

# A high-quality genome of the dobsonfly *Neoneuromus ignobilis* reveals molecular convergences in aquatic insects

Xing-Zhou Ma <sup>a</sup>, Zi-Qi Wang <sup>a</sup>, Xi-Qian Ye <sup>a,b,c</sup>, Xing-Yue Liu <sup>d</sup>, Pu Tang <sup>a,b,c</sup>, Xingxing Shen <sup>a,b,c</sup>, Xue-Xin Chen <sup>a,b,c,\*</sup>

<sup>a</sup> Institute of Insect Science, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China

<sup>b</sup> Ministry of Agriculture Key Lab of Molecular Biology of Crop Pathogens and Insect Pests, Zhejiang University, Hangzhou 310058, China

<sup>c</sup> State Key Lab of Rice Biology, Zhejiang University, Hangzhou 310058, China

<sup>d</sup> Department of Entomology, China Agricultural University, Beijing 100193, China

## ARTICLE INFO

### Keywords:

Aquatic insects  
Genome  
Convergent evolution  
Amino acid substitution  
Megaloptera

## ABSTRACT

*Neoneuromus ignobilis* is an archaic holometabolous aquatic predatory insect. However, a lack of genomic resources hinders the use of whole genome sequencing to explore their genetic basis and molecular mechanisms for adaptive evolution. Here, we provided a high-contiguity, chromosome-level genome assembly of *N. ignobilis* using high coverage Nanopore and PacBio reads with the Hi-C technique. The final assembly is 480.67 MB in size, containing 12 telomere-ended pseudochromosomes with only 17 gaps. We compared 42 hexapod species genomes including six independent lineages comprising 11 aquatic insects, and found convergent expansions of long wavelength-sensitive and blue-sensitive opsins, thermal stress response TRP channels, and sulfotransferases in aquatic insects, which may be related to their aquatic adaptation. We also detected strong nonrandom signals of convergent amino acid substitutions in aquatic insects. Collectively, our comparative genomic analysis revealed the evidence of molecular convergences in aquatic insects during both gene family evolution and convergent amino acid substitutions.

## 1. Introduction

The adaptation of insects to freshwater ecosystems has clearly been successful. Inland waters cover <1% of the Earth's surface but harbour nearly 100,000 species, representing >6% of all insect species [15]. Aquatic insects spend one or more stages of their life cycle in water; in the majority of these species, the egg and larval stages occur in water, after which adults move to terrestrial habitats. These species play important ecological roles in aquatic and terrestrial realms as primary consumers, detritivores, predators, and pollinators [36]. All aquatic insect groups originated from the invasion of freshwater systems by terrestrial groups [108]. Aquatic insects belong to 12 orders and may represent >50 separate invasions. Ephemeroptera, Odonata, Plecoptera, Trichoptera, and Megaloptera are almost exclusively restricted to freshwater by an aquatic larval stage [15]. Due to their sensitivity to different pollutants, aquatic insects have been widely adopted as indicators for freshwater biomonitoring [7]. Additionally, some blood-feeding aquatic Diptera, such as mosquitoes and midges, act as vectors of many human diseases [35].

The secondary adaptation of terrestrial animals to water has long been a hot spot of convergent evolution studies. Many studies have been carried out to reveal the molecular basis of the convergent evolution of aquatic amniotes, especially marine mammals, including the investigation of gene family evolution [19,53,65,112], amino acid substitutions [22,70,114,116], and evolutionary rate shifts [11,79,100]. Similar to marine mammals, aquatic insects are characterized by a set of convergent adaptive traits, including gills, a dorsoventrally flattened body shape, unique cuticle characteristics, and ammoniotelic nitrogen excretion [27,104]. In contrast to marine vertebrates, most aquatic insects have solved the problem of breathing underwater, but adults must still mate and reproduce in the air. The earliest fossils (mayflies) of aquatic pterygotan insects come from the early Permian, when the first group of aquatic amniotic animals (*Mesosaurus*) arose, much earlier than the first appearance of marine mammals [30,83,91,108].

The possibility of molecular convergence decreases with evolutionary distance because of epistasis, which increases the divergence of the gene context and may cause the same mutation to produce different phenotypes [28,119,120]. It is still controversial whether nonrandom

\* Corresponding author at: Institute of Insect Science, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China.

E-mail address: [xxchen@zju.edu.cn](mailto:xxchen@zju.edu.cn) (X.-X. Chen).

molecular convergence can occur between distantly related species [72,115]. The aquatic invasions of most living insect groups occurred in the Permian and Triassic, earlier than the appearance of mammals but after the divergence of fish and tetrapods. Therefore, aquatic insects can provide good examples for studying convergent evolution in species separated by moderately long evolutionary distances. In contrast to aquatic amniotes, there are only a few gene-level studies on the molecular convergence of aquatic insects [75,90]. It was largely due to the paucity of high-quality genomes of aquatic insects [41]. Recent years have seen a series aquatic insect genome sequenced including chironomids [96], mayflies [2], damselflies [9], aquatic fireflies [20], and caddisflies [67], and genome-wide comparative analysis of the molecular convergence in aquatic insects have yet to be performed.

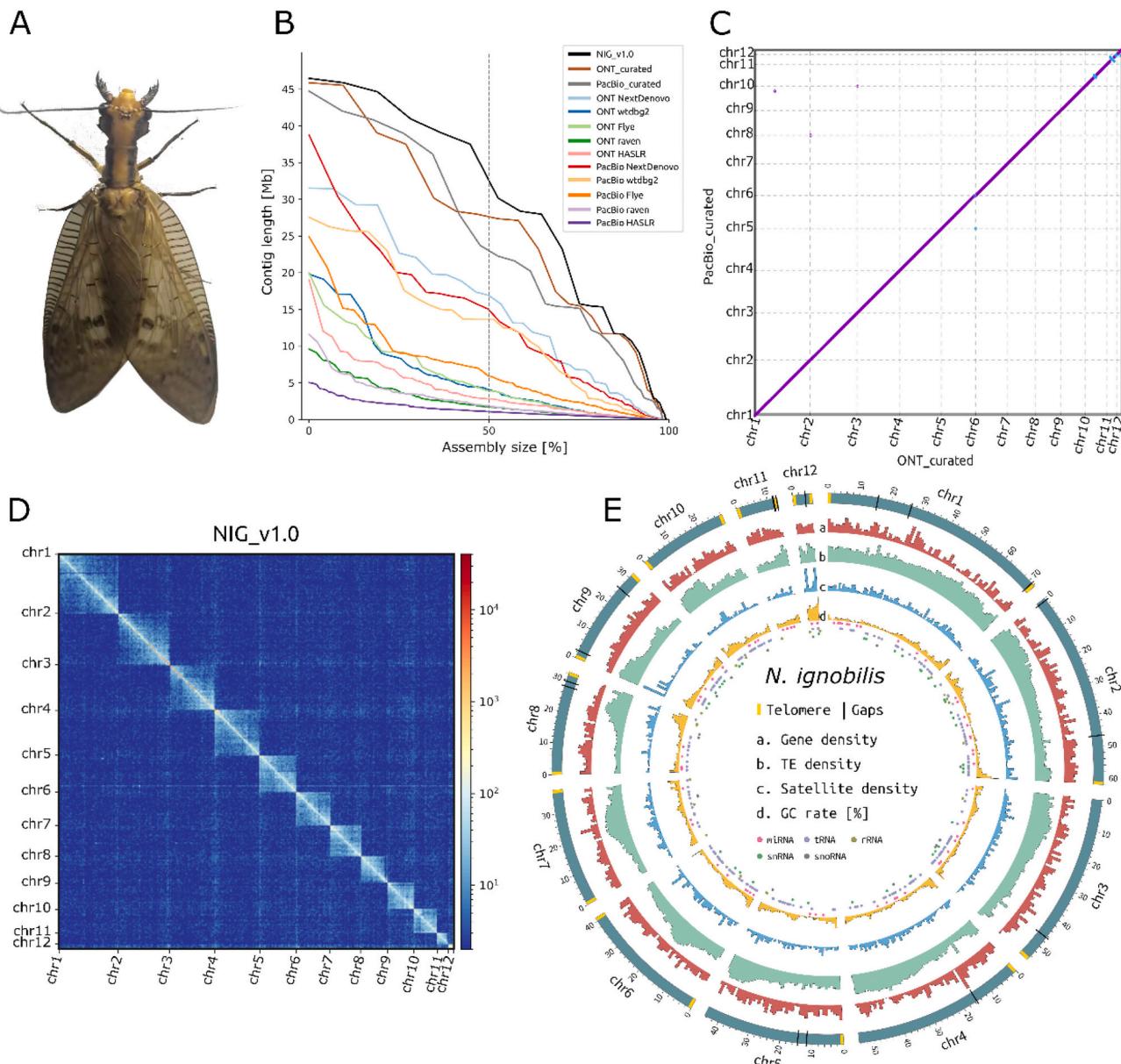
In this study, we conducted comparative genomic analyses using 42 hexapod species, including 11 aquatic insects representing six independent aquatic lineages, to study the molecular convergence of aquatic

insects. Due to the absence of a chromosome-level genome for the order Megaloptera, we first generated a high-quality genome for the Oriental dobsonfly *Neoneuromus ignobilis* Navás, 1932 (Corydalidae: Corydalinae) which is the most widespread megalopteran species in East Asia (Fig. 1A) [63,64,110]. Finally, we carried out analyses of gene family evolution and convergent amino acid substitutions to investigate convergent evolution in aquatic insects.

## 2. Results

### 2.1. High-contiguity genome assemble for *N. ignobilis*

We first compare the genome assemblies of the dobsonfly *N. ignobilis* using Oxford Nanopore Technologies (ONT) long reads, PacBio continuous long reads (CLRs) using five different assemble software, Flye, Wtdbg2, NextDenovo, HASLR, and Raven, to find the best assembly. It



**Fig. 1.** The genome assembly, annotation, and phylogeny of *N. ignobilis*. (A) The female adult of *N. ignobilis* (collected in Gutian Mountain, Zhejiang, China). (B) The contiguity of *N. ignobilis* assemblies produced by ONT and PacBio long reads and different software. The x-axis represents the cumulative length of contigs. The y-axis represents the size of the individual contig. (C) Synteny plot of the whole genome alignment of curated ONT (ONT\_curated) and PacBio (Pacbio\_curated) assembly shows high consistency. (D) Hi-C heatmap of curated ONT and PacBio assembly. (E) Circos plot of *N. ignobilis* genome annotation.

yields ten assemblies. After short reads polishing, Hi-C technology scaffolding, all assemblies reached chromosome-level with 11 or 12 pseudochromosomes (Fig. S1). The contig N50 ranged from 1.12 MB in Pacbio\_Haslr to 17.38 MB in Ont\_Nextdenovo (Fig. 1B, Table S1). NextDenovo performs best for PacBio and ONT reads with the highest contig N50, mounting size (the size of genome mounted to pseudochromosomes), BUSCO completeness, and the lowest gap number.

The NextDenovo assemblies of ONT and PacBio reads were selected for manual curation and gap filling. We closed 146 gaps for ONT\_NextDenovo and 144 gaps for PacBio\_NextDenovo manually using corrected long reads. The curated version of Nanopore (ONT\_curated) and PacBio (PacBio\_curated) assemblies contained 23 and 26 gaps in pseudochromosome scaffolds with contig N50 values of 27.93 and 23.71 MB, respectively. The whole-genome alignment showed these two assemblies were highly consistent (Fig. 1C). We selected the ONT\_curated assembly as a template and filled another six gaps using PacBio\_curated assembly. The final assembly (NIG\_v1.0) was 480.67 MB (Contig/Scaffold N50 = 37.51/54.40 MB), and contained 12 highly continuous scaffolds representing pseudochromosomes with only 17 gaps, of which three scaffolds representing chr6, chr7, and chr10 were telomere to telomere with no gap inside. Tandem repeats with identical sequences surrounded all the remaining gaps, suggesting the read length was the limitation for further continuity. Telomeric (TGACC)n repeats appeared at one or both ends of all the 12 pseudochromosomes, and eight of them had telomeres at both ends (Table S2). The chromosome ends without telomeric repeats were terminated with tandem repeats. Hi-C map showed no obvious mis-assembly in our final assembly (Fig. 1D). The BUSCO completeness of the assembly was 98.9%. Our final assembly of the nonmodel insect *N. ignobilis* presented comparable or even higher continuity and BUSCO completeness than traditional model insects (Table 1).

