G3, 2023, **13(9)**, jkad146
https://doi.org/10.1093/g3journal/jkad146
Advance Access Publication Date: 3 July 2023
Genome Report

Characteristics of the spiny dogfish (*Squalus acanthias*) nuclear genome

C. Isabel Wagner,[1,*] Martina E.L. Kopp,[1] James Thorburn,[2,3] Catherine S. Jones,[4] Galice Hoarau,[1] Leslie R. Noble[1,*]

[1]Faculty of Biosciences and Aquaculture, Nord University, 8026 Bodø, Norway
[2]School of Biology, University of St Andrews, St Andrews, KY16 9ST, UK
[3]School of Applied Sciences, Edinburgh Napier University, Edinburgh, EH11 4BN, UK
[4]School of Biological Sciences, University of Aberdeen, Aberdeen, AB24 3FX, UK

*Corresponding authors: Email: isabel.wagner@nord.no; leslie.r.noble@nord.no

Abstract

Sequenced shark nuclear genomes are underrepresented, with reference genomes available for only four out of nine orders so far. Here, we present the nuclear genome, with annotations, of the spiny dogfish (*Squalus acanthias*), a shark of interest to biomedical and conservation efforts, and the first representative of the second largest order of sharks (Squaliformes) with nuclear genome annotations available. Using Pacific Biosciences Continuous Long Read data in combination with Illumina paired-end and Hi-C sequencing, we assembled the genome de novo, followed by RNA-Seq-supported annotation. The final chromosome-level assembly is 3.7 Gb in size, has a BUSCO completeness score of 91.6%, and an error rate of less than 0.02%. Annotation predicted 33,283 gene models in the spiny dogfish's genome, of which 31,979 are functionally annotated.

Keywords: *Squalus acanthias*, nuclear genome, de novo assembly, Selachii, shark

Introduction

Despite intense interest in sharks, many important areas of their biology remain largely unexplored. Although biology is in the era of genomics, only twelve of the over 500 described shark species (Fricke *et al.* 2023, accessed 15.02.2023) have sequenced nuclear genomes, and of those only nine have genome annotation information connected to them (Read *et al.* 2017; Hara *et al.* 2018; Marra *et al.* 2019; Weber *et al.* 2020; Zhang *et al.* 2020; Rhie *et al.* 2021; Nishimura *et al.* 2022; Sayers *et al.* 2022; Stanhope *et al.* 2023).

We report the sequencing, assembly, and annotation of the thirteenth shark genome, that of the spiny dogfish (*Squalus acanthias*). This expands the number of shark orders with available genome annotation information from three to four out of nine. Furthermore, Squaliformes is the second-largest shark order, making the genome annotations of *S. acanthias* the closest related to the 140 species within that order (Fricke *et al.* 2023), which could facilitate genome research for all of them. In particular, this resource could assist annotation of the nuclear genome of *Squalus suckleyi* (Ebert *et al.* 2010), the publicly available nuclear reference genome of which (Sayers *et al.* 2022) awaits annotation.

S. acanthias, a medium sized shark, occupies all oceans except for the North Pacific (Ebert *et al.* 2010). It has attracted interest from a biomedical perspective [e.g. as a source of the antibiotic squalamine (Moore *et al.* 1993)]. Furthermore, it was once dubbed possibly the most abundant extant shark but has suffer rapid and well documented, fisheries-induced population declines (Compagno 1984; Ellis *et al.* 2015, 2016; Finucci *et al.* 2020). Conservation of this species will benefit from better understanding and characterization of markers for genomic regions, enabling more direct associations between gene function and environmental parameters. Therefore, we anticipate genome characterization will advance scientific endeavors in these and other areas, allowing further genomic exploration and conservation of this species.

To sequence the genome of *S. acanthias*, we non-lethally sampled skin, muscle, and blood from a female in the North-East Atlantic. We then employed Pacific Biosciences (PacBio) Continuous Long Reads (CLRs) in combination with Illumina paired end (PE) and Hi-C sequencing for de novo assembly, followed by annotation using publicly available transcriptome datasets (Chana-Munoz *et al.* 2017). From this, we generated a high-quality, annotated draft genome, which allowed a first view of the unique characteristics comprising the nuclear genome of *S. acanthias*.

Material and methods

Sampling

A female spiny dogfish (total length 71 cm) was caught by rod and line with a baited, barbless hook in the Lynn of Lorn, UK, at 56°28′22″N 5°25′30″W, August 2019. Two tissue samples (muscle and skin) of ø 5 and 2 ml of whole blood, split into two 1.3 ml lithium heparin tubes, were sampled. All samples were immediately flash frozen in liquid nitrogen; subsequent storage was between −78.5 and −80°C. The individual was released alive.

Sampling was conducted under the Animals (Scientific Procedures) Act 1986, Project License #P05E95C50.

Received: February 28, 2023. **Accepted:** June 12, 2023

DNA extraction for PacBio and Illumina short read sequencing

High molecular weight (HMW) DNA was extracted from frozen whole blood with the MagAttract HMW DNA kit (QIAGEN, Venlo, Netherlands), following 10× Genomics (Pleasanton, USA) recommendation with additional modifications and adjusted for the DNA content of nucleated blood cells in sharks (Saunders 1966; Hardie and Hebert 2003).

Separate extractions, each using 3, 5, 7.5, 10, 15, 20, 25, or 50 µl of whole blood, were performed. Briefly, whole blood was added to Proteinase K and mixed with RNase A and Buffer AL by pulse-vortexing. MagAttract Suspension G was then added to the mix, followed by Buffer MB. Two washing steps were performed with Buffer MW1, followed by two washing rounds with Buffer PE, and two rounds of washing with nuclease-free water. Final HMW DNA was eluted twice, first with 150 µl AE buffer, and again with 50 µl. Extracts were stored at −20°C and shipped on dry ice. See Supplementary File 1 for detailed protocol.

Genome sequencing

Pacific Biosciences (PacBio) and short-read Illumina sequencing were performed by the Functional Genomics Laboratory and Vincent J. Coates Genomics Sequencing Laboratory, California Institute for Quantitative Biosciences (QB3), University of California, Berkeley, USA. PacBio CLR libraries were prepared with the SMRTbell Express Template Prep Kit 2.0, Sequel II Binding Kit 1.0 and Sequel II Internal Control Complex 1.0 and sequenced on two SMRTcells on a PacBio Sequel II machine (Pacific Biosciences of California, Inc., Menlo Park, USA).

