Optimization of the *de novo* assembly of the transcriptome of the venom gland of *Pamphobeteus verdolaga*, prospecting novel bioactive peptides

Cristian Salinas-Restrepo
Universidad de Antioquia
Elizabeth Misas
Corporación para Investigaciones Biológicas: Corporacion para Investigaciones Biologicas
Sebastian Estrada-Gomez
Universidad de Antioquia
Juan Quintana
Universidad Cooperativa de Colombia: Universidad Cooperativa de Colombia
Fanny Guzman
Pontifical Catholic University of Valparaíso: Pontificia Universidad Catolica de Valparaiso
Juan Camilo Calderon
Universidad de Antioquia
Marco Antonio Giraldo
Universidad de Antioquia
Cesar Hernando Segura (cesar.segura@udea.edu.co)
Universidad de Antioquia https://orcid.org/0000-0002-9979-8416

Research article

Keywords: Spider, tarantula, transcriptomics, theraphosid, peptide prospection, venom gland, toxins

Posted Date: November 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-99776/v1

Abstract

Background: Spiders are among the most venomous animals in nature. Their venom constitutes a source of novel and innovative peptides and proteins with medicinal and biotechnological interest. However, their potential as antimicrobial, anti-cancerous, anti-hypertensive and even in the modulation of nociception is under-studied, mainly because handling the venom is technically challenging and there is paucity of next-generation-sequencing (NGS) data. Due to the increasing evidence of underestimation of the number of genes by the use of a single transcriptome assembler, we re-assembled and optimized the *de novo* transcriptome of the venom gland of the recently described Colombian spider *P. verdolaga*, by using three free access algorithms: Trinity, Soapdenovo and SPAdes. All the assemblies were evaluated by statistical parameters (e.g. contigs, GC%, max and min length and N50), by applying BUSCO´s terms retrieval against the arthropod data set to determine the best assembly for each software.

Results: Our analyses show that while approximately 54% of all the assembled and structurally annotated sequences could be found in all three algorithms, around 23% of these were unique for Trinity and 21% were unique for SPAdes. The non-redundant merge of all three assemblies' output permitted the annotation of 8640 sequences; at least 15% more when compared to each software separately, and an increase of 20% when compared to a previous *P. verdolaga* assembly. Analysis of the annotated genes allowed the identification of unreported lectins, kinins and over 200 peptides and proteins with potential antimicrobial and protease inhibition activities. Furthermore, homology search against the Arachnoserver database and the EROP knowledgebase allowed the identification of 135 novel theraphotoxins of biotechnological interest.

Conclusion: Transcriptomic data is of utmost importance for spiders, as it is one of the more feasible and scalable ways to characterize these organisms. However, the use of a single *de novo* assembler implies an under representation of the expressed sequences, as it has been shown here. In the generation of data for non-model organisms as well as in the search for novel peptides and proteins with biotechnological interest, it is highly recommended that at least two different assemblers are employed.

Background

Spiders evolved from the rest of the arachnids around 300 million years ago during the carboniferous period; since then, they have become the most specious venomous animal, with more than 48,000 species reported (1). Their highly complex venom, which typically contains up to 200 different peptides, and therefore over 9.6 million potential bioactive molecules for the whole order, is comprised of low (< 1 kDa; acylpolyamines and salts), medium (1–10 kDa; neurotoxic disulfide bridged, or DBPs, alpha helicoidal cytolytic and antimicrobial peptides, or NDBPs) and high molecular weight (> 10 kDa; proteins and enzymes) peptides and proteins (2, 3). The most abundant components in the venom: the neurotoxic and cytolytic/antimicrobial peptides, show much promise in the development of new tools for research, insecticidal products and therapeutic applications due to interesting activities, such as modulation of K^+, Ca^{2+} and Cl^- ion channels and anti-inflammatory and anti-cancer properties (4–7). Nonetheless, one of

the most interesting activities is that of antimicrobial peptides, especially because of the phenomena of antimicrobial resistance (8, 9).

The secondary structure of DBPs is predominantly formed by β-sheets that are stabilized by 5–10 disulfide bridges, known as the Inhibitory Cysteine Motif (ICK) (10). Inside the ICK, two other motifs can be identified: The Primary Structural Motif (PSM) and the Extra Structural Motif (ESM), which are characterized by bridges in the form C1 × 6 C2 …. C3 C4 and C5 X C6 … C7 X C8, respectively. The presence of the PSM is sufficient to assume that the peptide has neurotoxical activity as it binds ion channels (4), which may give it pain-relief properties. Peptides like Huentoxin-XVI (*Ornithoctonus huwena*) and ProTx-II (*Thrixopelma pruriens*) interrupt nociception in various models, while peptides like Ph1β and PnPP-19 from the venom of *Phoneutria nigriventer* are around 180 times more potent that ziconotide (11–13). Moreover, DBPs have potential in other pathologies like ischemic stroke, as shown by the peptide PcTx1 (*Psalmopoeus cambridgei*), in epilepsy, as shown by the peptide Hm3a (*Heteroscrodra maculate*), and in neurodegenerative diseases, as shown by the peptide PhKv (*Phoneutria nigriventer*) in Alzheimer's and Guangxitoxin-1E in Parkinson's models (12). As a notable exception for a DBP peptide, Gomesin from *Acanthoscurria gomesiana* shows activity against gram positive and gram negative bacteria, fungi, yeast, malaria and leishmanial, while retaining anti cancerous activity in a murine melanoma model (14).

On the other hand, the peptides termed NDBPs or Non-Disulfide Bridged Peptides, present a predominant -helix secondary structure with an abundance of cationic amino acids, such as lysine and arginine, and hydrophobic amino acids such as leucine, isoleucine and valine. Being cationic and amphipathic, these peptides tend to interact with the partially anionic phosphatidyl glycerol (PG) and the cardiolipin enriched bacterial membranes. Interestingly, transformed tumor cells, also rich in PG, are targeted by NDBPs, giving the latter antitumoral and antimicrobial properties (7, 15). There are four major families of NDBPs described in spiders: Lycotoxins, Latarcins, Cupeinnins and Oxyopinins (15). Most of these toxins display antimicrobial activities in the micro molar range. However, some toxins display activity against microorganisms of clinical interest: the peptide Lycotoxin I from spiders of the genera *Lycosa* displays activity against Methicillin Resistant *Staphylococcus aureus* (MRSA), and the peptide CIT1a from *Lachesana tarabaevi* displays activity against the yeast *Chlamydia trachomatis* (16). Furthermore, the peptide Lycosin-I (from *Lycosa singnoriensis*) exhibits activity against fungi of the genera *Penicillium* and *Aspegillus*, and is also able to induce the apoptosis of prostate cancer cells while the peptide Latarcin-3a from *Lachesana tarabaevi* is able to induce pores in the envelope of the HIV virus (12).

This broad display of activities shows the enormous biotechnological potential of spider venoms, yet, less than 1% of the hypothetical potential novel molecules has been reported in the literature (17). Proteomic studies have been quite limited, regarding the amount of data that they can provide, due to the limited amounts of venom recovered after the milking process (18). This fact makes transcriptomics a preferable tool for the prospection of bioactive molecules from arachnids (19, 20). Next Generation Sequencing (NGS) data from tissues, such as the venom gland, enable the prediction of antimicrobial peptides by *i)* sequence homology with previously reported primary sequences from antimicrobials or *ii)*

through the identification of known motifs that are key to the activities of interest. This can be done through a bioinformatics approach, applying tools like BLAST and HMMER (19, 21–23). However, in non-model organisms, such as the members from the *Araneae* order, there is a lack of reported nucleotide information at the genome and transcriptome level. Thereby, the assembly and annotation of the transcriptome becomes an important endeavor to guarantee that most of the nucleotide information is retrieved. Rarely, a single assembly method permits the retrieval of all genes and sequences within a transcriptome (24). In venoms, their inherit complexity and the presence of highly homologous and paralogous sequences might distort the quantity and quality of the information recovered. Therefore, the usage of various assemblers may be recommended as taxon specific as well as tissue specific biases and issues may arise, especially in these non-model organisms (24).

In this study, we optimized the assembly of the transcriptome of the venom gland of the Colombian spider *Pamphobeteus verdolaga* (*Araneae:Theraphosidae*). In order to maximize the amount of information obtained from the transcriptome, we assembled the reads with three different *de novo* assembly algorithms and assessed their performance through “classical assembly statistics” as well as BUSCO terms retrieval, and compared the number of genes annotated when only one assembly output was used and when a merge of the three assembler results were used as input for the annotation with OmicsBox. The file with the merged predicted genes from the three assemblers showed an increase of at least 15% in the number of annotated genes than in each of the individual assemblies. The improvement in annotation allowed the identification of 135 novel toxins that may hold biotechnological potential, unveiling several of new bioactive peptides with potential interest in biomedicine.

Methods

Venom gland transcriptome data

The transcriptome’s raw reads from the venom gland of *Pamphobeteus verdolaga* were obtained from the European Nucleotide Archive (ENA) under accessions numbers: PRJEB21288/ERS1788422/ERX2067777-ERR2008012. As explained by Estrada and co-workers (25), two female specimens were collected from the province of Antioquia, Colombia under the contract 155 signed by the University of Antioquia and the Environmental Ministry of Colombia, and the venom glands were extirpated. The total RNA was obtained through TRIzol® reagent (ThermoFisher Scientific, MA, USA) while the purification process of mRNA and the library creation was carried out with the Illumina mRNA TruSeq kit v2, as indicated by the manufacturer. The library 100 bp pair-ended reads was sequenced in an Illumina Hiseq 2500 instrument (illumina Inc, San Diego USA).

Transcriptome assembly

The raw reads obtained from the Illumina platform were subjected to a process of assembly optimization. Firstly, low quality reads and possible adaptor sequences were eliminated by using the programs TrimGalore v0.6.3 as well Trimmomatic v0.38, with default options. The cleaned reads were

subject to an assembly with Trinity v2.1.1 using two k-mer lengths: 25 and 32, thus, four different transcriptome assemblies were obtained. Each assemblies' statistics were obtained by using the TrinityStats.pl script. The reads edited with TrimGalore were also assembled with SOAPdenovo v2.04 and with SPAdes v3.13.1, with k-mer lengths 31 and 63 for a total of eight different transcriptome assemblies. The statistics from these assemblies were obtained with the QUAST v5.0.2 software. For Trinity, SOAPdenovo and SPAdes, only the k-mer length was varied.

Assessing transcriptome completion with BUSCO

A first measure of the completeness of the transcriptomes was evaluated using the Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.1.2 software. For this, each of the assemblies was analyzed against BUSCO's own arthropoda dataset. The data from the number of complete, duplicated, fragmented and missing BUSCO terms was extracted. In the case of the assemblies obtained from SOAPdenovo and SPAdes, a separation of transcripts with lengths higher and lower than 200 bp was done with the software SeqKit v0.12.0 before BUSCO.