In summary, we found NextDenovo performs best in both PacBio and Nanopore reads for insect genome assembly. The continuity of the assembly was elevated after manual curation and gap filling.

## 2.2. Genome annotation and phylogeny

Using RepeatMasker, a total of 175.0 MB of the *N. ignobilis* genome was annotated as repetitive regions, accounting for 36.41% of the assembly. The transposable element (TE) distribution was uneven in *N. ignobilis* chromosomes (Fig. 1E). A higher TE density can be detected in the middle of some chromosomes, corresponding to the centromere region. A total of 14,263 protein-coding genes (PCGs) were found with 18,873 transcripts, 11,073 3'-UTRs, and 11,141 5'-UTRs. The BUSCO completeness of our PCGs was 99.0%, suggesting that our PCG annotation was in high-quality. We also identified 95 microRNAs based on a small RNA-seq library and homologous searches.

**Table 1**  
Genome assembly of *N. ignobilis* and traditional model insects.

Features	<i>A. mellifera</i>	<i>T. castaneum</i>	<i>B. mori</i>	<i>D. melanogaster</i>	<i>N. ignobilis</i>
Version	Amel_HAv3.1	Tcas5.2	Bmori_2016v1.0	dmel_r6.39	NIG_v1.0
Order	Hymenoptera	Coleoptera	Lepidoptera	Diptera	Megaloptera
Sequencing approach	PB + 10× + IL + Hi-C + Bionano	IL + Bionano	PB + IL + BAC + formid	sanger + BAC + FISH mapping	ONT+ PB + IL + Hi-C
<sup>a</sup>					
Assembly size [Mb]	225.25	152.43	460.34	142.57	480.67
Contig N50 [Mb]	5.38	0.084	12.2	21.49	37.51
Scaffold N50 [Mb]	13.62	15.27	16.80	25.29	54.50
Chromosome number	16	10	28	5	12
GC content [%]	32.53	33.86	38.55	42.0	24.72
Gap number <sup>b</sup>	67	4341	30	277	17
BUSCO [%] c	99.5	99.2	98.0	99.9	98.9
Reference	Wallberg et al. 2019 [121]	Herndon et al. 2020 [123]	Kawamoto et al. 2019 [122]	Flybase 2021 [124]	this study

<sup>a</sup> IL, Illumina sequencing; 10×, 10× Genomics linked reads; HiC, high-throughput chromatin conformation capture; PB Pacbio long reads; ONT, Oxford Nanopore; BAC bacterial artificial chromosome.

<sup>b</sup> Gap number on pseudo-chromosome scaffold.

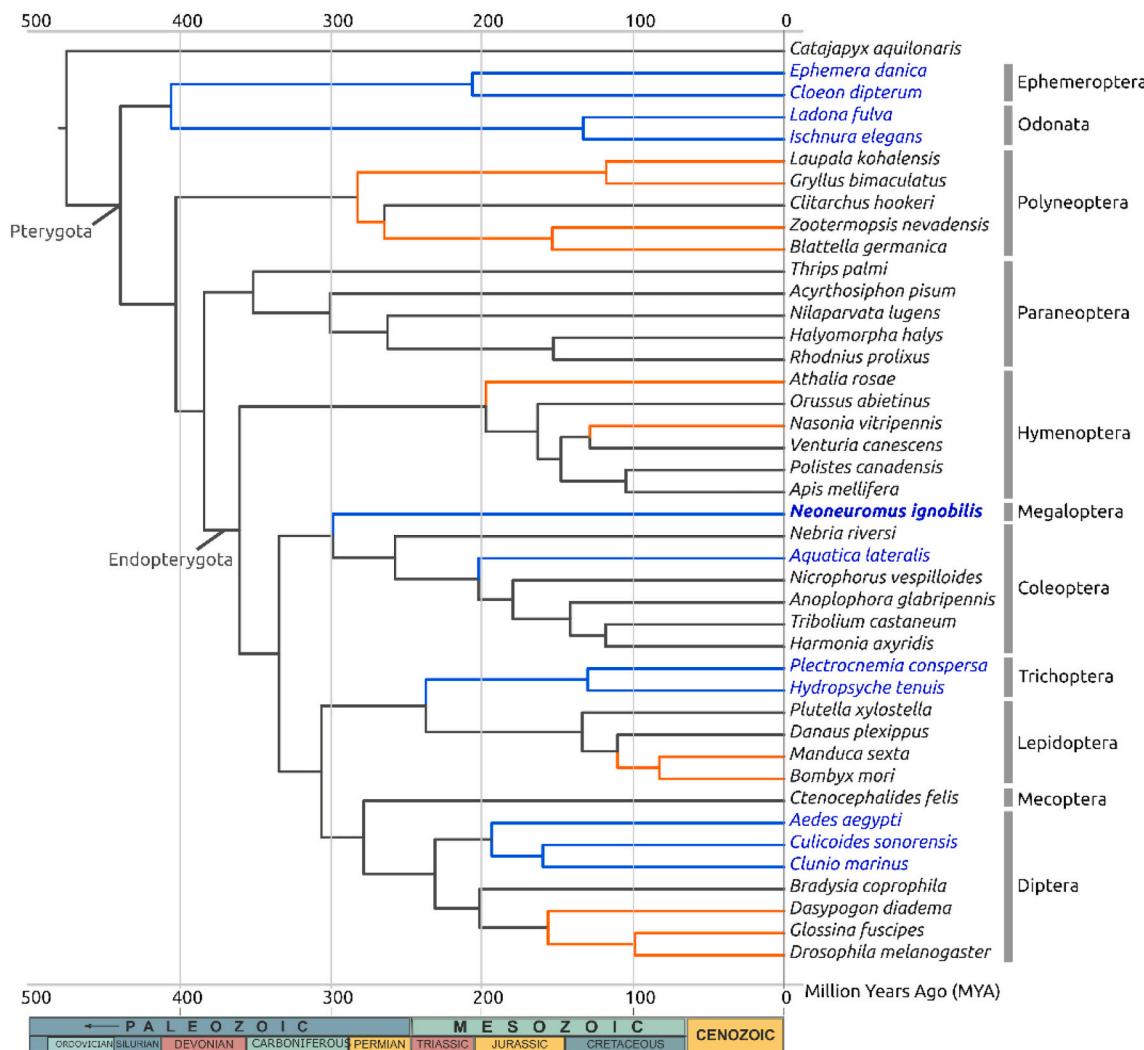
We selected 42 hexapods, including 11 aquatic insect species representing six independent aquatic insect lineages (Odonata, Ephemeroptera, Megaloptera, Trichoptera, aquatic firefly (Coleoptera, Lampyridae), and Callimorpha Diptera) for subsequent comparative genomic analyses (Table S3). Using Orthofinder, we obtained a total of 31,037 orthologous groups (OGs) present in more than two of the selected species. We selected 174 strictly “1:1:1” OGs with 45,997 AAs for phylogenetic tree construction. Our tree supported the sister-group relationship between Megaloptera and Coleoptera (Fig. 2).

## 2.3. Convergent evolution of gene families in aquatic insects

To identify the convergent expanded gene families of aquatic insects, we first compared the gene number in each orthologous group in the aquatic insect species with those in the species from the corresponding terrestrial sister groups. We defined an OG found in one aquatic species with a gene number >1.5 times the mean and median numbers in the corresponding terrestrial sister groups as an expansion, and we selected OGs that were expanded in more than three aquatic insect species as putative convergent expanded gene families, yielding 49 OGs (Supplementary File 1). We found that these OGs included serine peptides, chitin-binding domain-containing proteins, and the CPR cuticle protein family, which may reflect the changes in the digestive system and cuticle structure of aquatic insects. The expanded OGs also included opsins, transient receptor potential (TRP) channels, and sulfotransferases. We compared the evolution of these families between aquatic and terrestrial insects.

We identified opsin genes in all 42 species and constructed a phylogenetic tree (Fig. S2). The phylogenetic tree suggested that an expansion of long wavelength-sensitive (LWS) and blue-sensitive opsins (Fig. 3A). LWS opsin genes were expanded in Ephemeroptera, Odonata, and aquatic Diptera. Multiple copies of blue-sensitive opsins were found in the dragonfly *L. fulva*, the mayflies *E. danica*, and *C. dipterum*, and the blood-feeding midge *C. sonorensis*, while all of the terrestrial insects included in this study showed a single copy or zero copies. The expansion of opsins did not occur in all aquatic insects. Megaloptera and the aquatic firefly *A. lateralis* had only two LWS opsins and one LWS opsin, respectively, and blue-sensitive opsin was absent in both Megaloptera and Coleoptera, suggesting that the loss of blue-sensitive opsin occurred before the split of these groups. Selection pressure analysis suggested that LWS opsins presented significantly higher dN/dS ratios ( $\omega$ ) in Ephemeroptera, Odonata, and aquatic Diptera than in other insects (Table 2). The RELAX test showed no relaxed selection ( $P = 0.472$ ,  $k = 1.022$ ), and the aBSREL test identified three foreground branches under positive selection (Table S4), suggesting that LWS opsins in these three clades underwent positive rather than relaxed selection after expansion.

