Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A multi-tissue transcriptome and expression atlas of *Astacus leptodactylus*: assembly and annotation

Berk Sağlam
Hacettepe University
Mehmet Akgun
Hacettepe University
Bora Ergin
Intergen Genetics and Rare Diseases Diagnosis Center
Nuhan Purali
npurali@hacettepe.edu.tr

Hacettepe University https://orcid.org/0000-0002-0604-2303

Data Note

Keywords: Transcriptome assembly, genome assembly, annotation, expression atlas, crayfish

Posted Date: July 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10355696/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

We present a multi-tissue transcriptome of the narrow-clawed crayfish *Astacus leptodactylus*, assembled from 37 paired-end RNA-seq samples spanning 12 tissue types including ganglia, sensory neuron, and the abdominal receptor organ — a mechanosensory preparation comprising two mechanoreceptor neurons and associated receptor muscle fibers. A draft genome assembled from PacBio HiFi long reads supported genome-guided assembly.

Genome-guided, Trinity, and rnaSPAdes assemblies were constructed independently and subjected to sequential redundancy reduction using EvidentialGene tr2aacds and MMseqs2 clustering. The final merged transcriptome comprises 93,345 transcripts (N50 = 2,498 bp, mean = 1,656 bp, GC = 45.4%), achieving BUSCO completeness of C:89.5%[S:82.6%,D:7.0%],F:7.7%,M:2.8% (crustacea_odb12, n = 1,536). Functional annotation identified 25,729 annotated transcripts with 14,260 unique Gene Ontology terms. Tissue-specific expression profiling revealed distinct clustering of sensory, digestive, and musculoskeletal tissues. Functional completeness was independently validated using ten ion channel sequences previously cloned from this species: all ten were recovered at 84–100% nucleotide identity, with SwissProt annotation and tissue-specific expression patterns fully consistent with established channel physiology. This resource provides a foundation for functional genomics and ion channel research in non-model crustaceans.

Background & Summary

The narrow-clawed crayfish, *Astacus leptodactylus* (Eschscholtz, 1823), is a freshwater decapod crustacean widely distributed across Eastern Europe, Turkey, and Western Asia. It represents one of the most economically and ecologically important freshwater invertebrates in the region, supporting both commercial aquaculture and wild fisheries, and serving as a keystone species in freshwater ecosystems. Despite its ecological relevance and long history as a model organism in crustacean physiology and neuroscience, genomic and transcriptomic resources for *Astacus leptodactylus* remain severely limited compared to other model arthropods such as *Drosophila melanogaster* (Lnenicka, 2020) or other crustaceans including *Daphnia pulex* (Colbourne et al., 2011), *Penaeus vannamei* (Liao et al., 2025), *Procambarus clarkii* (Liao et al., 2024), and *Cherax quadricarinatus* (Liu et al., 2024).

Crustacean transcriptomes are inherently challenging to assemble due to large genome sizes, high heterozygosity, and the complexity of alternative splicing, particularly in neuronal tissues (Kao et al., 2016; Hyde et al., 2019; Zhang et al., 2019; Zhang et al., 2014). De novo transcriptome assembly without a high-quality reference genome requires careful integration of multiple assembly strategies, expression-based filtering, and redundancy reduction to produce a biologically meaningful and computationally tractable reference. Previous transcriptomic studies in crustaceans have been limited in scope, typically targeting a single tissue or specific gene families, leaving a comprehensive multi-tissue reference absent from public databases (Northcutt et al., 2016). Recently, a full-length transcriptome for *Pontastacus leptodactylus* based on PacBio Iso-Seq sequencing of eight tissues was reported (Chen et al., 2026);

however, that study did not include nervous system tissues and was limited to isoform-level characterization without a multi-sample expression atlas. The absence of neural tissue sampling has functional consequences: voltage-gated sodium channels, which are the molecular basis of action potential generation and are exclusively expressed in excitable cells, were not represented in that dataset but are captured in our assembly at 99–100% nucleotide identity to a previously cloned *Astacus leptodactylus* Nav sequence (KT187402.1), with expression confirmed predominantly in sensory neuron, ganglia, eye tissues.

Here we present a multi-tissue transcriptome of *Astacus leptodactylus* assembled from 37 paired-end RNA-seq samples representing 12 tissue types: antenna, antennule, eye, ganglia, gill, heart, hepatopancreas, muscle, stomach, testis, receptor organ, and sensory neuron. The tissue panel was deliberately designed to provide broad coverage of sensory and nervous system gene expression. In particular, the abdominal receptor organ — a classical crustacean mechanosensory preparation comprising two mechanoreceptor neurons (slowly and rapidly adapting) and associated receptor muscle fibers (Alexandrowicz, 1951, Purali, 2005) — was included as a dedicated source of mechanosensory transcripts. To support genome-guided assembly, a draft genome was constructed from PacBio HiFi long reads using HIFIASM (Cheng et al., 2021).

Genome-guided (GGTA), Trinity, and rnaSPAdes assemblies were constructed independently and subjected to sequential redundancy reduction using EvidentialGene tr2aacds and MMseqs2 clustering. The rnaSPAdes and GGTA assemblies were selected for merging based on lower redundancy and higher short-read mapping rate. The final transcriptome comprises 93,345 transcripts with a total length of 154.5 Mb, mean length of 1,656 bp, and N50 of 2,498 bp, achieving a BUSCO completeness score of C:89.5%[S:82.6%,D:7.0%],F:7.7%,M:2.8% (crustacea_odb12, n = 1,536; Manni et al., 2021). Functional annotation using BLASTP against UniProt Swiss-Prot and hmmscan against Pfam identified 25,729 annotated transcripts, with 14,260 unique Gene Ontology terms assigned. Tissue-specific expression profiling across all 12 tissues revealed distinct clustering of sensory, digestive, and musculoskeletal tissues. The functional validity of the assembly was independently confirmed using ten ion channel sequences previously cloned and deposited in GenBank by our laboratory: all ten were recovered at 84–100% nucleotide identity, with SwissProt protein annotation and tissue-specific expression patterns in full agreement with established channel physiology — neural channels predominating in sensory neuron and ganglia, and excitation-contraction coupling channels predominating in muscle. This resource provides a foundation for functional genomics and ion channel research in non-model crustaceans.

Methods

Tissue collection, HMWgDNA extraction

Tissue samples were obtained from abdominal flexor muscle with careful handling to prevent contamination. High-molecular-weight genomic DNA (HMWgDNA) was extracted using two independent kits in parallel: the Nanobind PanDNA Kit (PacBio) and the Monarch HMW DNA Extraction Kit (New

England Biolabs). Both kits yielded comparable results; the Monarch kit product was selected for PacBio HiFi sequencing. The extracted DNA had a concentration of 180 ng/μL with A260/A280 and A260/A230 ratios of 2.11 and 0.77, respectively. Pulsed-field capillary electrophoresis confirmed high molecular weight with fragment size averages of 44,916 bp and 61,119 bp.

HIFI long Read data

HMWgDNA sample was processed and sequenced at Maryland Genomics, Institute for Genome Sciences, University of Maryland. The same sample was sequenced independently in two runs on a PacBio Revio system. The first run produced 4,356,481 reads with a mean read length of 22,885 bp, and the second run produced 4,266,544 reads with a mean read length of 23,094 bp. Median read quality was Q32. The combined total sequence output from both runs was 199 Gbp.

Genome assembly from HIFI reads.

HiFi reads from both sequencing runs were assembled jointly using HIFIASM (v0.25.0-r726) on the TRUBA HPC system, producing separate primary (15 Gb) and alternate contig (5.4 Gb) files. The primary assembly comprised 129,019 contigs with a total length of 16.05 Gb, an N50 of 153,636 bp, and a GC content of 44.6% (Shen et al., 2016).

To assess the effect of haplotype duplication on assembly quality, Purge_Dups (Guan et al., 2020) was applied at two cutoff settings: aggressive [10 30 150 250 10 465] and moderate [5 13 40 60 5 26], derived from k-mer-based estimates of true and observed coverage, respectively. As shown in Table 1, aggressive purging reduced assembly size to 6.97 Gb but caused a substantial loss of BUSCO completeness (C:22.2%), while moderate purging retained more sequence (9.31 Gb) at a similar completeness cost (C:25.7%). Merging the primary and alternate contig sets without purging yielded the highest BUSCO completeness (C:33.5%), and this combined assembly was therefore used as the reference for genome-guided transcriptome assembly.

Table 1
BUSCO completeness assessment (Crustacea_odb12, n = 1,536) of the Astacus leptodactylus genome assembly at sequential purging stages. Primary: unpurged primary contigs; Combined: merged primary and alternate contigs without purging.

