 26th June 2023. Upon collection, the dogwhelks were immediately fixed and stored in 70% ethanol to preserve the tissue for subsequent genomic analysis. A single male specimen was later processed for sequencing.

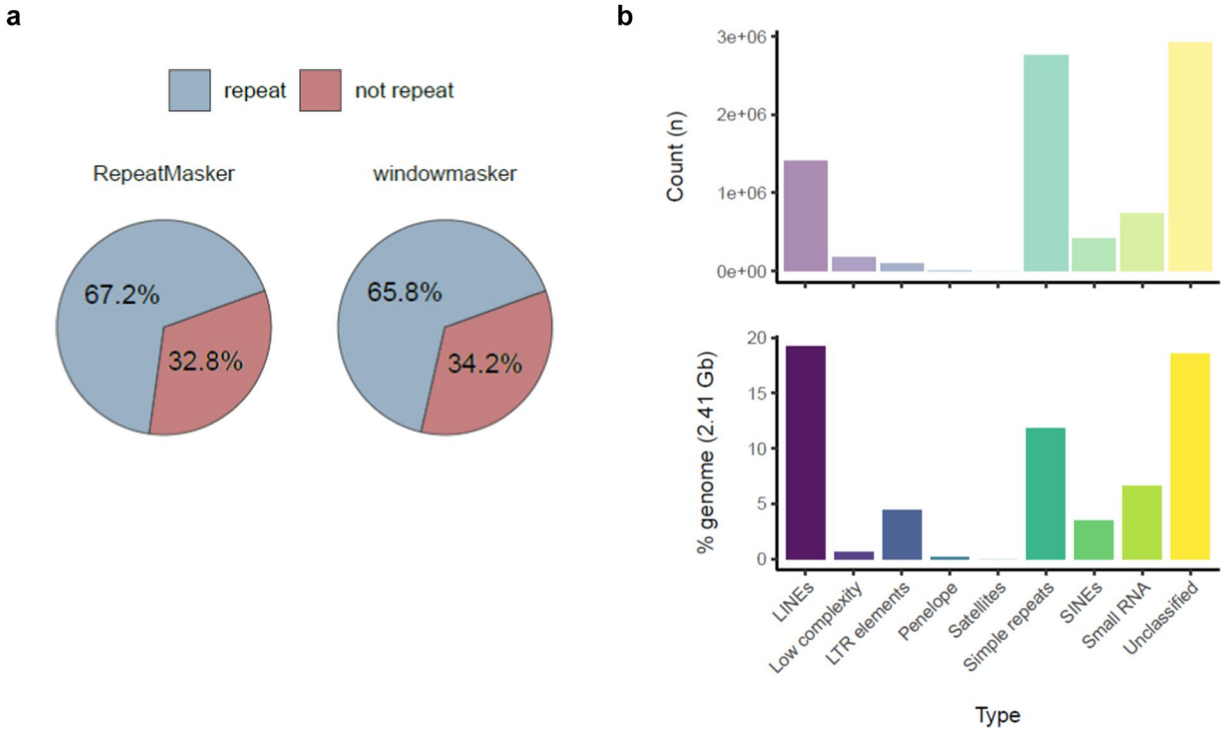

Fig. 1 The *N. lapillus* genome is rich in repeat sequences. (**a**) Proportion of the bases in the 2.41 Gb genome sequence masked by windowmasker or RepeatMasker. (**b**) The different families of repeat sequences identified by RepeatMasker, their frequency in the genome and the proportion of the genome they represent.

	***N. lapillus* (this study)**	***R. venosa***	***S. haemastoma***
Total length (Gb)	2.414	2.300	2.236
Contigs (n)	11,397	5,242	17,357
Contigs N50 (Kb)	338	433	200
Genes (n)	47,238	29,649	34,863
Exons (n)	263,697	180,204	168,274
Exons per gene (n)	5.6	6.1	4.8
BUSCO proteins (5925 orthologs, mollusca_odb10)	76.0% [S:59.9%, D:16.1%]	81.8% [S:69.9%, D:11.9%]	41.3% [S:37.3%, D:4.0%]

Table 2. Comparison of the assembly and annotation produced for *N. lapillus* with other assemblies available for the Muricidae family of gastropods.

