Ticks and Tick-borne Diseases 14 (2023) 102209

Contents lists available at ScienceDirect

Ticks and Tick-borne Diseases

journal homepage: www.elsevier.com/locate/ttbdis

Original article

A comparison of Illumina and PacBio methods to build tick salivary gland transcriptomes confirms large expression of lipocalins and other salivary protein families that are not represented in available tick genomes

Melina Garcia Guizzo [a], Ben Mans [b,c,d], Ronel Pienaar [b], Jose M.C. Ribeiro [a,*]

[a] Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, Rockville, MD, 20852, USA
[b] Epidemiology, Parasites and Vectors, Agricultural Research Council-Onderstepoort Veterinary Research, Onderstepoort, South Africa
[c] The Department of Veterinary Tropical Diseases, University of Pretoria, Pretoria, South Africa
[d] Department of Life and Consumer Sciences, University of South Africa, Pretoria, South Africa

ARTICLE INFO

Keywords:
Tick
Salivary glands
Polymorphism
Lipocalin
Transcriptome

ABSTRACT

Tick saliva helps blood feeding by its antihemostatic and immunomodulatory activities. Tick salivary gland transcriptomes (sialotranscriptomes) revealed thousands of transcripts coding for putative secreted polypeptides. Hundreds of these transcripts code for groups of similar proteins, constituting protein families, such as the lipocalins and metalloproteases. However, while many of these transcriptome-derived protein sequences matches sequences predicted by tick genome assemblies, the majority are not represented in these proteomes. The diversity of these transcriptome-derived transcripts could derive from artifacts generated during assembly of short Illumina reads or derive from polymorphisms of the genes coding for these proteins. To investigate this discrepancy, we collected salivary glands from blood-feeding ticks and, from the same homogenate, made and sequenced libraries following Illumina and PacBio protocols, with the assumption that the longer PacBio reads would reveal the sequences generated by the assembly of Illumina reads. Using both *Rhipicephalus zambeziensis* and *Ixodes scapularis* ticks, we have obtained more lipocalin transcripts from the Illumina library than the PacBio library. To verify whether these unique Illumina transcripts were real, we selected 9 uniquely Illumina-derived lipocalin transcripts from *I. scapularis* and attempted to obtain PCR products. These were obtained and their sequences confirmed the presence of these transcripts in the *I. scapularis* salivary homogenate. We further compared the predicted salivary lipocalins and metalloproteases from *I. scapularis* sialotranscriptomes with those found in the predicted proteomes of 3 publicly available genomes of *I. scapularis*. Results indicate that the discrepancy between the genome and transcriptome sequences for these salivary protein families is due to a high degree of polymorphism within these genes.

1. Introduction

Saliva of hematophagous animals assist with the acquisition of the blood meal by its inherent anti-hemostatic, immunomodulatory and anti-inflammatory activities (Ribeiro, 1995). Salivary gland transcriptomes of these animals revealed the complexity of their sialomes: sand flies have ~ 50 different salivary polypeptides, mosquitoes and kissing bugs have 100–150 different polypeptides and hard ticks have over 1000 (Ribeiro et al., 2010; Santiago et al., 2020). The larger number found for hard ticks probably reflects the duration of the meal, which lasts from several days to weeks, while blood-sucking insects feed for a few minutes (Francischetti et al., 2009).

A constant finding in the sialomes of hematophagous animals is the presence of extended families of proteins, such as the many peptides of the odorant binding family in mosquitoes, the yellow protein family in sand flies and the lipocalins in triatomine bugs and ticks (Francischetti et al., 2009; Ribeiro et al., 2010; Santiago et al., 2020). In all these very different families, these proteins function primarily as kratagonists of hemostasis and inflammation by binding agonists of inflammation or hemostasis, such as histamine, serotonin, ADP and prostanoids (Andersen and Ribeiro, 2017). While blood suckers evolved, gene duplications initially lead to an increased salivary concentration of the kratagonists, but later also created the opportunity for the acquisition of diverse functions. For example, the kissing bug *Rhodnius prolixus*

* Corresponding author.
E-mail address: jribeiro@niaid.nih.gov (J.M.C. Ribeiro).

https://doi.org/10.1016/j.ttbdis.2023.102209
Received 5 January 2023; Received in revised form 29 May 2023; Accepted 30 May 2023
Available online 14 June 2023

nitrophorin 2, in addition to scavenge histamine, it carries nitric oxide, and inhibits blood clotting (Ribeiro, 1995). In ticks, some lipocalins acquired an anti-complement function such as *Ornithodoros moubata* complement inhibitor (OMCI) and CirpA (Braunger et al., 2022; Nunn et al., 2005).

