Genomics Proteomics Bioinformatics 20 (2022) 1053–1065

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Genomes of Two Flying Squid Species Provide Novel Insights into Adaptations of Cephalopods to Pelagic Life

Min Li [1,3,4,#], Baosheng Wu [2,#], Peng Zhang [1], Ye Li [2], Wenjie Xu [2], Kun Wang [2], Qiang Qiu [2], Jun Zhang [1], Jie Li [1], Chi Zhang [5], Jiangtao Fan [1], Chenguang Feng [2,6,*], Zuozhi Chen [1,3,4,*]

[1] *South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou 510300, China*
[2] *School of Ecology and Environment, Northwestern Polytechnical University, Xi'an 710072, China*
[3] *Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou 511458, China*
[4] *Key Laboratory for Sustainable Utilization of Open-Sea Fishery, Ministry of Agriculture and Rural Affairs, Guangdong Provincial Key Laboratory of Fishery Ecology and Environment, Guangzhou 510300, China*
[5] *Qinghai Provincial Key Laboratory of Crop Molecular Breeding, CAS Key Laboratory of Adaptation and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810008, China*
[6] *CAS Key Laboratory of Aquatic Biodiversity and Conservation, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China*

Received 25 July 2021; revised 25 August 2022; accepted 28 September 2022
Available online 7 October 2022

Handled by Yu Jiang

KEYWORDS

Cephalopoda;
Evolution;
Flying squid;
Genome;
Photophore

Abstract Pelagic cephalopods have evolved a series of fascinating traits, such as excellent visual acuity, high-speed agility, and **photophores** for adaptation to open pelagic oceans. However, the genetic mechanisms underpinning these traits are not well understood. Thus, in this study, we obtained high-quality **genomes** of two purpleback **flying squid** species (*Sthenoteuthis oualaniensis* and *Sthenoteuthis* sp.), with sizes of 5450 Mb and 5651 Mb, respectively. Comparative genomic analyses revealed that the *S*-crystallin subfamily *SL20-1* associated with visual acuity in the purpleback flying squid lineage was significantly expanded, and the **evolution** of high-speed agility for the species was accompanied by significant positive selection pressure on genes related to energy metabolism. These molecular signals might have contributed to the evolution of their adaptative predatory and anti-predatory traits. In addition, the transcriptomic analysis provided clear indications of

* Corresponding authors.
E-mail: chenzuozhi@scsfri.ac.cn (Chen Z), fengcg@nwpu.edu.cn (Feng C).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.09.009

the evolution of the photophores of purpleback flying squids, especially the recruitment of new genes and energy metabolism-related genes which may have played key functional roles in the process.

Introduction

Cephalopods are a group of marine mollusk with remarkable morphology and behavior that play key ecological roles, are commercially important, and have been intensively studied [1]. Large populations of cephalopods inhabit depths ranging from shallow to abyssal [2,3]. They are preyed upon by various apex predators (such as billfish, tuna, sharks, and cetaceans) that are generally the fastest and most efficient in the ocean [4–6]. In response to the predation, most cephalopods have developed excellent visual acuity [7–9], strong muscles, and morphological traits [10] that enable them to evade danger rapidly. They also have high metabolic levels and hence constantly high levels of energy supplies for their muscles [11]. The continuous adaptations and counter-adaptations induced by interactions between prey and predators — the 'arms race', is one of the most intense forms of co-evolution. This 'arms race' is particularly pronounced between the pelagic cephalopods (*e.g.*, cuttlefishes and squids) and their predators [12]. These cephalopods have extremely strong muscles with obliquely-striated, quickly contractible, highly aerobic fibers, and morphological features, such as fins, which enable powerful swimming [10]. Some species of pelagic cephalopods have also repeatedly evolved photophores that assist in escape [13], predation [14], and mating [15]. However, although it has long been accepted that adaptative evolution has resulted in a series of fascinating traits in cephalopods [7,16–18], the genetic mechanisms involved are much less clear.

Common pelagic cephalopods include members of the Ommastrephidae (Teuthida, Decapodiformes) called flying squids, which can jump out of the water and in some cases glide more than 30 m in the air [19]. They can achieve the fastest recorded speed (∼ 8 m/s) of any aquatic invertebrates [20,21]. Important taxa with such capabilities include the purpleback flying squids (*Sthenoteuthis* spp.) [20,22,23]. These are the most abundant large squids in the tropical and subtropical Indo-Pacific ocean, found at depths from the surface to more than 600 m [2]. Moreover, the purpleback flying squids have a high degree of adaptation to their pelagic life and at least five morphological and ecological forms in terms of body size and possession or absence of photophores [2]. Thus, the purpleback flying squids are ideal models for studying the genetic mechanisms involved in the evolution of pelagic cephalopods' adaptive traits.

In this study, we constructed high-quality genomes for two 'forms' of purpleback flying squids. One (*S. oualaniensis*) is the 'medium' or 'typical' form, which has a dorsal mantel length (at maturity) of > 120 mm and spherical photophores forming an oval patch in the anterodorsal mantle musculature [2,24] (Figure S1). The other (*Sthenoteuthis* sp.) is the 'dwarf form', which is smaller and lacks the dorsal photophore patch (Figure S1). The dwarf form was previously treated as *S. oualaniensis*, but is now considered an undescribed species [25,26]. Through comparative genomic analyses, we deeply profiled genomic features of these two purpleback flying squids and investigated molecular signals associated with adaptations of pelagic life in cephalopods, such as their excellent visual system, high behavioral flexibility, and photophore. The results suggest that the genomes of these two purpleback flying squids are important resources that can facilitate research not only on cephalopods but also on adaptive evolution and molecular genetics more generally.