For paired-end Illumina sequencing, DNA was fragmented on a Bioruptor Pico (Diagenode, Seraing, Belgium) and libraries prepared with the KAPA Hyper Prep kit for DNA (F. Hoffmann-La Roche AG, Basel, Switzerland), using six Polymerase Chain Reaction (PCR) cycles and a mean insert size of 430 bp. Library quality was evaluated on a Fragment Analyzer (AATI, now Agilent, Santa Clara, USA), molarity was assessed via quantitative PCR on a CFX Connect thermal cycler (BioRad, Hercules, USA), using the Kapa Biosystems Illumina Quant qPCR Kits (F. Hoffmann-La Roche AG, Basel, Switzerland). Libraries were pooled according to molarity and sequenced on a NovaSeq6000 150PE S4 flow cell (Illumina, Inc., San Diego, USA). The raw sequencing data was transferred into fastq-format via bcl2fastq2 (v. 2.20, Illumina Inc. 2019).

Hi-C sequencing was performed by the Norwegian Sequencing Centre, University of Oslo, Oslo, Norway. Two libraries were prepared from 0.08 g muscle tissue with the Dovetail Omni-C kit and Omni-C proximity Ligation Assay (v. 1.0, Dovetail Genomics, Scotts Valley, USA), and sequenced on an Illumina NovaSeq 150PE S4 flow cell (Illumina, Inc., San Diego, USA).

Genome size estimation

Size of the nuclear genome was estimated with Jellyfish (v. 2.2.8, Marçais and Kingsford 2011) in combination with GenomeScope (online v. 1.0, Vurture *et al.* 2017).

Paired-end Illumina reads were first trimmed for adapters and low quality base calls with Trimmomatic (v. 0.39, Bolger *et al.* 2014), in PE mode for internally provided TruSeq3-PE-2 adapters. Seed mismatches were set to 2, palindrome and simple clip threshold to 30 and 10, respectively, and the minimum adapter length to be removed set to 1 bp, with both reads being retained after adapter trimming. Trimming of read ends was performed with a quality threshold of 3, followed by sliding window trimming with a window size of 4 bp, and a required base quality of 15. Next, reads were trimmed for poly-G tails, with cutadapt (v. 2.10, Martin 2011), with a phred threshold of 20, and a minimum read length of 1 bp to be retained.

The trimmed, paired reads were then fed to Jellyfish to count canonical 21-mers using a hash size of 140.961 Gbp, followed by construction of a count histogram with a maximum count value of 10,000,000. GenomeScope then used the histogram, adjusted for 21-mers, a read length of 151 bp, and a maximum k-mer coverage of 1,000,000, to model the genome size.

As organellar sequences can confound nuclear genome size estimates (Vurture *et al.* 2017), in a second approach the trimmed reads were first mapped to the mitochondrial genome of *S. acanthias* (Rasmussen and Arnason 1999; Sayers *et al.* 2022) as well as Phi X control sequences (Illumina, Inc., San Diego, USA). To allow mapping at both ends of the circular mitochondrial genome, the first 151 bases of the mitochondrial, fasta-formated sequence were duplicated at its opposite end. Read pairs were then mapped with BWA-MEM (v. 0.7.17-r1188, Li 2013); mapped reads as well as their respective read mate were discarded using SAMtools fastq (flag -f 13, v. 1.14, Danecek *et al.* 2021). Genome size was then modeled following the procedure described above.

Assembly

The assembly process followed a modified version of the DNAnexus VGP assembly pipeline (v. 1.6), by Rhie *et al.* 2021. In the first step, raw PacBio subreads were assembled using the long-read assemblers Canu, Flye, and wtdbg2 (Koren *et al.* 2017; Kolmogorov *et al.* 2019; Ruan and Li 2020).

Canu assembler (v. 2.0, Koren *et al.* 2017) was adjusted for an estimated genome size of 5.7 Gb (Hardie and Hebert 2003). Longest reads were corrected and used up to a genome coverage of 200. Bogart was used for unitig constriction, with allowed standard deviations of read dissimilarity set to 3 for contig construction and bubble detection; and to 1 for repeat detection. Furthermore, heuristics for contig construction at repeats with multiple possible paths in the assembly graph were set to require a minimum of 500 bp or 50% larger overlap in the chosen path than in alternative paths.

For wtdbg2 (v. 2.5, Ruan and Li 2020), the estimated genome size was again set to 5.7 Gb. Following the authors' recommendation, only the longest subreads were used, and all reads shorter than 5,000 bp discarded. Consensus was called with wtpoa-cns (v. 2.5), part of wtdbg2. Flye (v. 2.8-b1674, Kolmogorov *et al.* 2019) was run with default settings for raw PacBio CLR data. The assemblies generated with Flye and Canu were chosen for downstream analysis.

Purging haplotypes

The assemblies produced by Canu and Flye were purged for uncollapsed haplotigs, using purge_dups (v. 1.2.5, Guan *et al.* 2020) in combination with minimap2 (v. 2.17-r941, Li 2018).

The purge_dups pipeline was run manually step by step, with a RAM threshold of 10 Gbp for minimap2. The assembly produced by Flye was purged in one round, with manually set cutoffs for the lower, middle, and upper read depth bounds of 5, 43, and 255, respectively. The Canu-derived assembly was purged in two consecutive rounds, first with manually set lower, middle, and upper bounds for read depths of 5, 21, and 126, respectively. In the second round, automatic cutoffs were used.

First scaffolding

Illumina-derived Hi-C reads were used for scaffolding the primary assembly using the Arima-HiC Mapping Pipeline (v. 02, https://github.com/ArimaGenomics/mapping_pipeline, Arima Genomics, Inc., San Diego, USA) and Salsa (v. 2.3, Ghurye *et al.* 2017, 2019).

PE reads were first trimmed with Trimmomatic (v. 0.39, Bolger *et al.* 2014), followed by cutadapt (v. 3.4, Martin 2011, settings see Genome size estimation). Briefly for mapping, reads were first aligned with BWA-MEM (v. 0.7.17, Li 2013), followed by filtering of chimeric reads, pairing of read pairs, and filtering for a mapping quality threshold of 10. PCR duplicates were removed using Picard (v. 2.26.2, Broad Institute 2019). Lastly, the mapped reads of both libraries were merged, before scaffolding with Salsa, with settings for Omni-C data (Ghurye *et al.* 2017, 2019).

Polishing

Scaffolds were polished with long-read data via Arrow (Chin *et al.* 2013), followed by polishing with Illumina short read data with Pilon (v. 1.24, Walker *et al.* 2014), one round each.

PacBio reads were aligned to the assembly with pbmm2 (v. 1.7.0, SMRT Link v. 10.2, Pacific Biosciences of California, Inc., Menlo Park, USA, Li 2018), and then used for polishing the assembly with the Arrow algorithm implemented in gcpp (v. 2.0.2, Pacific Biosciences of California, Inc., Menlo Park, USA, Chin *et al.* 2013). Previously trimmed Illumina reads (see Genome size estimation) were then mapped to the pre-polished assembly with BWA-MEM (v. 0.7.17, Li 2013), and used by pilon (Walker *et al.* 2014) to polish the assembly a second time, in diploid mode with manually assigned blocks. Each block used the read-mapping to the whole genome, but polished only sub-parts of the assembly, overcoming the issue of single-threading in pilon.