While the extended salivary families in hematophagous insects are usually under 10–50 members per species, the lipocalin family in ticks have been estimated in the several hundreds (Mans et al., 2016). Recently, the genome sequencing of several tick species has been accomplished (Jia et al., 2020), but most of the lipocalins deducted from transcriptome work have not been found in such genomes (Ribeiro et al., 2022; Tirloni et al., 2020). These discrepancies between the number of transcriptome-derived coding sequences and genomic-derived coding sequences were also verified for other large tick salivary families such as the metalloproteases and 8.9 kDa families. What is the source of these discrepancies? Could it derive from the high polymorphism of these sequences, or is it an artifact due to chimeric coding sequences being generated by the assembly of short reads?

Here we compared the assemblies of tick salivary transcriptomes originating from Illumina (length = 150 nt, error rate ~ 0.1%) and PacBio HiFI isoseq (average length ~ 2500 nt, error rate ~0.1%) reads (Byrne et al., 2019; Rhoads and Au, 2015). It was expected that the longer PacBio reads could lead to identification of possibly mis-assembled contigs deriving from the Illumina reads. A first comparison was done between assemblies originating from *Rhipicephalus zambeziensis* ticks, a species without a known genome sequence. A second comparison was done between libraries deriving from the more studied *Ixodes scapularis* tick, which has a high-quality genome assembly (De et al., 2023). With the *I. scapularis* libraries we have also conducted PCR/amplicon sequencing experiments to validate the expression of predicted coding sequences identified by Illumina sequencing but not PacBio.

2. Material and methods

2.1. Ticks

Rhipicephalus zambeziensis was reared under standard laboratory and tick-rearing protocols at Onderstepoort Veterinary Research Institute (Heyne et al., 1987) under animal ethics AEC12.11. The colony was initiated from ticks collected from vegetation in the Marakele National Park, Limpopo Province, South Africa, in 2010. Adult ticks were fed in feeding bags fastened with glue to shaved regions on the backs of disease-free Hereford (*Bos taurus*) cattle. Roughly 100 female and 100 male ticks were fed per bag.

Ixodes scapularis adults were purchased from Oklahoma State University tick facility. These were artificially blood-fed by modifying slightly the protocol described by Oliver et al. (2016) (Oliver et al., 2016). Briefly, the feeding chambers were prepared by attaching 80-μm-thick silicone-and-rayon membranes to polycarbonate tubes fitting readily 6-well cell culture plates. Adult females and males were enclosed together ($n = 12$) into the chambers, and the females were fed to partial or full engorgement on manually defibrinated rabbit blood (Lampire, PA, USA) treated with gentamicin (10 μg/ml blood) (Sigma-Aldrich, MO, USA). Blood (5 ml) was exchanged every 12 h for a period of 10 days.

2.2. RNA extraction and library construction

Adult females of *R. zambeziensis* feeding on a Hereford ($n = 5$ ticks, 10 salivary glands for each time point) had their salivary glands dissected at 0, 6, 12, 18, 24, 36, 48, 72, 96 and 108 h and transferred to 1.0 ml RNAlater Solution (Invitrogen, Lithuania). Salivary glands were pooled together and the vials were sent to NovoGene (Beijing) for RNA extraction and sequencing. PolyA mRNA was used to construct both the 150 nt paired end Illumina NovaSeq and the Iso-Seq PacBio libraries.