We constructed a phylogenetic tree of all of the TRP channel genes of



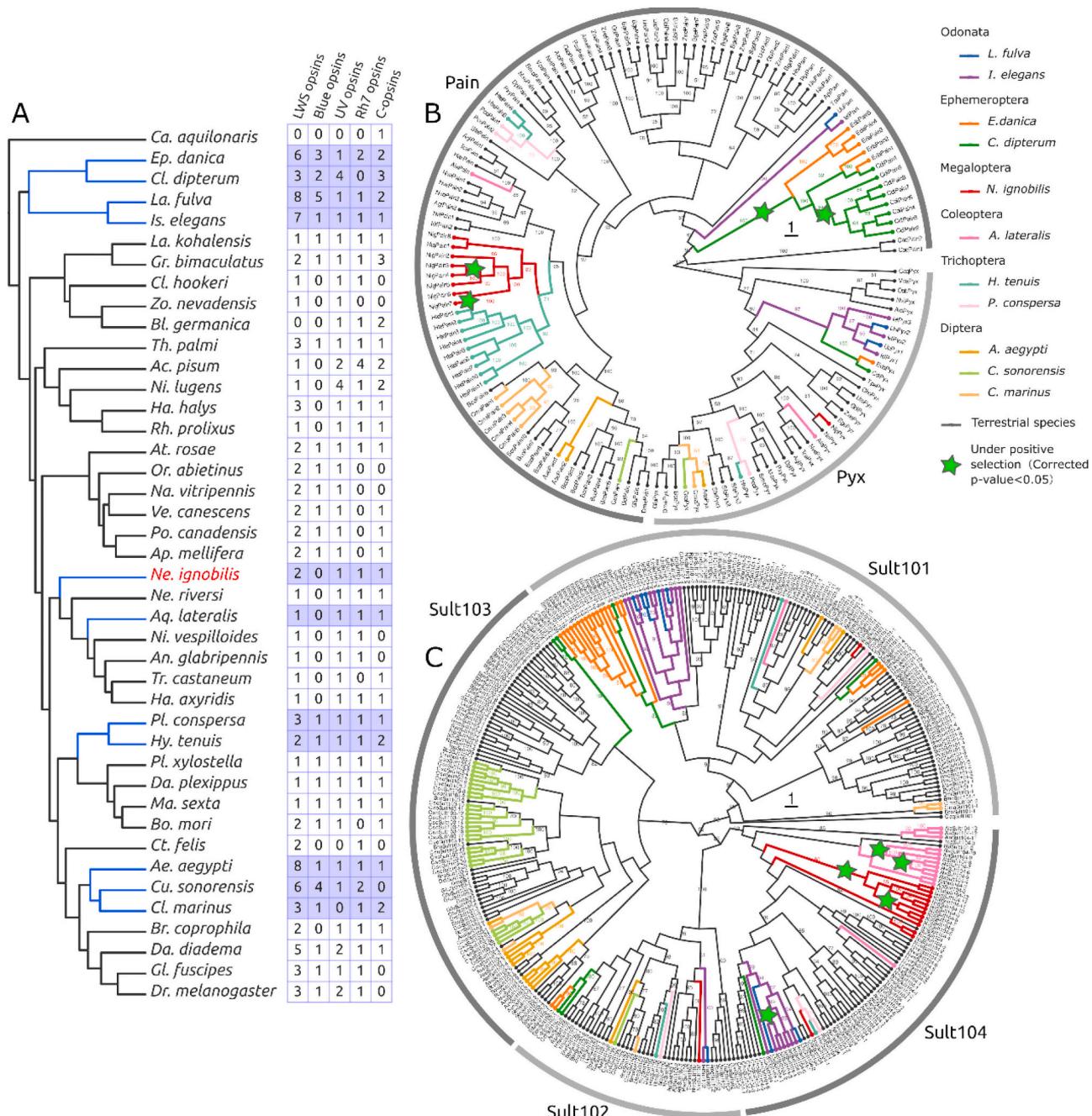
**Fig. 2.** Phylogenetic tree constructed by 174 strict “1:1:1” orthologous genes of 42 hexapods. Blue branches represent the aquatic insect clades. Orange branches represent the six control terrestrial insect clades. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

42 hexapods and confirmed that an expansion occurred in the clade orthologous to the *Drosophila* thermal stress response TRP gene *Painless* (*Pain*). A subtree consisting of the *Pain* and *Pyrexia* (*Pyr*, paralogous to *Pain*, used as outgroup) genes of 42 hexapods was used for subsequent analyses (Fig. 2B). *Pain* showed expansion in five aquatic insects, *N. ignobilis* (Megaloptera, 8 *Pain*), *H. tenuis* (Trichoptera, 11 *Pain*), the mayflies *E. danica* (5 *Pain*) and *C. dipterum* (8 *Pain*), and the midge *C. marinus* (5 *Pain*). Among terrestrial insects, most species had one or two *Pain* genes. The expansion of *Pain* among the nonaquatic insects occurred in Blattaria (*Z. nevadensis* and *B. germanica*) and the dipteran species *B. coprophila*. The tree of *Pain* demonstrated that most expansions in aquatic insects were lineage specific (Fig. 3B). Although Odonata had a single copy of *Pain*, two species, *L. fulva* and *I. elegans*, had two and three copies *Pyr*, respectively; *Pyr* is another important thermal response gene present in a single copy in most other insects. Selection pressure analysis suggested that the  $\omega$  value was significantly higher in five aquatic insects with *Pain* expansion (Table 2). The RELAX test suggested significant relaxed selection ( $P = 0$ ,  $k = 0.74924$ ), and the ABSREL test detected four foreground branches under positive selection (Fig. 3B, Table S4), indicating that both relaxed selection and positive selection acted on *Pain* genes after their expansion in aquatic insects.

A phylogenetic tree of sulfotransferases (SULTs), a phase II detoxification gene family, was also constructed for the 42 hexapods. The tree

formed four major clades, SULT101 ~ 104 (Fig. 3C), and the expansion of SULTs in different aquatic insects occurred in different clades. SULT101 was expanded in the mayfly *E. danica* and Odonata; SULT104 was expanded in *N. ignobilis*, the aquatic firefly *A. lateralis* and the damselfly *I. elegans*; and SULT103 was expanded in the mosquito *A. aegypti* and the blood-feeding midge *C. sonorensis*. By comparing the selection pressure on these three SULT clades, we discovered that the  $\omega$  value of SULT104 was significantly higher in expanded aquatic branches (Table 2). The RELAX test of SULT104 showed significant relaxed selection ( $P = 0$ ,  $k = 0.38792$ ), and the ABSREL test identified five foreground branches under positive selection (Fig. 3C, Table S4).

To determine whether chemosensory genes have degenerated in aquatic insects, as observed in many aquatic vertebrates, we compared the numbers of odourant-binding protein (OBP) and odourant receptor (OR) genes in aquatic and terrestrial insects. We identified OBP and OR genes in the 42 selected insects (Fig. S3) and found that OBP numbers were not significantly lower in aquatic insects ( $U$  test,  $P = 0.852$ ) than in terrestrial insects. A phylogenetic generalized least-squares regression (PGLS) analysis showed no negative correlation between gene numbers and aquatic habits (Table S5). However, OBP numbers were more variable in aquatic insects than in terrestrial insects ( $Std_{\text{aqu}} = 47.4$  vs.  $Std_{\text{ter}} = 20.2$ ). Odonata showed the smallest OBP family size across all insects, while blood-feeding Diptera and the aquatic firefly *A. lateralis* presented



**Fig. 3.** Convergent expanded gene families in aquatic insects. A. Gene number of five opsin clades in 42 insects. LWS opsin: long wave-sensitive opsin; UV opsin: ultraviolet-sensitive opsin; c-opsin: ciliary opsin. The blue branch and dark boxes represent aquatic species. B. Phylogenetic tree of *Painless* (*Pain*) and *Pyrexia* (*Pyx*). C. Phylogenetic tree of sulfotransferases (SULTs). Colored branch: aquatic insects; grey branch: terrestrial insects; star symbol: branches under positive selection. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the largest OBP family size. Mayflies and caddisflies showed similar OBP gene numbers to terrestrial insects. The dobsonfly *N. ignobilis* had fewer OBPs than terrestrial insects. Most *N. ignobilis* OBPs were more highly expressed in adults than in the aquatic stage of larvae (Fig. S4), indicating that they mainly function in the air-living stage. Regarding ORs, blood-feeding aquatic Diptera showed a larger OR family size than terrestrial insects. The gene number of ORs in aquatic insects excluding blood-feeding Diptera was significantly smaller than that in terrestrial insects (U test,  $P = 0.0056$ ). Similar to the findings regarding OBPs, Odonata presented the smallest OR family size among the studied insects. The dobsonfly *N. ignobilis* and the aquatic firefly *A. lateralis* exhibited 28 and 24 ORs, respectively, fewer than were found in

terrestrial Coleoptera.

#### 2.4. Convergent amino acid substitution in aquatic insects

We selected single-copy orthologous genes to search for convergent amino acid (AA) substitutions in aquatic insects. An OG dataset of single-copy genes was constructed and included 2513 OGS and 894,232 aligned amino acid sites from the 42 hexapod species.

Based on the ancestral reconstruction, a total of 42,401 convergent amino acid substitution sites in 2463 genes were identified in at least two aquatic insect clades, accounting for 4.7% of all amino acid sites. We used the simulation method to evaluate noise. A total of 30 simulations

**Table 2**

Selection pressure analysis of LWS opsin, Pain and SULTs using Branch model in codeml.

Gene	model	Foreground branches	$\omega_0$	$\omega_1$	LR	P
LWS opsin	M0	–	0.076	0.076	–	–
	M2	Ephemeroptera, Odonata & aquatic Diptera	0.060	0.113	155.322	0.000
Pain	M0	–	0.115	0.115	–	–
	M2	<i>N. ignobilis</i> , <i>H. tenuis</i> , <i>E. danica</i> , <i>C. dipterum</i> & <i>C. marinus</i>	0.051	0.226	999.536	0.000
SULT101	M0	–	0.093	0.093	–	–
	M2	<i>E. Danica</i> & Odonata	0.091	0.104	3.327	0.068
SULT102	M0	–	0.040	-0.040	–	–
	M2	<i>A. aegypti</i> & <i>C. sonorensis</i>	0.043	0.028	0.5100	0.475
SULT104	M0	–	0.143	0.143	–	–
	M2	<i>N. ignobilis</i> , <i>A. lateralis</i> & <i>I. elegans</i>	0.123	0.191	38.004	7.060e-10

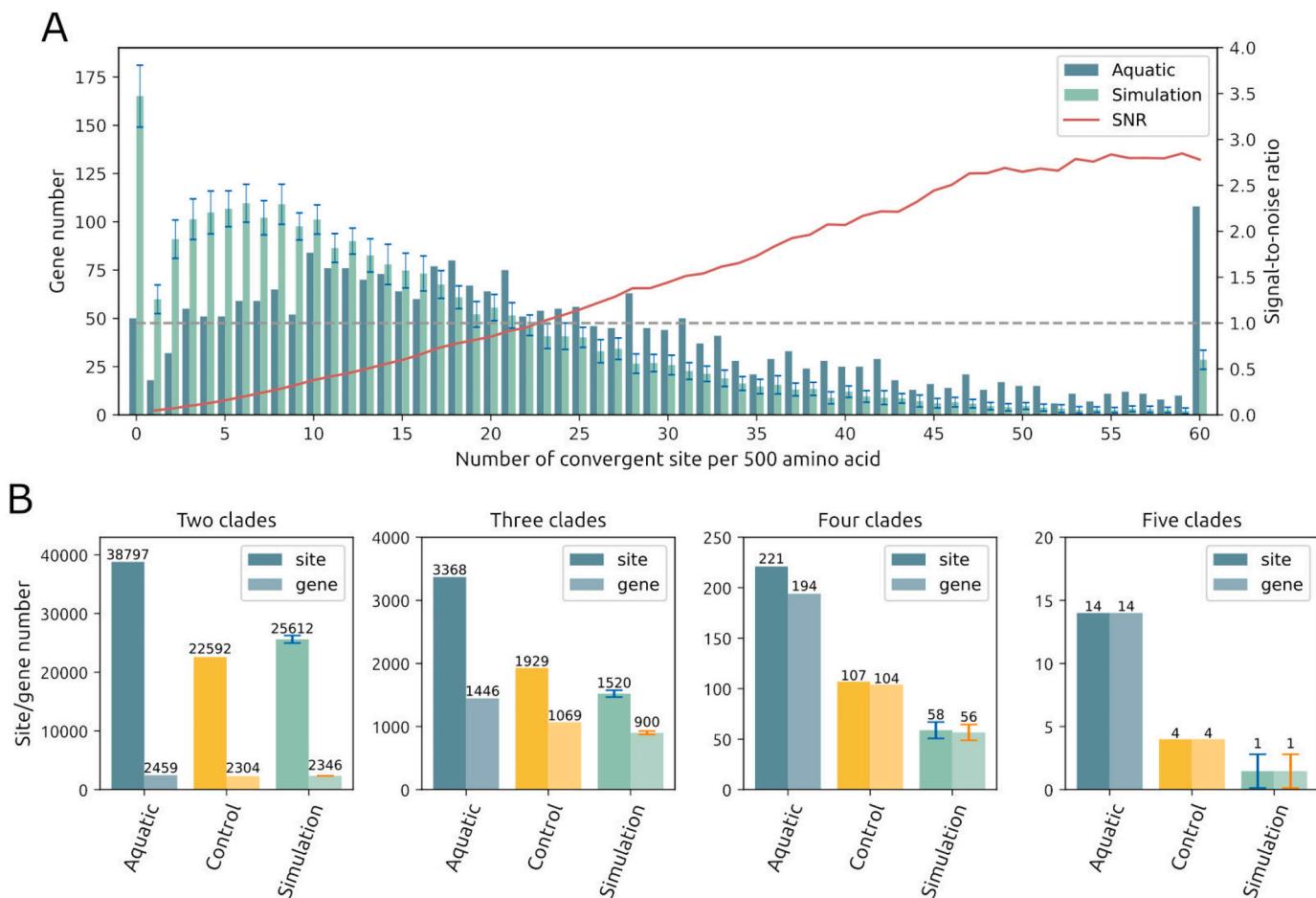
Foreground branches: The branches belong to listed species or taxon were selected as foreground, while the rest branches were selected as background. Model M0: one ratio model, assumes one  $\omega$  value for all branches; M2: two ratio branch model, allows foreground and background branch to have different  $\omega$  value;  $\omega_1$ : dN/dS ratios of foreground branches.  $\omega_0$ : dN/dS ratios of background branches. LR: likelihood ratio.

were performed. We defined noise (N) as the mean number of convergent sites in the 30 simulations and the signal (S) as the number of convergent sites in the aquatic branch (Fig. 2) minus the noise. The signal-to-noise ratio (SNR) can be estimated as (A-N)/N. We found that the average number of convergent sites in the simulations was 27,153.7, so the SNR of the convergent sites was 0.56, suggesting strong signals of nonrandom convergent substitution in aquatic insects.