Contamination filtering

The polished assembly was filtered for possible contaminants of foreign species and mitochondrial genomes in a three-step approach, using BLAST+ (v. 2.12.0, Altschul *et al.* 1990; Camacho *et al.* 2009).

First, all scaffolds were submitted to a nucleotide-nucleotide search optimized for highly similar matches (megablast) against the NCBI nucleotide database (nt, accessed 18.01.2022, Sayers *et al.* 2021), limited to hits that passed an expectation value threshold of 1e-4, and a maximum of five target sequences to be retained in the output. In a second round, all scaffolds without a hit in the previous search were submitted to another nucleotide-nucleotide search, this time optimized for somewhat similar matches (blastn, database: nt). As before, hits were limited to those that passed an expectation value threshold of 1e-4, with a maximum of five target sequences retained. All scaffolds with a hit outside the class Chondrichthyes were subsequently removed from the data set.

To filter for possible mitochondrial genomes contained in the assembly, all surviving scaffolds were submitted to a nucleotide-nucleotide search optimized for highly similar matches (megablast) against the mitochondrial reference genome of *S. acanthias* (Rasmussen and Arnason 1999), again filtered for matches that passed an expectation value threshold of 1e-4. Any scaffolds of completely mitochondrial origin were discarded.

Second scaffolding

Following polishing and decontamination of the scaffolded assembly, the new scaffolding tool YaHS emerged (Zhou *et al.* 2023), and was therefore used to re-scaffold the polished and decontaminated scaffolds produced with Salsa (Ghurye *et al.* 2017, 2019), which had not reached chromosome-level lengths. The Salsa-derived scaffolds were re-scaffolded using the same trimmed Illumina-derived Hi-C reads as in the first scaffolding, which were mapped to the scaffolds with BWA-MEM (v. 0.7.17, Li 2013), then cleaned and merged using the Arima-HiC Mapping Pipeline (v. 02, https://github.com/ArimaGenomics/mapping_pipeline, Arima Genomics, Inc., San Diego, USA) as described before. Scaffolds were finally re-scaffolded with YaHS (v. 1.2a.1, Zhou *et al.* 2023). A Hi-C contact map was generated with Pre from Juicer Tools (version distributed with YaHS and stand-alone version 2.13.06, Durand, Shamim, *et al.* 2016), and visualized with Juicebox (v. 1.11.08, Durand, Robinson, *et al.* 2016).

Annotation

Repeat masking

The YaHS-derived spiny dogfish genome assembly was soft-masked with RepeatMasker (v. 4.1.2-p1, Altschul *et al.* 1990; Benson 1999; Camacho *et al.* 2009; Smit *et al.* 2015), using species-specific repeat libraries from the Extensive de novo TE Annotator pipeline (EDTA, v. 2.0.0, Xu and Wang 2007; Ellinghaus *et al.* 2008; Xiong *et al.* 2014; Ou and Jiang 2018, 2019; Ou *et al.* 2019; Shi and Liang 2019; Su *et al.* 2019; Zhang *et al.* 2022) and RepeatModeler (v. 2.0.3, Benson 1999; Bao and Eddy 2002; Price *et al.* 2005; Flynn *et al.* 2020), combined with two short interspersed nuclear elements (SINEs) previously identified in higher elasmobranchs or *S. acanthias* itself (Ogiwara *et al.* 1999; Nishihara *et al.* 2006).

In preparation of repeat library construction, coding DNA sequences (CDS) were identified via genome-guided transcriptome assemblies. Transcriptome data derived from four tissues (brain, liver, kidney, and ovary), previously published by Chana-Munoz *et al.* (2017, retrieved 14.02.022 from the European Nucleotide Archive, Cummins *et al.* 2022), was analyzed for quality and adapter contamination via FastQC (v. 0.11.9, Babraham Bioinformatics 2010), and then adapter and quality trimmed via fastp (v. 0.23.2, Chen *et al.* 2018). In fastp, first read correction was conducted by PE read overlap. Low quality bases at the 5' end of the read were dropped, with a phred score threshold of 20 within a 4 b sliding window. Following this, read pruning was conducted, starting again from the 5' end with a phred score threshold of 20 within a 4 b sliding window, dropping the right part of the read if base quality sank below the set threshold. Adapters (automatically detected, Nextera, TruSeq2 and TruSeq3 PE) were trimmed as well as poly-X tails, and finally reads were filtered for a minimum length of 2 bp.

Next, trimmed reads, paired as well as unpaired, were mapped to the Salsa-scaffolded genome with HISAT2 (v. 2.2.1, Kim *et al.* 2015, 2019). Each tissue-specific data set was individually mapped with settings for downstream transcriptome assembly, in a non-deterministic manner, and sorted via SAMtools (v. 1.14, Danecek *et al.* 2021). Genome-guided, tissue-specific transcriptome assemblies were then conducted and combined using StringTie2 (v. 2.2.1., Pertea *et al.* 2015), with a minimal transcript length threshold of 30 bp for initial assembly construction, followed by merging of the four assemblies with default parameters. Finally, CDS were extracted via the TransDecoder pipeline (v. 5.5.0, Haas 2018), in default mode.

CDS were fed to EDTA (v. 2.0.0, Xu and Wang 2007; Ellinghaus *et al.* 2008; Xiong *et al.* 2014; Ou and Jiang 2018, 2019; Ou *et al.* 2019; Shi and Liang 2019; Su *et al.* 2019; Zhang *et al.* 2022) for purging of gene sequences from a repeat library produced in default mode. A second species-specific repeat library was constructed using RepeatModeler (v. 2.0.3, Benson 1999; Bao and Eddy 2002; Price *et al.* 2005; Flynn *et al.* 2020), using seven rounds and sampling 1.1 Gb of the genome for repeat detection. This time, protein coding sequences were purged by querying the sequences against the UniProtKB/Swiss-Prot database (accessed: 30.5.2022, The UniProt Consortium 2021) in a translated-nucleotide to protein search with BLAST+ (blastx, v. 2.12.0, Altschul *et al.* 1990; Camacho *et al.* 2009) with an *e*-value threshold of 1e-3, and removal of aligning sequences from the repeat library.

Both libraries were combined, and two known SINEs were added (Ogiwara *et al.* 1999; Nishihara *et al.* 2006). "SacSINE1" from Nishihara *et al.* is species-specific for *S. acanthias,* but as the sequence of SINE "HE1" from Ogiwara *et al.* is not, it was queried against the NCBI nucleotide database via the nucleotide-nucleotide BLAST+ web interface (blastn, nt data base accessed: 05.04.2022, Altschul *et al.* 1990; Johnson *et al.* 2008; Camacho *et al.* 2009; Sayers *et al.* 2021) with default settings. One sequence matching in *S. acanthias* was then included in the repeat library. The final repeat library was implemented for soft masking the YaHS-derived *S. acanthias* genome assembly with RepeatMasker (Altschul *et al.* 1990; Benson 1999; Camacho *et al.* 2009; Smit *et al.* 2015), run in sensitive mode, using NCBI BLAST+ modified for RepeatMasker as the search engine and omitting the masking of low complexity DNA and simple repeats.

Gene prediction

BRAKER2 was used for gene prediction in the YaHS-scaffolded genome assembly, using both RNA-Seq and protein evidence (v. 2.1.6, Lomsadze *et al.* 2014, 2005; Stanke *et al.* 2006, 2008; Gotoh 2008; Li *et al.* 2009; Barnett *et al.* 2011; Iwata and Gotoh 2012; Buchfink *et al.* 2015; Hoff *et al.* 2016, 2019; Brůna *et al.* 2020, 2021), followed by TSEBRA (v. 1.0.3, Gabriel *et al.* 2021) to combine the results of different gene prediction approaches.

For support with RNA-Seq data, transcriptome data from Chana-Munoz *et al.* (2017), corrected and trimmed as described earlier, was tissue-specifically mapped to the soft masked genome with HISAT2 (v. 2.2.1, Kim *et al.* 2015, 2019) and sorted with SAMtools (v. 1.14, Danecek *et al.* 2021). BRAKER2 was then run with the combined data as input, skipping all parameter training and using the human BRAKER2 pre-trained parameter set. The human parameter set was chosen because parameter sets trained specifically for *S. acanthias,* or provided by BRAKER2 for other organisms more closely related to our target species than *Homo sapiens,* resulted in much lower BUSCO gene set completeness for the finally predicted gene set. In a second approach, the Vertebrata section of OrthoDB v10 (Kriventseva *et al.* 2019), modified by declaring all selenocysteines to be amino acids of unknown identity, was used by BRAKER2 as protein evidence, again with the human parameter set and skipping all parameter training. In a third approach, BRAKER2 was run combining the RNA-Seq alignments plus the protein evidence from the two previous runs, again with the human parameter set, skipping all parameter training.

All three approaches were then amalgamated in various combinations via TSEBRA (Gabriel *et al.* 2021), using either default parameters which exclude all genes predicted without extrinsic supporting evidence, or with developer-provided configuration parameters that also retain ab initio predicted genes. However, after evaluation of gene set completeness with BUSCO (v. 5.2.2, Manni *et al.* 2021), the initial BRAKER2 run with RNA-Seq evidence only showed the highest completeness and was thus chosen for downstream analysis.

Functional annotation

For functional annotation, proteins predicted by BRAKER2 (v. 2.1.6, Lomsadze *et al.* 2014, 2005; Stanke *et al.* 2006, 2008; Gotoh 2008; Li *et al.* 2009; Barnett *et al.* 2011; Iwata and Gotoh 2012; Buchfink *et al.* 2015; Hoff *et al.* 2016, 2019; Brůna *et al.* 2020, 2021) were queried against the vertebrata UniProtKB/Swiss-Prot database (accessed: 19.4.2023, The UniProt Consortium 2021) in a protein to protein search with BLAST+ (blastp, v. 2.13.0, Altschul *et al.* 1990; Camacho *et al.* 2009) with an *e*-value threshold of 1e-6 and the output restricted to the maximum of a single target sequence and High Scoring Pair per query. Furthermore, the annotated protein sequences were queried against the InterPro data base (accessed: 20.4.2023, Blum *et al.* 2021) with InterProScan (v. 5.61-93.0, Jones *et al.* 2014), with the precalculated match lookup service disabled, but including the lookup of Gene Ontology and Pathway annotations. Biological information was then attached to genome features with the script agat_sp_manage_functional_annotation.pl in AGAT (v. 1.0.0, Dainat 2022).

Evaluation

Sequence statistics [assembly size, N50 (the weighted median length of the assembled sequence length), fragment number, and the length of the longest fragment], sequence completeness levels and error rates of the different stages of the genome assembly were assessed via a custom python3 script using the Biopython package (Supplementary File 2, Python3 v. 3.8.5, Biopython v. 1.78, Cock *et al.* 2009), via BUSCO (v 4.1.4 - 5.2.2, Manni *et al.* 2021) and via Merqury involving meryl (both v. 1.3, Miller *et al.* 2008; Koren *et al.* 2017; Rhie *et al.* 2020). For the predicted gene sets, completeness was assessed via BUSCO only (v. 5.2.2, Manni *et al.* 2021).

BUSCO was run in genome mode with the vertebrata reference gene set (vertebrata_odb10, $n = 3,354$, Manni *et al.* 2021). In Merqury, the most appropriate k-mer size was determined for a potential genome size of 2.0 Gb (haploid), 11.0 Gb (diploid), and 14.4 Gb (diploid). A custom 21-mer database was then built from Illumina data trimmed as described earlier ("Genome size estimation"), using meryl (v. 1.3, Miller *et al.* 2008; Koren *et al.* 2017; Rhie *et al.* 2020), and counting the occurrence of canonical 21-mers for each data set individually before merging by summing them. Merqury was then run with the same meryl database for all stages of the genome assembly, in default mode.

For the predicted gene sets, completeness was assessed via BUSCO (v. 5.2.2, Manni *et al.* 2021), in protein mode with the vertebrata reference gene set (vertebrata_odb10, $n = 3,354$).

Results and discussion

Genome sequencing provided a total of 297.9 Gb of PacBio CLR data, 716.3 Gb Illumina PE data and 589.2 Gb Hi-C data, covering the genome 71 times, 171 times, and 141 times, respectively (Table 1), based on an estimated genome size of 4,178,143,881 bp (see below).

Genome size estimation varied only marginally between the two computational approaches presented here, with a size of 4,178,143,881 bp for the data excluding the mitochondrial genome of the spurdog as well as potential Phi X contamination. The full data generated a genome size estimate of 4,178,415,829 bp, only 271,948 bp larger than the cleaned data set. Therefore, we conclude that the genome of the spiny dogfish should be around 4.18 Gb in size.

The reported genome is rich in repetitive regions, with both estimates by GenomeScope reporting a uniqueness of 36.9%, and RepeatMasker concordantly masking over 70% of the genome. Heterozygosity was estimated to a rate of 0.632%, again in both estimates by GenomeScope.

Assembling the genome with three different assemblers gave results of varying quality and quantity (Table 2). The Canu assembly was the largest (8.4 Gb), followed by that of wtdbg2 (5.2 Gb) and Flye (3.9 Gb), making the Canu assembly more than twice than that of Flye. However, when compared to a benchmarked reference set of single-copy genes in vertebrates (BUSCO, Seppey *et al.* 2019; Manni *et al.* 2021), duplication levels were comparable between wtdbg2 and Flye (<2%), whereas Canu had a duplication level of over 50% according to BUSCO scores. This can be attributed to the different approaches taken by the assemblers: wtdbg2 and Flye usually collapse haplotypes, whereas Canu was run trying to separate the two haplotypes.