Artificially membrane fed *I. scapularis* adult females of 19 varied weights (ranging 2.3–160 mg) (Supplemental Figure S1) were collected during the feeding course, and had the salivary glands dissected out and stored individually in RNAlater Solution (Invitrogen, Lithuania). Pair of salivary glands were washed in sterile PBS (KD Medical, MD, USA) pooled together and used for RNA isolation with the RNeasy® Micro kit (Qiagen, Germany). A sample of the total RNA was converted to cDNA using the SuperScript® III First-Strand kit (Invitrogen, CA, USA) for PCR experiments. Each 20 μl reaction consisted of 500 ng of total RNA. The remainder RNA was sent to Novogene (Beijing) to be used in RNA library preparation and transcriptome sequencing. PolyA mRNA was used to construct both the 150 nt paired end Illumina NovaSeq and the Iso-Seq PacBio libraries.

2.3. RNA quality control

Nanodrop was used to check RNA purity (OD260/OD280). Agarose gel electrophoresis checked for RNA degradation and potential contamination. Finally, Qubit was used to quantify RNA concentration.

2.4. PacBio workflow

After total RNA quality control, mRNA was enriched using oligo (dT) beads. The first strand cDNA was synthesized using the Clontech SMARTer PCR cDNA Synthesis Kit. The CDS Primer IIA annealed to the PolyA+ tail of transcripts, followed by first-strand synthesis with SMARTScribeTM Reverse Transcriptase. The first-strand product was diluted with Elution Buffer (EB) to an appropriate volume and subsequently used for large-scale PCR amplification. After cDNA amplification, a portion of product was used directly as a non-size selected SMRTbell library. In parallel, the rest of amplification was first selected using either BluePippin or SageELF, and then used to construct a size-selected SMRTbell library after size fractionation. DNA damage and ends were then repaired, followed by hairpin adaptor ligation. cDNAs were then transformed into circularized molecules, SMRTbell templates. Finally, sequencing primers and polymerase were annealed to SMRTbell templates. The library was then sequenced on a Sequel system.

2.5. Illumina workflow

After the QC procedures, mRNA was enriched using oligo(dT) beads. First, the mRNA was fragmented randomly by adding fragmentation buffer, then the cDNA is synthesized by using mRNA template and random hexamers primer, after which a custom second-strand synthesis buffer (Illumina), dNTPs, RNase H and DNA polymerase I are added to initiate the second-strand synthesis. Second, after a series of terminal repair, A ligation and sequencing adaptor ligation, the double-stranded cDNA library was completed through size selection and PCR enrichment. The library was then sequenced on a Illumina system.

2.6. Bioinformatic procedures

Illumina reads were stripped of primer sequences and low quality values (less than 20) using the program TrimGalore (Krueger, 2016). Reads were normalized with the program “insilico_read_normalization.pl” from the Trinity package (Haas et al., 2013), allowing a kmer maximum coverage of 50 and a minimum coverage of 2. The programs Trinity and Abyss (Simpson et al., 2009), with -k values 25, 35…95 were run on the normalized reads. The resulting assemblies were combined using the program CD-Hit-EST (Li and Godzik, 2006). PacBio reads were assembled with the pipeline smrt-analysis software from Pacific Biosciences downloaded from https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/. Resulting predicted coding sequences (CDS) were compared by blastx to an Ixodida database (from available protein sequences from Ixodida found on the NCBI NR and TSA databases) and to the UniprotKB (Boutet et al., 2016) databases. The

Table 1
Illumina and PacBio library reads specifications for *Rhipicephalus zambeziensis* and *Ixodes scapularis* salivary gland libraries.