Results and discussion

Assemblies and genomic characteristics

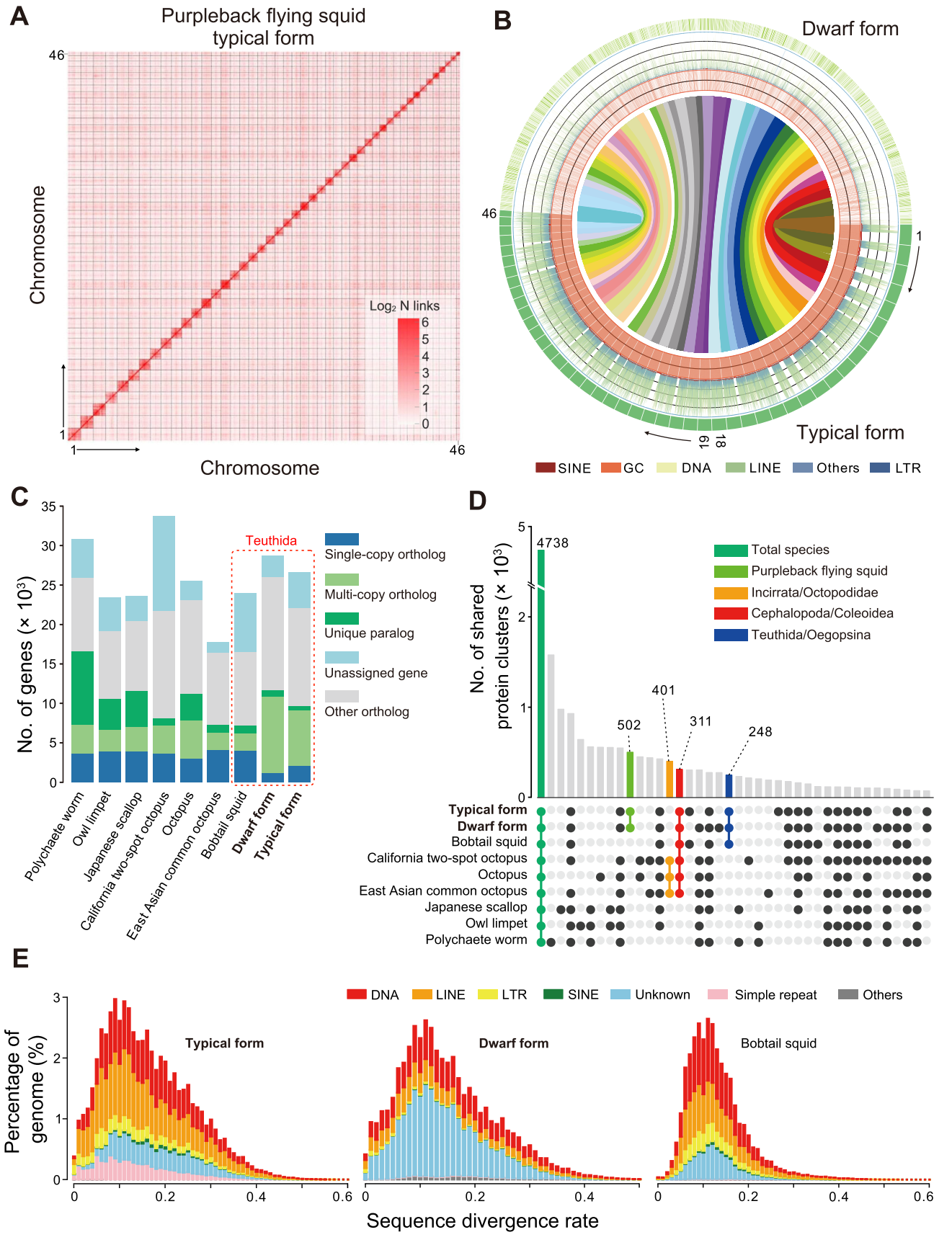
Using a combination of PacBio long reads, 10× Genomics short reads, RNA sequencing (RNA-seq), and Hi-C approaches, we obtained high-quality genomes of the two species of purpleback flying squids of the Ommastrephidae (*S. oualaniensis* and *Sthenoteuthis* sp.) (Tables S1–S4; Figures S2 and S3). The sizes of the genome assemblies were 5450 Mb and 5651 Mb for *S. oualaniensis* and *Sthenoteuthis* sp., respectively (**Table 1**), close to the genome sizes estimated by *k*-mer analysis (Figure S4). The *S. oualaniensis* genome was assembled to chromosome level, with 95.72% of contigs anchored to the 46 chromosomes (**Figure 1**A). The contig N50 and scaffold N50 values of the *S. oualaniensis* assembly were 1.52 Mb and 118.8 Mb, respectively, while the contig N50 value for the *Sthenoteuthis* sp. assembly was 1.3 Mb. The *S. oualaniensis* genome is the only squid genome that has been assembled to chromosome level (Figure 1A; Table 1). GC contents of the acquired *S. oualaniensis* and *Sthenoteuthis* sp. genomes were 33.70% and 33.00%, respectively, similar to those of other relatives (Table S5). Results of the Benchmarking Universal Single-Copy Orthologs (BUSCO) [27] assessment showed that the *S. oualaniensis* and *Sthenoteuthis* sp. assemblies had 89.8% and 93.6% genomic integrity, respectively (Table 1). RNA-seq analysis of *S. oualaniensis* and *Sthenoteuthis* sp. yielded 389,954 and 410,364 transcripts with total lengths of 200–25,632 bp and 200–31,105 bp, respectively (Table S6).

Based on the high-quality assemblies, we predicted a total of 26,646 and 28,715 protein-coding genes in the genomes of *S. oualaniensis* and *Sthenoteuthis* sp., respectively (Table S5). Of these, 19,913 (74.73%) and 21,853 (76.10%) are supported by corresponding transcripts, and 24,845 (93.24%) and 27,897

Table 1 Summary of the assemblies and annotations of the two *Sthenoteuthis* species

	Sthenoteuthis oualaniensis	*Sthenoteuthis* sp.
Genome size (Mb)	5450	5651
Contig N50 (Mb)	1.52	1.3
Scaffold N50 (Mb)	118.8	-
GC content	33.07%	33.00%
Chromosome number	46	-
Gene number	26,646	28,715
BUSCO for genomes	89.4%	93.6%
BUSCO for genes	90.4%	94.3%

Note: BUSCO, Benchmarking Universal Single-Copy Orthologs.


A
Purpleback flying squid
typical form
Chromosome
Chromosome
Log2 N links
B
Dwarf form
Typical form
SINE
GC
DNA
LINE
Others
LTR
C
No. of genes (× 10³)
Teuthida
Single-copy ortholog
Multi-copy ortholog
Unique paralog
Unassigned gene
Other ortholog
Polychaete worm
Owl limpet
Japanese scallop
California two-spot octopus
Octopus
East Asian common octopus
Bobtail squid
Dwarf form
Typical form
D
No. of shared protein clusters (× 10³)
4738
502
401
311
248
Total species
Purpleback flying squid
Incirrata/Octopodidae
Cephalopoda/Coleoidea
Teuthida/Oegopsina
Typical form
Dwarf form
Bobtail squid
California two-spot octopus
Octopus
East Asian common octopus
Japanese scallop
Owl limpet
Polychaete worm
E
DNA
LINE
LTR
SINE
Unknown
Simple repeat
Others
Percentage of genome (%)
Typical form
Dwarf form
Bobtail squid
Sequence divergence rate

(97.13%) predictions could be functionally annotated using entries in at least one database (Figures S5 and S6). Results of BUSCO analysis suggested that the assembly completeness of *S. oualaniensis* and *Sthenoteuthis* sp. were 89.4% and 93.6%, respectively, and the annotation completeness of them were 90.4 and 94.3%, respectively (Table 1, Table S7). Basic metrics for the protein-coding genes of these two purpleback flying squids, including gene number/length, exon number/length, and codon usage, are close to those of other Mollusca (Table S5). Synteny analysis, based on the coding genes, identified 743 blocks with at least five syntenic genes between *S. oualaniensis* and *Sthenoteuthis* sp. and demonstrated the good colinearity between the two genomes, which also indicated that the two genomes were well assembled and annotated (Figure 1B). Finally, 22,162 (83%) *S. oualaniensis* genes and 25,991 (90%) *Sthenoteuthis* sp. genes were clustered into 16,235 gene clusters, with sizes close to those of other relatives (Figure 1C and D).

Repeats analysis indicated that close to half of the genomes of both *S. oualaniensis* and *Sthenoteuthis* sp. are composed of repetitive sequences: 53.90% and 42.37%, respectively (Tables S8 and S9), which is comparable to that of other cephalopods [28,29]. Identification and classification of the transposable elements (TEs) of the two genomes (Figure 1E; Tables S8 and S9) showed that DNA and long interspersed nuclear element (LINE) were the two most abundant types of TEs in the *S. oualaniensis* genome (accounting for 25.47% and 22.05% of the total, respectively), but only accounted for 14.13% and 10.00% of the TEs, respectively, in the *Sthenoteuthis* sp. genome. There was a markedly smaller proportion of simple tandem repeats in the *S. oualaniensis* genome (6.33%) than in the *Sthenoteuthis* sp. genome (25.86%) (Tables S8 and S9). Thus, despite the very close relationship between the two purpleback flying squid species, they might have distinct patterns of TE activity.

Phylogenetic status and species validity of two purpleback flying squids

Maximum-likelihood (ML) gene tree and species tree analyses based on 334 one-to-one genes yielded consistent topologies (Figure S7). The Octopodiformes and Decapodiformes species each clustered into monophyletic lineages, which apparently diverged around 366.5 million years ago (MYA) (**Figure 2**; Figure S7). The two purpleback flying squid species are the most closely related of the included taxa, having diverged at approximately 41.0 MYA. These two forms of purpleback flying squids, *S. oualaniensis* and *Sthenoteuthis* sp., were considered to be members of the same species for a long time, but this was recently questioned [25,26]. Morphological studies provided evidence that *Sthenoteuthis* sp. was distinguishable from *S. oualaniensis* concerning external features, including variables of the head, carcass, and arms, as well as the shape and size of fins [30,31]. Our results corroborate the conclusion that *Sthenoteuthis* sp. should be regarded as a distinct species, as we detected clear differences in their genomes and derived a substantial divergence time.

