TYPE Original Research
PUBLISHED 07 December 2022
DOI 10.3389/fgene.2022.1003771

Discovering novel reproductive genes in a non-model fly using *de novo* GridION transcriptomics

Mrinalini Walter* and Nalini Puniamoorthy

Department of Biological Sciences, National University of Singapore, Singapore, Singapore

OPEN ACCESS

EDITED BY
Igor B. Rogozin,
National Institutes of Health (NIH),
United States

REVIEWED BY
Mariana Wolfner,
Cornell University, United States
Yasir Ahmed-Braimah,
Syracuse University, United States
Olga Postnikova,
United States Department of Agriculture
(USDA), United States

*CORRESPONDENCE
Mrinalini Walter,
dbsmrin@nus.edu.sg

SPECIALTY SECTION
This article was submitted to Genomic Assay Technology,
a section of the journal
Frontiers in Genetics

RECEIVED 26 July 2022
ACCEPTED 16 November 2022
PUBLISHED 07 December 2022

CITATION
Walter M and Puniamoorthy N (2022), Discovering novel reproductive genes in a non-model fly using *de novo* GridION transcriptomics.
Front. Genet. 13:1003771.
doi: 10.3389/fgene.2022.1003771

Gene discovery has important implications for investigating phenotypic trait evolution, adaptation, and speciation. Male reproductive tissues, such as accessory glands (AGs), are hotspots for recruitment of novel genes that diverge rapidly even among closely related species/populations. These genes synthesize seminal fluid proteins that often affect post-copulatory sexual selection—they can mediate male-male sperm competition, ejaculate-female interactions that modify female remating and even influence reproductive incompatibilities among diverging species/populations. Although *de novo* transcriptomics has facilitated gene discovery in non-model organisms, reproductive gene discovery is still challenging without a reference database as they are often novel and bear no homology to known proteins. Here, we use reference-free GridION long-read transcriptomics, from Oxford Nanopore Technologies (ONT), to discover novel AG genes and characterize their expression in the widespread dung fly, *Sepsis punctum*. Despite stark population differences in male reproductive traits (e.g.: Body size, testes size, and sperm length) as well as female re-mating, the male AG genes and their secretions of *S. punctum* are still unknown. We implement a *de novo* ONT transcriptome pipeline incorporating quality-filtering and rigorous error-correction procedures, and we evaluate gene sequence and gene expression results against high-quality Illumina short-read data. We discover highly-expressed reproductive genes in AG transcriptomes of *S. punctum* consisting of 40 high-quality and high-confidence ONT genes that cross-verify against Illumina genes, among which 26 are novel and specific to *S. punctum*. Novel genes account for an average of 81% of total gene expression and may be functionally relevant in seminal fluid protein production. For instance, 80% of genes encoding secretory proteins account for 74% total gene expression. In addition, median sequence similarities of ONT nucleotide and protein sequences match within-Illumina sequence similarities. Read-count based expression quantification in ONT is congruent with Illumina's Transcript per Million (TPM), both in overall pattern and within functional categories. Rapid genomic innovation followed by recruitment of *de novo* genes for high expression in *S. punctum* AG tissue, a pattern observed in other insects, could be a likely mechanism of evolution of these genes. The study also

Abbreviations: AG, Accessory Gland; cDNA, Complementary DNA; ONT, Oxford Nanopore Technologies; PacBio, Pacific Biosciences; PCR, Polymerase Chain Reaction; SMRT, Single-Molecule Real Time; TPM, Transcript per Million.

demonstrates the feasibility of adapting ONT transcriptomics for gene discovery in non-model systems.

KEYWORDS

gene expression, GridION, Illumina, novel gene, Oxford Nanopore Technologies (ONT), reproduction, *Sepsis punctum*, sexual selection

1 Introduction

Insects are hyper-diverse not only in terms of species richness, but also in many aspects of their reproductive biology. Most female insects mate multiple times and possess specialized organs for long-term sperm storage. This sets the stage for post-copulatory sexual selection that involves complex morphological and biochemical interactions between sperm, seminal plasma, and the female. There can also be interactions between ejaculates from competing males as they vie for limited storage space and for fertilization. Such sperm competition often leads to the evolution of extraordinarily diverse reproductive traits, including exaggerated sperm morphologies as well as rapidly diversifying seminal fluid compositions (Lupold et al., 2016; Gasparini et al., 2020; Simmons and Wedell, 2020). Some seminal fluid proteins and genes are known to evolve rapidly even among closely related species of insects (Swanson et al., 2001; Mueller et al., 2005; Findlay et al., 2008; Goenaga et al., 2015; Abry et al., 2017; Bayram et al., 2017; Bayram et al., 2019; Mrinalini et al., 2021), and they can play a crucial role in post-copulatory sexual selection in the female by forming mate plugs, affecting differential sperm storage, and even influencing female receptivity to remating (Avila et al., 2011; Wigby et al., 2020). Rapid divergence in genes underlying ejaculate-female interactions among populations can result in barriers to gene flow that generate reproductive incompatibilities and mediate incipient speciation within a species (Goenaga et al., 2015; Yamane et al., 2015).

The male reproductive tissues of insects typically consist of testes that are responsible for sperm production and closely associated accessory glands (AGs) that synthesize seminal fluid proteins (Figure 1). Analyses of several insect AG secretions have revealed the presence of protease inhibitors, c-type lectins, cysteine-rich secretory proteins, prohormones, antimicrobial proteins, as well as small peptides (Avila et al., 2011; Bayram et al., 2017; Mrinalini et al., 2021). However, the most striking aspect of insect AG secretions is the ubiquitous presence of dozens of newly-evolved proteins that show no similarity to proteins from other organisms and whose functions are largely unknown (Parthasarathy et al., 2009; Bayram et al., 2017; Bayram et al., 2019; Mrinalini et al., 2021). Furthermore, AG proteins are known to evolve rapidly even at the primary sequence level (Findlay and Swanson, 2010; Avila et al., 2011). Many studies on AG secretions have focused on model systems with well-characterized reference genomes like *Drosophila melanogaster,* disease vectors like *Aedes* mosquitoes, as well as agricultural pests like *Tribolium castaneum* and *Callosobruchus maculatus* beetles (Parthasarathy et al., 2009; Avila et al., 2011; Goenaga et al., 2015; Yamane et al., 2015; Ahmed-Braimah, 2016; Sayadi et al., 2016; Bayram et al., 2017; Bayram et al., 2019; Degner et al., 2019; Wigby et al., 2020). A handful of other insect groups that include crickets, moths, bees, and ants have also been explored (Braswell et al., 2006; Gorshkov et al., 2015; Fuessl et al., 2018; Gotoh et al., 2018; Saraswathi et al., 2020), but overall, studies on non-model insect reproductive gene discovery are still sparse. In a recent study on dung beetle, we discovered that AGs are hotspots for recruitment of completely novel genes for reproductive function whilst the testes generally express more conserved genes involved in sperm production (Mrinalini et al., 2021). We found that rapid evolution at the genomic level, driven by the birth of novel genes and their subsequent recruitment for high-expression in the AGs, underpins the starkly divergent reproductive AG gene repertoires even in closely-related dung beetle species (Mrinalini et al., 2021).

While early studies on insect AGs were limited to investigations of single AG genes and their gene products, recent advances in high-throughput transcriptomics and proteomics have facilitated the discovery of AG transcripts and proteins on a much larger scale (Wei et al., 2015; Sayadi et al., 2016; Bayram et al., 2017; Gotoh et al., 2018; Bayram et al., 2019; Saraswathi et al., 2020; Wigby et al., 2020; Mrinalini et al., 2021). *De novo* RNA sequencing (RNAseq) transcriptomics has allowed us to explore transcript libraries from any non-model species, even in the absence of a reference genome, because these technologies do not require *a priori* sequence knowledge. Among several short-read technologies developed for *de novo* RNAseq, Illumina is an established market leader due to its high base calling accuracy (>99.9%), high data yield, and the availability of well-established bioinformatics tools and best practices for data analysis (Conesa et al., 2016; Hölzer and Marz, 2019; Corchete et al., 2020). Third Generation long-read sequencing of transcriptomes, such as PacBio and Oxford Nanopore Technologies (ONT), eliminate the need for contig assembly and offer many advantages including long reads (>10 kbp), end-to-end transcripts, structural variants, isoform-level resolution of genes and expression. ONT long-read transcriptomics have proven useful for *de novo* and reference-based transcript sequencing and expression analysis in both model and non-model organisms (Hargreaves and Mulley, 2015); (Weirather et al., 2017; Bayega et al., 2018; Sessegolo et al., 2019; Soneson et al., 2019; Workman et al., 2019; Halstead et al., 2021; Lin et al., 2021). Despite these advances, discovering novel AG genes in non-model species can still be a challenge due to difficulties arising from lack of genomic/transcriptomic reference databases, the presence of novel species-specific genes, and the complete lack of protein homology to any other species.

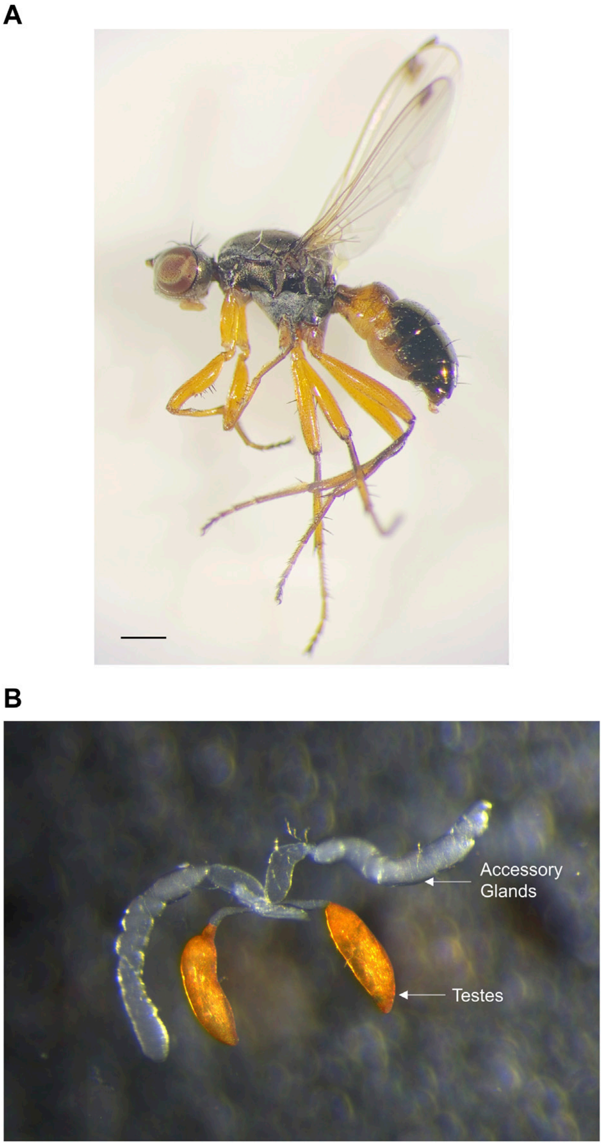

FIGURE 1
(A) Habitus of a male *Sepsis punctum* (1 mm scale bar) **(B)** Testes and accessory glands. Sepsid males can be easily distinguished from females based on foreleg modifications, which are absent in the latter. Males possess paired accessory glands that are translucent and are closely associated with the testes that are pigmented. Accessory glands of insects are hotspots for the recruitment of novel, species-specific reproductive genes, and this underpins the divergence of male ejaculatory proteins that play a crucial role in post-copulatory sexual selection.