Genomic DNA extraction. To prepare the specimen for lysis, the outer hard shell of the dogwhelk was first removed, and the soft tissue obtained was thoroughly washed three times with cold phosphate buffered saline to eliminate debris that could interfere with downstream processes. The entire organism (370 mg) was then dissected into small fragments using a sterile scalpel and homogenized with a sterile plastic pestle (NEB). High molecular weight genomic DNA was extracted from this using the QIAGEN Genomic-tip 100/G Kit following the manufacturer's protocol for tissue extraction with the following modifications. To ensure complete tissue lysis, the homogenized tissue was lysed by overnight incubation (~16 hours) at 50 °C in 9.5 mL lysis buffer G2 containing Proteinase K (1.25 mg/mL) and RNase A (200 µg/mL). Following lysis, the sample was centrifuged at 3000 × g for 5 minutes to remove unwanted debris. The supernatant containing the genomic DNA was then subjected to gravity flow extraction and clean-up using the recommended Genomic-tip protocol. All pipetting steps were performed using wide-bore tips to minimise fragmentation of the DNA. The extracted DNA was stored at −80 °C until further use. 370 mg of starting material yielded 17 µg (67 ng/µl) genomic DNA with A260/A280 1.85 and A260/A230 2.22. The concentration was determined using the Qubit dsDNA High Sensitivity Assay Kit on a Qubit Fluorometer (Thermo Fisher), and the UV absorbance ratios were measured using a Nanodrop spectrophotometer. DNA quality and fragment size were evaluated via TapeStation analysis (Agilent) and visualized on a 0.7% agarose gel (Fig. 3).

Oxford nanopore sequencing. The DNA libraries were prepared using ONT's Ligation Sequencing Kit V14 (SQK-LSK114) following the manufacturer's protocol. 1000 ng of genomic DNA was transferred into a 1.5 ml Eppendorf DNA LoBind tube and adjusted to a final volume of 47 µl with nuclease-free water. The DNA was

Species	Assembly	Source
Nucella lapillus	Quickmerge of the PacBio hifiasm assembly with the ONT flye assembly	This study
Rapana venosa	na	molluscDB2.0
Stramonita haemastoma	na	molluscDB2.0
Pomacea canaliculata	GCF_003073045.1_ASM307304v1	NCBI
Aplysia californica	GCF_000002075.1_AplCal3.0	NCBI
Haliotis rufescens	GCF_023055435.1_xgHalRufe1.0.p	NCBI
Littorina saxatilis	GCA_037325665.1_US_GU_Lsax_2.0	NCBI
Haliotis rubra	GCF_003918875.1_ASM391887v1	NCBI
Conus ventricosus	na	molluscDB2.0
Octopus vulgaris	GCA_951406725.2_xcOctVulg1.2	NCBI
Crassostrea gigas	GCF_963853765.1_xbMagGiga1.1	NCBI
Biomphalaria glabrata	GCF_947242115.1_xgBioGlab47.1	NCBI
Elysia chlorotica	GCA_003991915.1_ElyChl2.0	NCBI
Lottia gigantea	GCF_000327385.1_Helro1	NCBI
Achatina immaculata	na	molluscDB2.0

Table 3. Reference genomes used in this study. Assemblies indicated as being from NCBI in the source column were downloaded from https://ftp.ncbi.nlm.nih.gov/genomes, and those from molluscDB2.0 were obtained from http://mgbase.qnlm.ac/page/download/genomeDownload.

	C. ventricosus	*L. saxatilis*	*P. canaliculata*
Total length (Gb)	3.59	1.256	0.44
Contigs (n)	19399	4070	23
Contigs N50 (Kb)	114	270	1072
Genes (n)	32675	25144	24194
Exons (n)	146286	213470	495007
Exons per gene (n)	4.5	8.5	20.5
BUSCO proteins (5925 orthologs, mollusca_odb10)	19.4% [S:18.8%, D:0.6%]	82.0% [S:79.5%, D:2.5%]	93.7% [S:91.6%, D:2.1%]

Table 4. Assembly and annotation statistics for species from the Caenogastropoda subclass of gastropods, for comparison with the data for *N. lapillus* presented in Table 2.