Sequence coverage relative to the genome of *Parasteatoda tepitadorium*

The assembled contigs were aligned to the genome of the common house spider: *Parasteatoda tepitadorium* in order to assess the level of coverage. The alignment was done with the software minimap2 v2.17, using the splice setting due to the large size of the contigs. The resulting alignment coverage and its abundance was graphed with the software GraphPad Prism v7.00 for MacOS (GraphPad Software, La Jolla California, USA).

Structural and functional annotation of the transcriptome assemblies

The transcriptome assemblies from Trinity, SPAdes and Soapdenovo that showed the best performance evaluated through classical statistics, BUSCO terms retrieval and sequence coverage, were annotated with the software Augustus v 3.3.3 for the prediction of ORFs and with the software OmicsBox v 1.2.4 (BioBam Bioinformatics S.L., Valencia, Spain) (previously known as Blast2GO) for the functional annotation. Briefly, the structural annotation with Augustus used the genome of *P. tepitadorium* as a template, retrieving information such as intron length and possible splicing signals in order to predict the possible ORFs from each of the assemblies. Only complete ORFs were reported. The functional annotation was carried out with the software OmicsBox using the UniProtKB database for homology search as well as the InterPro database for GO terms retrieval. Only those sequences with hits from both databases were considered as annotated. For the redundancy analysis the software CD-HIT v4.8.1 was employed with a cutoff of 85% homology.

Prediction of proteins and enzymes

All annotated sequences' GO terms were scanned for the keywords: toxin, antimicrobial, nociception, antiparasitic, antiarrhythmic, cytolytic, hemolytic, hyaluronidase, inflammatory, kinin, lectin, necrosis, neurotoxin, neurotransmitter hydrolysis, presynaptic neurotoxin, protease and protease inhibitor. All

resulting sequences were extracted and manually curated using PubMed to check whether there was experimental evidence for their activity or were just related to an associated process. A further prospection was carried using the over 1800 peptides from the Arachnoserver database (http://www.arachnoserver.org) and the over 8000 peptides from the Erop knowledge base (http://erop.inbi.ras.ru). Homology to the assembled transcriptome was checked using local BLASTP v2.10.0. All of the retrieved sequences with a e-value higher than $1x10^{-3}$ were eliminated. If a gene had multiple hits, only the one with the smallest e-value and highest identity percentage was taken as the best and probable result. The associated mature toxin was identified by elimination of the possible signal peptide and pro-peptide using the SpiderP software (http://www.arachnoserver.org/spiderP.html). A global alignment of the identified mature toxin and the identified homologue from BLAST was carried through ClustalW (https://www.genome.jp/tools-bin/clustalw). Those sequence pairs with global identity smaller than 20% were discarded as possible homologues.

Results

Changing the assembly software brings more variability than changes in the pre-processing of reads.

The 24 million, 101 pb pair-ended raw sequences of the venom gland of *P. verdolaga* were first assembled using the most popular de novo assembly software: Trinity; using two pre-processing software: Trimmomatic and TrimGalore, and two different k-mer lengths: k-mer 25 and k-mer 32 (Fig. 1a). This allowed to choose the best pre-processing method and the best Trinity assembly. Next, the reads were assembled with the SPAdes and Soapdenovo software using the chosen pre-processing method and two k-mer lengths: k-mer 31 for comparison with Trinity and k-mer 63, due to the ability to assemble at higher lengths of both software (Fig. 1b).

In the first part of the optimization, neither changing the pre-processing method nor the k-mer number, changed the number of transcripts, %GC, N50 or the median and average length of the assembled transcripts by more than 3% (Table 1 A). BUSCOs, which are a means to assess the completeness of an assembly in terms of the gene content, provide a complementary view to statistics such as the N50. Each of the four assemblies was evaluated against the arthropod data set from BUSCO in order to determine their completeness. Again, a variation of less than 2% was observed among all of the presets. On average, 78% of the arthropod dataset genes were identified completely; however, on average 23% of the retrieved terms were duplicated (Table 1 B). Finally, all of the mappings showed poor quality but, unlike the previous statistics, the assembly with TrimGalore and k-mer 25 showed higher average length of the mapped reads as well as higher coverage (Table 1 C). A histogram of the alignment lengths shows that the highest variability is found in contigs with length between 100,000 to 200,000 bp (Figure 2). Whilst the differences between the presets cannot be taken as significant, the reads pre-processed with TrimGalore and assembled with a k-mer of 25, have a better balance between the number of transcripts, the N50 value, the number of completed and missing BUSCOs, the average length and total coverage of the mapped reads, when compared to the other evaluated presets.

In the second part of the optimization, all of the stats were analyzed with Quast instead of TrinityStats.pl due to compatibility issues with the outputs from SPAdes and Soapdenovo. Unlike the first part, changes in the assembly algorithm added more variance than changes in the k-mer length or the pre-processing method. Around 29-42% and 74%-95% of the assembled contigs from SPAdes and Soapdenovo respectively, have less than 200 bp. Removal of these sequences shows that SPAdes has a higher average N50 than Trinity, with a smaller number of total contigs. Soapdenovo has the smallest N50 of the three and the number of total contigs assembled is almost 30% smaller than that of Trinity (Table 2 A). The BUSCO analysis was carried out for sequences higher and smaller than 200 bp. No complete BUSCOs can be found in sequences smaller than 200 bp; however 3% and 6-16% of fragmented genes can be found in SPAdes and Soapdenovo k-mer 31/k-mer 63, respectively .The number of complete genes identified in sequences larger than 200 bp are close to that of Trinity for SPAdes (76% on average), and around half of the BUSCO dataset can be identified on the Soapdenovo assembly. Surprinsignly, only 8% and 2% of the retrieved BUSCO terms for SPAdes and Soapdenovo, respectively, are duplicated, which is four to ten times smaller than the duplicated average number found in Trinity (Table 2 B). The alignment of the reads to the genome of *P. tepitadorium* shows that, in general, the longest average length and coverage is presented by Trinity, followed by SPAdes and Soapdenovo. The total coverage is about 40% smaller on SPAdes and 82% smaller on Soapdenovo when compared to Trinity. Even with such small coverage, Soapdenovo k-mer 31 has the best alignment quality in all of the eight coverage analysis, followed by SPAdes k-mer 31 (Table 2 C). The distribution of the aligned sequences is similar between SPAdes and Trinity, while Soapdenovo greatly differs in the 50,000 to 100,000 bp range. Noticeably, both k-mer 63 assemblies show more sequences with length smaller than 10,000 bp when compared to Trinity; which is approximately 75% of all the contigs (Figure 3). Again, although not drastic, the differences between the presets show a tendency for the reads assembled with SPAdes k-mer 31, and Soapdenovo k-mer of 63 to have a better balance between the numbers of transcripts, the N50, the number of completed and missing BUSCOs and the coverage analysis. Even when the Soapdenovo k-mer 31 assembly to the *P. tepitadorium* genome has a higher average length and quality, the total coverage favors the Soapdenovo k-mer 63 alignment. . From the above results, it is concluded that the reads cleaned with TrimGalore and assembled with Trinity with k-mer 25, SPAdes with k-mer 32 and Soapdenovo with k-mer 63, showed the best performance.

The merging of the assembled transcriptomes increases the annotation performance

The assembled transcriptomes that showed the best overall performance were annotated. The output from Augustus shows that the biggest number of predicted ORFs comes from Trinity, followed by Spades and Soapdenovo. The average and maximum ORF length followed the same pattern. Due to the high number of duplicated genes observed in the previous BUSCO analysis, all of the genes with an identity higher than 85% within the Augustus output were eliminated with CD-HIT. Around 25% of all of the ORFs in the Trinity assembly were duplicated, while this number was only 12% for the SPAdes assembly and, surprisingly, 0.1% for Soapdenovo (Table 3).

The described results leave SPAdes with the highest number of non-redundant ORFs, followed by Trinity and Soapdenovo. The distribution of the non-redundant sequences shows that SPAdes has the highest percentage (76%) of sequences smaller than 200 amino acids, followed by Soapdenovo (71%) and Trinity (69%). Trinity and Soapdenovo have the same number of sequences, between 200 and 1000 amino acids (28%), followed by SPAdes (22%) (Figure 4). In order to increase the total number of sequences annotated, a merge of all of the predicted ORFs was created. The non-redundant merge file is composed of approximately 23% sequences unique to Trinity, 21% sequences unique to SPAdes and 1% sequences unique to Soapdenovo; the other 54% of the sequences are shared between all of the three assemblies. This implies an increase in the number of ORFs by 31% when compared to Trinity, 32% when compared to SPAdes and 96% when compared to Soapdenovo.

The predicted ORFs were annotated with help of the UniProtKB database (Swiss-prot) and the InterproScan database in the software OmicsBox, with default settings. Just like in all of the previous analysis, the biggest number of annotated ORFs came from the Trinity assembly, with a total of 7478 sequences (43% without annotation) (Figure 5 a), followed by SPAdes with 7019 sequences (49% without annotation) (Figure 5 b) and by Soapdenovo with 5134 sequences (Only 39% without annotation) (Figure 5 c). The non-redundant version (homology smaller than 85% in CD-HIT) had a total of 18,322 sequences from which 8640 were annotated (52% without annotation) (Figure 5 d). The merge increased the total number of annotated sequences when compared to Trinity, by 15%, and to Spades, by 23%.

The fourth level GO terms show similarity between all of the components found in the four annotations; however, there is an increase in the merge in GO terms for sequences associated with the metabolism of nitrogen (Supl. Figures). The merge also implied an increase in the number of sequences annotated and related to functions such as protein binding, protein homo-dimerization, DNA binding, H3-K4 and H3-K36 methylation (Supl. Figures). A higher number of enzymes with activities such as endonuclease, hydrolase and one-carbon transferases were annotated in the merge file when compared to Trinity (Table 4 A and 4 B).

The *P. verdolaga* transcriptome is enriched in proteins with potential biotechnological application

The non-redundant merged transcriptome of *P. verdolaga* was screened for potentially bioactive peptides using the set of keywords of interest described in Methods. This allowed the identification of 431 genes within the venom gland of *P. verdolaga* that were related to the aforementioned keywords. After manual curation, around 53% of the sequences were found to be related to biological pathways and not to the molecule itself, as identified through the PubMed database (data not shown). Within the 202 identified annotated genes, 52% were related to a protease or protease inhibitor (Kunitz domains), 25% were related to reported arachnid toxins and 8% related to vasodilator or smooth muscle contraction (kinin) activity. The average identity to their respective homologue was 60% (Table 5, Table S1).