R. zambeziensis Illumina file	Format	Number of reads	Sum Lengths	Minimum length	Average length	Maximum length	Error (%)	Q20 (%)	Q30 (%)
Rzam_1.fq.gz	FASTQ	44,514,867	6677,230,050	150	150	150	0.02	97.99	94.63
Rzam_2.fq.gz	FASTQ	44,514,867	6677,230,050						
Total bases			1.34E+10						
PacBio file	**Format**	**Number of reads**	**Sum Lengths**	**Minimum length**	**Average length**	**Maximum length**			
Rzam-pacbio.fa	FASTA	9982,978	22,279,509,461	50	2231.70	119,909			
Total bases:			2.23E+10						
I. scapularis									
Illumina file	**Format**	**Number of reads**	**Sum Lengths**	**Minimum length**	**Average length**	**Maximum length**	**Error (%)**	**Q20 (%)**	**Q30 (%)**
Iscap_1.fq.gz	FASTQ	22,130,929	3319,639,350	150	150	150	0.02	98.12	94.79
Iscap_2.fq.gz	FASTQ	22,130,929	3319,639,350						
Total bases			6.64E+09						
PacBio file	**Format**	**Number of reads**	**Sum Lengths**	**Minimum length**	**Average length**	**Maximum length**			
Iscap-pacbio.fa	FASTA	96,272	271,743,802	96	2822.70	11,231			
Total bases:			2.72E+08						

CDS covering more than 2/3 of their matching sequences were selected and their translated sequence was recovered. We also translated all CDS and recovered all peptides with 40 or more amino acids (aa) starting with a methionine. These peptides were submitted to the signalP program (v. 3.0) (Nielsen, 2017). If two or more 40mer peptides were found in the same CDS, the result of the most proximal to the 5′ location on the CDS was used to identify the CDS as coding for a putative secreted peptide/. The CDS containing the predicted signal indicative of a secreted protein were translated and added to the predicted peptides fasta file. We thus obtained CDS based on homology to known proteins, and those that had a predicted signal peptide. To remove redundancy, the peptide fasta file was clustered at 98% identity with the program CD-Hit (Li and Godzik, 2006). The RSEM software (Li and Dewey, 2011) was used to map the reads to the assembled databases and obtain the Transcripts per Million (TPM) value for each CDS. The TickSialoFam (TSF) database (Ribeiro and Mans, 2020) version 2.0 (Mans et al., 2022) was used to identify the transcripts associated with secreted salivary products including lipocalins and metalloproteases, using a threshold of acceptance of $>$ 90% coverage and minimal e value of $1e^{-4}$. The ribosomal protein class was identified by those CDS having best match to the PFAM databases sequences annotated as "Ribosomal proteins" and having a coverage larger than 90% of the blast reverse position model. The KOG "Posttranslational modification, protein turnover, chaperones" were identified by matches having over 90% coverage to the KOG class thus annotated.. The BUSCO (Benchmarking Universal Single-Copy Orthologs) program version 5.0.0 was run to identify the completeness of the transcriptomes using as input the predicted peptide coding sequences (clusterized to 98% identity – see above) and using the arachnida_odb10 lineage dataset. The program Tandem Repeat Finder (TRF) (Benson, 1999) was used to detect tandem repeats on the CDS. These were mapped to the spreadsheets indicating the repeat length, repeat size and sum of their scores.

The predicted protein sequences of the Wikel (WI) genome assembly of *I. scapularis* (Gulia-Nuss et al., 2016) were downloaded from VectorBase (Giraldo-Calderón et al., 2021), those from the Ise6 cell line genome assembly (I6) (Miller et al., 2018) and the Maryland University assembly of Feb 2021 (MD) (De et al., 2023) were downloaded from the NCBI genome site https://www.ncbi.nlm.nih.gov/genome/?term=txid6945%5bOrganism:noexp]. The predicted CDS from these three genomes were used as input to the program BUSCO (Simão et al., 2015) (version 5.0) against the data set of arachnida_obd10 to investigate the degree of genome completion.