Excellent vision of purpleback flying squids

Keen vision is regarded as a major product of evolutionary arms races in squids [7], which strongly contributes to the ability of purpleback flying squids (and others) to avoid predation and catch prey even in dim conditions [1,7,9]. Important components of their eyes include several soluble crystallins that play key roles in maintaining the transparency and optical clarity of the lens [32]. In particular, *S*-crystallins are present in the lenses of many cephalopods, and have refractive properties that strongly contribute to good vision (and hence cephalopod survival) in poor light [33,34]. *S*-crystallins are even claimed to provide a "perfect medium", forming gels of varying density, in the spherical lenses of cephalopods [34–36].

Cluster analysis detected 2489 gene families that were expanded in both purpleback flying squid species (under the criterion of both family-wide and viterbi P values < 0.01; Figure 2, Figures S8 and S9). The *SL20-1* gene subfamily which encodes *S*-crystallins was most significantly expanded, with 39 and 99 gene members in *S. oualaniensis* and *Sthenoteuthis* sp., respectively (**Figure 3**A and B). Transcriptomic analysis revealed that most of these expanded *SL20-1* genes were only highly expressed in eyes (Figure 3C, Figure S10), clearly indicating that expansion of this subfamily plays an important role in the emergence of purpleback flying squids' excellent vision.

High behavioral flexibility

The mobility of organisms strongly affects their predatory and anti-predatory abilities, and thus their evolutionary fitness [37]. Purpleback flying squids are highly successful in these terms, as they are the fastest and most mobile aquatic invertebrates, with very high metabolic levels that enable powerful output at all times [11,23].

Figure 1 Genome assemblies and annotations of two *Sthenoteuthis* species

S. oualaniensis and *Sthenoteuthis* sp. are referred to as the "typical form" and "dwarf form" of purpleback flying squids, respectively. **A.** Genome-wide Hi-C map of the 46 pseudo-chromosomes of *S. oualaniensis*. **B.** Syntenic comparison of *S. oualaniensis* and *Sthenoteuthis* sp. Numbers 1 to 46 refer to the chromosomes of *S. oualaniensis*. Each color block in the outermost layer represents a scaffold. Each line represents a syntenic block of five or more genes. Densities of specific kinds of TEs (ranging from 0 to 80%) were counted in 500-kb windows. **C.** Summary of gene clusters estimated from Orthofinder analysis based on sequences of eight Mollusca species and one annelid worm. **D.** UpSet plot of gene families. **E.** Divergence distribution of TEs of species in Teuthida. Latin binomials of the listed species are as follows: Polychaete worm, *Capitella teleta*; Owl limpet, *Lottia gigantea*; Japanese scallop, *Mizuhopecten yessoensis*; California two-spot octopus, *Octopus bimaculoides*; Octopus, *Octopus minor*; East Asian common octopus, *Octopus vulgaris*; Bobtail squid, *Euprymna scolopes*. TE, transposable element; LINE, long interspersed nuclear element; SINE, short interspersed nuclear element; LTR, long terminal repeat.

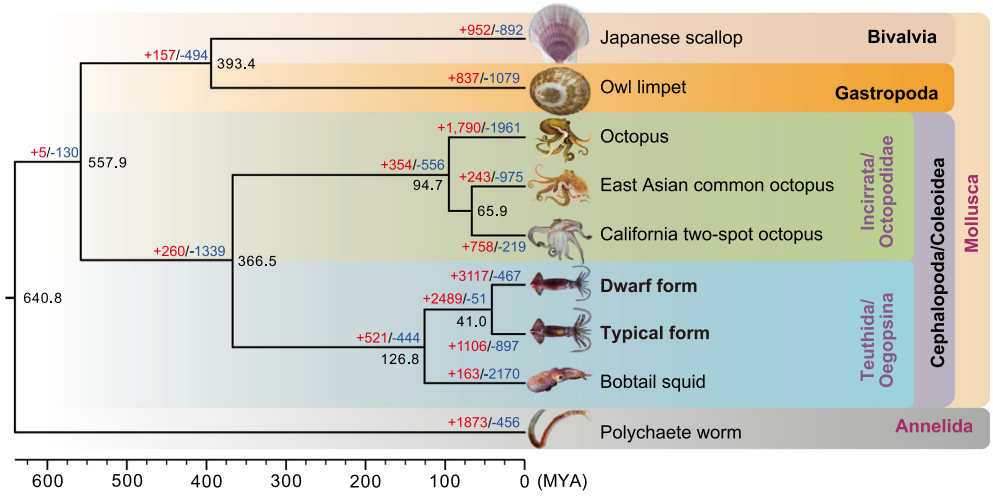


Figure 2 Coalescence tree of eight Mollusca species and one annelid worm based on 334 single-copy orthologs
Estimated divergence time and expanded/contracted gene families are marked at the nodes. Red, blue, and black numbers indicate the number of expanded gene families, the number of contracted gene families, and the estimated divergence time, respectively. MYA, million years ago.

Our analysis identified 66 positively selected genes (PSGs) in the *Sthenoteuthis* lineage [false discovery rate (FDR)-adjusted $P < 0.01$; Table S10]. Gene Ontology (GO) enrichment analyses (FDR < 0.01) indicated that 23, 42, and 30 of these PSGs were associated with cellular components, molecular functions, and biological processes, respectively (Figures S11 and S12). The PSGs involved in energy metabolism were apparently under significant evolutionary selection pressure (FDR-adjusted $P < 0.01$). "Phosphoglycerate kinase activity" (GO:0004618, *e.g.*, *pgk1*) and "[2Fe-2S] cluster assembly" (GO:0044571, *e.g.*, *IscS*) were the two most significant GO terms (FDR < 0.05; Tables S11 and S12). The "quinone binding" term (GO:0048038, *e.g.*, *ndufa6* and *ndufa7*) was also significantly enriched (FDR < 0.05; Table S11). Transcriptomic analysis demonstrated that these genes were highly expressed in all the investigated tissues, implying that they play important roles (**Figure 4**A).

Phosphoglycerate kinase 1 (PGK1), the first ATP-generating enzyme in the glycolytic pathway, both directly generates ATP and indirectly supplies fuel for the mitochondrial electron respiratory chain [38,39]. Phylogenetic analysis by maximum likelihood (PAML) detected positive selection signals at two extremely conserved regions of the *pgk1* gene (F85Y and F149Y; Figure 4B). Simulations showed that these two F-to-Y substitutions affect the structure of PGK1 protein resulting in the conversion of the original helix to a loop (Figure 4C and D). Thus, these substitutions were presumably highly important for purpleback flying squids. Similarly, a positive selection signal was detected at a conserved region of the *IscS* gene (Y67H; Figure S13). *IscS* participates in the synthesis of multiple iron-sulfur (Fe-S) proteins and the formation of Fe-S clusters in complex I (NADH: ubiquinone oxidoreductase; Figure 4E) [40]. Knockdown of the *IscS* gene leads to a decrease in mitochondrial activity [40,41]. We also found positive selection signals in *ndufa6* and *ndufa7*, which encode two important subunits of complex I that are essential for the catalytic activity of this complex (Table S10) [42,43]. Complex I plays a key role in ATP synthesis driven by the mitochondrial electron respiratory chain [44] (Figure 4E). Therefore, the PSGs mentioned above are likely to promote ATP synthesis, which is important for the maintenance of high metabolic levels and the high behavioral flexibility of purpleback flying squids. Furthermore, the presence of these positively selected sites in the pelagic *Architeuthis dux* (Figure S14), which also possess high metabolic levels [2,29], implies the importance of these selection signals for pelagic cephalopods. However, we should note that these results come from a small gene pool and therefore have some limitations.

Photophore transcriptome

Bioluminescence is a common feature of cephalopods, especially pelagic species [14,45]. At least 63 genera of squid and cuttlefish have repeatedly evolved photophores that play important roles in defense, predation, and communication [13–15]. One of the species included in this study, *S. oualaniensis*, has a dorsal photophore patch, but not *Sthenoteuthis* sp. (Figure S1). Therefore, they are ideal models for studying the evolution of photophores. Principal component analysis (PCA) showed that *S. oualaniensis* photophores clustered with the *pseudo*-photophores ('muscle tissue' corresponding to the position of the photophore of *S. oualaniensis*) of *Sthenoteuthis* sp. (**Figure 5**A), and highly expressed genes of photophores had similar expression patterns (Figure 5B). Therefore, the *pseudo*-photophores of *Sthenoteuthis* sp. and photophores of *S. oualaniensis* seem to be homologous organs, and the *pseudo*-photophores may have some essential functions similar to those of the photophores.