The prospection of peptides and proteins with potential biotechnological interest was also carried through homology using peptides from the Arachnoserver and EROP databases as templates. A total of

171 sequences with an e-value higher than $1x10^{-3}$ were identified. After curation, 79% of the sequences were kept. Out of the 122 sequences left from the Arachnoserver database, 40% corresponded to presynaptic neurotoxins, 20% to peptides and proteins with related protease or protease inhibition activities, and 15% to neurotoxins. For the EROP knowledge base 100% of the sequences were kept and 42% of them were related to arachnid toxins, 36% to neurotoxins and 16% to antimicrobial peptides. The average identity of the alignments was smaller for the Arachnoserver database, as 99% of the identified sequences had an identity to their hits smaller than 85%. For EROP, the average identity was of 67% (Table 5). Out of the total sequences identified from the annotation and Arachnoserver/EROP databases, only 36 were previously reported for *P. verdolaga*.

Processing of the peptides and proteins with SpiderP (http://www.arachnoserver.org/spiderP.html) allowed the identification of the mature toxins from both the annotation file and those coming from the Arachnoserver/EROP prediction (202 and 135 unique sequences respectively) (Table S1-S2). Clustering of the sequences at 30% identity revealed the presence of 87 clusters within the Arachnoserver and EROP predictions, allowing the identification of 87 new theraphotoxins: U-Theraphotoxin-Pv4/91. The possible activities reported, are those of all toxins that had a match at a significant e-value and identity percentage (Table S2).

Discussion

Humanity faces many challenges in the search for novel active molecules that help solve animal, human health as well as environmental issues. Recombinant DNA technology and optimization of the solid phase chemical synthesis have allowed peptides and proteins to become a feasible and scalable solution to these problems. Peptides entering clinical trials are now common (26) and venom-derived new products are increasingly available (3,27). Spiders, being the venomous animal with the greatest number of species described to date, are becoming protagonists for the prospection of bioactive molecules (1,17). In the absence of annotated genomes of arachnids (28), the *de novo* assembly of transcriptomes has become the main tool for the molecular characterization of non-model organisms (29,30). Still, indels of nucleotides, assembly of chimeras, under representation of isoforms, presence of partial transcripts, biases in highly and lowly expressed genes and elimination of contigs that contain highly repeated sequences, are some of the current challenges that assemblers' algorithms face (30,31). Moreover, the bias of some algorithms towards the assembly of certain types of proteins and performance variation in different tissues and/or organisms (29,31–33), may be overcome with the use of tools such as CEGMA or BUSCO, as the lack of consensus on the accuracy of the average length or the N50 value as estimators of the assembler's performance also rises, and the use of multiple assemblies with different software (29,31,34).

The quality of the assemblies

All four assemblies obtained with Trinity showed and average of 97,000 contigs, N50 value and average lengths of 750 and 570 bp respectively. These results are in agreement with reported spider

transcriptomes, which are highly variable and contain between 5000 up to 201,000 unigenes, and average lengths around 600 bp (35). Surprisingly, comparison between the assembly presets showed that the slight increase in the k-mer number, produced the increase of the N50 value, the median and average length, while it reduced the total number of contigs and assembled nucleotides, evidencing a potential increase in the assembly quality. However, BUSCO results show a drop in the number of retrieved conserved genes and an increase in the fragmented, missing and duplicated genes. Therefore, for our dataset, higher k-mer lengths assemblies done with Trinity have a lower possibility of correctly retrieving conserved arthropod genes (Table 1 a, b). Alignment of the assembled transcriptomes to the genome of *P. tepitadorium* showed an average length of the aligned sequences above the reported average and maximum length values of the assembled contigs. This implies the inclusion of gaps, mainly due to the phylogenetic distance between *Theraphosidae* (*Mygalomorphae*) and *Theriidae* (*Araneae*) which are estimated to have diverged between 240-300 million years ago (35,36). Around 70% of the aligned sequences have lengths smaller than 500 bp (Figure 2); therefore, the possibility of multiple alignments to different sections of the *P. tepitadorium* genome helped reducing the alignment quality as observed (37). Still, while the comparison between *P. tepitadorium* and *P. verdolaga* is not ideal, the coverage analysis results behave as the BUSCO results, and this indicates that for our dataset, smaller k-mer numbers give a better assembly for the *P. verdolaga* transcriptome with Trinity.

To date, Trinity is regarded as the best *de novo* assembler for various species (29,38); however, biases in the prediction of certain types of proteins suggest the use of other assemblers (24,33). Assembly of the transcriptome of *P. verdolaga* with the software SPAdes and Soapdenovo showed a reduction in the overall number of assembled contigs as well as the N50 statistic in sequences over 200 bp, as expected. SPAdes' assemblies show apparent worse performance as k-mer number increases, which is evidenced in worse assembly statistics and in BUSCO results, showing correlation between both metrics and indicating that a smaller k-mer produces a higher quality assembly (Table 2 a, b). Still, while not as superior in number as Trinity's results, the quality of SPAdes' results are comparable as it has been previously reported (33). However, the smaller number of duplicated genes present in SPAdes' BUSCO terms indicate again that assembly stats fail to describe the whole picture. On the other hand, Soapdenovo's assemblies, while showing the worst performance of all the three software, show a better performance at higher k-mer numbers (Table 2 a, b). Assembly of venom gland transcriptomes for snakes of the *Crotalus* genus as well as scorpions of the *Centruroides* and *Hadrurus* genera with SPAdes and Soapdenovo, show an inverse correlation between assembler performance and increased k-mer number, as percentages of missing and fragmented BUSCO terms increase from 30 to 60% for snakes and from 3 to 10% on scorpions with k-mer numbers from 31 to 137 respectively, due to the possible loss of transcripts with low abundance (24). While this behavior applies to both Trinity and SPAdes, Soapdenovo's results seem to contradict this tendency as it assembles over 1.3 million transcripts with lengths smaller than 200 bp at k-mer length 31, evidence of a high fragmentation of *P. verdolagas*' transcriptome that clearly punishes both assembly stats and BUSCO metrics. Soapdenovo is known to assemble a high proportion of small contigs in other datasets (32). For the prospection of novel bioactive peptides and proteins, the relation between k-mer length and read length in our dataset and for Soapdenovo, should be higher as observed.

This again suggests that, ideally for each organism and/or tissue, optimization of the assembly conditions is highly recommended. Still, the presence of over 160 fragmented core arthropod genes in sequences smaller than 200 bp, suggests the potential presence of genomic information that might be of interest to certain biological questions, since it has been reported that genes that are restricted to one genus or species tend to be shorter than sequences that are shared between different taxa (39).

We observed up to 20% of ORFs (complete genes) in relationship with the initial amount of contigs (Table 3), a number somewhat lower than the 36% reported for transcriptomes of other spiders (35). When we also included partial transcripts in Trinity analyses, we increased to 33% the percentage of complete genes in relation to the number of contigs (not shown); however, the inclusion of incomplete transcripts might negatively affect subsequent studies as annotation of incomplete genes leads to errors in transcript quantification as well as errors in gene expression profiles (40,41). A previous *P. verdolaga* assembly and annotation made by Estrada and collaborators with Trinity and TransDecoder, reported 78,000 contigs from which 20% (16,030) were identified as ORFs and only 45% were identified as non-redundant genes, showing a final 9% (7173 sequences) of the total predicted contigs as genes (25). Our methods classified only 0.07 to 25.2% of the ORFs as redundant sequences and functional annotation of these sequences represented an increase of up to 20% more complete genes in the non-redundant merge, when compared to the previous work.

The low abundance of spider-specific genes in databases contributes to the low level of annotation found. (28,39). Therefore, in some instances, alignments with e-values higher than of $1x10^{-3}$ might be considered as positive hits (42). Approximately 45.5% of the not annotated sequences have IPS hits that provide insight into the possible function of these genes (Table S3) and might classify them as probable “orphan genes”. As it has been observed in other spider transcriptomes, up to 20% of genes are not present within other taxa due to lineage-specific environmental adaptations, while showing the presence of signal peptide, IPS or Pfam hits (39,43). These findings become especially relevant as phylogenomic analysis of new world theraphosids place them as a separate clade group from other arachnids (44) and some reports show that there is higher genetic diversification found in neo-tropical spiders as well as in those spiders that lack orb webs (36). A low percentage of annotated sequences is also observed in spiders of clinical importance, such as those from the *Loxoceles* and *Latrodectus* genera, as only 39 to 54% of the described proteins from transcriptomes in these species are correctly identified (45,46).

Potential new toxins in the venom gland of *P. verdolaga*

The non-redundant merged transcriptome showed an increase of approximately 1200 more annotated genes when compared to the Trinity annotated sequences (highest in total numbers among the three) as well as the overall identification of a higher number of GO terms as observed in Table 4. 1152 of the extra annotated proteins were related to the SPAdes assembly, while 56 were traced to the Soapdenovo assembly, i.e. 27% of the unique sequences belonging to both SPAdes and Soapdenovo were correctly annotated. This increase in the total numbers of ORFs also carried the inclusion of approximately 3200 sequences which were not annotated, decreasing the percentage of correctly annotated sequences when

compared to Trinity by almost 10%, as expected (29,31). However, this reduction in annotation statistics is compensated by the extra sequences that were identified, as 42% of the sequences missing annotation show IPS matches that classify them as possible orphan genes, leaving a total potential of 12,000 identified genes (Table S3). The annotation of the non-redundant merged transcriptome identifies 35% of the annotated sequences as related to cellular processes. There is abundance of GO terms related to nucleotide binding and H3-K4/H3-K36 histones, which reflects the high level of transcriptomic activity that is expected from the venom gland (47). BLAST, IPS hits as well as GO terms, allowed the identification of 113 proteins as proteases or protease inhibitors, 62 toxins with either cytolytic, necrotic, neurotoxic or hemolytic activity and one protein with hyaluronidase activity. Venoms like those from spiders of the genera *Latrodectus* and *Loxoceles* are primarily composed of phospholipases and proteases, at 10 to 20% of the identified sequences. Other enzymes, like hyaluronidases, are found with low abundance (45,48). *Theraphosidae* venoms contain approximately 40% of proteins related to cellular processes and about 5 to 30% of proteins that evidence the ICK motif (possible toxins), while showing abundance in proteins related to redox activity, histones and proteases that also contribute to toxin diversity (49,50). Alignment of the non-redundant merged transcriptome to the Arachnoserver and EROP databases allowed the identification of 171 sequences with biological activities of interest; 36 of these were previously reported (25). All 135 remaining sequences matched to 427 different proteins within the aforementioned databases; however, the best hit was restricted to only 32 proteins at an average e-value of $1x10^{-5}$ and identities smaller than 85% in 98% of the sequences. The CD-HIT analysis showed a high level of homology within *P. verdolagas'* 135 predicted proteins, as 87 clusters were found with global identities higher than 30% (51). All identified proteins were clustered and named as per the nomenclature proposed by King and collaborators, were U stands for the unknown function and Pv for *Pamphobeteus verdolaga*; theraphotoxins-Pv1 to Pv3 were described elsewhere (25,52).