combined with 7 µl of NEBNext Formalin-Fixed Parafin-Embedded (FFPE) DNA Repair Buffer, 2 µl of NEBNext FFPE DNA Repair Mix, and 3 µl of Ultra II End Prep Enzyme Mix to a total reaction volume of 60 µl. The use of FFPE reagents was recommended by ONT to repair chemical and mechanical DNA damage that can occur during the extraction process, even in non-FFPE samples. The reaction was mixed thoroughly by pipetting and briefly spinning down. The DNA was incubated in a thermal cycler at 20 °C for 5 minutes followed by 65 °C for 5 minutes. After incubation, 60 µl of resuspended AMPure XP beads (Beckman Coulter) were added to the sample and mixed by flicking the tube. The sample was incubated at room temperature for 5 minutes on a Hula mixer, and then placed on a magnetic rack to pellet the beads. The supernatant was discarded, and the beads were washed twice with 200 µl of freshly prepared 80% ethanol. Residual ethanol was removed, and the beads were air-dried for 30 seconds. The DNA was eluted by resuspending the bead pellet in 61 µl of nuclease-free water, followed by a 2-minute incubation at room temperature. The beads were pelleted again on a magnet, and 61 µl of the clear eluate was transferred to a clean tube. DNA concentration was quantified using a Qubit Fluorometer (Thermo Fisher), and the sample was immediately taken forward for adapter ligation or stored at 4 °C. For adapter ligation, the DNA library (60 µl) was mixed with 5 µl of Ligation Adapter (LA), 25 µl of Ligation Buffer (LNB), and 10 µl of Salt-T4 DNA Ligase (New England Biolabs). The reaction was incubated at room temperature for 10 minutes, AMPure XP beads (40 µl) were added to the ligation mixture, and the sample was incubated for 5 minutes on a Hula mixer. The beads were pelleted on a magnetic rack, and the supernatant was discarded. The beads were washed twice with Long Fragment Buffer (LFB) and finally, DNA was eluted in 25 µl of Elution Buffer following a 10-minute incubation. The final DNA library was quantified using a Qubit Fluorometer and prepared for sequencing on the GridION P2 Solo platform. Approximately 35–40 fmol of the library was loaded onto the PromethION Flow Cell on the GridION P2 Solo device for a 72-hour sequencing run. The flow cells were washed approximately every 8–18 hours to increase sequencing output.

PacBio HiFi sequencing. Due to the presence of fragmentation in the genomic DNA and in order to enrich the sample with longer genome fragments, 10 µg of DNA were size selected prior to library preparation using a BluePippin 0.75% gel (Sage Science) using a marker S1 6–10 kb v3 High Pass protocol, with a lower cut off point of 7 kb. The elution wells were rinsed with 40 µL buffer and Tween solution. All eluate was consolidated on 1:1 v/v SMRTbell beads and incubated at room temperature for 15 minutes, then placed on a magnetic rack to pellet the beads. The supernatant was discarded, and the beads were washed twice with 200 µl of freshly prepared 80%