Metric	Scenario 1 (Aggressive)	Scenario 2 (Moderate)	Primary	Combined
Assembly Size	6.97 Gb	9.31 Gb	16.05 Gb	21.80 Gb
Number of Contigs	41,385	64,136	129,019	229,863
N50	224 kbp	189 kbp	153 kbp	119 kbp
Purging Strategy	Aggressive	Moderate	None	None
Complete BUSCOs (C)	22.2%	25.7%	30.7%	33.5%
— Single-copy (S)	21.4%	24.9%	29.4%	30.1%
— Duplicated (D)	0.8%	0.8%	1.4%	3.4%
Fragmented (F)	19.1%	20.7%	29.9%	32.6%
Missing (M)	58.7%	53.6%	39.4%	33.9%
Complete + Fragmented	41.3%	46.4%	60.6%	66.1%

Tissue collection, RNA extraction, cDNA library construction and quality control

Biological samples were collected as described previously (Ergin et al., 2023; Coşkun & Purali, 2016; Ergin & Purali, 2018; Arslan et al, 2023) from twelve tissue types: antenna, antennule, eye, ganglia, gill, heart, hepatopancreas, muscle, stomach, testis, receptor organ (RO), and sensory neuron (SN). The receptor organ is a classical crustacean mechanosensory preparation comprising two mechanoreceptor neurons (slowly and rapidly adapting) and associated receptor muscle fibers (Alexandrowicz, 1951; Purali 2005). Owing to the limited tissue yield, 5–6 receptor organ preparations were pooled per sample and treated as a single biological sample without replicates. Total RNA was used for cDNA library synthesis using the Repli-G WTA Kit (Qiagen) as described previously (Ergin et al., 2022; Saglam et al., 2023; Sağsöz et al., 2023). Libraries were sequenced on Illumina (NovaSeq 6000) and MGI (DNBSEQ-T7) platforms. Parallel sequencing of the same sample on both platforms indicated no significant difference in efficiency or accuracy between the two technologies. A total of 37 paired-end RNA-seq samples from 12 tissues yielded 1.875 billion read pairs (Table 2).

Table 2
Depth of tissue specific PE short read data

Tissue	Replicates	Read Pairs (M)
Antenna	2	34.7
Antennule	2	37.4
Eye	2	39.6
Ganglia	4	262.0
Gill	2	44.1
Heart	3	84.4
Hepatopancreas	4	67.5
Muscle	4	244.0
RO	1	107.0
SN	8	872.8
Stomach	3	41.3
Testis	2	40.8
TOTAL	37	1875.5

Genome guided transcriptome assembly

A genome-guided transcriptome assembly (GGTA) was constructed using the *Astacus leptodactylus* draft genome as the reference. Primary and alternate contig assemblies generated by HIFIASM were merged to create a combined reference genome, maximising gene space representation.

A genome index was constructed using HISAT2 (v2.2.1) with default parameters. The 37 paired-end RNA-seq samples were aligned individually to this index in downstream transcriptome assembly mode (--dta), enabling splice-aware alignment. Alignment statistics were recorded using SAMtools flagstat, and SAM files were converted to coordinate-sorted BAM format and indexed using SAMtools (Danecek et al., 2021).

Transcripts were assembled independently for each sample using StringTie (v3.0.3) with a minimum transcript length of 200 bp, a minimum coverage depth of 1.5×, and a minimum isoform fraction of 0.05 (Pertea et al., 2015). Per-sample GTF files were merged into a non-redundant consensus annotation using StringTie --merge with the same filtering thresholds. Final transcript sequences were extracted from the combined draft genome using gffread, yielding 341,838 transcript sequences.

The raw assembly was subjected to sequential quality filtering and redundancy reduction. Transcripts shorter than 200 bp were removed, and the length-filtered assembly was processed with EvidentialGene tr2aacds (Gilbert, 2019), which selects the best open reading frame (ORF) per transcript based on length and completeness, collapses identical protein sequences, and performs BLASTN-based clustering to select representative transcripts. Coding sequences were predicted from the EvidentialGene okay set using TransDecoder (--single_best_only) to retain one representative protein per transcript. Predicted proteins were further clustered at 70% sequence identity and 80% coverage using MMseqs2 easy-cluster in greedy incremental mode (--cluster-mode 2, --cov-mode 1), retaining the longest sequence per cluster as the representative (Steinegger & Söding, 2018). BUSCO completeness (crustacea_odb12, n = 1,536) was assessed at each major filtering stage (Table 3).

The completeness of the genome-guided assembly was further assessed by calculating the Ortholog Hit Ratio (OHR) for the MMseqs2 representative set (O'Neil & Emrich, 2013). Each transcript was searched against the UniProt Swiss-Prot database using BLASTP (e-value ≤ 1×10^{-5}, best hit per transcript). OHR was calculated as the ratio of the BLASTP alignment length to the matched Swiss-Prot reference protein length, capped at 1.0. OHR distributions were summarized at thresholds of ≥ 0.50, ≥ 0.75, and ≥ 0.90 to characterize the proportion of full-length, near-complete, and highly complete coding sequences, respectively. Of 30,775 transcripts, 13,024 (42.3%) aligned to Swiss-Prot entries with a mean OHR of 0.34 (Fig. 1), consistent with the fragmented nature of the draft genome reference.

Table 3
Comparison of size and completeness at sequential steps of GGTA pipeline.

Step	**Transcripts/Proteins**	**BUSCO Score (Crustacea_odb12, n:1536)**
Raw StringTie (341,838 → 200bp filter)	341,835	C:28.0% [S:20.7%,D:7.3%],F:29.2%,M:42.8%
EvidentialGene okay set	63,743	C:27.5% [S:26.8%,D:0.8%],F:28.2%,M:44.3%
TransDecoder proteins	34,795	—
GGTA (MMseqs2 70% representatives)	30,771	C:27.1% [S:27.0%,D:0.2%],F:26.9%,M:46.0%

GGTA quality was further assessed using rnaQUAST (v2.3.2) against the draft *Astacus leptodactylus* genome. Of 30,771 representative transcripts, 29,703 (96.5%) mapped uniquely to the genome reference with only 17 (0.06%) flagged as misassembly candidates. The average aligned fraction per transcript was 0.998, the average alignment length was 1,672 bp, and the mean number of exonic blocks per alignment was 2.58, consistent with multi-exon gene structures. The NA50 (aligned N50) was 2,606 bp.

De novo assembly

Thirty-seven paired-end RNA-seq samples from 12 tissues were used for de novo transcriptome assembly using two independent platforms, Trinity and rnaSPAdes. Each sample was individually normalized using khmer digital normalization (-k 21, -c 20, -M 60e, --paired; Crusoe et al., 2015), and the normalized reads were divided into 3–4 chunks to facilitate parallel computation.

TRINITY assembly

De novo transcriptome assembly was performed using Trinity (v2.15.2; Grabherr et al., 2011) executed within a Singularity container. Khmer-normalized paired-end reads from all tissue samples were provided as input. Internal read normalization was disabled (--no_normalize_reads) as reads had been pre-normalized. Butterfly phase memory was controlled using --bflyHeapSpaceMax 50G, --bflyHeapSpaceInit 1G, and --bflyGCThreads 6, with a minimum contig length of 200 bp. The --full_cleanup flag was used to remove intermediate files upon completion. Assembly was performed on the TRUBA HPC cluster (smp partition, orkinos1 node, 224 CPUs).

TPM-based expression filtering was initially evaluated for redundancy reduction. Applying TPM ≥ 1 and TPM ≥ 2 thresholds to the largest raw Trinity assembly (3.43 M transcripts) reduced transcript numbers to 579,100 and 309,700, respectively; however, BUSCO completeness dropped dramatically from C:91.2% [S:17.0%,D:74.2%],F:6.6%,M:2.2% to C:38.1%[S:22.2%,D:15.9%],F:14.8%,M:47.1% and C:24.1% [S:14.9%,D:9.2%],F:10.5%,M:65.4%, respectively. TPM-based filtering was therefore abandoned.

EvidentialGene tr2aacds (Gilbert, 2019) was adopted as the primary redundancy reduction method. The tool scans all reading frames per transcript, selects the best ORF based on length and completeness, collapses 100% identical protein sequences, and performs internal BLASTN-based clustering at 98% identity to select the most complete representative per gene locus — rendering separate sequence identity clustering steps redundant. As shown in Table 4, EvidentialGene filtering reduced transcript number many folds while BUSCO completeness decreased by only few percentages and duplication rates were substantially suppressed.

Residual redundancy in the okay sets was further reduced using MMseqs2 easy-cluster (v14; Steinegger & Söding, 2018) applied to TransDecoder-predicted proteins, with a minimum sequence identity of 70% and target coverage of 80% (--cov-mode 1, --cluster-mode 2), retaining the longest sequence per cluster as the representative. Transcript numbers were reduced substantially with a BUSCO completeness loss of only 0.1–0.5% (Table 4). The four Trinity chunk assemblies were finally merged using a single EvidentialGene tr2aacds run, yielding the highest BUSCO completeness score among all individual assemblies in this study.

Table 4
Assembly pipeline progression and BUSCO assessment (crustacea_odb12, n = 1,536)Three-assembly ceiling represents the uncollapsed merge of Trinity, rnaSPAdes, and GGTA assemblies prior to any redundancy filtering, providing the theoretical maximum gene space coverage achievable from these data. The final assembly (93,345 transcripts) retains 96.9% of this ceiling BUSCO score.

Assembly	Transcripts	C%	S%	D%	F%	M%
De novo — Trinity						
Trinity four-batch EvidentialGene okay	1,062,747	86.6	76.7	9.9	11.7	1.8
De novo — rnaSPAdes						
rnaSPAdes four-batch EvidentialGene okay	131,370	88.6	80.0	8.6	9.0	2.4
Genome-guided — GGTA						
GGTA raw (> 200 bp)	341,835	28.0	20.7	7.3	29.2	42.8
GGTA EvidentialGene okay	63,743	27.5	26.8	0.8	28.2	44.3
GGTA MMseqs2 70% representative	30,771	27.1	27.0	0.2	26.9	46.0
Hybrid merge pipeline						
rnaSPAdes + GGTA EvidentialGene okay	119,051	90.1	82.7	7.4	7.7	2.1
rnaSPAdes + GGTA TPM ≥ 1 filtered	101,252	90.0	82.8	7.2	7.6	2.4
ASTLE transcriptome (final)	93,345	89.5	82.6	7.0	7.7	2.8
Three-assembly theoretical ceiling	296,469	92.4	83.1	9.2	6.4	1.2

rnaSPAdes assembly

De novo transcriptome assembly was additionally performed using rnaSPAdes (v3.15.5, SPAdes package; Bushmanova et al., 2019) on the same khmer-normalized RNA-seq data used for Trinity assembly. Paired-end reads from each sample were provided as independent library inputs. The resulting transcript sequences were filtered to retain sequences ≥ 200 bp, followed by redundancy reduction using EvidentialGene tr2aacds and MMseqs2 clustering at 70% sequence identity and 80% coverage as described above. The four rnaSPAdes chunk assemblies were merged using a final EvidentialGene tr2aacds run, yielding a transcriptome of 131,370 transcripts (Table 4).

Comparison and merging of the assemblies

Both Trinity and rnaSPAdes assemblies achieved comparable BUSCO completeness scores, while the rnaSPAdes assembly contained substantially fewer transcripts. A three-way merge of Trinity, rnaSPAdes, and GGTA assemblies using EvidentialGene tr2aacds yielded 296,469 transcripts with a BUSCO score of

C:92.4%[S:83.1%,D:9.2%],F:6.4%,M:1.2%, with Trinity, rnaSPAdes, and GGTA contributing 71.1%, 24.3%, and 4.7% of transcripts, respectively. However, the large transcript number could not be reduced without significant loss of completeness, with Trinity identified as the primary source of redundancy due to isoform inflation at individual gene loci.

To objectively compare assembly quality, Salmon quantification (Patro et al., 2017) was performed on four representative RNA-seq samples — sensory neuron, hepatopancreas, muscle, and ganglia — mapped independently to the Trinity and rnaSPAdes assemblies. Mean mapping rates to rnaSPAdes assembly were 42.7%, 43.3%, 50.1%, and 38.7%, compared to 39.8%, 26.3%, 42.9%, and 32.4% for Trinity, representing an average improvement of 8.4 percentage points in favor of rnaSPAdes assembly. Trinity assembly was therefore excluded from the final merge due to its larger size and lower mapping rate, and the rnaSPAdes and GGTA assemblies were merged using EvidentialGene tr2aacds, yielding 119,051 transcripts with a BUSCO score of C:90.1%[S:82.7%,D:7.4%],F:7.7%,M:2.1%.

A TPM ≥ 1 expression filter, requiring expression support in at least one sample, reduced the transcript number to 101,252 with negligible change in BUSCO completeness. Subsequent screening for sequencing adapter and vector sequences (UniVec), ribosomal RNA contamination (SILVA database), and foreign biological contamination (NCBI foreign contamination screen) reduced the final transcriptome to 93,345 transcripts. The final assembly comprised 154.5 Mb total length, a mean length of 1,656 bp, median length of 1,070 bp, minimum length of 256 bp, N50 of 2,498 bp, and maximum length of 47,035 bp. This assembly was used for all downstream analyses.

Quality of the final transcriptome was assessed using rnaQUAST (v2.3.2) against the *draft Astacus leptodactylus* genome as reference. Of 93,345 transcripts, 85.4% showed genome alignment, with 56.8% mapping uniquely and 14.6% remaining unaligned. The average aligned fraction per transcript was 0.875, the average alignment length was 1,081 bp, and 22,704 structural misassemblies were detected. The relatively low alignment rate and elevated misassembly count are expected characteristics of de novo short-read transcriptome assemblies evaluated against a fragmented draft genome. A proportion of unaligned transcripts likely represent genuine expressed sequences from genomic regions absent or incomplete in the current draft assembly.

Tissue-specific expression was quantified from Salmon output at a TPM ≥ 1 threshold using two complementary approaches: mean TPM across replicates to assess expression consistency (expression breadth; Fig. 2), and maximum TPM across replicates to assess presence (Fig. 3). Ganglia, receptor organ, and sensory neuron samples showed the highest expression breadth values, reflecting the transcriptional complexity of nervous system tissues. Muscle ranked fourth in expression breadth but dropped to tenth by maximum TPM, suggesting consistent but moderate expression levels. 24.4% of transcripts were expressed exclusively in a single tissue, while ubiquitous expression across 10–12 tissues was observed for 18.0% of transcripts. Testis showed the lowest expression breadth and transcript count among all tissues (Figs. 2 and 3).

Annotation.

TransDecoder-predicted protein sequences were searched against the UniProt Swiss-Prot database (release December 2025, 274 MB, 570,000 manually curated entries) using BLASTP (v2.17.0; Camacho et al., 2009) with an E-value threshold of 1×10^{-3} and a maximum of 5 hits per transcript. For each transcript, the best hit was selected based on the lowest E-value among the top 5 reported hits. To assess the effect of identity-based filtering on transcriptome completeness, three thresholds were applied: lenient (≥ 30% identity, ≥ 50 aa alignment, $E \leq 1\times10^{-3}$), moderate (≥ 50% identity, ≥ 100 aa, $E \leq 1\times10^{-10}$), and stringent (≥ 70% identity, ≥ 150 aa, $E \leq 1\times10^{-20}$). BUSCO completeness (crustacea_odb12, n = 1,536) was assessed on each filtered subset (Table 5). The full unfiltered assembly retained the highest BUSCO completeness (C:89.5%), while stringent filtering reduced completeness dramatically, demonstrating that identity-based filtering introduces strong taxonomic bias against non-model invertebrate sequences.

Table 5
Assessment of protein identity thresholds on transcriptome size and completeness

Setting	Transcripts	Min. Identity	Max. E-value	Min. Alignment	BUSCO Complete	Missing
Full assembly	93,345	—	—	—	89.5%	2.8%
Lenient	29,610	30%	1×10^{-3}	50 aa	88.0%	4.9%
Moderate	10,681	50%	1×10^{-10}	100 aa	55.7%	38.9%
Stringent	4,159	70%	1×10^{-20}	150 aa	17.3%	80.0%

Swiss-Prot contains fewer than 500 reviewed crustacean entries among approximately 570,000 manually curated records (UniProt Consortium, 2023), making homology-based filtering strongly biased against genuinely expressed crustacean-specific transcripts. Swiss-Prot homology filtering was therefore not applied as a quality criterion; instead, the full unfiltered transcriptome was retained for downstream analyses.

Predicted protein sequences were searched against the Pfam-A database (v35.0; Mistry et al., 2021) using hmmscan (HMMER v3.3.2; Eddy, 2011) with an E-value threshold of 1×10^{-5} to identify transcripts encoding recognizable protein domains. Protein IDs were mapped back to their parent transcript IDs to yield a Pfam-supported transcript subset. As shown in Table 6, the majority of transcripts (67,616; 72.4%) lacked annotation in either database.

Table 6
Comparison of BUSCO scores of the transcriptome subsets, extracted in reference to the annotation source.

Subset	Transcripts	BUSCO Scores				
		C%	S%	D%	F%	M%
Full assembly	93,345	89.5	82.6	7.0	7.7	2.8
Any annotation (BLAST or Pfam)	25,729	88.9	82.3	6.6	7.7	3.3
Swiss-Prot annotated	23,276	88.5	81.9	6.6	7.7	3.8
Pfam annotated	19,426	87.6	81.1	6.5	7.2	5.2
Unannotated	67,616	1.2	1.0	0.1	1.8	97.0

To confirm the genuine organismal origin of unannotated transcripts, all transcriptome subsets were aligned to the draft *Astacus leptodactylus* genome using minimap2 (v2.30) with the splice-aware preset (-ax splice), recognizing canonical GT-AG splice sites and allowing intron gaps up to 200 kb (Li 2018). Alignments were coordinate-sorted and indexed using SAMtools, and mapping statistics were obtained with SAMtools flagstat (Danecek et al., 2021). As shown in Table 7, the unannotated subset (67,616 transcripts) showed the highest primary genome alignment rate (89.22%), exceeding both the full assembly (86.72%) and annotated subset (80.17%), confirming that unannotated transcripts are genuine expressed sequences of *Astacus leptodactylus* rather than assembly artefacts, and likely represent crustacean-specific genes with no current homologs in public databases.

Table 7
Comparison of mapping rate of the transcriptome subsets to draft genome

Subset	Transcripts	Primary mapped	Primary %	Total Mapped %
Full assembly	93,345	80,953	86.72%	95.87%
Annotation-based (BLAST or Pfam)	25,729	20,626	80.17%	93.16%
Unannotated	67,616	60,327	89.22%	96.77%

Gene Ontology (GO) terms were assigned based on Swiss-Prot hits filtered to metazoan organisms and Pfam domain-derived GO annotations. 17,426 transcripts (18.7% of the assembly) received at least one GO annotation, yielding 183,810 transcript-GO annotation pairs distributed across Biological Process (83,949; 45.7%), Molecular Function (47,268; 25.7%), and Cellular Component (52,593; 28.6%) (Fig. 6).

Expression quantification and multivariate analysis

Salmon index was constructed from the final 93,345-transcript assembly and all 37 paired-end RNA-seq samples across 12 tissues were quantified independently. Raw counts were imported into R using

tximport and subjected to variance stabilizing transformation (VST; DESeq2, blind=TRUE) prior to dimensionality reduction. Principal component analysis of VST-normalized expression across all 37 samples (75,179 transcripts passing expression filters) revealed that PC1 (33.6% variance) separated sensory neuron samples from all other tissues, reflecting their highly divergent transcriptional profile. Along PC3 (8.2% variance), hepatopancreas and stomach samples separated downward reflecting their shared digestive function, while muscle samples formed a distinct upward cluster. Ganglia, gill, eye, antennule, heart, and testis samples formed a central cluster, with receptor organ co-localizing in this group (Fig. 4).

Hierarchical clustering of the 2,000 most variable transcripts (Ward.D2 linkage, Euclidean distance) confirmed the strongly distinct transcriptional identity of sensory neuron samples, which formed a tight isolated cluster with a large block of highly upregulated transcripts relative to all other tissues (Fig. 5). Hepatopancreas and stomach samples co-clustered together reflecting their shared digestive function. Muscle samples formed a distinct intermediate cluster, while receptor organ co-localized with testis and heart samples. Antenna and antennule grouped together, with gill, eye, and ganglia samples forming a broader somatic cluster on the opposite branch. The consistent separation of sensory neuron, digestive, and somatic tissue clusters in both PCA and hierarchical clustering validates the breadth and biological coherence of the 12-tissue expression atlas.

Validation using previously cloned ion channel sequences

To provide independent functional validation of transcriptome completeness and annotation accuracy, ten ion channel sequences previously cloned from *Pontastacus leptodactylus* in our laboratory and deposited in the NCBI GenBank database were used as reference queries. The sequences represent a functionally diverse panel of ion channels: voltage-gated sodium channel (KT187402.1), otopetrin proton channel (OR735052.1), sarcoplasmic/endoplasmic reticulum Ca^{2+}-ATPase (OQ786992.1), voltage-gated calcium channel (ON661560.1), transmembrane channel-like protein (ON604003.1), ryanodine receptor (ON420719.1), Piezo-like mechanosensitive channel (OK513276.1), inward rectifier potassium channel (MN816260.1), calcium-activated potassium channel (MN816259.1), and sodium-calcium exchanger (KT697616.1).

Each sequence was used as a BLASTN query against the final 93,345-transcript assembly using a local BLAST nucleotide database constructed with makeblastdb. Searches were performed with an E-value threshold of 1×10^{-10}, retaining up to 5 hits per query (BLAST+ v2.17.0). For each cloned sequence, the best-matching transcript was identified based on E-value and alignment length. Percentage nucleotide identity, query coverage, alignment length, and E-value were recorded for each best hit. SwissProt BLASTP annotation of the best-matching transcripts was extracted from the annotation table, providing an independent protein-level confirmation of transcript identity. Mean TPM values per tissue, derived from Salmon quantification as described above, were used to assess tissue-specific expression of each recovered channel transcript. Results are summarized in Fig. 7.

• Data Records

All the sequencing RNA-Seq data samples and the assembled multi-tissue transcriptome have been submitted to the National Center for Biotechnology Information (NCBI) under BioProject accession number PRJNA1175850; BioSample accession numbers SAMN60680959-95; SRA Run accession numbers SRR39447606-42.

PacBio HIFI long read data and the assembled draft genome have been submitted to the National Center for Biotechnology Information (NCBI) under BioProject accession number PRJNA1481542; BioSample accession number SAMN61093799; SRA Run accession number SRR39353963.

Figure scripts and associated files together with NCBI_Metadata files with SRA accessions are deposited at the Figshare repository

https://doi.org/10.6084/m9.figshare.32781303 .

https://doi.org/10.6084/m9.figshare.32789988

• Data Overview

The final transcriptome comprises 93,345 transcripts (N50 = 2,498 bp, total length 154.5 Mb, GC = 45.4%) with a BUSCO completeness of C:89.5%[S:82.6%,D:7.0%],F:7.7%,M:2.8% (crustacea_odb12, n = 1,536). Functional annotation identified 25,729 transcripts (27.6%) with homology to UniProt Swiss-Prot entries or recognizable Pfam domains, with 14,260 unique Gene Ontology terms assigned. Expression profiling across 37 samples and 12 tissues revealed that sensory neuron and ganglia tissues express the broadest range of transcripts, while testis showed the most restricted expression profile.

Although the draft genome achieved modest BUSCO completeness (C:33.5%, crustacea_odb12), transcript-to-genome mapping confirmed that 86.72% of assembled transcripts align to the draft genome, rising to 89.22% for unannotated transcripts — exceeding the alignment rate of the annotated subset (80.17%). This discrepancy reflects the well-documented fragmentation of crustacean genomes rather than absence of gene content, and confirms that unannotated transcripts represent genuine expressed sequences of *Astacus leptodactylus*, likely encoding crustacean-specific genes absent from current databases. Together, the transcriptome assembly, functional annotation table, expression quantification data, and draft genome constitute a coherent multi-layer genomic resource for *Astacus leptodactylus*.

• Technical Validation

The Astacus leptodactylus transcriptome was assembled using a hybrid strategy combining genome-guided and de novo approaches, leveraging a draft genome assembled from high-quality (Q ≥ 32) PacBio HiFi long reads and RNA-seq data from 37 samples spanning 12 tissues.

The draft genome primary assembly achieved BUSCO completeness of C:30.7% [S:29.4%,D:1.4%],F:29.9%,M:39.4% (crustacea_odb12, n = 1,536), improving to C:33.5% [S:30.1%,D:3.4%],F:32.6%,M:33.9% when primary and alternate contigs were combined. These scores reflect the highly fragmented nature of the draft assembly rather than genuine gene space deficiency, as crustacean genomes are characteristically large and repeat-rich. Contamination screening by alignment of the 50 longest and 50 shortest contigs against the NCBI nucleotide database and the Procambarus clarkii genome (GenBank: JBFOCG00000000038) revealed no evidence of foreign sequence contamination.

The genome-guided transcriptome assembly (GGTA), produced by mapping all 37 RNA-seq samples to the draft genome using HISAT2 followed by transcript reconstruction with StringTie, yielded 30,771 transcripts with BUSCO completeness of C:27.1%[S:27.0%,D:0.2%],F:26.9%,M:46.0% (crustacea_odb12, n = 1,536), consistent with the fragmented genome reference. Despite this, rnaQUAST analysis confirmed that 99.99% of GGTA transcripts in total mapped back to the draft genome with 96.5% mapping uniquely, confirming high assembly fidelity and absence of spurious sequences.

De novo assembly using Trinity and rnaSPAdes yielded near-complete transcriptomes with BUSCO scores of C:86.6%[S:76.7%,D:9.9%],F:11.7%,M:1.8% and C:88.6%[S:80.0%,D:8.6%],F:9.0%,M:2.4% (crustacea_odb12, n = 1,536), respectively. rnaSPAdes was selected for merging with GGTA based on lower redundancy and higher short-read mapping rate. Merging the two assemblies followed by EvidentialGene protein homology filtering progressively reduced redundancy while preserving gene space coverage: the merged assembly after EvidentialGene filtering reduced to 119,051 sequences and achieved a score of C:90.1%[S:82.7%,D:7.4%],F:7.7%,M:2.1%. After expression-based filtering at TPM ≥ 1, transcript number reduced to 101,252 and completeness score to C:90.0% [S:82.8%,D:7.2%],F:7.6%,M:2.4%. The assembled transcriptome was screened for vector and ribosomal RNA contamination using UniVec and SILVA databases, as well as NCBI's foreign contamination screen; 4,749 sequences (4.8%) were removed and 24 additional sequences were trimmed to remove contaminating regions. The final assembly after contamination screening comprises 93,345 transcripts with a BUSCO score of C:89.5%[S:82.6%,D:7.0%],F:7.7%,M:2.8% (crustacea_odb12, n = 1,536), representing 97.2% of all 1,536 crustacean BUSCO genes when fragmented hits are included. Quality of the final transcriptome was assessed using rnaQUAST (v2.3.2) against the draft Astacus leptodactylus genome as reference. Of 93,345 transcripts, 85.4% showed genome alignment, with 56.8% mapping uniquely and 14.6% remaining unaligned — consistent with the 86.72% primary alignment rate obtained by splice-aware minimap2 mapping (Table 7). The average aligned fraction per transcript was 0.875, the average alignment length was 1,081 bp, and 22,704 structural misassemblies were detected. The relatively low alignment rate and elevated misassembly count are expected characteristics of de novo short-read transcriptome assemblies evaluated against a fragmented draft genome. A proportion of unaligned transcripts likely represent genuine expressed sequences from genomic regions absent or incomplete in the current draft assembly.

Expression breadth analysis using a TPM ≥ 1 threshold across all 37 samples revealed that 77,253 transcripts (82.8%) were expressed in at least one tissue. Of these, 22,749 transcripts (24.4%) showed tissue-specific expression (exactly one tissue), while 16,826 transcripts (18.0%) were expressed ubiquitously across 10–12 tissues. Ganglia, receptor organ, and sensory neuron samples showed the highest expression breadth values, reflecting the transcriptional complexity of nervous system tissues. Principal component analysis (PCA) and hierarchical clustering of the 2,000 most variable transcripts grouped samples consistently by tissue type, confirming the reproducibility of tissue-specific expression patterns across biological replicates.

Functional annotation using BLASTP against the UniProt Swiss-Prot database and hmmscan against the Pfam database identified 25,729 annotated transcripts, which retained a BUSCO score of C:88.9% [S:82.3%,D:6.6%],F:7.7%,M:3.3% — nearly identical to the full assembly. The unannotated subset achieved C:1.2%[S:1.0%,D:0.1%],F:1.8%,M:97.0%, confirming that BUSCO-detectable conserved genes are concentrated in the annotated fraction and that unannotated transcripts represent genuinely novel or highly divergent sequences. Genome mapping confirmed that the unannotated subset had the highest alignment rate (89.22%) to the species-specific draft genome, exceeding those of the full assembly (86.72%) and annotated subset (80.17%), demonstrating their genuine organismal origin independent of sequence homology to known proteins. GO annotation was performed using metazoan-only Swiss-Prot entries to avoid taxonomic bias, as the Swiss-Prot database is heavily weighted towards vertebrate model organisms.

To further validate the functional completeness of the transcriptome and demonstrate the critical contribution of sensory neural tissue sampling, we performed BLASTN searches using ten ion channel sequences previously cloned from *Pontastacus leptodactylus* in our laboratory and deposited in NCBI GenBank (KT187402.1, OR735052.1, OQ786992.1, ON661560.1, ON604003.1, ON420719.1, OK513276.1, MN816260.1, MN816259.1, KT697616.1). All ten cloned sequences were recovered in the transcriptome at 84–100% nucleotide identity. Importantly, independent SwissProt BLASTP annotation of the corresponding translated sequences confirmed the identity of each channel at the protein level: the Nav transcript (ASTLE_010799) matched sodium channel protein Para (SCNA_DROME, 79.1% amino acid identity), KCa matched the slowpoke calcium-activated potassium channel (SLO_DROME, 77.4%), SERCA matched a sarcoplasmic Ca^{2+}-ATPase (ATC1_ANOGA, 77.2%), and the ryanodine receptor matched RYR_DROME (65.6%), among others. The concordance between nucleotide-level recovery of our cloned sequences and independent protein-level SwissProt annotation provides strong cross-validation of annotation accuracy. Expression profiling across 12 tissues revealed a biologically coherent tissue-specificity pattern fully consistent with established physiology (Fig. 7): channels essential for neuronal excitability — the voltage-gated sodium channel Nav (ASTLE_010799, 37 TPM in sensory neuron), the calcium-activated potassium channel KCa (ASTLE_069470, 591 TPM) and the inward rectifier potassium channel Kir (ASTLE_098202, 362 TPM) — showed predominant expression in sensory neuron tissue with near-zero expression in non-excitable tissues such as gill, hepatopancreas and muscle. Conversely, channels mediating excitation-contraction coupling — SERCA (ASTLE_052940, 280 TPM in muscle), the ryanodine receptor (ASTLE_063423, 117 TPM) and the Na^{+}/Ca^{2+} exchanger NCX (ASTLE_045588, 52

TPM) — showed highest expression in muscle and receptor organ. This dual concordance — between cloned nucleotide sequences and transcriptome assembly, and between SwissProt protein annotation and tissue-specific expression — confirms both the technical accuracy of the assembly and quantification pipeline, and underscores the biological necessity of including dedicated sensory neural tissues to capture the full repertoire of excitable-cell transcripts.

Usage Notes

The final transcriptome assembly (ASTLE_transcriptome, 93,345 transcripts) is provided with identifiers in the format ASTLE_XXXXXX and is suitable for direct use in downstream analyses including differential expression, functional annotation, and phylogenetic studies.

The accompanying functional annotation table provides SwissProt BLAST hits, Pfam domain assignments, GO terms, and quality tier classifications for each transcript. Users are advised to apply tier-based filtering appropriate to their downstream application: the SwissProt or Pfam annotated subset (25,729 transcripts, BUSCO C:88.9%) is recommended for comparative analyses requiring high annotation confidence, while the full assembly should be used for expression quantification to avoid discarding genuinely expressed unannotated transcripts.

Expression quantification should be performed using Salmon against the full 93,345-transcript assembly. Raw RNA-seq reads from all 37 samples are available under BioProject PRJNA1175850 and can be requantified using Salmon with the --validateMappings and --gcBias flags as described in the Methods section.

The draft genome assembly (combined primary and alternate contigs, 229,863 sequences, N50 = 119 kbp) is provided as a supplementary resource to support genome-guided analyses and transcript-to-genome mapping. Given the fragmented nature of the assembly, it is not recommended as a standalone reference genome but serves as a valuable scaffold for splice-site validation and exon boundary identification using splice-aware aligners such as minimap2.

Declarations

Code Availability

All custom scripts used in this study, including those for functional annotation, expression breadth analysis, and figure generation, are available alongside the annotation table in the Figshare repository (https://doi.org/10.6084/m9.figshare.32781303; https://doi.org/10.6084/m9.figshare.32789988).

Author Contributions

N.P. conceived and planned the study, participated in sample collection, performed quality control of sequencing data, carried out all bioinformatic analyses of genomic and transcriptomic data, and wrote the manuscript. B.S. participated in sample collection, cDNA library construction, and quality

assessment of next-generation sequencing data. B.E. performed next-generation sequencing of the cDNA libraries. M.A. assisted with sample collection.

Competing Interests

The authors declare no competing interests.

Acknowledgements

The authors thank the TRUBA high-performance computing infrastructure (TUBITAK ULAKBIM) for providing computational resources used in this study. We would like to thank Efe Dalli from INTERGEN (Ankara-TURKEY) for his valuable help in data transfer.

Funding

This study was supported by The Scientific and Technological Research Council of Turkey (TUBITAK), grant number 125S130, and Hacettepe University Research Foundation, grant number TSA-19692.

Ethics Statement

Experimental animals (*Astacus leptodactylus*) were collected from fisheries located at Hirfanlı Lake, Ankara, Turkey. As an invertebrate species, formal ethical approval was not required under Turkish national regulations. All procedures were conducted in accordance with national and institutional guidelines for the use of animals in research.

References

1. Alexandrowicz JS (1951) Muscle receptor organs in the abdomen of *Homarus vulgaris* and *Palinurus vulgaris*. Q J Microsc Sci 92:163–199
2. Arslan K, Saglam B, Beyatli NC, Taskiran EZ, Bastug T, Purali N (2023) Cloning and functional characterization of a TMC-like channel gene and protein in the crayfish *Astacus leptodactylus*. J Crustac Biol 43:ruad032
3. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD (2019) rnaSPAdes: a de novo transcriptome assembler and its application to RNA-seq data. GigaScience 8:giz100
4. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421
5. Chen Y, Hou Y, Jia S, Gong K, Ren X, Xu W, Ma W, Lv J (2026) A comprehensive Iso-Seq transcriptome of the narrow-clawed crayfish *Pontastacus leptodactylus*. Sci Data. https://doi.org/10.1038/s41597-026-07503-w
6. Cheng H, Concepcion GT, Feng X, Zhang H, Li H (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 18:170–175
7. Colbourne JK et al (2011) The ecoresponsive genome of *Daphnia pulex*. Science 331:555–561

8. Coşkun C, Puralı N (2016) Cloning and molecular characterization of a putative voltage-gated sodium channel gene in the crayfish. Invertebr Neurosci 16:2
9. Crusoe MR et al (2015) The khmer software package: enabling efficient nucleotide sequence analysis. F1000Res 4:900
10. Danecek P et al (2021) Twelve years of SAMtools and BCFtools. GigaScience 10, giab008
11. Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195
12. Ergin B, Puralı N (2018) Cloning and molecular characterization of a putative voltage-gated potassium channel gene in the crayfish. Invertebr Neurosci 18:5
13. Ergin B, Saglam B, Taskiran EZ, Bastug T, Purali N (2022) De novo cloning and functional characterization of potassium channel genes and proteins in the crayfish *Astacus leptodactylus*. J Crustac Biol 42:ruac018
14. Ergin B, Saglam B, Arslan K, Beyatli NC, Taskiran ZE, Bastug T, Purali N (2023) De novo cloning and functional characterization of a mechanosensitive Piezo-like ion channel in the crayfish. Cell Physiol Biochem 57:226–237
15. Gilbert DG, EvidentialGene (2019) mRNA transcript assembly software. figshare https://doi.org/10.6084/m9.figshare.7391387
16. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol 29:644–652
17. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R (2020) Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36:2896–2898
18. Haas BJ et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512
19. Hyde CJ, Fitzgibbon QP, Elizur A, Ventura T (2019) A crustacean annotated transcriptome (CAT) database. BMC Genomics 20:531
20. Kao D et al (2016) The genome of the crustacean *Parhyale hawaiensis*. eLife 5, e20062
21. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915
22. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100
23. Liao J et al (2025) A high-quality chromosome-level genome assembly of Pacific whiteleg shrimp (*Penaeus vannamei*). Sci Data 12:340
24. Liao M et al (2024) The chromosome-level genome assembly of the red swamp crayfish *Procambarus clarkii*. Sci Data 11:885
25. Liu Z et al (2024) Genome assembly of redclaw crayfish (Cherax quadricarinatus) provides insights into its immune adaptation and hypoxia tolerance. BMC Genomics 25:746
26. Lnenicka GA (2020) Crayfish and Drosophila NMJs. Neurosci Lett 732:135110
27. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550

28. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38:4647–4654
29. Mistry J et al (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419
30. Northcutt AJ et al (2016) Deep sequencing of transcriptomes from the nervous systems of two decapod crustaceans to characterize genes important for neural circuit function and modulation. BMC Genomics 17:868
31. O'Neil ST, Emrich SJ (2013) Assessing de novo transcriptome assembly metrics for consistency and utility. BMC Genomics 14:465
32. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14:417–419
33. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295
34. Purali N (2005) Structure and function relationship in the abdominal stretch receptor organs of the crayfish. J Comp Neurol 488:369–383
35. Sağlam B, Ergin B, Beyatli NC, Arslan K, Bastug T, Purali N (2023) Cloning and in silico investigation of a putative voltage-gated calcium channel gene and protein in *Astacus leptodactylus*. Turk J Biochem 48:675–681
36. Sağsöz ME, Sağlam B, Arslan K, Baştuğ T, Çavuş M, Puralı N (2024) Structural, functional and molecular dynamics examination of a de novo cloned otopetrin-like proton channel in crayfish. Cell Biochem Biophys 82:2029–2036
37. Soneson C, Love MI, Robinson MD (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res 4:1521
38. Steinegger M, Söding J (2018) Clustering huge protein sequence sets in linear time. Nat Commun 9:2542
39. UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 51:D523–D531
40. Zhang X et al (2019) Genome-wide analysis of alternative splicing in the Pacific white shrimp *Penaeus (Litopenaeus) vannamei*. Front Genet 10:754
41. Zhang Y et al (2014) An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. J Neurosci 34:11929–11947

Figures

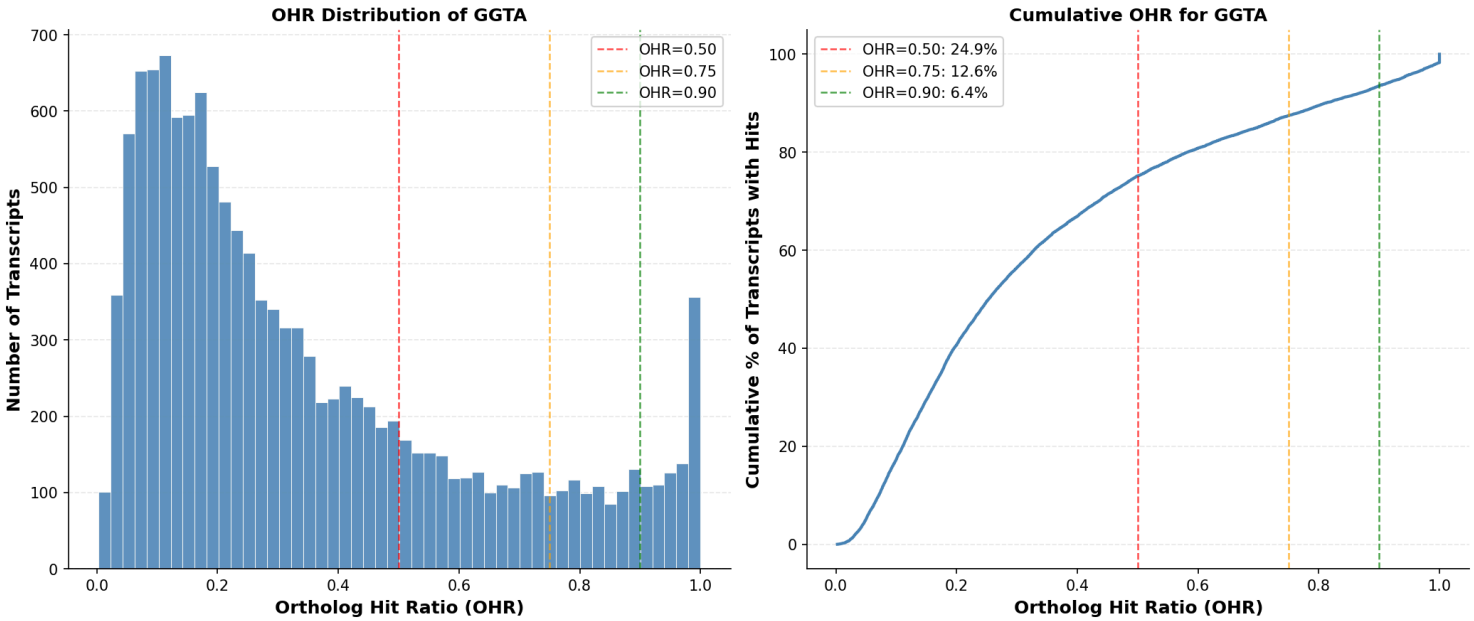


Figure 1

Ortholog hit ratio for genome-guided transcriptome assembly. OHR distribution (left) and Cumulative OHR (right).

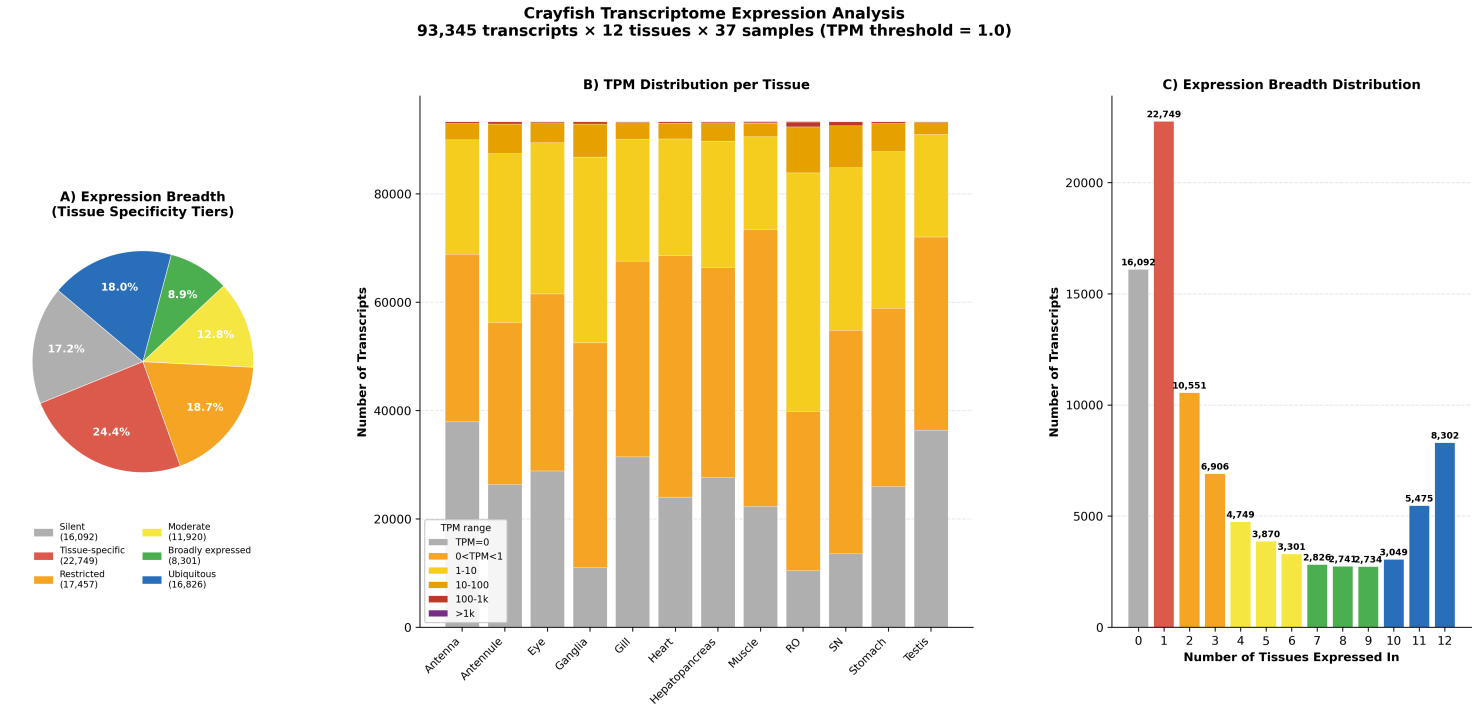


Figure 2

Transcriptome-wide expression analysis of *Astacus leptodactylus* transcripts across 12 tissues and 37 samples. (A) Pie chart showing expression breadth categories based on TPM ≥ 1 threshold across all 12 tissues. (B) Stacked bar chart showing the TPM range distribution of all 93,345 transcripts per tissue. (C)

Expression breadth histogram showing the number of transcripts expressed in 0 to 12 tissues (TPM ≥ 1, mean across replicates). The color code is as in A.

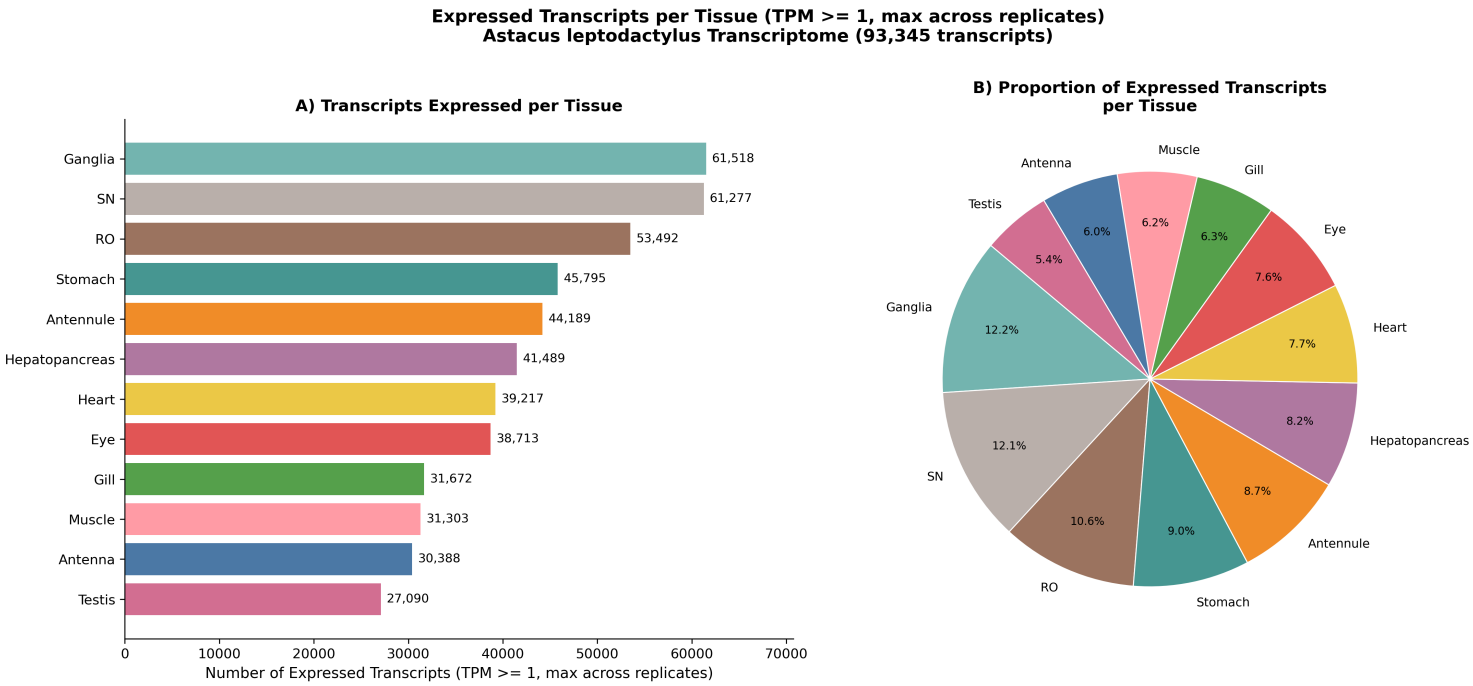


Figure 3

Number of expressed transcripts per tissue in the *Astacus leptodactylus* transcriptome (93,345 transcripts). (A) Bar charts show the number of transcripts with TPM ≥ 1 (maximum across replicates) in each of the 12 tissues. (B) Pie charts show the proportional contribution of each tissue to the total expressed transcript pool.

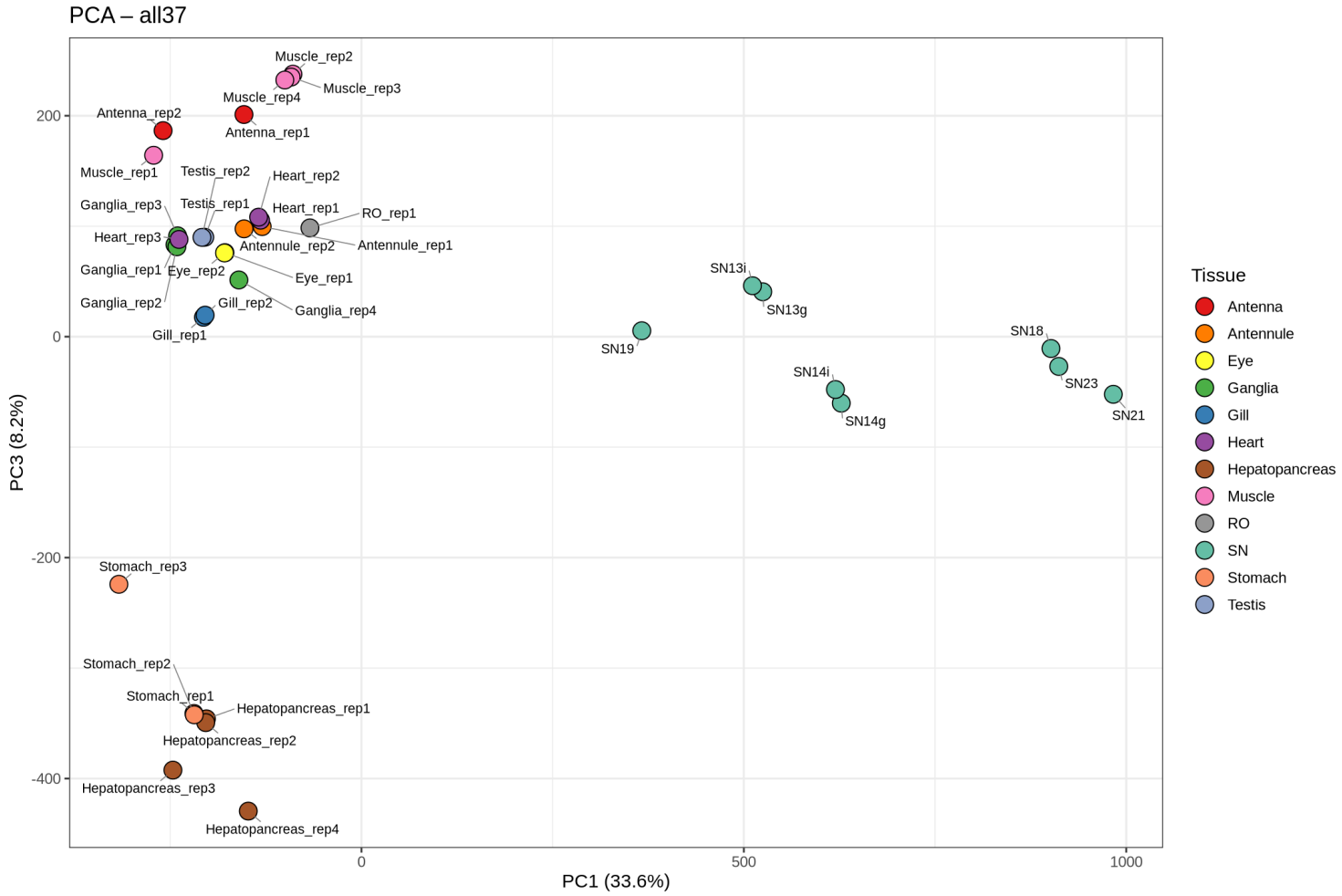


Figure 4

Principal component analysis (PCA) of VST-normalized transcript expression across 37 samples from 12 tissues of *Astacus leptodactylus*. PC1 (33.6% variance) separates sensory neuron samples from all other tissues. PC3 (8.2% variance) resolves digestive tissues (hepatopancreas, stomach) from musculoskeletal tissues (muscle). Each point represents one sample, colored by tissue type. Labels indicate individual sample identifiers.

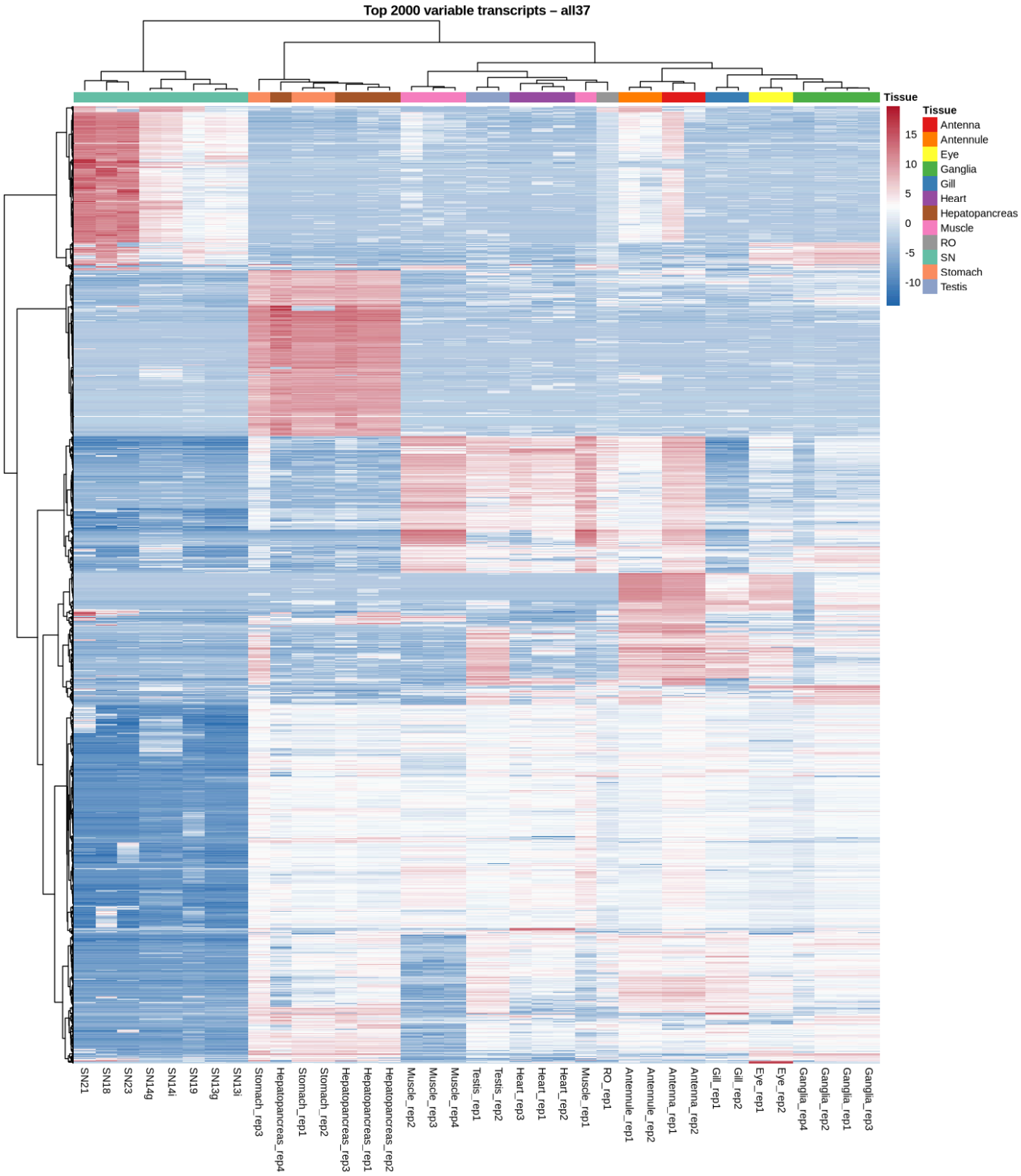


Figure 5

Hierarchical clustering heatmap of the 2,000 most variable transcripts across 37 samples from 12 tissues of *Astacus leptodactylus*. Clustering was performed using Ward.D2 linkage with Euclidean distance on VST-normalized expression values centered by row mean. Color scale represents centered $\log_2$ expression (red: high, blue: low). Tissue identity is indicated by the color bar above each column.