Various patterns could be observed between *P. verdolagas'* predicted sequences and their best score homologs. First, all best score homologs were proteins that within the arachnoserver database are among the longest (over 100 kDa in weight). Second, an inverse relationship was observed between the predicted protein length and its global identity to the homolog protein, as those with the shortest lengths had global identities over 20% and ORFs with similar lengths to the homolog had on average global identities below 20%. Third, on average all proteins showed hits to proteins that were either isoforms to the best score homolog or to homolog proteins described in other species (data not shown). Fourth and last, all 32 proteins were described in 11 species, of which 30% were spiders from the genera *Latrodectus* or *Loxoceles,* evidencing again the bias in arachnid research. All the smaller *P. verdolagas'* ORFs are hypothesized to be then smaller isoforms of the best score homolog proteins, as the Augustus prediction algorithm identified them as complete transcripts and the global identity was above 20% (51). The varying lengths of these ORFs, which represent 7 to 27% of the best score homolog proteins primary structure, could be indicators of splicing events within the venom gland of spiders, as it has been described before in other venomous species (53,54), and therefore could also imply similar biological function. On the other hand, ORFs with the smallest global identity show between 58 and 164% of the

best homolog protein's length, implying enough variability within the primary structure as to assume probable different biological activities to those of their homologs (51).

Forty *P. verdolagas'* toxins, distributed in 29 clusters, matched α-Latrocrustotoxin. This toxin, from *Latrodectus tredecimguttatus*, affects crustaceans by promoting the release of neurotransmitters by various mechanisms, including pore formation (55). α and δ-Latroinsectotoxin also induce neurotransmitter release by pore formation; however, these are found to be only active in arthropods making them potential insecticidal proteins (55,56). 8 and 9 of the described *P. verdolaga* toxins are related to α and δ-Latroinsectotoxin respectively. α-Latrotoxin-Lt1a is a 150kDa toxin that shares 93% of its amino acid sequence to the toxin α-Latrotoxin-Lh1a from *Latrodectus hasselti*. These α-Latrotoxins show toxicity to vertebrates and in some situations can represent a fatal danger to mammals due to the formation of membrane pores that lead to neurotransmitter release (57). 12 and 7 proteins were found to match to α-Latrotoxin-Lt1a and Lha1a respectively. Presence of 19 sequences within the transcriptome of *P. verdolaga* that share identity to these toxins might indicate probable noxious effects of this theraphosid venom to mammals; however, the pore formation activity could be exploited in the development of various cytolytic drugs that may prove to be either antimicrobial or anti-cancerous. *P. verdolagas'* ORFs are also related to toxins described from the *Loxoceles* genera. Such is the case of the toxin Hayluronidase-1 from the spider *Loxoceles intermedia*, which is related to skin necrosis (58) and therefore indicative of possible necrotic activity in the bite of *P. verdolaga.*

Seven of the described toxins are related to the peptide κ-theraphotoxin-Gr1d from the spider *Grammostola rosea*, which is also homolog to the Hanatoxin (HaTx) from the same species. These toxins are known to modulate voltage-gated potassium channels, specifically Kv2.1 and Kv4.2 (59), which could pose them as valuable tools for ion channel research. However, unlike previously described toxins (and due to their short length), these toxins are also probable homologs to other *Grammostola rosea* toxins such as the antimicrobial peptide M-theraphotoxin-Gr1a (60), the analgesic and antiarrhythmic toxin κ-theraphotoxin-Gr2c (as stated by US patents US5756663 and US5877026), and the lectin-like toxin U5-theraphotoxin-Hs1a from the spider *Haplopelma schmidti* (61), making them some of the most interesting peptides found in the venom gland transcriptome of *P. verdolaga*. Other ORFs related to theraphosid venoms are two toxins related to *H. schmidti's* κ-theraphotoxin-Hsb1, a potassium channel and protease inhibitor (62,63), and three ORFs related to U15-theraphotoxin-Hhn1a/e/h, three protease inhibitors from the venom of the spider *Haplopelma hainanum* (63,64).

Other ORFs related to proteins from other species were the Acetylcholinesterase-1, Cysteine rich secretory protein-1 (CRISP-1) and Neprilysin-1 from the spider *Trittame loki*, homologs to 23 of the *P. verdolagas'* predicted proteins and distributed in 1, 3 and 10 clusters respectively. These proteins are known to be related to the envenomation process, as acetylcholinesterase cleaves acetylcholine in neuron synapses and Neprilysin-1 is a metalloprotease that is involved in the degradation of proteins from the extracellular matrix; CRISP-1 is to be related to vespid allergens (65). And the toxin U1-filistatoxin-Kh1b from the venom of the spider *Kukulcania hibernalis*, an insecticidal toxin as described by patent number US5457178, which is homolog to two of the *P. verdolagas'* ORFs.

Some of the hits can also be traced to organisms outside the *Araneae* order. Such examples are the homology of one of the predicted sequences to the Kazal-type serine protease inhibitor from the lobster *Sagmariasus verreauxi* (66), four ORFs homolog to species of the genera *Scolopendra*: Scolopendin 1 and 2, both antimicrobial and antifungal peptides (67,68), the toxin SLPTX10-4 of unknown function (69) and lastly Scolopendrasin VII, an anticancer peptide (70). Finally, 14 ORFs also showed homology to the protein Techylectin-1 from the *Phoneutria nigriventer*, which is also homolog to the toxin Techylectin-5B, an antimicrobial lectin from the horseshoe crab *Tachypleus tridentatus* (71).

Conclusions

The *de novo* transcriptome assembly has become the main method of obtaining genomic information in non-model organisms, especially in spiders, were there is an evident lack of nucleotide information in contrast to the number of reported species to date. As these arthropods become coveted for the richness of bioactive peptides in their venoms, the completeness in the resulting assembled transcriptomes also gains importance. Here, the utilization of three distinct assembly algorithms allowed the identification and annotation of over 1200 more proteins (Over 4000 considering the probable orphan genes) when compared to the v1.0 of *P. verdolagas'* transcriptome assembly previously done in our group. This supports the need for the merging of the output of various assemblers as a means to obtain more complete transcriptomes. Furthermore, the 8460 annotated proteins will serve as guidance for the description and annotation of posterior spider proteins, especially in the absence of nucleotide sequences of new world theraphosids. On the other hand, the prospection of bioactive peptides allowed the identification 135 new theraphotoxins with exciting hypothetic activities that range from probable new antimicrobials to novel insecticidal proteins that may have a place in the development of new products in the medium term. These activities are then left to be tested biologically.

Abbreviations

%GC: Guanidine-Cytosine percentage

BUSCO: Benchmarking Universal Single-Copy Orthologs

CEGMA: Core Eukaryotic Genes Mapping Approach

CRISP: Cysteine rich secretory protein

DBPs: Disulfide Bridged Peptides

ENA: European Nucleotide Archive

EROP: Endogenous Regulatory Oligo Peptide knowledgebase

ESM: Extra Structural Motif

GO: Gene Ontology

ICK: Inhibitory Cysteine Knot

IPS: InterPro Scan

MRSA: Methicillin Resistant *Staphylococcus aureus*

NDBPs: Non-disulfide Bridged Peptides

NGS: Next Generation Sequencing.

ORFs: Open reading frame

PG: Phosphatidyl Glycerol

PSM: Primary Structural Motif

Main Findings

1. Pre-processing of reads carries less variation in the quality of the obtained contigs when compared to changes in the assembly algorithm.
2. Trinity, as reported, is the most used *de novo* assembler, and this is reflected by its performance, as measured by assembly stats and BUSCO terms retrieval. However, SPAdes and Soapdenovo were able to assemble contigs that Trinity could not, evidencing the need for the use of various assemblers when assembling transcriptomes from non-model organisms.
3. The merge of all three assemblies allowed the increment of up to 15% more annotated sequences when compared to the annotation of Trinity alone. When compared to a previous *verdolaga* annotation, the merge allowed the identification of about 20% more proteins.
4. This second version of the transcriptome of the venom gland of *verdolaga* also allowed the identification of 135 new theraphotoxins that share sequence identity, and therefore possible homology to toxins already reported in the Arachnoserver database.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All raw *Pamphobeteus verdolaga's* reads have been submitted previously to the European Nucleotide Archive under accessions numbers: PRJEB21288/ERS1788422/ERX2067777-ERR2008012 (25). All of the assemblies have been deposited at DDBJ/EMBL/GenBank under the accession GIUY00000000. The version described in this paper is the first version, GIUY01000000. The mature toxins and their functional annotation, as well as the identified probable orphan genes are available within the article in the documents described below:

- pdf: Illustrates the top 20 GO terms found for each category, in the annotation of the transcriptomes and the non-redundant merge. The top functions of proteins associated with the non-redundant merge and Trinity annotation are also shown:
- xlsx: List of mature toxins identified from the OmicsBox annotation.
- xlsx: List of the mature toxins identified from the Arachnoserver and EROP databases.
- xlsx: List of probable orphan genes identified in the non-redundant merged transcriptome

Competing interests

The authors declare that they have no competing interests

Funding

This work was done with the help of Ministerio de Ciencia y Tecnología (MinCiencias) of Colombia through the project 111577757673 executed through the contract 761 of 2017 with the Universidad de Antioquia in Medellín, Colombia. All authors declare that the funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author's contributions

Cristian Salinas-Restrepo conducted, analyzed and processed all the bioinformatic data and wrote the first draft. Elizabeth Misas gave important advisory and guidance for the transcriptome assembly and pre-processing, as well as software troubleshooting. Sebastian Estrada-Gomez provided all information regarding *P. verdolaga*. Juan Quintana gave advisory regarding spider physiology/biology as well as the manuscript's proofreading. Fanny Guzmán gave advisory regarding protein and peptide bioinformatic analysis. Juan Camilo Calderón, Marco Antonio Giraldo and Cesar Segura participated in writing and financing the project and did intellectual contributions to the manuscript and its proofreading. Cesar Segura was also the team coordinator and resource manager.

Acknowledgements

Special thanks to Orville Hernandez, Oscar Mauricio Gómez and to Grupo Biología Molecular from Corporación para Investigaciones Biológicas in Medellin, Colombia for their support with the infrastructure for this work. And to Agencia de Educación Superior de Medellín Sapiencia for their support through their program "Extendiendo fronteras educativas".