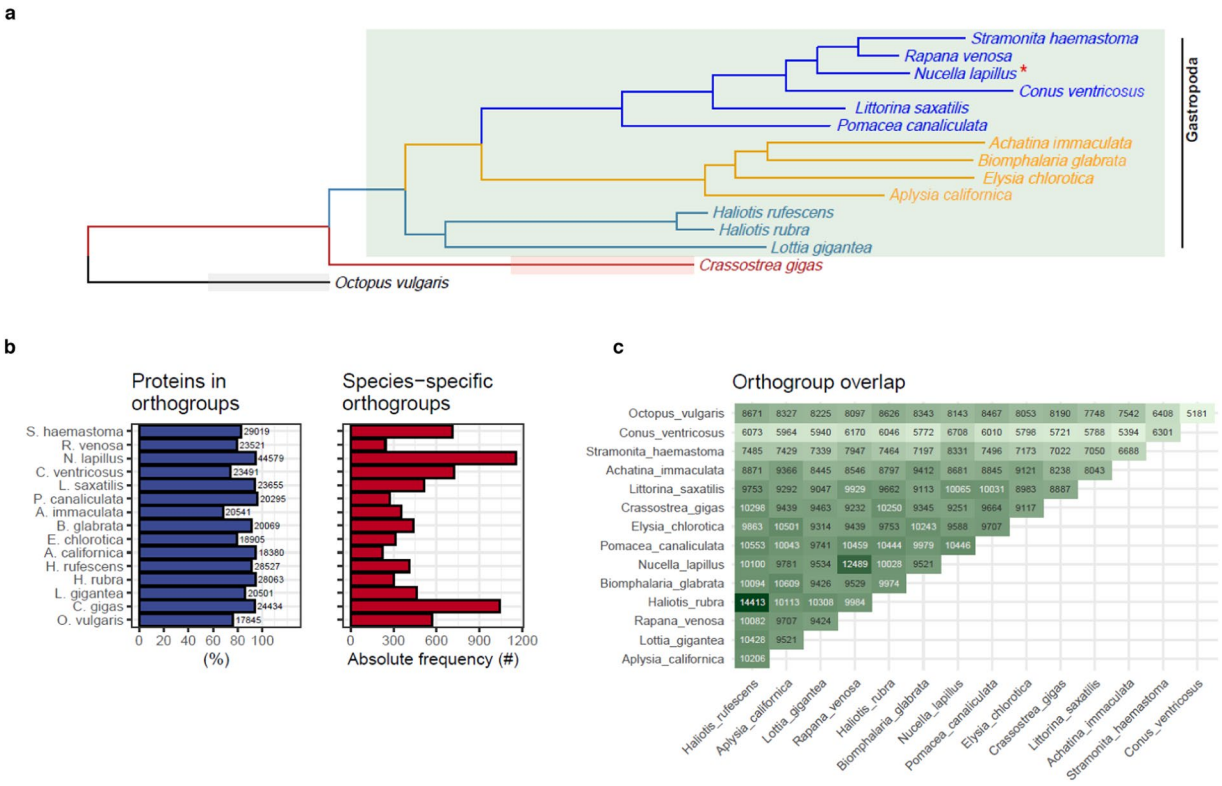

Fig. 2 Comparison of the proteins predicted from annotation of the *N. lapillus* genome assembly with the proteomes from 14 other related marine species using OrthoFinder. (**a**) Phylogenetic rooted species tree generated using the multiple sequence alignment method. Key to text label colours: black = Cephalapoda; red = Bivalvia; green = Vetigastropoda; yellow = Heterobranchia; blue = Caenogastropoda. (**b**) The % of proteins assigned to orthologous groups from each species (with absolute numbers given as text labels) (left panel), and the numbers of orthogroups found only in a particular species (right panel). (**c**) The number of orthogroups shared between each species-pair, where more closely related species tend to share more orthogroups.

ethanol. The beads were air-dried for 30 seconds to remove any residual ethanol before elution in 60 µL of low TE buffer. 5 µg of recovered genomic DNA were subjected to a DNA repair procedure using an NEB Next kit (New England Biolabs) by addition of 3.5 µL of NEBNext FFPE DNA Repair buffer and 2 µL of NEBNext FFPE Repair Mix, followed by incubation in a thermal cycler (Bio-Rad Laboratories) at 20 °C for 1 hour then 65 °C for 30 minutes.

After incubation, the DNA was cleaned up with 65 µL of SMRTbell beads as previously described, eluting in 47 µL of low TE buffer. The resulting DNA was subjected to SMRT bell library preparation using the PacBio SPK 3.0 kit (Pacific Biosciences of California) and following the manufacturers protocol with no deviations (protocol REV04 Apr2024). 1.03 µg of library was obtained after the final elution step and this was subject to further size selection on BluePippin as described previously using 8 kb as the lower cut off point. This second size selection step excludes any smaller SMRTbell templates that can arise for technical reasons (eg. incomplete or partial ligation of hairpin adapters to the ends of DNA fragments; degradation during handling), and yielded 432 ng of library with a peak size of 17603 bp.

The library was prepared for two PacBio Revio SMRT cells (Pacific Biosciences of California), with target on plate loading concentration (OPLC) of 235 nM and 200 nM, respectively, using a Revio Polymerase Binding kit and following the instructions generated by SMRTlink v.13.1.0.221970. The Revio run was set up without adaptive loading and had a duration of 30 hours, using Instrument Control SW version 13.1.0.221972 and Instrument Chemistry Bundle 13.1.0217683.

