scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of the White-spotted spinefoot *Siganus canaliculatus*

Xiaolin Huang[1,2,3,6], Yanke Lu[4,6], Hui Zhang[4,6], Lin Xian[1,2,5,6], Shiting Huang[4], Yukai Yang[1,2,3], Lei Wang[4], Dianchang Zhang[1,2,3] ✉ & Chao Li[4] ✉

The White-spotted spinefoot *S. canaliculatus*, is an economically important marine fish in South China and featured by possessing poisonous glands in its fin spines. However, the unavailability of the *S. canaliculatus* genome has been a serious obstacle to genetic breeding as well as basic researches such as uncovering genomic basis underlying its toxigenic glands. Here, we presented a chromosome-level genome assembly coupled with good annotation of *S. canaliculatus* using multiple omics technologies. The assembled genome size was 547.39 Mb, with a contig N50 and scaffold N50 length of 21.41 Mb and 21.79 Mb, respectively. Approximately 95.32% (521.76 Mb) of assembled sequences were placed into 24 pseudochromosomes with the support of Hi-C contact map. Furthermore, around 16.37% of the genome was composed of repetitive elements. The quality of the assembly assessed using BUSCO showed that 98.6% of BUSCO genes were identified as complete. 25,323 protein-coding genes were predicted after integration of three kinds of evidence, of which 96.96% were functionally annotated in at least one of nine protein databases. In sum, the chromosome-level genome assembly and annotation provide fundamental resources for genetic breeding and molecular mechanism related studies of *S. canaliculatus*.

Background & Summary

The family Siganidae (also known as rabbitfish), are small and medium-sized marine fish. Rabbitfish inhabit nearshore reef areas and are found in the Indo-Pacific from the Red Sea and the coast of eastern Africa through the Pacific Ocean as far as Pitcairn Island[1]. As a group of perciform fishes, rabbitfish only includes one genus, namely *Siganus* Forsskål 1775 and currently 28 species are recognized[2]. However, natural hybridization are also found between both close related species or morphs and distantly related ones within rabbitfish[3], making taxonomy and phylogenetic studies of this taxa a little difficult and complicated. Rabbitfish are herbivorous and feed on benthic algae, consisting of a important community in coral reef ecosystem. Due to this feeding characteristic, they are usually introduced in culture ponds to clean net cages[4]. In aquaculture, there are several species (e.g., *S. canaliculatus, S. guttatus* and *S. fuscescens*) that are heavily explored because of their high protein content and delicious meat[4]. In addition, some species in Siganidae are very popular in the Indo-Pacific and Mediterranean regions as ornamental fishes due to their gorgeous appearance, such as *S. vermicularisi* and *S. corallinus*[5]. In

[1]Chinese Academy of Fishery Sciences, Key Laboratory of South China Sea Fishery Resources Exploitation and Utilization, Ministry of Agriculture and Rural Affairs, South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, 510300, China. [2]Sanya Tropical Fisheries Research Institute, Hainan Engineering Research Center of deep-sea aquaculture and processing, Sanya, 572018, China. [3]National Fishery Resources and Environment Dapeng Observation and Experimental Station, Shenzhen Base of South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Shenzhen, 518121, China. [4]Guangzhou Key Laboratory of Subtropical Biodiversity and Biomonitoring, Guangdong Provincial Key Laboratory for Healthy and Safe Aquaculture, Guangdong Provincial Engineering Technology Research Center for Environmentally Friendly Aquaculture, School of Life Sciences, South China Normal University, Guangzhou, China. [5]State Key Laboratory of Agricultural Genomics, BGI-Shenzhen, Shenzhen, 518083, China. [6]These authors contributed equally: Xiaolin Huang, Yanke Lu, Hui Zhang, Lin Xian. ✉e-mail: zhangdch@scsfri.ac.cn; 2015021118@m.scnu.edu.cn

Library type	Library size (bp)	Raw data (Gb)	Clean data (Gb)	Depth (×)†	Mean length/N50 (bp)
HiFi	20,000	25.14	—	45.02	16,243/16,338
Hi-C	350	101.66	96.75	173.26	—/149
Iso-seq	—	96.30	—	—	3,159/3,410
RNA-seq	350	18.14	16.96	30.37	—/149

Table 1. Sequencing data for *Siganus canaliculatus* genome assembly. †Estimated by a contig-level assembly (genome size: 558.39 Mb).

China, 14 Siganidae species are formally described or recorded with a distribution across South China Sea to East China Sea[5].

Among these species, the White-spotted spinefoot *S. canaliculatus* (synonym of *S. oramin*), is an important member for various reasons. First, *S. canaliculatus* is a common commercial fish in the family Siganidae and widely distributed in tropical and subtropical areas of the Indo-Pacific Ocean[1]. It is especially abundant in the wild along the coast of South China. Most of the rabbitfish have beautiful body color and appearance while *S. canaliculatus* has many small oblong yellow spots on the head and side of the body, which are relatively unremarkable[5]. Interestingly, its color can change sharply when inspired by external stimulus. As other species in this genus, *S. canaliculatus* is also featured by possessing poisonous glands in its dorsal and pelvic fin spines. The toxins likely originate from its food resource such as algae. However, its muscle is nontoxic and full of unsaturated fatty acids as well as minerals and trace elements[4]. The large gallbladder could be responsible for this special phenomenon (equal to 30% of its body length). These above valuable traits have made *S. canaliculatus* as one of the most important marine aquaculture species in the past decades in China costal provinces. For example, in Fujian province, more than 1000 tons have been reported for the annual production of this fish[5].

Meanwhile, as a saltwater fish, *S. canaliculatus* has the characteristics as freshwater fish. In general, the fertilized eggs of freshwater fish are heavy and sticky, while the fertilized eggs of marine fish are floating (caused by differences between the density of freshwater and seawater). However, as a true marine fish, *S. canaliculatus* is unusual by laying heavy and sticky fertilized eggs[6]. Moreover, freshwater fish usually have the ability to synthesize highly unsaturated fatty acids (HUFAs) while seawater fish generally lack or are poor at this ability. Their demands for HUFAs mainly depend on direct food intake, so the diet of seawater fish are highly dependent on fish oil. *S. canaliculatus* is the first seawater fish that has been found to possess the ability to convert linolenic acid and linoleic acid into HUFAs[7]. The *elovl* gene family was shown to function underlying biosynthesis of HUFAs[8,9].

Apart from nutrition studies, in recent years, there are many investigations of *S. canaliculatus* covering divers topics. For instance, morphology[6], genetic structures[10,11], phylogenetics[3,12], reproduction[13], net cage culture[14] as well as disease control[15]. However, our knowledge of *S. canaliculatus* have still been limited due to lack of genetic resources and genomic information. The advancements of third-generation sequencing and high-throughput chromatin conformation capture (Hi-C) technologies have provided an unprecedented opportunity for producing high quality and chromosome-level genomes for various organisms on the earth.

In this study, we employed an integrated strategy of HiFi long reads, Hi-C, Iso-seq and RNA-seq sequencing technologies to assemble a high-quality genome of *S. canaliculatus*. This genome was 547.39 Mb with contig N50 of 21.41 Mb and scaffold N50 of 21.79 Mb. Approximately 95.32% (521.76 Mb) of assembled sequences were placed into 24 pseudochromosomes with the support of Hi-C contact map. 25,323 protein-coding genes were predicted and 96.96% were functionally annotated. BUSCOs assessment of the assembly showed 3589 (98.6%) BUSCOs was complete. This high-quality *S. canaliculatus* reference genome will provide an important genomic resource for genetic breeding and molecular mechanism related studies.

Methods

Ethics statement. The fish in our experiments were collected from Shenzhen City, Guangdong Province, China. Furthermore, the methods used in this work are strictly in accordance with the Guidelines for the Care and Use of Laboratory Animals and approved by Laboratory Animal Ethics Committee of South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences (permit reference number No. 2024-MRB-00-001). Fish was collected for experiment utilization only and sacrificed using MS-222 (Sigma).

Sample collection and DNA extraction. A wild female *S.canaliculatus* (body mass: 250.2 g) was collected from Da Peng, Shenzhen, Guangdong, China (22°38′32.31″N; 114°24′40.87 E). The muscle was isolated and flash-frozen for ~30 minutes. Total DNA was extracted using QIAGEN Genomic DNA extraction kit and was used for PacBio sequencing and Hi-C sequencing. The extracted high molecular weight was assessed by 1% agarose gel and Qubit 3.0 Fluorometer (Invitrogen, USA).

Library construction and DNA sequencing. a SMRTbell Express Template Prep Kit 2.0 was used to generate a 20 kb long library for PacBio HiFi sequencing. The library was then sequenced on a PacBio Revio System (Pacifc Biosciences, Menlo Park, CA, USA). HiFi reads were obtained using the CCS module in SMRT Link v9.0[16]. After HiFi reads calling, 25.14 Gb PacBio HiFi reads were generated (N50: 20.47 kb, 45.02× in depth) (Table 1).

For Hi-C sequencing, a GrandOmics Hi-C kit with DpnII enzyme (GrandOmics, China) was used to construct libraries following the standard manufacturer's protocol. The resulted Hi-C libraries were sequenced on

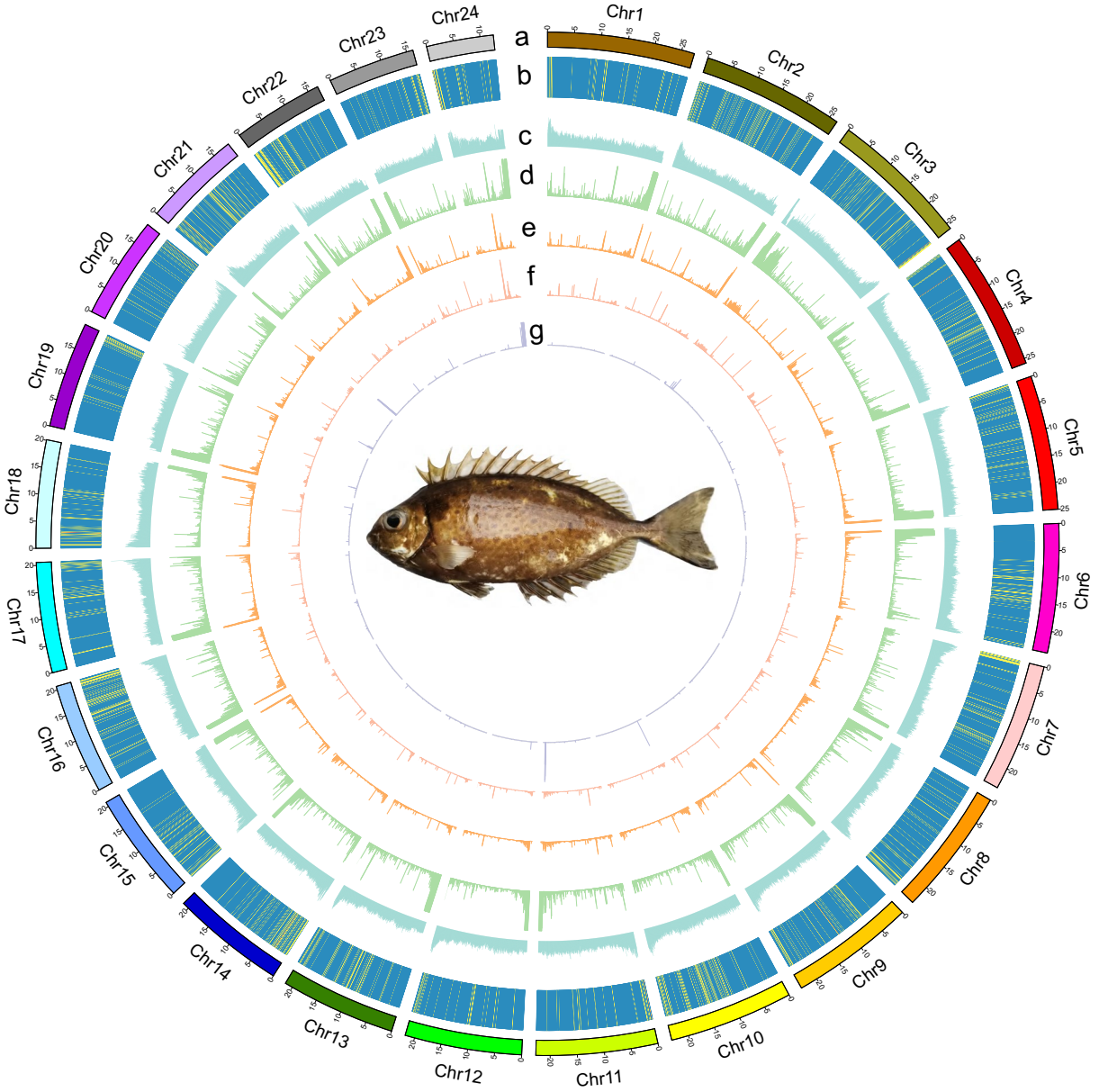

Fig. 1 Circos plot of *Siganus canaliculatus* genome. (**a**) chromosome sizes, (**b**) gene density, (c) GC density, (**d**) repeat elements abundance, (**e**) DNA transposons, (**f**) LTRs, and (**g**) ncRNAs.

a MGISEQ-2000 platform (MGI, BGI Shenzhen, China). 101.66 Gb raw reads were produced. These raw reads were filtered by using fastp v0.19.5[17] to filter low quality reads. 96.75 Gb (173.26 × in depth) clean reads were obtained in total. This clean Hi-C data was subsequently used for placing contigs onto psedochromosomes.

RNA extraction and sequencing. Both RNA-seq and Iso-seq were employed to assist RNA evidence based gene prediction. Seven tissues (skin, fin, heart, liver, gill, muscle and gonad) from the same individual as DNA extraction were equally mixed and extracted by using a TRIZOL Kit (Invitrogen, Carlsbad, CA, USA) following the manufacturer's instructions. RNA integrity and quality was checked by the Nanodrop 2000 spectrophotometer and the Agilent 2100 Bioanalyzer System (Agilent Technologies, Santa Clara, CA, USA). RNA with RIN (RNA integrity number) $\geq$7.0 were selected for library construction. Procedures described in our previous study[18] were performed for Iso-seq. Briefly, the extracted RNA was used for cDNA synthesis followed by a large-scale PCR amplification step. PCR products were purified and subjected to the construction of SMRTbell template libraries. Finally, SMRT cells were sequenced on a PacBio Revio platform. For RNA-seq, cDNA libraries with insert sizes of ~350 bp were constructed and sequenced on a MGISEQ-2000 platform (MGI, BGI Shenzhen, China). 96.30 Gb and 18.14 Gb raw data were generated from Iso-seq and RNA-seq, respectively (Table 1).

Genome assembly and telomere identification. HiFi reads were first assembled using hifiasm v0.19.5-r587[19] with default parameters to generate a contig-level assembly which had a size of 558.39 Mb with

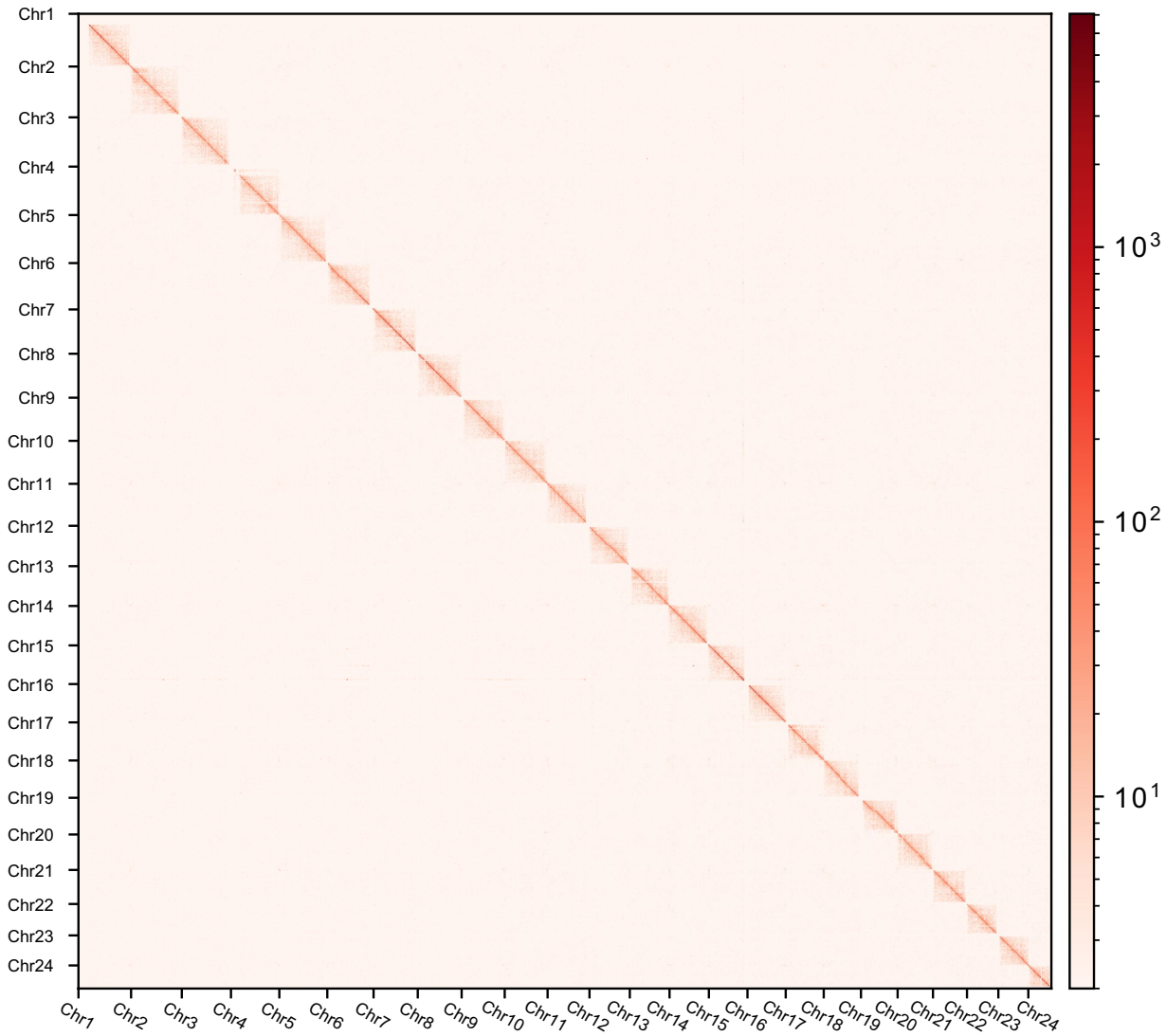

Fig. 2 Chromosome heatmaps of Hi-C data of *Siganus canaliculatus*. The bar beside indicates chromatin interactions quantified based on the count of Hi-C reads.

108 contigs (N50: 21.41 Mb). The mitochondrial sequences were removed in this step. After hifiasm assembly, purge_dups v1.2.6[20] was used to remove haplotigs and contig overlaps based on read depth following the standard pipeline. AutoHiC v1.3.3[21] was then used to scaffold these contigs using deep learning-based methods for automatic error correction. Briefly, this newly developed software utilizes Hi-C reads and input draft reference assembly to generate a candidate assembly. With built-in AutoHiC deep learning models, AutoHiC can automatically correct errors during genome assembly and generate a chromosome-level genome. The resulted draft genome was then polished by NextPolish v1.4.1[22] to fix base errors (SNV/Indel) with HiFi long reads. Telomere sequences at ends of each chromosome was identified quarTeT v1.2.5[23]. The size of the final assembly version was 547.39 Mb, of which 95.32% (521.76 Mb) were placed onto 24 chromosomes with Hi-C heat map support (Figs. 1, 2; Table 4). 70 sequences were presented in the final assembly with N50 length of 21.79 Mb. The length of 24 chromosome-level sequences ranged from 12.47 Mb to 27.41 Mb. The 24 chromosome numbers suggested by the Hi-C heat map was identical with a karyotype study of *S. canaliculatus*[24]. Telomere sequences were found to be presented at both ends of three chromosomes while only single telomere sequences were identified at one end of 20 chromosomes (Table 4).

Repeat elements annotation. EDTA pipeline[25] was used to annotate repeat elements in the *S. canaliculatus* genome. This pipeline was developed for automated whole-genome *de-novo* TE annotation. It first utilizes LTR-FINDER v1.0.6[26], LTRharvest[27], HelitronScanner[28] and TIR-Learner[29] to predict LTR, TIR and Helitron, respectively. Then, LTR_retriever v3.0.3[30] was used to filter false positive results of LTR. Subsequently, basic and advance filter in EDTA were applied to do additional filtering and resulted in raw TE library. This raw library was used for RepeatMasker v4.1.2-p1[31] to mask the target genome followed by RepeatModeler v2.0.3[32] to predict the

Class	Subclass	Repeat size (bp)	Percentage of genome
LTR			
	Copia	64758	0.01%
	Gypsy	5121211	0.94%
	unknown	8902636	1.63%
TIR			
	CACTA	17167767	3.14%
	Mutator	4966423	0.91%
	PIF_Harbinger	384112	0.07%
	Tc1_Mariner	364779	0.07%
nonLTR			
	DIRS_YR	84371	0.02%
	LINE_element	1925551	0.35%
	Penelope	55820	0.01%
nonTIR			
	helitron	3151857	0.58%
repeat_region		43821918	8.01%
Total		89597434	16.37%

Table 2. Statistics of repetitive sequences.

Method	Software	Species	Gene number
Ab initio	Augustus	—	38789
	GeneMark-ET	—	38161
Homology-based	blastn/blastx/exonerate	*Danio rerio*	37191
		Oreochromis niloticus	49829
		Oryzias latipes	37635
		Scatophagus argus	43500
		Takifugu rubripes	47202
Transcriptome-based	stringtie/taco (RNA-seq)	—	30416
	gmap (Iso-seq)	—	35972
Integration	maker	—	25323

Table 3. Statistics of gene prediction.

remaining TE in the genome. The results showed 89,597,434 bp (16.37%) was identified to be repetitive sequences (Table 2), in which LTR accounting for 2.58%, TIR 4.19%, nonLTR 0.38%, nonTIR 0.58% and repeat_region 8.1%.

Gene structure prediction and functional annotation. The masked genome generated in the repeat annotation step was used as an input for gene structure prediction. Three approaches which were commonly adopted was employed in this study: (1) *Ab initio* prediction: AUGUSTUS v3.5.0[33] and GeneMark-ET[34] were performed to do *ab initio* prediction; (2) Homology-based prediction: Protein sequences from five representative species (*Danio rerio, Oreochromis niloticus, Oryzias latipes, Scatophagus argus, Takifugu rubripes*) were download from the NCBI database. Using these data as references, gene structures in the *S. canaliculatus* genome were predicted using blastx v2.2.26[35] and exonerate v2.2[36]; (3) Transcriptome-based: for RNA-seq based predictions, raw RNA-seq reads were filtered using fastp[17] (-a auto --adapter_sequence_r2 auto --dedup --dup_calc_accuracy 3). After filtering, 16.96 Gb clean reads were mapped onto the *S. canaliculatus* genome using HISAT2 v2.2.1[37] and stringtie v2.2.1[38] and merged with TACO v0.7.3[39]. For Iso-seq based predictions, raw Iso-seq read was processed using isoseq pipeline[40]. GMAP[41] was introduced to align cDNA to the *S. canaliculatus* genome. Finally, gene structures predicted from above three methods were integrated by MAKER v3.01.03[42]. Genes with a Annotation Edit Distance (AED) ≤ 1 were retained in the final dataset.

For functional annotation of predicted genes, protein sequences were extracted from the *S. canaliculatus* genome and blasted against nine commonly used protein databases (NR, Swissprot, KEGG, KOG, GO, Pfam, TrEMBL, eggNOG, InterPro) using DIAMOND v0.9.25[43] with an *E* value of $1e^{-5}$ and InterProscan v5.59-91.0[44].

Non-coding RNA (ncRNAs, i.e., tRNAs, rRNAs, miRNAs, snRNAs and snoRNAs) in the *S. canaliculatus* genome were also annotated. We first utilized tRNAscan-SE v1.3.1[45] to predict tRNAs in the assembly. For the rRNA genes, RNAmmer v1.2[46] was used (-S euk -m lsu,ssu,tsu -gff). MiRNAs, snRNAs and snoRNAs were searched by CMSAN v1.1.2[47] against the Rfam v14.10 database[48] (--cut_ga --rfam --nohmmonly --tblout --fmt 2).

For *ab initio* prediction, AUGUSTUS v3.5.0[33] and GeneMark-ET[34] found 38789 and 38161 genes in the *S. canaliculatus* genome, respectively. Homology-based approach predicted 37191 to 49829 genes depending on reference genomes. RNA-seq based evidence predicted 30416 genes while Iso-seq based evidence found 35972

Chromosome	Size (Mb)	Gap number	Telomere number	GC%	Gene	Protein
Chr1	27.41	0	1	42.24%	1036	3193
Chr2	26.58	0	1	42.38%	1415	3818
Chr3	26.14	0	2	43.02%	1274	3297
Chr4	25.91	0	0	42.48%	1273	3532
Chr5	25.02	0	2	42.80%	1289	3543
Chr6	24.02	0	2	42.48%	951	2327
Chr7	23.89	0	1	42.72%	1231	3325
Chr8	23.72	0	1	42.65%	1125	2892
Chr9	23.26	0	1	42.66%	1183	3276
Chr10	23.19	0	1	42.97%	1211	3208
Chr11	22.64	0	1	42.83%	963	2589
Chr12	21.79	0	1	42.72%	929	2618
Chr13	21.47	0	1	42.51%	1025	3053
Chr14	21.27	0	1	42.76%	1213	3391
Chr15	20.86	0	1	43%	990	2534
Chr16	20.67	0	1	43.17%	1303	3405
Chr17	20.55	0	1	42.48%	982	2728
Chr18	20.12	0	1	43.09%	1053	2814
Chr19	19.81	0	1	42.74%	732	2071
Chr20	19.11	0	1	43.18%	966	2448
Chr21	18.37	0	1	43.50%	1028	2798
Chr22	16.95	0	0	42.80%	880	1987
Chr23	16.2	0	1	43.38%	670	1816
Chr24	12.47	0	1	44.77%	572	1461
unplaced	25.63	0	—	40.13%	29	33
Total	547.39	0	—	42.74%	25323	68157

Table 4. Statistics of gene numbers predicted across each chromosome.

Database	Annotated number (Percentage)	300 < =length < 1000	length > = 1000
NR	65891 (96.68%)	38368 (56.29%)	11093 (16.28%)
Swissprot	62791 (92.13%)	37411 (54.89%)	10994 (16.13%)
TrEMBL	65902 (96.69%)	38370 (56.30%)	11093 (16.28%)
GO	56342 (82.67%)	33636 (49.35%)	9733 (14.28%)
KEGG	48701 (71.45%)	29793 (43.71%)	8108 (11.90%)
KOG	50701 (74.39%)	30967 (45.43%)	9356 (13.73%)
eggNOG	64612 (94.80%)	38027 (55.79%)	11072 (16.24%)
Pfam	60502 (88.77%)	36312 (53.28%)	10613 (15.57%)
InterPro	63662 (93.40%)	37738 (55.37%)	11016 (16.16%)
Total	66083 (96.96%)	38400 (56.34%)	11096 (16.28%)

Table 5. Statistics of gene functional annotation.

Type	Number
miRNA	1352
tRNA	1551
rRNA	2968
snRNA	260
snoRNA	209

Table 6. Statistics of non-coding genes.

genes (Table 3). After integrated by MAKER v3.01.03[42], 25323 protein-coding genes were finally annotated with a range from 572 to 1415 genes across each chromosome (Table 4). Functional annotation results showed 71.45% to 96.68% of proteins can be blasted in one of nine databases (Fig. 3). After removing redundancy, 96.96% proteins had at least one database hits (Table 5). For ncRNA annotation, 1352 miRNA, 1551 tRNA, 2968 rRNA, 260 snRNA and 209 snoRNA were predicted in the *S. canaliculatus* genome (Table 6).

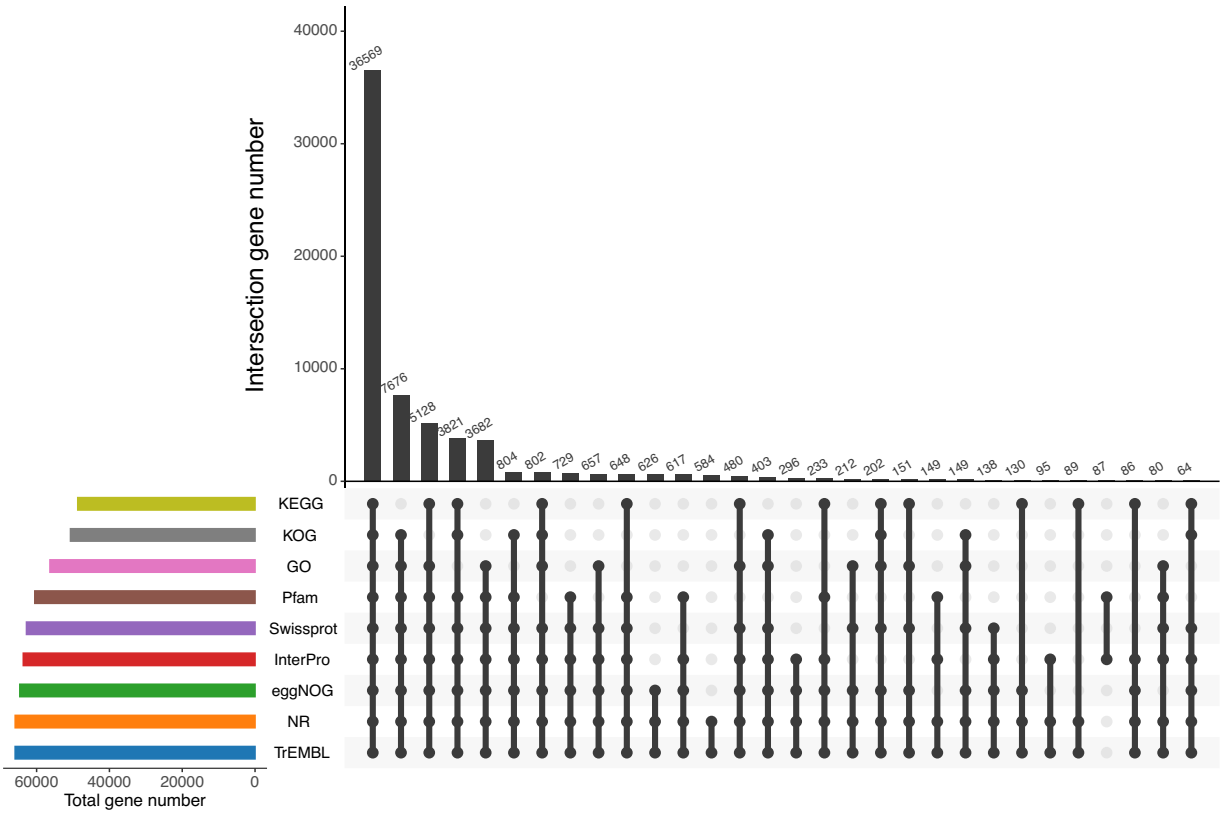

Fig. 3 Upset plot showing protein sequences of *Siganus canaliculatus* annotated in nine databases. Only the first 30 intersections have been shown.

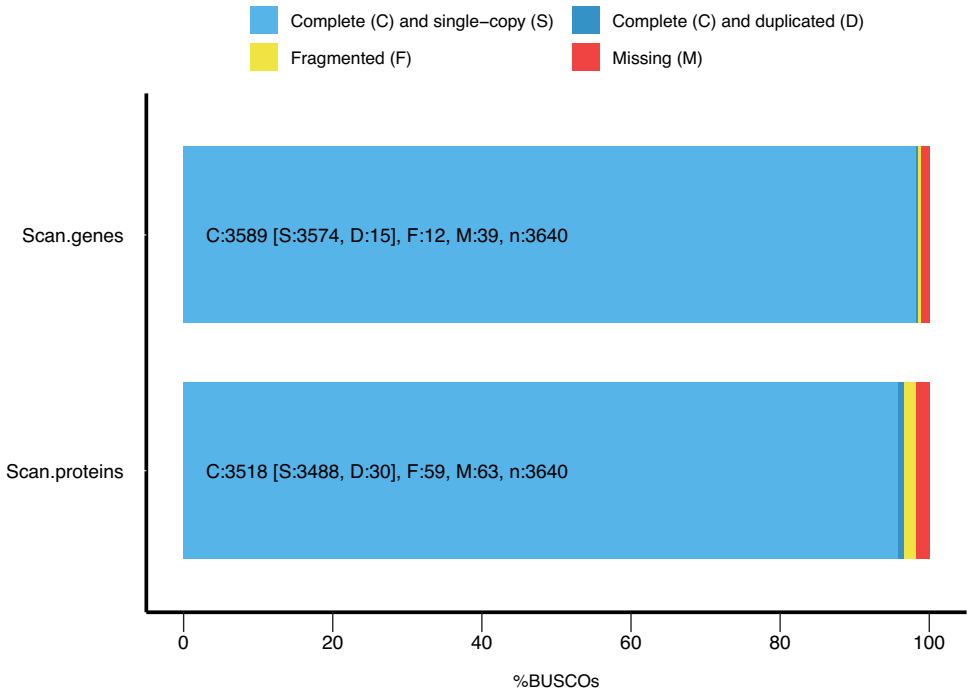

Fig. 4 BUSCO assessment results of Siganus canaliculatus gene and protein sequences.

Data Records

Raw reads sequenced in this study have been submitted to the National Genomics Data Center (https://ngdc.cncb.ac.cn/, BioProject number: PRJCA029961[49], Run IDs: CRR1288946-CRR1288949). The genome sequences

and annotation files were deposited at figshare (https://doi.org/10.6084/m9.figshare.27117169[50]) and NCBI (accession number: JBLRWB000000000[51]).

Technical Validation

The quality of the assembly was assessed using BUSCO v5.5.0[52] with the actinopterygii_odb10 database (3,640 BUSCOs). The BUSCO assessment showed that 3589 (98.6%) BUSCOs were identified as complete, of which 3574 (98.2%) and 15 (0.4%) were single-copy and duplicated, respectively. Chromosome numbers of the *S. canaliculatus* genome were confirmed by the Hi-C heat map (Fig. 2). Completeness assessment of proteins showed that a total of 3518 (96.6%) BUSCOs were identified as complete. Of these, 3488 (95.8%) were single-copy and 30 (0.8%) were duplicated BUSCOs (Fig. 4). Taking all above results and quality assessment metrics together, we concluded that the *S. canaliculatus* genome was high quality and has good annotations.

Code availability

No new scripts or pipelines were developed for this study. Software for raw data quality control, genome assembly and annotation, quality assessment have been described in the method part of this paper with parameters specified if applicable.

Received: 11 October 2024; Accepted: 17 March 2025;
Published online: 23 March 2025

References

1. Froese, R. & Pauly, D. Family Siganidae. *FishBase* (2023).
2. Randall, J. E. & Kulbicki, M. *Siganus woodlandi*, new species of rabbitfish (Siganidae) from New Caledonia. *Cybium* **29**, 185–189 (2005).
3. Kuriiwa, K., Hanzawa, N., Yoshino, T., Kimura, S. & Nishida, M. Phylogenetic relationships and natural hybridization in rabbitfishes (Teleostei: Siganidae) inferred from mitochondrial and nuclear DNA analyses. *Mol Phylogenet Evol* **45**, 69–80, https://doi.org/10.1016/j.ympev.2007.04.018 (2007).
4. Yang, Y. *et al*. Comparative analysis of nutritional composition of muscle from *Siganus oramin* living in different habitats (in Chinese). *South China Fisheries Science* **19**, 128–134 (2023).
5. Ma, Q. & Lu, J. Introduction and prospect of the systematics study of Siganidae in China (in Chinese). *South China Fisheries Science* **2** (2006).
6. Huang, X. *et al*. Morphology and growth of larval, juvenile and young *Siganus oramin* (in Chinese). *South China Fisheries Science* **14**, 88–94 (2018).
7. Li, Y. *et al*. Vertebrate fatty acyl desaturase with Delta4 activity. *Proc Natl Acad Sci USA* **107**, 16840–16845, https://doi.org/10.1073/pnas.1008429107 (2010).
8. Li, Y. *et al*. Genome wide identification and functional characterization of two LC-PUFA biosynthesis elongase (elovl8) genes in rabbitfish (*Siganus canaliculatus*). *Aquaculture* **522** https://doi.org/10.1016/j.aquaculture.2020.735127 (2020).
9. Wen, Z., Li, Y., Bian, C., Shi, Q. & Li, Y. Characterization of two kcnk3 genes in rabbitfish (*Siganus canaliculatus*): Molecular cloning, distribution patterns and their potential roles in fatty acids metabolism and osmoregulation. *Gen Comp Endocrinol* **296**, 113546, https://doi.org/10.1016/j.ygcen.2020.113546 (2020).
10. Huang, X. *et al*. Genetic variations among *Siganus oramin* populations in coastal waters of southeast China based on mtDNA control region sequences (in Chinese). *Journal of Tropical Oceanography* **37**, 45–51, https://doi.org/10.11978/2017109 (2018).
11. Peng, M. *et al*. Genetic diversity analysis of different geographical populations of *Siganus canaliculatus* along the South China Coast (in Chinese). *Journal of Hydroecology* **43**, 127–133, https://doi.org/10.15928/j.1674-3075.202104280127 (2022).
12. Huang, X. *et al*. Phylogenetic information analysis of mitochondrial genome sequences in *Siganus* (Perciformes: Siganidae) (in Chinese). *Journal of Biology* **35**, 33–36 (2018).
13. Huang, X. *et al*. Gonadal development of first sexual maturation of *Siganus oramin* cultured in pond (in Chinese). *South China Fisheries Science* **16**, 99–107, https://doi.org/10.12131/20200051 (2020).
14. Feng, G. *et al*. Feeding habit and growth characteristics of *Siganus canaliculatus* cultured in sea net cage (in Chinese). *Marine Fisheries* **30**, 37–42 (2008).
15. Jiang, B. *et al*. Transcriptome analysis provides insights into molecular immune mechanisms of rabbitfish, Siganus oramin against Cryptocaryon irritans infection. *Fish Shellfish Immunol* **88**, 111–116, https://doi.org/10.1016/j.fsi.2019.02.039 (2019).
16. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289, https://doi.org/10.1016/j.gpb.2015.08.002 (2015).
17. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
18. Li, C. *et al*. Full-Length Transcriptome Data for the White Cloud Mountain Minnow (Tanichthys albonubes) From a Wild Population Based on Isoform Sequencing. *Frontiers in Marine Science* **9** https://doi.org/10.3389/fmars.2022.831148 (2022).
19. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
20. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
21. Jiang, Z. *et al*. A deep learning-based method enables the automatic and accurate assembly of chromosome-level genomes. *Nucleic Acids Res* https://doi.org/10.1093/nar/gkae789 (2024).
22. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255, https://doi.org/10.1093/bioinformatics/btz891 (2020).
23. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res* **10**, uhad127, https://doi.org/10.1093/hr/uhad127 (2023).
24. Shu, H., Huang, C., Zhang, H. & Wang, Y. Studies on the karyotype of *Siganus canaliculatus* (in Chinese). *Journal of Guangzhou University (Natural Science Edition)* **9**, 90–93 (2010).
25. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20**, 275, https://doi.org/10.1186/s13059-019-1905-y (2019).
26. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**, W265–W268 (2007).
27. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18, https://doi.org/10.1186/1471-2105-9-18 (2008).

28. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc Natl Acad Sci USA* **111**, 10263–10268, https://doi.org/10.1073/pnas.1410068111 (2014).
29. Su, W., Gu, X. & Peterson, T. TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. *Mol Plant* **12**, 447–460, https://doi.org/10.1016/j.molp.2019.02.008 (2019).
30. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol* **176**, 1410–1422, https://doi.org/10.1104/pp.17.01310 (2018).
31. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **25**, 4.10. 11–14.10. 14 (2009).
32. Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009).
33. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–439, https://doi.org/10.1093/nar/gkl200 (2006).
34. Lukashin, A. & Borodovsky, M. GeneMark. hmm: new solutions for gene finding. *Nucleic acids research* **26**, 1107–1115 (1998).
35. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, https://doi.org/10.1186/1471-2105-10-421 (2009).
36. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31, https://doi.org/10.1186/1471-2105-6-31 (2005).
37. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
38. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
39. Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M. & Iyer, M. K. TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat Methods* **14**, 68–70, https://doi.org/10.1038/nmeth.4078 (2017).
40. PacificBiosciences. IsoSeq. *github*, https://github.com/PacificBiosciences/IsoSeq?tab=readme-ov-file (2024).
41. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875, https://doi.org/10.1093/bioinformatics/bti310 (2005).
42. Cantarel, B. L. *et al*. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**, 188–196, https://doi.org/10.1101/gr.6743907 (2008).
43. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
44. Quevillon, E. *et al*. InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, W116–120, https://doi.org/10.1093/nar/gki442 (2005).
45. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol Biol* **1962**, 1–14, https://doi.org/10.1007/978-1-4939-9173-0_1 (2019).
46. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100–3108, https://doi.org/10.1093/nar/gkm160 (2007).
47. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
48. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* **49**, D192–D200, https://doi.org/10.1093/nar/gkaa1047 (2021).
49. Chao, L. White-spotted spinefoot genome data archieve. *National Genomics Data Center* https://bigd.big.ac.cn/gsa/browse/CRA018870 (2024).
50. Chao, L. Chromosome-level genome assembly and annotation of the White-spotted spinefoot *Siganus canaliculatus*. *figshare* https://doi.org/10.6084/m9.figshare.27117169 (2024).
51. Chao, L. White-spotted spinefoot genome. *GenBank* https://identifiers.org/ncbi/insdc:JBLRWB000000000 (2025).
52. Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol* **38**, 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021).

Acknowledgements

This study was financially supported by the Core Technology Research Project for Suitable Species of Modern Marine Ranch in Guangdong Province (2024-MRB-00-001), Central Public-interest Scientific Institution Basal Research Fund (CAFS2023TD58). Chao Li was funded by the Natural Science Foundation of China (32300366), Guangdong Basic and Applied Basic Research Foundation (2023A1515010991;2022A1515110391), Guangzhou Basic and Applied Basic Research Foundation (2024A04J00318), China Postdoctoral Science Foundation (2022M711218), Open Project of Institute of Zoology, Guangdong Academy of Sciences (GIZ-KF202302).

Author contributions

C.L., X.H. and D.Z. conceived this project; H.Z., Y.L. and S.H. collected and identified the samples; C.L., Y.L., L.W. and X.H. did the genome assembly and annotation. C.L., H.X., Y.L. and L.X. wrote the manuscript. All authors have read and approved the final manuscript for publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.Z. or C.L.