RESEARCH Open Access

Systematic dissection of genomic features determining the vast diversity of conotoxins

Jian-Wei Zheng[1,2†], Yang Lu[1†], Yu-Feng Yang[1], Dan Huang[1], Da-Wei Li[1], Xiang Wang[1], Yang Gao[3], Wei-Dong Yang[1], Yuanfang Guan[4] and Hong-Ye Li[1*]

Abstract

Background *Conus*, a highly diverse species of venomous predators, has attracted significant attention in neuroscience and new drug development due to their rich collection of neuroactive peptides called conotoxins. Recent advancements in transcriptome, proteome, and genome analyses have facilitated the identification of conotoxins within *Conus*' venom glands, providing insights into the genetic features and evolutionary patterns of conotoxin genes. However, the underlying mechanism behind the extraordinary hypervariability of conotoxins remains largely unknown.

Results We analyzed the transcriptomes of 34 *Conus* species, examining various tissues such as the venom duct, venom bulb, and salivary gland, leading to the identification of conotoxin genes. Genetic variation analysis revealed that a subset of these genes (15.78% of the total) in *Conus* species underwent positive selection ($Ka/Ks > 1$, $p < 0.01$). Additionally, we reassembled and annotated the genome of *C. betulinus*, uncovering 221 conotoxin-encoding genes. These genes primarily consisted of three exons, with a significant portion showing high transcriptional activity in the venom ducts. Importantly, the flanking regions and adjacent introns of conotoxin genes exhibited a higher prevalence of transposon elements, suggesting their potential contribution to the extensive variability observed in conotoxins. Furthermore, we detected genome duplication in *C. betulinus*, which likely contributed to the expansion of conotoxin gene numbers. Interestingly, our study also provided evidence of introgression among *Conus* species, indicating that interspecies hybridization may have played a role in shaping the evolution of diverse conotoxin genes.

Conclusions This study highlights the impact of adaptive evolution and introgressive hybridization on the genetic diversity of conotoxin genes and the evolution of *Conus*. We also propose a hypothesis suggesting that transposable elements might significantly contribute to the remarkable diversity observed in conotoxins. These findings not only enhance our understanding of peptide genetic diversity but also present a novel approach for peptide bioengineering.

Keywords *Conus*, Conotoxin, Gene feature, Transposon element, Introgression

[†]Jian-Wei Zheng and Yang Lu authors contributed equally to this work.

*Correspondence:
Hong-Ye Li
thyli@jnu.edu.cn

[1]Key Laboratory of Aquatic Eutrophication and Control of Harmful Algal Blooms of Guangdong Higher Education Institute, College of Life Science and Technology, Jinan University, Guangzhou 510632, China
[2]College of Food Science and Engineering, Foshan University of Science and Technology, Foshan 528231, China
[3]Gulou Hospital, Nanjing University, Nanjing, China
[4]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

Background

Marine cone snails, belonging to the genus *Conus*, comprise approximately 700 species, and they are typically classified into three main groups: vermivorous, molluscivorous, and piscivorous [1]. These carnivorous predators utilize conotoxins, found in their milked venom, for prey hunting. Conotoxins are intricate combinations of small-molecule active polypeptides known as conopeptides, which generally consist of 30 to 200 amino acid residues and can adopt diverse disulfide structures [2, 3].

The precursor of conopeptides typically consists of three domains: an N-terminal signal peptide, a propeptide, and a mature peptide located near the C-terminal. Among these domains, the signal peptide exhibits high conservation within the gene superfamily [4]. In contrast, the mature peptides display significant variability among conopeptides. Analysis of mature peptides reveals an accelerated rate of nucleotide substitution and a predominance of nonsynonymous substitutions, suggesting that the targeted mutators in the mature peptide region and diversifying selection may account for the hypervariability observed in conopeptides [5]. Similarly, *C. bullatus* demonstrates high structural diversity and a high single-nucleotide polymorphism (SNP) rate in conopeptides, supporting the hypothesis of diversifying selection in conopeptides [6].

Furthermore, targeted sequencing of venom genes from 32 *Conus* species has revealed a wide range of conotoxin gene copies, varying from 120 to 859. Notably, exons encoding the mature toxin region exhibit higher divergence, indicating positive selection acting on conotoxin genes [7]. However, the precise factors influencing the genetic hypervariability of conotoxin genes remain unclear.

The genomes of four *Conus* species have been released, including *C. tribblei* [8], *C. consors* [9], *C. betulinus* [10], and *C. ventricosus* [11]. However, the genomes of *C. tribblei* and *C. consors* suffer from severe fragmentation, which limits further analysis. On the other hand, the high-quality genome of *C. betulinus* provides valuable insights into the fundamental genetic principles governing conopeptides. Notably, it reveals a primary genetic relationship known as the "central dogma" of conopeptides, where the ratio of genes to transcripts to proteins to conopeptides is approximately 1:1:1:10. This observation suggests that post-translational modifications, such as alternative cleavage sites, highly variable N- and C-terminal truncations, and post-translational modifications, may play a significant role in generating the extensive diversity of conopeptides derived from a limited set of conotoxin genes [10, 12]. These findings significantly advance our understanding of conopeptide diversity at the translational and post-translational modification levels.

However, the genetic evolution of conotoxin genes remains a subject of ongoing investigation. Moreover, the lack of available genome annotation for the published *C. betulinus* genome hinders its comprehensive utilization in further analysis. Additionally, noncoding regions of the genome have been shown to contribute to gene diversity. Introns, which are prevalent in Metazoan genomes, facilitate frequent alternative splicing and promote the diversification of gene families through exon recombination [13]. Similarly, transposon elements (TEs) play crucial roles in genome evolution and function, including genome mutations, rearrangements, and the generation of new genes [14]. The chromosome-level genome of *C. ventricosus* suggests that conotoxin genes are located within repetitive regions, and a whole-genome duplication event has been identified [11]. These findings indicate a potential association between genome features and the diversity of conotoxin genes. In summary, while extensive research has been conducted on conotoxins, the molecular mechanisms underlying the genetic diversity of conotoxin genes remain largely unresolved. Further investigation is warranted to explore the intricate relationships between genome features and the hypervariability observed in conotoxin genes.

In this study, we conducted transcriptome assembly for 34 *Conus* species to analyze the genetic evolution of conotoxin genes. However, due to the unavailability of publicly accessible genome annotation for the published *C. betulinus* genome, which limits comprehensive analysis of its genomic structure, we performed additional reassembly of the complete genome of *C. betulinus* using publicly available genome sequencing datasets from Peng et al. [10]. Through this process, we annotated repetitive elements and protein-coding genes, enabling us to investigate whole genome duplication events, structural characteristics, expression patterns, and alternative splicing processes of conotoxin genes in *C. betulinus*. Moreover, we explored the presence of transposable elements in the flanking regions of conotoxin genes and in the introns adjacent to the highly variable mature-peptide coding sequences. Additionally, we assessed the occurrence of introgressive hybridization of conotoxin genes among various *Conus* species, utilizing publicly available conotoxin gene targeted sequencing datasets from Phuong et al. [7].

Results

Transcriptome assembly and evaluation of 34 *Conus* species

RNA sequencing datasets of 34 *Conus* species were retrieved from NCBI (Table S1), and transcripts of each species were assembled. The number of unigenes among different species varied greatly, as shown in Table 1, ranging from 20,062 to 235,341. Similarly, N50

Table 1 Summary of transcriptome assembly of 34 *Conus* species

Species	No. of Unigenes	N50 (bp)	BUSCO (%) [a]	Species	No. of Unigenes	N50 (bp)	BUSCO (%) [a]
C. abbreviatus	64,205	469	45.5	*C. magus*	95,345	917	77.4
C. arenatus	19,861	584	29.7	*C. maioensis*	100,467	890	72.9
C. aristophanes	73,258	489	53.3	*C. marmoreus*	89,231	818	78.4
C. bayani	135,227	549	78.0	*C. miliaris*	123,758	695	80.2
C. betulinus	235,341	677	95.4	*C. mordeirae*	83,999	628	67.6
C. chaldaeus	123,872	557	64.9	*C. purpurascens*	181,074	1,128	95.7
C. consors	179,498	1,213	94.2	*C. quercinus*	25,760	617	36.1
C. coronatus	31,974	511	41.5	*C. rattus*	23,807	616	40.7
C. ebraeus	46,323	583	49.7	*C. regonae*	77,176	587	62.4
C. episcopatus	71,367	629	21.6	*C. sp.* f AW-2021	91,192	290	34.2
C. ermineus	104,136	1,029	78.5	*C. sponsalis*	20,062	582	27.6
C. gloriamaris	178,627	512	69.3	*C. striatus*	108,738	854	86.4
C. imperialis	116,677	707	80.6	*C. terebra*	50,768	622	49.4
C. judaeus	80,425	324	46.5	*C. textile*	66,371	426	47.1
C. lenavati	203,012	542	84.2	*C. tribblei*	182,538	776	87.9
C. litteratus	88,210	428	58.0	*C. ventricosus*	119,838	1,028	79.7
C. lividus	25,887	575	38.9	*C. virgo*	106,353	892	87.3

a: Percentage of complete and fragmented BUSCOs.

Table 2 Identified conotoxins in 34 *Conus* species

Species	No. of conotoxins	Species	No. of conotoxins
C. abbreviatus	253	*C. magus*	135
C. arenatus	129	*C. maioensis*	154
C. aristophanes	282	*C. marmoreus*	188
C. bayani	119	*C. miliaris*	127
C. betulinus	65	*C. mordeirae*	153
C. chaldaeus	83	*C. purpurascens*	51
C. consors	57	*C. quercinus*	73
C. coronatus	206	*C. rattus*	94
C. ebraeus	111	*C. regonae*	178
C. episcopatus	34	*C. sp.* f AW-2021	123
C. ermineus	58	*C. sponsalis*	159
C. gloriamaris	89	*C. striatus*	107
C. imperialis	102	*C. terebra*	81
C. judaeus	124	*C. textile*	99
C. lenavati	81	*C. tribblei*	61
C. litteratus	145	*C. ventricosus*	86
C. lividus	105	*C. virgo*	199

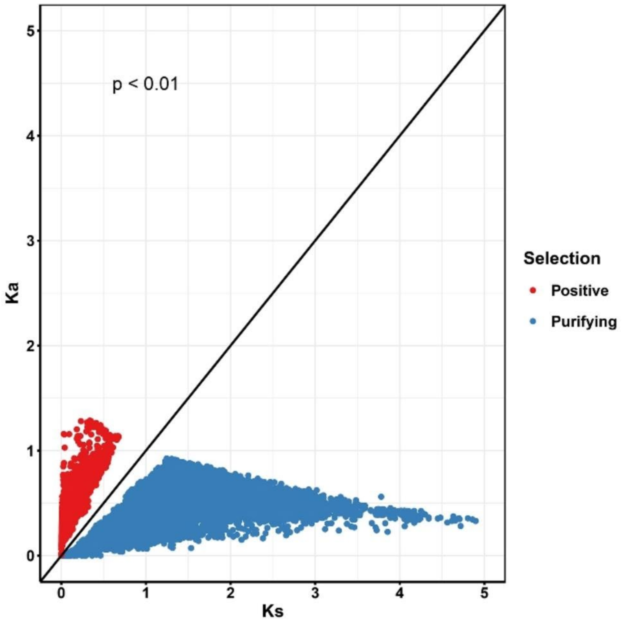

Fig. 1 Ka/Ks calculation of conotoxin genes orthogroups in 34 *Conus* species

and completeness of transcripts in each species were also diverse in the range from 290 bp to 1,213 bp in N50 and 21.6–95.7% in completeness (percentage of complete and fragmented BUSCOs), respectively. In addition, there were 21 species of *Conus* for which the "complete and fragmented BUSCOs" were higher than 50%.

Identification and genetic variation of conotoxins in *Conus*

Conotoxins were predicted in each species of *Conus*. A total of 4,111 conotoxins were identified in 34 *Conus* species transcriptomes from published data (Table S1), and the number of identified conotoxins in each species varied widely, ranging from 34 to 282 (Table 2). Subsequently, homology clustering was performed on the total conotoxin genes, and 91 orthogroups were archived. Among these orthogroups, there were 22 orthogroups that each contained more than 80% of the species represented, including 4 orthogroups that contained all species represented. Genetic variation analysis of conotoxins within orthogroups showed that most conotoxins were under purifying selection (Ka/Ks < 1) (Fig. 1). However, positive selection (Ka/Ks > 1, total of 15.78%) and significant genetic variation were also observed in conotoxins.

Positive: positive selection (Ka/Ks > 1). **Purifying**: purifying selection (Ka/Ks < 1).

Genome reassembly and annotation of *C. betulinus*

A total of 239.7 Gb of raw reads generated by PacBio Sequel were used to assemble the genome of *C. betulinus* and polished with cleaned Illumina sequencing reads, resulting in an assembly of 51,913 scaffolds with a total length of 2.67 Gb. The assembly genome results showed that the maximum and average scaffold lengths were 2.66 Mb and 51.5 kb, respectively, with an N50 length of 127 kb (Table 3). Meanwhile, in addition to simple repeats, LINEs (long interspersed nuclear elements, 9.71% of the genome) had the highest proportion and were the main types of transposon elements (Table S2). LTRs (long terminal repeats, 6.88% of the genome) were also abundant in the genome of *C. betulinus.* It has been reported that both LINEs and LTRs are retrotransposons [14].

Combined with *de novo*, homology, and RNA-seq methods, we finally predicted 24,308 protein-coding genes in *C. betulinus.* Overall, the transcription of 97.7% of the protein-coding genes was supported by the transcriptomes of multiple specimens. Moreover, BUSCO with the metazoa_odb10 database was employed to evaluate the completeness of predicted genes. It showed that the predicted genes contained 688 (72.1%) single-copy and 112 (11.7%) duplicated complete genes, as well as 49 (5.1%) fragmented genes. Compared with the published genome of *C. betulinus* [10], which contained 763 (78.0%) single-copy and 115 (11.8%) duplicated complete genes and 31 (3.17%) fragmented BUSCOs, genome completeness was similarly high in the present study.

Subsequently, as shown in Fig. 2A, the distribution of synonymous substitution rate (Ks) between paralog pairs in *C. betulinus* was calculated. And a second Ks peak was observed, suggesting similar divergence between paralogs after whole genome duplication (WGD). Furthermore, significantly conserved homologous gene blocks were identified among the scaffolds of the *C. betulinus* genome (Fig. 2B). Similarly, conserved homologous gene blocks among the pseudo-chromosomes of the *C. ventricosus* genome were also observed (Fig. 2C), for instance, between pseudo-chromosomes 1 and 2, and between pseudo-chromosomes 7 and 8.

(A) Distribution of synonymous substitution rate (Ks) between paralog pairs in *C. betulinus*. The presence of second Ks peak suggests the similar divergence between paralogs after WGD. **(B)** Conserved homologous gene blocks between scaffolds in *C. betulinus* derived from ortholog proteins. **C.** Conserved homologous gene blocks between chromosome-level scaffolds in *C. ventricosus* derived from ortholog proteins.

Table 3 Summary of the reassembly genome of *C. betulinus*

Genome evaluation	This study (reassembled)	Peng et al. [10]
Scaffold number	51,913	41,426
Total bases (bp)	2,673,840,836	3,430,828,710
Max sequence length (bp)	2,659,028	2,850,889
Average sequence length (bp)	51,506.19	82,815.21
Median sequence length (bp)	20,689	31,036
N50 (bp)	127,191	232,607
Ns (%)	1.53	0.87
Protein-coding gene number	24,308	22,698
Complete BUSCO score (%)	83.8	89.8

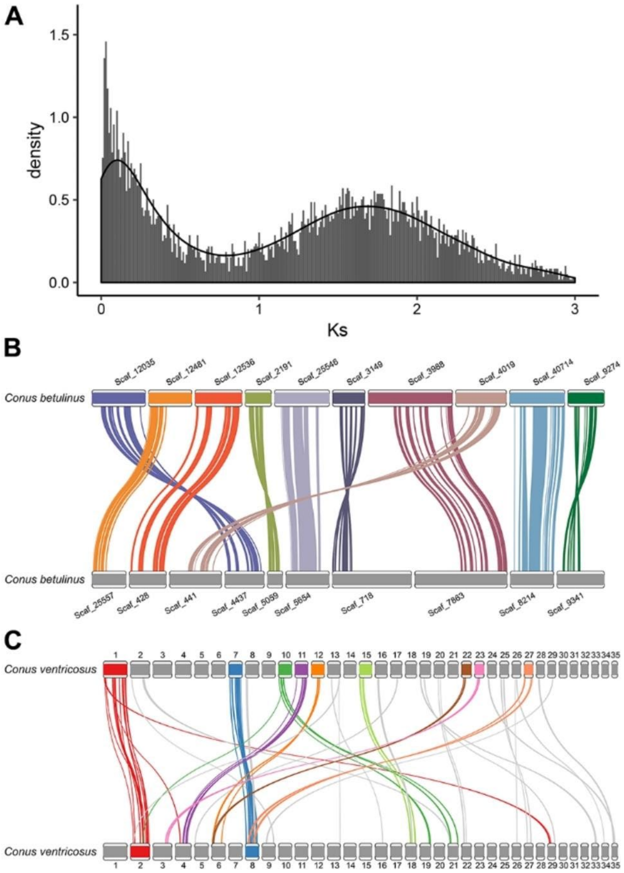

Fig. 2 Whole genome duplication (WGD) and conserved homologous gene blocks analysis of *C. betulinus* and *C. ventricosus*