Previous studies based on gene expression profiles have suggested that the massive recruitment of pre-existing gene modules plays an important role in the formation of cephalopod photophores [46,47]. Here we specifically investigated the highly expressed genes in the "photophores" of these two squids, which represent genetic factors associated with the

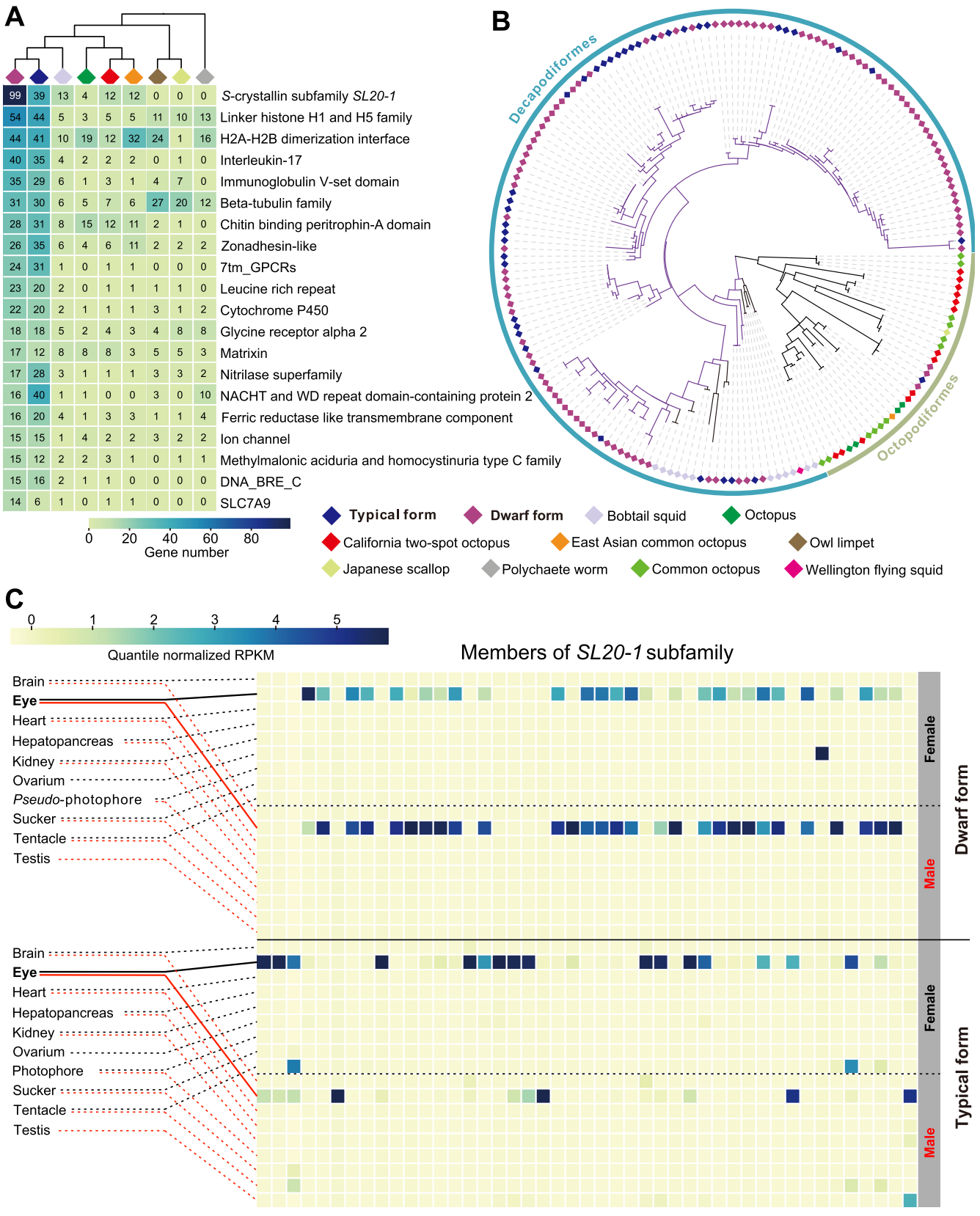


Figure 3 Expanded gene families in the *Sthenoteuthis* lineage

A. The top 20 expanded gene families. **B.** Unrooted maximum-likelihood tree of the massive expansion of the *S*-crystallin subfamily *SL20-1*, with 39 and 99 gene members in *S. oualaniensis* and *Sthenoteuthis* sp., respectively. **C.** Expression patterns of coexisting gene members of the *SL20-1* subfamily in *S. oualaniensis* and *Sthenoteuthis* sp. Only those expressed gene members were shown. These gene IDs were detailed in Figure S10. Most of the significantly expanded genes were highly expressed in the eyes. Commen octopus, *Octopus sinensis*; Wellington flying squid, *Sepia pharaonis*.