Genome assembly. Several assembly pipelines were tested and compared, as listed in Table 1. The final assembly was generated using the following procedure. The reads from ONT sequencing were assembled using flye[20] v2.9.4 with the default settings, and a separate assembly of the PacBio HiFi reads was generated using hifasm[21] v0.19. These two individual assemblies were then merged using Quickmerge[22,23] v0.3, with the PacBio assembly used as the base and the ONT flye assembly being merged in. Possible contaminants in the assembly were identified and removed using fcs-gx[24] v0.5.4 from NCBI, and repeat sequences in the genome were then soft-masked using windowmasker[25] v1.0. The number of contaminants identified was low, with 65 contigs potentially of bacterial origin being removed from the initial assembly total of 11,397 contigs. The other assemblies in Table 1 correspond to: (1) A hybrid assembly using hifasm, supplying all PacBio HiFi reads plus ONT long

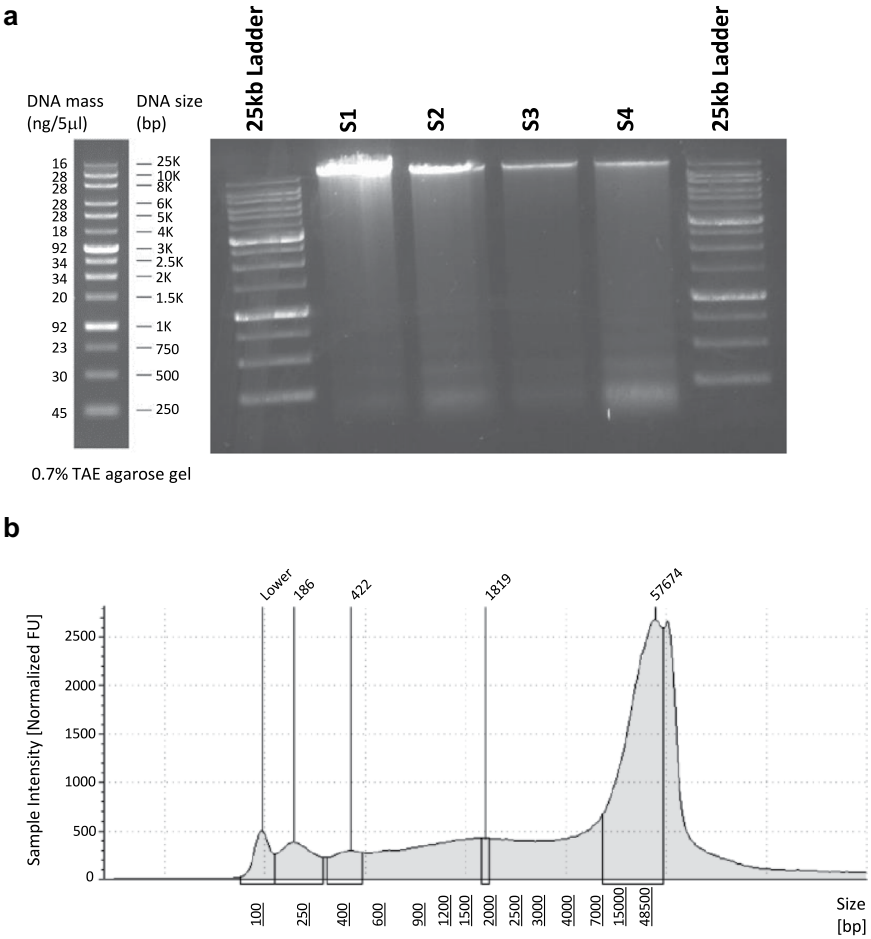

Fig. 3 Quality assessment of genomic DNA extraction. (**a**) Agarose Gel Electrophoresis of DNA Integrity and Size Agarose gel electrophoresis (0.7%) demonstrates the size distribution and integrity of DNA extracted using the QIAGEN Genomic-tip 100/G Kit. Samples labeled S1-S4 were derived from the same biological replicate to ensure consistency. A total of 6 µl of each DNA sample was loaded onto the gel. The gel image illustrates distinct bands, confirming the presence of high molecular weight and intact DNA across all samples that are >25kB. (**b**) Tape Station Analysis of DNA Fragment Size Tape station analysis of DNA extracted using the QIAGEN Genomic-tip 100/G Kit. The analysis indicates that the majority of the extracted DNA fragments exceed 22–40 kb in length, demonstrating the high molecular weight and integrity of the isolated DNA. This profile is indicative of successful extraction for subsequent genomic applications.