Expression characteristics of conotoxin genes in *C. betulinus*

In the present study, 221 conotoxin genes were identified in the reassembled *C. betulinus* genome. Whereas 133 conotoxin genes are identified in the published genome [10]. Homology clustering results on these two gene sets showed that 117 and 43 conotoxin genes were uniquely identified in the reassembled and published genomes, respectively (Table S3). The identified conotoxin genes in the present study were classified into 12 known superfamilies, and the M- and O- superfamilies were the most abundant. Meanwhile, 17 cysteine frameworks were also classified. However, 142 conotoxin genes were unclassified into the known superfamilies. The expression level (TPM) of those 221 conotoxin genes in the venom bulbs and venom ducts from different body lengths of *C. betulinus*, namely small, middle, and big, was calculated and used to profile the expression patterns of conotoxin genes. It showed that the expression level of conotoxin genes in venom ducts was dramatically higher than that in the venom bulb (Fig. 3A), and in venom ducts, 169 out of 221 (76.47%) conotoxin genes had an average TMP higher than 10. Meanwhile, although conotoxin genes were highly expressed in all venom duct tissues, the expression characteristics of conotoxin genes in the venom ducts of individuals with different body lengths were significantly diverse. As shown in Fig. 3B, the expression level (TPM) of most genes in the venom ducts of small and big individuals was relatively consistent, especially the non-conotoxin genes. However, the expression level of conotoxin genes in the venom ducts between individuals with different body lengths, including small, middle, and big, was significantly diverse (Fig. 3C and D).

Additionally, alternative splicing has been observed in the transcription process of conotoxin genes among individuals with different body lengths, which may increase the diversity of conotoxins in *C. betulinus*. However, the frequency of alternative splicing of conotoxin genes between different individuals was limited, with only 24 conotoxin genes detected.

Structure features of conotoxin genes in *C. betulinus*

The distribution of the length of protein-coding genes in *C. betulinus* showed that the length of conotoxin and non-conotoxin genes was relatively similar; the

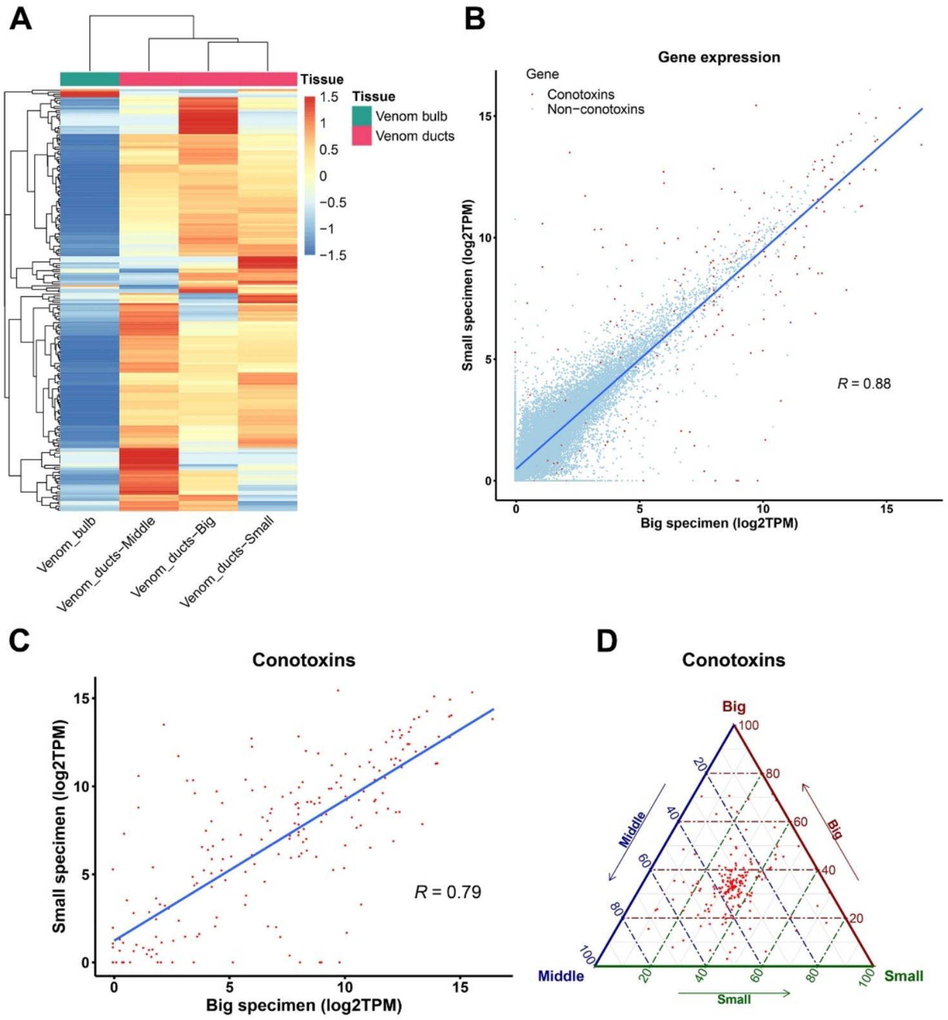

Fig. 3 Expression characteristics of conotoxin genes in different tissues and specimens of *C. betulinus*. **A.** Expression (TPM, transcripts per kilobase million) of conotoxin genes in different tissues (venom bulb and venom ducts) of *C. betulinus*. TPM was normalized by z-score. **B.** Expression of conotoxin and non-conotoxin genes in venom ducts between different body length individuals (small and big) of *C. betulinus*. Red color: conotoxin genes, blue color: non-conotoxin genes. **C - D.** Expression of conotoxin genes in venom ducts between different body length individuals (small, middle and big) of *C. betulinus*

average length of conotoxin and non-conotoxin genes was 12,014 bp and 14,063 bp, respectively (Fig. 4A). In addition, the statistical results of the exon number of protein-coding genes showed that the conotoxin genes were mainly composed of three exons, while the exon number of non-conotoxin genes showed greater fluctuation (Fig. 4B). Furthermore, the statistical results of the exon length of protein-coding genes in *C. betulinus* showed that the exon length of non-conotoxin genes was approximately double that of conotoxin genes, with the average exon length of conotoxin and non-conotoxin genes being 92 and 177 bp, respectively (Fig. 4C). It's worth noting that the intron length of conotoxin genes in *C. betulinus* was significantly longer than that of non-conotoxin genes (Fig. 4D); however, the effects of introns, such as providing mutational hotspots [15] or affecting gene size or structure expansion [16], on the diversity of conotoxin genes remained largely unknown.

In the present study, gene family expansion and contraction analysis of *C. betulinus* showed that two reverse transcriptases and a transposable element-derived protein had expanded rapidly. TE abundance in the genome of *C. betulinus* was counted with a sliding window of 20 kb. Combined with the distribution of conotoxin genes in the *C. betulinus* genome, we proposed that there may be a high density of TEs around conotoxin genes (Fig. 5A). Considering that the genome of *C. betulinus* assembled in the present study was still fragmented and might affect the statistical results, the chromosome-level genome of *C. ventricosus* published recently [11] was also included in the analysis. It showed that conotoxin genes in *C. ventricosus* tended to be distributed in regions with high TE density (Fig. 5B), suggesting that TEs may be related to the genetic diversity of conotoxin genes.

Subsequently, the content of TEs in the upstream and downstream flanking regions (100 kb) of protein-coding genes in the *C. betulinus* genome was analyzed. As shown in Fig. 6A, there was no significant difference in the content of TEs in the upstream flanking regions between conotoxin and non-conotoxin genes in *C. betulinus*. In contrast, the content of TEs in the downstream flanking regions of conotoxin genes was significantly higher than that in non-conotoxin genes (Fig. 6B, Table S4). Considering that the genome of *C. betulinus* assembled in the present study was still fragmented, the analysis was also carried out on the chromosome-level genome of *C.*

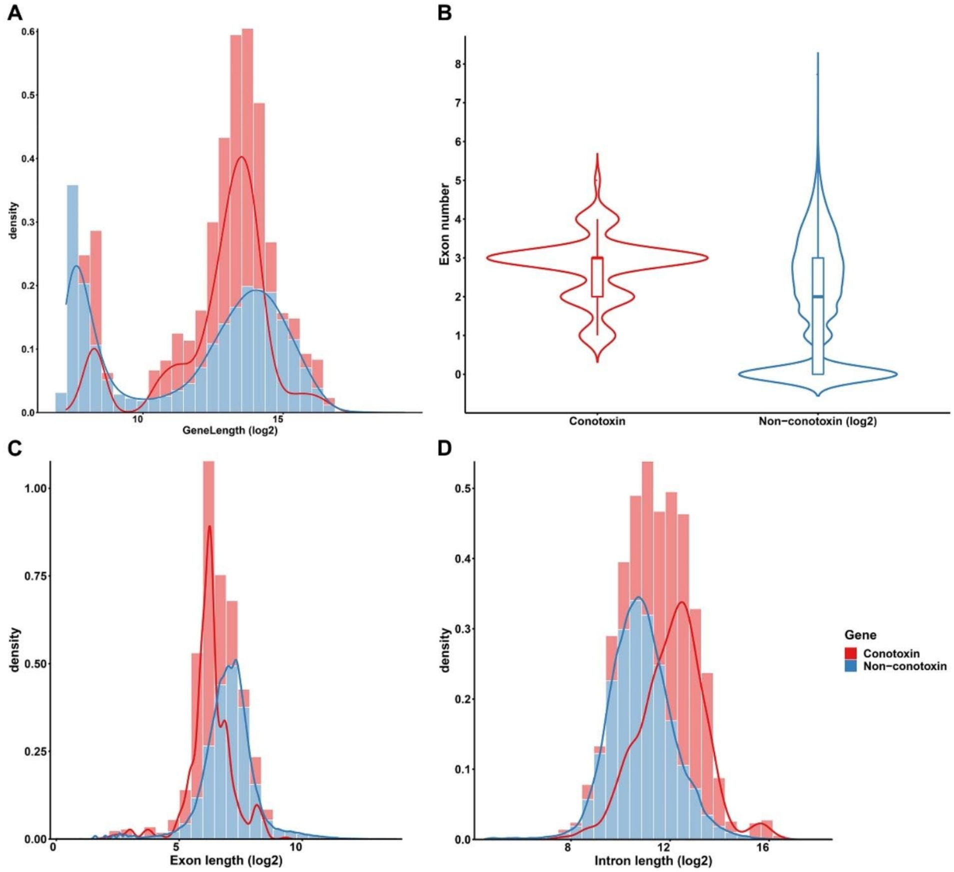

Fig. 4 Structure features of protein-coding genes in *C. betulinus*. **A.** Distribution of protein-coding genes length. **B.** Statistics of exon numbers in protein-coding genes. **C.** Distribution of exon length. **D.** Distribution of intron length. **Red color**: conotoxin genes. **Blue color**: non-conotoxin genes

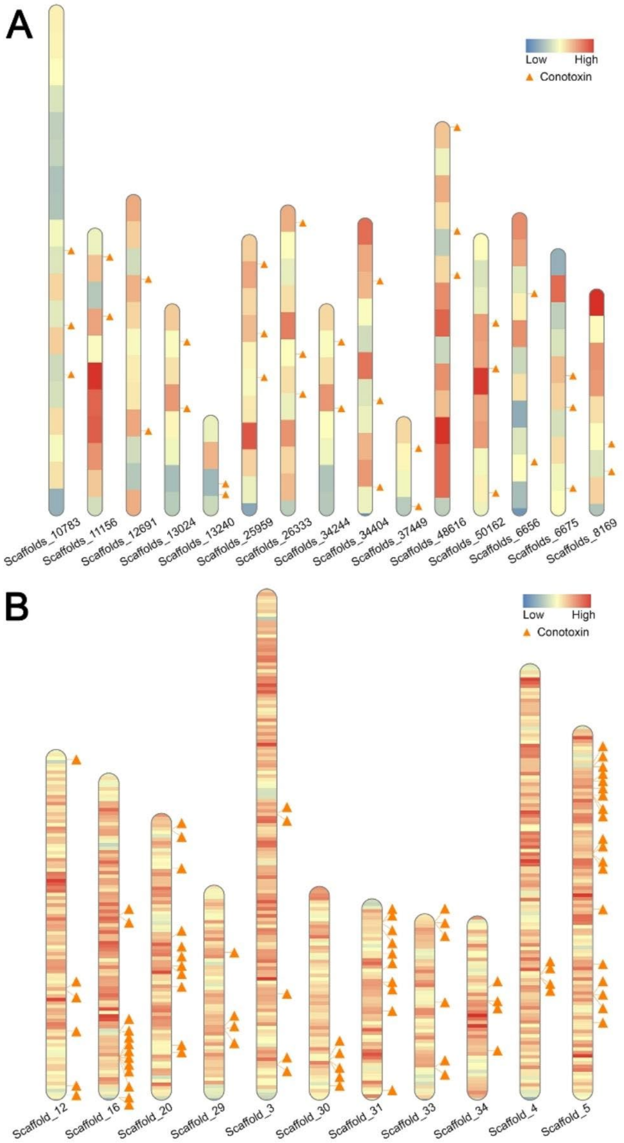

Fig. 5 Distribution of conotoxin genes and density of transposable elements (TEs) in parts of the scaffolds and chromosomes of *C. betulinus* and *C. ventricosus*. **A.** Distribution of parts of conotoxin genes and density of TEs in some of *C. betulinus* scaffolds. **B.** Distribution of parts of conotoxin genes and density of TEs in some of *C. ventricosus* chromosomes. **Triangle**: Conotoxin genes located in genome. **Heatmap**: Density of TEs in genome from high (red) to low (blue)

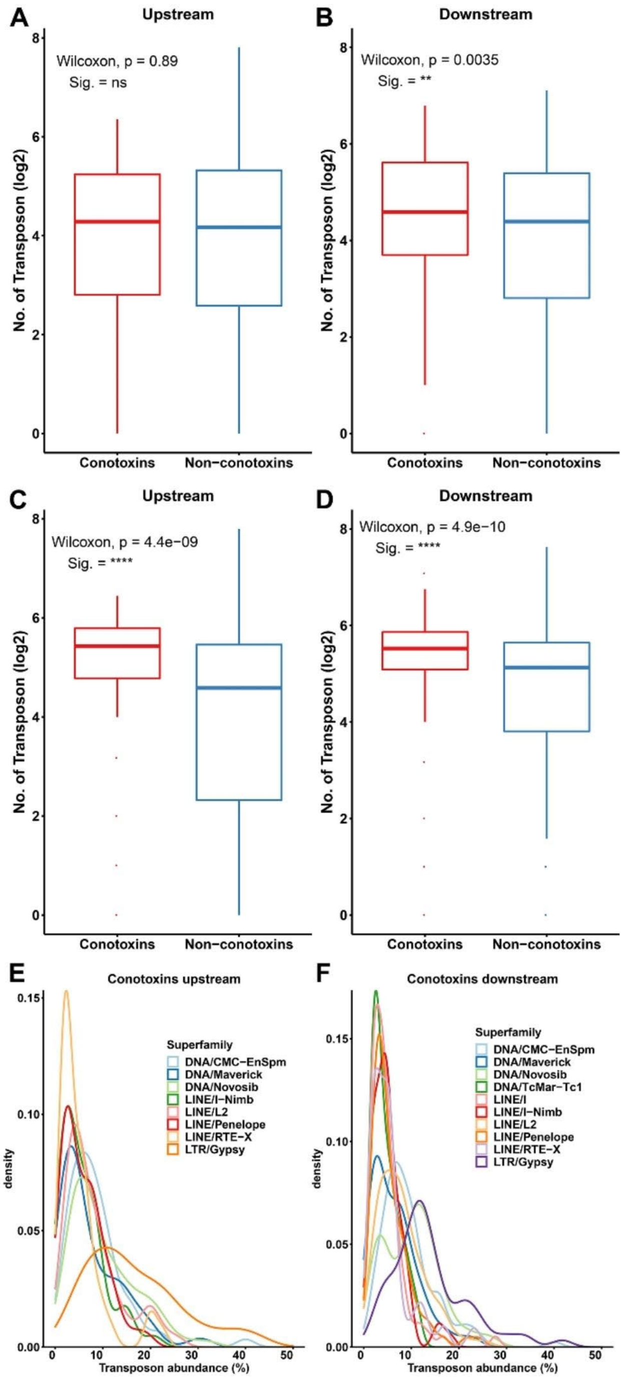

Fig. 6 Content of TEs in flanking regions (100 kb) of protein-coding genes in *C. betulinus* and *C. ventricosus*. **A - B.** Content of TEs in the upstream and downstream flanking regions of protein-coding genes in *C. betulinus*. **C - D.** Content of TEs in the upstream and downstream flanking regions of protein-coding genes in *C. ventricosus*. **E - F.** Abundance of different type of TEs in the upstream and downstream flanking regions of conotoxin genes in *C. ventricosus*. Significant differences were performed by Wilcoxon method, and indicated at $p < 0.01$ (**) or $p < 1e\text{-}5$ (****)

ventricosus. It showed that the content of TEs in both the upstream and downstream flanking regions of conotoxin genes was significantly higher than that in non-conotoxin genes in *C. ventricosus* (Fig. 6C and D, Table S5). Moreover, type I retrotransposons (LINE and LTR) and type II DNA transposons are the main TEs in the flanking regions (both upstream and downstream) of conotoxin genes (Fig. 6E F, Table S5). Furthermore, the downstream flanking region of the conotoxin gene had more TEs than the upstream region. In addition, the Gypsy superfamily of LTR retrotransposons has a higher proportion in both upstream and downstream flanking regions of protein-coding genes, ranging from 8.13 to 10.43% in *C. betulinus* and 14.83–16.70% in *C. ventricosus* (Table S4 and Table S5).

As shown in Fig. 4D, conotoxin genes have longer intron structures compared to those of non-conotoxin genes in *C. betulinus*. The content of TEs in introns adjacent to the highly variable mature peptide of conotoxins was significantly higher than that of introns adjacent to the conserved pro-peptide (Fig. 7A and B, Table S6). Like the main types of TEs in the flanking regions (upstream and downstream) of conotoxin genes, the content of the Gypsy and unclassified families that belong to LTRs in introns adjacent to the mature peptide was also markedly higher than that of introns adjacent to the pro-peptide (Fig. 7C and D, Table S6). However, there was no significant difference between mature peptide and pro-peptide for the retrotransposon of LINEs (data not shown).

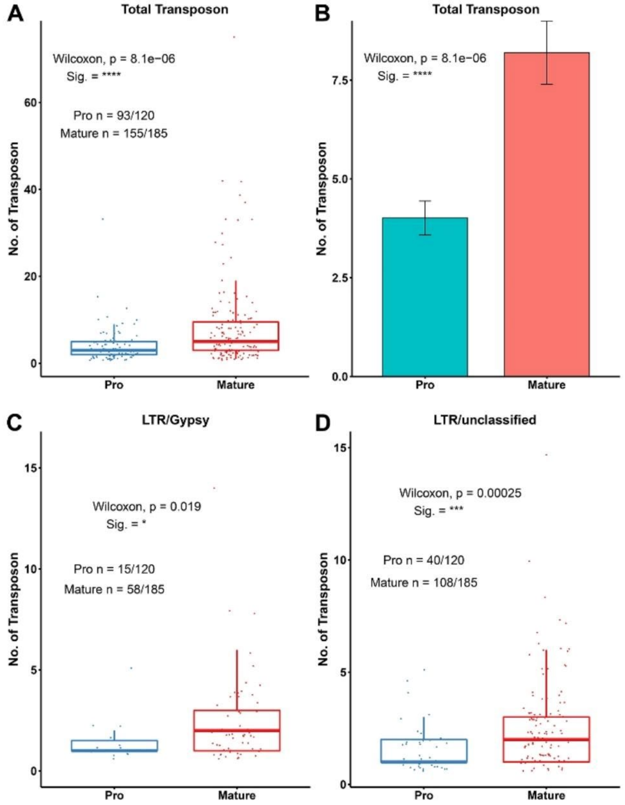

Fig. 7 Content of TEs in introns of conotoxin genes in *C. betulinus*. **A - B.** Content of TEs in introns that adjacent to the conotoxin pro- and mature peptide. **C - D.** Content of LTRs in introns that adjacent to the conotoxin pro- and mature peptide. **Pro**: pro-peptide of conopeptides. **Mature**: mature peptide of conopeptide. Significant differences were performed using Wilcoxon method, and indicated at $p<0.05$ (*), $p<0.001$ (***) or $p<1e\text{-}5$ (****)

Introgressive hybridization of conotoxin genes in *Conus*

Introgression among conotoxin genes was detected based on the exon capture targeting sequencing of conotoxin gene loci from 32 *Conus* species [7]. The introgression signals detected by D-statistic are shown in Fig. 8. Most of the *Conus* species pairs showed significant introgression signals. For instance, strong evidence of introgression was detected in *C. capitaneus* and *C. virgo* ($D=0.4403$, *p*-value = 1.55e-10), *C. capitaneus* and *C. marmoreus* ($D=0.3868$, *p*-value = 5.50e-11), *C. imperialis* and *C. papilliferus* ($D=0.3510$, *p*-value = 2.53e-09), and *C. arenatus* and *C. quercinus* ($D=0.3396$, *p*-value = 1.91e-07). These findings strongly suggested that there was significant gene flow of conotoxin genes between *Conus* species.

Discussion

In this study, we conducted transcriptome assembly for 34 *Conus* species. However, some species exhibited lower completeness of BUSCOs, which could be attributed to variations in sequencing depth coverage or the tissue types used for sequencing (Table S1). Additionally, the detection of conotoxins varied significantly among the different *Conus* species (Table 2), consistent with previous findings [17–19]. In *C. quercinus*, significant differences in the classes of conotoxin gene superfamilies between the venom duct, venom bulb, and salivary gland were observed, and the transcript activity of conotoxins was lower in both the venom bulb and salivary gland, suggesting that the venom duct is the primary site of conotoxin production [20]. Moreover, significant variations in conotoxins were identified among different individuals of *C. magus*, highlighting the highly diverse nature of conotoxins [21]. In the present study, it also showed that the expression characteristics of conotoxin genes in the venom ducts of *C. betulinus* with different body lengths were significantly diverse, suggesting that the different transcriptional expression or regulation patterns of conotoxin genes during the developmental phases may be one of the factors causing the diversification of conotoxins. Furthermore, we assessed Ka (nonsynonymous nucleotide substitutions) and Ks (synonymous nucleotide substitutions) values for conotoxin genes across the 34 *Conus* species. These parameters play a crucial role in molecular evolutionary analysis, with Ka/Ks > 1, Ka/Ks = 1, and Ka/Ks < 1 generally indicating positive selection, neutral mutation, and purifying selection, respectively [22]. In our study, we found that certain conotoxin genes from the 34 *Conus* species exhibited Ka/Ks values > 1 ($p<0.01$, Fig. 1), indicating positive selection acting on these genes. This finding aligns with the hypervariability observed in conopeptides [5]. Positive selection facilitates the spread of advantageous mutations, while purifying selection prevents the propagation

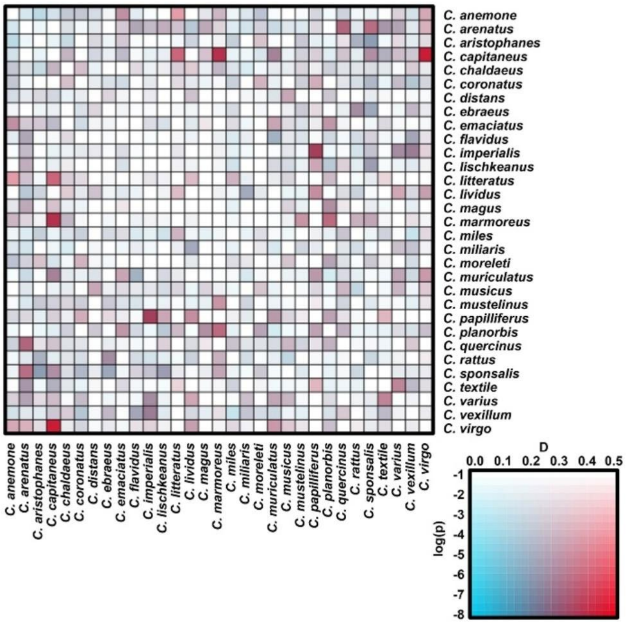

Fig. 8 Paterson's *D* (ABBA-BABA) statistic test of introgression of conotoxin genes in *Conus*. **Legend heatmap**: *D*-statistic value in abscissa and *p* value that transformed by logarithm in ordinate. Redder colors in grid indicate higher introgression level

of detrimental mutations [23]. Therefore, positive selection may contribute to the genetic diversity of conotoxin genes. Likewise, the exons coding the mature peptide of conotoxins exhibited approximately three times higher divergence than their flanking non-coding regions [7].

Conotoxin genes in *Conus* typically consist of 1–6 exons [7]. Our findings revealed that conotoxin genes in *C. betulinus* exhibited a range of exon numbers from 1 to 5, with a predominant composition of 3 exons, similar to *C. ventricosus* [11]. However, it is important to note that these three exons may not align precisely with the three structural domains of conopeptides, namely the signal peptide, pro-peptide, and mature peptide [11]. Interestingly, we observed a longer intron structure in conotoxin genes of *C. betulinus* compared to non-conotoxin genes. It has been observed that Metazoa genomes are enriched with introns, which can provide additional binding sites for transcriptional regulatory elements and facilitate gene diversification through exon recombination [13]. In fungal mitochondria, self-splicing introns have been implicated in increasing the genetic diversity of exons flanking them, suggesting that intron mobility directly influences host gene diversity [15]. Moreover, genes expressed abundantly in the nervous system often exhibit intron and gene size expansion, implying that the unique attributes of neurons may facilitate the evolution of neuronal genes [16]. Consequently, it is worth investigating whether introns in conotoxin genes influence the diversity of conotoxins.

Our results indicated a rapid expansion of two reverse transcriptases and one transposable element-derived protein in *C. betulinus*. Reverse transcriptase is a key enzymatic domain found in all autonomous retrotransposons, as it catalyzes the process of reverse transcription effectively [24]. Class I retrotransposons, one of the major classes of transposable elements (TEs), rely on the activity of reverse transcriptases and integrases [25]. Interestingly, in concurrence with the expansion of TE-related gene families, TEs were found to be highly prevalent in introns adjacent to the hypervariable mature peptide of conotoxins. This suggests that TE hotspots in these specific introns may contribute to the high variability observed in the mature peptide of conotoxins. Similarly, conotoxin genes in *C. ventricosus* are typically found in regions that harbor Class I retrotransposons (Gypsy, Penelope, etc.) and Class II DNA transposons (Tc1-Mariner, etc.) [11]. Consistent with previous studies, our findings indicated that the content of TEs in the flanking regions (upstream and downstream) of conotoxin genes in both *C. betulinus* and *C. ventricosus* was significantly higher compared to non-conotoxin genes (Table S4 and Table S5). Notably, the proportion of the Gypsy superfamily in the flanking regions (combined with upstream and downstream) of conotoxin genes was particularly prominent (9.45% of major TEs in *C. betulinus* and of that, 15.18% in *C. ventricosus*, Table S4 and Table S5), and it belonged to the LTRs of Class I retrotransposons. Given the critical roles of TEs in genome evolution and function, including genome mutations, rearrangements, and the promotion of new gene formation [14], we hypothesize that TE hotspots in crucial regions may be associated with the high diversity and hypervariability observed in conopeptides.

Furthermore, extensive research has revealed that gene flow between genetically distinct populations is a common occurrence in nature. In fact, it has been observed that introgressive hybridization between species can confer selective advantages to the receiving population, serving as a driving force behind the evolution of adaptive phenotypes [26]. The genus *Conus*, estimated to comprise approximately 700 species [1], has undergone rapid speciation through adaptive radiation, which likely promotes the occurrence of introgressive hybridization among different *Conus* lineages [27]. This notion is supported by studies documenting hybridization and introgression in various genetic regions, including mitochondrial genomes, nuclear gene regions, and conotoxin loci, within several *Virroconus* species. These findings strongly suggest that introgressive hybridization plays a significant role in the adaptive radiation of Conidae [28]. Similarly, our study provides clear evidence of introgressive hybridization involving conotoxin genes among multiple *Conus* species, indicating that introgressive hybridization is a frequent phenomenon within the genus *Conus*. This phenomenon is likely a contributing factor to the observed genetic diversity in conotoxin genes.

Conclusions

Our study provides valuable insights into the genetic diversity and evolution of conotoxin genes in *Conus* species. We observed species-specific variations, evidence of positive selection, and higher divergence in conotoxin coding regions. Notably, our investigation uncovered transposable element hotspots in the flanking regions (both upstream and downstream) of conotoxin genes, as well as in the introns adjacent to the highly diverse mature peptide of conotoxins. It implies that these transposable element-rich regions play a crucial role in driving the extensive diversity observed in conopeptides. Additionally, our study detected robust signals of introgressive hybridization involving conotoxin genes across numerous species of *Conus*, highlighting the significant impact of introgressive hybridization on the genetic diversity of conotoxin genes and the overall evolution of the *Conus* genus. These findings contribute to our understanding of the molecular mechanisms underlying the hypervariability of conotoxin genes.

Materials and methods

Assembly of *Conus* transcriptomes

Transcriptome sequencing datasets of 34 *Conus* species were downloaded from NCBI (Table S1). Adapters and low quality bases of the sequencing reads were filtered using Trimmomatic-v0.36 [29] with as parameters: "LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:80", and the quality of reads were checked by FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Transcripts of each species were assembled using Trinity-v2.6.6 [30] with as parameters: "--min_kmer_cov 2 --min_glue 3 --no_normalize_reads", followed by clustering of the predicted transcripts into unigenes by using Corset-v1.09 [31] with default parameters, respectively. The N50 of each resulting transcriptome species were calculated using a homemade Perl script and the completeness evaluation was performed using BUSCO-v5.4.7 [32] with default parameters using metazoa_odb10. TransDecoder-v5.3.0 (https://github.com/TransDecoder/) was used for prediction of ORF in unigenes, followed by functional annotation using eggNOG-mapper-v2.0.1 [33] with as parameters: "--target_orthologs all -m diamond".

Prediction and identification of conotoxins in *Conus* transcriptomes

Mapping of CDS sequences of 34 *Conus* species against the reference conotoxin peptides from ConoServer database [34] was performed by using Diamond-v0.9.22 [35] with as parameters: "-p 30 -k 10 -e 1e-5". Also, the Hidden Markov Model (HMM) method was used to identify the conotoxin candidates in each species of *Conus*. A profile HMM from previous research [36] was used as reference, and prediction of conotoxins was performed by HMMER-v3.1b2 [37] with as parameters: "-E 1e-5 --domE 1e-5". All candidate conotoxins predicted from homology and HMM search in each species of *Conus* were merged and subjected to redundancy removal. Furthermore, non-redundant candidate conotoxins were searched against the non-redundant protein sequences database (NR) using Diamond-v0.9.22, and non-conotoxins were filtered by manual curation. Finally, domains in each conotoxins were identified by ConoPrec [34].

Calculation of Ka/Ks in conotoxins

OrthoFinder-v2.2.1 [38] with the parameters: "-S diamond -og -t 30 -a 30" was used to infer orthogroup of the predicted conotoxins from 34 *Conus* species. And combined with the corresponding CDS of each conotoxins, ParaAT-v2.0 [39] was used to calculate the Ka/Ks within each of the conotoxins orthogroup pairs with as parameters: "-p proc -m mafft -f axt -g -t -k", followed by filtering with the threshold of p-value (Fisher) <0.01. The visualization of Ka/Ks results was performed using ggplot2 [40].

Genome reassembly of *C. betulinus*

Whole genome sequencing datasets of *C. betulinus* from recently published research [10], including Illumina and PacBio sequencing platform, were obtained from NCBI (PRJNA578609). For Illumina sequencing datasets, adapters and low quality bases of reads were filtered using Trimmomatic-v0.36 and checked with FastQC, followed by removing duplication by using Nubeam-dedup [41] with default parameters. In addition, 1 million reads in each sequencing dataset were randomly selected using seqtk (https://github.com/lh3/seqtk), and against the non-redundant nucleotide database (NT) using blastn [42] to make taxonomic assignments. A homemade Perl script was used to summarize the taxa of reads and analyze the potential contamination organisms, followed by removing contaminated reads using bbmap-v38.08 [43] with default parameters. For PacBio sequencing datasets, combined with the pretreatment of Illumina sequencing reads, FMLRC2-v0.1.3 [44] was used to perform the error correction of reads with as parameters: "-t 40 -C 10". 50,000 corrected reads in each dataset were randomly selected by a homemade Perl script, and blastn was used against the NT database. Contaminated organisms were summarized, followed by removal using minimap2-v2.11 [45] with default parameters.

PacBio raw reads were used to perform genome assembly using wtdbg2-v2.5 [46] with as parameters: "-A -S 3 -X

50 -l 5000 -x sq -g 2.5 g -t 40", and consensus sequences were generated and polished using wtpoa-cns. Furthermore, contigs were scaffolded using LRScaf-v1.1.11 [47], and polished using NextPolish-v1.3.1 [48] with corrected Illumina reads. Finally, sequences that were longer than 1 kb were retained, and blastn was used against the NT database and the mitochondrion genome of *C. betulinus* (MG924728.1) to make taxonomic assignments. Possible contamination, such as bacteria or mitochondrion, was manually filtered.

Repeats and genome annotation of *C. betulinus*

De novo repeat library was constructed using RepeatModeler-v1.0.11 [49]. Additionally, LTR_Finder-v1.07 [50] with as parameters: "-D 20000 -d 1000 -L 7000 -l 100 -p 20 -M 0.9 -C", and LTRharvest [51] with as parameters: "-similar 90 -vic 10 -seed 20 -seqids yes -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6" was used to identify the LTR retrotransposons, respectively. And a high-quality LTR library was generated using LTR_retriever-v2.9.0 [52] with default parameters. Subsequently, combined with the results of RepeatModeler, LTR_retriever and Repbase database [53], transposable elements (TEs) were identified and classified by performing with RepeatMasker-v4.0.7 [54]. Furthermore, gene structures were firstly predicted using MAKER2 [55] together by *de novo*, homology and RNA-seq methods. In addition, PASA-v2.3.3 [56] and Stringtie-v1.3.4d [57] were used to optimize the predicted gene structures, respectively. Finally, gene models predicted from MAKER2, PASA and Stringtie were integrated by EVidenceModeler [58] into a comprehensive and non-redundant set of gene structures.

Conotoxin genes in *C. betulinus* were identified as described above. Additionally, gene structures, especially the coordinates of exons of conotoxin genes, were manually checked and revised using Exonerate [59]. Finally, completeness evaluation of all protein-coding genes was performed using BUSCO-v5.4.7 [32] using metazoan_odb10.

Whole genome duplication and gene family analysis of *C. betulinus*

WGDdetector-v1.1 [60] was used to perform the whole genome duplication (WGD) analysis of *C. betulinus*. Meanwhile, conserved homologous gene blocks in *C. betulinus* was detected using MCScanX [61], and visualized using R package RIdeogram [62]. Additionally, recently published genome of *C. ventricosus* (GCA_018398815.1) [11] was also used for WGD and conserved homologous gene blocks analysis.

Orthogroups were identified among 8 selected species of gastropods, namely, *Achatina fulica* [63], *Aplysia californica* (GCF_000002075.1), *Biomphalaria glabrata* (GCA_000457375.1), *Chrysomallon squamiferum* [64], *Elysia chlorotica* (GCA_003991915.1), *Lottia gigantea* (GCF_000327385.1), *Pomacea canaliculata* (GCA_003073045.1) and *C. betulinus* using OrthoFinder-v2.2.1 [38] with the parameters of "-S diamond -M msa -t 30 -a 30 -T fasttree". Subsequently, the absolute rates of molecular evolution and divergence times were inferred using r8s-v1.81 [65] with as parameters: "-s 961030 -p 'Achatina_fulica,Pomacea_canaliculata' -c '421'", followed by identifying gene family expansion and contraction using CAFÉ-v4.2.1 [66] with default parameters. Finally, rapidly evolving family genes were functional annotated using Diamond-v0.9.22 [35] against with KEGG (Kyoto Encyclopedia of Genes and Genomes), NR and UniProt databases with the threshold of E-value 1e-5.

Structure and expression of conotoxin genes in *C. betulinus*

Gene full length, exon and intron length, and exon number in conotoxin and non-conotoxin genes were summarized using a homemade Perl script. Visualization of statistical results was performed using ggplot2 [40].

Transcriptomes of *C. betulinus* with multiple specimens and tissues were obtained from NCBI (PRJNA290540) [36]. Three of the samples are from venom duct tissue of specimens with different body lengths, namely small, middle, and big; another sample is from venom duct that mixed with these tissues; and the last one is from venom bulb tissue of the middle specimen. Quality control of sequencing reads was performed as described above. Clean reads of each sample were mapped into genome of *C. betulinus* using HISAT2-v2.1.0 [67] with default parameters, respectively. Subsequently, paired read counts were quantified using featureCounts-v1.6.2 [68], and TPM (Transcripts Per Kilobase Million) method was used to normalize and calculate the expression of genes using a homemade Perl script. The TPM values of conotoxin genes were used to perform hierarchical clustering using pheatmap with z-score normalization and to compare the expression of conotoxin genes between different specimens using ggtern [69]. Meanwhile, alternative splicing between different specimens and tissues was performed using LeafCutter-v0.2 [70] and filtered with $p<0.05$ for significantly alternative splicing sites, followed by visualization using ggsashimi [71].

Introgression analysis of conotoxin genes in *Conus*

Targeted sequencing data sets of conotoxin genes from 32 cone snails (Conidae) that were published in previous research [7] were obtained from NCBI (PRJNA437715). Quality control of reads was performed as described above. Clean reads from each cone snail were mapped into the chromosome-level genome of *C. ventricosus* [11] using BWA-v0.7.17 [72]. SNPs in each cone snail were identified using GATK-v4.0.5.2 [73] and filtered with the following parameters: “QD<2.0 || MQ<40.0 || FS>60.0 || SOR>3.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0”. Finally, introgression between cone snails was performed using Dsuite-v0.4 [74].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09689-4.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Supplementary Material 7

Supplementary Material 8

Supplementary Material 9

Acknowledgements

Not applicable.

Author contributions

Jian-Wei Zheng: Conceptualization, Investigation, Validation, Visualization, Writing - original draft. Yang Lu: Conceptualization, Formal analysis, Data curation. Yu-Feng Yang: Conceptualization, Data curation. Dan Huang: Conceptualization, Data curation. Da-Wei Li: Conceptualization, Data curation. Xiang Wang: Conceptualization, Data curation. Yang Gao: Conceptualization, Supervision. Wei-Dong Yang: Conceptualization, Supervision. Yuanfang Guan: Conceptualization, Supervision. Hong-Ye Li: Conceptualization, Supervision, Writing - review & editing, Funding acquisition. All authors have read and approved the final manuscript.

Funding

This work was supported by the Natural Science Foundation of Guangdong Province (No. 2019B1515120062) and the National Natural Science Foundation of China (No. 42006125).

Data Availability

Transcriptome sequencing datasets of 34 *Conus* species are listed in Table S1 (Supplementary File 1). Whole genome sequencing datasets of *C. betulinus* used in this study are available from published research (PRJNA578609) [10], and targeted sequencing of conotoxin genes from 32 Conidae genomes used in this study are available from published research (PRJNA437715) [7]. All datasets are freely available on NCBI. The predicted protein-coding genes and conotoxins from the reassembled genome of *C. betulinus* in the present study are available in Supplementary Files 7 to 9. Homology clustering results for conotoxin genes that were identified from the reassembled and published genomes of *C. betulinus* are listed in Table S3 (Supplementary File 3). The annotation of repeat elements in the reassembled genome of *C. betulinus* are listed in Table S2 (Supplementary File 2). The content of major transposon elements in flanking regions of protein-coding genes and introns in conotoxin genes in *C. betulinus* and *C. ventricosus* was summarized in Table S4 to Table S6 (Supplementary File 4 to 6).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors have declared no competing interests.

Received: 13 February 2023 / Accepted: 20 September 2023
Published online: 09 October 2023

References

1. Gao B, Peng C, Yang J, Yi Y, Zhang J, Shi Q. Cone snails: a big store of conotoxins for novel drug discovery. Toxins. 2017;9(12):397. https://doi.org/10.3390/toxins9120397.
2. Jin AH, Muttenthaler M, Dutertre S, Himaya SWA, Kaas Q, Craik DJ, et al. Conotoxins: chemistry and biology. Chem Rev. 2019;119(21):11510–49. https://doi.org/10.1021/acs.chemrev.9b00207.
3. Olivera BM, Showers Corneli P, Watkins M, Fedosov A. Biodiversity of cone snails and other venomous marine gastropods: evolutionary success through neuropharmacology. Annu Rev Anim Biosci. 2014;2:487–513. https://doi.org/10.1146/annurev-animal-022513-114124.
4. Robinson SD, Norton RS. Conotoxin gene superfamilies. Mar Drugs. 2014;12(12):6058–101. https://doi.org/10.3390/md12126058.
5. Conticello SG, Gilad Y, Avidan N, Ben-Asher E, Levy Z, Fainzilber M. Mechanisms for evolving hypervariability: the case of conopeptides. Mol Biol Evol. 2001;18(2):120–31. https://doi.org/10.1093/oxfordjournals.molbev.a003786.
6. Hu H, Bandyopadhyay PK, Olivera BM, Yandell M. Characterization of the *Conus bullatus* genome and its venom-duct transcriptome. BMC Genom. 2011;12:60. https://doi.org/10.1186/1471-2164-12-60.
7. Phuong MA, Mahardika GN. Targeted sequencing of venom genes from cone snail genomes improves understanding of conotoxin molecular evolution. Mol Biol Evol. 2018;35(5):1210–24. https://doi.org/10.1093/molbev/msy034.
8. Barghi N, Concepcion GP, Olivera BM, Lluisma AO. Structural features of conopeptide genes inferred from partial sequences of the *Conus tribblei* genome. Mol Genet Genomics. 2016;291(1):411–22. https://doi.org/10.1007/s00438-015-1119-2.
9. Andreson R, Roosaare M, Kaplinski L, Laht S, Kõressaar T, Lepamets M, et al. Gene content of the fish-hunting cone snail *Conus consors*. bioRxiv. 2019. https://doi.org/10.1101/590695.
10. Peng C, Huang Y, Bian C, Li J, Liu J, Zhang K, et al. The first *Conus* genome assembly reveals a primary genetic central dogma of conopeptides in *C. betulinus*. Cell Discov. 2021;7(1):11. https://doi.org/10.1038/s41421-021-00244-7.
11. Pardos-Blas JR, Irisarri I, Abalde S, Afonso CML, Tenorio MJ, Zardoya R. The genome of the venomous snail *Lautoconus ventricosus* sheds light on the origin of conotoxin diversity. GigaScience. 2021;10(5):giab037. https://doi.org/10.1093/gigascience/giab037.
12. Dutertre S, Jin AH, Kaas Q, Jones A, Alewood PF, Lewis RJ. Deep venomics reveals the mechanism for expanded peptide diversity in cone snail venom. Mol Cell Proteomics. 2013;12(2):312–29. https://doi.org/10.1074/mcp.M112.021469.
13. Grau-Bové X, Torruella G, Donachie S, Suga H, Leonard G, Richards TA, et al. Dynamics of genomic innovation in the unicellular ancestry of animals. eLife. 2017;6:e26036. https://doi.org/10.7554/eLife.26036.
14. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19(1):199. https://doi.org/10.1186/s13059-018-1577-z.
15. Repar J, Warnecke T. Mobile introns shape the genetic diversity of their host genes. Genetics. 2017;205(4):1641–8. https://doi.org/10.1534/genetics.116.199059.
16. McCoy MJ, Fire AZ. Intron and gene size expansion during nervous system evolution. BMC Genom. 2020;21(1):360. https://doi.org/10.1186/s12864-020-6760-4.
17. Yao G, Peng C, Zhu Y, Fan C, Jiang H, Chen J, et al. High-throughput identification and analysis of novel conotoxins from three vermivorous cone snails by transcriptome sequencing. Mar Drugs. 2019;17(3):193. https://doi.org/10.3390/md17030193.
18. Abalde S, Tenorio MJ, Afonso CML, Zardoya R. Comparative transcriptomics of the venoms of continental and insular radiations of West African cones. Proc R Soc B. 2020;287(1929):20200794. https://doi.org/10.1098/rspb.2020.0794.
19. Barghi N, Concepcion GP, Olivera BM, Lluisma AO. Comparison of the venom peptides and their expression in closely related *Conus* species: insights into

adaptive post-speciation evolution of *Conus* exogenomes. Genome Biol Evol. 2015;7(6):1797–814. https://doi.org/10.1093/gbe/evv109.

20. Gao B, Peng C, Zhu Y, Sun Y, Zhao T, Huang Y, et al. High throughput identification of novel conotoxins from the vermivorous oak cone snail (*Conus quercinus*) by transcriptome sequencing. Int J Mol Sci. 2018;19(12):3901. https://doi.org/10.3390/ijms19123901.
21. Pardos-Blas JR, Irisarri I, Abalde S, Tenorio MJ, Zardoya R. Conotoxin diversity in the venom gland transcriptome of the Magician's cone, *Pionoconus magus*. Mar Drugs. 2019;17(10):553. https://doi.org/10.3390/md17100553.
22. Yang Z. Computational molecular evolution. Oxford University Press; 2006. https://doi.org/10.1093/acprof:oso/9780198567028.001.0001.
23. Page RDM, Holmes EC. Molecular evolution: a phylogenetic approach. Wiley; 2009.
24. Gabriel A. Retrotransposons and human disease. World Scientific; 2022.
25. Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. Mol Ecol. 2019;28(6):1537–49. https://doi.org/10.1111/mec.14794.
26. Svedberg J, Shchur V, Reinman S, Nielsen R, Corbett-Detig R. Inferring adaptive introgression using hidden markov models. Mol Biol Evol. 2021;38(5):2152–65. https://doi.org/10.1093/molbev/msab014.
27. Seehausen O. Hybridization and adaptive radiation. Trends Ecol Evol. 2004;19(4):198–207. https://doi.org/10.1016/j.tree.2004.01.003.
28. Wood AW, Duda TF. Jr. Reticulate evolution in Conidae: evidence of nuclear and mitochondrial introgression. Mol Phylogen Evol. 2021;161:107182. https://doi.org/10.1016/j.ympev.2021.107182.
29. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
30. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2011;29(7):644–52. https://doi.org/10.1038/nbt.1883.
31. Davidson NM, Oshlack A, Corset. Enabling differential gene expression analysis for *de novo* assembled transcriptomes. Genome Biol. 2014;15(7):410. https://doi.org/10.1186/s13059-014-0410-6.
32. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2. https://doi.org/10.1093/bioinformatics/btv351.
33. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021;38(12):5825–9. https://doi.org/10.1093/molbev/msab293.
34. Kaas Q, Yu R, Jin A-H, Dutertre S, Craik DJ. ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. Nucleic Acids Res. 2012;40(D1):D325–D30. https://doi.org/10.1093/nar/gkr886.
35. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2014;12:59–60. https://doi.org/10.1038/nmeth.3176.
36. Peng C, Yao G, Gao BM, Fan CX, Bian C, Wang J, et al. High-throughput identification of novel conotoxins from the Chinese tubular cone snail (*Conus betulinus*) by multi-transcriptome sequencing. GigaScience. 2016;5:17. https://doi.org/10.1186/s13742-016-0122-9.
37. Eddy SR. Accelerated profile HMM searches. PLoS Comp Biol. 2011;7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
38. Emms DM, Kelly S, OrthoFinder. Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157. https://doi.org/10.1186/s13059-015-0721-2.
39. Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, et al. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. Biochem Biophys Res Commun. 2012;419(4):779–81. https://doi.org/10.1016/j.bbrc.2012.02.101.
40. Wickham H. ggplot2: elegant graphics for data analysis. New York: In.: Springer-Verlag; 2016.
41. Dai H, Guan Y. Nubeam-dedup: a fast and RAM-efficient tool to de-duplicate sequencing reads without mapping. Bioinformatics. 2020;36(10):3254–6. https://doi.org/10.1093/bioinformatics/btaa112.
42. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinform. 2009;10(1):421. https://doi.org/10.1186/1471-2105-10-421.
43. Bushnell B. BBMap: A fast, accurate, splice-aware aligner. Berkeley, CA (United States): In.: Lawrence Berkeley National Lab.(LBNL); 2014.
44. Wang JR, Holt J, McMillan L, Jones CD. FMLRC: hybrid long read error correction using an FM-index. BMC Bioinform. 2018;19(1):50. https://doi.org/10.1186/s12859-018-2051-3.
45. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191.
46. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17(2):155–8. https://doi.org/10.1038/s41592-019-0669-3.
47. Qin M, Wu S, Li A, Zhao F, Feng H, Ding L, et al. LRScaf: improving draft genomes using long noisy reads. BMC Genom. 2019;20(1):955. https://doi.org/10.1186/s12864-019-6337-2.
48. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 2020;36(7):2253–5. https://doi.org/10.1093/bioinformatics/btz891.
49. Smit AF, Hubley R. RepeatModeler Open-1.0. 2008http://www.repeatmaskerorg.
50. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35(suppl2):W265–W8. https://doi.org/10.1093/nar/gkm286.
51. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinform. 2008;9(1):18. https://doi.org/10.1186/1471-2105-9-18.
52. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018;176(2):1410–22. https://doi.org/10.1104/pp.17.01310.
53. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110(1–4):462–7. https://doi.org/10.1159/000084979.
54. Smit AF, Hubley R, Green P. RepeatMasker Open-4.0. 2013http://www.repeatmaskerorg.
55. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinform. 2011;12(1):491. https://doi.org/10.1186/1471-2105-12-491.
56. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr., Hannick LI, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31(19):5654–66. https://doi.org/10.1093/nar/gkg770.
57. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290. https://doi.org/10.1038/nbt.3122.
58. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008;9(1):R7. https://doi.org/10.1186/gb-2008-9-1-r7.
59. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinform. 2005;6(1):31. https://doi.org/10.1186/1471-2105-6-31.
60. Yang Y, Li Y, Chen Q, Sun Y, Lu Z. WGDdetector: a pipeline for detecting whole genome duplication events using the genome or transcriptome annotations. BMC Bioinform. 2019;20(1):75. https://doi.org/10.1186/s12859-019-2670-3.
61. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40(7):e49. https://doi.org/10.1093/nar/gkr1293.
62. Hao Z, Lv D, Ge Y, Shi J, Weijers D, Yu G, et al. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Comput Sci. 2020;6:e251. https://doi.org/10.7717/peerj-cs.251.
63. Guo Y, Zhang Y, Liu Q, Huang Y, Mao G, Yue Z, et al. A chromosomal-level genome assembly for the giant African snail *Achatina fulica*. GigaScience. 2019;8(10). https://doi.org/10.1093/gigascience/giz124.
64. Sun J, Chen C, Miyamoto N, Li R, Sigwart JD, Xu T, et al. The scaly-foot snail genome and implications for the origins of biomineralised armour. Nat Commun. 2020;11(1):1657. https://doi.org/10.1038/s41467-020-15522-3.
65. Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics. 2003;19(2):301–2. https://doi.org/10.1093/bioinformatics/19.2.301.
66. Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 2013;30(8):1987–97. https://doi.org/10.1093/molbev/mst100.
67. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15. https://doi.org/10.1038/s41587-019-0201-4.
68. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923–30. https://doi.org/10.1093/bioinformatics/btt656.

69. Hamilton NE, Ferry M. ggtern: ternary diagrams using ggplot2. J Stat Softw Code Snippets. 2018;87(3):1–17. https://doi.org/10.18637/jss.v087.c03.
70. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet. 2018;50(1):151–8. https://doi.org/10.1038/s41588-017-0004-9.
71. Garrido-Martín D, Palumbo E, Guigó R, Breschi A. ggsashimi: sashimi plot revised for browser- and annotation-independent splicing visualization. PLoS Comp Biol. 2018;14(8):e1006360. https://doi.org/10.1371/journal.pcbi.1006360.
72. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.
73. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From fastq data to high confidence variant calls: the genome analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1110):11. https://doi.org/10.1002/0471250953.bi1110s43.
74. Malinsky M, Matschiner M, Svardal H. Dsuite - fast D-statistics and related admixture evidence from VCF files. Mol Ecol Resour. 2020;21(2):584–95. https://doi.org/10.1111/1755-0998.13265.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.