In this study, we investigate reproductive genes and quantify gene expression in the AGs of a non-model dung fly species, *Sepsis punctum* (Diptera; Sepsidae). This is an ecologically relevant insect, often found on decaying organic material such as cattle dung and is widespread across North America and Europe. Sepsid flies are emerging study systems in a range of disciplines such as eco-toxicology (Blanckenhorn et al., 2013a; Blanckenhorn et al., 2013b), biogeography (Giesen et al., 2017; Giesen et al., 2019; Blanckenhorn et al., 2021), evo-devo (Herath et al., 2015) as well as reproductive biology (Puniamoorthy et al., 2008; Puniamoorthy et al., 2009). In particular, *S. punctum* populations in North America and Europe differ significantly with respect to mating behaviour as well as male reproductive investments and female remating frequencies, making it an interesting model for sexual selection studies (Puniamoorthy et al., 2012a; Puniamoorthy et al., 2012b; Rohner et al., 2016; Blanckenhorn et al., 2021). However, with the exception of species from the genus *Themira,* which is distantly-related to *S. punctum*, there is generally a lack of genic, genomic, or transcriptomic data for sepsid species. Here, we use reference-free GridION transcriptomics, an ONT long-read transcriptomics method, to characterize AG genes and quantify their expression in *S. punctum.* We implement a *de novo* ONT transcriptome pipeline with error correction procedures, and we evaluate gene sequence and gene expression results from this emerging technology against high-quality Illumina short-read data.

2 Materials and methods

2.1 Dissection of accessory glands from *Sepsis punctum*

The sampling and maintenance of *S. punctum* cultures followed previously published work (Puniamoorthy et al., 2012a; Rohner et al., 2016). For this study, a North American population from Ottawa (45.43 °N, −75.67 °E) was used and adult flies were housed in plastic containers measuring 11 cm by 9 cm by 9 cm and reared at a temperature of 26°C with cattle dung, sugar, and water given *ad libitum.* A mixture of mated and unmated male flies that were four to 10 day old were aspirated from culture containers into plastic vials, cooled at −20°C for 10 min, and placed on ice until dissection. Each fly was transferred to a glass slide and the abdomen was dissected into 1X PBS. Paired accessory glands were isolated and collected into 1.5 ul microcentrifuge tube snap frozen on dry ice. For the purposes of this study, we excluded the testes, ejaculatory bulb, ejaculatory duct, and aedeagus. For ONT GridION, tissues from 80 flies were pooled to allow for protocol optimisation, and for Illumina HiSeq, tissues from 63 flies were pooled. For each sequencing technology two biological replicates of pooled tissues were collected and samples were stored at−80°C until RNA extraction.

2.2 RNA extraction

Total RNA was extracted using Aurum Total RNA Mini Kit (BIO-RAD Cat # 732–6820). Samples stored at −80°C were centrifuged at 13,000 rpm for 20 min at 4°C and placed on ice. 700 ml lysis solution was added to each sample and homogenized using PTFE pestles. The lysate was centrifuged for 3 min at 4°C and the supernatant was transferred to new tube. 700 ul of 60% ethanol was added and thoroughly mixed by vortexing for 2–3 min to make sure there was no visible bilayer. 700 ul of homogenized lysate was transferred into an RNA binding column inserted into a wash tube and the set up was centrifuged for 1 min. The filtrate was discarded, and the same wash was repeated a second time. 700 ul of low stringency wash was added to column and centrifuged for 1 min and filtrate discarded. 80 ul of DNase (5ul of DNase I solution +75 ul of DNase solution) was added to each column and incubated at room temperature for 25 min. The samples were washed two more times, first with 700 ul of high stringency wash solution and second with 700 ul of low stringency wash with centrifuging for 1 min and discarding of filtrate after each wash. The samples were spun for 3 min to remove residual wash solution and the RNA binding column was transferred to 1.5 ul microcentrifuge tubes. 40 ul of elution solution was added to the membrane of the binding column and after 1 minute of membrane saturation, the sample was centrifuged for 2 min to elute total RNA.

2.3 cDNA library preparation and RNAseq

2.3.1 ONT GridION

ONT GridION offers Direct RNA or Direct cDNA library preparation and sequencing options, however these technologies require high amounts of starting RNA input. The RNA quantities of our samples were inherently low given the small size of our study species *S. punctum* (2–7 mm in length) and even smaller size of reproductive tissues (Figure 1). Therefore, the PCR-cDNA (PCB109) protocol was used, which allows for lower RNA input. Total RNA samples were submitted to Genome Institute of Singapore, Singapore, for ONT GridION long-read RNAseq. Nanodrop 2000 spectrophotometer (NanoDrop, Wilmington, DE) was used to determine RNA concentration, and quality check was performed using Agilent RNA Screentape kit with Agilent Tapestation 4200 (Agilent, Santa Clara, CA). 100 ng total RNA was used for cDNA synthesis and strand switching of full-length poly AAA tail. cDNA was amplified with 5’ barcoded primers and sequencing adapter annealing. Thirteen cycles of amplification were performed allowing an extension time of 10 min to amplify transcripts up to 12 kbp. Barcoded libraries were multiplexed by pooling at 100 fmol based on average Agilent DNA 12000 size, and sequencing was performed on one FLO-MIN106D (R9.4.1) GridION flowcell. Guppy v4.0.11 was used in high accuracy (hac) mode to perform live basecalling on GridION.

2.3.2 Illumina HiSeq

For Illumina short-read sequencing, total RNA was shipped to Genomics Research Center at University of Rochester, New York, for cDNA library preparation and sequencing. Total RNA concentration was determined with NanoDrop 1000 spectrophotometer (NanoDrop, Wilmington, DE) and RNA quality assessed with the Agilent Bioanalyzer (Agilent, Santa Clara, CA). TruSeq RNA Sample Preparation Kit V2 was used for library construction as per manufacturer’s protocols. Briefly, mRNA was purified from 100 ng total RNA with oligo-dT magnetic beads and then fragmented. First-strand cDNA was synthesized with random hexamer priming followed by second-strand cDNA synthesis. End repair and 3′ adenylation was performed on the double stranded cDNA. Illumina adaptors were ligated to both ends of the cDNA, purified by gel electrophoresis and amplified with Polymerase Chain Reaction (PCR) primers specific to the adaptor sequences to generate amplicons of approximately 200–500 bp in size. Libraries were loaded at a concentration of 8 p.m. per lane and Paired End reads of length 125 bp were sequenced on HiSeq 2500 v4 platform.

2.4 *De novo* transcriptome pipelines and gene expression analysis

Two different bioinformatics pipelines were employed, with some steps commonly implemented in both pipelines, for *de novo* transcriptome construction and analysis of ONT long-reads and Illumina short-reads (Figures 2A,B). For ONT, *de novo* gene clustering, consensus sequence calling, and gene polishing were used to derive error corrected gene sequences (Sahlin and Medvedev, 2019; Sahlin et al., 2021). For Illumina, a *de novo* transcriptome assembly approach was used to reconstruct contigs from which full length coding DNA sequences (CDS) could be derived. The two methods differed in gene expression quantification in that, read count was used as a proxy for gene expression in the case of ONT, whereas Transcript per Million (TPM) was used for Illumina based gene expression calculation (Figure 2).

2.5 *De novo* ONT pipeline

A suite of bioinformatics tools, including standalone tools and those developed by ONT, were used for *de novo* transcriptome analysis of ONT long-reads. Read quality filtering, orientation, and trimming was performed using *Pychopper v2* (https://github.com/nanoporetech/pychopper) with default parameters, and read statistics were analysed using *NanoPlot 1.32.1* (De Coster et al., 2018). A non-hybrid

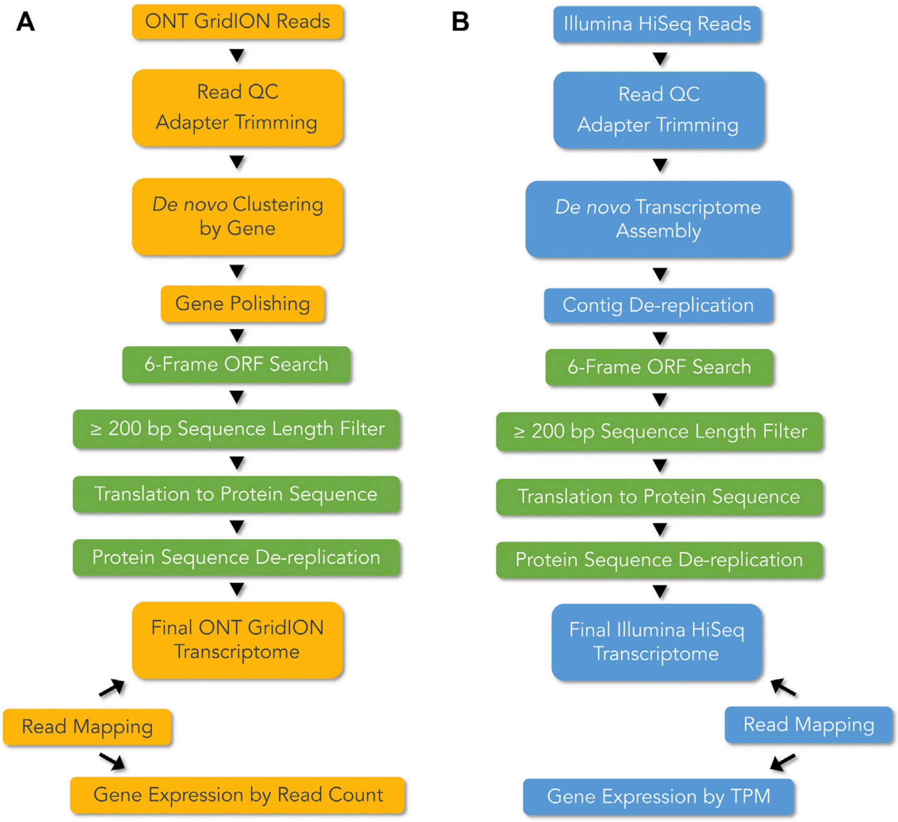
A
ONT GridION Reads
Read QC
Adapter Trimming
De novo Clustering by Gene
Gene Polishing
6-Frame ORF Search
≥ 200 bp Sequence Length Filter
Translation to Protein Sequence
Protein Sequence De-replication
Final ONT GridION Transcriptome
Read Mapping
Gene Expression by Read Count
B
Illumina HiSeq Reads
Read QC
Adapter Trimming
De novo Transcriptome Assembly
Contig De-replication
6-Frame ORF Search
≥ 200 bp Sequence Length Filter
Translation to Protein Sequence
Protein Sequence De-replication
Final Illumina HiSeq Transcriptome
Read Mapping
Gene Expression by TPM

FIGURE 2
De novo transcriptome pipelines for **(A)** ONT long-read technology, and **(B)** Illumina short-read technology. mRNA from the accessory glands of *Sepsis punctum*, was used for cDNA library preparation and RNAseq using ONT long-read and Illumina short-read technologies. ONT transcripts were generated by *de novo* gene clustering, consensus generation, and gene polishing, whereas for Illumina short-reads were *de novo* assembled to reconstruct contigs. Preliminary transcriptomes were filtered using common steps (shown in green), using six-frame Open Reading Frame (ORF) searches, ≥ 200 bp length, and dereplication at 100% protein identity, to derive final transcriptomes. Gene expression was quantified using read count as a proxy for ONT and Transcript per Million (TPM) for Illumina.