References

1. Natural History Museum Bern. World Spider Catalog [Internet]. World Spider Catalog v 20.0. 2019 [cited 2020 Feb 8]. Available from: http://wsc.nmbe.ch
2. Dubovskii P V., Vassilevski AA, Kozlov SA, Feofanov A V., Grishin E V., Efremov RG. Latarcins: Versatile spider venom peptides. Vol. 72, Cellular and Molecular Life Sciences. Springer Basel; 2015. p. 4501–22.
3. King GF, Hardy MC. Spider-Venom Peptides: Structure, Pharmacology, and Potential for Control of Insect Pests. Annu Rev Entomol. 2013;58(1):475–96.
4. Vassilevski AA, Kozlov SA, Grishin E V. Molecular diversity of spider venom. Biochem. 2009;74(13):1505–34.
5. King GF. Modulation of insect Cav channels by peptidic spider toxins. Vol. 49, Toxicon. 2007. p. 513–30.
6. Rodrigues EG, Dobroff ASS, Cavarsan CF, Paschoalin T, Nimrichter L, Mortara RA, et al. Effective Topical Treatment of Subcutaneous Murine B16F10-Nex2 Melanoma By the Antimicrobial Peptide Gomesin. Neoplasia. 2008;10(1):61–8.
7. Santibáñez-López CE, Possani LD. Overview of the Knottin scorpion toxin-like peptides in scorpion venoms: Insights on their classification and evolution. Toxicon. 2015;107:317–26.
8. Chellat MF, Raguž L, Riedl R. Targeting Antibiotic Resistance. Angew Chemie - Int Ed. 2016;55(23):6600–26.
9. WHO. Antimicrobial resistance. Global report on surveillance. 1st ed. Cadman H, Martinez L, editors. Vol. 61, World Health Organization. Geneva: World Health Organization; 2014. 12–28 p.
10. Windley MJ, Herzig V, Dziemborowicz SA, Hardy MC, King GF, Nicholson GM. Spider-venom peptides as bioinsecticides. Toxins (Basel). 2012;4(3):191–227.
11. Wu T, Wang M, Wu W, Luo Q, Jiang L, Tao H, et al. Spider venom peptides as potential drug candidates due to their anticancer and antinociceptive activities. J Venom Anim Toxins Incl Trop Dis. 2019;25(June 2018):1–13.
12. Saez NJ, Herzig V. Versatile spider venom peptides and their medical and agricultural applications. Toxicon. 2019;158(December 2018):109–26.
13. Rapôso C. Scorpion and spider venoms in cancer treatment: state of the art, challenges, and perspectives. J Clin Transl Res. 2017;3(2):233–49.
14. Tanner JD, Deplazes E, Mancera RL. The Biological and Biophysical Properties of the Spider Peptide Gomesin. Molecules. 2018;23(7):6–9.
15. Wang X, Wang G. Insights into Antimicrobial Peptides from Spiders and Scorpions. Protein Pept Lett. 2016;23(8):707–21.
16. Santos DM, Reis P V, Pimenta AMC. Antimicrobial Peptides in Spider Venoms. Spider Venoms. 2015;1–15.

17. Pineda SS, Chaumeil PA, Kunert A, Kaas Q, Thang MWC, Le L, et al. ArachnoServer 3.0: An online resource for automated discovery, analysis and annotation of spider toxins. Bioinformatics. 2018;34(6):1074–6.
18. Estrada-Gomez S, Vargas Muñoz LJ, Quintana Castillo JC. Extraction and partial characterization of venom from the Colombian spider Pamphobeteus aff. nigricolor (Aranae:Theraphosidae). Toxicon. 2013;76:301–9.
19. Oldrati V, Koua D, Allard PM, Hulo N, Arrell M, Nentwig W, et al. Peptidomic and transcriptomic profiling of four distinct spider venoms. PLoS One. 2017;12(3):1–18.
20. Luna-Ramírez K, Quintero-Hernández V, Juárez-González VR, Possani LD. Whole transcriptome of the venom gland from urodacus yaschenkoi scorpion. PLoS One. 2015;10(5):1–33.
21. Bouzid W, Verdenaud M, Klopp C, Ducancel F, Noirot C, Vétillard A. De Novo sequencing and transcriptome analysis for tetramorium bicarinatum: A comprehensive venom gland transcriptome analysis from an ant species. BMC Genomics. 2014;15(1):1–18.
22. Chetia H, Kabiraj D, Singh D, Mosahari PV, Das S, Sharma P, et al. De novo transcriptome of the muga silkworm, Antheraea assamensis (Helfer). Vol. 611, Gene. 2017. 54–65 p.
23. Gupta SK, Kupper M, Ratzka C, Feldhaar H, Vilcinskas A, Gross R, et al. Scrutinizing the immune defence inventory of Camponotus floridanus applying total transcriptome sequencing. BMC Genomics. 2015;16(1):1–21.
24. Holding ML, Margres MJ, Mason AJ, Parkinson CL, Rokyta DR. Evaluating the performance of de novo assembly methods for venom-gland transcriptomics. Toxins (Basel). 2018;10(6):1–21.
25. Estrada-Gomez S, Cardoso FC, Vargas-Muñoz LJ, Quintana-Castillo JC, Gómez CMA, Pineda SS, et al. Venomic, Transcriptomic, and Bioactivity Analyses of Pamphobeteus verdolaga Venom Reveal Complex Disulfide-Rich Peptides That Modulate Calcium Channels Sebastian. Toxins (Basel). 2019;11(496):21.
26. Lau JL, Dunn MK. Therapeutic peptides: Historical perspectives, current development trends, and future directions. Bioorganic Med Chem. 2018;26(10):2700–7.
27. Bende NS, Dziemborowicz S, Herzig V, Ramanujam V, Brown GW, Bosmans F, et al. The insecticidal spider toxin SFI1 is a knottin peptide that blocks the pore of insect voltage-gated sodium channels via a large β-hairpin loop. FEBS J. 2015;282(5):904–20.
28. NIH. NCBI refseq [Internet]. [cited 2020 Jun 11]. Available from: https://www.ncbi.nlm.nih.gov/refseq/
29. Rana SB, Zadlock FJ, Zhang Z, Murphy WR, Bentivegna CS. Comparison of de Novo transcriptome assemblers and k-mer strategies using the killifish, fundulus heteroclitus. PLoS One. 2016;11(4):1–16.
30. Freedman AH, Clamp M, Sackton TB. Error, noise and bias in de novo transcriptome assemblies. Molecular Ecology Resources. 2020. 0–3 p.
31. Cabau C, Escudié F, Djari A, Guiguen Y, Bobe J, Klopp C. Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. PeerJ. 2017;2017(2):e2988.

32. Sadat-Hosseini M, Bakhtiarizadeh MR, Boroomand N, Tohidfar M, Vahdati K. Combining independent de novo assemblies to optimize leaf transcriptome of Persian walnut. PLoS One. 2020;15(4):1–17.
33. Hölzer M, Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience. 2019;8(5):1–16.
34. Honaas LA, Wafula EK, Wickett NJ, Der JP, Zhang Y, Edger PP, et al. Selecting superior de novo transcriptome assemblies: Lessons learned by leveraging the best plant genome. PLoS One. 2016;11(1):1–42.
35. Garrison NL, Rodriguez J, Agnarsson I, Coddington JA, Griswold CE, Hamilton CA, et al. Spider phylogenomics: Untangling the Spider Tree of Life. PeerJ. 2016;2016(2).
36. Fernández R, Kallal RJ, Dimitrov D, Ballesteros JA, Arnedo MA, Giribet G, et al. Phylogenomics, Diversification Dynamics, and Comparative Transcriptomics across the Spider Tree of Life. Curr Biol. 2018;28(9):1489-1497.e5.
37. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
38. He B, Zhao S, Chen Y, Cao Q, Wei C, Cheng X, et al. Optimal assembly strategies of transcriptome related to ploidies of eukaryotic organisms. BMC Genomics. 2015;16(1):1–10.
39. Carlson DE, Hedin M. Comparative transcriptomics of Entelegyne spiders (Araneae, Entelegynae), with emphasis on molecular evolution of orphan genes. PLoS One. 2017;12(4).
40. Morillon A, Gautheret D. Bridging the gap between reference and real transcriptomes. Genome Biol. 2019;20(1):1–7.
41. Hsieh PH, Oyang YJ, Chen CY. Effect of de novo transcriptome assembly on transcript quantification. Sci Rep. 2019;9(1):1–12.
42. Chen J, Zhao L, Jiang L, Meng E, Zhang Y, Xiong X, et al. Transcriptome analysis revealed novel possible venom components and cellular processes of the tarantula Chilobrachys jingzhao venom gland. Toxicon. 2008;52(7):794–806.
43. Cheng TC, Long RW, Wu YQ, Guo YB, Liu DL, Peng L, et al. Identification and characterization of toxins in the venom gland of the Chinese bird spider, Haplopelma hainanum, by transcriptomic analysis. Insect Sci. 2016;23(3):487–99.
44. Foley S, Lüddecke T, Cheng DQ, Krehenwinkel H, Künzel S, Longhorn SJ, et al. Tarantula phylogenomics: A robust phylogeny of deep theraphosid clades inferred from transcriptome data sheds light on the prickly issue of urticating setae evolution. Mol Phylogenet Evol. 2019;140(April):106573.
45. Haney RA, Ayoub NA, Clarke TH, Hayashi CY, Garb JE. Dramatic expansion of the black widow toxin arsenal uncovered by multi-tissue transcriptomics and venom proteomics. BMC Genomics. 2014;15(1):1–18.
46. Gremski LH, Da Silveira RB, Chaim OM, Probst CMA, Ferrer VP, Nowatzki J, et al. A novel expression profile of the Loxosceles intermedia spider venomous gland revealed by transcriptome analysis. Mol Biosyst. 2010;6(12):2403–16.

