DATA RELEASE

The genome of a giant (trevally): 
Caranx ignobilis

Brandon D. Pickett1, Jessica R. Glass2,3, Timothy P. Johnson4, Perry G. Ridge1 and John S. K. Kauwe1,5,*

1 Department of Biology, Brigham Young University, Provo, Utah, USA
2 South African Institute for Aquatic Biodiversity, Makhanda, South Africa
3 College of Fisheries and Ocean Sciences, University of Alaska Fairbanks, Fairbanks, Alaska, USA
4 Tim Johnson Gallery, Mesa, Arizona, USA
5 Brigham Young University – Hawai‘i, Laie, Hawai‘i, USA

ABSTRACT

Caranx ignobilis, commonly known as giant kingfish or giant trevally, is a large, reef-associated apex predator. It is a prized sportfish, targeted throughout its tropical and subtropical range in the Indian and Pacific Oceans. It also gained significant interest in aquaculture due to its unusual freshwater tolerance. Here, we present a draft assembly of the estimated 625.92 Mbp nuclear genome of a C. ignobilis individual from Hawaiian waters, which host a genetically distinct population. Our 97.4% BUSCO-complete assembly has a contig NG50 of 7.3 Mbp and a scaffold NG50 of 46.3 Mbp. Twenty-five of the 203 scaffolds contain 90% of the genome. We also present noisy, long-read DNA, Hi-C, and RNA-seq datasets, the latter containing eight distinct tissues and can help with annotations and studies of freshwater tolerance. Our genome assembly and its supporting data are valuable tools for ecological and comparative genomics studies of kingfishes and other carangoid fishes.

Subjects  Genetics and Genomics, Evolutionary Biology, Marine Biology

CONTEXT

The “genomic revolution” continues to rapidly advance our understanding of human evolution and the evolution of non-model organisms [1]. Comparative genomic approaches using whole-genome datasets allow for discoveries at every scale: from genome to chromosome to organism to entire clades of organisms. Genomic datasets of non-model marine teleost fishes (the most diverse clade of vertebrates) are invaluable for investigating evolutionary questions relating to adaptation, selection, genome duplication, and phylogenetic conservatism in vertebrates.

Here, we present a draft genome assembly of a marine teleost, giant trevally (Caranx ignobilis; Carangiformes: Carangoidei; Figure 1). This assembly is a valuable resource for the fields of evolutionary biology, ecology, and phylogenetics. Caranx ignobilis is a member of the Carangini clade, the most specious subclade within Carangoidei. Carangoid fishes are known for their extreme diversity in morphology and ecology [2, 3]. The giant trevally, specifically, is known to be highly tolerant of freshwater environments. This feature renders this species highly interesting for aquaculture [4–6] and makes it an ideal candidate species to investigate linkages between genotype and phenotype in the context of the freshwater adaptation of marine fishes [7, 8]. Caranx ignobilis is an apex predator in tropical and subtropical reefs and coastal environments in the Indian and Pacific
Figure 1. **Giant trevally (Caranx ignobilis) adult and juvenile.**
The quantitative morphological data for this illustration of *C. ignobilis* were obtained primarily from Smith-Vaniz (1999) [12]. These were evaluated by the artist who selected the specific values for meristic traits represented in the adult illustration, including the number of lateral-line scutes (32), the number of dorsal-fin rays (20) and spines (9), and the number of anal-fin rays (16) and spines (1*). Each value represents the centermost whole number in the corresponding range reported (within 1). The ratio of body depth to fork length, as depicted (1:2.85), is also at the center of the reported range (1:2.5–3.2). While the literature provides limited physical descriptions (such as the general shape, the color, and the presence of a posterior adipose eyelid), the artist found great benefit in some excellent photographs of live specimens caught and identified by Dr. J. S. K. Kauwe and others. A full-resolution version can be viewed at [https://www.timjohnsongallery.com/caranx-ignobilis-illustration](https://www.timjohnsongallery.com/caranx-ignobilis-illustration) (also archived at the Internet Archive ([https://web.archive.org](https://web.archive.org)) on 30th August 2022).

*Smith-Vaniz [12] and others report *C. ignobilis* as having three anal-fin spines (two anterior to and one connected to the lobe of the anal fin), and these are represented in the juvenile illustration. However, the adult is presented here without anterior anal fin spines; they are fully embedded and, therefore, not visible. Although anal-fin-spine embedment (or a corresponding change in spine count) among the adults of *C. ignobilis* is not reported by Smith-Vaniz [12], or indeed by any of the other descriptions we found, it has been reported to often happen to the detached anal-fin spines of Carangids as they grow to adulthood [13]. More importantly, the large majority of the source photos of adult *C. ignobilis* identified and provided by Dr. Kauwe clearly showed the absence of visible, unembedded anterior anal-fin spines. The illustration was rendered accordingly. It may be advisable to update current morphological data sets on *C. ignobilis* to reflect the apparently common phenomenon of anal-fin spine embedment, including the corresponding change in visible spine count. It would also be important to include this information in future published descriptions to prevent confusion or error in identifying adults of the species that would otherwise not match the reported meristic characteristic.

Oceans [9], and is heavily targeted by small-scale and recreational fisheries throughout its range. Understanding its evolutionary and ecological role in the ecosystem structure and function is important for fisheries management and the protection of reef and coral ecosystems. Importantly, new putative populations of *C. ignobilis* in the Indian and Pacific Oceans have recently been described using genomic datasets [10].