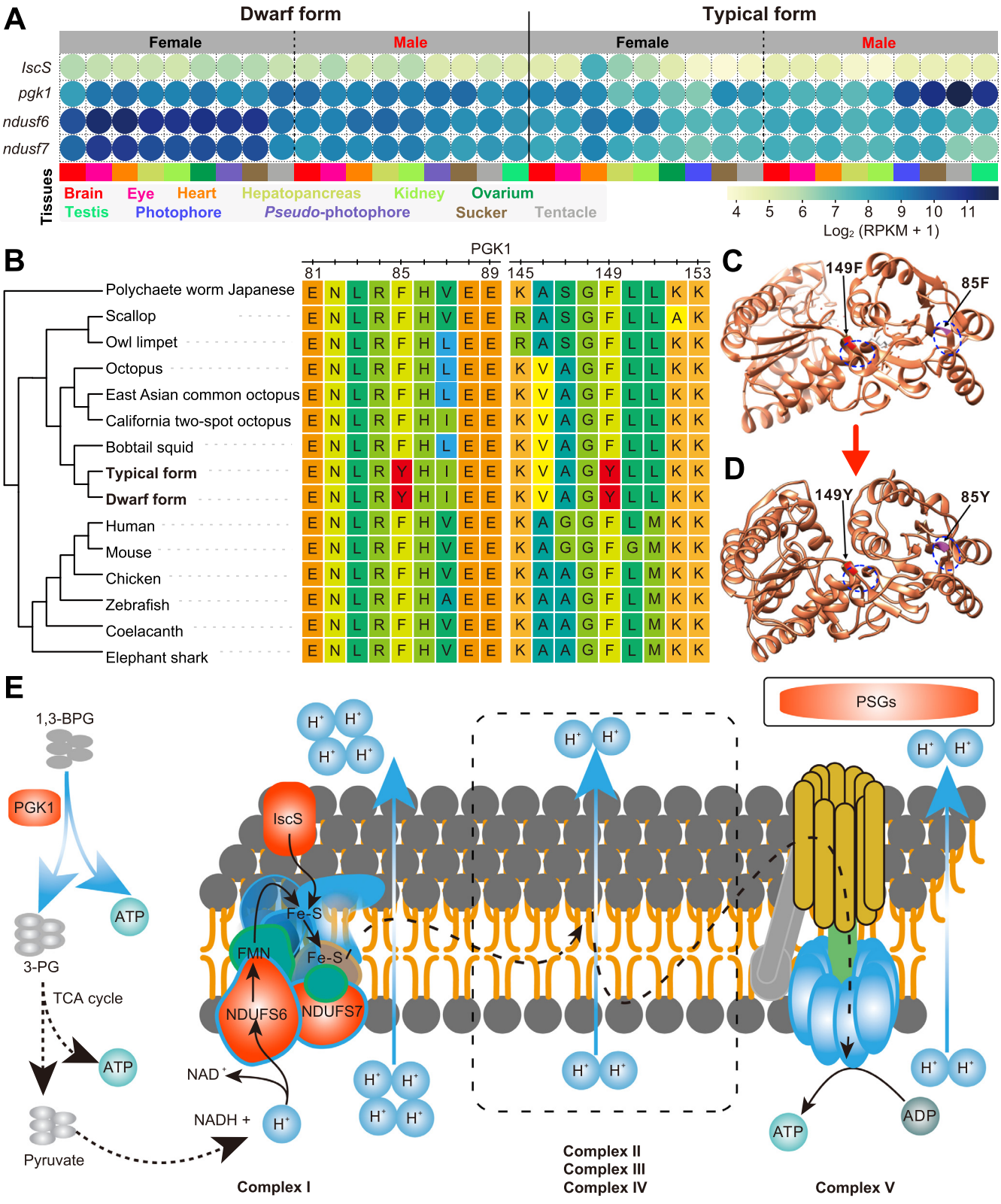


Figure 4 Diagram of PSGs associated with energy metabolism

A. Expression patterns of four PSGs involved in energy metabolism that are highly expressed in all investigated tissues. **B.** Positively selected signals in two extremely conserved regions of PGK1. **C.** Three-dimensional structure of mouse PGK1 protein downloaded from the PDB database. Substructures of 85F and 149F are highlighted. **D.** Three-dimensional structure simulated by a homologous approach of mouse PGK1 with F85Y and F149Y substitutions. Structures adjacent to the substitute sites in (C) and (D) are signaled by the blue dashed circles. **E.** Schematic diagram of the glycolysis pathway and respiratory electron chain. PSG, positively selected gene; PGK1, phosphoglycerate kinase 1; PDB, Protein Data Bank; RPKM, reads per kilobase million.

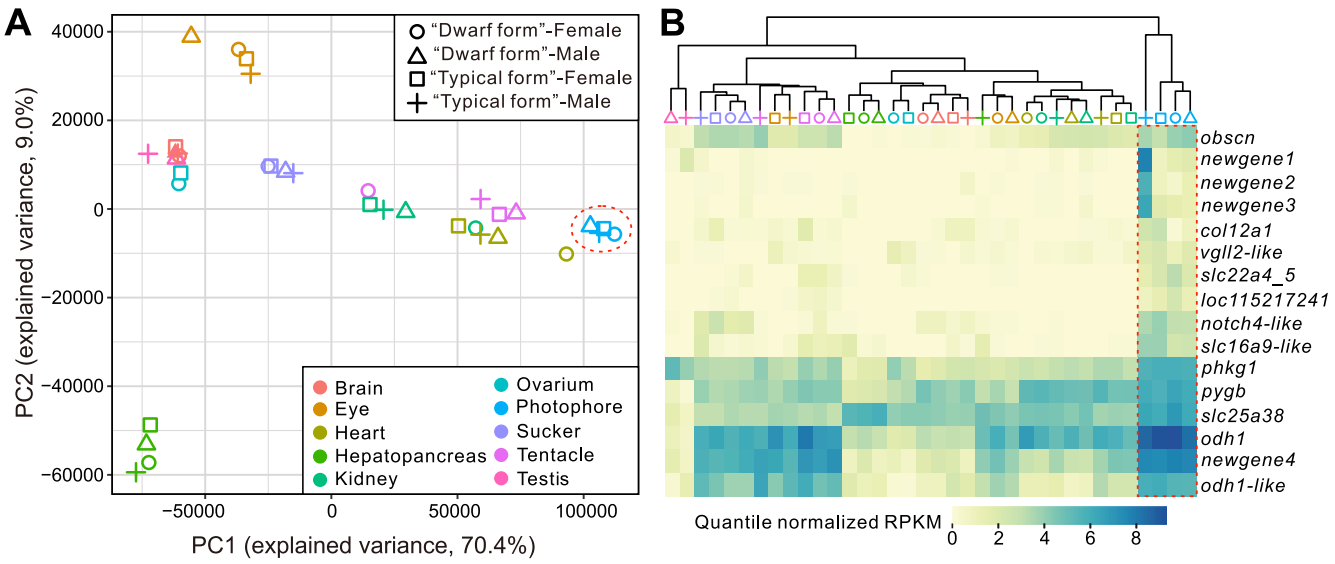


Figure 5 Results of photophore transcriptome analysis

A. PCA plot of 36 RNA-seq samples based on expression profiling of 23,082 orthologs. *Pseudo*-photophores of *Sthenoteuthis* sp. and photophores of *S. oualaniensis* clustered together (red dotted circle). **B.** Whole-tissue expression patterns of 16 highly expressed genes in the four "photophore" tissues. PCA, principal component analysis; PC, principal component.

essential function of this organ. Among the 16 highly expressed genes, four were new genes (*newgene1–4*; Figure 5B). Although functional assignations of these genes were not supported by homologs in public databases, their specific expression patterns suggest that they might participate in the formation and some of the basic functions of luminophores. In addition, four of the highly expressed genes (*phkg1*, *pygb*, *odh1*, and *obh1*-like) are energy metabolism-related genes. All of these genes are involved in glucose metabolism. Both *phkg1* and *pygb* are glycogen phosphorylases that regulate the catabolism of glycogen and provide glucose used in glycolysis [48,49]. Octopine dehydrogenase (encoded by *odh1*), mainly found in mollusks [50], has a major function similar to that of lactate dehydrogenase, providing an important reducing agent for the glycolytic process [51]. These observations imply that these new genes and recruited energy metabolism-related genes might provide important support for the photophores of purpleback flying squids.

Conclusion

In this study, we generated high-quality genomes of the two purpleback flying squid species. Comparative genomic analyses indicated that expansion of the *S*-crystallin subfamily *SL20-1* and locus variation and expression patterns of genes related to energy metabolism are associated with adaptations of purpleback flying squids (such as excellent vision, high behavioral flexibility, and photophore) that have played important roles in the 'arms race' and other pelagic adaptations among marine organisms. Moreover, the study supports the validity of treating *Sthenoteuthis* sp. as a separate species at the genomic level. These findings advance our understanding of the genetic basis of pelagic cephalopods associated with predatory and anti-predatory traits and suggest that the two genomes could be important resources for studying not only cephalopods, but also co-evolution, bioluminescence, and other broader aspects of molecular genetics.

Materials and methods

Sampling and sequencing

Squid samples were caught by a commercial fishing vessel using a lit falling net at night in the South China Sea. One male individual and one female individual of both medium (*S. oualaniensis*) and dwarf (*Sthenoteuthis* sp.) purpleback flying squids were collected for sequencing. Muscles from the males were used for DNA extraction and genomic library preparation. A PacBio Sequel device was used for sequencing the long reads. Short-insert paired-ends libraries were prepared and sequenced according to the Illumina sequencing protocol. Sample indexing and partition barcoded libraries were prepared using a chromium genome reagent kit (Catalog No. PN-120229, 10× Genomics, CA) and sequenced by an Illumina HiSeq X-Ten system for Hi-C analysis of *S. oualaniensis*. To explore gene expression patterns of the species and aid gene annotation, RNA was extracted from nine tissues, including tentacle, brain, eye, heart, kidney, sucker, hepatopancreas, ovarium/testis, and photophore (for *Sthenoteuthis* sp., muscle tissue corresponding to the position of the photophore of *S. oualaniensis* was obtained), from each of the four *S. oualaniensis* and *Sthenoteuthis* sp. individuals for library preparation and sequenced using the Illumina HiSeq 2000 platform.

Estimation of genome sizes of the two *Sthenoteuthis* species

The genome sizes of the two *Sthenoteuthis* species were estimated by *k*-mer analysis using filtered Illumina reads. We used SOAPec v2 [52] to estimate the distribution of 17-mer depth,

and then estimated the genome size from the total base and peak values of 17-mer depth.

Genome assembly

Based on the estimated genome sizes, we first assembled genomes of the two purpleback flying squid species to contig level with wtdbg2 v2.4.1 [53] and standard parameters. The arrow algorithm was used to polish the two draft genomes with the filtered PacBio reads. Then the filtered short paired-end reads were aligned to the draft genome by BWA-MEM v.0.7.12-r1039 [54] with standard parameters, and Pilon [55] was used to further polish the genomes in two rounds using the sorted bam files. Finally, the genome assembly of *S. oualaniensis* was anchored with the Hi-C reads by 3D-DNA [56] and Juicer v1.5 [57]. To improve the quality of the chromosome assembly, we used Juicebox assembly tools [58] to remove potential assembly errors. BUSCO v3.02 [27] with the "metazoa_odb9" library was used to evaluate the completeness of the two assemblies.

Repetitive sequence annotation

After obtaining a high-quality genome assembly, we used a combination of *de novo* and homologous predictions to annotate repetitive sequences, including tandem repeats and TEs. Firstly, tandem repeat finder v4.07 [59] was used to scan the tandem repeat elements with the parameter settings "2 7 7 80 10 50 500 -d -h -ngs". Then we used RepeatModeler v1.0.8 [60] to build a *de novo* repeat library, and RepeatMasker v3.3.0 [61] to detect homologous repeat elements. After integrating the results of *de novo* and homologous predictions, Jukes-Cantor distances were calculated, and the R8s algorithm [62] was used to calculate rates of evolution from them.

Protein-coding gene prediction

A combination of *ab initio*, homologous, and transcript-based gene predictions was used to integrate the two genomes. The gene prediction pipeline was as follows. First, AUGUSTES v3.2.1 [63], GlimmerHMM v3.02 [64], and GeneID v1.4 [65] were used for *de novo* gene prediction. Second, we downloaded the non-redundant proteomes of *Lottia gigantea* (GCF_000327385.1), *Mizuhopecten yessoensis* (GCF_002113885.1), *Octopus bimaculoides* (GCF_001194135.1), *Euprymna scolopes* (GCA_004765925.1), *O. minor* (http://dx.doi.org/10.5524/100503), and *O. vulgaris* (GCA_003957725.1) for homologous gene prediction. We used TBLASTN v2.9 [66] to align the proteomes of the six relatives to the two purpleback flying squid genomes and extended 10,000 bp in both directions from the start and end of every TBLASTN hit. Then, all non-redundant transcripts of all tissues were aligned to the genome with TBLASTN and a 1000-bp extension was applied. Genewise v2.4 [67] was then used to resolve the gene structure according to the aforementioned hits. Next, we integrated results of the *ab initio*, homologous, and transcript-based predictions with 1:4:5 weights using EvidenceModeler [68]. Finally, for further functional annotation of these two gene sets, we scanned public databases, including Swiss-Prot, KOG, Nr, KEGG, GO, and Pfam to detect the best matches using Interproscan v5 [69].

Gene family clustering analysis

In addition to the two predicted gene sets and six Mollusca species mentioned above, we downloaded the genome of *Capitella teleta* (GCA_000328365.1) as an outgroup. Proteomes of these nine species were subjected to an all-vs-all BLAST search (E-value $\leq$ 1E−6) and then clustered by OrthoFinder [70] with default parameters. The shared gene clusters were visualized by the R package UpSetR [71]. The expanded and contracted gene families were investigated by CAFÉ v4.0.1 [72] using the result of the clustering analysis under the criterion of both family-wide and viterbi P values < 0.01.

Phylogeny and divergence time estimation

Based on the cluster analysis of the aforementioned nine species, the protein-coding sequences and corresponding codon sequences of 334 one-to-one homologous genes were picked out and aligned using MAFFT v7 [73]. The bad alignments were removed by trimAl [74]. Finally, we used RAxML v8.2.4 [75] with "-m GTRGAMMA -f a -x 271828 -N 100 -p 54321" parameter settings to construct phylogeny trees and ASTRAL [76] to infer a species tree. MCMCtree in the PAML package [77] was used to estimate divergence times in conjunction with two softbound calibration points from https://www.timetree.org, *O. bimaculoides–C. teleta* (585–679 MYA) and *O. bimaculoides–L. gigantea* (531–582 MYA).

Synteny between the genomes of two *Sthenoteuthis* species

To evaluate the conservation and quality of the two assemblies of *Sthenoteuthis* species, we used the *S. oualaniensis* assembly as a reference and *Sthenoteuthis* sp. assembly as the query in alignment analysis by LAST v942 [78] with the "-E 0.05" parameter setting. We also calculated the densities of repetitive elements and GC contents in 500-kb windows of the genomes. Finally, the results of these analyses were integrated into a circular layout by CIRCOS v0.69 [79].

Positive selection analysis

To evaluate the evolutionary pressure on *Sthenoteuthis*, we used the one-to-one homologous genes of the nine species listed above (see the "Gene family clustering analysis" section for details) to identify PSGs. We aligned codons of all the one-to-one homologous genes by PRANK v140603 [80] with "-codon -f = fasta". All the gaps generated by alignments were removed by Gblocks v0.91b [81] with "-t = c". Then we used an in-house Perl script to convert the aligned sequence to PAML format for use in PAML analysis. Finally, PAML 4.9i [77] was used to analyze the selection pressure on each gene with the ML method under the branch-site model. A species tree constructed from ASTRAL analysis was used as the input tree. The two *Sthenoteuthis* species were selected as foreground and the other seven species as background. The significance of the alternative model (estimated omega) against the null model (fixed omega) was assessed by likelihood ratio tests (LRTs), in which twice the log-likelihood difference (2DL) values were calculated and compared to a chi-squared distribution. Genes with $P < 0.01$ (with FDR correction) and

carrying at least one site under positive selection with a Bayes empirical Bayes (BEB) posterior probability > 0.8, were identified as candidate PSGs.

Protein structure simulation

A homology-based approach was used to simulate structures of proteins encoded by PSGs. We first sought matches to the *S. oualaniensis* PGK1 protein sequence in the Protein Data Bank (PDB) database (https://www.rcsb.org/) and selected the hit with the highest score as a potential template for the simulation of the PGK1 protein structure. The corresponding positively selected sites (F85 and F149) of mouse PGK1 protein were replaced by those of *Sthenoteuthis* (85Y and 149Y). Then, the modified sequence was submitted to Phyre2 [82] for structure simulation. Finally, the Phyre2 result with the highest score was selected as the final structure and visualized by UCSF Chimera [83].

Transcriptomic analysis

Raw reads obtained from RNA-seq of the nine mentioned tissues of the four sampled individuals were filtered using fastp [84] with default parameters. The low-quality reads were removed by Sickle v1.33 (https://github.com/najoshi/sickle) with default parameters except for the "pe" setting. The cleaned reads were mapped to the reference genome with HISAT2 [85]. Then Trinity [86] was used to assemble these reads into transcripts. Next, we used TransDecoder [87] with default parameters to predict gene structures of the transcripts and CD-HIT [88] to remove redundant predictions. The numbers of reads and reads per kilobase of transcript per million mapped reads (RPKM) values for all genes in the 36 tissues were calculated by StringTie v2.1.4 [89] using the output of HISAT2 [85] analysis. Each gene with an RPKM value greater than 1 was considered a validly expressed gene. The *Tau* value of genes in all tissues was calculated using an in-house Perl script. Genes expressed in the photophore with *Tau* values $\geq$ 0.8 and higher RPKM values than in other tissues were regarded as being highly specifically expressed in the photophore. All the RNA-seq data for the 36 tissues were subjected to PCA, using normalized and logarithmically transformed RPKM values. Seaborn [90] was used to cluster and visualize clusters of the highly expressed genes in the photophore. Those genes shared only by the two *Sthenoteuthis* species and supported by transcripts were identified as new genes. To investigate the expression pattern of expanded genes of the *SL20-1* subfamily across organs, RNA-seq reads of both species were mapped to the genomes of *Sthenoteuthis* sp. and subsequently assessed by RPKM values.

Code availability

The custom scripts related to this work have been deposited in BioCode at the National Genomics Data Center (NGDC), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS) / China National Center for Bioinformation (CNCB), and are publicly accessible at https://ngdc.cncb.ac.cn/biocode/tools/BT007297/releases/1.0.

Data availability

The genomes and annotations of these two purpleback flying squid species, raw sequencing data, and acquired RNA-seq data have been deposited in the China National GeneBank (CNGB: CNP0001958), and are publicly accessible at https://db.cngb.org/cnsa. The raw sequencing data have also been deposited in the Genome Sequence Archive [91] at the NGDC, BIG, CAS / CNCB (GSA: CRA004867), and are publicly accessible at https://ngdc.cncb.ac.cn/gsa. The genomes and annotations of these two purpleback flying squid species have also been deposited in the Genome Warehouse [92] at the NGDC, BIG, CAS / CNCB (GWH: GWHBECU00000000 for *S. oualaniensis*; GWHBFHL00000000 for *Sthenoteuthis* sp.), and are publicly accessible at https://ngdc.cncb.ac.cn/gwh.