47. Fernandes-Pedrosa M de F, Junqueira-de-Azevedo I de LM, Gonçalves-de-Andrade RM, Kobashi LS, Almeida DD, Ho PL, et al. Transcriptome analysis of Loxosceles laeta (Araneae, Sicariidae) spider venomous gland using expressed sequence tags. BMC Genomics. 2008;9:1–12.
48. Regina D, Nunes F, Otto G, Marcelo A, Helena L, Trevisan-silva D, et al. Brown spider (Loxosceles genus) venom toxins : Evaluation of biological conservation by immune cross-reactivity. 2015;108:154–66.
49. Zhang YY, Huang Y, He QZ, Luo J, Zhu L, Lu SS, et al. Structural and functional diversity of peptide toxins from tarantula Haplopelma hainanum (Ornithoctonus hainana) venom revealed by transcriptomic, peptidomic, and patch clamp approaches. J Biol Chem. 2015;290(22):14192–207.
50. Zhang Y, Huang Y, He Q, Liu J, Luo J, Zhu L, et al. Toxin diversity revealed by a transcriptomic study of Ornithoctonus huwena. PLoS One. 2014;9(6).
51. Pearson WR. An introduction to sequence similarity ("homology") searching. Curr Protoc Bioinforma. 2013;
52. King GF, Gentz MC, Escoubas P, Nicholson GM. A rational nomenclature for naming peptide toxins from spiders and other venomous animals. Toxicon. 2008;52(2):264–76.
53. Ogawa T, Oda-Ueda N, Hisata K, Nakamura H, Chijiwa T, Hattori S, et al. Alternative mRNA splicing in three venom families underlying a possible production of divergent venom proteins of the habu snake, protobothrops flavoviridis. Toxins (Basel). 2019;11(10).
54. Zeng XC, Luo F, Li WX. Characterization of a novel cDNA encoding a short venom peptide derived from venom gland of scorpion Buthus martensii Karsch: Trans-splicing may play an important role in the diversification of scorpion venom peptides. Peptides. 2006;27(4):675–81.
55. Rohou A, Nield J, Ushkaryov YA. Insecticidal toxins from black widow spider venom. Toxicon. 2007;49(4):531–49.
56. Ashton AC, Rahman MA, Volynski KE, Manser C, Orlova E V., Matsushita H, et al. Tetramerisation of α-latrotoxin by divalent cations is responsible for toxin-induced non-vesicular release and contributes to the Ca2+-dependent vesicular exocytosis from synaptosomes. Biochimie. 2000;82(5):453–68.
57. Ushkaryov YA. α-Latrotoxin and Its Receptors. Handb Exp Pharmacol. 2008;184(184):1–33.
58. Ferrer VP, de Mari TL, Gremski LH, Trevisan Silva D, da Silveira RB, Gremski W, et al. A Novel Hyaluronidase from Brown Spider (Loxosceles intermedia) Venom (Dietrich's Hyaluronidase): From Cloning to Functional Characterization. PLoS Negl Trop Dis. 2013;7(5):1–12.
59. Shiau YS, Lin T Bin, Liou HH, Huang PT, Lou KL, Shiau YY. Molecular simulation reveals structural determinants of the hanatoxin binding in Kv2.1 channels. J Mol Model. 2002;8(8):253–7.
60. Jung HJ, Kim P Il, Lee SK, Lee CW, Eu YJ, Lee DG, et al. Lipid membrane interaction and antimicrobial activity of GsMTx-4, an inhibitor of mechanosensitive channel. Biochem Biophys Res Commun. 2006;340(2):633–8.
61. Liang SP, Pan X. A lectin-like peptide isolated from the venom of the chinese bird spider Selenocosmia huwena. Toxicon. 1995;33(7):875–82.

62. Liang S. An overview of peptide toxins from the venom of the Chinese bird spider Selenocosmia huwena Wang [=Ornithoctonus huwena (Wang)]. Toxicon. 2004;43(5):575–85.
63. Yuan CH, He QY, Peng K, Diao JB, Jiang LP, Tang X, et al. Discovery of a distinct superfamily of kunitz-type toxin (KTT) from Tarantulas. PLoS One. 2008;3(10).
64. Tang X, Zhang Y, Hu W, Xu D, Tao H, Yang X, et al. Molecular diversification of peptide toxins from the tarantula haplopelma hainanum (Ornithoctonus hainana) venom based on transcriptomic, peptidomic, and genomic analyses. J Proteome Res. 2010;9(5):2550–64.
65. Undheim EAB, Sunagar K, Herzig V, Kely L, Low DHW, Jackson TNW, et al. A Proteomics and Transcriptomics investigation of the venom from the Barychelid spider Trittame loki (brush-foot trapdoor). Toxins (Basel). 2013;5(12):2488–503.
66. Ventura T, Cummins SF, Fitzgibbon Q, Battaglene S, Elizur A. Analysis of the central nervous system transcriptome of the Eastern rock lobster Sagmariasus verreauxi reveals its putative neuropeptidome. PLoS One. 2014;9(5).
67. Choi H, Hwang JS, Lee DG. Identification of a novel antimicrobial peptide, scolopendin 1, derived from centipede Scolopendra subspinipes mutilans and its antifungal mechanism. Insect Mol Biol. 2014;23(6):788–99.
68. Lee H, Hwang JS, Lee J, Kim J Il, Lee DG. Scolopendin 2, a cationic antimicrobial peptide from centipede, and its membrane-active mechanism. Biochim Biophys Acta - Biomembr. 2015;1848(2):634–42.
69. Ward MJ, Rokyta DR. Venom-gland transcriptomics and venom proteomics of the giant Florida blue centipede, Scolopendra viridis. Toxicon. 2018;152:121–36.
70. Lee JH, Kim IW, Kim SH, Kim MA, Yun EY, Nam SH, et al. Anticancer activity of the antimicrobial peptide scolopendrasin VII derived from the centipede, scolopendra subspinipes mutilans. J Microbiol Biotechnol. 2015;25(8):1275–80.
71. Kawabata S ichiro, Iwanaga S. Role of lectins in the innate immunity of horseshoe crab. Dev Comp Immunol. 1999;23(4–5):391–400.

Tables

Table 1. Optimization part 1: While the differences between the presets cannot be taken as significant, there is a tendency for the reads pre-processed with TrimGalore and assembled with a k-mer of 25 to have better performance than the other presets.

1 A. Summary of the assembly statistics for the Trinity optimization.

	Trinity + Trimmomatic		Trinity + TrimGalore	
	K-mer 25	K-mer 32	K-mer 25	K-mer 32
# of transcripts	99598	97769	99316	96821
% GC	39.6	39.8	39.5	39.8
N50	768	796	772	797
Median Length	339	345	339	345
Average contig	570.7	583.9	572.4	584.9
# nt	56840668.0	57084984.0	56851075.0	56627550.0

1 B. Summary of the BUSCO terms retrieved for the Trinity optimization

	Trinity + Trimmomatic		Trinity + TrimGalore	
	K-mer 25	K-mer 32	K-mer 25	K-mer 32
Complete buscos	79.54%	78.42%	79.36%	77.95%
fragmented buscos	9.01%	9.19%	9.1%	9.47%
missing buscos	11.44%	12.38%	11.53%	12.57%
duplicated buscos	22.87%	23.69%	24.23%	23.7%

1 C. Metrics of alignment between the *Parasteatoda tepitadorium* genome and the transcriptomes assembled through Trinity.

	Trinity + Trimmomatic		Trinity + TrimGalore	
	K-mer 25	K-mer 32	K-mer 25	K-mer 32
AVERAGE LENGTH	13154 BP	10724 BP	14083 BP	11069 BP
TOTAL COVERAGE	27979065 BP	24998738 BP	29236516 BP	25625827 BP
AVERAGE QUALITY	14	12	14	12

Table 2. Optimization part 2: While not drastic, the differences between the presets show a tendency for the reads assembled with SPAdes k-mer 31, and Soapdenovo k-mer of 63 to have better balance between

the number of transcripts, N50, the number of completed and missing BUSCOs and the coverage analysis in regard to the SPAdes k-mer 31 assembly. Even when the Soapdenovo k-mer 31 assembly to the P. tepitadorium genome has higher average length and quality, the total coverage favors the Soapdenovo k-mer 63 alignment.

2 A. Summary of the assembly statistics for the contigs obtained with Spades and Soapdenovo with k-mers 31 and 63.

	trinity	SPAdes		soapdenovo	
	K-mer 32	K-mer 31	K-mer 63	K-mer 31	K-mer 63
# of contigs bigger than 200 bp	96821	80030	81831	68700	59910
# of contigs smaller than 200 bp	0	58250	32694	1348391	169055
% GC	39.82	39.42	39.89	38.93	40.21
N50	797	882	722	468	504
# nt	56627550	49124765	44839329	29425929	26911362

2 B. Quality assessment of Spades and Soapdenovo's assemblies trough BUSCO.

	Spades				Soapdenovo			
	K-mer 31		K-mer 63		K-mer 31		K-mer 63	
Size	< 200 bp	> 200 bp	< 200 bp	> 200 bp	< 200 bp	> 200 bp	< 200 bp	> 200 bp
Complete buscos	0%	79,36%	0%	74,01%	0%	46,15%	0,1%	58,26%
fragmented buscos	3,65%	9,01%	3,28%	12,76%	16,51%	32,18%	6,38%	22,42%
missing buscos	96,34%	11,63%	96,72%	13,23%	83,49%	21,67%	93,52%	19,32%
duplicated buscos	0%	8,98%	0%	6,46%	0%	1,42%	0%	2,25%

2 C. Metrics of the the alignment between the *Parasteatoda tepitadorium* genome and the transcriptomes assembled through SPAdes and Soapdenovo.

	Trinity	SPAdes		Soapdenovo	
	K-mer 32	K-mer 31	K-mer 63	K-mer 31	K-mer 63
AVERAGE LENGTH	11069	11269	7819	7091	5050
TOTAL COVERAGE	25625827	17635778	15521127	3148352	6004386
AVERAGE QUALITY	12	15	11	20	11
Maximum length	247989	247989	247989	106194	168652
Minimum length	40	40	40	40	40

Table 3. Summary of the structural annotation stats for the assembled transcripts from Trinity k-mer 25, SPAdes k-mer 31 and Soapdenovo k-mer 63. Trinity stats show better performance when compared to the SPAdes and Soapdenovo assemblies. However, as over 25% of the encountered ORFs are duplicated at 85% homology, SPAdes show a better number of predicted genes.

	# ORFs	AVERAGE LENGTH	MAXIMUM LENGTH	TOTAL AA NUMBER	% redundant genes	# non-redundant genes
Trinity	17,706	287.5	4,077	5,089,960	25.20	13,244
SPADES	15,858	239.3	3,777	3,795,493	12.31	13,905
SOAP	8,465	217.7	1,790	1,842,643	0.07	8,459

Table 4. Summary of the merge file annotation versus the Trinity assembly annotation stats, regarding the top 10 GO terms a), number of enzymes annotated b) and types of transferases identified c). Overall, the merge of the three transcriptomes shows an increment in the number of annotated genes, as expressed by the number of GO terms found. Biggest change observed was in the amount of terms related to the transferases activity.