We found that the SNR increased with the number of convergent sites per gene (Fig. 4A). Using the Poisson exact test, we identified 20 genes that contained significantly more convergent sites than were observed in

the simulations after FDR correction ( $q < 0.05$ , BH method, Supplementary File 2). GO enrichment analysis (Supplementary File 3) showed that these genes were enriched in the GO term response to stress (GO: 0006950,  $p = 0.01$ ), accounting for 8 of the 20 genes.

We found that the SNR increased with the number of clades in which convergent substitutions occurred (Fig. 4B). The SNR of the convergent sites that occurred in two clades was 0.51, which increased to 1.22 for sites found in three clades, 2.75 for sites found in four clades, and 8.55 for sites found in five clades. Only aquatic insects showed a convergent substitution site (in the glycosyltransferase gene Rumi) that occurred in



**Fig. 4.** Signal of convergent substitutions in aquatic insects. A. The number of genes corresponding to the convergent sites number (per 500 amino acids) in aquatic insect clades and simulation. Redline denotes the Signal-to-noise ratio, calculated by the genes having convergent sites exceeds the number at the x axes. The error bar represents the standard deviations (SDs) of 30 simulations. B. The number of convergent site/genes occurred in different clade numbers in aquatic insects, control and simulation. The error bar represents the SDs of 30 simulations.

all six clades. We selected six terrestrial insect clades as controls and predicted their convergent substitution sites (Fig. 1D). The number of convergent sites in aquatic insects was significantly higher than that in the terrestrial control (G-test, Table S6). We noted that the number of sites that were convergent in more than three clades was also higher in the control group of terrestrial insects than in the simulations (Fig. 4B), suggesting that in addition to random noise, the convergence of uninteresting traits could account for some portion of the adaptive convergent signals. Our results suggested that aquatic insects present strong signals of amino acid convergence despite being distantly evolutionarily related.

### 2.5. Significant convergent profile change sites in aquatic insects

To obtain the convergent substitution site more relevant to the aquatic adaptive evolution, we searched the convergent site in more species using NCBI Transcriptome Shotgun Assembly (TSA) data and analysed their amino acid frequency. Here, we defined the amino acid profile as the site's amino acid frequency in a clade. We selected the convergent site whose amino acid profile in the clades under convergent evolution was convergent and significantly different (Fisher's exact test,  $P < 0.05$ ) from the non-convergent sister group clades. These sites underwent dramatic shifts in amino acid preferences during environmental change, and we referred to these sites as significant convergent profile change (SCPC) sites. We chose Megaloptera and Trichoptera for study and compared them with their terrestrial sister groups, terrestrial Neuroptera and nonditrysian Lepidoptera (Table S7). We also required the sites located in genes expressed at higher levels in the larvae of *N. ignobilis* to show an increased relationship with aquatic adaptation.

We identified 62 SCPC sites in 51 genes (Table S8). The enriched GO terms (Supplementary File 4) of the 51 genes harbouring SCPC sites included endoplasmic reticulum (GO:0005783,  $P = 0.00172$ ), endo-membrane system (GO:0012505,  $P = 0.00120$ ), and glycoprotein metabolic process (GO:0009100,  $P = 0.02334$ ). Genes annotated to the endoplasmic reticulum GO term included *Rumi* and *Botv* (glycosylation), *UFL1* (ufmylation), *PDIA3* (disulfide bond formation), and *PIGT* and *PGAP2* (phosphatidylinositol glycan anchor biosynthesis), which encode proteins that function as covalent modifiers of target proteins. To our surprise, only ten convergent sites in seven genes, *Nostrin*, *Botv*, *Mybbp1a*, *Pwp2*, *Pex1*, *Srp72*, and *Ebna1bp2*, were strict convergent in two aquatic clades with the derived amino acid type present in all the Megaloptera and Trichoptera species but in none of the sister group terrestrial species (Table 3).

Interestingly, we found that the conservation scores of these 62 SCPC sites were higher in the two terrestrial sister groups than in Megaloptera and Trichoptera (Fig. 5). Additionally, the convergent sites were less conserved than their surrounding sites in the terrestrial sister groups (Fig. 5). This pattern also existed in the 42 hexapod species (Fig. S5).

**Table 3**  
The strict convergent sites between Megaloptera and Trichoptera.

Gene	Position <sup>a</sup>	Amino acid substitution	Convergent clades
<i>Pwp2</i>	678	E678K	Megaloptera & Trichoptera
<i>Pex1</i>	893	L893F	Megaloptera & Trichoptera
<i>Srp72</i>	222	R222K	Megaloptera & Trichoptera
<i>Botv</i>	562	T/V562I	Megaloptera & Trichoptera & Aquatic fireflies
<i>Ebna1bp2</i>	100	K100A	Megaloptera & Trichoptera
<i>Nostrin</i>	465	V/S465A	Megaloptera & Trichoptera
<i>Nostrin</i>	678	M/V678L	Megaloptera & Trichoptera
<i>Nostrin</i>	524	K524R	Megaloptera & Trichoptera & Ephemeroptera & Aquatic fireflies
<i>Mybbp1a</i>	125	D125E	Megaloptera & Trichoptera
<i>Mybbp1a</i>	192	V192L	Megaloptera & Trichoptera & Aquatic fireflies

<sup>a</sup> The position was based on longest peptide of the genes in *N. ignobilis*.

These results suggested that many convergent sites in aquatic insects may result from the convergent selection pressure increasing at sub-conserved sites.

## 3. Discussion

### 3.1. Near-complete genomes for nonmodel insects

This study provides a high-quality genome for a nonmodel aquatic insect (order Megaloptera), the oriental dobsonfly *N. ignobilis*. Previous study suggest manual curation can significantly improve the quality of the genome assembly [42,80,85,113]. Our final version of the nanopore assembly after manual curation included only 17 gaps on 12 telomere-ended pseudochromosomes, thus exceeding the quality of the assemblies of most current nonmodel insects. In our study, Nanopore technology produced greater assembly continuity but a slightly lower BUSCO value than PacBio technology. Because Nanopore sequencing is lower in cost, when combined with the Hi-C technique and manual curation, it could be a promising method for generating additional high-quality genomes for nonmodel insects. Also, using newly developed R10.3 flow cells and the Bonito base-calling method, the accuracy of Nanopore assemblies can be improved in future studies.

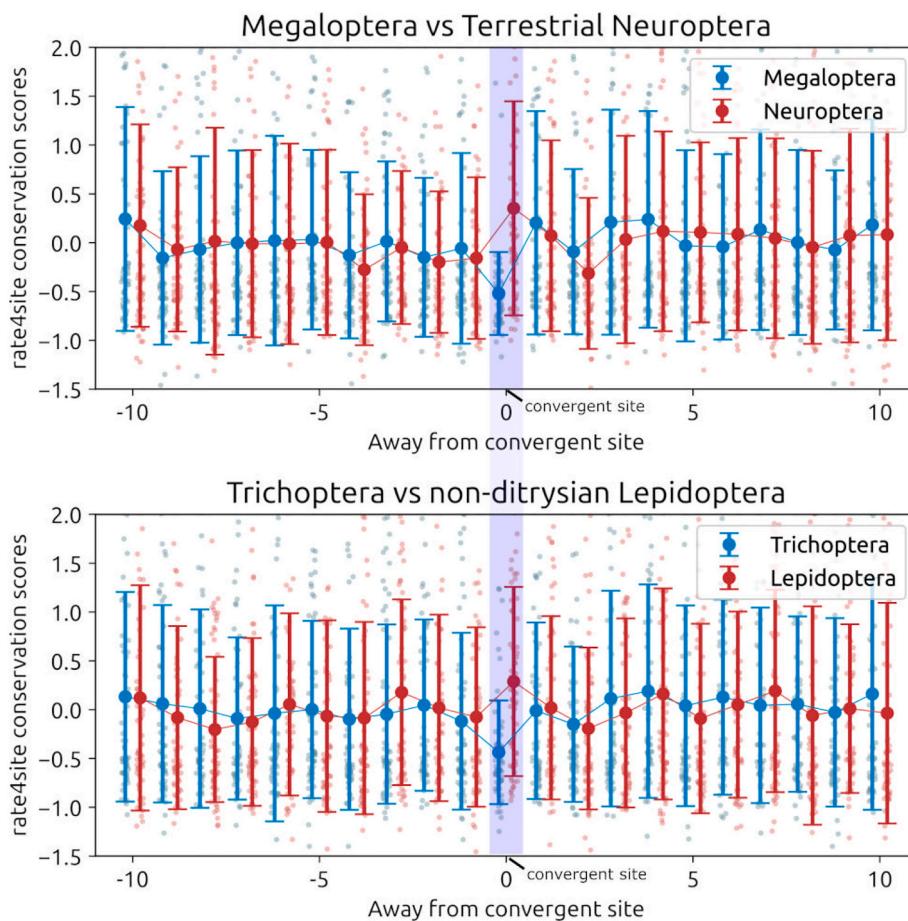
### 3.2. Convergent evolution of gene families in aquatic insects

We identified the convergent expansion of LWS and blue opsins, the thermal stress response TRP channel *Pain*, and sulfotransferases in aquatic insects. The observed expansions of opsins in Odonata, Ephemeroptera, and mosquitos are consistent with previous reports [2,25,37]. Among aquatic insects, the adults of these species (Megaloptera and aquatic firefly) are mainly nocturnal and show no expansion of opsin genes. A previous study suggested that many expanded opsin genes in Odonata were more highly expressed in adults. The habits of adults probably drove the convergent expansion of opsin genes. Aquatic insect adults can detect polarized light reflected by water to find mating and oviposition sites [106]. Additionally, when aquatic insects fly above the water, they must distinguish the light reflected from objects (such as prey and predator) and water, which requires good vision to detect different types of light.

The transition to freshwater requires the adaptation of thermal regulation mechanisms [15]. We identified the convergent expansion of *Pain* genes, encoding thermal stress response TRP channels, followed by positive selection in aquatic insects. *Pain* encodes an important thermosensor that helps insects sense high-temperature stress ([99], [109]). Because the thermal conductivity of water is higher than that of air, the dissolved oxygen concentration of water decreases with increasing temperature. The expansion of thermal stress response TRP channels may increase the ability of aquatic insects to detect and avoid high-temperature stress.