CRediT author statement

Min Li: Conceptualization, Validation, Formal analysis, Investigation, Project administration, Writing - original draft. **Baosheng Wu:** Formal analysis, Data curation, Visualization, Writing - original draft. **Peng Zhang:** Validation, Resources, Investigation, Writing - review & editing. **Ye Li:** Formal analysis, Data curation, Visualization. **Wenjie Xu:** Formal analysis, Data curation, Visualization. **Kun Wang:** Writing - review & editing. **Qiang Qiu:** Writing - review & editing. **Jun Zhang:** Resources, Writing - review & editing. **Jie Li:** Resources, Investigation. **Chi Zhang:** Methodology, Software, Data curation. **Jiangtao Fan:** Resources, Writing - review & editing. **Chenguang Feng:** Conceptualization, Methodology, Validation, Project administration, Writing - original draft. **Zuozhi Chen:** Supervision, Project administration, Funding acquisition, Writing - review & editing. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This study was supported by grants from the Guangdong Major Project of Basic and Applied Basic Research, China (Grant No. 2019B030302004), the National Key R&D Program of China (Grant No. 2018YFC1406502), the Key Special Project for Introduced Talents Team of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), China (Grant No. GML2019ZD0605), the Financial Fund of the Ministry of Agriculture and Rural Affairs of China (Grant No. NFZX2018), the Central Public-interest Scientific Basal Research Fund, Chinese Academy of Fishery Sciences (Grant Nos. 2019HY-XKQ03, 2020TD05, and 2021SD18), the China Postdoctoral Science Foundation (Grant No. 2021M693342), the Hubei Postdoctoral Innovation Post Project, China, the 1000 Talent Project of Shaanxi Province, China, and the Research Funds for Interdisciplinary subject of Northwestern Polytechnical University, China

(Grant No. 19SH030408). We gratefully acknowledge colleagues at BGI-Shenzhen for data analyses. We also thank Prof. Shuigui Jiang at South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences for project administration and Dr. Shuai Zhang for reviewing the manuscript.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2022.09.009.

ORCID

ORCID 0000-0002-5551-0134 (Min Li)
ORCID 0000-0002-9442-3738 (Baosheng Wu)
ORCID 0000-0002-8371-242X (Peng Zhang)
ORCID 0000-0002-2734-2700 (Ye Li)
ORCID 0000-0001-6240-8472 (Wenjie Xu)
ORCID 0000-0001-6059-6529 (Kun Wang)
ORCID 0000-0002-9874-271X (Qiang Qiu)
ORCID 0000-0002-4683-4735 (Jun Zhang)
ORCID 0000-0002-7495-8559 (Jie Li)
ORCID 0000-0002-6281-4187 (Chi Zhang)
ORCID 0000-0003-1383-3055 (Jiangtao Fan)
ORCID 0000-0002-4566-6848 (Chenguang Feng)
ORCID 0000-0003-4045-1212 (Zuozhi Chen)

References

[1] Hoving HJT, Perez JAA, Bolstad KSR, Braid HE, Evans AB, Fuchs D, et al. The study of deep-sea cephalopods. Adv Mar Biol 2014;67:235–359.
[2] Jereb P, Roper CFE. Myopsid and oegopsid squids. Cephalopods of the world. An annotated and illustrated catalogue of cephalopod species known to date. Rome: FAO; 2010.
[3] Jereb P, Roper CFE. Chambered nautiluses and sepioids (Nautilidae, Sepiidae, Sepiolidae, Sepiadariidae, Idiosepiidae and Spirulidae). Cephalopods of the world. An annotated and illustrated catalogue of cephalopod species known to date. Rome: FAO; 2005.
[4] Lindberg DR, Pyenson ND. Things that go bump in the night: evolutionary interactions between cephalopods and cetaceans in the tertiary. Lethaia 2007;40:335–43.
[5] Allain V. What do tuna eat? A tuna diet study. SPC Fish Newsl 2005:20–2.
[6] Rosas-Luis R, Loor-Andrade P, Carrera-Fernández M, Pincay-Espinoza JE, Vinces-Ortega C, Chompoy-Salazar L. Cephalopod species in the diet of large pelagic fish (sharks and billfishes) in Ecuadorian waters. Fish Res 2016;173:159–68.
[7] Partridge JC. Sensory ecology: giant eyes for giant predators? Curr Biol 2012;22:R268–70.
[8] Nilsson DE, Warrant EJ, Johnsen S, Hanlon R, Shashar N. A unique advantage for giant eyes in giant squid. Curr Biol 2012;22:683–8.
[9] Thomas KN, Robison BH, Johnsen S. Two eyes for two purposes: *in situ* evidence for asymmetric vision in the cockeyed squids *Histioteuthis heteropsis* and *Stigmatoteuthis dofleini*. Philos Trans R Soc Lond B Biol Sci 2017;372:20160069.
[10] Hochachka PW, Moon TW, Mustafa T, Storey KB. Metabolic sources of power for mantle muscle of a fast swimming squid. Comp Biochem Physiol B 1975;52:151–8.
[11] Shulman GE, Chesalin MV, Abolmasova GI, Yuneva TV, Kideys A. Metabolic strategy in pelagic squid of genus *Sthenoteuthis* (Ommastrephidae) as the basis of high abundance and productivity. An overview of the Soviet investigations. Bull Mar Sci 2002;71:815–36.
[12] Doubleday ZA, Prowse TA, Arkhipkin A, Pierce GJ, Semmens J, Steer M, et al. Global proliferation of cephalopods. Curr Biol 2016;26:R406–7.
[13] Young Richard E, Roper Clyde FE. Bioluminescent countershading in midwater animals: evidence from living squid. Science 1976;191:1046–8.
[14] Johnsen S, Balser EJ, Fisher EC, Widder EA. Bioluminescence in the deep-sea cirrate octopod *Stauroteuthis syrtensis* Verrill (Mollusca: Cephalopoda). Biol Bull 1999;197:26–39.
[15] Robison BH, Young RE. Bioluminescence in pelagic octopods. Pac Sci 1981;35:39–44.
[16] York CA, Bartol IK, Krueger PS, Thompson JT. Squids use multiple escape jet patterns throughout ontogeny. Biol Open 2020;9:bio054585.
[17] Hanlon RT, Messenger JB. Cephalopod behaviour. Cambridge: Cambridge University Press; 2018.
[18] York CA, Bartol IK. Anti-predator behavior of squid throughout ontogeny. J Exp Mar Biol Ecol 2016;480:26–35.
[19] Muramatsu K, Yamamoto J, Abe T, Sekiguchi K, Hoshi N, Sakurai Y. Oceanic squid do fly. Mar Biol 2013;160:1171–5.
[20] Hou TG, Yang XB, Wang TM, Liang JH, Li SW, Fan YB. Locomotor transition: how squid jet from water to air. Bioinspir Biomim 2020;15:036014.
[21] Vogel S. Flow-assisted mantle cavity refilling in jetting squid. Biol Bull 1987;172:61–8.
[22] O'Dor R, Stewart J, Gilly W, Payne J, Borges TC, Thys T. Squid rocket science: how squid launch into air. Deep-Sea Res Part II-Topical Stud Oceanogr 2013;95:113–8.
[23] Zuyev G, Nigmatullin C, Chesalin M, Nesis K. Main results of long-term worldwide studies on tropical nektonic oceanic squid genus *Sthenoteuthis*: an overview of the soviet investigations. Bull Mar Sci 2002;71:1019–60.
[24] Young RE. A brief review of the biology of the oceanic squid, *Symplectoteuthis oualaniensis* (Lesson). Comp Biochem Physiol B 1975;52:141–3.
[25] Nesis KN. Population structure of oceanic ommastrephids, with particular reference to *Sthenoteuthis oualaniensis*: a review. In: Okutani T, O'Dor RK, Kubodera T, editors. Recent advances in fisheries biology. Tokyo: Tokai University Press; 1993, p. 375–83.
[26] Staaf DJ, Ruiz-Cooley RI, Elliger C, Lebaric Z, Campos B, Markaida U, et al. Ommastrephid squids *Sthenoteuthis oualaniensis* and *Dosidicus gigas* in the eastern Pacific show convergent biogeographic breaks but contrasting population structures. Mar Ecol Prog Ser 2010;418:165–78.
[27] Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31:3210–2.
[28] Li F, Bian L, Ge J, Han F, Liu Z, Li X, et al. Chromosome-level genome assembly of the East Asian common octopus (*Octopus sinensis*) using PacBio sequencing and Hi-C technology. Mol Ecol Resour 2020;20:1572–82.
[29] da Fonseca RR, Couto A, Machado AM, Brejova B, Albertin CB, Silva F, et al. A draft genome sequence of the elusive giant squid, *Architeuthis dux*. Gigascience 2020;9:giz152.
[30] Chembian J, Mathew S. Population structure of the purpleback squid *Sthenoteuthis oualaniensis* (Lesson, 1830) along the southwest coast of India. Indian Journal of Fisheries 2014;61:20–8.
[31] Zhao C, Kang B, He X, Yan Y. Morphological, molecular, and ecological evidence in population determination and fishery management of purpleback flying squid *Sthenoteuthis oualaniensis* in the South China Sea. Taiwania 2021;66:241–50.