4 A. Summary of the top 10 GO terms for the non-redundant merge when compared to Trinity.

GO terms	Merge	Trinity	Change
Protein binding	2580	2232	Up 16%
Protein homodimerization activity	762	594	Up 28 %
Identical protein binding	715	810	Down 13%
ATP binding	625	608	Up 3%
DNA binding	615	405	Up 52%
Single Stranded DNA binding	443	209	Up 112%
Metal ion binding	436	447	Down 3%
Double Stranded DNA binding	392	156	Up 151%
H3-K4	372	139	Up 168%
H3-K36	362	128	Up 183%
Endonuclease	124	58	Up 114%
Hydrolase	95	76	Up 25%

4 B. Summary of the annotated enzymes in the non-redundant merge and Trinity, classified by their activity.

Type of enzyme	Merge	Trinity	Change
Oxidoreductases	325	346	Down 6%
Trasferases	1171	884	Up 32%
Hydrolases	1006	1036	Down 3%
Lyases	102	104	Down 2%
Isomerasas	93	97	Down 4%
Ligases	91	87	Up 5%
Translocases	75	75	No change

Table 5. Summary of the prospection of bioactive peptides and proteins of biotechnological interest within the annotated sequences and homology through BLASTP using the Arachnoserver database and the EROP knowledge base. 202 proteins of interest were identified through the functional annotation and

135 novel theraphotoxins were identified by contrasting the transcriptome to the Arachnoserver and EROP databases.

	Annotation	Arachnoserver	EROP
Unique Genes	431	158	13
After Curation	202	122	13
Antimicrobial	4.11%	3.64%	16.13%
Nociception	0%	0.53%	0%
Antiparasitic	0%	0%	0%
Antiarrhythmic	0%	1.04%	0%
Cytolytic	1.37%	1.04%	0%
Hemolytic	0.91%	1.04%	3.22%
Hyaluronidase	0%	0.52%	0%
Inflammatory	1.37%	0.52%	0%
Kinin	8.22%	0%	0%
Lectin	6.39%	14.58%	0%
Necrosis	0.91%	1.04%	0%
Neurotoxin	0%	15.10%	35.48%
Neurotransmitter hydrolysis	0%	2.08%	0%
Presynaptic neurotoxin	0%	39.59%	0%
Protease – Protease inhibitor	52%	19.27%	3.22%
Toxin (General)	25%	NA	41.93%
Average identity	60.34%	37.68%	66.96%
Min. identity	38.31%	20.82%	44.73%
Max. identity	97.75%	98.27%	100%
With identity < 50%	16.44%	88.02%	6.45%
With identity < 85%	97.72%	99.45%	87.09%
Annotated	100%	96.35%	87.09%

Figures

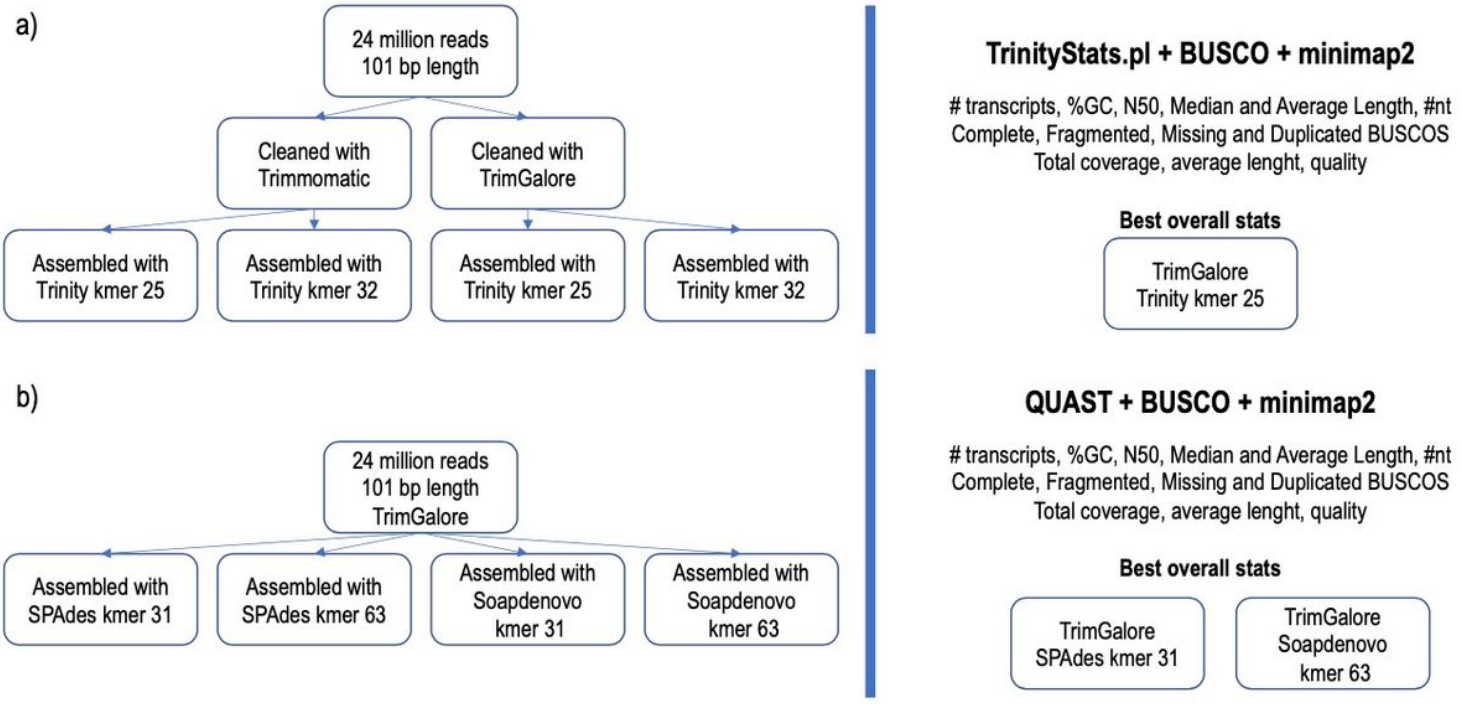

Figure 1

Summary scheme of the optimization of the assembly of the venom gland transcriptome of Pamphobeteus verdolaga. The optimization was carried out in two stages: first obtaining the best assembly with the most used de novo assembly software: Trinity; then, obtaining the best assembly with two other free software, SPAdes and Soapdenovo. The decision was taken based on stats obtained through the TrinityStats.pl script and QUAST software, as well as the number of retrieved terms from BUSCO and the coverage obtained when aligned to the genome of the common house spider Parasteatoda tepitadorium.

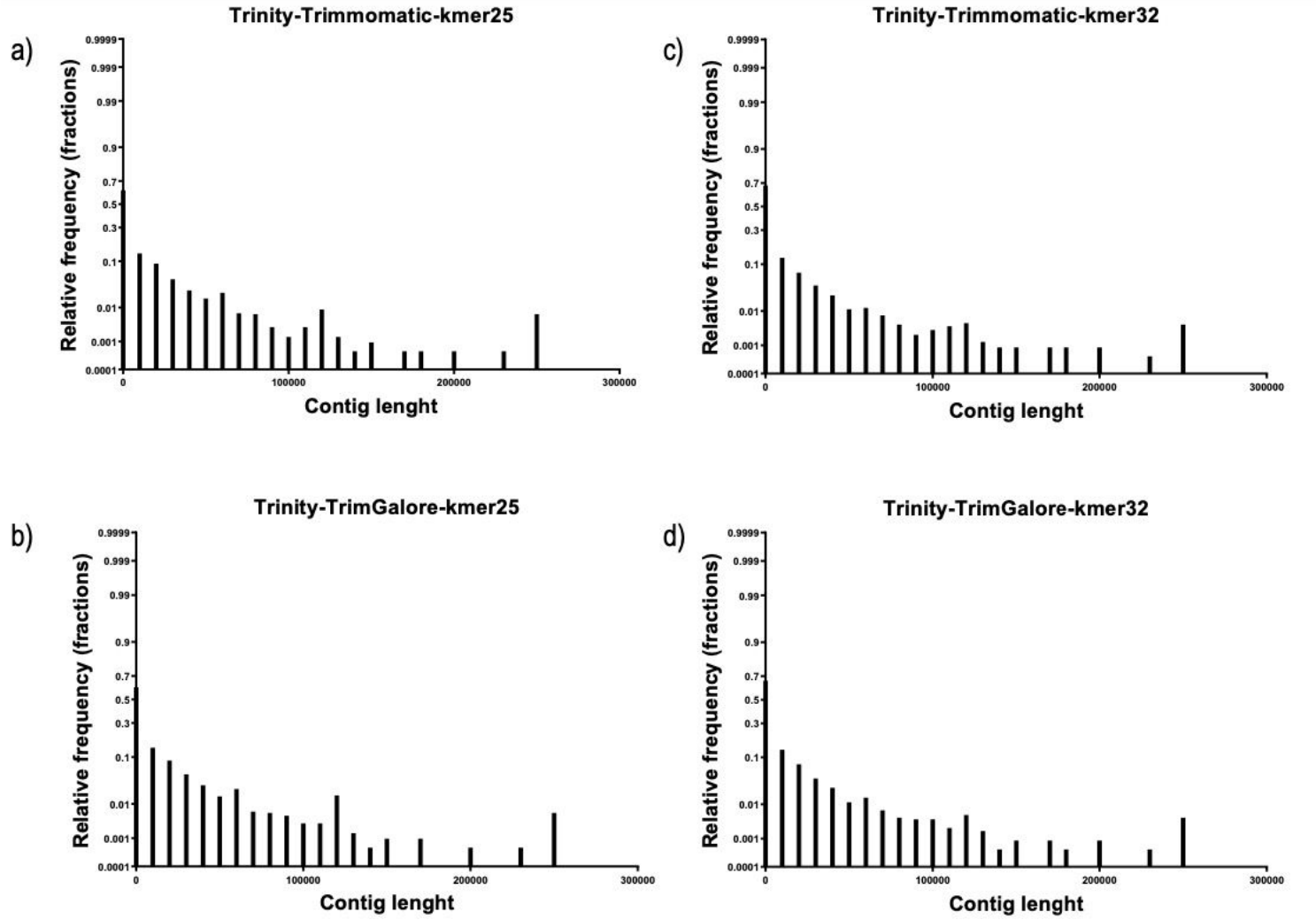

Figure 2

Optimization part 1: Graphical summary of the coverage analysis done with minimap2. Distribution of the length of the alignments between the Parasteatoda tepitadorium genome and the assembly preprocessed with Trimmomatic and a k-mer of 25 a), k-mer of 32 b). Assembly preprocessed with TrimGalore and a k-mer of 25 c), k-mer of 32 d). Smaller k-mer numbers give a smaller percentage of alignments with longitude smaller than 10000 bp. However, the highest variability can be observed between 100,000 and 200,000 bp with changes between 1 to 0.1%, making the distribution of the alignment lengths similar between all presets.