We identified the convergent expansion of sulfotransferase (SULT) genes in aquatic insects. Sulfotransferases are phase II detoxification enzymes encoded by a supergene family that catalyse the sulfonate conjugation of xenobiotics [95]. A previous study showed that a large proportion of 3-trifluoromethyl-4-nitrophenol was conjugated into sulfated derivatives in the aquatic midge *Chironomus tentans* [49]. The expansion of SULT may reflect a less-well-known detoxification mechanism in aquatic insects. Interestingly, our analysis failed to identify a significant change in selection pressure on SULT101 and SULT102 after expansion, suggesting that their expansion in aquatic insects may be related to a dosage increase, rather than functional innovation.

A reduction in chemosensory genes that occurred in parallel with secondary aquatic invasion has been identified in many vertebrate clades, such as platypuses, marine mammals, and sea snakes [45,53,54,65,117]. We did not identify a reduction in OBPs (a major chemosensory gene family in insects) in all aquatic insect clades. Unlike marine mammals, aquatic insects still need to mate and lay eggs in



**Fig. 5.** Rate4site conservation score of significant profile change sites in Megaloptera, terrestrial Neuroptera, Trichoptera, and non-Ditrysian Lepidoptera.

terrestrial environments. Highly specialized antennae for receiving chemical signals, such as ramified antennae, appear in many aquatic insect lineages, including dobsonflies and caddisflies [26]. Aquatic life may reduce the minimum number of chemosensory genes (those required in larvae), as our study showed that the insects with the smallest OBP and OR repertoires belonged to Odonata. The number of chemosensory genes in aquatic insects may depend more heavily on their adult lifestyle. The small numbers of OBPs and ORs in Odonata are consistent with their degenerated adult antennae, as dragonflies and damselflies rely more strongly on their highly developed vision for predation [14]. Blood-feeding Diptera show the greatest numbers of OBPs and ORs for host and spawning site searching observed among insects [14].

### 3.3. Convergent amino acid substitutions in aquatic insects

Based on the ancestral reconstruction of orthologous genes in 42 hexapod species, we identified strong signals of nonrandom amino acid convergence in aquatic insects, despite the distant evolutionary relationships of these species. We found the genes harbouring SCPC sites, those involved in covalent protein modification were enriched in the endoplasmic reticulum GO term. These convergent substitutions may affect protein modification in aquatic environments. One example is the Notch glycosyltransferase *Rumi*. In insects, *Rumi* glycosylates epidermal growth factor (EGF) repeats in the extracellular domain of Notch [1,97]. The convergent substitution in *Rumi* may alter the conformation of the Rumi-EGF binding site, thus changing the glycosylation efficiency of *Rumi* to adapt to the aquatic environment.

Interestingly, we found an increase in site conservation at convergent sites in the species under convergent evolution. This pattern is

associated with selective sweeps in microevolution, whereby a pre-existing beneficial mutation increases in frequency and becomes fixed. Recent studies have suggested that selective sweeps of standing variation play a vital role in repeated adaptive evolution, such as the freshwater adaptation of stickleback fish [12,46,93]. In macroevolution, assuming that a pre-existing AA type in an ancestral species is nearly neutral but is less affected by epistasis during evolution, this AA type may be convergently fixed in distantly related species by convergent selective sweeps, leading to the pattern of convergent substitution in aquatic insects.

In summary, our study provided a high-quality genome for the dobsonfly *Neoneuromus ignobilis*. Our comparative genomic analysis revealed the evidence of molecular convergences in aquatic insects during both gene family evolution and convergent amino acid substitution, and these molecular convergences may contribute to the ecological adaptation of aquatic insects.

## 4. Materials and methods

### 4.1. Samples collection and sequencing

For genome sequencing, *N. ignobilis* samples used for genome sequencing were collected in Gutian Mountain (Zhejiang, China) from 2019 to 2021. One adult female was used for PacBio CLR sequencing, and the abdomen was removed before DNA extraction to reduce the microbe-contaminations. DNA extraction, library construction, and sequencing were conducted by Novogene (Beijing, China). A total of 42.3 Gb PacBio CLR subreads ( $87 \times$ , reads length N50 = 16,964 nt) were produced. Short reads (42.4 Gb,  $87 \times$ ) were also sequenced from the same individual for polishing and genome survey using the Illumina

Hiseq platform. Another female was kept in the laboratory until egg laying. The eggs were incubated at 25 °C, and the newly hatched larvae were sent to NextOmics Biosciences (Wuhan, China) for nanopore, Illumina, and Hi-C sequencing. A total of 54.3 Gb ultra-long Oxford nanopore (ONT) reads (112 $\times$ , reads length N50 = 43,487 nt) were generated by a PromethION sequencer by R9.4 flow cells. The base-calling was performed by Guppy. Hi-C library (digested by DpnII) was sequencing via the Illumina Novaseq, and a total of 39.9Gb (82 $\times$ ) Hi-C reads were produced. Short reads (35.2 Gb 75 $\times$ ) of newly hatched larvae were also generated from an Illumina Novaseq platform for polishing.

A whole female adult collected in Tianmu Mountain (Zhejiang, China) during 2020 was used for transcriptome sequencing. One full-length cDNA library was constructed and sequenced by a PacBio Sequel platform. A total of 20.87Gb long polymerase reads were produced. One RNA-seq and one small RNA-seq library were sequenced by the Illumina platform, obtained 6.56 Gb short RNA-seq reads and 11,324,628 small RNA reads. Novogene (Beijing, China) carried out the RNA extraction, library construction, and sequencing of these libraries. In addition, two RNA-seq libraries from newly hatched larvae and one large size larva (5 cm in length, collected in Gutian Mountain, identified using COI) were sequenced from whole body using MGI-SEQ 2000 platform, producing 19.76 Gb and 17.08 Gb short RNA-seq reads. RNA extraction, library construction, and sequencing were conducted by NextOmics Biosciences (Wuhan, China).

#### 4.2. Genome assemblies

We first obtain the mitochondrial genome of *N. ignobilis*. We downloaded the mitochondrial genome of *Tribolium castaneum* from NCBI (NC 003081.2) as the reference. The PacBio CLR reads were aligned to the *T. castaneum* mitochondria using Minimap2 v2.1 [62]. The aligned reads were self-corrected using Canu v2.1.1 [56]. The corresponding mitochondrial reads were manually assembled into a circular mitochondrion and polished using Nextpolish v1.3.1 by Illumina short reads [43].

To assemble the nuclear genome of *N. ignobilis*, we started with two types of long reads. The PacBio and ONT reads were assembled separately. Five assembly software, Wtdbg2 v2.5 [86], Flye v2.8.3 [55], NextDenovo v2.3 (<https://github.com/Nextomics/NextDenovo>), Raven v1.5 [101], and HASLR v0.8 [34], were used for genome assembly with the default parameters. We obtained ten assemblies. Because mitochondria can cause over-polishing to the nuclear mitochondrial (NUMT) region, it is important to include the mitochondrial genome before genome polishing [42,85]. The mitochondrial sequence was tandemly replicated twice and added as a single contig to the end of each assembly. We performed two rounds of long reads correction using Racon v1.4.2 [102] and one round of short reads correction using Nextpolish. The mitochondrial contig was discarded after polish finished. One round of haplotig removal was carried out using Purge\_dups v1.2.5 [31]. Next, we followed the 3D de novo assembly (3D-DNA) pipeline [16] and assembled each draft into a candidate chromosome-length assembly. Juicebox v1.22 viewed the HiC-map of each assembly [17]. Benchmarking Universal Single-Copy Orthologs (BUSCOs v4.1.4) [87] with Insecta\_Odb10.2020-09-10 was used to evaluate the completeness of the assemblies.

We selected the NextDenovo assemblies of PacBio and ONT reads for subsequent manual curation because it has the highest contig N50, mounting size, BUSCO completeness, and the lowest gap number for both reads type (Table S1). The manual curation pipeline includes (1) Fixing mis-assembly and mis-orientation using Juicebox based on Hi-C map, (2) Adjusting local mis-orientation contigs by comparing raw draft contigs, (3) Filling gaps on Hi-C produced scaffolds by aligning raw draft contigs and corrected-long reads to the ends of the gaps and find single or overlap contig/reads that cover the gap, and (4) Align correct long reads to the ends of each chromosome to discover telomere.

For gap-filling and telomere discovery, the error-prone long reads were corrected first by Fmlrc2 [103] with short reads and then self-corrected using Canu v2.1.1. The corrected PacBio reads >30 KB and ONT reads >50 KB were selected. The non-repeat region of surrounding sequences of both sides of the gap were used as a query to search the raw draft assembly and the corrected long reads using blastn and MUMmer v3.5 [59]. The ends of super scaffolds were queried to the corrected reads to find reads containing telomeric repeat. We used mummerplot v3.5 [59] to perform whole-genome alignment on the two cured assemblies to check their consistency. Finally, we used the curated Nanopore assembly (ONT\_curated) as template and filled the remaining gaps using the Pacbio assembly. Hi-C heatmap of our final assemblies (NIG\_v1.0) was plotted and visualized by hicPlotMatrix v3.6 [107].

#### 4.3. Genome annotation

We first masked tandem repeats using tandem repeat finder [5]. Extensive de-novo TE Annotator (EDTA) v1.9.4 [78] and RepeatModeler v2.02 [21] were used to de-novo predict transposable elements (TEs). The results were combined with insect transposons in Dfam3.2, and Petersen's insert TE repertoire [82], and cd-hit v4.8.1 [24] were used to remove the redundancy. This data set was then used as a library to search the TEs on *N. ignobilis* using RepeatMasker v4.1.2 [98].

We integrated the gene models generated from transcript alignment, homologous searching, and *ab initio* prediction for annotation protein-coding gene (PCG). For transcript evidence, we ran Iso-Seq analysis in Smrtlink v10.1 for PacBio Iso-Seq subreads to generate full-length cDNAs. The cDNAs were mapped to the genome using minimap2, and Stringtie v2.7.1 [81] predicted the transcript structure. The RNA-Seq short reads were used to predict transcript structure using hisat2 v2.2.1 [52] and Stringtie. The transcript structures of long and short reads were integrated using Pasa v2.3.3 [32]. For homologous searching, the annotation of *D. melanogaster*, *A. pisum*, *A. mellifera*, *T. castaneum*, and *B. mori* from the NCBI Refseq database as queries. Exonerate v2.4.0 [88], GenomeThreader v1.7.3 [29], and GeMoMa v1.7 [50] were used to find the homologous gene model in *N. ignobilis*. For *ab initio* prediction, the high-quality transcripts were used to generate gene models for training and testing, followed by Hoff's pipeline using the scripts in PASA and Augustus [39]. We used Augustus v3.3.2 [92], GlimmerHMM [69], Genemark-ES v4.57 [66], and Snap v2006-07-28 [57] to predict the *ab initio* gene model after training from high-quality gene models. Finally, the gene models from transcript alignment, homologous searching, and *ab initio* prediction were integrated to generate final PCG annotation using EVM v1.1.1 [33]. The isoforms, 5'-UTRs, and 3'-UTRs were annotated using Pasa. The completeness of the annotation was assessed using BUSCO. The function of PCGs was annotated using Eggnog v5.0.2 [44]. The domains of peptides were annotated using interproscan v5.52 [47]. The expression levels of genes in *N. ignobilis* adult and larvae were estimated using RSEM v1.3.3 [61].

For non-coding RNA annotation, tRNA were predicted by tRNAscan-SE v2.0.5 [8], and rRNA, snRNA, and snoRNA were searched using cmscan v1.1.3 [76]. The small RNA-seq library was used to search miRNAs using miRDeep2 v0.1.3 [23], combined with homologous searching identified miRNAs by blastn using the insect miRNA repertoires as query [68].