A highly-continuous genome allows for the inference of demographic history, genomic signals of selection and adaption, and comparative genomic studies with other Carangoid fishes, such as the hybridization with the closely related bluefin trevally, *Caranx melampygus* [11].

For our *C. ignobilis* assembly, we present the results derived from 58.25 Gbp of Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing data. The Illumina paired-end sequencing data were also generated with libraries for both RNA-seq and Hi-C, totaling 347.6 Gbp. Both datasets were used for scaffolding purposes and are valuable
individually. The estimated genome size is 625.92 Mbp \cite{14, 15}, of which 96.7% is covered by known bases in the primary haploid assembly. In addition to being highly contiguous, our genome assembly contains complete, unduplicated copies of >95% of the expected single-copy orthologs, suggesting the assembly is reasonably complete. This draft assembly and the supporting sequencing datasets are sufficiently high-quality to serve as valuable resources for a variety of prospective comparative and population genomics studies.

**METHODS**

An overview of the methods used in this study is provided here. Where appropriate, additional details, such as the code of custom scripts and the commands used to run software tools, are provided in a file in GigaDB \cite{16}.

**Sample acquisition and sequencing**

Blood, brain, eye, fin, gill, heart, kidney, liver, and muscle tissues from one *C. ignobilis* (NCBI:txid376895; Fishbase ID: 1985) individual were collected off the coast of O’ahu (near Kaneohe, Hawai‘i, USA) in April 2019. The blood sample was preserved in ethylenediaminetetraacetic acid (EDTA), and the other tissue samples were flash-frozen in liquid nitrogen. All samples were packaged in dry ice for transportation to Brigham Young University (BYU; Provo, Utah, USA) and stored at −80 °C until sequencing. The blood sample was used to create the Omni-C dataset. All the non-blood tissue samples were used for short-read RNA sequencing; the heart tissue was also used for long-read DNA sequencing.

DNA was prepared for long-read sequencing with a Pacific Biosciences (PacBio; Menlo Park, California, USA; https://www.pacb.com) SMRTbell Library kit, adhering to the following protocol: “Procedure & Checklist – Preparing gDNA Libraries Using the SMRTbell Express Template Preparation Kit 2.0” \cite{17}. The continuous long-read (CLR) sequencing was performed on seven SMRT cells for a 10-h movie on the PacBio Sequel at the BYU DNA Sequencing Center (DNASC; https://dnasc.byu.edu), a PacBio Certified Service Provider. The RNA libraries were prepared with Roche (Basel, Switzerland; https://sequencing.roche.com) KAPA Stranded RNA-seq kits, following recommended protocols. The paired-end sequencing was performed in High Output mode for 125 cycles, with the eight samples across two lanes, on an Illumina (San Diego, California, USA; https://www.illumina.com) Hi-Seq 2500 (RRID:SCR_016383) at the DNASC. Finally, the “Omni-C Proximity Ligation Assay Protocol” version 1.0 was followed using a Dovetail Genomics Omni-C kit to prepare the DNA for Illumina Paired-end sequencing. Adapters were obtained from Integrated DNA Technologies, and sequencing proceeded in Rapid Run mode for 250 cycles in one lane on an Illumina Hi-Seq 2500.

**Sequence assembly, duplicate purging, and scaffolding**

The PacBio CLR reads were self-corrected (Figure 2) and assembled with Canu v1.8 (RRID:SCR_015880) \cite{18}. To get a haploid representation of the genome, duplicates were purged with purge_dups v1.2.5 (RRID:SCR_021173) \cite{19}. The primary set of 329 contigs was selected for scaffolding with Omni-C data, which required reads to be mapped to the assembly before determining how to order and orient the contigs. The Omni-C reads were aligned following the Arima Genomics (San Diego, California, USA; https://arimagenomics.com) Mapping Pipeline commit #2e74ea4 (https://github.com/ArimaGenomics/mapping_pipeline), which relied on BWA-MEM2 v2.1
Figure 2. Frequency of Pacific Biosciences Read Lengths.
The read length distributions before and after correction. The dramatic shift from raw to corrected reads is evident. Reads were corrected by consensus using the correction phase of Canu v1.8.

Genome assembly validation
At each phase of the assembly process, continuity statistics (e.g., N50 and auNG [36, 37]) were calculated with caln50 commit #3e1b2be (RRID:SCR_022015) (https://github.com/lh3/caln50) and a custom Python script (Table 3). The genome size (625.92 Mbp) provided to Canu and used for the computation of the assembly statistics was based on the C-value of 0.64 from Hardie and Hebert [14], as recorded in the Animal Genome Size Database [15]. The assembly completeness was also assessed at each phase using single-copy orthologs from the Actinopterygii set of OrthoDB v10 (RRID:SCR_011980) [38], as identified by BUSCO v5.3.2 (RRID:SCR_015008) [39, 40] (Table 4). The scaffolds were visually inspected using a Hi-C contact matrix (Figure 4) created with PretextView v0.1.4 (https://github.com/wtsi-hpag/PretextView) (RRID:SCR_022024) and...
Figure 3. **Area Under the NG curve for each Assembly Step.**

The NG curve and the area under it are plotted for the contigs and scaffolds. This visually demonstrates an increase in continuity from contigs to scaffolds. Scaffolding with RNA-seq data – which has minimal effect on its own (data not shown) – further increases the scaffold-level continuity. This plot also shows that duplicate purging and fixing misassemblies slightly reduced the contig-level continuity, as expected.

**DATA VALIDATION AND QUALITY CONTROL**

**Sequencing**

CLR sequencing (PacBio) generated 3.74 M reads with a total of 58.25 Gbp, which is approximately 93× physical coverage of the genome. The mean and N50 read lengths were 15,591.278 and 27,441 bp, respectively. The longest read was 129,643 bp. The read length
In the context of scaffolding, Hi-C contact matrices show how correct the scaffolds are based on Hi-C alignment evidence. The longest 26 scaffolds are shown, ordered by descending length from top-left to bottom-right; the grey lines show the scaffold boundaries. Off-diagonal marks, especially dark and large ones, are possible evidence of mis-assembly and/or incorrect scaffolding. Regions with sharp edges similar to where the grey lines appear, but without the grey lines (e.g., three such locations occur in the top-left square), are joins between contigs in that scaffold that lack Hi-C evidence. The lack of Hi-C alignment evidence could suggest that these joins are invalid; however, evidence for these joins does exist from the RNA-seq alignments. The detection of any spurious joins would, at a minimum, require manual curation. Such curation would enable additional adjustments to fix the minor issues evidenced in the contact matrix.

---

**Figure 4.** Hi-C Contact Matrix.

In the context of scaffolding, Hi-C contact matrices show how correct the scaffolds are based on Hi-C alignment evidence. The longest 26 scaffolds are shown, ordered by descending length from top-left to bottom-right; the grey lines show the scaffold boundaries. Off-diagonal marks, especially dark and large ones, are possible evidence of mis-assembly and/or incorrect scaffolding. Regions with sharp edges similar to where the grey lines appear, but without the grey lines (e.g., three such locations occur in the top-left square), are joins between contigs in that scaffold that lack Hi-C evidence. The lack of Hi-C alignment evidence could suggest that these joins are invalid; however, evidence for these joins does exist from the RNA-seq alignments. The detection of any spurious joins would, at a minimum, require manual curation. Such curation would enable additional adjustments to fix the minor issues evidenced in the contact matrix.

distribution is plotted in Figure 2. A summary of the results for the sequencing run is available in Table 1. This genome is the second for the *Caranx* genus and ranks highly in terms of N50 among the available carangoid genomes [49, 51].

The RNA-seq from the eight tissues (i.e., brain, eye, fin, gill, heart, kidney, liver, and muscle) generated 435.99 M pairs of reads totaling 108.30 Gbp. Across all eight tissues, the mean and N50 read lengths were 124.21 and 125 bp, respectively. The combined results from all eight tissues are provided in Table 1, while the results from each tissue are available in Table 2. Omni-C sequencing generated 80.92 Gbp of data across 169.1 M read pairs. The N50 and mean read length were respectively 250 and 239.3 bp. The Omni-C results are also provided in Table 1 with the PacBio and RNA-seq data. The RNA-seq and Omni-C reads were not corrected, but the quality was assessed using fastqc [54].
Table 1. Sequencing information.

| Company | Illumina | Illumina | PacBio |
|---------|----------|----------|--------|
| Instrument | Hi-Seq 2500 | Hi-Seq 2500 | Sequel I |
| Mode | High Output | Rapid Run | NA |
| Sequencing type | PE | Omni-C, PE | SMRT, CLR |
| Duration | 125 cycles | 250 cycles | 10 h |
| Specimen | 1 | 1 | 1 |
| Tissues | Brain, Eye, Fin, Gill, Heart, Kidney, Liver, Muscle | Blood | Heart |
| Molecule | RNA | DNA | DNA |
| Millions of read (Pair)s | 435.99 | 169.11 | 3.74 |
| Mean read length (bp) | 124.21 | 239.26 | 15,591.28 |
| Read N50 (bp) | 125 | 250 | 27,441 |
| Nucleotides (Gbp) | 108.30 | 80.92 | 58.25 |

The results from each type of DNA and RNA sequencing from *Caranx ignobilis*. PE = Paired-end reads; SMRT = single-molecule real-time sequencing; CLR = continuous long-reads.

Table 2. RNA sequencing details per tissue.

| Tissue | Millions of read pairs | Mean read length | Read N50 | Nucleotides (Gbp) |
|--------|------------------------|------------------|----------|-------------------|
| Brain  | 45.59                  | 124.17           | 125      | 11.32             |
| Eye    | 52.02                  | 124.26           | 125      | 12.93             |
| Fin    | 50.13                  | 124.16           | 125      | 12.45             |
| Gill   | 55.56                  | 124.22           | 125      | 13.80             |
| Heart  | 57.87                  | 124.29           | 125      | 14.39             |
| Kidney | 58.73                  | 124.16           | 125      | 14.58             |
| Liver  | 58.25                  | 124.23           | 125      | 14.47             |
| Muscle | 57.84                  | 124.16           | 125      | 14.36             |
| All    | 435.99                 | 124.21           | 125      | 108.30            |

Results of the RNA sequencing of each tissue from one *Caranx ignobilis* individual. The eight tissues were spread across two lanes and run on an Illumina Hi-Seq 2500 in Rapid Run mode for 250 cycles to generate paired-end reads. Unless otherwise specified, lengths of nucleotide sequences are measured in base pairs (bp).

PacBio CLR error correction

The correction process reduced the number of reads from 3.74 M to 656 K and the total number of bases from 58.3 Gbp to 23.9 Gbp, for an approximate physical coverage of 38.3×. The mean and N50 read lengths changed from 15,591 and 27,441 bp to 36,475 and 40,065 bp, respectively. The longest read was 126,321 bases. The distribution of the corrected read lengths can be viewed together with the raw read lengths in Figure 2.

Genome assembly, duplicate purging, and scaffolding

The initial assembly generated by Canu comprised 1.8 K contigs for a total assembly size of 758 Mbp. That was a diploid assembly: both haplotypes were present and intermixed, separated whenever a bubble in the assembly graph prevented a single, reasonable contig. The duplicate purging to get a haploid representation of the genome (albeit with inevitable haplotype switching) and fixing misassemblies using evidence from Hi-C data yielded 343 contigs with a total assembly size of 605 Mbp. The mean contig length, N50, NG50 [55], and maximum contig length were 1.8 Mbp, 7.7 Mbp, 7.3 Mbp, and 19.6 Mbp, respectively. The L50 was 23, and the LG50 was 25. The area under the NG-curve (auNG) was 8.55 M. These values show modest reductions from the original Canu assembly (as expected), and they can be visualized through the auNG as shown in Figure 3 (also see Table 3).
Table 3. Continuity statistics.

|                          | Contigs (purge_dups) | Contigs (purge_dups + SALSA) | Scaffolds (SALSA) | Scaffolds (SALSA + Rascaf) | Scaffolds |
|--------------------------|----------------------|-------------------------------|-------------------|---------------------------|-----------|
| Sequences                | 1,804                | 329                           | 343               | 240                       | 209       | 203       |
| Known bases              | 757.523 Mbp          | 605.140 Mbp                   | 605.140 Mbp       | 605.140 Mbp               | 605.115 Mbp |
| Mean length              | 0.420 Mbp            | 1.839 Mbp                     | 1.764 Mbp         | 2.521 Mbp                 | 2.895 Mbp | 2.981 Mbp |
| Max. length              | 23.990 Mbp           | 23.990 Mbp                    | 19.607 Mbp        | 32.157 Mbp                | 89.251 Mbp | 89.251 Mbp |
| NG50                     | 7.412 Mbp            | 7.412 Mbp                     | 7.261 Mbp         | 23.385 Mbp                | 46.318 Mbp | 46.303 Mbp |
| NG90                     | 1.097 Mbp            | 0.950 Mbp                     | 0.700 Mbp         | 1.386 Mbp                 | 1.410 Mbp | 1.410 Mbp |
| LG50                     | 24                   | 24                            | 25                | 12                        | 5         | 5         |
| LG90                     | 103                  | 105                           | 114               | 39                        | 25        | 25        |
| auNG                     | 9.090 M              | 9.051 M                       | 8.549 M           | 19.716 M                  | 42.606 M  | 42.600 M  |
| Sequences with gaps      | -                    | -                             | -                 | 40                        | 35        | 35        |
| Gaps                     | -                    | -                             | -                 | 103                       | 134       | 133       |
| Unknown bases            | -                    | -                             | -                 | 51,500                    | 52,027    | 13,300    |
| Mean Gap length          | -                    | -                             | -                 | 500                       | 388,261   | 100       |

Continuity statistics for the *Caranx ignobilis* genome assembly at the contig and scaffold level. Note that the auNG value is the area under the NG curve, not the N curve. The final set of scaffolds (far right column) is the same as “Scaffolds (SALSA + Rascaf)” except that the identified contaminants were manually removed from the assembly and the gaps were unified to 100 Ns. Unless otherwise specified, all nucleotide sequences are measured in base pairs (bp).

Paired-end Illumina reads, such as those produced from Hi-C or RNA-seq libraries, can provide information to order and orient contigs into scaffolds. However, they contain insufficient information for gap-filling procedures. Accordingly, the result of the assembly statistics should increase lengths, decrease the number of sequences, and leave the number of known bases unchanged. This pattern was evident in the assembly statistics from our iterative scaffolding procedure (Table 3). It is important to note that SALSA and Rascaf introduce gaps of unknown size, using fixed runs of 500 and 17 Ns, respectively, to represent such gaps. For submission to NCBI, these gaps were converted to a fixed length of 100 Ns; the evidence for whether the joins were supported by Hi-C or RNA-seq data was submitted in an accompanying file in AGP format (https://www.ncbi.nlm.nih.gov/assembly/agp/AGPSpecification). The NCBI submission process also identified minor contaminants in some sequences, which were manually removed. The final set of scaffolds had an NG50 of 46.3 Mbp and an auNG of 42.6 M (Figure 3; Table 3). All joins were represented in a contact matrix (Figure 4), showing the Hi-C evidence for the assembly. Some joins were poorly supported by the Hi-C evidence, which was not surprising as some joins were based on RNA-seq evidence instead. Without manual curation, it is difficult to ascertain whether any individual join is spurious.

The assembly completeness, as assessed with single-copy orthologs, was also evaluated at the contig and scaffold level (Table 4). The results suggest that the modifications made to the primary contig assembly from scaffolding did not significantly impact the complete assembly of single-copy orthologs. The final set of scaffolds had 3,545 complete single-copy orthologs (97.4% of 3,640 from the OrthoDB10 Actinopterygii set). Of these, 98.4% (3,488) were present in the assembly only once, and 1.6% (57) were present more than once. Fifteen (0.4%) and 80 (2.3%) single-copy orthologs were fragmented and missing from the assembly, respectively. Approximately 16.7% of the genome was comprised of repetitive elements (Table 5), similar to other Carangoid genomes: 16.9% for *Caranx melampygus* [11], 12.8% for *Pseudocaranx georgianus* [56], and 20.3% for *Trachinotus ovatus* [49].
Comparison between the genomes of the giant trevally and other carangoids

We compared the *C. ignobilis* genome with the published genomes of other carangoids spanning the carangoid phylogeny, including the live sharksucker (*Echeneis naucrates*) [43, 44], the golden pompano (*Trachinotus ovaus*) [48, 49], the yellowtail (*Seriola quinqueradiata*) [45, 46], the longfin yellowtail (*Seriola rivoliana*) [47], the greater amberjack (*Seriola dumerili*) [57, 58], the Atlantic horse mackerel (*Trachurus trachurus*) [50–52], and the closely-related bluefin trevally (*Caranx melampygus*) [11]. We generated dot plots to visualize the genome alignments and look for general similarities between the genomes (Figure 5). Some structural variations can be seen, but overall there do not appear to be regions of significant variation (e.g., inversions or frameshifts) between *C. ignobilis* and other carangoid species. We similarly compared the same genomes by visualizing the grouping of single-copy orthologs plotted along the assemblies (Figure 6). Large groupings of orthologs consistently appear between genomes, suggesting orthology not just between genes but also between larger genomic regions. However, at this scale and by comparing several genomes at once, it is difficult to make more refined inferences on the evolution of specific orthologs within Carangoidei. Additional information could be gleaned if all genomes were assembled at the chromosome scale and the sequences were ordered based on similarity.
Figure 5. Dot Plot Comparisons with other Carangiformes (Carangoidei) Genomes.
The dot plots show the relative continuity of the various segments of two genomes. The purple dots show segments that align in the positive orientation, blue in the negative. The x-axis is the *Caranx ignobilis* genome; the y-axes of each plot are the genomes of other carangoids. Dots off the diagonal indicate the structural variation between the genome assemblies. For assemblies that did not have duplicates purged to reduce the assembly to pseudohaplotypes (*Caranx melampygus* and *Seriola* spp.), the extra dots are presumably due to the alignment to the secondary copy.
Figure 6. Single-copy Ortholog Comparisons with other Carangiformes (Carangoides) Fishes. Single-copy orthologs from the Vertebrata set of OrthoDB v9 were identified with BUSCO v3.0.6 and visualized using ChrOrthLink. “Chromosomes” (usually contigs or scaffolds) are ordered based on length. Note that the sizes of the “chromosomes” are only relative to the other “chromosomes” in the same genome and cannot be compared between genomes. Chromosome-scale assemblies are marked with an asterisk. Colors are assigned based on the *E. naucrates* chromosomes, and individual lines are drawn tracking the placement of individual single-copy orthologs through each genome. Provided there are no structural rearrangements between different species’ genomes and the genomes are all of reliable quality, large blocks of colored lines should consistently appear together on single chromosomes across the various genomes. Sections of color appearing in blocks on more than one chromosome indicate regions where either chromosome rearrangements occurred or where there were scaffolding errors.
Table 6. Database information for raw sequences.

| Specimen | Tissue | BioSample number | Sequencing type | SRA accession |
|----------|--------|------------------|-----------------|---------------|
| 1        | Blood  | SAMN16629462     | Dovetail Omni-C | SRR13036356   |
| 1        | Brain  | SAMN1651519      | Illumina RNA-seq| SRR13036383   |
| 1        | Eye    | SAMN16515620     | Illumina RNA-seq| SRR13036362   |
| 1        | Fin    | SAMN1651521      | Illumina RNA-seq| SRR13036361   |
| 1        | Gill   | SAMN1651522      | Illumina RNA-seq| SRR13036360   |
| 1        | Heart  | SAMN1651523      | Illumina RNA-seq| SRR13036359   |
| 1        | Kidney | SAMN1651524      | PacBio CLR WGA  | SRR13036357   |
| 1        | Liver  | SAMN1651525      | Illumina RNA-seq| SRR13036354   |
| 1        | Muscle | SAMN1651526      | Illumina RNA-seq| SRR13036355   |

All samples were collected from the same *Caranx ignobilis* specimen in April 2019 off the coast of O'ahu (near Kaneohe, Hawai'i, USA). They are combined under the BioProject PRJNA670456. The genome assembly is deposited in GenBank under accession JAFHLA000000000 with the “container” BioSample SAMN18021194.

Specific patterns become difficult to inspect at the genome scale when the contigs and scaffolds are small. We observed that the longest scaffolds in the *C. ignobilis* assembly have many single-copy orthologs for more than one chromosome from chromosome-scale assemblies like *E. naucrates*. This observation suggests that an investigation of the validity of some of the *C. ignobilis* scaffolding joins should be performed before inferences are drawn about those regions. The joins based on Hi-C evidence are reasonably trustworthy. However, some joins based on RNA-seq data can be spurious under certain conditions — such as when RNA-seq reads split across introns and the mapping software mistakenly assigns each end to different genes with similar sequences (e.g., from duplication events or gene families). The true structure of the genome can be further elucidated by karyotype analysis, additional sequencing data (e.g., Ultra-long Nanopore (Oxford, England, UK)), and one-on-one comparisons with high-quality, chromosome-scale assemblies from related species. Ultimately, this genomic dataset is useful for future comparative studies on genome structure and evolution within Carangiformes and, more broadly, marine teleosts.

**DATA AVAILABILITY**

Raw reads have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) [59–68] under BioProject PRJNA670456 [69] and BioSamples SAMN16516519–SAMN16516526 and SAMN16629462 [70–78]. The genome assembly is associated with the same BioProject under the “container” BioSample SAMN18021194 [79] and can be found in GenBank under the accession JAFHLA000000000. See Table 6 for a complete list of the datasets and their mapping to BioSamples. The contigs, the scaffolds resulting from Hi-C evidence, and the scaffolds resulting from Hi-C or RNA-seq evidence are also available from the Center for Open Science's (https://www.cos.io) Open Science Framework [80]. Snapshots of the code and other results files are available in the GigaDB repository [16].

**AVAILABILITY OF SUPPORTING SOURCE CODE AND REQUIREMENTS**

No significant computer programs were generated in this work. The commands used to generate these data are available as a shell script via the archived Bioinformatics Methods file in GigaDB [16]. Custom scripts referenced therein are available via GitHub.
• Project name: Caranx-ignobilis assembly-paper misc-scripts
• Project home page: https://github.com/pickettbd/caranx-ignobilis_assembly-paper_misc-scripts
• Operating system(s): Platform independent
• Programming language: Python
• License: MIT.

LIST OF ABBREVIATIONS
auNG: area under the NG-curve; BYU: Brigham Young University; CLR: continuous long-read; DNASC: DNA Sequencing Center; NCBI: National Center for Biotechnology Information; SMRT: single-molecule real-time; SRA: Sequence Read Archive.

AUTHOR CONTRIBUTIONS
JRG: Funding Acquisition; Writing - Original Draft Preparation; Writing - Review & Editing. TPJ: Visualization. JSKK: Conceptualization; Funding Acquisition; Investigation; Supervision; Resources; Writing - Review & Editing. BDP: Conceptualization; Data Curation; Formal Analysis; Funding Acquisition; Investigation; Methodology; Software; Visualization; Writing - Original Draft Preparation; Writing - Review & Editing. PGR: Funding Acquisition; Supervision; Resources; Writing - Review & Editing.

FUNDING
Illumina (United States) and Brigham Young University DNA Sequencing Center, Illumina Pilot Award, BDP; Illumina (United States) and Brigham Young University DNA Sequencing Center, Illumina Pilot Award, JRG; Illumina (United States) and Brigham Young University DNA Sequencing Center, Illumina Pilot Award, JSKK.

COMPETING INTERESTS
The authors declare no competing interests.

ETHICS APPROVAL
The authors declare that ethical approval was not required for this type of research.

ACKNOWLEDGEMENTS
We thank the Brigham Young University (BYU; Provo, Utah, USA) DNA Sequencing Center (https://dnasc.byu.edu) and Office of Research Computing (https://rc.byu.edu) for their continued support of our research. We thank Chul Lee of Seoul National University (Seoul, Republic of South Korea), and Ann Mc Cartney and Arang Rhie of the National Institutes of Health – National Human Genome Research Institute (Bethesda, Maryland, USA) for helpful discussions about ChrOrthLink and general assembly validation.

REFERENCES
1 Koonin EV, Aravind L, Kondrashov AS. The impact of comparative genomics on our understanding of evolution. Cell, 2000; 101(6): 573–576. doi:10.1016/s0092-8674(00)80867-3.
2 Price SA, Claverie T, Near TJ et al. Phylogenetic insights into the history and diversification of fishes on reefs. Coral Reefs, 2015; 34(4): 997–1009. doi:10.1007/s00338-015-1326-7.
3 Frédéric B. Marrama G. Carnevall G et al. Non-reef environments impact the diversification of extant jacks, remoras and allies (Carangoidae, Percomorpha). Proc. Royal Soc. B: Biol. Sci., 2016; 283: 20161556. doi:10.1098/rspb.2016.1556.

4 Abdussamad EM, Kassim HM, Balasubramanian TS. Distribution, biology and behaviour of the giant trevally, Caranx ignobilis - a candidate species for mariculture. Bangladesh J. Fish. Res., 2008; 12(1): 89–94.

5 Kappen DC, Kaipilly D, N.D. D. Pioneer attempt on cage culture of Giant Trevally, Caranx Ignobilis through farmer participatory approach in Thiruthipuram backwaters, Kochi, Kerala, India. Ambient Sci., 2018; 5(2): 6–8. doi:10.21276/ambi.2018.05.2a02.

6 Mutia MTM, Muyot FB, Magistrado ML et al. Induced spawning of Giant Trevally, Caranx ignobilis (Forsskål, 1775) using human Chorionic Gonadotropin (hCG) and Luteinising Hormone-releasing Hormone Analogue (LHRHa). Asian Fish. Sci., 2020; 33(2): 118–127. doi:10.33997/afs.2020.33.2.004.

7 Cossins AR, Crawford DL. Fish as models for environmental genomics. Nat. Rev. Genet., 2005; 6(4): 324–333. doi:10.1038/nrg1590.

8 Kültz D. Physiological mechanisms used by fish to cope with salinity stress. J. Exp. Biol., 2015; 218(12): 1907–1914. doi:10.1242/jeb.1081695.

9 Glass JR, Daly R, Cowley PD et al. Spatial trophic variability of a coastal apex predator, the giant trevally Caranx ignobilis, in the western Indian Ocean. Mar. Ecol. Prog. Ser., 2020; 641: 195–208.

10 Glass JR, Santos SR, Kauwe JSk et al. Phylogeography of two coastal marine predators (Caranx ignobilis and Caranx melampygus) across the Indo-Pacific. Bull. Mar. Sci., 2021; 97(2): 257–280. doi:10.5343/bms.2019.0114.

11 Pickett BD, Glass JR, Ridge PG et al. De novo genome assembly of the marine teleost, Bluefin Trevally (Caranx ignobilis). G3: Genes Genom. Genet., 2021; 11(10): jkab229. doi:10.1093/g3journal/jkab229.

12 Smith-Vaniz WF. Family carangidae. In: Carpenter KE, Niem VH (eds), FAO Species Identification Guide for Fishery Purposes: The Living Marine Resources of the Western Central Pacific, Volume 4: Bony fishes part 2 (Mugilidae to Carangidae). Rome, Italy: Food and Agricultural Organization of the United Nations, 1999; pp. 2659–2756.

13 Gunn JS. A revision of selected genera of the family Carangidae (Pisces) from Australian waters. Rec. Aust. Mus., Suppl., 1990; 12: 1–77. doi:10.3853/j.0812-7387.12.1990.92.

14 Hardie DC, Hebert PDN. Genome-size evolution in fishes. Can. J. Fish. Aquat. Sci., 2004; 61(9): 1636–1646. doi:10.1139/F04-106.

15 Gregory TR. Animal genome size database. 2018; http://www.genomeseize.com.

16 Pickett BD, Glass JR, Ridge PG et al. Supporting data for “Genome of a Giant (Trevally): Caranx ignobilis”. GigaScience Database, 2022; http://dx.doi.org/10.5524/102248.

17 Pacific Biosciences. Procedure & Checklist - Preparing gDNA Libraries Using the SMRTbell® Express Template Preparation Kit 2.0. 1 ed. 2019; https://www.pacb.com/documentation/procedure-checklist-preparing-gdna-libraries-using-the-smrtbell-express-template-preparation-kit-2-0/.

18 Koren S, Walenz BP, Berlin K et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res., 2017; 27(5): 722–736. doi:10.1101/gr.215087.116.

19 Guan D, McCarthy SA, Wood J et al. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics, 2020; 36(9): 2896–2898. doi:10.1093/bioinformatics/btaa025.

20 Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013; https://arxiv.org/abs/1303.3997.

21 Vasimuddin M, Misra S, Li H et al. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), Rio de Janeiro, Brazil, 20–24 May 2019. Institute of Electrical and Electronics Engineers (IEEE): 2019; pp. 314–324.

22 Broad Institute. Picard toolkit. GitHub. 2019; https://github.com/broadinstitute/picard.

23 Danecek P, Bonfield JK, Liddle J et al. Twelve years of SAMtools and BCFtools. Gigascience, 2021; 10(2): giab008. doi:10.1093/gigascience/giab008.

24 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 2010; 26(6): 841–842. doi:10.1093/bioinformatics/btq033.
25 Ghurye J, Rhee A, Walenz BP et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.*, 2019; 15(8): e1007273. doi:10.1371/journal.pcbi.1007273.

26 Ghurye J, Pop M, Koren S et al. Scaffolding of long read assemblies using long range contact information. *BMC Genom.*, 2017; 18(1): 1–11. doi:10.1186/s12864-017-3879-z.

27 Song L, Shankar DS, Florea L. Rascaf: improving genome assembly with RNA sequencing data. *Plant Genome*, 2016; 9(3): 1–12. doi:10.3835/plantgenome2016.03.0027.

28 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/nmeth3317.

29 Smit AFA, Hubley R, Green P. RepeatMasker. 2021; https://repeatmasker.org. Accessed 22 May 2021.

30 Camacho C, Coulouris G, Avagyan V et al. BLAST+: architecture and applications. *BMC Bioinform.*, 2009; 10: 421. doi:10.1186/1471-2105-10-421.

31 Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, 1999; 27: 573–580. doi:10.1093/nr/nar.27.2.573.

32 Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 2013; 29(19): 2487–2489. doi:10.1093/bioinformatics/btt403.

33 Storer J, Hubley R, Rosen J et al. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, 2021; 12(1): 2. doi:10.1186/s13100-020-00230-y.

34 Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 2015; 6(1): 11. doi:10.1186/s13100-015-0041-9.

35 Jurka J. Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.*, 1998; 8(3): 333–337. doi:10.1016/S0959-440X(98)80067-5.

36 Li H. auN: a new metric to measure assembly contiguity. *Heng Li Blog*, 2020; http://lih.github.io/2020/04/08/a-new-metric-on-assembly-contiguity.

37 Salzberg SL, Phillippy AM, Zimin A et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.*, 2012; 22(3): 557–567. doi:10.1101/gr.131833.111.

38 Kriventseva EV, Kuznetsov D, Tegenfeldt F et al. 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(1): D807–D811. doi:10.1093/nar/gky1053.

39 Simão FA, Waterhouse RM, Ioannidis P et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 2015; 31(19): 3210–3212. doi:10.1093/bioinformatics/btv351.

40 Manni M, Berkeley MR, Seppey M et al. 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.

41 Jain C, Koren S, Dilthey A et al. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics*, 2018; 34(17): i748–i756. doi:10.1093/bioinformatics/bty597.

42 Kriventseva EV, Tegenfeldt F, Petty TJ et al. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.*, 2015; 43(D1): D250–D256. doi:10.1093/nar/gku1220.

43 Echeneis naucrates Genome Assembly fEchNa1.1 GCF_900963305.1. 2019; https://identifiers.org/insdc.gca:GCF_900963305.1.

44 Vertebrate Genomes Project: *Echeneis naucrates*, Live Shark sucker, 2019; https://vgp.github.io/genomeark/Echeneis_naucrates. Accessed 1 February 2021.

45 Seriola quinqueradiata Genome Assembly Squ_2.0 GCA_002217815.1. 2017; https://identifiers.org/insdc.gca:GCA_002217815.1.

46 Yasuiki M, Iwasaki Y, Nishiki I et al. The yellowtail (*Seriola quinqueradiata*) genome and transcriptome atlas of the digestive tract. *DNA Res.*, 2018; 25(5): 547–560. doi:10.1093/dnares/dsy024.

47 Seriola rivoliana Genome Assembly GCA_002994505.1. 2018; https://identifiers.org/insdc.gca:GCA_002994505.1.

48 Trachinotus ovatus Genome Assembly GCA_900607315.1. 2018; https://identifiers.org/insdc.gca:GCA_900607315.1.
Zhang D-C, Guo L, Guo H-Y et al. Chromosome-level genome assembly of golden pompano (*Trachinotus ovatus*) in the family Carangidae. *Sci. Data*, 2019; 6: 216. doi:10.1038/s41597-019-0238-8.

*Trachurus trachurus* Genome Assembly fTraTra1 GCA_905171665.1. 2021; https://identifiers.org/insdc.gca:GCA_905171665.1.

Vertebrate Genomes Project: *Trachurus trachurus*, Atlantic Horse Mackerel. 2020; https://vgp.github.io/genomeark/Trachurus_trachurus. Accessed 1 February 2021.

Genner M, Rupert C. The genome sequence of the Atlantic horse mackerel, *Trachurus trachurus* (Linnaeus 1758) [version 1; peer review: 1 approved]. *Wellcome Open Res.*, 2022; 7: 118. doi:10.12688/wellcomeopenres.17813.1.

Babraham Bioinformatics Group. FASTQC: A quality control tool for high throughput sequence data. *Babraham Institute*. 2015.

Earl D, Bradnam K St, John J et al. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res.*, 2011; 21(12): 2224–2241. doi:10.1101/gr.126599.111.

Catanach A, Ruigrok M, Bowatte D et al. The genome of New Zealand trevally (Carangidae: *Pseudocaranx georgianus*) uncovers a XY sex determination locus. *BMC Genom.*, 2021; 22(1): 785. doi:10.1186/s12864-021-08102-2.

Araki K, Aoki C-Y, Kawase J et al. Whole genome sequencing of greater amberjack (*Seriola dumerili*) for SNP identification on aligned scaffolds and genome structural variation analysis using parallel resequencing. *Int. J. Genomics*, 2018; 2018: 7984292. doi:10.1155/2018/7984292.

*Seriola dumerili* Genome Assembly GCF_002260705.1. 2017; https://identifiers.org/insdc.gca:GCF_002260705.1.

SRR13036353. 2021; https://identifiers.org/ncbi/insdc.sra:SRR13036353.

SRR13036354. 2021; https://identifiers.org/ncbi/insdc.sra:SRR13036354.

SRR13036355. 2021; https://identifiers.org/ncbi/insdc.sra:SRR13036355.

SRR13036356. 2021; https://identifiers.org/ncbi/insdc.sra:SRR13036356.

SRR13036357. 2021; https://identifiers.org/ncbi/insdc.sra:SRR13036357.

SRR13036359. 2021; https://identifiers.org/ncbi/insdc.sra:SRR13036359.

SRR13036360. 2021; https://identifiers.org/ncbi/insdc.sra:SRR13036360.

SRR13036361. 2021; https://identifiers.org/ncbi/insdc.sra:SRR13036361.

SRR13036362. 2021; https://identifiers.org/ncbi/insdc.sra:SRR13036362.

SRR13036363. 2021; https://identifiers.org/ncbi/insdc.sra:SRR13036363.

PRJNA670456. 2021; https://identifiers.org/bioproject:PRJNA670456.

SAMN16629462. 2021; https://identifiers.org/biosample:SAMN16629462.

SAMN16516519. 2021; https://identifiers.org/biosample:SAMN16516519.

SAMN16516520. 2021; https://identifiers.org/biosample:SAMN16516520.

SAMN16516521. 2021; https://identifiers.org/biosample:SAMN16516521.

SAMN16516522. 2021; https://identifiers.org/biosample:SAMN16516522.

SAMN16516523. 2021; https://identifiers.org/biosample:SAMN16516523.

SAMN16516524. 2021; https://identifiers.org/biosample:SAMN16516524.

SAMN16516525. 2021; https://identifiers.org/biosample:SAMN16516525.

SAMN16516526. 2021; https://identifiers.org/biosample:SAMN16516526.

SAMN18021194. 2021; https://identifiers.org/biosample:SAMN18021194.

Pickett B. Giant Trevally Genome Assemblies. OSF. 2021; https://osf.io/v6yua.