The Canu and Flye assemblies had comparable BUSCO completeness levels of over 80%, whereas the wtdbg2 output has a completeness score of <75%. Nevertheless, N50 scores (the weighted median length of the assembled sequence length) were comparable between Canu (0.3 Mb) and wtdbg2 (0.2 Mb), and clearly surpassed by the Flye assembly (1.3 Mb). The Canu assembly can be expected to contain many poorly assembled genome fragments of the alternative haplotype, degrading its apparent success. Furthermore, it contained the longest contig

Table 1. Sequencing data generated to assemble the nuclear genome of *Squalus acanthias*. Pacific Biosciences (PacBio) Continuous Long Reads (CLRs) were used in combination with Illumina paired end (PE) and Hi-C sequencing. Data characteristics were derived via a custom Python3 script (Supplementary File 2). N50 describes the weighted median length of the assembled sequence length. Genome coverage was calculated based on a genome size of 4,178,143,881 bp (see below).

Data type	Total data [b]	Sequence number	N50 [b]	Maximum sequence length [b]	Genome coverage
PacBio CLR	297,896,106,484	15,790,308	33,006	193,081	71.30
Illumina PE	716,327,706,496	4,743,892,096	151	151	171.45
Hi-C	589,216,490,262	3,902,095,962	151	151	141.02

Table 2. Characteristics of three assemblies for the nuclear genome of *Squalus acanthias*. N50 describes the weighted median length of the assembled sequence length, BUSCO scores are C(omplete) and S(ingle), C(omplete) and D(uplicated), F(ragmented) and M(issing).

Assembly	Assembly size	N50	Longest fragment	Fragment number	BUSCO scores	Per-base error rate
wtdbg2, raw	5.2 Gb	0.2 Mb	7.3 Mb	85,859	C: 74.5% [S: 72.9%, D: 1.6%], F: 10.7%, M: 14.8%	1.20%
Flye, raw	3.9 Gb	1.3 Mb	10.8 Mb	37,143	C: 85.7% [S: 83.8%, D: 1.9%], F: 6.2%, M: 8.1%	0.05%
Canu, raw	8.4 Gb	0.3 Mb	19.4 Mb	54,009	C: 83.7% [S: 33.6%, D: 50.1%], F: 7.6%, M: 8.7%	0.03%
Flye, purged	3.6 Gb	1.5 Mb	10.8 Mb	19,423	C: 85.6% [S: 83.8%, D: 1.8%], F: 6.4%, M: 8.0%	0.04%
Canu, purged	3.7 Gb	1.4 Mb	19.4 Mb	10,017	C: 82.7% [S: 81.2%, D: 1.5%], F: 8.4%, M: 8.9%	0.02%

Table 3. Characteristics of the nuclear genome assembly of *Squalus acanthias*. N50 describes the weighted median length of the assembled sequence length, BUSCO scores are C(omplete) and S(ingle), C(omplete) and D(uplicated), F(ragmented) and M(issing).

Assembly step	Assembly size	N50	Longest fragment	Fragment number	BUSCO scores	Per-base error rate
Raw	8.4 Gb	0.3 Mb	19.4 Mb	54,009	C: 83.7% [S: 33.6%, D: 50.1%], F: 7.6%, M: 8.7%	0.03%
Purged	3.7 Gb	1.4 Mb	19.4 Mb	10,017	C: 82.7% [S: 81.2%, D: 1.5%], F: 8.4%, M: 8.9%	0.02%
Scaffolded with Salsa	3.7 Gb	10.5 Mb	90.6 Mb	6,090	C: 90.7% [S: 88.3%, D: 2.4%], F: 4.0%, M: 5.3%	0.02%
Polished with long-read data	3.7 Gb	10.5 Mb	90.7 Mb	6,090	C: 91.0% [S: 88.6%, D: 2.4%], F: 3.9%, M: 5.1%	0.02%
Fully polished	3.7 Gb	10.5 Mb	90.6 Mb	6,090	C: 91.0% [S: 88.6%, D: 2.4%], F: 3.9%, M: 5.1%	< 0.02%
Contamination-free	3.7 Gb	10.7 Mb	90.6 Mb	5,672	C: 91.0% [S: 88.6%, D: 2.4%], F: 3.9%, M: 5.1%	< 0.02%
Scaffolded with YaHS	3.7 Gb	124.1 Mb	266.4 Mb	3,899	C: 91.6% [S: 89.2%, D: 2.4%], F: 3.7%, M: 4.7%	< 0.02%

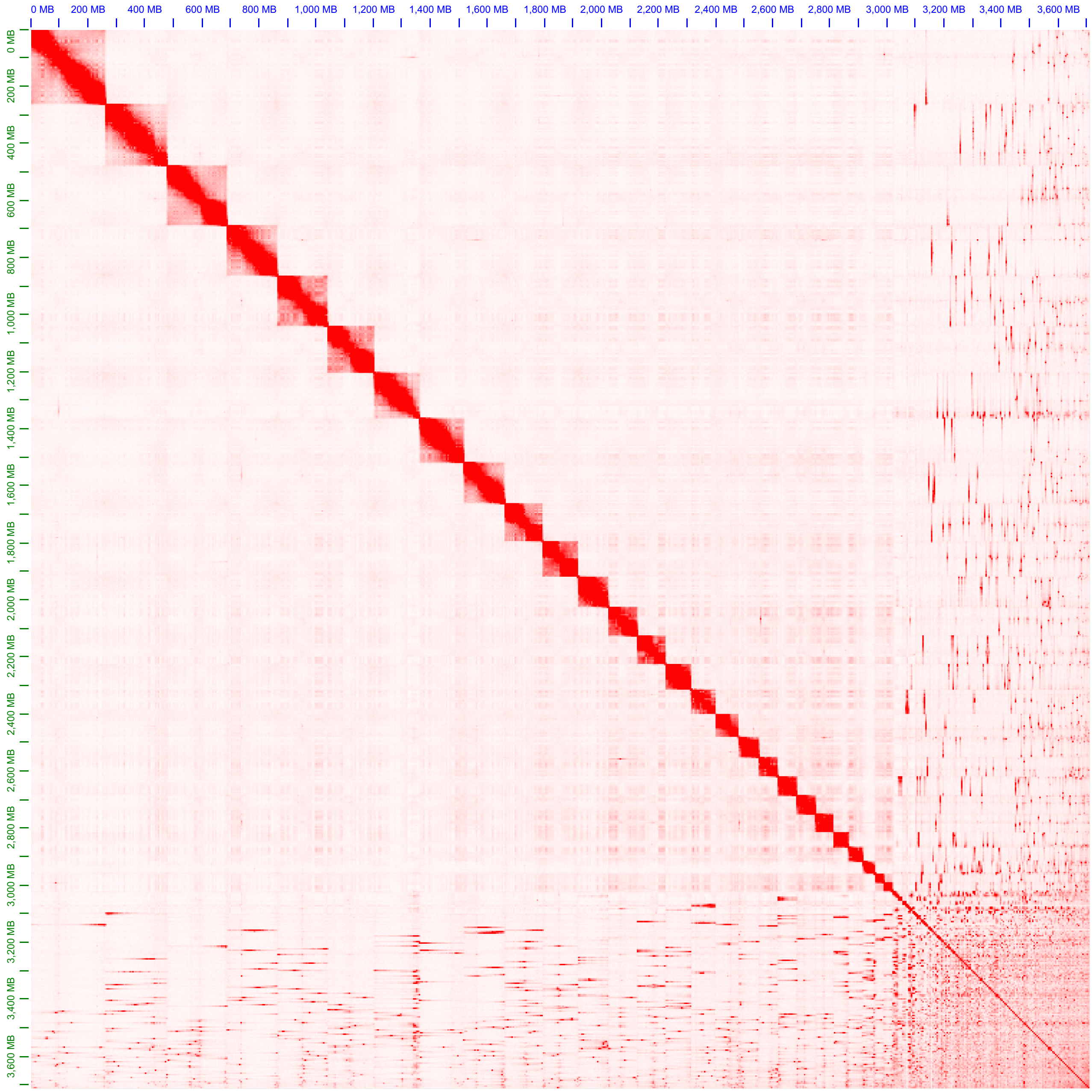

Fig. 1. Hi-C contact map for the nuclear genome assembly of *Squalus acanthias*, scaffolded with YaHS (Zhou *et al.* 2023). Total assembly length is 3.7 Gb. The map was visualized with Juicebox (v. 1.11.08, Durand *et al.* 2016a).

of all three assemblies (19.4 Mb), suggesting a rather successful assembly. Finally, per-base error rates estimated by Merqury were lowest in Canu (0.03%) and Flye (0.05%), when compared to wtdbg2 (1.20%). Therefore, considering all assembly characteristics, both the Canu- and Flye-derived assemblies were chosen for further processing.

Purging of haplotigs led to an increase in some assembly quality parameters for both assemblies, but decreased others. The error rates improved by 0.01% for both assemblies, and the N50 increased by 1.1 Mb for the Canu assembly and 0.2 Mb for the Flye assembly. Duplication levels sank below 2%, for the Flye assembly after one round and for the Canu assembly after two rounds of purging. However, the rate of complete BUSCOs decreased for both assemblies, more strongly for the Canu (1.0%) than for the Flye assembly (0.1%). In both cases, parts of this can be explained by an increase of fragmented BUSCOs, however, the Canu assembly lost true genomic information during purging, as can be seen from an increase (0.2%) of missing BUSCOs.

In total, after purging the Canu assembly had lower error and genome duplication rates than the Flye assembly but was surpassed by the Flye assembly with a higher N50 and BUSCO completeness score. As a higher quality assembly, with lower duplication levels and error rates, should benefit the scaffolding process, the Canu assembly was selected for downstream analysis.

Scaffolding of the Canu assembly with Hi-C data and Salsa (Ghurye *et al.* 2017, 2019) increased the rate of complete BUSCOs to over 90%, and the sequence N50 from 1.4 Mb to 10.5 Mb (Table 3). However, only 1.3 Gb of the assembly were contained in the 30 and 31 longest scaffolds, the expected haploid karyotype of *S. acanthias* (Nygren *et al.* 1971; Nygren and Jahnke 1972; Schwartz and Maddock 1986; but see Stingo and Rocco 2001). The scaffolded and polished assembly, cleared from 417 scaffolds containing foreign organism contamination and one scaffold completely of mitochondrial origin, was scaffolded a second time with the tool YaHS (Zhou *et al.* 2023). This time, the assembly reached an N50 of 124.1 Mb, and the 30 longest scaffolds accumulated to 3.07 Gb, 82.78% of the total assembly length (Fig. 1). Upon manual investigation of the Hi-C contact map (Fig. 1), one of the longest 30 scaffolds ("scaffold_29", length: 13,596,185 bp) appears to be part of another larger scaffold ("scaffold_20"). We thus conclude that our final assembly reached pseudo-chromosomal level, identifying 29 out of 30 to 31 putative chromosomes, but can be improved further in the future.

Our final assembly has an N50 of 124.1 Mb, and is 91.6% complete according to BUSCO scores. The error rate is 38.07 in phred score or 0.01559%, according to Merqury.

RepeatMasker was used to soft-mask 73.79% of the genome (Supplementary File 3). Based on RNA-Seq evidence, BRAKER2 predicted a total of 37,280 genes in the masked genome, with a protein BUSCO completeness score of 88.8% (Complete and single copy: 72.3%, Complete and duplicated: 16.5%, Fragmented: 5.8%, Missing: 5.4%). High duplication levels can be attributed to multiple protein sequences per gene being included in the analysis. Protein evidence from other vertebrate genomes did not lead to higher gene set completeness (data not shown). Functional annotation attached biological information to 31,979 of these genes [Supplementary File 4 (raw results BLAST+) and Supplementary File 5 (raw results InterProScan)]. Together with the gene models that received full or partial RNA-Seq support during the structural annotation process (Supplementary File 6), this resulted in 33,283 gene models with external support. Due to their external support, we considered these (Supplementary File 7) to be more reliable than the rest of the gene model set. We acknowledge that our annotation approach can only be considered as a first version, as gene numbers are around 10,000–15,000 above what might be expected following the gene numbers found in high quality genome annotation of other shark species (especially Rhie *et al.* 2021; Sayers *et al.* 2022).

Conclusion

We report the nuclear draft genome, and its annotation, of the spiny dogfish (*S. acanthias*). Together with the existing interest in this shark's biomedical characteristics, and its ecological importance, this assembled genome will facilitate further, more focused research on a variety of topics in this species. Furthermore, we expect that this resource will facilitate genomic research in other shark species, for example assisting reference-guided genome or transcriptome assemblies, or their annotation, as well as comparative genomics or phylogenomic analysis in other sharks.

Data availability

For a detailed bench protocol for high-molecular weight DNA extractions and a Python3 script for assembly statistics see Supplementary Files 1 and 2. The raw sequencing data (Hi-C, PacBio CLR, and Illumina short reads) and final assembly can be found on NCBI under BioProject PRJNA978993. Repeat content information are included in Supplementary File 3, and annotation information in Supplementary File 8; further information regarding annotation can be found in Supplementary Files 4 to 7. Supplementary Files are available on the GSA figshare: https://doi.org/10.25387/g3.23260280.

Supplemental material available at G3 online.

Acknowledgements

The authors would like to thank Aurélien Delaval, Sven Gust, and Clive Fox who supported the field work of this study, and Lars Martin Jakt for support with bioinformatics.

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

Funding

This project was funded by Nord University, Faculty of Biosciences and Aquaculture, Norway. The Norwegian Sequencing Centre, University of Oslo, Norway, which performed some of the DNA sequencing for this study, is supported by both the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the South-Eastern Norway Regional Health Authorities

Conflicts of interest statement

The author(s) declare no conflict of interest.

Author contributions

C.I.W., L.R.N., G.H., and C.S.J. designed the study, C.I.W. planned and conducted the bioinformatical work, C.I.W. and M.E.L.K. contributed to the laboratory work, J.T. provided samples, C.I.W. drafted the manuscript, C.S.J., G.H., and L.R.N. led the study, and all authors contributed to the manuscript.

Literature cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–410. doi:10.1016/S0022-2836(05)80360-2.

Babraham Bioinformatics. 2010 FastQC. https://www.bioinformatics.babraham.ac.uk/projects.

Bao Z, Eddy SR. Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res. 2002;12(8):1269–1276. doi:10.1101/gr.88502.

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. Bamtools: A C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27(12):1691–1692. doi:10.1093/bioinformatics/btr174.

Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–580. doi:10.1093/nar/27.2.573.

Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, *et al.* The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021;49(D1):D344–D354. doi:10.1093/nar/gkaa977.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–2120. doi:10.1093/bioinformatics/btu170.

Broad Institute. Picard Toolkit. https://broadinstitute.github.io/picard 2019.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 2021;3(1):lqaa108. doi:10.1093/nargab/lqaa108.

Brůna T, Lomsadze A, Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genomics Bioinforma. 2020;2(2):lqaa026. doi:10.1093/nargab/lqaa026.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60. doi:10.1038/nmeth.3176.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421. doi:10.1186/1471-2105-10-421.

Chana-Munoz A, Jendroszek A, Sønnichsen M, Kristiansen R, Jensen JK, Andreasen PA, Bendixen C, Panitz F. Multi-tissue RNA-seq and transcriptome characterisation of the spiny dogfish shark (*Squalus acanthias*) provides a molecular tool for biological research and reveals new genes involved in osmoregulation. PLoS One. 2017;12(8):e0182756. doi:10.1371/journal.pone.0182756.

Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–i890. doi:10.1093/bioinformatics/bty560.

Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563–569. doi:10.1038/nmeth.2474.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, *et al.* Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–1423. doi:10.1093/bioinformatics/btp163.

Compagno LJV. FAO species catalogue. In: Fischer W, Nauen CE, editors. Sharks of the World. An Annotated and Illustrated Catalogue of Sharks species Known to Date. Part 1. Hexanchiformes to Lamniformes. Vol. 4. Rome, Italy: FAO; 1984. p. 111–113.

Cummins C, Ahamed A, Aslam R, Burgin J, Devraj R, Edbali O, Gupta D, Harrison PW, Haseeb M, Holt S, *et al.* The European Nucleotide Archive in 2021. Nucleic Acids Res. 2022;50(D1):D106–D110. doi:10.1093/nar/gkab1051.

Dainat J. AGAT: Another Gff analysis toolkit to handle annotations in any GTF/GFF format. https://www.doi.org/10.5281/zenodo.3552717 2022.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.* Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008. doi:10.1093/gigascience/giab008.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3(1):99–101. doi:10.1016/j.cels.2015.07.012.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3(1):95–98. doi:10.1016/j.cels.2016.07.002.

Ebert DA, White WT, Goldman KJ, Compagno LJV, Daly-Engel TS, Ward RD. Resurrection and redescription of *Squalus suckleyi* (Girard, 1854) from the North Pacific, with comments on the *Squalus acanthias* subgroup (Squaliformes: Squalidae). Zootaxa. 2010;2612(1):22–40. doi:10.11646/zootaxa.2612.1.2.

Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics. 2008;9(1):18. doi:10.1186/1471-2105-9-18.

Ellis J, Soldo A, Dureuil M, Fordham S. *Squalus acanthias* (Europe assessment). The IUCN Red List of Threatened Species 2015. e.T91209505A48910866.

Ellis JR, Soldo A, Dureuil M, Fordham S. *Squalus acanthias* (Mediterranean assessment). The IUCN Red List of Threatened Species. e.T91209505A16527761 2016.

Finucci B, Cheok J, Chiaramonte GE, Cotton CF, Dulvy NK. *Squalus acanthias*. The IUCN Red List of Threatened Species. e.T91209505A124551959 2020.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. Repeatmodeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117(17):9451–9457. doi:10.1073/pnas.1921046117.

Fricke R, Eschmeyer WN, van der Laan R. Eschmeyer's Catalog of Fishes: Genera, Species, References 2023. http://researcharchive.calacademy.org/research/ich.

Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M. TSEBRA: transcript selector for BRAKER. BMC Bioinformatics. 2021;22(1):566. doi:10.1186/s12859-021-04482-0.

Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using long range contact information. BMC Genomics. 2017;18(1):527. doi:10.1186/s12864-017-3879-z.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019;15(8):e1007273. doi:10.1371/journal.pcbi.1007273.

Gotoh O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Res. 2008;36(8):2630–2638. doi:10.1093/nar/gkn105.

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

Haas BJ. TransDecoder 2018. https://github.com/TransDecoder/TransDecoder.

Hara Y, Yamaguchi K, Onimaru K, Kadota M, Koyanagi M, Keeley SD, Tatsumi K, Tanaka K, Motone F, Kageyama Y, *et al.* Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates. Nat. Ecol. Evol. 2018;2(11):1761–1771. doi:10.1038/s41559-018-0673-5.

Hardie DC, Hebert PDN. The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. Genome. 2003;46(4):683–706. doi:10.1139/g03-040.

Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32(5):767–769. doi:10.1093/bioinformatics/btv661.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. In: Kollmar M, editor. Gene Prediction. Methods in Molecular Biology. New York, NY: Humana; 2019. p. 65–95.

Iwata H, Gotoh O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res. 2012;40(20):e161. doi:10.1093/nar/gks708.

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res. 2008;36(Web Server):W5–W9. doi:10.1093/nar/gkn201.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, *et al.* Interproscan 5:

genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–1240. doi:10.1093/bioinformatics/btu031.

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–360. doi:10.1038/nmeth.3317.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–915. doi:10.1038/s41587-019-0201-4.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5): 540–546. doi:10.1038/s41587-019-0072-8.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res. 2017;27(5): 722–736. doi:10.1101/gr.215087.116.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019;47(D1):D807–D811. doi:10.1093/nar/gky1053.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr. arXiv:1303 2013.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–3100. doi:10.1093/bioinformatics/bty191.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014;42(15):e119. doi:10.1093/nar/gku557.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005;33(20):6494–6506. doi:10.1093/nar/gki937.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647–4654. doi:10.1093/molbev/msab199.

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

Marra NJ, Stanhope MJ, Jue NK, Wang M, Sun Q, Pavinski Bitar P, Richards VP, Komissarov A, Rayko M, Kliver S, *et al.* White shark genome reveals ancient elasmobranch adaptations associated with wound healing and the maintenance of genome stability. Proc Natl Acad Sci U S A. 2019;116(10):4446–4455. doi:10.1073/pnas.1819778116.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 2011;17(1):10–12. doi: 10.14806/ej.17.1.200.

Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics. 2008;24(24): 2818–2824. doi:10.1093/bioinformatics/btn548.

Moore KS, Wehrli S, Roder H, Rogers M, Forrest JN, McCrimmon D, Zasloff M. Squalamine: an aminosterol antibiotic from the shark. Proc Natl Acad Sci U S A. 1993;90(4):1354–1358. doi:10.1073/pnas.90.4.1354.

Nishihara H, Smit AFA, Okada N. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res. 2006;16(7):864–874. doi:10.1101/gr.5255506.

Nishimura O, Rozewicki J, Yamaguchi K, Tatsumi K, Ohishi Y, Ohta T, Yagura M, Niwa T, Tanegashima C, Teramura A, *et al.* Squalomix: shark and ray genome analysis consortium and its data sharing platform. F1000Res. 2022;11:1077. doi:10.12688/f1000research.123591.1.

Nygren A, Jahnke M. Microchromosomes in primitive fishes. Swedish J. Agric. Res. 1972;2:229–238.

Nygren A, Nilsson B, Jahnke M. Cytological studies in hypotremata and pleurotremata (pisces). Hereditas. 1971;67(2):275–281. doi: 10.1111/j.1601-5223.1971.tb02380.x.

Ogiwara I, Miya M, Ohshima K, Okada N. Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranchs. Mol Biol Evol. 1999;16(9):1238–1250. doi:10.1093/oxfordjournals.molbev.a026214.

Ou S, Jiang N. LTR_Retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018;176(2):1410–1422. doi:10.1104/pp.17.01310.

Ou S, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA. 2019;10(1):48. doi:10.1186/s13100-019-0193-0.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20(1):275. doi:10.1186/s13059-019-1905-y.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-Seq reads. Nat Biotechnol. 2015;33(3):290–295. doi:10.1038/nbt.3122.

Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics. 2005;21(Suppl 1):i351–i358. doi:10.1093/bioinformatics/bti1018.

Rasmussen A-S, Arnason U. Phylogenetic studies of complete mitochondrial DNA molecules place cartilaginous fishes within the tree of bony fishes. J Mol Evol. 1999;48(1):118–123. doi:10.1007/PL00006439.

Read TD, Petit RA, Joseph SJ III, Alam MT, Weil MR, Ahmad M, Bhimani R, Vuong JS, Haase CP, Webb DH, *et al.* Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: *Rhincodon typus* Smith 1828. BMC Genomics. 2017;18(1):532. doi:10.1186/s12864-017-3926-9.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, *et al.* Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592(7856):737–746. doi:10.1038/s41586-021-03451-0.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21(1):245. doi:10.1186/s13059-020-02134-9.

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

Saunders DC. Elasmobranch blood cells. Copeia. 1966;1966(2): 348–351. doi:10.2307/1441146.

Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, Comeau DC, Funk K, Kim S, Klimke W, *et al.* Database resources of the national center for biotechnology information. Nucleic Acids Res. 2021;49(D1):D10–D17. doi:10.1093/nar/gkaa892.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, *et al.* Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2022;50(D1):D20–D26. doi:10.1093/nar/gkab1112.

Schwartz FJ, Maddock MB. Comparisons of karyotypes and cellular DNA contents within and between major lines of elasmobranch. In: Uyeno T, Arai R, Tuniuchi T, Matsuura K, editors. Indo-Pacific Fish Biology. Tokyo, Japan: Ichthyological Society of Japan; 1986. p. 148–157.

Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M, editor. Gene Prediction. New York, NY: Humana; 2019. p. 227–245.

Shi J, Liang C. Generic repeat finder: a high-sensitivity tool for genome-wide *de novo* repeat detection. Plant Physiol. 2019; 180(4):1803–1815. doi:10.1104/pp.19.00386.

Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0 2015. http://www.repeatmasker.org.

Stanhope MJ, Ceres KM, Sun Q, Wang M, Zehr JD, Marra NJ, Wilder AP, Zou C, Bernard AM, Pavinski-Bitar P, *et al.* Genomes of endangered great hammerhead and shortfin mako sharks reveal historic population declines and high levels of inbreeding in great hammerhead. iScience. 2023;26(1):105815. doi:10.1016/j.isci.2022.105815.

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

Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006;7(1):62. doi:10.1186/1471-2105-7-62.

Stingo V, Rocco L. Selachian cytogenetics: a review. Genetica. 2001; 111(1/3):329–347. doi:10.1023/A:1013747215866.

Su W, Gu X, Peterson T. TIR-Learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. Mol Plant. 2019;12(3):447–460. doi:10.1016/j.molp.2019.02.008.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480–D489. doi:10.1093/nar/gkaa1100.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. Genomescope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33(14): 2202–2204. doi:10.1093/bioinformatics/btx153.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963. doi:10.1371/journal.pone.0112963.

Weber JA, Park SG, Luria V, Jeon S, Kim H-M, Jeon Y, Bhak Y, Jun JH, Kim SW, Hong WH, *et al.* The whale shark genome reveals how genomic and physiological properties scale with body size. Proc Natl Acad Sci U S A. 2020;117(34):20662–20671. doi:10.1073/pnas.1922576117.

Xiong W, He L, Lai J, Dooner HK, Du C. Helitronscanner uncovers a large overlooked cache of helitron transposons in many plant genomes. Proc Natl Acad Sci U S A. 2014;111(28):10263–10268. doi:10.1073/pnas.1410068111.

Xu Z, Wang H. LTR-FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35(Web Server):W265–W268. doi:10.1093/nar/gkm286.

Zhang Y, Gao H, Li H, Guo J, Ouyang B, Wang M, Xu Q, Wang J, Lv M, Guo X, *et al.* The white-spotted bamboo shark genome reveals chromosome rearrangements and fast-evolving immune genes of cartilaginous fish. iScience. 2020;23(11):101754. doi:10.1016/j.isci.2020.101754.

Zhang R-G, Li G-Y, Wang X-L, Dainat J, Wang Z-X, Ou S, Ma Y. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic Res. 2022;9:uhac017. doi:10.1093/hr/uhac017.

Zhou C, McCarthy SA, Durbin R. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 2023;39(1):btac808. doi:10.1093/bioinformatics/btac808.

Editor: R. Mallarino