[32] de Jong WW, Hendriks W, Mulders JW, Bloemendal H. Evolution of eye lens crystallins: the stress connection. Trends Biochem Sci 1989;14:365–8.
[33] Tan WH, Cheng SC, Liu YT, Wu CG, Lin MH, Chen CC, et al. Structure of a highly active cephalopod *S*-crystallin mutant: new molecular evidence for evolution from an active enzyme into lens-refractive protein. Sci Rep 2016;6:31176.
[34] Madl T. Patchy proteins form a perfect lens. Science 2017;357:546–7.
[35] Chatterjee A, Cerna Sanchez JA, Yamauchi T, Taupin V, Couvrette J, Gorodetsky AA. Cephalopod-inspired optical engineering of human cells. Nat Commun 2020;11:2708.
[36] Cai J, Townsend JP, Dodson TC, Heiney PA, Sweeney AM. Eye patches: protein assembly of index-gradient squid lenses. Science 2017;357:564–9.
[37] Bro-Jorgensen J. Evolution of sprint speed in African savannah herbivores in relation to predation. Evolution 2013;67:3371–6.
[38] Bernstein BE, Hol WG. Crystal structures of substrates and products bound to the phosphoglycerate kinase active site reveal the catalytic mechanism. Biochemistry 1998;37:4429–36.
[39] Li X, Jiang Y, Meisenhelder J, Yang W, Hawke DH, Zheng Y, et al. Mitochondria-translocated PGK1 functions as a protein kinase to coordinate glycolysis and the TCA cycle in tumorigenesis. Mol Cell 2016;61:705–19.
[40] Urbina HD, Silberg JJ, Hoff KG, Vickery LE. Transfer of sulfur from IscS to IscU during Fe/S cluster assembly. J Biol Chem 2001;276:44521–6.
[41] Smid O, Horakova E, Vilimova V, Hrdy I, Cammack R, Horvath A, et al. Knock-downs of iron-sulfur cluster assembly proteins IscS and IscU down-regulate the active mitochondrion of procyclic *Trypanosoma brucei*. J Biol Chem 2006;281:28679–86.
[42] Angerer H, Radermacher M, Mankowska M, Steger M, Zwicker K, Heide H, et al. The LYR protein subunit NB4M/NDUFA6 of mitochondrial complex I anchors an acyl carrier protein and is essential for catalytic activity. Proc Natl Acad Sci U S A 2014;111:5207–12.
[43] Shi X, Zhang Y, Chen R, Gong Y, Zhang M, Guan R, et al. *ndufa7* plays a critical role in cardiac hypertrophy. J Cell Mol Med 2020;24:13151–62.
[44] Zhu J, Vinothkumar KR, Hirst J. Structure of mammalian respiratory complex I. Nature 2016;536:354–8.
[45] Munk O. The escal photophore of ceratioids (Pisces; Ceratioidei) – a review of structure and function. Acta Zoologica 1999;80:265–84.
[46] Belcaid M, Casaburi G, McAnulty Sarah J, Schmidbaur H, Suria Andrea M, Moriano-Gutierrez S, et al. Symbiotic organs shaped by distinct modes of genome evolution in cephalopods. Proc Natl Acad Sci U S A 2019;116:3030–5.
[47] Pankey MS, Minin VN, Imholte GC, Suchard MA, Oakley TH. Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid. Proc Natl Acad Sci U S A 2014;111:E4736–42.
[48] Philips KB, Kurtoglu M, Leung HJ, Liu H, Gao N, Lehrman MA, et al. Increased sensitivity to glucose starvation correlates with downregulation of glycogen phosphorylase isoform PYGB in tumor cell lines resistant to 2-deoxy-D-glucose. Cancer Chemother Pharmacol 2014;73:349–61.
[49] Burwinkel B, Hu B, Schroers A, Clemens PR, Moses SW, Shin YS, et al. Muscle glycogenosis with low phosphorylase kinase activity: mutations in *PHKA1*, *PHKG1* or six other candidate genes explain only a minority of cases. Eur J Hum Genet 2003;11:516–26.
[50] Muller A, Janssen F, Grieshaber MK. Putative reaction mechanism of heterologously expressed octopine dehydrogenase from the great scallop, *Pecten maximus* (L). FEBS J 2007;274:6329–39.
[51] Philipp EE, Wessels W, Gruber H, Strahl J, Wagner AE, Ernst IM, et al. Gene expression and physiological changes of different populations of the long-lived bivalve *Arctica islandica* under low oxygen conditions. PLoS One 2012;7:e44621.
[52] Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience 2012;1:18.
[53] Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods 2020;17:155–8.
[54] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
[55] Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 2014;9:e112963.
[56] Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 2017;356:92–5.
[57] Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 2016;3:95–8.
[58] Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst 2016;3:99–101.
[59] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;27:573–80.
[60] Saha S, Bridges S, Magbanua ZV, Peterson DG. Empirical comparison of *ab initio* repeat finding programs. Nucleic Acids Res 2008;36:2284–94.
[61] Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics 2009;Chapter 4:4.10.1–14.
[62] Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 2003;19:301–2.
[63] Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 2005;33:W465–7.
[64] Allen JE, Majoros WH, Pertea M, Salzberg SL. JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. Genome Biol 2006;7:S9.1–13.
[65] Alioto T, Blanco E, Parra G, Guigo R. Using geneid to identify genes. Curr Protoc Bioinformatics 2018;64:e56.
[66] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;10:421.
[67] Birney E, Clamp M, Durbin R. GeneWise and genomewise. Genome Res 2004;14:988–95.
[68] Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using evidence-modeler and the program to assemble spliced alignments. Genome Biol 2008;9:R7.
[69] Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 2014;30:1236–40.
[70] Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 2015;16:157.
[71] Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics 2017;33:2938–40.
[72] De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics 2006;22:1269–71.
[73] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30:772–880.

[74] Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 2009;25:1972–3.
[75] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–3.
[76] Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 2014;30:i541–8.
[77] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24:1586–91.
[78] Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res 2011;21:487–93.
[79] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res 2009;19:1639–45.
[80] Loytynoja A. Phylogeny-aware alignment with PRANK. Methods Mol Biol 2014;1079:155–70.
[81] Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 2007;56:564–77.
[82] Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 2015;10:845–58.
[83] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem 2004;25:1605–12.
[84] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34:i884–90.
[85] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 2015;12:357–60.
[86] 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;29:644–52.
[87] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 2013;8:1494–512.
[88] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9.
[89] Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protoc 2016;11:1650–67.
[90] Waskom ML. Seaborn: statistical data visualization. J Open Source Softw 2021;6:3021.
[91] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021;19:578–83.
[92] Chen M, Ma Y, Wu S, Zheng X, Kang H, Sang J, et al. Genome Warehouse: a public repository housing genome-scale data. Genomics Proteomics Bioinformatics 2021;19:584–9.