Sensory neuron samples form a distinct isolated cluster with a large block of highly upregulated transcripts relative to all other tissues.

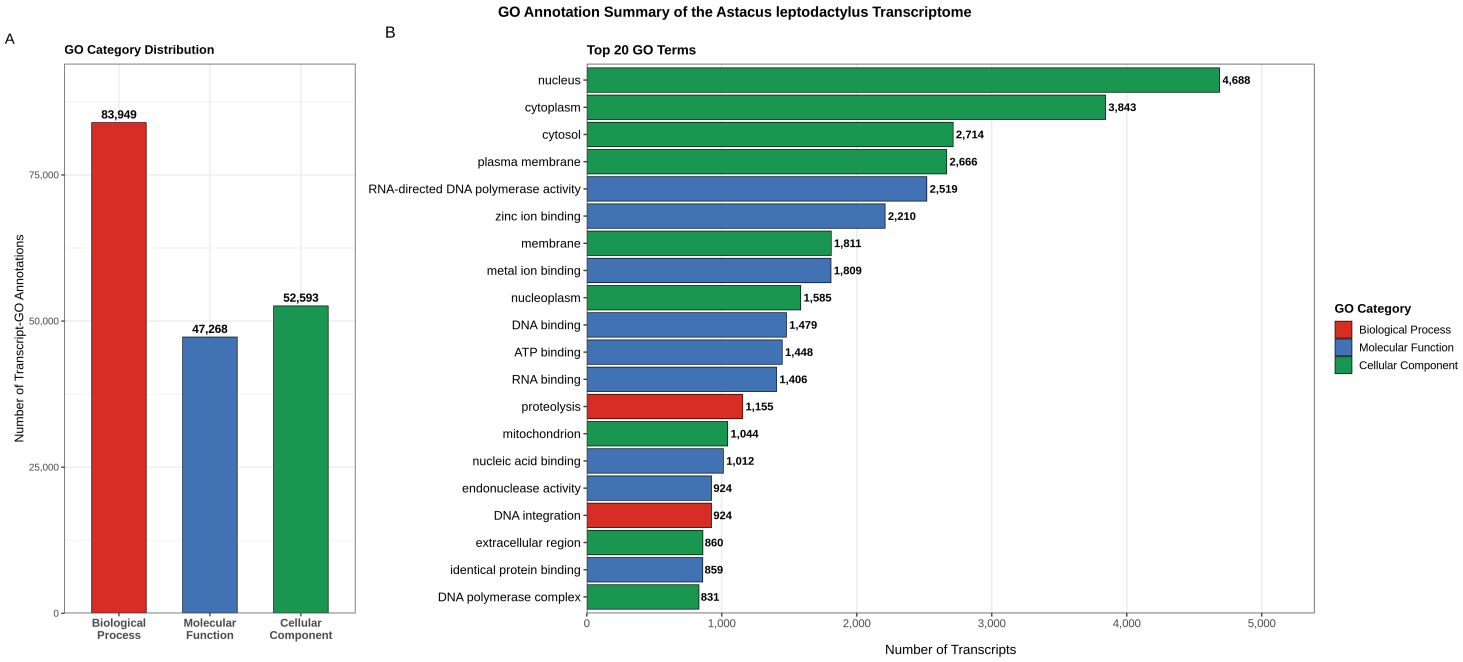


Figure 6

Distribution of the GO terms of the assembled transcriptome.

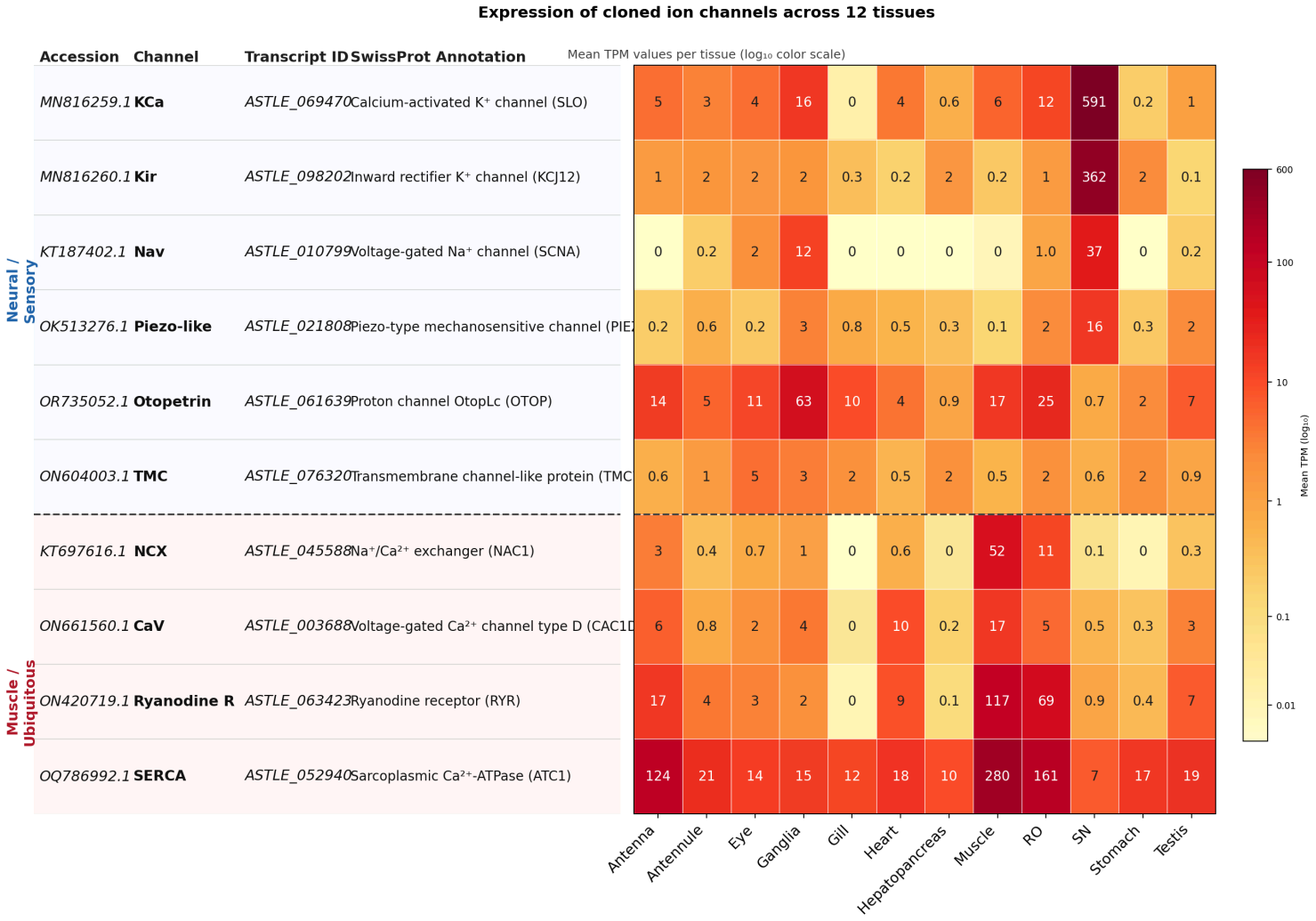


Figure 7

Expression of previously cloned ion channels across 12 tissues of *Pontastacus leptodactylus*. Mean TPM values per tissue are shown on a $\log_{10}$ colour scale (yellow: low, dark red: high). Rows represent ten ion channel transcripts identified by BLASTN against the transcriptome using GenBank sequences cloned in our laboratory. Channels are grouped into neural/sensory (upper, blue) and muscle/ubiquitous (lower, red) functional categories separated by a dashed line. All ten transcripts were independently confirmed by SwissProt BLASTP annotation at 35–79% amino acid identity to their respective arthropod orthologs. Neural channels (Nav, KCa, Kir, Piezo-like) show predominant expression in sensory neuron (SN), ganglia and eye tissues, while excitation-contraction coupling channels (SERCA, Ryanodine receptor, NCX, CaV) are most highly expressed in muscle and receptor organ (RO), consistent with their established physiological roles.