approach of using ONT long-reads for error correction was employed, as our ONT and Illumina read data were derived from separate biological samples. Error correction of each ONT transcriptome was performed using ONT long-reads within the same biological sample. *IsONclust2,* implemented in *pipeline-nanopore-denovo-isoforms* (https://github.com/nanoporetech/isONclust2) was used for *de novo* clustering of ONT long-reads and one sequence cluster was generated for each gene (Sahlin and Medvedev, 2019; Sahlin et al., 2021). Consensus sequences were called for each sequence cluster to generate one consensus sequence per gene. The consensus gene sequences were further polished using raw reads in *medaka 1.2.5* (https://github.com/nanoporetech/medaka). Open Reading Frames (ORFs) were derived by translating polished sequences in all six frames using *getorf* provided with *EMBOSS:6.6.0* (https://www.bioinformatics.nl/cgi-bin/emboss/getorf). ORF sequences ≥200 bp were translated into protein sequences using *transeq* provided with *EMBOSS:6.6.0* (https://www.bioinformatics.nl/cgi-bin/emboss/transeq), and a final dereplication was performed at 100% protein sequence identity using *CD-HIT v4.7* (Li and Godzik, 2006; Fu et al., 2012). For gene expression quantification, long-reads were mapped back to filtered *de novo* ONT transcriptomes using *minimap2*, excluding any secondary alignments (Li, 2018). *Samtools 1.7* (Li et al., 2009) was used to further filter aligned reads, with any supplementary and secondary alignments discarded. Only reads aligning on ≥ 80% of their length were counted towards gene expression quantification.

2.6 *De novo* Illumina pipeline

Raw reads were processed in *Trimmomatic-0.36* (Bolger et al., 2014) for adapter removal and quality trimming. A sliding window quality score cut-off of Q30 was applied and reads of minimum 100 bp in length were retained. For each sample, cleaned reads were *de novo* assembled into contigs using *Trinity v2.8.6* (Grabherr et al., 2011), and the resulting contigs were de-replicated at 100% identity at nucleotide level using *CDHIT 4.7* (Li and Godzik, 2006; Fu et al., 2012). The remaining contigs were translated in all six frames to search for ORF prediction using *getorf* (*EMBOSS:6.6.0*) (https://www.bioinformatics.nl/cgi-bin/emboss/getorf), and all sequences containing ORFs of ≥200 bp were retained. The contigs with ORFs were translated into protein sequences using *transeq* (*EMBOSS:6.6.0*) (https://www.bioinformatics.nl/cgi-bin/emboss/transeq), and a final dereplication was performed at 100% protein sequence identity. Reads were mapped back to the filtered transcriptome assembly using an alignment-free method in *salmon v1.0.0* (Patro et al., 2017) to generate TPM values that represent gene expression.

2.7 Sequence curation and gene orthology

Sepsis punctum lacks species-specific reference databases to compare our *de novo* transcriptome constructions. While genomic/transcriptomic data are available for *Themira sp.,* it is not the ideal species for verifying the accuracy of *S. punctum* AG transcripts because it is from a basal, distantly related genus and insect AG genes and protein compositions vary at species and even population levels (Swanson et al., 2001; Goenaga et al., 2015; Abry et al., 2017; Bayram et al., 2019; Mrinalini et al., 2021). Therefore, building on our *de novo* approach transcriptomics, a reference-free approach was taken for transcript curation and an extensive manual curation of *S. punctum* AG genes was performed.

A gene expression cut-off was applied, and the top 100 highest expressed transcripts were selected from each of the four *de novo* transcriptomes (ONT-1, ONT-2, ILL-1, ILL-2) since transcriptome-based quantification of gene expression generally shows a steep drop after the first few transcripts. The subset of 400 sequences was further examined to filter out chimeric and contaminant bacterial and nematode sequences with BLASTX in *nr* using *DIAMOND v 0.8* (Buchfink et al., 2014; Buchfink et al., 2021). Using the cleaned sequence set from each sample, putatively orthologous genes were established in the *de novo* transcriptomes of the remaining three samples by a reciprocal BLASTP with an *e-value* cut-off of 1e-5. These putative orthologs were further curated by manually examining end-to-end alignments. Finally, a set of 40 high-confidence and high-quality accessory gland genes were derived with orthologs established in all four samples and used for the downstream analyses.

2.8 Evolutionary novelty and functional classification of curated genes

To examine whether these 40 genes included novel AG genes that have evolved in *S. punctum*, we used a two-step approach: First, we conducted BLASTP and BLASTX searches against *nr* that contains protein sequences from broad taxonomic categories; and second, using BLASTN and TBLASTX against the genome of the closest sepsid relative, *Themira minor* (GenBank Accession No. GCA_001014575.1), to potentially identify unannotated genic regions in genomic scaffolds. A cut-off *e-value* of 1e-5 and at least 50% identity over 70% of the sequence was applied to call BLAST hits. Genes with no hits to both *nr* and *Themira minor* genome were defined as novel genes that have *de novo* evolved in *S. punctum*. Functional annotation was also performed for the 40 genes and broad functional categories were assigned based on BLASTP results against *nr*. Secretory proteins were identified by translating coding DNA sequences in MEGAX and analysing the resulting proteins in SignalP 5.0 webserver (https://services.healthtech.dtu.dk/service.php?SignalP-5.0) to check for the presence of signal peptides.

2.9 Evaluation of ONT transcriptome

ONT long-reads are prone to high error rates, therefore the usefulness of our *de novo* ONT transcriptomics pipeline and error correction procedures in mitigating the effects of sequencing errors was evaluated (Figure 2A). Given that our study is reference-free, and Illumina short-reads are of high quality, a sequence similarity analysis was performed by comparing ONT gene sequences to Illumina gene sequences. The Illumina sample, ILL-1, was designated as the control sample and the remaining samples ONT-1, ONT-2, as well as ILL-2 were the test samples. Comparing ILL-2 to ILL-1 control sample allowed for within-Illumina assessment that can uncover effects of tissue pooling and natural genetic variation in the population. Sequence similarity values were generated by performing BLASTN of nucleotide sequences and BLASTP of translated protein sequences from 40 genes of three test samples against 40 genes from ILL-1 control. Percent sequence similarities were summarized, including median sequence similarity and percentage of sequences with 100% match to ILL-1 control. Similarities of all 40 genes were plotted as a heatmap for visualization at individual gene level.

3 Results

3.1 *De novo* transcriptome statistics

After implementing two separate *de novo* transcriptome pipelines for ONT and Illumina read data (Figure 2), we generated summary statistics for the transcriptomes. Table 1 provides the statistics for experimental details, cDNA library preparation, RNAseq reads, and filtered *de novo* transcriptomes.

TABLE 1 Statistics from ONT long-read and Illumina short-read cDNA library preparation, RNAseq reads filtering, and final *de novo* transcriptomes.

	ONT		Illumina	
	ONT-1	ONT-2	ILL-1	ILL-2
Experimental Details				
No. of pooled individuals	80	80	63	63
Total RNA used for library prep (ng/ul)	100	100	100	100
RNAseq Read Statistics				
No. of Reads (Million)	4.48	5.28	29.64	32.96
Total No. of Bases (Gbp)	2.05	2.14	7.41	8.24
Max. Read Length (bp)	11,837	9,370	125	125
Mean Read Length (bp)	456	405	125	125
Min. Read Length (bp)	58	50	125	125
Mean Read Quality	11.7	11.6	35.49	35.52
GC Content (%)	39.5	39.2	46	46
Filtered *De Novo* Transcriptome Assembly Statistics				
No. of Transcripts	44,958	33,564	44,523	45,681
Largest Transcript (bp)	4,422	2,964	13,932	12,372
Mean Transcript Length (bp)	358	342	513	492
GC Content (%)	54.3	53.9	54.4	53.7

3.2 RNAseq read statistics

ONT long-read sequencing generated 4.48 M and 5.28 M reads for samples ONT-1 and ONT-2 respectively, whereas Illumina sequencing generated 29.64 M and 32.96 M reads for ILL-1 and ILL-2 respectively. Long-read lengths range from 50 bp to a maximum of 11,837 bp for ONT-1 and 50 bp to 9,370 bp for ONT-2. Average read quality, represented by Phred scores, are 11.7 and 11.6 for ONT-1 and ONT-2 respectively. For Illumina, read qualities were much higher at 35.5 on average from the two sample. Filtered RNAseq read GC content was 39.5% and 39.2% for ONT-1 and ONT-2, whereas GC content for both Illumina samples was 46%.

After quality filtering, most long-reads were found to be distributed within the 4000 bp range, with an average base quality ≤19 (minimum base quality 7) for both ONT-1 and ONT-2 (Figures 3A,B). A histogram of read length distribution shows that the maximum number of long-reads occur in the range of 250–450 bp for both samples and read N50 were 448 bp and 407 bp for ONT-1 and ONT-2 respectively (Figures 3C,D; Table 1).

3.3 Filtered *de novo* transcriptome statistics

After *de novo* gene clustering, gene consensus calling, gene polishing and subsequent filtering (Figure 2A), *de novo* transcriptomes contained 44,958 and 33,564 transcripts in ONT-1 and ONT-2 respectively (Table 1). For Illumina, *de novo* transcriptome assembly followed by filtering steps showed 44,523 and 45,681 assembled contigs for ILL-1 and ILL-2 respectively (Table 1). ONT long-read transcript N50 were 363 and 345 bp, with the largest transcripts at 4,422 and 2,964 bp, for ONT-1 and ONT-2 respectively (Table 1). For Illumina short-read transcripts, both values were much higher, with N50 values at 702 and 645 bp and longest transcript lengths at 13,932 and 12,372 bp, for ILL-1 and ILL-2 respectively. Although GC content of filtered RNAseq reads varied between ONT and Illumina, the GC content of the final *de novo* transcriptomes from both technologies were comparable at c. 54% (Table 1).