#### 4.4. Ortholog groups identification and phylogenetic analysis

We selected 42 species for the subsequent comparative genomic analyses (Table S3). The genome and annotation of the selected species were downloaded from the NCBI RefSeq database, BCM-HGSC, and Insectbase 2.0 [71]. The longest coding sequence was selected to represent a gene. Orthofinder v2.5.2 [18] was used to cluster all the genes of 42 species into OGs with default parameters.

The strict 1:1:1 OGs present in >70% of species were selected for phylogenetic tree construction. Mafft v7.48 [48] was used to carry out

the multi-sequence alignment for the peptide sequence of each orthologous group (`-localpair -maximal 1000`), followed by using Aliscore v02.2 and Alicut v2.31 [58] to remove the sites with weak phylogenetic signals. The aligned OGS were concatenated to form a supermatrix, and iqtree2 v2.1.3 [73] was used to construct the phylogenetic tree (each gene a partition, find the best model using the Model Finder Plus option; `-m MFP -B 1000 -Bnni`). We selected the divergence time inferred from Misof [74] as calibration points, and the divergent time was estimated using mcmctree program (WAG+F + G6, burnin = 50,000, sampfreq = 50, nsample = 550,000) in PAML (v4.9) package [111].

#### 4.5. Gene family analysis

We compared the gene number of OGS and selected the OGS whose gene number is 1.5 times higher in aquatic insects than the mean and median in their terrestrial sister group as expansions. We chose Polyneoptera to compare with Odonata and Ephemeroptera, terrestrial Coleoptera to compare with Megaloptera and *A. lateralis*, Lepidoptera to compare with Trichoptera, and the terrestrial Diptera to compare with aquatic Diptera. OGS expanded in more than three aquatic species were selected to find the convergent expanded gene family.

We identified opsin, TRP channel and SULT genes in 42 hexapods. The peptides of these families downloaded from Flybase (<http://flybase.org/>) were used as query and search the protein annotations of 42 species using blastp (e-value = 1e-20). The obtained target sequence was annotated using Interproscan. The sequence containing domains (Table S9) of respective gene families were selected. For each family, all 42 species' sequence was aligned using mafft (`--globalpair`), and we used iqtree2 phylogenetic tree construction (`-m LG + I + G4 + F -B 1000 -Bnni -bcor 0.99`). We divided SULTs family into four subfamilies SULT101 ~ 104. The nomination followed the previously reported SULT101A1 in *Spodoptera frugiperda* [6].

For OBP and OR identification, we ran Interproscan for all the genes in 42 species. The domain of ORs and OBPs (Table S9) were used to search these genes. PGLSs analysis was used to study the correlation between the gene number and aquatic or terrestrial habitats and was performed using the R package "capper" [77]. U test was performed using the "scipy" package in python.

#### 4.6. Selection pressure analysis for gene families

We set the branches belonging to aquatic insects that underwent gene expansion as foreground, and the rest branches as background. We used codeml in the PAML package and tested one ratio (M0) model, which assumes one  $\omega$  value for all branches, and two-ration branch model (M2), which allows the foreground and background branch to have different  $\omega$ . The likelihood ratio test (LRT) was used to compare M0 and M2 models. If M2 is significantly better fits the data ( $P < 0.05$ ) and  $\omega$  is higher in foreground branches, we subsequently tested the aBSREL and relaxed model in Hyphy v2.5.8 used the foreground branches as test [89,105]. For the aBSREL test, the branch with the corrected  $P$ -value  $<0.05$  was considered under positive selection.

#### 4.7. Data preparation for analysing convergent substitutions

We selected single-copy genes for convergent amino acid substitution analysis. OGS present in  $>32$  species and single copied in  $>90\%$  of the species were selected. If an orthologous group contained more than two gene copies in one species, we removed all the copies of this species and treated as missing in this species. Mafft was used to align each orthologous group. We removed the sites with  $<30$  no-gap amino acids. Rate4site v2.01 [84] was used to calculate the site conservation score. The sites with the conservation score  $> 2.5$  were removed. The species tree was pruned using the ete3 package in python by the present species to produce gene tree for each orthologous group. The aaml program in the PAML package with the JTT + F + 4G model to reconstruct the

ancestral state and estimate the branch length.

#### 4.8. Convergent substitution sites identification

We defined a "convergent substitution site" as a site independently changed from any other ancestral AA type to the same derived AA type in any two or more clades in our six aquatic insect clades—this definition included "parallel" and "convergent" sites defined by Storz [94]. The substitution predicted by the aaml program above was used to identify convergent substitution sites under this definition. Meanwhile, if one aquatic insect clade we selected contained more than one species, we required the convergent substitution site should have the convergent AA type in every species in this clade. We also selected six terrestrial insect clades showed in Fig. 1F as control. The convergent sites in control clades were identified using the same method. G-test was performed using the python model "scipy" to compare the convergent site number between control and aquatic clades.

#### 4.9. Molecular evolution simulation

We conducted simulations using the evolver program in the PAML package. For each orthologous group, the length and amino acid frequency of the reconstructed ancestral sequence, and the topology and branch length of gene tree were used. The substitution model and gamma categories for simulation were the same as the ancestral reconstruction (JTT + F + 4G). After the simulation, the gaps in the raw alignment were added to the alignment of the simulated sequences. We ran simulation 30 times for all the selected OGS. Aaml program was used for simulations for ancestral reconstruction. For each orthologous group, Poisson-test was carried out using the "poisson.estest" package in python to compare the convergent sites number between the test and the average number of simulations. FDR correction (BH-method) was performed using the python model "statsmodels."

#### 4.10. Search convergent site in transcriptome data

The sites convergent in Megaloptera and Trichoptera and on the genes with higher FPKM value in the large size larvae than the adult in *N. ignobilis* were selected. The protein sequences of *N. ignobilis* containing these sites were used as a query. The transcriptomes of selected species in the NCBI TSA database were downloaded (Table S7). Tblastn were used to search the homologous sites in other Megaloptera, Trichoptera, terrestrial Neuroptera, and non-ditrysian Lepidoptera species. We set the cutoff identity to 60% to prevent inclusion of paralogous genes. The sites obtained in more than three species in all four clades were selected. The frequency of amino acids was counted in different clades.

Using Fisher's exact test, we defined the sites with the convergent AA frequency higher than 80% in both Megaloptera and Trichoptera and significantly higher ( $P < 0.05$ ) than terrestrial Neuroptera, and non-ditrysian Lepidoptera as *significant convergent profile change* (SCPC) sites. The conservation score (Rate4score) generated by Rate4site was used to evaluate the site conservation. We also extended to search these sites to transcriptomes of the species in Hymenoptera, Coleoptera, Lepidoptera, and Diptera to compare their amino acid profile. The aquatic or terrestrial habits of all the species with transcriptome data were manually verified.

#### 4.11. GO enrichment analysis

We first performed GO annotation for all the single-copied OGS. The sequence of *D. melanogaster* or other species (if the sequence belonging to *Drosophila* is not present) was annotated using EggNOG. The obtained GO terms were regarded as the terms of this orthologous group. The GO term of the 2513 single-copied OGS was used as background. The enrichment analysis of interested OGS was performed using TBtools

v0.665 [10].

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.genome.2022.110437>.

## Availability of data and materials

The genome of dobsonfly *Neoneurotoma ignobilis* and trace data are available in NCBI BioProject under the accession PRJNA801769.

## Formatting of funding sources

This research was funded by the Key Project of Laboratory of Lingnan Modern Agriculture (NT2021003), the Key International Joint Research Program of National Natural Science Foundation of China (31920103005), the General Program of National Natural Science Foundation of China (31702035), the National Key Research and Development Plan (2017YFD0200101, 2019YFD0300104), the Fundamental Research Funds for the Central Universities, and the Special Research Fund for Distinguished Scholars of Zhejiang Province, China (2018R51004).

## Author contributions

Xue-xin Chen conceived the project and designed the research. Xingzhou Ma assembled and annotated the genome. Xing-zhou Ma conducted comparative genomic analysis. Zi-qi Wang conducted gene family analysis. Xi-qian Ye set up the computer for analysis. Xue-xin Chen, Xing-zhou Ma, Zi-qi Wang, Xi-qian Ye, Pu Tang, and Xingxing Shen discuss the results. Xing-zhou Ma and Xue-xin Chen wrote the manuscript. Xue-xin Chen, Xin-yue Liu, and Xingxing Shen revised the manuscript.

## CRediT authorship contribution statement

**Xing-Zhou Ma:** Methodology, Investigation, Visualization, Writing – original draft. **Zi-Qi Wang:** Formal analysis, Investigation. **Xi-Qian Ye:** Methodology, Visualization. **Xing-Yue Liu:** Validation, Writing – review & editing. **Pu Tang:** Data curation, Visualization. **Xingxing Shen:** Writing – review & editing. **Xue-Xin Chen:** Supervision, Conceptualization, Funding acquisition, Writing – review & editing.

## Acknowledgements

We thank the help of Ph.D. student Boying Zhen, graduate students Zekai Li and Ruizhong Yuan, and undergraduate student Yiyang Chen for the dobsonfly specimen collection.