reads filtered to remove reads shorter than 3 Kb; (2) A hybrid assembly using verkko[26] v2.1, supplying all PacBio HiFi reads plus ONT long reads corrected using DeChat v1.0.1 (https://doi.org/10.21203/rs.3.rs-4384428/v1); (3) A merged assembly using Quickmerge using the flye assembly of ONT long reads as the base, and the hifasm assembly of the PacBio HiFi reads for merging in. Assemblies were assessed using BUSCO[15] v5.7.1 and QUAST[27] v5.2.0. Prior to assembly, k-mer frequencies within raw sequencing reads were analyzed using jellyfish[28] and GenomeScope v2.0[29] to estimate size, heterozygosity, and repetitiveness.

Genome annotation. Annotation of protein-coding genes in the cleaned and masked genome assembly was performed using GALBA v1.0.11, an automated pipeline that uses proteins from a closely related species to assist in the training of gene prediction using AUGUSTUS[30]. Proteins from *R. venosa* were provided for this purpose, and the miniprot option was used to perform the protein-to-genome alignments[31]. Functional annotation of predicted protein-coding genes was performed using eggNOG-mapper[32] v2.1.12, and additionally annotated with best hit BLAST results (v2.16.0) against the proteomes of the following marine gastropod species: *Rapana venosa*, *Littorina. saxatilis*, *Pomocea canaliculata, Stramonita haemastoma* and *Haliotis rufescens*.

Comparative genomics and phylogenetic analysis. The protein sequences predicted from the annotation of the *N. lapillus* genome were compared to those available for the 14 marine species listed in Table 3, including 12 representatives from three different subclasses of Gastropoda. All proteomes were processed to retain only the single longest protein sequence per gene, then compared using OrthoFinder[17,33] v2.5.5. OrthoFinder is designed to determine the phylogenetic relationship between gene sequences, identifying duplication events, orthologs and paralogs while being robust to processes such as gene duplication, loss, incomplete lineage sorting, and gene tree inaccuracies.

Assessment of repeat sequences. The types of repeat sequences in the genome assembly were analysed using RepeatModeler and RepeatMasker[34], versions 2.0.5 and 4.1.7 respectively.

Data Records

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JBJHGG000000000[35]. Raw sequence is available from NCBI Sequence Read Archive via the accession number SRP546914[36]. The genome annotation is available from *Zenodo*[37].

Technical Validation

As described above, the completeness of the assembly and annotation was assessed using BUSCO with the mollusca_odb10 database. The assembly sequence is estimated to be 84% complete, and the predicted annotation of 47,238 proteins is estimated to be 75% complete. Phylogenetic analyses was used to validate the precision of our genome assembly and annotation. The longest protein-coding gene sequences predicted for each locus annotated in the *N. lapillus* genome were compared to those available for 14 marine species, including 12 from three subclasses within the class Gastrapoda (Vetigastropoda, Heterobranchia, Caenogastropoda), using OrthoFinder 2.5.5. This analysis as expected places *N. lapillus* within the Caenogastropoda subclass of Gastropoda, and closest to relatives from the Muricidae family (Fig. 2a). Completeness in terms of length was also assessed in the context of the estimates made from the reported mass of a haploid copy of the *N. lapillus genome* (2.8 pg)[38], and a kmer analysis of the sequencing reads using GenomeScope[29]. Given the average molecular weight of a base pair of DNA is 650 Daltons, we calculated an expected length of approximately 2.59 Gb, while GenomeScope estimates a size range of 1.78–2.56 Gb using 19-mers. The 2.41 Gb assembly reported here represents 93% of the highest of these estimates. Genome sequences are subject to iterative rounds of assembly and annotation as new data becomes available, and it is expected that the first version of the genome presented here will be revised in subsequent versions. Transcriptome data will be particularly useful to define the annotation of gene locations.