3.4 Evolutionary and functional characterisation

The functional aspects of *S. punctum* AG genes were characterized using gene expression analysis, BLASTP against NCBI non-redundant database (*nr*) Version 2.2.26 for functional annotation, and signal peptide analysis of the translated protein sequences (Figure 4). Ranking of accessory gland genes from high expression (1) to low expression (40) showed high level of congruence in overall gene expression pattern across all four samples despite using two different methods of quantification for the two technologies, i.e., read count for ONT samples and TPM for Illumina samples (Figure 4A). The highest expressed transcript showed read count of 72,731 and 68,120 for ONT-1 and ONT-2 respectively and TPM of 62,424 and 76,279 for ILL-1

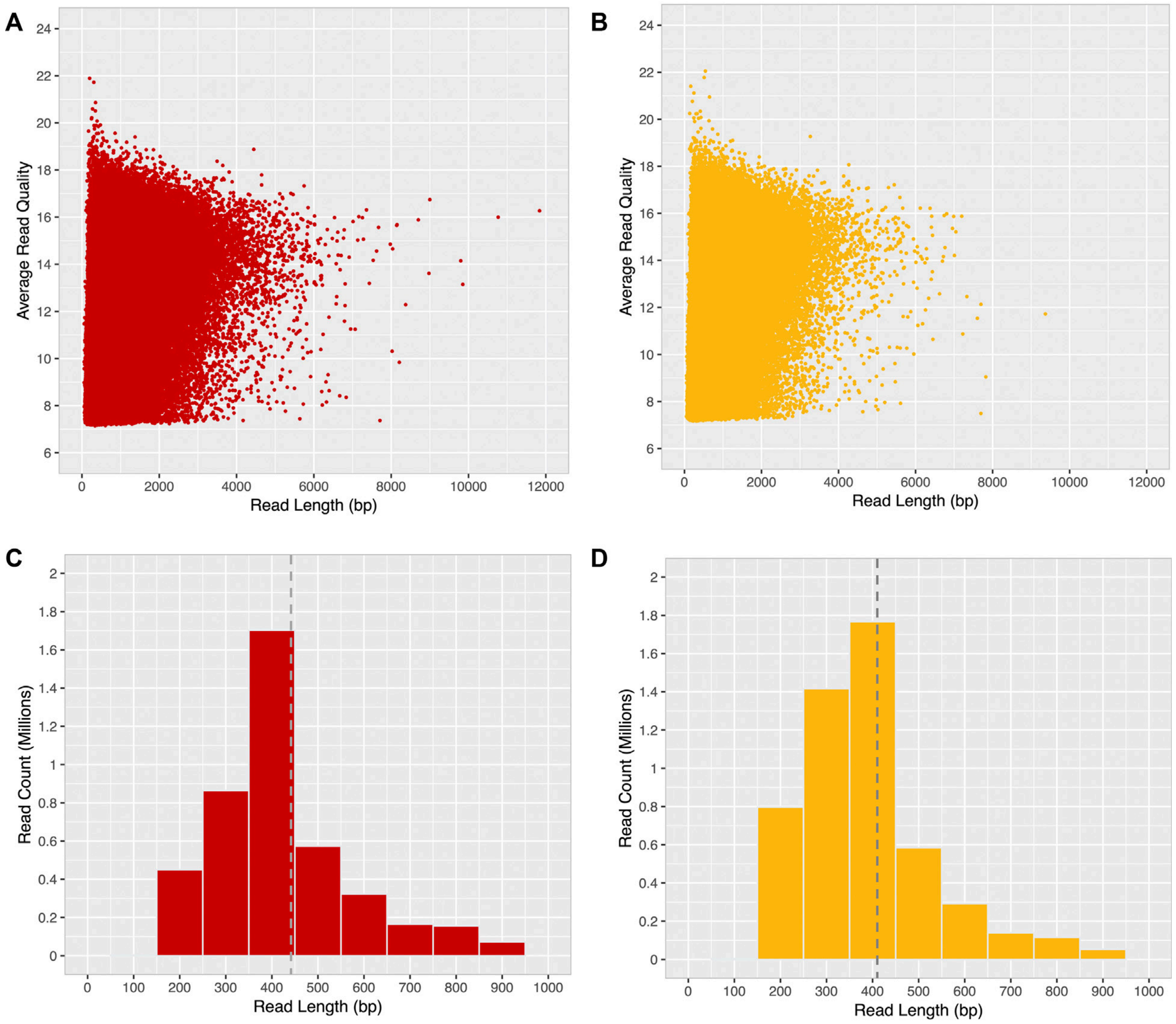

FIGURE 3
Read quality and read length distribution of filtered ONT long-reads. After a minimum base quality cut-off of 7, a majority of long-reads for both samples **(A)** ONT-1, and **(B)** ONT-2 were within the 4 kbp range, with an average base quality ≤19. Most long-reads were 400 bp in length for both **(C)** ONT-1, and **(D)** ONT-2, with N50 values of 448 bp and 407 bp represented as grey dotted lines.

and ILL-2 samples respectively. In both ONT and Illumina, a steep drop in gene expression occurred and expression levels tapered off from transcripts ranked at 12 and 13 and reached negligible levels by transcript 40.

Of the 40 genes analysed, 65% (26) were novel genes that evolved specifically in the genome of *S. punctum* as they did not return any hits against *nr* or against the genome of the closest sepsid relative, *Themira minor*. The remaining 14 genes were found in other insect species, some of which were interestingly identified from the male AGs of fly species as well (e.g., *Drosophila sp.*). Functional annotation showed that *S. punctum* AG genes fall into four main categories: *unknown function*; *protease inhibitors*; *C-type lectins*; and *cystatins*. Incidentally, all novel genes that *de novo* evolved in *S. punctum* also fell into the category of *unknown function* due to lack of even partial BLASTP homology in their protein sequences. However, these genes accounted for most of the gene expression in the accessory glands, i.e., 87% and 90% of total expression in ONT-1 and ONT-2 respectively, and 78% and 71% of total expression in ILL-1 and ILL2 respectively (Figure 4C). Of the remaining 14 genes, 10 genes were *protease inhibitors,* and they were the second highest expressed gene category, with 10% and 8% in ONT-1 and ONT-2, and 17% and 23% in ILL-1 and ILL-2 respectively (Figures 4B,C). Two genes each for *cystatins* and *C-type lectin* were found, and both gene families were the least expressed,

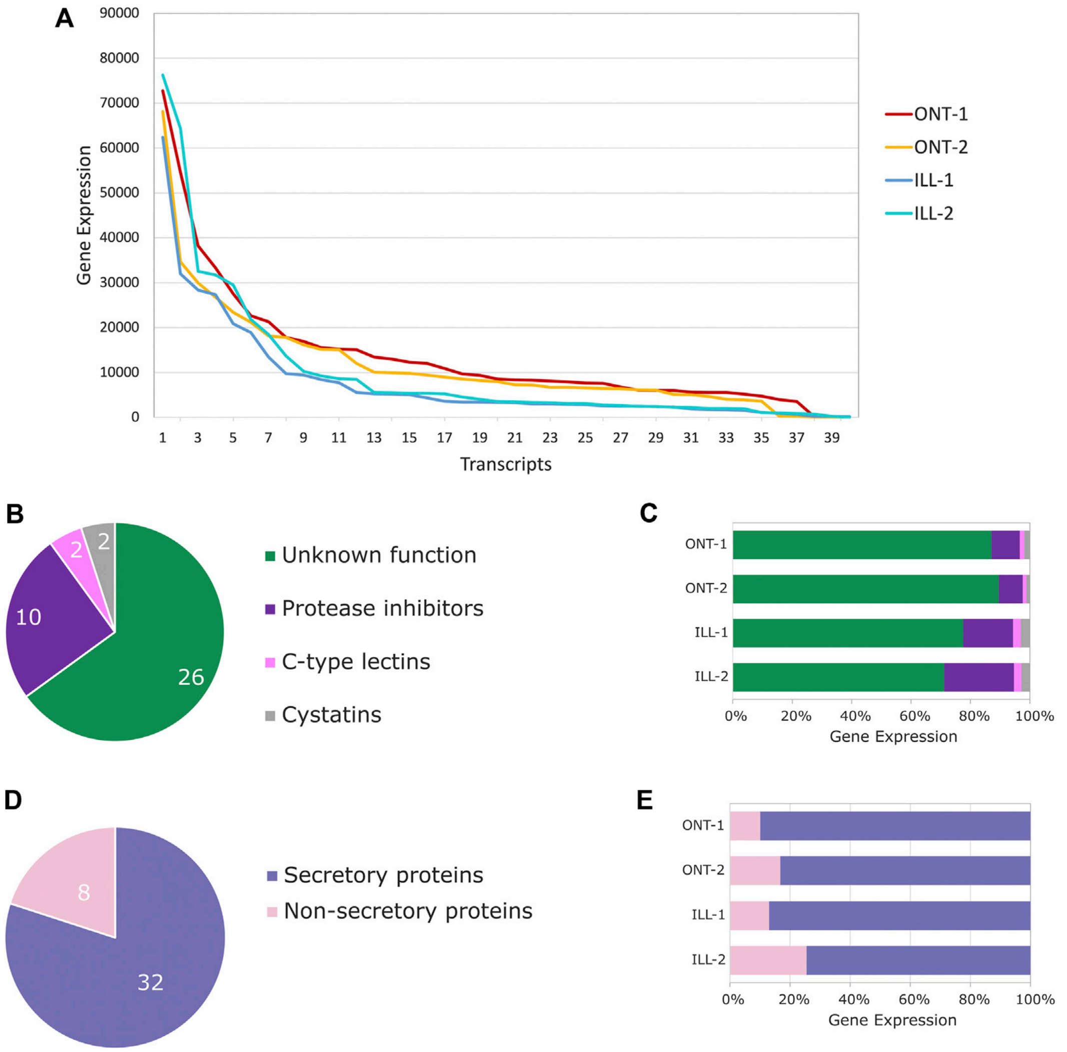

FIGURE 4
Gene expression and gene function in *S. punctum* accessory gland genes. Analysis of 40 highest expressed genes in *S. punctum* AGs showed **(A)** a high level of congruence in overall gene expression pattern between ONT read counts and Illumina Transcripts per Million (TPM) **(B)** 26 genes of *unknown function*, 10 *protease inhibitors,* and 2 each of *C-type lectins* and *cystatins,* **(C)** genes of *unknown function* were the highest expressed gene category, followed by *protease inhibitors,* in both ONT and Illumina samples **(D)** 32 genes encoding secretory proteins and eight genes encoding non-secretory proteins, and **(E)** genes encoding secretory proteins accounted for the majority of expression in both ONT and Illumina samples.

with 1%–3% of total expression in both ONT and Illumina (Figures 4B,C). Sequence analysis for the presence of secretory signals showed that 80% (32) of 40 AG genes synthesize proteins that were secretory in nature (Figure 4B). Moreover, genes encoding secretory proteins also accounted for the majority AG gene expression, with 90% and 83% of total expression in samples ONT-1 and ONT-2 and 87% and 75% in samples ILL-1 and ILL-2 respectively.

Gene expression quantification of individual genes in all the four samples are represented as a heatmap, with genes grouped by functional categories (Figure 5A). Novel genes of unknown function were among the highest expressed genes in *S. punctum* accessory glands in both ONT and Illumina samples (Figure 5A). Expression levels of several individual genes were largely comparable across samples and across sequencing technologies, with some variation between and even within technologies (Figure 5A).

3.5 Evaluation of ONT transcriptomes

Summary statistics of nucleotide and protein sequence similarities within each technology and comparing GridION samples (ONT-1 and ONT-2) against a representative Illumina control sample (ILL-1) are shown in Table 2. Median sequence similarities within each technology were congruent

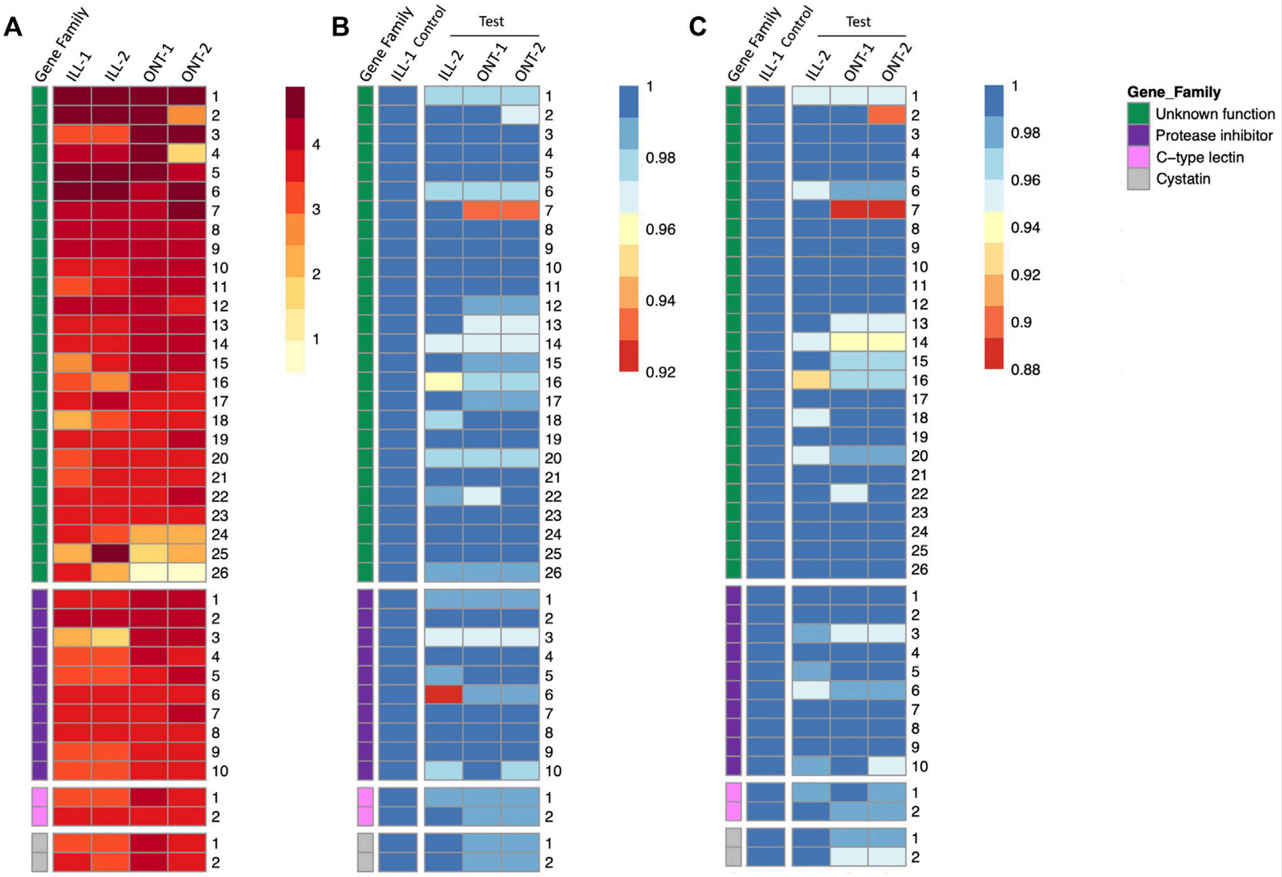

FIGURE 5
Heatmaps of gene expression and sequence similarity from 40 highest expressed accessory gland genes in *Sepsis punctum*. **(A)** Gene expression quantified by read count for ONT samples and TPM for Illumina samples and genes are grouped by functional categories. Gene of *unknown function* were among the highest expressed in both ONT and Illumina samples. Sequence similarities at **(B)** nucleotide level, and **(C)** protein level are plotted by comparing test samples ONT-1, ONT-2, and ILL-2 against a designated control sample ILL-1.

TABLE 2 Summary statistics of sequence similarity at nucleotide and protein level. Sequence similarities were generated within each technology and across technologies by comparing GridION test samples (ONT-1 and ONT-2) against a representative Illumina control sample (ILL-1).

	Within technology		Illumina control v/s GridION	
	ONT	Illumina	ONT-1	ONT-2
Nucleotide Level				
Median sequence similarity of 40 genes	100.00%	99.65%	99.60%	99.53%
Sequences with 100% similarity	93%	65%	53%	50%
Protein Level				
Median sequence similarity of 40 genes	100.00%	99.31%	98.99%	98.93%
Sequences with 100% similarity	90%	43%	43%	45%

with each other, with Illumina showing 99.65% and 99.31% at nucleotide and protein levels respectively and GridION showing 100% for both. Median sequence similarities across technologies were comparable to within technology values as well. The two GridION test samples ONT-1 and ONT-2, when compared against Illumina control, showed median sequence similarities of 99.60% and 99.53% at nucleotide level and 98.99% and 98.93% at protein level respectively.

In GridION, over 90% of both nucleotide and protein sequences showed 100% similarity across the two samples. In contrast, a much lower proportion of sequences with 100% similarity were seen among the two Illumina samples, with 65% among nucleotide sequences and 43% among protein sequences. Further, 53% and 50% of GridION sequences were 100% similar to Illumina control at nucleotide level, whereas 43% and 45% of GridION sequences were 100% similar to Illumina control proteins. Sequences similarities of individual genes are represented as heatmaps in Figures 5B,C. No trend is observed within or among the four functional categories, i.e., novel genes, *protease inhibitors, C-type lectins,* and *cystatins.*

4 DISCUSSION

4.1 Adopting ONT long-read transcriptomics in non-model systems

Illumina has been the technology of choice for RNAseq transcriptomics given high data quality and well-established bioinformatics pipelines (Conesa et al., 2016; Hölzer and Marz, 2019; Corchete et al., 2020). However, short CDS reads still need to be stitched together *in silico* using a reference database or *via de novo* concatenated assembly to create contiguous sequences (contigs), from which full-length, protein-coding transcripts are derived. In non-model organisms with no reference databases, partially or spuriously assembled CDS and chimeric transcripts are common pitfalls (Conesa et al., 2016; Freedman et al., 2021). Moreover, in evolutionarily closely-related genes and in gene families consisting of multiple isoforms, it is difficult to resolve CDS and quantify gene expression to isoform level using short-reads (Steijger et al., 2013; Conesa et al., 2016).

Recent inventions in Third Generation Sequencing technologies return high throughput, long-read sequences that provide us with end-to-end transcripts, thereby eliminating the need for assembling contigs. Oxford Nanopore Technologies (ONT), arguably the leader in long-read transcriptomics, offer short to ultra-long DNA/RNA molecules longer than 10 kbp in length. Single-use cartridges with pre-loaded reagents that can be easily used with portable, bench-top instruments makes ONT a convenient platform. Further, ONT offers cDNA sequencing, in both polymerase chain reaction (PCR) based and PCR-free formats, and direct RNA sequencing that bypasses the need for converting RNA into cDNA. Despite these advantages, a major obstacle to the adoption of ONT long-read transcriptomics is the high error rate reported in both cDNA and direct RNA sequencing (Workman et al., 2019). Moreover, although over 555 tools are available for long-read analysis (https://long-read-tools.org/) (Amarasinghe et al., 2020; Amarasinghe et al., 2021), no clear bioinformatics pipelines have been established for *de novo* ONT transcriptomics and we are still in early stages of reference-free ONT transcriptomics. One way to mitigate sequencing errors is by adopting a reference-based approach, and as mentioned earlier, well-characterized, species-specific reference databases can go a long way in resolving transcript sequences and their expression levels, even to the level of isoforms (Sessegolo et al., 2019; Workman et al., 2019; Dong et al., 2021). This approach with ONT is often used in species with good quality reference databases such as humans (Weirather et al., 2017; Soneson et al., 2019; Workman et al., 2019), mice (Sessegolo et al., 2019), cattle (Halstead et al., 2021), fruit flies (Bayega et al., 2018), viruses (Boldogkői et al., 2018), and well-studied plants (Cui et al., 2020; Wang et al., 2021). However, in non-model species without reference databases, the uptake of ONT long-read transcriptomics is still not widespread.

We present the first study to adapt ONT long-read RNAseq in a completely *de novo* and reference-free approach to characterize novel genes in any animal species. Our study shows that ONT can be successfully used to sequence mRNA transcriptomes of even minute tissues such as insect AGs, with more than two dozen novel AG genes discovered from *S. punctum* (Figure 4B). Thus far, novel insect reproductive genes and gene products have been primarily discovered using Illumina based RNAseq and microarrays (Vibranovski et al., 2009; Sayadi et al., 2016; Bayram et al., 2017; Vedelek et al., 2018; Mrinalini et al., 2021), as well as using traditional protein or proteomics analyses (Peferoen and De loof, 1984; Parthasarathy et al., 2009; Goenaga et al., 2015; Gorshkov et al., 2015; Wei et al., 2015; Yamane et al., 2015). Gene sequences from our assembly-free and reference-free ONT long-read pipeline were of high quality, with sequence similarity levels comparable to genes derived from Illumina short-read assembly (Table 2; Figures 5B,C). Error-prone ONT long-reads, when combined with effective read quality filtering and rigorous error correction methods, can be a reliable new technology for end-to-end gene discovery and expression quantification that eliminates the need for transcriptome assemblies or reference databases. Moreover, gene expression quantification for ONT was highly congruent with Illumina (Figures 4A,D,E). Finally, we were able to achieve results comparable to Illumina without the need for transcriptome assembly, and using only a quarter the number of sequenced bases than that of Illumina sequencing (Table 1). ONT long-reads can therefore be a reliable, resource-friendly, and costeffective alternative that can achieve end-to-end sequencing of novel genes from non-model species, even in the absence of a reference database.

4.2 *De novo* gene discovery and gene expression quantification in a non-model species: ONT GridION v/s Illumina

Our *de novo* ONT transcriptome pipeline, incorporating read quality filtering and rigorous post-sequencing error correction procedures, successfully mitigated high error rates in ONT long-reads (Figure 2A, 4B) (Sahlin and Medvedev, 2019; Sahlin et al., 2021). Using ONT GridION, we successfully discovery novel AG genes, which constituted 65% of 40 high-quality and high-confidence gene set and cross-verified against evidence from Illumina transcriptomes (Figure 4B). In our analysis of sequence similarity, median sequence similarities of genes and proteins in ONT (to their respective orthologs in the Illumina control sample) are comparable to within-Illumina median sequence similarities (Table 2). Further, although two different methods of gene expression quantification were used for ONT and Illumina pipelines (Figure 2A), there was a high degree of congruence between the two technologies with respect to overall expression pattern (Figure 4A) and in terms of proportion of expression in each functional category (Figures 4D,E). At the level of individual *S. punctum* genes, expression patterns were less congruent, and could be attributed to variation in biological replicates (Figure 5A).

When comparing sequences across technologies at the nucleotide level, ONT samples (ONT-1 and ONT-2) had 53% and 50% sequences that were 100% similar to Illumina control orthologs (Table 2). At the protein level, ONT samples show 43% and 45% of sequences are 100% similar to the Illumina control orthologs (Table 2). This is comparable to within Illumina assessment, where 65% of nucleotide sequences and 43% of protein sequences were found to be 100% similar to each other among the two samples. Due to the small size of *S. punctum* and the minute size of our tissue, our samples were collected by pooling tissues dissected from multiple flies collected at different time points. Therefore, the lack of 100% sequence similarity among a majority of sequences even within Illumina suggests that rather than sequencing errors, natural genetic variation together with differential representation and incorporation of transcripts during *de novo* transcriptome construction could be the likely source of sequence variation. This is corroborated by the much larger proportion (over 90%) of genes that are 100% similar within ONT at both nucleotide and protein levels. ONT uses a consensus calling and polishing approach which could iron out high levels of differences, be it sequencing errors or sequence variation.