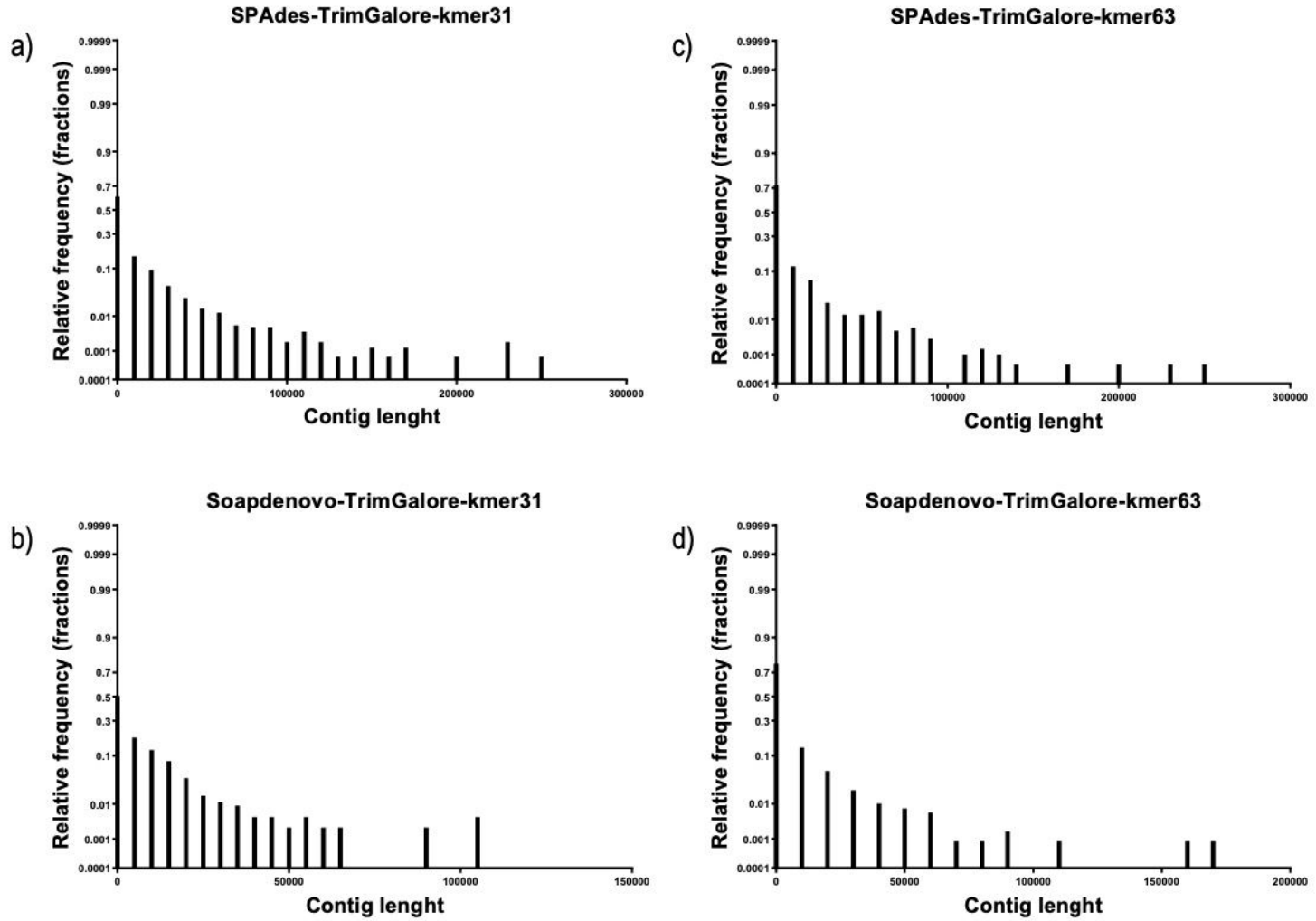

Figure 3

Optimization part 2: Graphical summary of the coverage analysis done with minimap2 for the SPAdes and Soapdenovo assembled transcripts. The length distribution is shown for the the alignments between the Parasteatoda tepitadorium genome and the assemblies from SPAdes k-mer of 31 a), SPAdes k-mer of 32 b), Soapdenovo k-mer of 31 c) and Soapdenovo k-mer of 63 d). As seen in Figure 2, smaller k-mer numbers result in a smaller percentage of alignments with longitude smaller than 10,000 bp. The distribution between SPAdes and Trinity is similar, while Soapdenovo results in a different distribution with significant gaps in the 50,000 to 100,000 bp region.

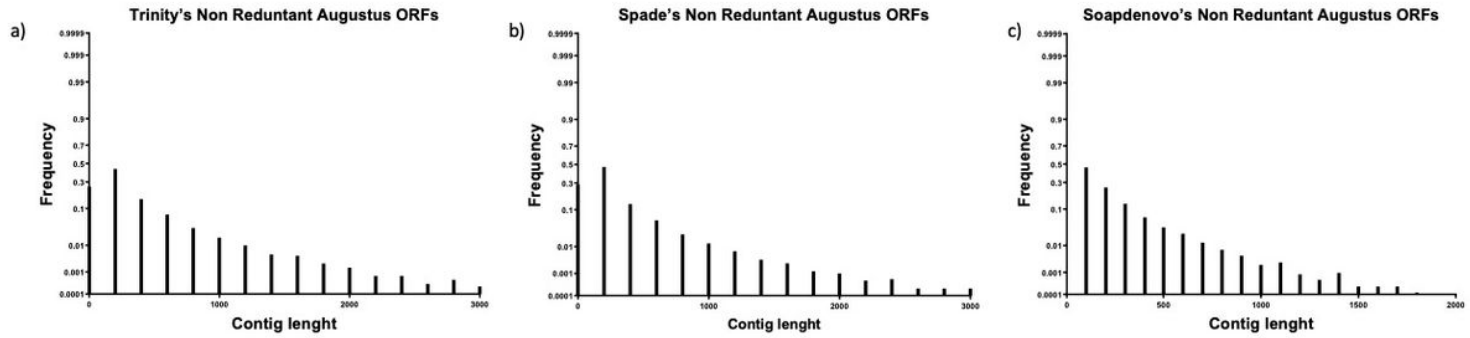

Figure 4

Histogram of the length distribution of the predicted ORFs for Trinity a), SPAdes b) and Soapdenovo c). It can be observed that the majority of predicted ORFs have less than 500 amino acids. A similar distribution is found between Trinity and SPAdes, while Soapdenovo's ORFs show a significant smaller number of ORFs with lengths between 1000 and 2000 amino acids.

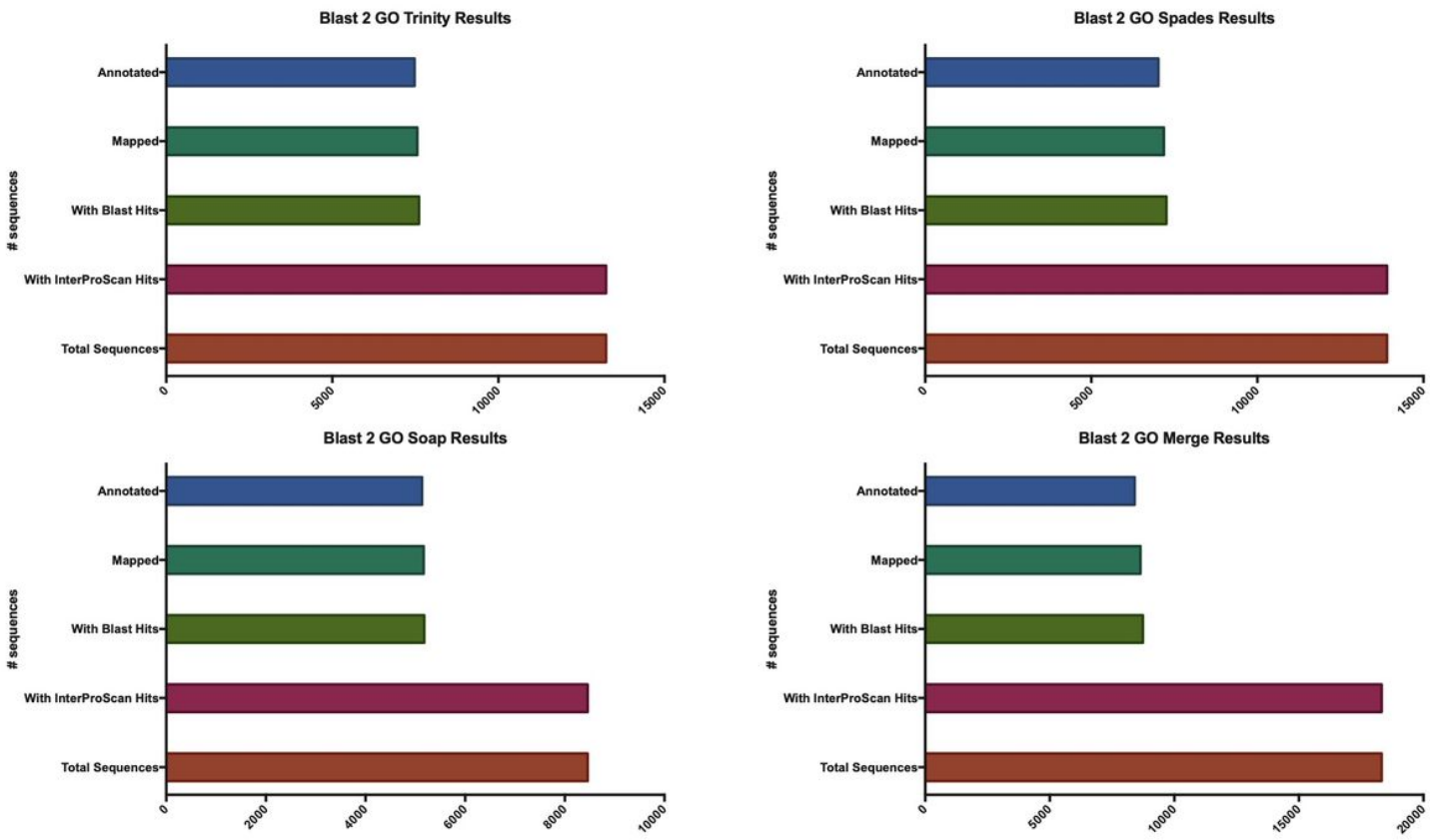

Figure 5

Summary of the annotation of the three assembled transcriptomes with the software Trinity a), SPAdes b) and Soapdenovo c); as well as the merge of the three assemblies with the elimination of the redundant sequences d). Most sequences obtained InterProScan hits; however, the annotation was highly limited by BLAST hits against the Swiss-prot database. In total numbers, the highest amount of annotated proteins corresponds to the non-redundant merge followed by Trinity, SPAdes and Soapdenovo.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.pdf
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx