Genome Assembly of the Roundjaw Bonefish (*Albula glossodonta*), a Vulnerable Circumtropical Sportfish

Brandon D. Pickett, Ph.D.* | Department of Biology, Brigham Young University, Provo, Utah, USA

Sheena Talma* | Department of Ichthyology and Fisheries Science, Rhodes University, Makhanda, South Africa

Jessica R. Glass, Ph.D. | South African Institute for Aquatic Biodiversity, Makhanda, South Africa

Daniel Ence, Ph.D. | School of Forest Resources and Conservation, University of Florida, Gainesville, Florida, USA

Paul D. Cowley, Ph.D. | South African Institute for Aquatic Biodiversity, Makhanda, South Africa

Perry G. Ridge, Ph.D. | Department of Biology, Brigham Young University, Provo, Utah, USA

John S. K. Kauwe III, Ph.D.† | Department of Biology, Brigham Young University, Provo, Utah, USA and Brigham Young University - Hawai‘i, Laie, Hawai‘i, USA. E-mail: kauwe@byu.edu

* These authors contributed equally to this work
† Corresponding author

ABSTRACT

Background

Bonefishes are cryptic species indiscriminately targeted by subsistence and recreational fisheries worldwide. The roundjaw bonefish, *Albula glossodonta* is the most widespread bonefish species in the Indo-Pacific and is listed as vulnerable to extinction by the IUCN's Red List due to anthropogenic activities. Whole-genome datasets allow for improved population and species delimitation, which – prior to this study – were lacking for *Albula* species.

Results

We generated a high-quality genome assembly of an *A. glossodonta* individual from Hawai'i, USA. The assembled contigs had an NG50 of 4.75 Mbp and a maximum length of 28.2 Mbp. Scaffolding yielded an NG50 of 14.49 Mbp, with the longest scaffold reaching 42.29 Mbp. Half the genome was contained in 20 scaffolds. The genome was annotated with 28.3 K protein-coding genes. We then analyzed 66 *A. glossodonta* individuals and 38,355 SNP loci to evaluate population genetic connectivity between six atolls in Seychelles and Mauritius in the Western Indian Ocean. We observed genetic homogeneity between atolls in Seychelles and evidence of reduced gene flow between Seychelles and Mauritius. The South Equatorial Current could be one mechanism limiting gene flow of *A. glossodonta* populations between Seychelles and Mauritius.

Conclusions

Quantifying the spatial population structure of widespread fishery species such as bonefishes is necessary for effective transboundary management and conservation. This population genomic dataset mapped to a high-quality genome assembly allowed us to discern shallow population structure in a widespread species in the Western Indian Ocean. The genome assembly will be useful for addressing the taxonomic uncertainties of bonefishes globally.

KEYWORDS

bonefish; *Albula glossodonta*; *Albula*; genome assembly; genome; genome annotation; population genetics; conservation

INTRODUCTION

Bonefishes (*Albula* spp.) are popular and economically important sportfishes found in the tropics around the globe. In the Florida Keys (Florida, USA) alone, $465 million of the annual economy is attributed to sportfishing tourism for bonefish and other fishery species inhabiting coastal flats [1]. Considering only bonefish, the sportfishing industry generates $169 million annually in the Bahamas [2, 3]. Unfortunately, population declines of bonefish have been observed around the globe, raising questions about how best to conserve bonefish and manage the associated fisheries [4]. *Albula* contains many morphological cryptic species, which, when combined with baseline data gaps, creates a significant hurdle to effective management [5-7].

All bonefish species were historically synonymized to a single species, *Albula vulpes* (Linnaeus 1758) [8], by 1940 [9-11], except for the threadfin bonefish, *A. nemoptera* (Fowler 1911) [12], which is morphologically distinct [12, 13]. Molecular testing in the last several decades has enabled specific distinctions that were not previously possible [6, 9, 14-16]. Presently, three species complexes (*A. argentea*, *A. nemoptera*, and *A. vulpes* complexes) contain the twelve putative albulid species, although identification remains difficult in most cases [4]. The roundjaw bonefish (Fig. 1), *A. glossodonta* (Forsskål 1775) [17], is one of seven species in the *A. vulpes* complex.

Most of the species in the *A. vulpes* complex can be found in the Caribbean Sea and Atlantic Ocean. By contrast, *A. glossodonta* can be found throughout the Indian and Pacific Oceans; this range overlaps slightly with *A. koreana* (Kwun and Kim 2011) [18] from the *A. vulpes* complex and drastically with each species in the *A. argentea* complex [4]. *Albula glossodonta* may be distinguished genetically from other species, but morphological identification based on its more-rounded jaw and larger average size is difficult for non-experts

[4, 19]. This difficulty, alongside underregulated fisheries and anthropogenic habitat loss, poses significant threats to the future of this species. In point of fact, *A. glossodonta* has been evaluated as "Vulnerable" on the International Union for the Conservation of Nature's (IUCN) Red List of Threatened Species™ [7], and several incidents of overexploitation, including regional extirpation, have been reported [20-24].

The threat to *A. glossodonta* and other bonefish species will persist unless identification is made easier and population genomics techniques are employed to understand and identify evolutionarily significant units, areas of overlap between species, presence and extent of hybridization, and life-history traits, especially migration and spawning [4]. Genetic identification has hitherto been accomplished using only a portion of the mitochondrial cytochrome b gene and some microsatellite markers [6, 9, 15, 18, 25-32], which likely provide an insufficient taxonomic history [4, 33-35]. To contribute to a more robust capacity for identification and enable more complex genomics-based analyses, we present a high-quality genome assembly of an *A. glossodonta* individual. A transcriptome assembly was also created and was used alongside computational annotation methods to create structural and functional annotations for the genome assembly. Additionally, we present results from a population genomic analysis of *A. glossodonta* populations in Seychelles and Mauritius, two island nations that support lucrative bonefish fly fishing industries. The raw data, assembly, and annotations are available on the National Center for Biotechnology Information (NCBI) website under BioProject Accessions PRJNA668352 and PRJNA702254.

METHODS

An overview of the methods used in this study is provided here. Where appropriate, additional details, such as the code for custom scripts and the commands used to run software, are provided in the Supplementary Bioinformatics Methods [see Additional File 1].

Tissue Collection and Preservation

Blood, gill, heart, and liver tissues from one *A. glossodonta* individual were collected off the coast of Molokaʻi (near Kaunakakai, Hawaiʻi, USA) in February 2016. Heart tissue from a second individual was also collected at the same location in September 2017. Tissue samples were flash-frozen in liquid nitrogen, and blood samples were preserved in EDTA. 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 from the first individual was used for short-read DNA sequencing. The gill, heart, and liver samples from the same individual were used for short-read RNA sequencing. The heart sample from the second individual was used for long-read sequencing and Hi□C sequencing.

For population genomic analyses, tissues (dorsal muscle samples or fin clips) were collected by fly fishing charter operators from 96 individuals of *A. glossodonta* from six coral atolls in the Southwest Indian Ocean (SWIO; Fig 1; Table S1 [Additional File 2]). All tissues were preserved in 95% EtOH at -20℃ until sequencing, and thereafter cataloged and preserved in -80℃ in the tissue biobank of South African Institute for Aquatic Biodiversity (Makhanda, South Africa) [36].

Sequencing

DNA Sequencing

DNA was prepared for long-read sequencing with Pacific Biosciences (PacBio; Menlo Park, California, USA) [37] SMRTbell Library kits, following the protocol "Procedure & Checklist – Preparing >30 kb SMRTbell Libraries Using Megaruptor Shearing and BluePippin Size-Selection for PacBio RS II and Sequel Systems". Continuous long-read (CLR) sequencing was performed on thirteen SMRT cells for a 10-hour movie on the PacBio Sequel at the BYU DNA Sequencing Center (DNASC) [38], a PacBio Certified Service Provider. Short-read sequencing was performed in Rapid Run mode for 250 cycles in one lane on the Illumina (San Diego, California, USA) [39] Hi-Seq 2500 at the DNASC after sonication with Covaris (Woburn, Massachusetts, USA) [40] Adaptive Focus Acoustics technology and preparation with New England Biolabs (Ipswich, Massachusetts, USA) [41] NEBNext Ultra II End Repair and Ligation kits with adapters from Integrated DNA Technologies (Coralville, Iowa, USA) [42].

mRNA Sequencing

RNA was prepared with Roche (Basel, Switzerland) [43] KAPA Stranded RNA-Seq kit, following manufacturer recommendations. Paired-end sequencing was performed in High Output mode for 125 cycles on the three samples together in one lane on the Illumina Hi-Seq 2500 at the DNASC.

Hi□C Sequencing

DNA was prepared with Phase Genomics (Seattle, Washington, USA) [44] Proximo Hi□C Kit (Animal) using the Sau3AI restriction enzyme (cut site: GATC) following recommended protocols. Paired-end sequencing was performed in Rapid Run mode for 250 cycles in one lane on the Illumina Hi-Seq 2500 at the DNASC.

ddRAD Library Preparation and Sequencing

We employed double digest restricted site-associated (ddRAD) sequencing to measure intraspecific genetic variation across six sampling localities in the SWIO. We extracted total DNA using Qiagen DNeasy Tissue kits per the manufacturer's protocol (Qiagen, Inc., Valencia, California, USA) [45]. We examined the quality of DNA extractions visually using gel electrophoresis and by quantifying isolated DNA using a Qubit fluorometer (Life Technologies, Carlsbad, California, USA) [46].

We modified a protocol developed by Peterson et al. [47] to prepare samples for ddRAD sequencing. We used the rare cutter *PstI* (5´-CTGCAG-3´ recognition site) and common cutter *MspI* (5´-CCGG-3´ recognition site). We carried out double digests of 150 – 200 ng total DNA per sample using the two enzymes in the manufacturer's supplied buffer (New England Biolabs) for 8 hours at 37℃. We randomly distributed samples from different localities across the sequencing plate to minimize bias during library preparation. We visually examined samples using gel electrophoresis to determine digestion success and then ligated barcoded Illumina adapters to DNA fragments [47]. After ligation, we pooled samples into 12 libraries and performed a clean-up using the QIAquick PCR Purification Kit. We then performed PCR using Phusion *Taq* (New England Biolabs) and Illumina indexed primers [47]. Library DNA concentration was checked using a Qubit fluorometer, followed by normalization, a second round of pooling into four libraries, and an additional QIAquick cleanup step. We then re-measured DNA concentration using a Qubit and combined equal amounts from each of the four pools into one. We analyzed this final pool using a BioAnalyzer (Agilent, Santa Clara, California, USA) [48] and performed size-selection using a Pippin Prep (Sage Science, Beverly,

Massachusetts, USA) [49], selecting for fragments between 300 – 500 bp. This was followed by a final measure of concentration using a BioAnalyzer. We sent the library to the University of Oregon Genomics and Cell Characterization Core Facility [50] where concentrations were verified via qPCR before 100 bp single-end sequencing on an Illumina Hi-Seq 4000.

Read Error Correction

Illumina DNA

Illumina whole-genome sequencing (WGS) reads were corrected using Quake v0.3.5 [51], which depended upon old versions of R (v3.4.0) [52] and the R package VGAM (v0.7-8) [53, 54]. Quake attempts to automatically choose a k□mer cutoff, traditionally based on k□mer counts provided by Jellyfish [55]. To generate q□mer counts instead of k□mer counts, BFCounter v0.2 [56] was used. Quake suggested a q□mer cutoff of 2.33, which was subsequently used by the correction phase of Quake. Unlike the WGS reads, the Illumina DNA reads created with the Hi□C library preparation were not corrected.

Illumina RNA

Illumina RNA-seq reads underwent a correction procedure using Rcorrector v1.0.2 [57]. Rcorrector automatically chooses a k□mer cutoff based on k□mer counts provided by Jellyfish [55], which Rcorrector runs automatically for the user. Alternately, Jellyfish can be run externally or bypassed with an alternate k□mer counting program, and counts can subsequently be provided to Rcorrector, which may be started at what it calls “stage 3”. We bypassed Jellyfish by using BFCounter v0.2 [56] to count k□mers. Note that Rcorrector made no changes to the reads.

PacBio CLRs

Several methods were attempted for the correction of the PacBio CLRs. The corrected reads from each method that did not fail were assembled, and the assembly results were used to choose the correction strategy. Ultimately, a hybrid correction strategy was employed. First, the reads were self-corrected using Canu v1.6 [58]. Second, the self-corrected reads were further corrected using Illumina short-reads (previously corrected with Quake) using CoLoRMap downloaded April 2018 [59].

Genome Size Estimation

Genome size was estimated using a k□mer analysis on the corrected Illumina WGS reads. First, the k□mer coverage was estimated using ntCard v1.0.1 [60]. The k□mer coverage histogram was computationally processed to calculate the area under the curve and identify the peak to determine genome size according to the following equation: $a / p = s$, where a is the area under the curve, p is the number of times the k□mers occur (the x-value) at the peak, and s is the genome size.

Genome Assembly, Polishing, and Scaffolding

Multiple assemblies were generated from various correction strategies. The final assembly was based on a hybrid correction strategy as described in the previous section “PacBio CLRs”. The assembly was created using Canu v1.6 [58]. The assembly underwent two rounds of polishing with the corrected Illumina WGS reads using RaCon v1.3.1 [61]. The polished contigs were scaffolded in a stepwise fashion using two types of long-range information: Hi-C and

RNA-seq reads. Both scaffolding steps required read mapping to the contigs before determining how to order and orient contigs. The Hi-C data alignments were performed following the Arima Genomics (San Diego, California, USA) [62] Mapping Pipeline [63], which relied on bwa v0.7.17-r1998 [64], Picard v2.19.2 [65], and SAMtools v1.6 [66]. BEDTools v2.28.0 [67] was used to prepare the Hi-C alignments for scaffolding. The RNA-seq data were aligned using HiSat v0.1.6-beta [68]. Scaffolding was performed for the Hi-C and RNA-seq data, respectively, with SALSA, downloaded 29 May 2019 [69, 70], and Rascaf, downloaded June 2018 [71]. Assembly continuity statistics, e.g., N50 and auNG [72], were calculated with caln50 downloaded 10 April 2020 [73] and a custom Python [74] script. Assembly completeness was assessed using single-copy orthologs with BUSCO v4.0.6 [75] and OrthoDB v10 [76] (Table S2 [Additional File 2]).

Transcriptome Assembly

The transcriptome was assembled from Illumina RNA-seq reads from all three tissues (i.e., gill, heart, and liver). The raw reads were used because Rcorrector did not modify any bases, thus making the raw reads and the "corrected" reads identical. The transcripts were assembled using Trinity v2.6.6 [77]. Assembly completeness was assessed using single-copy orthologs with BUSCO v4.0.6 [75] and OrthoDB v10 [76] (Table S2 [Additional File 2]).

ddRAD Sequence Assembly and Filtering

We assembled all ddRAD data using the program *ipyrad* v0.9.31 [78]. The input parameters for *ipyrad* are included in the supplementary materials (Table S3 [Additional File 2]). All *A. glossodonta* reads were mapped to the genome assembly. In step one of the *ipyrad* workflow, we demultiplexed sequences by identifying individual sample barcode sequences and

restriction overhangs. During step two, we trimmed barcodes and adapters from reads, which were then hard-masked using a *q*-score threshold of 20 and filtered for a maximum number of undetermined bases per read. In step three we clustered reads with a minimum depth of coverage of six to retain clusters in the ddRAD assembly. During step four, we jointly estimated sequencing error rate and heterozygosity from site patterns across the clustered reads assuming a maximum of two consensus alleles per individual. In step five, we determined consensus base calls for each allele using the parameters from step four and filtered for a maximum number of undetermined sites per locus. During step six, we clustered consensus sequences and aligned reads for each sample. During step seven, we filtered the data by the maximum number of alleles per locus, the maximum number of shared heterozygous sites per locus, and other criteria [78] and formatted output files for downstream analyses. We included all loci shared by at least 10 individuals.

We performed additional filtering steps after running *ipyrad* to account for missing data and rare alleles. Using VCFtools v0.1.16 [79] and BCFtools v1.6 [66], we removed individuals missing more than 98% of genotype calls. We retained only biallelic single nucleotide polymorphisms (SNPs) and removed (i) indels, (ii) loci with minor allele frequencies < 0.05 to exclude singletons and false polymorphic loci due to potential sequencing errors, (iii) alleles with a minimum count < 2, and (iv) loci with high mean depth values (> 100). We then implemented an iterative series of filtering steps based on missing data and genotype call rates to maximize genomic coverage per individual (Table S4 [Additional File 2]) [80]. Thereafter, we removed loci out of Hardy-Weinberg Equilibrium to filter for excess heterozygosity. We then used PLINK v1.9 [81] to perform linkage disequilibrium pruning by calculating the squared coefficient of

correlation (r^2) on all SNPs within a 1 kb window [82]. We removed all SNPs with an r^2 value greater than 0.6.

Computational Annotation of Assembled Genome

The MAKER v3.01.02-beta [83] pipeline was used to annotate the assembly. With minor modifications (see Supplementary Bioinformatics Methods, Additional File 1), annotation proceeded according to the process described in the most recent Maker Wiki tutorial [84]. A custom repeat library was created using RepeatModeler v1.0.11 [85]. The transcriptome assembly, genome assembly, and proteins from UniProtKB Swiss-prot [86, 87] were used as input to MAKER to create initial annotations. Gene models based on these annotations were used to train the following *ab initio* gene predictors: AUGUSTUS v3.3.2 [88, 89] and SNAP downloaded 3 June 2019 [90]. AUGUSTUS was trained using BUSCO [75] as a wrapper; SNAP was trained without a wrapper. Genemark-ES v4.38 [91-93] was also trained on the assembled genome. These models were all provided to MAKER for a second round of structural annotation. The gene models based on those annotations were filtered with gFACs v1.1.1 [94] and again provided to AUGUSTUS and SNAP. As Genemark-ES does not accept initial gene models, it did not need to be run again. The gene models from the *ab initio* gene predictors were again provided to MAKER for a third and final round of annotation. Functional annotations were added using MAKER accessory scripts, the BLAST+ Suite v2.9.0 [95, 96], and InterProScan v5.45-80.0 [97, 98]. The annotations in GFF3 format were validated with GenomeTools v1.6.1 [99] and manually curated to adhere to GenBank submission guidelines.

Statistical Analysis of Population Genomic Data

Detection of Loci under Selection

Before conducting population genomic analyses, we performed outlier tests to identify loci putatively under selection, which are generally identified by a significant difference in allele frequencies between populations [100]. Specifically, we implemented two outlier detection methods that accommodate missing data: *pcadapt* v4.1.0 [100] and BayeScan v2.1 [101]. The assumption behind *pcadapt* is that loci associated with population structure, ascertained via principal component analysis (PCA), are under selection [100]. *pcadapt* is advantageously fast and able to handle large numbers of loci. The number of principal components (*K*) was chosen based on visualization of a scree plot of the eigenvalues of a covariance matrix. Once the *K*-value was chosen, the Mahalanobis distance (*D* test statistic) was calculated using multiple linear regression of the number of SNPs versus *K* [100, 102]. To account for false discovery rates, the *p*-values generated using the Mahalanobis distance *D* were transformed to q-values using the R v3.6.3 [52] *q-value* package v2.15.0 [103] with the cut-off point (α) set to 10% (0.1).

BayeScan measures allele frequencies between different populations and identifies loci that are perceived to be undergoing natural selection based on their F_{ST} values [104, 105]. The method applies linear regression to generate population- and locus-specific F_{ST} estimates and calculates subpopulation F_{ST} coefficients by taking the difference in allele frequency between each population and the common gene pool. BayeScan incorporates uncertainties in allele frequencies due to small sample sizes, as well as imbalances in the number of samples between populations [101]. We assigned each of the six sampling localities as a population. Our analyses were based on 1:50 prior odds and included 100,000 iterations and a false discovery rate of 10%. We used the default values for the remaining parameters and visualized results in R v3.6.3 following the developer's manual [106]. After running both *pcadapt* and BayeScan, we used R

to assess the number of outliers identified by both programs and subsequently removed outlier loci to generate a neutral dataset for downstream analyses.

Population Structure and Genetic Differentiation

To examine population structure, we used a model-based clustering method to reconstruct the genetic ancestry of individuals using sparse nonnegative matrix factorization (sNMF) and least-squares optimization. Model-based analyses were performed using the package *LEA* v2.6.0 [107] in R. The sNMF function in *LEA* estimates the number of ancestral populations and the probability of the number of gene pools from which each individual originated by calculating an ancestry coefficient and investigating the model's fit through cross-entropy criterion [108]. We calculated and visualized cross-entropy scores of *K* population clusters ranging from 1–10 with 10 replicates. To complement sNMF, we also used principal component analysis (PCA), a distance-based approach based on variation in allele distributions, implemented in VCFtools v0.1.16 [79]. For sNMF and PCA analyses, no populations were assigned *a priori*. We assigned each of the six sampling localities as populations for subsequent visualization, grouped into four "island groups" based on the proximity of some of the atolls that comprised our sampling localities (Fig. 2). The five Seychelles atolls we sampled were spread amongst three separate clusters of islands that are commonly referred to as the "outer island groups" due to the geographic locations of these outlying coralline islands relative to the densely-populated, granitic "inner islands" of the Seychelles Archipelago. The island groups consisted of (i) Amirantes (St. Joseph's Atoll), (ii) Farquhar (Farquhar and Providence Atolls), (iii) Aldabra (Aldabra and Cosmoledo Atolls), as well as (iv) Mauritius (St. Brandon's Atoll; Table S1 [Additional File 2]). We computed summary statistics in R v3.6.3, including pairwise F_{ST} estimates (StAMPP v1.6.1

[109]), isolation by distance via the Mantel Rand test (adegenet v2.1.3 [110]), and expected and observed heterozygosity (hierfstat v0.5-7 [111]) to compare genetic diversity and differentiation between the four island groups.

RESULTS

Sequencing

DNA Sequencing

Paired-end, short-read sequencing (Illumina) yielded 109.5M pairs of reads comprised of 53.86Gbp. The mean and N50 read lengths were 245.981 and 250, respectively. Continuous long-read sequencing (PacBio) generated 9.5M reads with a total of 69.85Gbp. The mean and N50 read lengths were 7,352.726 and 13,831, respectively. The longest read was 103,889bp. The read length distribution is plotted in Figure 2. Result summaries for both sequencing runs are available in Table 1.

mRNA Sequencing

RNA-seq from the three tissues (i.e., gill, heart, and liver) generated 270.7M pairs of reads totaling 67.2Gbp. The gill tissue yielded 107.7M pairs of reads, with a total of 26.7Gbp. The heart tissue generated 19.6Gbp across 78.8M pairs of reads. The 84.2M pairs of reads from the liver tissue were comprised of 20.9Gbp. Across all three tissues, the mean and N50 read lengths were 124.122 and 125, respectively. The combined results from all three tissues are summarized in Table 1.

Hi□C Sequencing

Sequencing yielded 88.7M pairs of reads comprised of 44.28Gbp. The mean and N50 read lengths were 249.493 and 250, respectively. A summary of these results is presented in Table 1.

ddRAD sequencing

After data processing using *ipyrad*, we recovered a mean of 114,324 reads per individual for *A. glossodonta* and an average of 107,105 loci per individual. Following filtering for missing data, minor allele frequency, and linkage disequilibrium, the dataset contained 66 individuals and 38,355 SNPs. BayeScan, being a more conservative outlier detection method than *pcadapt*, did not identify any outliers; we thus used only outlier detection results from *pcadapt*. Subsequent removal of *pcadapt* outliers (N = 155) resulted in a neutral dataset containing 38,200 SNPs with 9% missing data.

Read Error Correction

Illumina DNA

When Quake corrects paired-end reads, three outcomes are possible for each pair of reads: (i) both reads are either already correct or correctable, (ii) one read is either correct or correctable and the other is low-quality, or (iii) both reads are low-quality. Of the original 218.96M reads (109.5M pairs of reads), Quake corrected 62.7M of them and removed 51.6M of them. 5.97M pairs of reads were discarded because both reads were rated as erroneous. 39.6M pairs of reads had one read that was correct or correctable and one read that was low-quality;

these were also discarded. The remaining 63.88M pairs of reads were either correct or correctable and were kept in the final read set containing 29.11Gbp of sequence.

Illumina RNA

No corrections were made to the RNA-seq reads by the error correction software.

PacBio CLRs

The dual-correction strategy (self-correction followed by hybrid-correction) reduced the number of reads from 9.5M to 2.79M and the total number of bases from 69.85Gbp to 36.79Gbp. The mean and N50 read lengths were changed from 7,354 and 13,831 to 13,193 and 15,483, respectively. The longest read was 63,271 bases. The distribution of read lengths can be viewed in Fig. 3.

Genome Size Estimation

The genome size was estimated to be approximately 0.933Gbp as a result of the k□mer analysis, which was consistent with the authors' expectations based on two closely related elopomorph species [112, 113].

Genome Assembly, Polishing, and Scaffolding

The initial assembly from Canu was comprised of 3.8K contigs with a total assembly size of 1.05Gbp. The mean contig length, N50, NG50, and maximum contig length were 276.2Kbp, 3.6Mbp, 4.7Mbp, and 28.2Mbp, respectively. The L50 was 57, and the LG50 was 43. The auNG was 8.17M. After two rounds of polishing these contigs with the corrected Illumina WGS reads

using RaCon, the assembly statistics changed only marginally. The number of contigs, L50, and LG50 were unchanged. The assembly size decreased by 318.7Kbp (0.03%). The mean contig length, N50, NG50, and maximum contig length were reduced by 83.8bp (0.03%), 1.3Kbp (0.04%), 1.5Kbp (0.03%), and 3.8Kbp (0.01%), respectively. The auNG decreased by 2Kbp (0.02%).

The scaffolding with the Hi-C data joined some polished contigs together, reducing the sequence count to 3.6K (-4.69%). The number of bases, excluding unknown bases (Ns), was unchanged; however, it is important to note that when SALSA creates gaps while ordering and orienting contigs, it always uses a gap size of 500bp. The result, in this case, was adding 116Kbp of Ns, which means 232 gaps were created. These gaps were spread across 113 scaffolds. No scaffold had more than six gaps (seven contigs ordered and oriented together). The mean scaffold length, scaffold N50, scaffold NG50, and maximum scaffold length increased by 13.6Kbp (4.92%), 3.8Mbp (106.25%), 5.79Mbp (121.90%), and 14.1Mbp (49.85%), respectively. Coupled with these increases were decreases of 29 (50.88%) and 22 (51.16%) in the L50 and LG50, respectively. The auNG increased to 14.1M (+72.81%). The quality of the Hi□C scaffolding can be visualized (Fig. 4) via a contact matrix generated by PretextMap [114] and PretextView [115].

The genome assembly was further improved by scaffolding with RNA-seq data. As expected, the magnitude of the changes between sets of scaffolds was smaller than what was observed between contigs and scaffolds. The total number of sequences was reduced by 176 to 3.4K (-4.69%). The number of known bases was again unchanged; however, it is important to note that when Rascaf orders and orients contigs (or other scaffolds) it always inserts a gap of 17bp to represent gaps of unknown size. Rascaf added 179 new gaps (3,043 unknown bases)

across 148 sequences. Three gaps (1,500 unknown bases) from SALSA were removed, but the rest remained unchanged. The most gaps added to a single sequence by Rascaf was five. The sequence with the most total gaps (from either source) had seven gaps (six from Hi-C), thus eight contigs were joined together.

This resulting set of scaffolds (which also includes all the contigs that were not joined to another contig in some way) had a mean length of 304.5Kbp (+5.11% from the Hi-C only value) and a maximum length of 42.29Mbp (+0.08%). The N50 and NG50 increased to 7.97Mbp (+7.04%) and 14.49Mbp (+37.58%), respectively. Decreases to 26 (-7.14%) and 20 (-4.76%) were observed for L50 and LG50, respectively. The auNG increased to 14.7M (+4.37%). Table 2 summarizes the assembly continuity statistics, and the area under the NG-curve (auNG) is visualized in Fig. 5.

The assembly completeness, as assessed with single-copy orthologs, was also evaluated at each stage (Table S2 [Additional File 2]). The results suggest that the modifications made to the primary Canu-based assembly from polishing and scaffolding did not significantly impact the correct assembly of single-copy orthologs. The final set of scaffolds had 3,481 complete single-copy orthologs (95.6% of 3,640 from the ODB10 Actinopterygii set). Of these 88.4% (3,076) were present in the assembly only once, and 11.6% (405) were present more than once. Twenty-five (0.7%) and 135 (3.7%) single-copy orthologs were fragmented in and missing from the assembly, respectively.

Transcriptome Assembly

The transcriptome assembly generated by Trinity was comprised of 455K sequences with a mean sequence length of 1,177bp. The N50 and L50 were 2.6Kbp and 56K, respectively. The

N90 and L90 were, respectively, 410bp and 270K. Of the 3,640 single-copy orthologs in the ODB10 Actinopterygii set, 86.4% (3,144) were complete; 39.5% (1,241) of which were present only once in the transcript set. 128 (3.5%) single-copy orthologs were fragmented in the transcript set, 368 (10.1%) were missing. (See Table S2 [Additional File 2])

Computational Genome Annotation

Computational structural and functional annotation yielded 28.3K protein-coding genes. Of these, 17.2K and 15.6K have annotated 5′ and 3′ UTRs, respectively. 1.8K tRNA genes were also identified. The annotations are available with the assembly on GenBank.

Population Genomic Analysis

Cross-entropy scores generated by the model-based population differentiation analysis, sNMF, provided support for a single population of *A. glossodonta* across all localities. However, individual ancestry plots generated by sNMF showed evidence of genetic differentiation in individuals from Mauritius (St. Brandon's Atoll), compared to the Seychelles sites (Fig. 6A). This differentiation was corroborated by PCA visualization of the first two principal components, where St. Brandon's Atoll individuals clustered separately from the four Seychelles island groups (Fig. 6B). Together, both population differentiation analyses indicated weak geographic population structure across all sampling localities, with reduced gene flow between St. Brandon's Atoll and the Seychelles sites.

Pairwise F_{ST} results also indicated greater genetic differentiation between St. Brandon's Atoll and all other island groups (Table 3). Estimates of observed and expected heterozygosity were similar across island groups (Table S5 [Additional File 2]), suggesting no differences in

genetic diversity between sampling localities and providing no evidence for distinguishing metapopulation processes such as inbreeding. A test of isolation by distance between sampling sites was not significant ($p = 0.1501$).

DISCUSSION

Albula glossodonta is an important fishery species in the Indo-Pacific for both subsistence and recreational purposes [20, 30, 116, 117]. Given this species' current "Vulnerable" IUCN status [7, 118] amidst recent taxonomic uncertainties [4], understanding patterns of gene flow and population structure in *A. glossodonta* is important for fisheries management [30, 119].

We observed a genetically homogenous population of *A. glossodonta* across five island atolls in the Seychelles Archipelago, with limited gene flow between Seychelles and Mauritius. Unlike highly migratory species such as eels (Anguillidae), which are close relatives of bonefishes, adult bonefishes are known for high site fidelity with relatively short migrations (~10-100 km) [117, 120, 121]. We hypothesized that adult bonefishes would not migrate between the Seychelles islands, or between the Seychelles and St. Brandon's Atoll in Mauritius, since these distances span 400–2,000 km. Consequently, the observed trend of genetic homogeneity across the Seychelles is likely not a result of adult long-distance migrations, but rather pelagic larval dispersal, the primary dispersal mechanism for bonefishes [32, 121-123]. Bonefish larvae, also referred to as leptocephali, have a long pelagic larval duration ranging from 41–72 days, which enables them to drift long distances with ocean currents [21, 124]. The estimated pelagic larval duration for *A. glossodonta* is 57 days, based on observations of individuals from French Polynesia in the South Pacific [21]. The Seychelles islands are located

in the South Equatorial Current, which flows westwards from the Indian Ocean towards the eastern coast of continental Africa, enabling larvae to be transported across the Seychelles islands, even across depths exceeding 4000 m (Fig. 2) [125, 126].

Genetic homogeneity is not always an outcome of long pelagic larval duration, as demonstrated by *Anguilla marmorata*, for which 2–5 stocks were identified in the Indo-Pacific [127, 128], and *A. glossodonta*, where putative stocks between the Indian and Pacific Oceans were suggested [119]. Indeed, we found evidence of restricted gene flow between the Seychelles sampling sites and St. Brandon's Atoll, Mauritius, which is ~1500–2000 km from the Seychelles Islands (Fig. 2). This genetic structuring was unexpected, given the long pelagic larval duration of *A. glossodonta*. However, there is evidence of limited gene flow between Seychelles and Mauritius in other marine fish species with pelagic larvae, such as *Lutjanid kasmira* [129], *Lethrinus nebulosus* [130], and *Pristipomoides filamentous* [131].

We attribute the observed genetic structure between Seychelles and St. Brandon's Atoll to the ocean currents in the southwestern Indian Ocean and their role in larval transport [132, 133]. St. Brandon's Atoll is in the direct path of one of the bifurcated arms of the South Equatorial Current as it passes through the Mascarene Plateau [125, 134]. The South Equatorial Current pushes water westward, which may create a barrier to gene flow to islands south of Seychelles such as Mauritius and Réunion [130, 131, 134]. Although there are currently no bonefish – or even elopomorph – larval dispersal models for the Indian Ocean, pelagic larval dispersal simulation models of coral species in the southwestern Indian Ocean corroborate the biogeographic break between Seychelles and Mauritius, suggesting connectivity is limited even when the pelagic larval duration is between 50–60 days [125, 134]. However, these models

considered coral larvae, which are completely reliant on currents for their dispersal [122, 134, 135]. Whilst the dispersal behavior of *A. glossodonta* larvae is unknown, we speculate that, similar to eels (Anguillidae; which also have long pelagic larval durations), bonefishes could disperse greater distances than passive corals by having the ability to swim (e.g., *Anguilla japonica* [136]) or may even take part in vertical migrations (e.g., *Anguilla japonica* [137, 138]). While officially undescribed, swimming ability in bonefish leptocephali has been observed [139], and vertical migrations have previously been theorized [122, 140].

Genome-wide datasets have enabled researchers to better-delineate population connectivity across seascapes for marine species where conventional markers (e.g., mtDNA, microsatellites) have not provided sufficient genomic resolution [127, 141, 142]. Such advances in genomic sequencing have altered our view of population connectivity in other marine fishes such as yellowfin tuna (*Thunnus albacores* [143]) and the American eel (*Anguilla rostrata* [144]). These studies, including ours, highlight the power of large genomic datasets for investigating connectivity in open-ocean environments containing few, if any, natural barriers that were traditionally thought to drive population structure. Although there has been a rapid increase in the number of studies using next-generation sequencing datasets for marine fishes, few studies to date have employed the use of genomic datasets on elopomorphs, and none on bonefish [144-146].

Conclusions

This is the first genome assembly and annotation for an albulid species, as well as the first use of a genome-wide single-nucleotide polymorphism dataset to investigate population structure for *Albula glossodonta* or any bonefish species in the Indian Ocean. Individuals of *A.*

glossodonta were genetically homogenous across four coralline island groups in the Seychelles Archipelago, but they showed evidence of genetic differentiation between the Seychelles and Mauritius (St. Brandon's Atoll). These patterns of connectivity are likely facilitated by pelagic larval dispersal, which is presumed to be strongly shaped by currents in the southwestern Indian Ocean. Only with high-resolution genomic data were we able to discern this pattern of population structure between Seychelles and Mauritius. Our dataset serves as a valuable resource for future genomic studies of bonefishes to facilitate their management and conservation.

DATA AVAILABILITY

The raw reads, genome assembly, and annotations are available under BioProject PRJNA668352 and BioSamples SAMN16516506-SAMN16516510 and SAMN17284271. The ddRAD reads are available under BioProject PRJNA702254, BioSamples SAMN18012541-SAMN18012606.

AUTHOR CONTRIBUTIONS

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

Writing - Review & Editing. **ST:** Investigation; Resources; Writing - Original Draft Preparation; Writing - Review & Editing.

ORCIDS

Paul D. Cowley, Ph.D.: 0000-0003-1246-4390

Daniel Ence, Ph.D.: 0000-0001-6099-9985

Jessica R. Glass, Ph.D.: 0000-0002-9843-1786

John S. K. Kauwe III, Ph.D.: 0000-0001-8641-2468

Brandon D. Pickett: 0000-0001-8235-4440

Perry G. Ridge, Ph.D.: 0000-0001-6944-2753

Sheena Talma: 0000-0003-2971-6523

ACKNOWLEDGEMENTS

We thank the artist, Tim Johnson [147], for creating the beautiful illustration (Fig. 1). We thank the Brigham Young University DNA Sequencing Center [38] and Office of Research Computing [148] for their continued support of our research. We thank Elizabeth M. Wallace, Clayton Ching, Josiah Ching, Derek Olthuis, Zachary Emig, Weston Gleave, and the fly fishing guides from FlyCastaway [149] and Alphonse Fishing Company [150], especially Daniel Hoenings and Matthieu Cosson, for the collection of samples in Hawai‘i and the western Indian Ocean. We are grateful to Taryn Bodill and Martinus Scheepers of the South African Institute for Aquatic Biodiversity [36] for laboratory assistance and Thomas Near of Yale University [151]

for the use of laboratory space, funding, and equipment. We also thank the Seychelles Fishing Authority [152], the Island Conservation Society [153], the Islands Development Company Ltd. [154], the Seychelles Islands Foundation [155], the Ministry of Agriculture, Climate Change and Environment [156], and Shane and Hafiza Talma for their logistical support.

FUNDING

BDP was supported by a Conservation Scholarship [157] from Fly Fishers International [158]. ST was supported by the South African Institute for Aquatic Biodiversity [36], the Mandela Rhodes Foundation [159], the Marine Research Grant [160] from the Western Indian Ocean Marine Science Association [161], and the Yale University Department of Ecology and Evolutionary Biology [162].

CONFLICT OF INTEREST

None declared.

ADDITIONAL FILES

Additional File 1: Supplementary Bioinformatics Methods (PDF).

Additional File 2: Supplementary Tables (PDF).

REFERENCES

1. Fedler T. *Economic Impact of the Florida Keys Flats Fishery*. 2013. Vero Beach, FL, USA: The Bonefish and Tarpon Trust.

2. Fedler T. *The Economic Impact of Flats Fishing in The Bahamas*. 2010. The Bahamian Flats Fishing Alliance.

3. Fedler T. *The 2018 Economic Impact of Flats Fishing in The Bahamas.* 2019. Miami, FL, USA: The Bonefish and Tarpon Trust.

4. Pickett BD, Wallace EM, Ridge PG and Kauwe JSK. Lingering Taxonomic Challenges Hinder Conservation and Management of Global Bonefishes. Fisheries. 2020; 45 7:347-58. doi:10.1002/fsh.10438.

5. Jörger KM and Schrödl M. How to describe a cryptic species? Practical challenges of molecular taxonomy. Frontiers in Zoology. 2013; 10 1:59. doi:10.1186/1742-9994-10-59.

6. Wallace EM. *Assessing Biodiversity, Evolution, and Biogeography in Bonefishes (Albuliformes): Resolving Relationships and Aiding Management*. Doctoral dissertation, University of Minnesota, St. Paul, MN, USA, 2014.

7. Adams AJ, Horodysky AZ, McBride RS, Guindon K, Shenker J, MacDonald TC, et al. Global conservation status and research needs for tarpons (Megalopidae), ladyfishes (Elopidae) and bonefishes (Albulidae). Fish and Fisheries. 2014; 15 2:280-311. doi:10.1111/faf.12017.

8. Linnaeus C. Systema Naturæ. 10 ed. Stockholm, Sweden 1758.

9. Colborn J, Crabtree RE, Shaklee JB, Pfeiler E and Bowen BW. The Evolutionary Enigma of Bonefishes (*Albula spp.*): Cryptic Species and Ancient Separations in a Globally Distributed Shorefish. Evolution. 2001; 55:807-20. doi:10.1111/j.0014-3820.2001.tb00816.x.

10. Bowen BW, Karl SA and Pfeiler E. Resolving Evolutionary Lineages and Taxonomy of Bonefishes (Albula spp.). In: Ault JS, editor. Biology and management of the world Tarpon and Bonefish fisheries. Boca Raton, FL, USA: CRC Press; 2008. p. 147-54.

11. Whitehead PJP. The Synonymy of *Albula vulpes* (Linnaeus, 1758) (Teleostei, Albulidae). Cybium. 1986; 10:211-30.

12. Fowler HW. A new albuloid fish from Santo Domingo. Proc Acad Nat Sci Philadelphia. 1911; 62:651-4.

13. Rivas LR and Warlen SM. Systematics and biology of the bonefish *Albula Nemoptera* (Fowler). Fishery Bulletin US Fish and Wildlife Services. 1967; 66 2:251-8.

14. Shaklee JB and Tamaru CS. Biochemical and Morphological Evolution of Hawaiian Bonefishes (*Albula*). Syst Zool. 1981; 30:125. doi:10.2307/2992412.

15. Seyoum S, Wallace EM and Tringali MD. PERMANENT GENETIC RESOURCES: Twelve polymorphic microsatellite markers for the bonefish, *Albula vulpes* and two congeners. Mol Eco Res. 2008; 8:354-6. doi:10.1111/j.1471-8286.2007.01954.x.

16. Wallace EM and Tringali MD. Identification of a novel member in the family Albulidae (bonefishes). J Fish Biol. 2010; 76:1972-83. doi:10.1111/j.1095-8649.2010.02639.x.

17. Forsskål P. Descriptiones Animalium: Avium, Amphibiorum, Piscium, Insectorum, Vermium. Hauniæ 1775.

18. Kwun HJ and Kim JK. A new species of bonefish, *Albula koreana* (Albuliformes: Albulidae) from Korea and Taiwan. Zootaxa. 2011; 63:57-63.

19. Donovan MK, Friedlander AM, Harding KK, Schemmel EM, Filous A, Kamikawa K, et al. Ecology and niche specialization of two bonefish species in Hawai‘i. Environmental Biology of Fishes. 2015; 98:2159-71. doi:10.1007/s10641-015-0427-z.

20. Filous A, Lennox RJ, Clua EEG and Danylchuk AJ. Fisheries selectivity and annual exploitation of the principal species harvested in a data-limited artisanal fishery at a remote atoll in French Polynesia. Ocean & Coastal Management. 2019; 178 1 August 2019:1-13. doi:10.1016/j.ocecoaman.2019.104818.

21. Filous A, Lennox RJ, Coleman RR, Friedlander AM, Clua EEG and Danylchuk AJ. Life□history characteristics of an exploited bonefish *Albula glossodonta* population in a remote South Pacific atoll. J Fish Biol. 2019; 95 2:562-74. doi:10.1111/jfb.14057.

22. Johannes RE and Yeeting B. I-Kiribati knowledge and management of Tarawa's Lagoon resources. Atoll Research Bulletin. 2000; 489:1-24. doi:10.5479/si.00775630.489.1.

23. Ram-Bidesi V. An economic assessment of destructive fishing methods in Kiribati: A case study of *te ororo* fishing in Tarawa. SPC Fisheries Newsletter. 2011; 135 May/August:21-7.

24. Ram-Bidesi V and Petaia S. *Socio-economic assessment of fishing practices by North and South Tarawa fishers in Kiribati*. 2010.

25. Pfeiler E, Colborn J, Douglas MR and Douglas ME. Systematic status of bonefishes (*Albula spp.*) from the eastern Pacific Ocean inferred from analyses of allozymes and mitochondrial DNA. Environmental Biology of Fishes. 2002; 63:151-9. doi:10.1023/A:1014263528547.

26. Pfeiler E. Resurrection of the name *Albula pacifica* (Beebe, 1942) for the shafted bonefish (Albuliformes: Albulidae) from the eastern Pacific. Rev Biol Trop. 2008; 56:839-44.

27. Pfeiler E, Bitler BG and Ulloa R. Phylogenetic Relationships of the Shafted Bonefish *Albula Nemoptera* (Albuliformes: Albulidae) from the Eastern Pacific Based on Cytochrome B Sequence Analyses. Copeia. 2006; 2006:778-84. doi:10.1643/0045-8511(2006)6[778:PROTSB]2.0.CO;2.

28. Kwun HJ, Kim JK, Doiuchi R and Nakabo T. Molecular and morphological evidence for the taxonomic status of a newly reported species of *Albula* (Albuliformes: Albulidae) from Korea and Taiwan. Animal Cells and Systems. 2011; 15:45-51. doi:10.1080/19768354.2011.555151.

29. Valdez-Moreno M, Vásquez-Yeomans L, Elías-Gutiérrez M, Ivanova NV and Hebert PDN. Using DNA barcodes to connect adults and early life stages of marine fishes from the Yucatan Peninsula, Mexico: potential in fisheries management. Marine and Freshwater Research. 2010; 61:655. doi:10.1071/MF09222.

30. Wallace EM. High intraspecific genetic connectivity in the Indo-Pacific bonefishes: implications for conservation and management. Environmental Biology of Fishes. 2015; 98:2173-86. doi:10.1007/s10641-015-0416-2.

31. Díaz-Viloria N, Sánchez-Velasco L, Perez-Enriquez R, Zárate-Villafranco A, Miller MJ and Jiménez-Rosenberg SPA. Morphological description of genetically identified Cortez bonefish (*Albula gilberti*, Pfeiler and van der Heiden 2011) leptocephali from the southern Gulf of California. Mitochondrial DNA Part A. 2017; 28:717-24. doi:10.3109/24701394.2016.1174226.

32. Wallace EM and Tringali MD. Fishery composition and evidence of population structure and hybridization in the Atlantic bonefish species complex (*Albula spp.*). Mar Biol. 2016; 163:142. doi:10.1007/s00227-016-2915-x.

33. Pamilo P and Nei M. Relationships between Gene Trees and Species Trees. Mol Biol Evol. 1988; 5 5:568-83. doi:0.1093/oxfordjournals.molbev.a040517.

34. Nichols R. Gene trees and species trees are not the same. Trends Eco Evol. 2001; 16:358-64. doi:10.1016/S0169-5347(01)02203-0.

35. Song H, Buhay JE, Whiting MF and Crandall KA. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proceedings of the National Academy of Sciences. 2008; 105 36:13486-91. doi:10.1073/pnas.0803076105.

36. The South African Insitute for Aquatic Biodiversity (SAIAB). https://www.saiab.ac.za. Accessed 1 February 2021.

37. Pacific Bioscienes. https://www.pacb.com. Accessed 1 February 2021.

38. Brigham Young University DNA Sequencing Center. https://dnasc.byu.edu. Accessed 1 February 2021.

39. Illumina. https://www.illumina.com. Accessed 1 February 2021.

40. Covaris. https://www.covaris.com. Accessed 1 February 2021.

41. New England Biolabs. https://www.neb.com. Accessed 1 February 2021.

42. Integrated DNA Technologies. https://www.idtdna.com. Accessed 1 February 2021.

43. Roche. https://sequencing.roche.com. Accessed 1 February 2021.

44. Phase Genomics. https://phasegenomics.com. Accessed 1 February 2021.

45. Qiagen. https://www.qiagen.com/. Accessed 1 February 2021.

46. Life Technologies. https://www.thermofisher.com. Accessed 1 February 2021.

47. Peterson BK, Weber JN, Kay EH, Fisher HS and Hoekstra HE. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE. 2012; 7 5:e37135. doi:10.1371/journal.pone.0037135.

48. Agilent. https://www.agilent.com. Accessed 1 February 2021.

49. Sage Science. https://sagescience.com. Accessed 1 February 2021.

50. University of Oregon Genomics and Cell Characterization Core Facility. https://gc3f.uoregon.edu. Accessed 1 February 2021.

51. Kelley DR, Schatz MC and Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 2010; 11:R116. doi:10.1186/gb-2010-11-11-r116.

52. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2021. https://www.r-project.org.

53. Yee TW and Wild CJ. Vector Generalized Additive Models. Journal of Royal Statistical Society, Series B. 1996; 58 3:481-93. doi:10.1111/j.2517-6161.1996.tb02095.x.

54. Yee TWM, Cleve. VGAM: Vector Generalized Additive Models. The Comprehensive R Archive Network. 2009; v0.7-8.

55. Marcais G and Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011; 27 6:764-70. doi:10.1093/bioinformatics/btr011.

56. Melsted P and Pritchard JK. Efficient counting of k-mers in DNA sequences using a bloom filter. BMC Bioinform. 2011; 12 333 doi:10.1186/1471-2105-12-333.

57. Song L and Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience. 2015; 4 48 doi:10.1186/s13742-015-0089-y.

58. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH and Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017; 27 5:722-36. doi:10.1101/gr.215087.116.

59. Haghshenas E, Hach F, Sahinalp SC and Chauve C. CoLoRMap: Correcting Long Reads by Mapping short reads. Bioinformatics. 2016; 32:i545-i51. doi:10.1093/bioinformatics/btw463.

60. Hamid M, Khan H and Birol I. ntCard: a streaming algorithm for the cardinality estimation of genomics data. Bioinformatics. 2017; 33 9:1324-30. doi:10.1093/bioinformatics/btw832.

61. Vaser R, Sović I, Nagarajan N and Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017; 27 5:737-46. doi:10.1101/gr.214270.116.

62. Arima Genomics. https://arimagenomics.com. Accessed 1 February 2021.

63. Arima Genomics Mapping Pipeline. https://github.com/ArimaGenomics/mapping_pipeline. Accessed 1 February 2021.

64. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013; 1303.3997.

65. Broad Institute. Picard Toolkit. Broad Institute, GitHub repository: Broad Institute, 2019. http://broadinstitute.github.io/picard.

66. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25 16:2078-9. doi:10.1093/bioinformatics/btp352.

67. Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26 6:841-2. doi:10.1093/bioinformatics/btq033.

68. Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015; 12 4:357-60. doi:10.1038/nmeth.3317.

69. Ghurye J, Pop M, Koren S, Bickhart D and Chin C-S. Scaffolding of long read assemblies using long range contact information. BMC Genomics. 2017; 18 1:1-11. doi:10.1186/s12864-017-3879-z.

70. Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, et al. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019; 15 8:e1007273. doi:0.1371/journal.pcbi.1007273.

71. Song L, Shankar DS and Florea L. Rascaf: Improving Genome Assembly with RNA Sequencing Data. Plant Genome. 2016; 9 3:1-12. doi:10.3835/plantgenome2016.03.0027.

72. Li H. auN: a new metric to measure assembly contiguity. *Heng Li's Blog*. 2020. http://lh3.github.io/2020/04/08/a-new-metric-on-assembly-contiguity.

73. Li H: calN50 GitHub Repository. https://github.com/lh3/calN50. Accessed 10 April 2020.

74. Python Programming Language. https://www.python.org. Accessed 1 February 2021.

75. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31 19:3210-2. doi:10.1093/bioinformatics/btv351.

76. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, 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 D1:D807-D11. doi:10.1093/nar/gky1053.

77. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29 7:644-52. doi:10.1038/nbt.1883.

78. Eaton DAR. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics. 2014; 30 13:1844-9. doi:10.1093/bioinformatics/btu121.

79. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, Depristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27 15:2156-8. doi:10.1093/bioinformatics/btr330.

80. O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM and Portnoy DS. These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. Mol Ecol. 2018; 27 16:3193-206. doi:10.1111/mec.14792.

81. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007; 81 3:559-75. doi:10.1086/519795.

82. Hill WG and Robertson A. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics. 1968; 38 6:226-31. doi:10.1007/BF01245622.

83. Holt C and Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinform. 2011; 12:491. doi:10.1186/1471-2105-12-491.

84. Holt C and Yandell M: MAKER Tutorial for WGS Assembly and Annotation Winter School 2018. http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_WGS_Assembly_and_Annotation_Winter_School_2018 (2018). Accessed 1 March 2018.

85. Smit AFA and Hubley R. RepeatModeler Open-1.0. 2008.

86. The Uniprot Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019; 47 D1:D506-D15. doi:10.1093/nar/gky1049.

87. Boutet E, Lieberherr D, Tognolli M, Schneider M and Bairoch A. UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. In: Edwards D, editor. Plant Bioinformatics: Methods and Protocols. Totowa, NJ: Humana Press; 2007. p. 89-112.

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

89. Stanke M and Waack S. Ggene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003; 19 Suppl. 2:ii215-ii25. doi:10.1093/bioinformatics/btg1080.

90. Korf I. Gene finding in novel genomes. BMC Bioinform. 2004; 5:59.

91. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO and Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005; 33 20:6964-506. doi:10.1093/nar/gki937.

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

93. Lomsadze A, Burns PD and 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.

94. Caballero M and Wegrzyn J. gFACs: Gene Filtering, Analysis, and Conversion to Unify Genome Annotations Across Alignment and Gene Prediction Frameworks. Genomics Proteomics Bioinformatics. 2019; 17 3:305-10. doi:10.1016/j.gpb.2019.04.002.

95. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinform. 2009; 10:421. doi:Artn 421\nDoi 10.1186/1471-2105-10-421.

96. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol. 1990; 215:403-10. doi:10.1016/S0022-2836(05)80360-2.

97. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014; 30 9:1236-40. doi:10.1093/bioinformatics/btu031.

98. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 2019; 47 D1:D351-D60. doi:10.1093/nar/gky1100.

99. Gremme G, Steinbiss S and Kurtz S. GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. IEEE/ACM Trans Comput Biol Bioinform. 2013; 10 3:645-56. doi:10.1109/TCBB.2013.68.

100. Luu K, Bazin E and Blum MGB. pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Eco Res. 2017; 17 1:67-77. doi:10.1111/1755-0998.12592.

101. Foll M and Gaggiotti O. A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. Genetics. 2008; 180 2:977-93. doi:10.1534/genetics.108.092221.

102. Martins H, Caye K, Luu K, Blum MGB and François O. Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. Mol Ecol. 2016; 25 20:5029-42. doi:10.1111/mec.13822.

103. Storey JD, Bass AJ, Dabney A and Robinson D. qvalue: Q-value estimation for false discovery rate control. The Comprehensive R Archive Network. 2017; v2.15.0.

104. Beaumont MA and Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol. 2004; 13 4:969-80. doi:10.1111/j.1365-294x.2004.02125.x.

105. Vitalis R, Dawson K and Boursot P. Interpretation of Variation Across Marker Loci as Evidence of Selection. Genetics. 2001; 158 4:1811-23. doi:10.1093/genetics/158.4.1811.

106. Foll M: BayeScan v2.1User Manual. http://cmpg.unibe.ch/software/BayeScan/files/BayeScan2.1_manual.pdf (2012). Accessed 1 February 2021.

107. Frichot E and François O. LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution. 2015; 6 8:925-9. doi:10.1111/2041-210x.12382.

108. Shryock DF, Havrilla CA, Defalco LA, Esque TC, Custer NA and Wood TE. Landscape genetic approaches to guide native plant restoration in the Mojave Desert. Ecological Applications. 2017; 27 2:429-45. doi:10.1002/eap.1447.

109. Pembleton LW, Cogan NOI and Forster JW. StAMPP: an R package for calculation of genetic differentiation and structure of mixed□ploidy level populations. Mol Eco Res. 2013; 13 5:946-52. doi:10.1111/1755-0998.12129.

110. Jombart T and Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011; 27 21:3070-1. doi:10.1093/bioinformatics/btr521.

111. Goudet J. hierfstat, a package for r to compute and test hierarchical F-statistics. Molecular Ecology Notes. 2005; 5 1:184-6. doi:10.1111/j.1471-8286.2004.00828.x.

112. Hardie DC and Hebert PDN. Genome-size evolution in fishes. Canadian Journal of Fisheries and Aquatic Sciences. 2004; 61 9:1636-46. doi:10.1139/F04-106.

113. Hinegardner RR, Donn Eric. Cellular DNA Content and the Evolution of Teleostean Fishes. The American Naturalist. 1972; 106 951:621-44.

114. High Performance Assembly Group - Wellcome Sanger Institute. PretextMap. 2020; 0.1.4.

115. High Performance Assembly Group - Wellcome Sanger Institute. PretextView. 2019; 0.0.1.

116. Kamikawa KT, Friedlander AM, Harding KK, Filous A, Donovan MK and Schemmel E. Bonefishes in Hawai‘i and the importance of angler-based data to inform fisheries management. Environmental Biology of Fishes. 2015; 98:2147-57. doi:10.1007/s10641-015-0421-5.

117. Moxham EJ, Cowley PD, Bennett RH and von Brandis RG. Movement and predation: a catch-and-release study on the acoustic tracking of bonefish in the Indian Ocean. Environmental Biology of Fishes. 2019; 102 2:365-81. doi:10.1007/s10641-019-00850-1.

118. Adams A, Guindon K, Horodysky A, MacDonald T, McBride R, Shenker J, et al. *Albula glossodonta, Shortjaw Bonefish*. Report no. T194299A2310398, 2012. The International Union for Conservation of Nature.

119. Williams CT, Mcivor AJ, Wallace EM, Lin YJ and Berumen ML. Genetic diversity and life□history traits of bonefish *Albula* spp. from the Red Sea. J Fish Biol. 2020:1-10. doi:10.1111/jfb.14638.

120. Larkin MF. *Assessment of South Florida's Bonefish Stock*. Dissertation, University of Miami, Coral Gables, Florida, USA, 2011.

121. Perez AU, Schmitter-Soto JJ, Adams AJ and Heyman WD. Connectivity mediated by seasonal bonefish (*Albula vulpes*) migration between the Caribbean Sea and a tropical estuary of Belize and Mexico. Environmental Biology of Fishes. 2019; 102 2:197-207. doi:10.1007/s10641-018-0834-z.

122. Zeng X, Adams A, Roffer M and He R. Potential connectivity among spatially distinct management zones for Bonefish (*Albula vulpes*) via larval dispersal. Environmental Biology of Fishes. 2019; 102:233-52. doi:10.1007/s10641-018-0826-z.

123. Danylchuk AJ, Cooke SJ, Goldberg TL, Suski CD, Murchie KJ, Danylchuk SE, et al. Aggregations and offshore movements as indicators of spawning activity of bonefish (*Albula vulpes*) in The Bahamas. Mar Biol. 2011; 158 9:1981-99. doi:10.1007/s00227-011-1707-6.

124. Friedlander A, Caselle JE, Beets J, Lowe CG, Bowen BW, Ogawa TK, et al. Biology and Ecology of the Recreational Bonefish Fishery at Palmyra Atoll National Wildlife Refuge with Comparisons to Other Pacific Islands. In: Ault JS, editor. Biology and management of the world Tarpon and Bonefish fisheries. Boca Raton, FL, USA: CRC Press; 2008. p. 27-56.

125. Crochelet E, Roberts J, Lagabrielle E, Obura D, Petit M and Chabanet P. A model-based assessment of reef larvae dispersal in the Western Indian Ocean reveals regional connectivity patterns — Potential implications for conservation policies. Regional Studies in Marine Science. 2016; 7 September:159-67. doi:10.1016/j.rsma.2016.06.007.

126. Badal MR, Rughooputh S, Rydberg L, Robinson IS and Pattiaratchi C. Eddy formation around South West Mascarene Plateau (Indian Ocean) as evidenced by satellite 'global ocean colour' data. Western Indian Ocean Journal of Marine Science. 2010; 8 2:139-45. doi:10.4314/wiojms.v8i2.56969.

127. Gagnaire P-A, Minegishi Y, Zenboudji S, Valade P, Aoyama J and Berrebi P. Within-population structure highlighted by differential introgression across semipermeable barriers to gene flow in *Anguilla marmorata*. Evolution. 2011; 65 12:3413-27. doi:10.1111/j.1558-5646.2011.01404.x.

128. Donovan S, Pezold F, Chen Y and Lynch B. Phylogeography of *Anguilla marmorata* (Teleostei: Anguilliformes) from the eastern Caroline Islands. Ichthyological Research. 2012; 59 1:70-6. doi:10.1007/s10228-011-0245-z.

129. Muths D, Gouws G, Mwale M, Tessier E and Bourjea J. Genetic connectivity of the reef fish *Lutjanus kasmira* at the scale of the western Indian Ocean. Canadian Journal of Fisheries and Aquatic Sciences. 2012; 69 5:842-53. doi:10.1139/f2012-012.

130. Healey AJE, Gouws G, Fennessy ST, Kuguru B, Sauer WHH, Shaw PW, et al. Genetic analysis reveals harvested Lethrinus nebulosus in the Southwest Indian Ocean comprise two cryptic species. ICES Journal of Marine Science. 2018; 75 4:1465-72. doi:10.1093/icesjms/fsx245.

131. Mzingirwa FA, Mkare TK, Nyingi DW and Njiru J. Genetic diversity and spatial population structure of a deepwater snapper, *Pristipomoides filamentosus* in the south-west Indian Ocean. Mol Biol Rep. 2019; 46 5:5079-88. doi:10.1007/s11033-019-04962-w.

132. Muths D, Grewe P, Jean C and Bourjea J. Genetic population structure of the Swordfish (*Xiphias gladius*) in the southwest Indian Ocean: Sex-biased differentiation, congruency between markers and its incidence in a way of stock assessment. Fish Res. 2009; 97 3:263-9. doi:10.1016/j.fishres.2009.03.004.

133. Obura D. The Diversity and Biogeography of Western Indian Ocean Reef-Building Corals. PLoS ONE. 2012; 7 9:e45013. doi:10.1371/journal.pone.0045013.

134. Gamoyo M, Obura D and Reason CJC. Estimating Connectivity Through Larval Dispersal in the Western Indian Ocean. Journal of Geophysical Research: Biogeosciences. 2019; 124 8:2446-59. doi:10.1029/2019JG005128.

135. Otwoma LM, Reuter H, Timm J and Meyer A. Genetic connectivity in a herbivorous coral reef fish (*Acanthurus leucosternon* Bennet, 1833) in the Eastern African region. Hydrobiologia. 2018; 806 1:237-50. doi:10.1007/s10750-017-3363-4.

136. Chang Y-LK, Miller MJ, Tsukamoto K and Miyazawa Y. Effect of larval swimming in the western North Pacific subtropical gyre on the recruitment success of the Japanese eel. PLoS ONE. 2018; 13 12:e0208704. doi:10.1371/journal.pone.0208704.

137. Kudo K. Larval vertical-migration strategy of Japanese eel. In: *MTS/IEEE Oceans 2001 An Ocean Odyssey* Honolulu, HI, USA, 5-8 November 2001 2001, pp.870-5. Escondido, CA, USA: Marine Technology Society.

138. Shinoda A, Aoyama J, Miller MJ, Otake T, Mochioka N, Watanabe S, et al. Evaluation of the larval distribution and migration of the Japanese eel in the western North Pacific. Reviews in Fish Biology and Fisheries. 2011; 21 3:591-611. doi:10.1007/s11160-010-9195-1.

139. Pfeiler E. Inshore migration, seasonal distribution and sizes of larval bonefish, *Albula*, in the Gulf of California. Environmental Biology of Fishes. 1984; 10 1/2:117-22. doi:10.1007/BF00001668.

140. Mojica RJ, Shenker JM, Harnden CW and Wagner DE. Recruitment of bonefish, *Albula vulpes*, around Lee Stocking Island, Bahamas. Fish Bull. 1994; 93 4:666-74.

141. Lemopoulos A, Prokkola JM, Uusi□Heikkilä S, Vasemägi A, Huusko A, Hyvärinen P, et al. Comparing RADseq and microsatellites for estimating genetic diversity and relatedness — Implications for brown trout conservation. Ecol Evol. 2019; 9 4:2106-20. doi:10.1002/ece3.4905.

142. Willette DA, Allendorf FW, Barber PH, Barshis DJ, Carpenter KE, Crandall ED, et al. So, you want to use next-generation sequencing in marine systems? Insight from the Pan-Pacific Advanced Studies Institute. Bull Mar Sci. 2014; 90 1:79-122. doi:10.5343/bms.2013.1008.

143. Mullins RB, McKeown NJ, Sauer WHH and Shaw PW. Genomic analysis reveals multiple mismatches between biological and management units in yellowfin tuna (*Thunnus albacares*). ICES Journal of Marine Science. 2018; 75 6:2145-52. doi:10.1093/icesjms/fsy102.

144. Babin C, Gagnaire P-A, Pavey SA and Bernatchez L. RAD-Seq Reveals Patterns of Additive Polygenic Variation Caused by Spatially-Varying Selection in the American Eel (*Anguilla rostrata*). Genome Biology and Evolution. 2017; 9 11:2974-86. doi:10.1093/gbe/evx226.

145. Benestan L, Quinn BK, Maaroufi H, Laporte M, Clark FK, Greenwood SJ, et al. Seascape genomics provides evidence for thermal adaptation and current-mediated population structure in American lobster (*Homarus americanus*). Mol Ecol. 2016; 25 20:5073-92. doi:10.1111/mec.13811.

146. Valenzuela-Quiñonez F. How fisheries management can benefit from genomics? Briefings in Functional Genomics. 2016; 15 5:352-7. doi:10.1093/bfgp/elw006.

147. Johnson T: Tim Johnson Gallery. https://timjohnsongallery.com. Accessed 8 March 2021.

148. Brigham Young University Office of Research Computing. https://rc.byu.edu. Accessed 1 February 2021.

149. FlyCastaway. https://www.flycastaway.com. Accessed 1 February 2021.

150. Alphonse Fishing Company. https://www.alphonsefishingco.com. Accessed 1 February 2021.

151. Yale University. https://www.yale.edu. Accessed 1 February 2021.

152. Seychelles Fishing Authority. http://www.sfa.sc. Accessed 1 February 2021.

153. Island Conservation Society. http://www.islandconservationseychelles.com. Accessed 1 February 2021.

154. Islands Development Company Ltd. https://www.idcseychelles.com. Accessed 1 February 2021.

155. Seychelles Islands Foundation. https://www.sif.sc. Accessed 1 February 2021.

156. Ministry of Agriculture, Climate Change and Environment. https://pcusey.sc/about-meecc. Accessed 1 February 2021.

157. Fly Fishers International Conservation Scholarship. https://flyfishersinternational.org/Conservation/Projects-Programs/Scholarship-Program. Accessed 1 February 2021.

158. Fly Fishers International. https://flyfishersinternational.org. Accessed 1 February 2021.

159. The Mandela Rhodes Foundation. https://www.mandelarhodes.org. Accessed 1 February 2021.

160. The Western Indian Ocean Marine Science Association Marine Research Grant. https://www.wiomsa.org/research-support/marg. Accessed 1 February 2021.

161. The Western Indian Ocean Marine Science Association. https://www.wiomsa.org. Accessed 1 February 2021.

162. Yale University Department of Ecology and Evolutionary Biology. https://eeb.yale.edu. Accessed 1 February 2021.

163. Hidaka K, Iwatsuki Y and Randall JE. A review of the Indo-Pacific bonefishes of the *Albula argentea* complex, with a description of a new species. Ichthyological Research. 2008; 55:53-64. doi:10.1007/s10228-007-0010-5.

FIGURE TITLES & CAPTIONS

Figure 1. Roundjaw Bonefish (*Albula glossodonta*) adult. Quantitative morphological data for this illustration of *A. glossodonta* were obtained primarily from two articles: Hidaka et al. 2008 [163] and Shaklee and Tamaru 1981 [14]. These were then evaluated by the artist, with assistance and input from the authors, to select specific values for details such as the number of pored lateral line scales (76) and the number of rays in the pectoral (18), dorsal (16), pelvic (10), and anal fins (9). Each of these was portrayed in the illustration to be at or near the middle of the ranges reported in the aforementioned articles. While some limited information was found in the literature describing coloration and general shape, the artist found particular benefit in some excellent, detailed photographs by Derek Olthuis of samples that were both personally caught in Hawai‘i and later genetically identified as *A. glossodonta* by Dr. J. S. K. Kauwe. Illustration Copyright: Tim Johnson, used with permission.

Figure 2. Sampling localities for *A. glossodonta* population genomic analysis. The upper panel shows the marine boundaries for the Seychelles and Mauritius in light blue. Locations of sampling sites are indicated by dark blue ovals. The lower panel shows the atolls comprising the four island groups: Amirantes, Farquhar, Aldabra, and Mauritius.

Figure 3. Frequency of Pacific Biosciences Read Lengths. The change in read length distribution is demonstrated as reads are corrected. The dramatic shift from raw to corrected reads is evident. The self-corrected (purple) data points are slightly larger than the dual-corrected (yellow) data points to make the purple distribution visible, the size has no meaning.

Figure 4. Hi-C Contact Matrix showing Scaffolding Correctness. In the context of scaffolding, Hi-C contact matrices show how correct the scaffolds are. Off-diagonal marks, especially those that are bright and large, are evidence of mis-assembly and/or incorrect scaffolding. The interpretations of the lighter and smaller off-diagonal marks in this image are ambiguous because the assembly is unphased with some relatively short contigs/scaffolds. Additional detail may be viewed by zooming in on the high-resolution image.

Figure 5. Area Under the N-curve (auNG) for each Assembly Step. The N-curve and the area under it are plotted for each major step of the assembly: contigs, polished contigs, scaffolds from only Hi-C data, and scaffolds from both Hi-C and RNA-seq data. The auNG for the polished contigs (green) is very similar to the contigs (yellow). Most of the curve was completely blocked by the contigs (yellow) curve. To show that the polished contigs (green) share nearly the same curve, the line was plotted more thickly so it can just barely be seen. Similarly, the Hi-C + RNA-seq scaffolds (purple) curve is very similar to the Hi-C only scaffolds (blue) curve. In this case, differences are more apparent. In certain places, e.g., at the highest peak, the Hi-C + RNA-seq scaffolds (purple) are plotted more thickly so it can be seen behind the Hi-C only scaffolds (blue).

Figure 6. Population Differentiation Analyses. Weak geographic population structure is present across all sampling localities, with reduced gene flow between St. Brandon's Atoll and the Seychelles sites. Island groups are colored as in Fig. 2. **(A)** Individual ancestry plots generated using sNMF, indicating $K = 2$ putative populations. **(B)** Principal component analysis biplot showing the first two principal components.

TABLES

Table 1. Sequencing Information

Company	Illumina	Illumina	Illumina	PacBio
Instrument	Hi-Seq 2500	Hi-Seq 2500	Hi-Seq 2500	Sequel I
Mode	Rapid Run	High Output	Rapid Run	NA
Sequencing Type	PE	PE	Hi-C, PE	SMRT, CLR
Duration	250 cycles	125 cycles	250 cycles	30 hours
Specimen	1	1	2	2
Tissues	Blood	Gill, Heart, Liver	Heart	Heart
Molecule	DNA	RNA	DNA	DNA
Millions of Read(Pair)s	109.5	270.7	88.7	9.5
Mean Read Length (bp)	246	124	245	7,353
Read N50 (bp)	250	125	250	13,831
Nucleotides (Gbp)	53.9	67.2	44.3	69.9

The results from each type of DNA and RNA sequencing from *Albula glossodonta*. PE=Paired-end reads. SMRT=Single-Molecule, Real-Time sequencing. CLR=Continuous Long-reads.

Table 2. Continuity Statistics

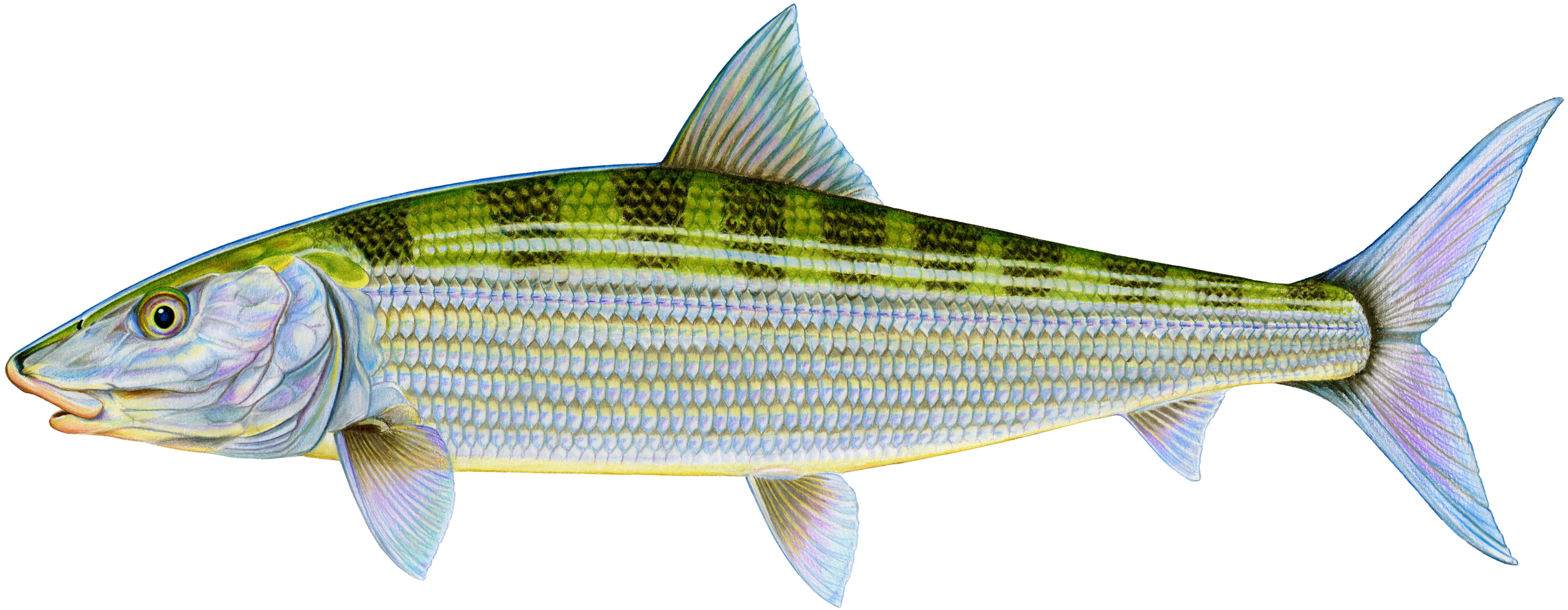
	Contigs	Polished Contigs	Scaffolds (Hi□C)	Scaffolds (Hi□C + RNA□seq)
Sequences	3,799	3,799	3,621	3,445
Known Bases	1.04935 Gbp	1.04903 Gbp	1.04903 Gbp	1.04903 Gbp
Mean Length	276,217.073	276,133.196	289,707.267	304,507.986
Max. Length	28,203,290	28,199,443	42,256,846	42,290,388
NG50	4,747,926	4,746,442	10,532,420	14,490,288
NG90	503,090	503,135	739,806	827,489
LG50	43	43	21	20
LG90	289	289	181	162
auNG	8,165,188	8,163,173	14,106,761	14,723,001
Sequences with Gaps	-	-	133	236
Gaps	-	-	232	408
Unknown Bases	-	-	116,000	117,543
Mean Gap Length	-	-	500.000	288.096

Continuity statistics for the *Albula glossodonta* genome assembly at various stages. Also note that when submitted to GenBank, the gaps were all converted to a length of 100 bp. Unless otherwise specified, all nucleotide sequences are measured in base pairs (bp).

Table 3. Pairwise F_{ST} comparisons by island group

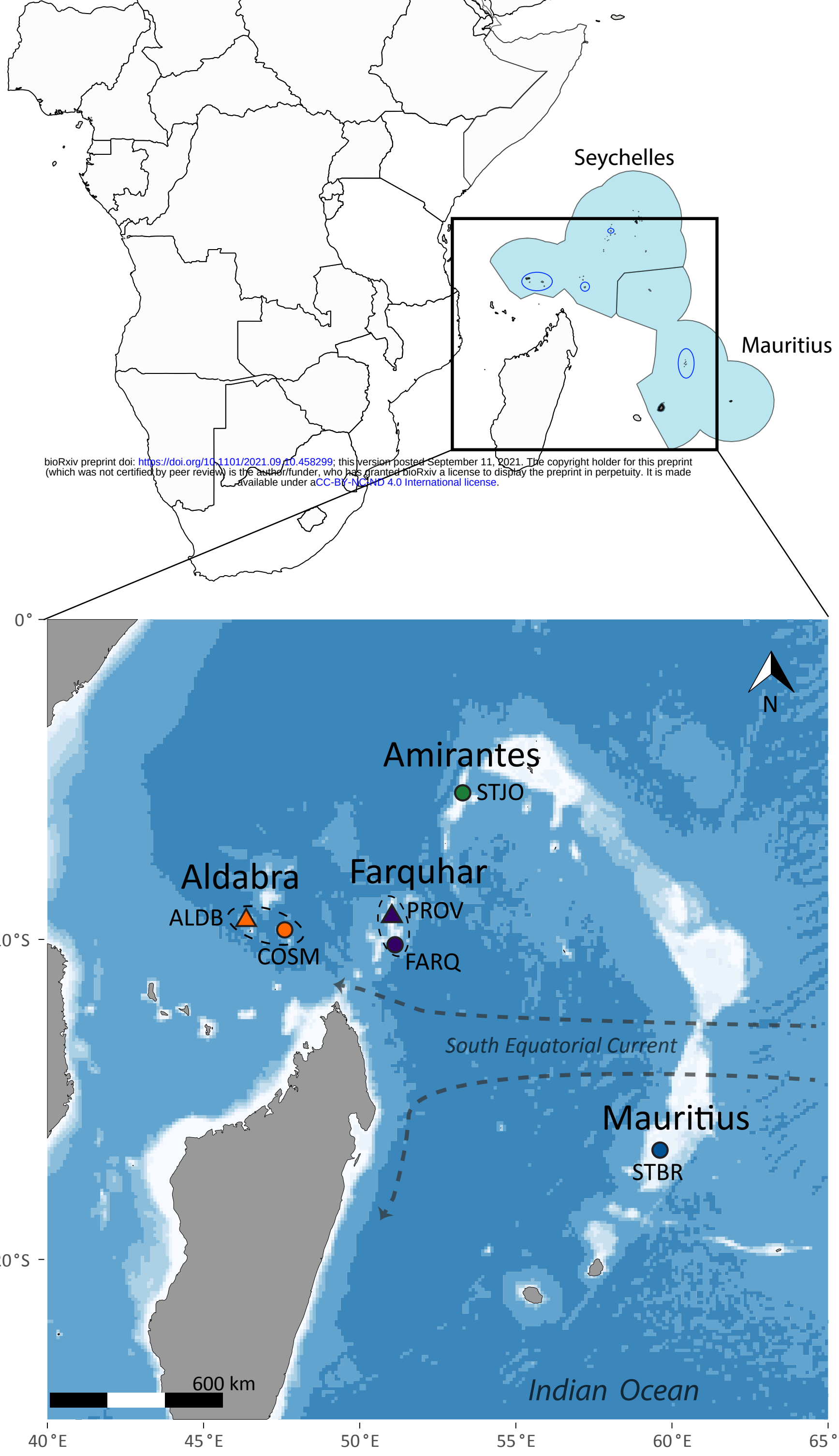
	Amirantes	Farquhar	Aldabra
Farquhar	0.0014*		
Aldabra	0.0005	0.0020*	
Mauritius	0.0034*	0.0043*	0.0040*

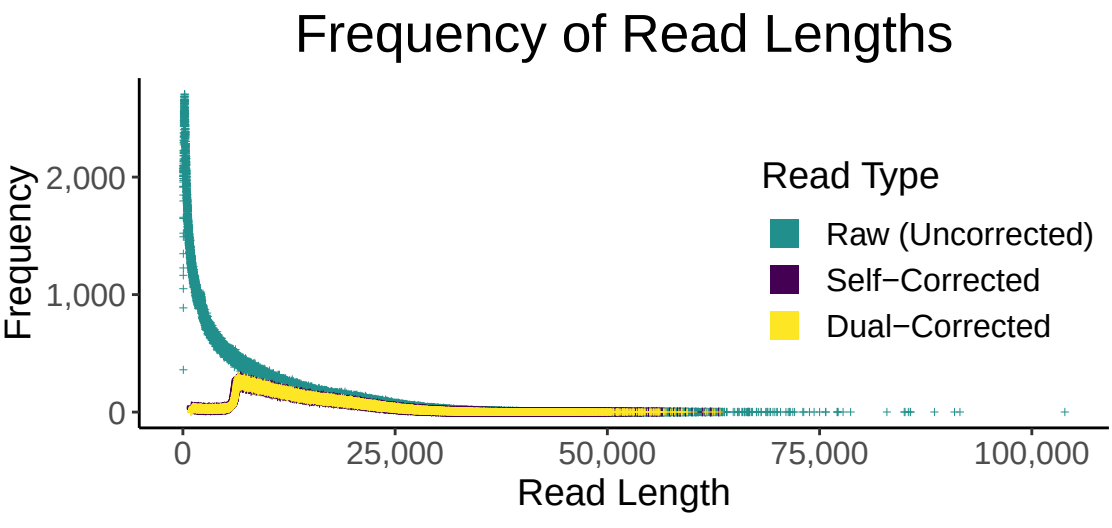
Significance ($p < 0.05$) is indicated with an asterisk

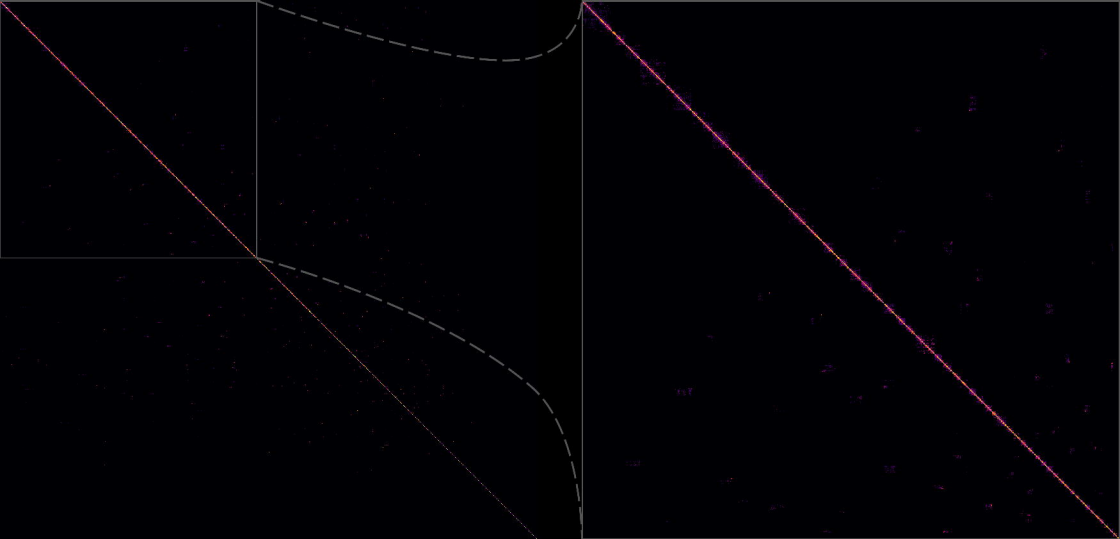


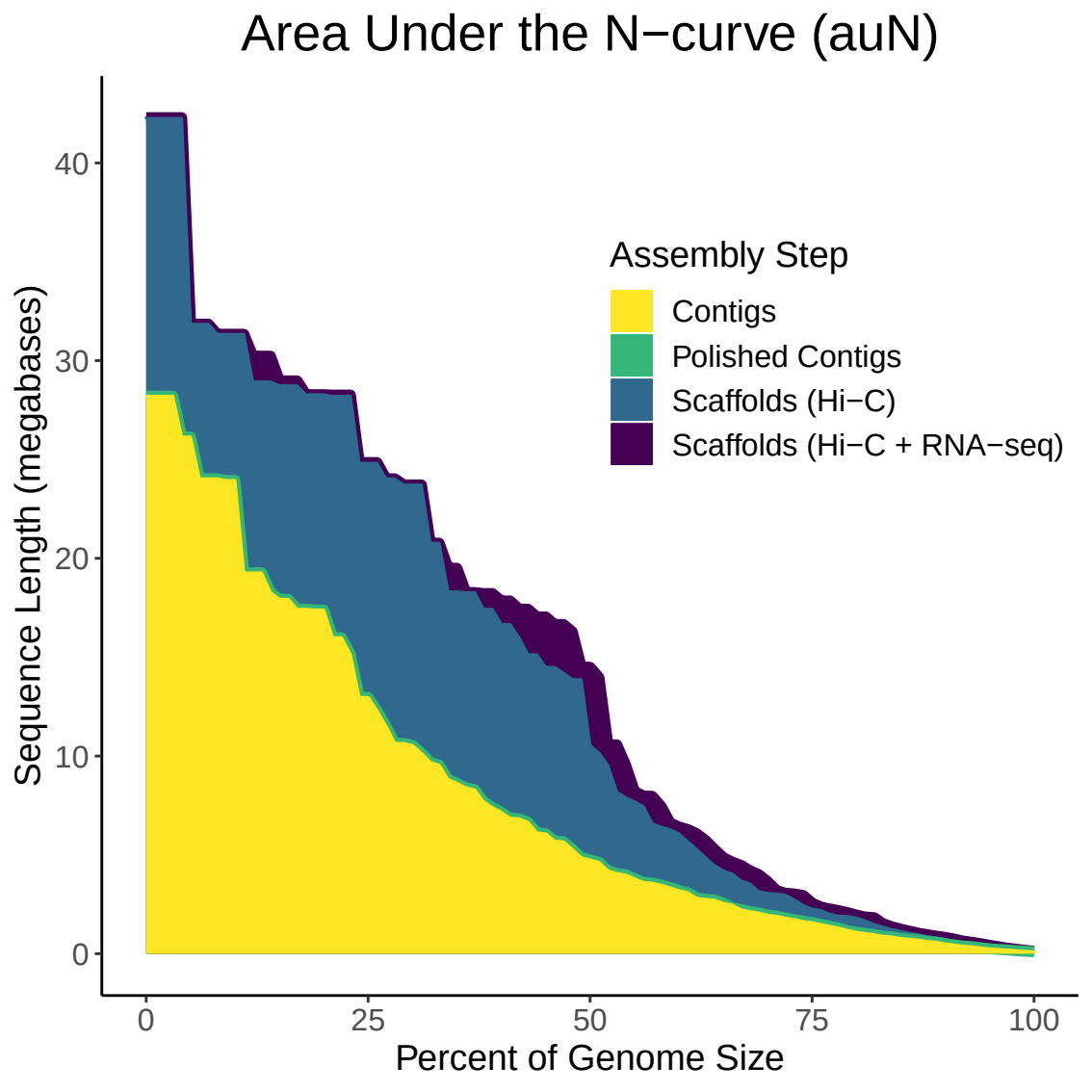
Albula glossodonta

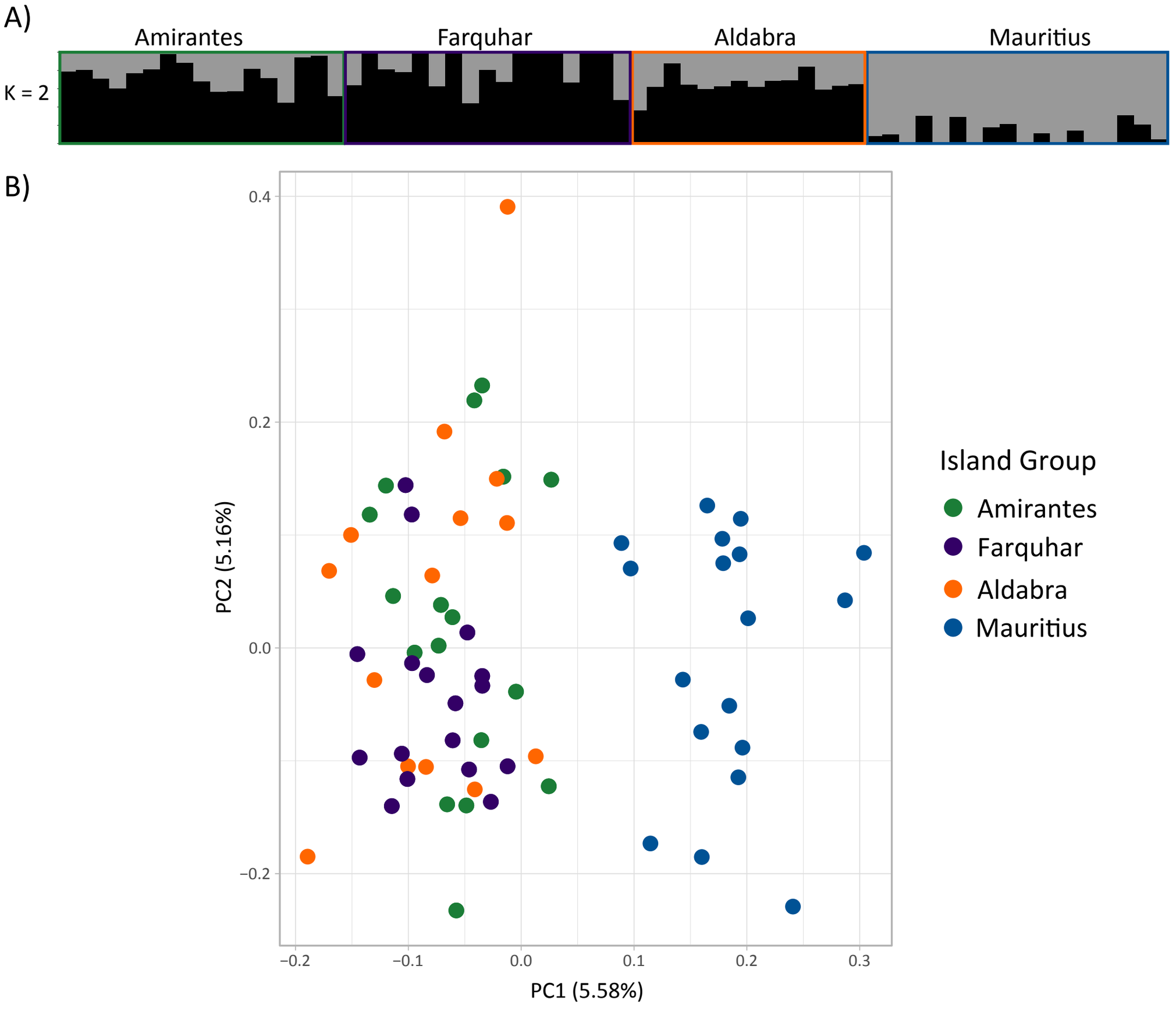
Jim Johnson '21




Frequency of Read Lengths
Frequency
2,000
1,000
0
0
25,000
50,000
75,000
100,000
Read Length
Read Type
Raw (Uncorrected)
Self−Corrected
Dual−Corrected




Area Under the N−curve (auN)
Assembly Step
Contigs
Polished Contigs
Scaffolds (Hi−C)
Scaffolds (Hi−C + RNA−seq)
Sequence Length (megabases)
40
30
20
10
0
0
25
50
75
100
Percent of Genome Size


A)
Amirantes
Farquhar
Aldabra
Mauritius
K = 2
B)
0.4
0.2
0.0
−0.2
PC2 (5.16%)
−0.2
−0.1
0.0
0.1
0.2
0.3
PC1 (5.58%)
Island Group
Amirantes
Farquhar
Aldabra
Mauritius