In order to find genes that are uniquely identifiable in either ONT or Illumina, we investigated samples within each technology. We found seven transcripts in total, of which six were uniquely present in ONT samples but not in Illumina, and one was uniquely present in Illumina samples but not in ONT. Among transcripts identified only in ONT, we found two potential genes, and both of these were 100% identical at the protein level when compared across the two samples. However, the functional relevance of both these genes is unclear as they do not show similarity to any other proteins in *nr,* they do not contain a signal peptide, and they do not show high expression levels in AGs of *S. punctum* either. Of the remaining four transcripts unique to ONT, one was found to be a chimera of poly-AAA tail sequence that had been missed by the filter. Three other transcripts showed poor sequence similarity, and were incongruent between samples either throughout the length of the sequence or at C-terminal ends of their protein sequence, therefore making identification of the correct sequence unreliable. In the single gene unique to Illumina samples, the C-terminal was partially absent in one of the samples and the functional significance was not clear in this case either.

4.3 Evolution of novel reproductive genes

We successfully discover 26 novel genes that are expressed in the male AGs of *Sepsis punctum* (Figure 4B). These novel AG genes are also of *unknown function* as we find no homology to any other proteins in *nr*. However, these genes are among the highest expressed in AG tissues, both as a gene category (81% of total expression) (Figure 4C) and at the level of individual genes (Figure 5A). Novel genes with high expression that also encode seminal fluid proteins may play a role in AG function and *S. punctum* reproduction (Figure 4C). Our results are similar to patterns of gene evolution and gene expression that have been reported in other insect species (Parthasarathy et al., 2009; Sayadi et al., 2016; Bayram et al., 2017; Gotoh et al., 2018; Bayram et al., 2019; Mrinalini et al., 2021). Much of what we know about the mechanisms underlying insect reproductive gene evolution comes from model species like *Drosophila melanogaster* (Chen et al., 2013; Sirot, 2019). It is generally accepted that novel genes often arise *via* neo- or sub-functionalization following duplication of existing protein coding genes (Wagstaff and Begun, 2007; Mancini et al., 2011; LaFlamme et al., 2012; Sirot et al., 2014; Garlovsky et al., 2020). However, novel genes may also evolve *de novo* from non-coding regions of a genome, and several studies have suggested that insect reproductive tissues are candidate sites for the recruitment of *de novo* evolved genes (Levine et al., 2006; Begun et al., 2007; Reinhardt et al., 2013; Mrinalini et al., 2021; Rivard et al., 2021). Novel AG genes have been shown to evolve and diversify rapidly through genomic innovation followed by recruitment for high expression in insect AG tissue (Mrinalini et al., 2021), as well as through rapid evolution at the primary sequence level (Findlay and Swanson, 2010; Avila et al., 2011).

Despite such evidence, the selective forces driving this rapid evolution and diversification of reproductive genes are not always well understood. For instance, adaptive evolution and positive selection have often been invoked to explain this phenomenon in various animal species (Birkhead and Pizzari, 2002; Swanson and Vacquier, 2002; Torgerson et al., 2002;

Jansa et al., 2003; Teng et al., 2017; Weber et al., 2017; Rowe et al., 2020). Specifically, adaptive evolution *via* post-copulatory sexual selection predicts strong directional selection for fast-evolving reproductive genes that would increase competitive fertilization among ejaculates or mediate ejaculate-female interactions. Molecular evolutionary studies comparing dN/dS rates of reproductive genes and non-reproductive genes among closely-related species also suggest that this can drive divergence between species (Ahmed-Braimah et al., 2017; Rowe et al., 2020). Others suggest that positive selection can be related to immune functions because seminal fluids may have antibacterial properties (Mueller et al., 2007; Wong et al., 2008). However, there is also evidence that most reproductive genes may not evolve adaptively since less than 12% of seminal fluid proteins in *D. melanogaster* actually display evidence of positive selection and that both slow *and* fast evolving genes have been found to be functionally important for insect reproduction (Findlay et al., 2008; Findlay and Swanson, 2010; Hurtado et al., 2021; Patlar et al., 2021). Therefore, relaxed selection has been proposed as an alternative hypothesis given that the expression of reproductive genes is sex-specific and selection is also limited to each sex (Dapper and Wade, 2020). Studies that assess dN/dS taking into account polymorphisms or variation at the population level suggest that more than 50% of AG genes are in fact under relaxed selection (Patlar et al., 2021). *Sepsis punctum* populations in North America and Europe exhibit stark variation in both female remating rates as well as male reproductive investments (Puniamoorthy et al., 2012a; Rohner et al., 2016), and comparative molecular evolutionary analysis across multiple populations may provide interesting insights on the selective forces behind this rapid evolution and diversification of reproductive genes.

4.4 Role of *protease inhibitors*, *C-type lectins*, and *cystatins*

In addition to novel genes, *S. punctum* AG transcriptomes revealed ten protease inhibitors and two genes each of C-type lectins and cystatins (Figure 4C). *Protease inhibitors* were the largest group and the second highest expressed functional category of genes in *S. punctum* (Figures 4B,C). Consisting of a large and diverse group of genes that synthesize many classes of proteins, *protease inhibitors* have been found in seminal fluids of *D. melanogaster* (Swanson et al., 2001; Mueller et al., 2008). They play a role in male sperm competitiveness and protect seminal fluids and sperm from proteolysis in the sperm storage organs of female flies post-mating (Park and Wolfner, 1995; Mueller et al., 2008). *C-type lectins* are immune-related genes that have been found in many insects that could be involved in protecting seminal fluids from microbial infections (Tian et al., 2017). *Cystatins* are also a large and diverse group of genes, that regulate the activity of cysteine and serine proteases. Although little is known about their specific functions in male reproductive tissues, *cystatins* have been found in seminal fluids and reproductive tissues of flies, flatworms and ticks (Sonenshine et al., 2011; Geadkaew et al., 2014; Garlovsky et al., 2020). They may be involved in spermatogenesis and fertilization (Geadkaew et al., 2014), and have been found to play a critical role in regulating programmed cell death during embryogenesis in plants (Zhao et al., 2014).

4.5 Genes encoding secretory proteins

Eighty percent (32) of AG genes analysed in *S. punctum* were likely to encode secretory proteins (Figure 4D). All 10 *protease inhibitors* and both *C-type lectins* and *cystatins* contain secretory signals, whereas among 26 novel *S. punctum* genes, 69% (18) encoded secretory proteins. These genes also accounted for the majority of AG gene expression, with an average of 84% of total expression in the four samples (Figure 4E). These patterns observed in *S. punctum* were similar to those found in other insects such as dung beetles, where 73% of AG genes analysed likely encoded secretory proteins, accounting for over 80% of total gene expression (Mrinalini et al., 2021). This supports that the primary function of male AG in most male insects is the synthesis of secretory proteins.

5 Conclusion

We discovered 26 novel reproductive genes that show high expression in the accessory glands of male *S. punctum.* We find that by implementing rigorous post-sequencing error correction procedures, error-prone ONT long-reads can produce gene sequence and gene expression data that are comparable to Illumina. Our study demonstrates that *de novo* ONT long-read transcriptomics is a reliable approach for novel gene discovery and gene expression analysis in the absence of reference databases. Gene discovery in non-model insects has important implications for understanding fundamental evolutionary processes such as phenotypic trait evolution, adaptation, and speciation. In particular, male reproductive genes of insects are known to synthesize seminal fluid proteins that interact with the female reproductive environment and thereby play a role in post-copulatory sexual selection. Hence, understanding rapid specialization and diversification of male reproductive genes in a species helps shed light on mechanisms of divergence of populations and the processes of incipient speciation.

Data availability statement

The datasets generated and analysed during the current study are available in GenBank at https://www.ncbi.nlm.nih.gov/and can be accessed with Project ID: PRJNA765219.

Author contributions

MW and NP conceptualized and designed the study. Both authors collected the samples, and MW generated transcriptome data and conducted all bioinformatics and data analyses. Both authors were involved in preparation of the manuscript.

Funding

This work was supported by Ministry of Education (Singapore) Tier-1 research grants awarded to NP (R154-000-A75-114; R154-000-C16-114).

Acknowledgments

We thank members of the Reprolab at the National University of Singapore as well as Centre for Reproductive Evolution at Syracuse University for their support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abry, M. F., Kimenyi, K. M., Masiga, D., and Kulohoma, B. W. (2017). Comparative genomics identifies male accessory gland proteins in five Glossina species. *Wellcome Open Res.* 2, 73. doi:10.12688/wellcomeopenres.12445.2

Ahmed-Braimah, Y. H. (2016). Multiple genes cause postmating prezygotic reproductive isolation in the *Drosophila virilis* group. *G3-Genes Genomes Genet.* 6 (12), 4067–4076. doi:10.1534/g3.116.033340

Ahmed-Braimah, Y. H., Unckless, R. L., and Clark, A. G. (2017). Evolutionary dynamics of male reproductive genes in the drosophila virilis subgroup. *G3-Genes Genomes Genet.* 7 (9), 3145–3155. doi:10.1534/g3.117.1136

Amarasinghe, S. L., Ritchie, M. E., and Gouil, Q. (2021). Long-Read-Tools.Org: An interactive catalogue of analysis methods for long-read sequencing data. *GigaScience* 10 (2), giab003–7. doi:10.1093/gigascience/giab003

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21 (1), 30–16. doi:10.1186/s13059-020-1935-5

Avila, F. W., Sirot, L. K., LaFlamme, B. A., Rubinstein, C. D., and Wolfner, M. F. (2011). Insect seminal fluid proteins: Identification and function. *Annu. Rev. Entomol.* 56, 21–40. doi:10.1146/ANNUREV-ENTO-120709-144823

Bayega, A., Oikonomopoulos, S., Zorbas, E., and Wang, Y. C. (2018). Transcriptome landscape of the developing olive fruit fly embryo delineated by Oxford Nanopore long-read RNA-Seq. *bioRxiv*, 478172. doi:10.1101/478172

Bayram, H., SayAdi, A., Goenaga, J., ImmonEn, E., and Arnqvist, G. (2017). Novel seminal fluid proteins in the seed beetle Callosobruchus maculatus identified by a proteomic and transcriptomic approach. *Insect Mol. Biol.* 26 (1), 58–73. doi:10.1111/imb.12271

Bayram, H., Sayadi, A., Immonen, E., and Arnqvist, G. (2019). Identification of novel ejaculate proteins in a seed beetle and division of labour across male accessory reproductive glands. *Insect biochem. Mol. Biol.* 104, 50–57. doi:10.1016/j.ibmb.2018.12.002

Begun, D. J., Lindfors, H. A., Kern, A. D., and Jones, C. D. (2007). Evidence for de Novo evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* clade. *Genetics* 176 (2), 1131–1137. doi:10.1534/GENETICS.106.069245

Birkhead, T. R., and Pizzari, T. (2002). Postcopulatory sexual selection. *Nat. Rev. Genet.* 3 (44), 262–273. doi:10.1038/nrg774

Blanckenhorn, W. U., Baur, J., Roy, J., Puniamoorthy, N., Busso, J. P., Schafer, M. A., et al. (2021). Comparative sexual selection in field and laboratory in a guild of sepsid dung flies. *Anim. Behav.* 175, 219–230. doi:10.1016/J.ANBEHAV.2021.03.001

Blanckenhorn, W. U., Puniamoorthy, N., Schäfer, M. A., Scheffczyk, A., and Rombke, J. (2013b). Standardized laboratory tests with 21 species of temperate and tropical sepsid flies confirm their suitability as bioassays of pharmaceutical residues (ivermectin) in cattle dung. *Ecotoxicol. Environ. Saf.* 89, 21–28. doi:10.1016/J.ECOENV.2012.10.020

Blanckenhorn, W. U., Puniamoorthy, N., Scheffczyk, A., and Rombke, J. (2013a). Evaluation of eco-toxicological effects of the parasiticide moxidectin in comparison to ivermectin in 11 species of dung flies. *Ecotoxicol. Environ. Saf.* 89, 15–20. doi:10.1016/j.ecoenv.2012.10.030

Boldogkői, Z., Moldovan, N., Szucs, A., and Tombacz, D. (2018). Data descriptor: Transcriptome-wide analysis of a baculovirus using nanopore sequencing. *Sci. Data* 5 (1), 180276–180310. doi:10.1038/sdata.2018.276

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi:10.1093/BIOINFORMATICS/BTU170

Braswell, W. E., Andres, J. A., Maroja, L. S., Harrison, R. G., Howard, D. J., and Swanson, W. J. (2006). Identification and comparative analysis of accessory gland proteins in Orthoptera. *Genome* 49 (9), 1069–1080. doi:10.1139/g06-06110.1139/G06-061

Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* 18 (44), 366–368. doi:10.1038/s41592-021-01101-x

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12 (11), 59–60. doi:10.1038/nmeth.3176

Chen, S., Krinsky, B. H., and Long, M. (2013). New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* 14 (9), 645–660. doi:10.1038/NRG3521

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17 (1), 13–19. doi:10.1186/s13059-016-0881-8

Corchete, L. A., Rojas, E. A., Alonso-Lopez, D., De Las Rivas, J., Gutierrez, N. C., and Burguillo, F. J. (2020). Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci. Rep.* 10 (1), 19737–19815. doi:10.1038/s41598-020-76881-x

Cui, J., Shen, N., Lu, Z., Xu, G., Wang, Y., and Jin, B. (2020). Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome. *Plant Methods* 16 (1), 85–13. doi:10.1186/s13007-020-00629-x

Dapper, A. L., and Wade, M. J. (2020). Relaxed selection and the rapid evolution of reproductive genes. *Trends Genet.* 36 (9), 640–649. doi:10.1016/J.TIG.2020.06.014

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* 34 (15), 2666–2669. doi:10.1093/bioinformatics/bty149

Degner, E. C., Ahmed-Braimah, Y. H., Borziak, K., Wolfner, M. F., Harrington, L. C., and Dorus, S. (2019). Proteins, transcripts, and genetic architecture of seminal fluid and sperm in the mosquito *Aedes aegypti*. *Mol. Cell. Proteomics* 18, S6–S22. doi:10.1074/MCP.RA118.001067

Dong, X., Tian, L., Gouil, Q., Kariyawasam, H., Su, S., De Paoli-Iseppi, R., et al. (2021). The long and the short of it: Unlocking nanopore long-read RNA sequencing data with short-read differential expression analysis tools. *Nar. Genom. Bioinform.* 3 (2), lqab028. doi:10.1093/nargab/lqab028

Findlay, G. D., and Swanson, W. J. (2010). Proteomics enhances evolutionary and functional analysis of reproductive proteins. *BioEssays* 32 (1), 26–36. doi:10.1002/BIES.200900127

Findlay, G. D., Yi, X., Maccoss, M. J., and Swanson, W. J. (2008). Proteomics reveals novel Drosophila seminal fluid proteins transferred at mating. *PLoS Biol.* 6 (7), e178. doi:10.1371/JOURNAL.PBIO.0060178

Freedman, A. H., Clamp, M., and Sackton, T. B. (2021). Error, noise and bias in de novo transcriptome assemblies. *Mol. Ecol. Resour.* 21 (1), 18–29. doi:10.1111/1755-0998.13156

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28 (23), 3150–3152. doi:10.1093/BIOINFORMATICS/BTS565

Fuessl, M., Santos, C. G., Hartfelder, K., Schrempf, A., and Heinze, J. (2018). Accessory gland proteins of males in the male-diphenic ant Cardiocondyla obscurior. *Physiol. Entomol.* 43 (4), 276–284. doi:10.1111/PHEN.12257

Garlovsky, M. D., Evans, C., Rosenow, M. A., Karr, T. L., and Snook, R. R. (2020). Seminal fluid protein divergence among populations exhibiting postmating prezygotic reproductive isolation. *Mol. Ecol.* 29 (22), 4428–4441. doi:10.1111/MEC.15636

Gasparini, C., Pilastro, A., and Evans, J. P. (2020). The role of female reproductive fluid in sperm competition. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375 (1813), 20200077. doi:10.1098/RSTB.2020.0077

Geadkaew, A., Kosa, N., Siricoon, S., Grams, S. V., and Grams, R. (2014). A 170 kDa multi-domain cystatin of Fasciola gigantica is active in the male reproductive system. *Mol. Biochem. Parasitol.* 196 (2), 100–107. doi:10.1016/J.MOLBIOPARA.2014.08.004

Giesen, A., Blanckenhorn, W. U., and Schäfer, M. A. (2017). Behavioural mechanisms of reproductive isolation between two hybridizing dung fly species. *Anim. Behav.* 132, 155–166. doi:10.1016/j.anbehav.2017.08.008

Giesen, A., Schäfer, M. A., and Blanckenhorn, W. U. (2019). Geographic patterns of postzygotic isolation between two closely related widespread dung fly species (Sepsis cynipsea and Sepsis neocynipsea; Diptera: Sepsidae). *J. Zool. Syst. Evol. Res.* 57 (1), 80–90. doi:10.1111/JZS.12239

Goenaga, J., Yamane, T., Ronn, J., and Arnqvist, G. (2015). Within-species divergence in the seminal fluid proteome and its effect on male and female reproduction in a beetle. *BMC Evol. Biol.* 15, 266. doi:10.1186/s12862-015-0547-2

Gorshkov, V., Blenau, W., Koeniger, G., Rompp, A., Vilcinskas, A., and Spengler, B. (2015). Protein and peptide composition of male accessory glands of *Apis mellifera* drones Investigated by mass spectrometry. *PLoS One* 10 (5), e0125068. doi:10.1371/journal.pone.0125068

Gotoh, A., Yamaguchi, K., KobayaShi, S., ItoF., and Tsuji, K. (2018). Transcriptome characterization of male accessory glands in ants to identify molecules involved in their reproductive success. *Insect Mol. Biol.* 27 (2), 212–220. doi:10.1111/IMB.12364

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 2929 (77), 644–652. doi:10.1038/nbt.1883

Halstead, M. M., Islas-Trejo, A., Goszczynski, D. E., Medrano, J. F., Zhou, H., and Ross, P. J. (2021). Large-scale multiplexing permits full-length transcriptome annotation of 32 bovine tissues from a single nanopore flow cell. *Front. Genet.* 12, 664260. doi:10.3389/fgene.2021.664260

Hargreaves, A. D., and Mulley, J. F. (2015). Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. *PeerJ* 2015 (11), e1441. doi:10.7717/peerj.1441

Herath, B., Dochtermann, N. A., Johnson, J. I., Leonard, Z., and Bowsher, J. H. (2015). Selection on bristle length has the ability to drive the evolution of male abdominal appendages in the sepsid fly Themira biloba. *J. Evol. Biol.* 28 (12), 2308–2317. doi:10.1111/JEB.12755

Hölzer, M., and Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-seq assemblers. *GigaScience* 8 (5), giz039–16. doi:10.1093/gigascience/giz039

Hurtado, J., Almeida, F. C., Belliard, S. A., Revale, S., and Hasson, E. (2021). Research gaps and new insights in the evolution of Drosophila seminal fluid proteins. *Insect Mol. Biol.* 31, 139–158. doi:10.1111/IMB.12746

Jansa, S. A., Lundrigan, B. L., and Tucker, P. K. (2003). Tests for positive selection on immune and reproductive genes in closely related species of the murine genus Mus. *J. Mol. Evol.* 56 (33), 294–307. doi:10.1007/S00239-002-2401-6

LaFlamme, B. A., Ravi Ram, K., and Wolfner, M. F. (2012). The *Drosophila melanogaster* seminal fluid protease "seminase" regulates proteolytic and post-mating reproductive processes. *PLoS Genet.* 8 (1), 1002435. doi:10.1371/JOURNAL.PGEN.1002435

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A., and Begun, D. J. (2006). Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. U. S. A.* 103 (26), 9935–9939. doi:10.1073/PNAS.0509809103

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25 (16), 2078–2079. doi:10.1093/BIOINFORMATICS/BTP352

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinforma. Oxf. Engl.* 34 (18), 3094–3100. doi:10.1093/BIOINFORMATICS/BTY191

Li, W., and Godzik, A. (2006). Cd-Hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22 (13), 1658–1659. doi:10.1093/bioinformatics/btl158

Lin, J., Guan, L., Ge, L., Liu, G., Bai, Y., and Liu, X. (2021). Nanopore-based full-length transcriptome sequencing of Muscovy duck (*Cairina moschata*) ovary. *Poult. Sci.* 100 (8), 101246. doi:10.1016/j.psj.2021.101246

Lupold, S., Manier, M. K., Puniamoorthy, N., Schoff, C., Starmer, W. T., Luepold, S. H. B., et al. (2016). How sexual selection can drive the evolution of costly sperm ornamentation. *Nature* 533 (7604), 7604533535–7604533538. doi:10.1038/nature18005

Mancini, E., Baldini, F., Tammaro, F., Calzetta, M., Serrao, A., George, P., et al. (2011). Molecular characterization and evolution of a gene family encoding male-specific reproductive proteins in the African malaria vector *Anopheles gambiae*. *BMC Evol. Biol.* 11 (1), 292. doi:10.1186/1471-2148-11-292

Mrinalini, Koh, C. Y., and Puniamoorthy, N. (2021). Rapid genomic evolution drives the diversification of male reproductive genes in dung beetles. *Genome Biol. Evol.* 13 (8), evab172. doi:10.1093/gbe/evab172

Mueller, J. L., Linklater, J. R., Ravi Ram, K., Chapman, T., and Wolfner, M. F. (2008). Targeted gene deletion and phenotypic analysis of the *Drosophila melanogaster* seminal fluid protease inhibitor Acp62F. *Genetics* 178 (3), 1605–1614. doi:10.1534/GENETICS.107.083766

Mueller, J. L., Page, J. L., and Wolfner, M. F. (2007). An ectopic expression screen reveals the protective and toxic effects of Drosophila seminal fluid proteins. *Genetics* 175 (2), 777–783. doi:10.1534/GENETICS.106.065318

Mueller, J. L., Ravi Ram, K., McGraw, L. A., Bloch Qazi, M. C., Siggia, E. D., Clark, A. G., et al. (2005). Cross-species comparison of Drosophila male accessory gland protein genes. *Genetics* 171 (1), 131–143. doi:10.1534/GENETICS.105.043844

Park, M., and Wolfner, M. F. (1995). Male and female cooperate in the prohormone-like processing of a *Drosophila melanogaster* seminal fluid protein. *Dev. Biol.* 171 (2), 694–702. Available at: https://reader.elsevier.com/reader/sd/pii/S0012160685713152?token=2DF93285F003B3C4AF390E23FA28D99E8682C41A0FD82C491189EE1BEF9FAE12ABB6B01074EE3B3230F14257E00A2A1A&originRegion=eu-west-1&originCreation=20210915184341 (Accessed on September 16, 2021). doi:10.1006/dbio.1995.1315

Parthasarathy, R., Sun, Z., Chen, Z., RankinM., and Palli, S. R. (2009). Juvenile hormone regulation of male accessory gland activity in the red flour beetle, *Tribolium castaneum*. *Mech. Dev.* 126 (7), 563–579. doi:10.1016/j.mod.2009.03.005

Patlar, B., Jayaswal, V., Ranz, J. M., and Civetta, A. (2021). Nonadaptive molecular evolution of seminal fluid proteins in Drosophila. *Evolution* 75 (8), 2102–2113. doi:10.1111/EVO.14297

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14 (4), 417–419. doi:10.1038/nmeth.4197

Peferoen, M., and De loof, A. (1984). Intraglandular and extraglandular synthesis of proteins secreted by the accessory reproductive glands of the Colorado potato beetle, *Leptinotarsa decemlineata*. *Insect Biochem.* 14 (4), 407–416. doi:10.1016/0020-1790(84)90097-0

Puniamoorthy, N., Blanckenhorn, W. U., and Schäfer, M. A. (2012a). Differential investment in pre- vs. post-copulatory sexual selection reinforces a cross-continental reversal of sexual size dimorphism in Sepsis punctum (Diptera: Sepsidae). *J. Evol. Biol.* 25 (11), 2253–2263. doi:10.1111/j.1420-9101.2012.02605.x

Puniamoorthy, N., Ismail, M. R. B., Tan, D. S. H., and MeieR, R. (2009). From kissing to belly stridulation: Comparative analysis reveals surprising diversity, rapid evolution, and much homoplasy in the mating behaviour of 27 species of sepsid flies (Diptera: Sepsidae). *J. Evol. Biol.* 22 (11), 2146–2156. doi:10.1111/J.1420-9101.2009.01826.X

Puniamoorthy, N., Schäfer, M. A., and Blanckenhorn, W. U. (2012b). Sexual selection accounts for the geographic reversal of sexual size dimorphism in the dung fly, sepsis punctum (diptera: Sepsidae). *Evolution* 66 (7), 2117–2126. doi:10.1111/J.1558-5646.2012.01599.X

Puniamoorthy, N., Su, K. F.-Y., and Meier, R. (2008). Bending for love: Losses and gains of sexual dimorphisms are strictly correlated with changes in the mounting position of sepsid flies (sepsidae: Diptera). *BMC Evol. Biol.* 8 (1), 155. doi:10.1186/1471-2148-8-155

Reinhardt, J. A., Wanjiru, B. M., Brant, A. T., Saelao, P., Begun, D. J., and Jones, C. D. (2013). De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet.* 9 (10), e1003860. doi:10.1371/JOURNAL.PGEN.1003860

Rivard, E. L., Ludwig, A. G., Patel, P. H., Grandchamp, A., Arnold, S. E., Berger, A., et al. (2021). A putative de novo evolved gene required for spermatid chromatin condensation in *Drosophila melanogaster*. *PLoS Genet.* 17 (9), e1009787. doi:10.1371/JOURNAL.PGEN.1009787

Rohner, P. T., Blanckenhorn, W. U., and Puniamoorthy, N. (2016). Sexual selection on male size drives the evolution of male-biased sexual size dimorphism via the prolongation of male development. *Evolution* 70 (6), 1189–1199. doi:10.1111/EVO.12944

Rowe, M., Whittington, E., Borziak, K., Ravinet, M., Eroukhmanoff, F., Sætre, G. P., et al. (2020). Molecular diversification of the seminal fluid proteome in a recently diverged passerine species pair. *Mol. Biol. Evol.* 37 (2), 488–506. doi:10.1093/MOLBEV/MSZ235

Sahlin, K., and Medvedev, P. (2019). De novo clustering of long-read transcriptome data using a greedy, quality-value based algorithm. *Lect. Notes Comput. Sci.* 11467, 227–242. doi:10.1007/978-3-030-17083-7_14

Sahlin, K., and Medvedev, P. (2021). Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nat. Commun.* 12 (1), 2. doi:10.1038/s41467-020-20340-8

Saraswathi, S., Chaitra, B. S., Tannavi, K., Mamtha, R., SowRabha, R., Rao, K. V., et al. (2020). Proteome analysis of male accessory gland secretions in Leucinodes orbonalis Guenee (Lepidoptera: Crambidae), a Solanum melongena L. pest. *Arch. Insect Biochem. Physiol.* 104 (3), e21672. doi:10.1002/arch.21672

Sayadi, A., Immonen, E., Bayram, H., and Arnqvist, G. (2016). The de novo transcriptome and its functional annotation in the seed beetle Callosobruchus maculatus. *PLoS One* 11 (7), e0158565. doi:10.1371/journal.pone.0158565

Sessegolo, C., Cruaud, C., Da Silva, C., Cologne, A., Dubarry, M., Derrien, T., et al. (2019). Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules. *Sci. Rep.* 9 (1), 14908. doi:10.1038/s41598-019-51470-9

Simmons, L. W., and Wedell, N. (2020). Fifty years of sperm competition: The structure of a scientific revolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375 (1813), 20200060. doi:10.1098/RSTB.2020.0060

Sirot, L. K., Findlay, G. D., Sitnik, J. L., Frasheri, D., Avila, F. W., and Wolfner, M. F. (2014). Molecular characterization and evolution of a gene family encoding both female- and male-specific reproductive proteins in Drosophila. *Mol. Biol. Evol.* 31 (6), 1554–1567. doi:10.1093/MOLBEV/MSU114

Sirot, L. K. (2019). On the evolutionary origins of insect seminal fluid proteins. *Gen. Comp. Endocrinol.* 278, 104–111. doi:10.1016/J.YGCEN.2019.01.011

Sonenshine, D. E., Bissinger, B. W., Egekwu, N., Donohue, K. V., Khalil, S. M., and Roe, R. M. (2011). First transcriptome of the testis-vas deferens-male accessory gland and proteome of the spermatophore from dermacentor variabilis (Acari: Ixodidae). *PLOS ONE* 6 (9), e24711. doi:10.1371/JOURNAL.PONE.0024711

Soneson, C., Yao, Y., Bratus-Neuenschwander, A., Patrignani, A., Robinson, M. D., and Hussain, S. (2019). A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* 10 (1), 3359–3414. doi:10.1038/s41467-019-11272-z

Steijger, T., Abril, J. F., Engstrom, P. G., Kokocinski, F., Hubbard, T. J., et al. RGASP Consortium (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10 (12), 1177–1184. doi:10.1038/nmeth.2714

Swanson, W. J., Clark, A. G., Waldrip-Dail, H. M., Wolfner, M. F., and Aquadro, C. F. (2001). Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* 98 (13), 7375–7379. doi:10.1073/pnas.131568198

Swanson, W. J., and Vacquier, V. D. (2002). The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* 3 (232), 137–144. doi:10.1038/nrg733

Teng, L., Fan, X., Xu, D., Zhang, X., Mock, T., and Ye, N. (2017). Identification of genes under positive selection reveals differences in evolutionary adaptation between brown-algal species. *Front. Plant Sci.* 8, 1429. doi:10.3389/fpls.2017.01429

Tian, C.-B., Wei, D., Xiao, L. F., Dou, W., Liu, H., and Wang, J. J. (2017). Comparative transcriptome analysis of three*Bactrocera dorsalis*(Diptera: Tephritidae) organs to identify functional genes in the male accessory glands and ejaculatory duct. *Fla. Entomol.* 100 (1), 42–51. doi:10.1653/024.100.010810.1653/024.100.0108

Torgerson, D. G., Kulathinal, R. J., and Singh, R. S. (2002). Mammalian sperm proteins are rapidly evolving: Evidence of positive selection in functionally diverse genes. *Mol. Biol. Evol.* 19 (11), 1973–1980. doi:10.1093/OXFORDJOURNALS.MOLBEV.A004021

Vedelek, V., Bodai, L., Grezal, G., Kovacs, B., Boros, I. M., Laurinyecz, B., et al. (2018). Analysis of *Drosophila melanogaster* testis transcriptome. *BMC Genomics* 19 (1), 697. doi:10.1186/s12864-018-5085-z

Vibranovski, M. D., Lopes, H. F., Karr, T. L., and Long, M. (2009). Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. *PLoS Genet.* 5 (11), e1000731. doi:10.1371/journal.pgen.1000731

Wagstaff, B. J., and Begun, D. J. (2007). Adaptive evolution of recently duplicated accessory gland protein genes in desert Drosophila. *Genetics* 177 (2), 1023–1030. doi:10.1534/GENETICS.107.077503

Wang, F., Chen, Z., Pei, H., Guo, Z., Wen, D., Liu, R., et al. (2021). Transcriptome profiling analysis of tea plant (Camellia sinensis) using Oxford Nanopore long-read RNA-Seq technology. *Gene* 769, 145247. doi:10.1016/j.gene.2020.145247

Weber, A. A. T., Abi-Rached, L., GaltierN., BernArd, A., Montoya-Burgos, J. I., and Chenuil, A. (2017). Positive selection on sperm ion channels in a brooding brittle star: Consequence of life-history traits evolution. *Mol. Ecol.* 26 (14), 3744–3759. doi:10.1111/MEC.14024

Wei, D., Li, H. M., Tian, C. B., Smagghe, G., Jia, F. X., Jiang, H. B., et al. (2015). Proteome analysis of male accessory gland secretions in oriental fruit flies reveals juvenile hormone-binding protein, suggesting impact on female reproduction. *Sci. Rep.* 5 (5), 16845. doi:10.1038/srep16845

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X. J., et al. (2017). Comprehensive comparison of pacific Biosciences and Oxford nanopore technologies and their applications to transcriptome analysis. *F1000Res.* 6 (6), 100. doi:10.12688/f1000research.10571.1

Wigby, S., Brown, N. C., Allen, S. E., Misra, S., Sitnik, J. L., Sepil, I., et al. (2020). The Drosophila seminal proteome and its role in postcopulatory sexual selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375 (1813), 20200072. doi:10.1098/RSTB.2020.0072

Wong, A., Turchin, M. C., Wolfner, M. F., and Aquadro, C. F. (2008). Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. *Mol. Biol. Evol.* 25 (3), 497–506. doi:10.1093/MOLBEV/MSM270

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., et al. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* 16 (12), 1297–1305. doi:10.1038/s41592-019-0617-2

Yamane, T., Goenaga, J., Ronn, J. L., and Arnqvist, G. (2015). Male seminal fluid substances affect sperm competition success and female reproductive behavior in a seed beetle. *PLoS One* 10 (4), e0123770. doi:10.1371/journal.pone.0123770

Zhao, P., Zhou, X. m., Zou, J., Wang, W., Wang, L., Peng, X. b., et al. (2014). Comprehensive analysis of cystatin family genes suggests their putative functions in sexual reproduction, embryogenesis, and seed formation. *J. Exp. Bot.* 65 (17), 5093–5107. doi:10.1093/JXB/ERU274