## References

- [1] M. Acar, H. Jafar-Nejad, H. Takeuchi, A. Rajan, D. Ibrani, N.A. Rana, H. Pan, R. S. Haltiwanger, H.J. Bellen, Rumi is a CAP10 domain glycosyltransferase that modifies notch and is required for notch signaling, *Cell* 132 (2008) 247–258.
- [2] I. Almudi, J. Vizueta, C.D. Wyatt, A. de Mendoza, F. Marletaz, P.N. Firbas, R. Feuda, G. Masiero, P. Medina, A. Alcaina-Caro, et al., Genomic adaptations to aquatic and aerial life in mayflies and the origin of insect wings, *Nat. Commun.* 11 (2020) 1–11.
- [3] G. Benson, Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Res.* 27 (1999) 573–580.
- [4] R.L. Blanchard, R.R. Freimuth, J. Buck, R.M. Weinshilboum, M.W. Coughtrie, A proposed nomenclature system for the cytosolic sulfotransferase (SULT) superfamily, *Pharmacogenetics* 14 (2004) 199–211.
- [5] N. Bonada, N. Prat, V.H. Resh, B. Statzner, Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches, *Annu. Rev. Entomol.* 51 (2006) 495–523.
- [6] P.P. Chan, B.Y. Lin, A.J. Mak, T.M. Lowe, tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes, *BioRxiv* (2021) 614032.
- [7] P. Chauhan, J. Swaege, R.A. Sánchez-Guillén, E.I. Svensson, M. Wellenreuther, B. Hansson, Genome assembly, sex-biased gene expression and dosage compensation in the damselfly *Ischnura elegans*, *Genomics* 113 (2021) 1828–1837.
- [8] C. Chen, H. Chen, Y. Zhang, H.R. Thomas, M.H. Frank, Y. He, R. Xia, TBtools: an integrative toolkit developed for interactive analyses of big biological data, *Mol. Plant* 13 (2020) 1194–1202.
- [9] M. Chikina, J.D. Robinson, N.L. Clark, Hundreds of genes experienced convergent shifts in selective pressure in marine mammals, *Mol. Biol. Evol.* 33 (2016) 2182–2192.
- [10] P.F. Colosimo, K.E. Hosemann, S. Balabhadra, G. Villarreal, M. Dickson, J. Grimwood, J. Schmutz, R.M. Myers, D. Schlut, D.M. Kingsley, Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles, *Science* 307 (2005) 1928–1933.
- [11] K. Del-Claro, R. Guillermo, *Aquatic Insects*, Springer, 2019.
- [12] K.-D.B. Dijkstra, M.T. Monaghan, S.U. Pauls, Freshwater biodiversity and aquatic insect diversification, *Annu. Rev. Entomol.* 59 (2014) 143–163.
- [13] O. Dudchenko, S.S. Batra, A.D. Omer, S.K. Nyquist, M. Hoeger, N.C. Durand, M. S. Shamim, I. Machol, E.S. Lander, A.P. Aiden, et al., De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds, *Science* 356 (2017) 92–95.
- [14] N.C. Durand, J.T. Robinson, M.S. Shamim, I. Machol, J.P. Mesirov, E.S. Lander, E. L. Aiden, Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom, *Cell Syst.* 3 (2016) 99–101.
- [15] D.M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics, *Genome Biol.* 20 (2019) 1–14.
- [16] G. Espregueira Themudo, L.Q. Alves, A.M. Machado, M. Lopes-Marques, R.R. Da Fonseca, M. Fonseca, R. Ruivo, L.F.C. Castro, Losing genes: the evolutionary remodeling of cetacean skin, *Front. Mar. Sci.* (2020) 912.
- [17] T.R. Fallon, S.E. Lower, C.-H. Chang, M. Bessho-Uehara, G.J. Martin, A.J. Bewick, M. Behringer, H.J. Debat, I. Wong, J.C. Day, et al., Firefly genomes illuminate parallel origins of bioluminescence in beetles, *Elife* 7 (2018), e36495.
- [18] J.M. Flynn, R. Hubley, C. Goubert, J. Rosen, A.G. Clark, C. Feschotte, A.F. Smit, RepeatModeler2 for automated genomic discovery of transposable element families, *Proc. Natl. Acad. Sci.* 117 (2020) 9451–9457.
- [19] A.D. Foote, Y. Liu, G.W. Thomas, T. Vinar, J. Alföldi, J. Deng, S. Dugan, C.E. van Elk, M.E. Hunter, V. Joshi, et al., Convergent evolution of the genomes of marine mammals, *Nat. Genet.* 47 (2015) 272–275.
- [20] M.R. Friedländer, S.D. Mackowiak, N. Li, W. Chen, N. Rajewsky, miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades, *Nucleic Acids Res.* 40 (2012) 37–52.
- [21] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics* 28 (2012) 3150–3152.
- [22] R. Futahashi, R. Kawahara-Miki, M. Kinoshita, K. Yoshitake, S. Yajima, K. Arikawa, T. Fukatsu, Extraordinary diversity of visual opsin genes in dragonflies, *Proc. Natl. Acad. Sci. U. S. A.* 112 (2015) E1247–E1256.
- [23] T. Gao, C. Shih, C.C. Labandeira, J.A. Santiago-Blay, Y. Yao, D. Ren, Convergent evolution of ramified antennae in insect lineages from the Early Cretaceous of Northeastern China, *Proc. R. Soc. B* 283 (2016) 20161448.
- [24] C. Gillott, *Entomology*, Springer Science & Business Media, 2005.
- [25] R.A. Goldstein, S.T. Pollard, S.D. Shah, D.D. Pollock, Nonadaptive amino acid convergence rates decrease over time, *Mol. Biol. Evol.* 32 (2015) 1373–1381.
- [26] G. Gremme, S. Steinbiss, S. Kurtz, GenomeTools: a comprehensive software library for efficient processing of structured genome annotations, *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 10 (2013) 645–656.
- [27] D. Grimaldi, M.S. Engel, M.S. Engel, *Evolution of the Insects*, Cambridge University Press, 2005.
- [28] D. Guan, S.A. McCarthy, J. Wood, K. Howe, Y. Wang, R. Durbin, Identifying and removing haplotypic duplication in primary genome assemblies. Valencia A, editor, *Bioinformatics* 36 (2020) 2896–2898.
- [29] B.J. Haas, A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith Jr., L.I. Hannick, R. Maiti, C.M. Ronning, D.B. Rusch, C.D. Town, et al., Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res.* 31 (2003) 5654–5666.
- [30] B.J. Haas, S.L. Salzberg, W. Zhu, M. Pertea, J.E. Allen, J. Orvis, O. White, C. R. Buell, J.R. Wortman, Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments, *Genome Biol.* 9 (2008) 1–22.
- [31] E. Haghshenas, H. Asghari, J. Stoye, C. Chauve, F. Hach, HASLR: fast hybrid assembly of long reads, *Iscience* 23 (2020), 101389.
- [32] C.W. Heckman, *Ecological Strategies of Aquatic Insects*, CRC Press, Taylor & Francis Group, Boca Raton, 2018.
- [33] A.E. Hershey, G.A. Lamberti, D.T. Chaloner, R.M. Northington, Aquatic insect ecology, in: *Ecology and Classification of North American Freshwater Invertebrates*, Elsevier, 2010, pp. 659–694.
- [34] C.A. Hill, A.N. Fox, R.J. Pitts, L.B. Kent, P.L. Tan, M.A. Chrystal, A. Cravchik, F. H. Collins, H.M. Robertson, L.J. Zwiebel, G protein-coupled receptors in *Anopheles gambiae*, *Science* 298 (2002) 176–178.
- [35] K.J. Hoff, M. Stanke, Predicting genes in single genomes with AUGUSTUS, *Curr. Protoc. Bioinformatics* 65 (2019), e57.
- [36] S. Hotaling, J.L. Kelley, P.B. Frandsen, Aquatic insects are dramatically underrepresented in genomic research, *Insects* 11 (2020) 601.
- [37] K. Howe, W. Chow, J. Collins, S. Pelan, D.-L. Pointon, Y. Sims, J. Torrance, A. Tracey, J. Wood, Significantly improving the quality of genome assemblies through curation, *GigaScience* 10 (2021) giaa153.
- [38] J. Hu, J. Fan, Z. Sun, S. Liu, NextPolish: a fast and efficient genome polishing tool for long-read assembly, Berger B, editor, *Bioinformatics* 36 (2020) 2253–2255.
- [39] J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S.K. Forslund, H. Cook, D.R. Mende, I. Letunic, T. Rattei, L.J. Jensen, et al., eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource

- based on 5090 organisms and 2502 viruses, *Nucleic Acids Res.* 47 (2019) D309–D314.
- [45] G.M. Hughes, E.S. Boston, J.A. Finarelli, W.J. Murphy, D.G. Higgins, E.C. Teeling, The birth and death of olfactory receptor gene families in mammalian niche adaptation, *Mol. Biol. Evol.* 35 (2018) 1390–1406.
- [46] F.C. Jones, M.G. Grabherr, Y.F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M.C. Zody, S. White, et al., The genomic basis of adaptive evolution in threespine sticklebacks, *Nature* 484 (2012) 55–61.
- [47] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, et al., InterProScan 5: genome-scale protein function classification, *Bioinformatics* 30 (2014) 1236–1240.
- [48] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Mol. Biol. Evol.* 30 (2013) 772–780.
- [49] J.A. Kawatsuki, M.A. Bittner, Uptake, elimination, and biotransformation of the lampicide 3-trifluoromethyl-4-nitrophenol (TFM) by larvae of the aquatic midge *Chironomus tentans*, *Toxicology* 4 (1975) 183–194.
- [50] J. Keilwagen, M. Wenk, J.L. Erickson, M.H. Schattat, J. Grau, F. Hartung, Using intron position conservation for homology-based gene prediction, *Nucleic Acids Res.* 44 (2016) e89.
- [52] D. Kim, J.M. Paggi, C. Park, C. Bennett, S.L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype, *Nat. Biotechnol.* 37 (2019) 907–915.
- [53] T. Kishida, Y. Go, S. Tatsumoto, K. Tatsumi, S. Kuraku, M. Toda, Loss of olfaction in sea snakes provides new perspectives on the aquatic adaptation of amniotes, *Proc. R. Soc. B* 286 (2019) 20191828.
- [54] T. Kishida, J. Thewissen, T. Hayakawa, H. Imai, K. Agata, Aquatic adaptation and the evolution of smell and taste in whales, *Zool. Lett.* 1 (2015) 9.
- [55] M. Kolmogorov, J. Yuan, Y. Lin, P.A. Pevzner, Assembly of long, error-prone reads using repeat graphs, *Nat. Biotechnol.* 37 (2019) 540–546.
- [56] S. Koren, B.P. Walenz, K. Berlin, J.R. Miller, N.H. Bergman, A.M. Phillippy, Canu: scalable and accurate long-read assembly via adaptive  $k$ -mer weighting and repeat separation, *Genome Res.* 27 (2017) 722–736.
- [57] I. Korf, Gene finding in novel genomes, *BMC Bioinformatics* 5 (2004) 1–9.
- [58] P. Kück, ALICUT: A Perlscript which Cuts ALISCORE Identified RSS, Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany, 2009 version 2.
- [59] S. Kurtz, A. Phillippy, A.L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S. L. Salzberg, Versatile and open software for comparing large genomes, *Genome Biol.* 5 (2004) 1–9.
- [61] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, *BMC Bioinformatics* 12 (2011) 1–16.
- [62] H. Li, Minimap2: pairwise alignment for nucleotide sequences.Biro I, editor, *Bioinformatics* 34 (2018) 3094–3100.
- [63] A. Lin, L. Cao, S. Wei, X. Liu, Phylogeography of the oriental dobsonfly, *Neoneurothemis ignobilis* (Navás), suggests Pleistocene allopatric isolation and glacial dispersal shaping its wide distribution, *Syst. Entomol.* 47 (1) (2021) 65–81, sven.12519.
- [64] A. Lin, S. Wei, L. Cao, X. Liu, First development and characterization of 27 novel microsatellite markers in the dobsonfly *Neoneurothemis ignobilis* (Megaloptera: Corydalidae) at genome-scale level, *Appl. Entomol. Zool.* 55 (2020) 149–158.
- [65] A. Liu, F. He, L. Shen, R. Liu, Z. Wang, J. Zhou, Convergent degeneration of olfactory receptor gene repertoires in marine mammals, *BMC Genomics* 20 (2019) 977.
- [66] A. Lomsadze, V. Ter-Hovhannyan, Y.O. Chernoff, M. Borodovsky, Gene identification in novel eukaryotic genomes by self-training algorithm, *Nucleic Acids Res.* 33 (2005) 6494–6506.
- [67] S. Luo, M. Tang, P.B. Frandsen, R.J. Stewart, X. Zhou, The genome of an underwater architect, the caddisfly *Stenopsyche tienmushanensis* Hwang (Insecta: Trichoptera), *GigaScience* 7 (2018), <https://doi.org/10.1093/gigascience/giy143/5202446> [Internet]. Available from.
- [68] X. Ma, K. He, Z. Shi, M. Li, F. Li, X.-X. Chen, Large-scale annotation and evolution analysis of MiRNA in insects, *Genome Biol.* 13 (2021) evab083.
- [69] W.H. Majors, M. Pertea, S.L. Salzberg, TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders, *Bioinformatics* 20 (2004) 2878–2879.
- [70] A. Marcovitz, Y. Turakhia, H.I. Chen, M. Gloudemans, B.A. Braun, H. Wang, G. Bejerano, A functional enrichment test for molecular convergent evolution finds a clear protein-coding signal in echolocating bats and whales, *Proc. Natl. Acad. Sci. U. S. A.* 116 (2019) 21094–21103.
- [71] Y. Mei, D. Jing, S. Tang, X. Chen, H. Chen, H. Duanmu, Y. Cong, M. Chen, X. Ye, H. Zhou, et al., InsectBase 2.0: a comprehensive gene resource for insects, *Nucleic Acids Res.* 50 (2022) D1040–D1045.
- [72] Z. Merényi, A.N. Prasanna, Z. Wang, K. Kovács, B. Hegedüs, B. Bálint, B. Papp, J. P. Townsend, L.G. Nagy, Unmatched level of molecular convergence among deeply divergent complex multicellular fungi, *Mol. Biol. Evol.* 37 (2020) 2228–2240.
- [73] B.Q. Minh, H.A. Schmidt, O. Chernomor, D. Schrempf, M.D. Woodhams, A. Von Haeseler, R. Lanfear, IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era, *Mol. Biol. Evol.* 37 (2020) 1530–1534.
- [74] B. Misof, S. Liu, K. Meusemann, R.S. Peters, A. Donath, C. Mayer, P.B. Frandsen, J. Ware, T. Flouri, R.G. Beutel, et al., Phylogenomics resolves the timing and pattern of insect evolution, *Science* 346 (2014) 763–767.
- [75] T.F. Mitterboeck, J. Fu, S.J. Adamowicz, Rates and patterns of molecular evolution in freshwater versus terrestrial insects.Golding B, editor, *Genome* 59 (2016) 968–980.
- [76] E.P. Nawrocki, S.R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches, *Bioinformatics* 29 (2013) 2933–2935.
- [77] D. Orme, R. Freckleton, G. Thomas, T. Petzoldt, S. Fritz, N. Isaac, W. Pearse, The caper package: comparative analysis of phylogenetics and evolution in R, in: *R Package Version 5*, 2013, pp. 1–36.
- [78] S. Ou, W. Su, Y. Liao, K. Chougule, J.R. Agda, A.J. Hellinga, C.S.B. Lugo, T. A. Elliott, D. Ware, T. Peterson, et al., Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline, *Genome Biol.* 20 (2019) 1–18.
- [79] R. Partha, A. Kowalezyk, N.L. Clark, M. Chikina, Robust method for detecting convergent shifts in evolutionary rates, *Mol. Biol. Evol.* 36 (2019) 1817–1830.
- [80] V. Peona, M.P. Blom, L. Xu, R. Burri, S. Sullivan, I. Bunikis, I. Liachko, T. Haryoko, K.A. Jonsson, Q. Zhou, et al., Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise, *Mol. Ecol. Resour.* 21 (2021) 263–286.
- [81] M. Pertea, G.M. Pertea, C.M. Antonescu, T.-C. Chang, J.T. Mendell, S.L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, *Nat. Biotechnol.* 33 (2015) 290–295.
- [82] M. Petersen, D. Armién, R.A. Gibbs, L. Hering, A. Khila, G. Mayer, S. Richards, O. Niehuis, B. Misof, Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects, *BMC Evol. Biol.* 19 (2019) 11.
- [83] G. Piñeiro, A. Ramos, C. Goso, F. Scarabino, M. Laurin, Unusual environmental conditions preserve a Permian mesosaur-bearing Konservat-Lagerstätte from Uruguay, *Acta Palaeontol. Pol.* 57 (2011) 299–318.
- [84] T. Pupko, R.E. Bell, I. Mayrose, F. Glaser, N. Ben-Tal, Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues, *Bioinformatics* 18 (2002) S71–S77.
- [85] A. Rhie, S.A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, J. Kim, et al., Towards complete and error-free genome assemblies of all vertebrate species, *Nature* 592 (2021) 737–746.
- [86] J. Ruan, H. Li, Fast and accurate long-read assembly with wtdbg2, *Nat. Methods* 17 (2020) 155–158.
- [87] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics* 31 (2015) 3210–3212.
- [88] G.S.C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison, *BMC Bioinformatics* 6 (2005) 1–11.
- [89] M.D. Smith, J.O. Wertheim, S. Weaver, B. Murrell, K. Scheffler, S.L. Kosakovsky Pond, Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection, *Mol. Biol. Evol.* 32 (2015) 1342–1353.
- [90] K.-Q. Song, H.-J. Xue, R.G. Beutel, M. Bai, D.-J. Bian, J. Liu, Y.-Y. Ruan, W.-Z. Li, F.-L. Jia, X.-K. Yang, Habitat-dependent diversification and parallel molecular evolution: water scavenger beetles as a case study, *Curr. Zool.* 60 (2014) 561–570.
- [91] P. Sroka, R.J. Godunko, N.D. Sinitshenkova, J. Prokop, Life history, systematics and flight ability of the Early Permian stem-mayflies in the genus *Misthodites* Sellards, 1909 (Insecta, Ephemeroidea, Permoleptoptera), *BMC Ecol. Evo* 21 (2021) 97.
- [92] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, B. Morgenstern, AUGUSTUS: ab initio prediction of alternative transcripts, *Nucleic Acids Res.* 34 (2006) W435–W439.
- [93] D.L. Stern, The genetic causes of convergent evolution, *Nat. Rev. Genet.* 14 (2013) 751–764.
- [94] J.F. Storz, Causes of molecular convergence and parallelism in protein evolution, *Nat. Rev. Genet.* 17 (2016) 239–250.
- [95] M. Suiko, K. Kurogi, T. Hashiguchi, Y. Sakakibara, M.-C. Liu, Updated perspectives on the cytosolic sulfotransferases (SULTs) and SULT-mediated sulfation, *Biosci. Biotechnol. Biochem.* 81 (2017) 63–72.
- [96] X. Sun, W. Liu, R. Li, C. Zhao, L. Pan, C. Yan, A chromosome level genome assembly of *Propolyceros akamusi* to understand its response to heavy metal exposure, *Mol. Ecol. Resour.* 21 (2021) 1996–2012.
- [97] H. Takeuchi, R.C. Fernández-Valdivia, D.S. Caswell, A. Nita-Lazar, N.A. Rana, T. P. Garner, T.K. Weldeghiorghis, M.A. Macnaughtan, H. Jafar-Nejad, R. S. Haltiwanger, Rumi functions as both a protein O-glucosyltransferase and a protein O-xylosyltransferase, *Proc. Natl. Acad. Sci.* 108 (2011) 16600–16605.
- [98] S. Tempel, Using and understanding RepeatMasker, in: *Mobile Genetic Elements*, Springer, 2012, pp. 29–51.
- [99] W.D. Tracey Jr., R.I. Wilson, G. Laurent, S. Benzer, *Painless*, a *Drosophila* gene essential for nociception, *Cell* 113 (2003) 261–273.
- [100] S. Treaster, J.M. Daane, M. Harris, Refining convergent rate analysis with topology in mammalian longevity and marine transitions, *Mol. Biol. Evol.* 38 (11) (2021) 5190–5203.
- [101] R. Vaser, M. Šikić, Time-and memory-efficient genome assembly with raven, *Nat. Comput. Sci.* 1 (2021) 332–336.
- [102] R. Vaser, I. Šović, N. Nagarajan, M. Šikić, Fast and accurate de novo genome assembly from long uncorrected reads, *Genome Res.* 27 (2017) 737–746.
- [103] J.R. Wang, J. Holt, L. McMillan, C.D. Jones, FMLRC: hybrid long read error correction using an FM-index, *BMC Bioinformatics* 19 (2018) 1–11.
- [104] D. Weihrauch, A. Donini, M.J. O'Donnell, Ammonia transport by terrestrial and aquatic insects, *J. Insect Physiol.* 58 (2012) 473–487.
- [105] J.O. Wertheim, B. Murrell, M.D. Smith, S.L. Kosakovsky Pond, K. Scheffler, RELAX: detecting relaxed selection in a phylogenetic framework, *Mol. Biol. Evol.* 32 (2015) 820–832.

- [106] H. Wildermuth, Dragonflies recognize the water of rendezvous and oviposition sites by horizontally polarized light: a behavioural field test, *Naturwissenschaften* 85 (1998) 297–302.
- [107] J. Wolff, L. Rabbani, R. Gilsbach, G. Richard, T. Manke, R. Backofen, B. A. Grüning, Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization, *Nucleic Acids Res.* 48 (2020) W177–W184.
- [108] R.J. Woottton, The historical ecology of aquatic insects: an overview, *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 62 (1988) 477–492.
- [109] S. Xu, C. Cang, X. Liu, Y. Peng, Y. Ye, Z. Zhao, A. Guo, Thermal nociception in adult *Drosophila*: behavioral characterization and the role of the painless gene, *Genes Brain Behav.* 5 (2006) 602–613.
- [110] F. Yang, W. Chang, F. Hayashi, J. Gillung, Y. Jiang, D. Yang, X. Liu, Evolutionary history of the complex polymorphic dobsonfly genus *Neoneuromus* (Megaloptera: Corydalidae): *Neoneuromus* systematics and phylogeny, *Syst. Entomol.* 43 (2018) 568–595.
- [111] Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol.* 24 (2007) 1586–1591.
- [112] H.-S. Yim, Y.S. Cho, X. Guang, S.G. Kang, J.-Y. Jeong, S.-S. Cha, H.-M. Oh, Lee Jae-Hak, E.C. Yang, K.K. Kwon, et al., Minke whale genome and aquatic adaptation in cetaceans, *Nat. Genet.* 46 (2014) 88–92.
- [113] J. Yoshimura, K. Ichikawa, M.J. Shoura, K.L. Artiles, I. Gabdank, L. Wahba, C. L. Smith, M.L. Edgley, A.E. Rougvie, A.Z. Fire, et al., Recompleting the *Caenorhabditis elegans* genome, *Genome Res.* 29 (2019) 1009–1022.
- [114] Y. Yuan, Y. Zhang, P. Zhang, C. Liu, J. Wang, Gao Haiyu, A.R. Hoelzel, I. Seim, M. Lv, M. Lin, et al., Comparative genomics provides insights into the aquatic adaptations of mammals, *Proc. Natl. Acad. Sci. U. S. A.* 118 (2021), e2106080118.
- [115] Y.-H. Zhang, V. Ravi, G. Qin, H. Dai, H.-X. Zhang, F.-M. Han, X. Wang, Y.-H. Liu, J.-P. Yin, L.-M. Huang, et al., Comparative genomics reveal shared genomic changes in syngnathid fishes and signatures of genetic convergence with placental mammals, *Natl. Sci. Rev.* 7 (2020) 964–977.
- [116] X. Zhou, I. Seim, V.N. Gladyshev, Convergent evolution of marine mammals is associated with distinct substitutions in common genes, *Sci. Rep.* 5 (2015) 16550.
- [117] Y. Zhou, L. Shearwin-Whyatt, J. Li, Z. Song, T. Hayakawa, D. Stevens, J. C. Fenelon, E. Peel, Y. Cheng, F. Pajpach, et al., Platypus and echidna genomes reveal mammalian biology and evolution, *Nature* 592 (2021) 756–762.
- [118] Z. Zou, J. Zhang, Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Mol. Biol. Evol.* 32 (2015) 2085–2096.
- [119] Z. Zou, J. Zhang, Gene tree discordance does not explain away the temporal decline of convergence in mammalian protein sequence evolution, *Mol. Biol. Evol.* 34 (2017) 1682–1688.
- [120] A. Wallberg, I. Bunikis, O.V. Pettersson, M.-B. Mosbech, A.K. Childers, J.D. Evans, A.S. Mikheyev, H.M. Robertson, G.E. Robinson, M.T. Webster, A hybrid de novo genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds, *BMC Genomics* 20 (2019) 1–19.
- [121] M. Kawamoto, A. Jouraku, A. Toyoda, K. Yokoi, Y. Minakuchi, S. Katsuma, A. Fujiyama, T. Kiuchi, K. Yamamoto, T. Shimada, High-quality genome assembly of the silkworm, *Bombyx mori*, *Insect Biochem. Mol. Biol.* 107 (2019) 53–62.
- [122] N. Herndon, J. Shelton, L. Gerischer, P. Ioannidis, M. Ninova, J. Dönitz, R. M. Waterhouse, C. Liang, C. Damm, J. Siemanowski, others, Enhanced genome assembly and a new official gene set for *Tribolium castaneum*, *BMC Genomics* 21 (2020) 1–13.
- [123] S.J. Larkin, G. Marygold, H. Antonazzo, G. Attrill, P.V. Dos Santos, J.L. Garapati, L.S. Goodman, G. Gramates, V.B. Millburn, Strelets, others, FlyBase: updates to the *Drosophila melanogaster* knowledge base, *Nucleic Acids Res.* 49 (2021) D899–D907.