Code availability

No custom code was used. All commands and pipelines used for data processing were executed according to the manuals and protocols of the corresponding bioinformatics software.

Received: 8 December 2024; Accepted: 5 March 2025;
Published online: 28 April 2025

References

1. D'Ambrosio, M., Goncalves, C., Calmao, M., Rodrigues, M. & Costa, P. M. Localization and Bioreactivity of Cysteine-Rich Secretions in the Marine Gastropod Nucella lapillus. *Mar Drugs* **19**, https://doi.org/10.3390/md19050276 (2021).
2. Song, H. *et al*. Chromosome-level genome assembly of the caenogastropod snail Rapana venosa. *Sci Data* **10**, 539, https://doi.org/10.1038/s41597-023-02459-7 (2023).
3. Farhat, S., Modica, M. V. & Puillandre, N. Whole Genome Duplication and Gene Evolution in the Hyperdiverse Venomous Gastropods. *Mol Biol Evol* **40**, https://doi.org/10.1093/molbev/msad171 (2023).
4. Schoyen, M. *et al*. Levels and trends of tributyltin (TBT) and imposex in dogwhelk (Nucella lapillus) along the Norwegian coastline from 1991 to 2017. *Mar Environ Res* **144**, 1–8, https://doi.org/10.1016/j.marenvres.2018.11.011 (2019).
5. Giltrap, M. *et al*. Use of caged Nucella lapillus and Crassostrea gigas to monitor tributyltin-induced bioeffects in Irish coastal waters. *Environ Toxicol Chem* **28**, 1671–1678, https://doi.org/10.1897/08-384.1 (2009).
6. Nicolaus, E. E. & Barry, J. Imposex in the dogwhelk (Nucella lapillus): 22-year monitoring around England and Wales. *Environ Monit Assess* **187**, 736, https://doi.org/10.1007/s10661-015-4961-0 (2015).
7. Harrison, T. D., Gilmour, G., McNeill, M. T., Armour, N. & McIlroy, L. Survey of imposex in Nucella lapillus as an indicator of tributyltin pollution in Northern Irish coastal waters, 2004 to 2017. *Mar Pollut Bull* **159**, 111474, https://doi.org/10.1016/j.marpolbul.2020.111474 (2020).
8. Gomes, D. M. *et al*. Long-term monitoring of Nucella lapillus imposex in Ria de Aveiro (Portugal): When will a full recovery happen? *Mar Pollut Bull* **168**, 112411, https://doi.org/10.1016/j.marpolbul.2021.112411 (2021).
9. Oliveira, I. B. *et al*. Spatial and temporal evolution of imposex in dogwhelk Nucella lapillus (L.) populations from North Wales, UK. *J Environ Monit* **11**, 1462–1468, https://doi.org/10.1039/b906766c (2009).
10. Castro, L. F. *et al*. Imposex induction is mediated through the Retinoid X Receptor signalling pathway in the neogastropod Nucella lapillus. *Aquat Toxicol* **85**, 57–66, https://doi.org/10.1016/j.aquatox.2007.07.016 (2007).
11. Morais, H., Arenas, F., Cruzeiro, C., Galante-Oliveira, S. & Cardoso, P. G. Combined effects of climate change and environmentally relevant mixtures of endocrine disrupting compounds on the fitness and gonads' maturation dynamics of Nucella lapillus (Gastropoda). *Mar Pollut Bull* **190**, 114841, https://doi.org/10.1016/j.marpolbul.2023.114841 (2023).
12. Galante-Oliveira, S. *et al*. Factors affecting RPSI in imposex monitoring studies using Nucella lapillus (L.) as bioindicator. *J Environ Monit* **12**, 1055–1063, https://doi.org/10.1039/b921834c (2010).
13. Minchin, A. & Davies, I. M. Imposex measurement in the dogwhelk Nucella lapillus (L.)–temporal aspects of specimen preparation. *J Environ Monit* **1**, 239–241, https://doi.org/10.1039/a902836f (1999).
14. Fehrenbach, G. W., Pogue, R., Carter, F., Clifford, E. & Rowan, N. Implications for the seafood industry, consumers and the environment arising from contamination of shellfish with pharmaceuticals, plastics and potentially toxic elements: A case study from Irish waters with a global orientation. *Sci Total Environ* **844**, 157067, https://doi.org/10.1016/j.scitotenv.2022.157067 (2022).
15. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
16. Yuan, J. *et al*. Simple sequence repeats drive genome plasticity and promote adaptive evolution in penaeid shrimp. *Commun Biol* **4**, 186, https://doi.org/10.1038/s42003-021-01716-y (2021).
17. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238, https://doi.org/10.1186/s13059-019-1832-y (2019).
18. Panova, M. *et al*. DNA Extraction Protocols for Whole-Genome Sequencing in Marine Organisms. *Methods Mol Biol* **1452**, 13–44, https://doi.org/10.1007/978-1-4939-3774-5_2 (2016).
19. Boughattas, S. *et al*. Whole genome sequencing of marine organisms by Oxford Nanopore Technologies: Assessment and optimization of HMW-DNA extraction protocols. *Ecol Evol* **11**, 18505–18513, https://doi.org/10.1002/ece3.8447 (2021).

20. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540–546, https://doi.org/10.1038/s41587-019-0072-8 (2019).
21. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
22. Solares, E. A. *et al*. Rapid Low-Cost Assembly of the Drosophila melanogaster Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3 (Bethesda)* **8**, 3143–3154, https://doi.org/10.1534/g3.118.200162 (2018).
23. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res* **44**, e147, https://doi.org/10.1093/nar/gkw654 (2016).
24. Astashyn, A. *et al*. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol* **25**, 60, https://doi.org/10.1186/s13059-024-03198-7 (2024).
25. Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141, https://doi.org/10.1093/bioinformatics/bti774 (2006).
26. Rautiainen, M. *et al*. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat Biotechnol* **41**, 1474–1482, https://doi.org/10.1038/s41587-023-01662-6 (2023).
27. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150, https://doi.org/10.1093/bioinformatics/bty266 (2018).
28. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
29. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
30. Bruna, T. *et al*. Galba: genome annotation with miniprot and AUGUSTUS. *BMC Bioinformatics* **24**, 327, https://doi.org/10.1186/s12859-023-05449-z (2023).
31. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, https://doi.org/10.1093/bioinformatics/btad014 (2023).
32. Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* **38**, 5825–5829, https://doi.org/10.1093/molbev/msab293 (2021).
33. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, 157, https://doi.org/10.1186/s13059-015-0721-2 (2015).
34. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
35. Hesketh, A., Kadiwala, J., De Weerd, H., Ritch, H. & Shakur, R. Nucella lapillus isolate BIG2024_Nlap_m_1, whole genome shotgun sequencing project. *GenBank* https://www.ncbi.nlm.nih.gov/nuccore/JBJHGG000000000 (2024).
36. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP546914 (2024).
37. Hesketh, A., Kadiwala, J., De Weerd, Ritch, H. & Shakur, R. Genome annotation file for a draft genome assembly for Nucella lapillus. *Zenodo* https://doi.org/10.5281/zenodo.14170281 (2024).
38. Pascoe, P. L., Jha, A. N. & Dixon, D. R. Variation of karyotype composition and genome size in some muricid gastropods from the northern hemisphere. *Journal of Molluscan Studies* **70**, 389–398, https://doi.org/10.1093/mollus/70.4.389 (2004).

Acknowledgements

We thank Dr Martin J. Willing, (Vice President), The Conchological Society of Great Britain and Ireland for collection of the specimen used for sequencing, and Dr Corina Ciocan, School of Applied Sciences, University of Brighton, for helping in arranging sample collection and for useful discussions. PacBio Revio sequencing was performed by Edinburgh Genomics at the University of Edinburgh, supported by BBSRC grant BB/X019586/1.

Author contributions

J.K. and H.R. performed experiments. A.H. and H.D.W. processed and analysed data. R.S. conceived and supervised the study. A.H., J.K. and R.S. drafted the manuscript, and H.D.W. and H.R. contributed to the draft.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.S.