2.7. Validation of Ixodes scapularis lipocalin sequences

The expression of lipocalins identified by Illumina (IL) but not by PacBio (PB) was validated by PCR. *Ixodes scapularis* ribosomal protein L6 (Rpl6) gene was used as an internal control to validate the cDNA synthesis and the PCR procedure. PCR was performed on 1 μl of the undiluted cDNA solution in 20 μl of DreamTaq Green PCR master mix 2x (Thermo Fisher Scientific, MA, USA) using primers designed for the different lipocalins (Supplementary Table 1). PCR was carried out in a 96 well Thermal Cycler (Apllied Biosystems, MA, USA) with an initial denaturation of 95 °C (3 min), followed by 35 cycles of 95 °C (30 s), 55 °C (30 s), and 72 °C (1 min), and a final extension of 72 °C (10 min). Because the target lipocalins were found to have low expression (TPM $<$ 50) a second PCR was performed using 1 μl of the first PCR product as the template. PCR products were electrophoresed on 1% agarose gel (Thermo Scientific, IL, USA) containing Sybr® Safe DNA gel stain (Invitrogen, CA, USA). Products were purified using the PureLink Quick Gel Extraction and PCR Purification Combo Kit (Invitrogen, CA, USA), and had the identity confirmed by sequencing (Eurofins genomics, KY, USA).

2.8. Determination of single nucleotide polymorphisms (SNPs) on Ixodes scapularis CDS

The reads provided on NCBI Bioproject PRJNA731189 were downloaded and concatenated to produce a single pair of fastq paired files containing a total of 9.00044 $E+11$ nt, representing a 401 coverage of the estimated genome size for *I. scapularis* of 2.24 Gb (Miller et al., 2018). These were mapped to the CDS obtained from the MD genome assembly (De et al., 2023) using bowtie2 (Langmead and Salzberg, 2012). The Variable Call Format (VCF) files containing the SNPs information were obtained with BCFtools (Danecek et al., 2021). The VCF files were parsed and hyperlinked to each CDS within the spreadsheet. A program written in visual basic extracted the SNPs that had a minimum coverage of 5 x, rewriting the CDS and determining whether the SNP lead to a synonymous or non-synonymous mutation. These results were also mapped to the spreadsheet.

2.9. Data availability

This project has been registered with the National Center for Biotechnology Information (NCBI) under project number PRJNA905811 (*Ixodes scapularis*), BioSample SAMN32356505 and sequence read archive (SRA) SRR22900767 (Illumina) and SRR22900766 (PacBio). The derived CDS from the assemblies were deposited to the Transcriptome shotgun annotation (TSA) database of the NCBI under accessions GKHW01000001-GKHW01005950. The data from the *Rhipicephalus zambeziensis* were deposited under project accession PRJNA905810, BioSample SAMN32356046, and SRA SRR22891217 (Illumina) and SRR22891216 (PacBio). The derived CDS from the assemblies were deposited to the Transcriptome shotgun annotation (TSA) database of the NCBI under accessions GKHV01000001-GKHV01003807.

Table 2
Comparison of Illumina (ILL) vs. PacBio (PB) transcriptome assembly from the salivary glands of *Rhipicephalus zambeziensis* according to the number of coding sequences (CDS) identified for each group of transcripts coding for secreted products (identified by the TSF2.0 database). The ratio of the number of identified CDS by the PB library compared to the ILL library is given in the column PB/ILL. The average length of the CDS for each group as well as the average sum of the Tandem Repeat Finder (TRF) scores and average RPKM's are also shown. Displayed are the results where the ILL assembly identified 10 or more CDS per group.

Group	Number of CDS Illumina	Number of CDS Pacbio	PB/ILL	Average CDS length	Average TRF scores	Average RPKM
Lipocalin	343	43	0.13	627	0.48	174.48
Kunitz	111	17	0.15	575	2.90	156.12
Metalloprotease	89	80	0.90	1488	7.08	46.21
8.9kDa	55	5	0.09	383	0.00	57.71
GRP	37	75	2.03	1114	161.14	1235.08
Cystatin	35	3	0.09	373	1.94	43.91
Ixodegrin	32	8	0.25	390	1.72	25.91
TIL	32	1	0.03	1697	9.50	210.88
BTSP	31	10	0.32	1112	11.77	250.93
Evasin	30	0	0.00	394	5.77	46.70
Mucin	26	14	0.54	798	0.00	101.38
Transposon	26	11	0.42	1545	2.08	11.23
Derf7/JHBP	25	3	0.12	672	0.00	14.28
Metalloproteoid	19	8	0.42	714	0.00	126.16
Defensin	19	2	0.11	296	0.00	343.95
Endothelin converting enzyme	18	15	0.83	2184	3.33	27.33
Serpin	18	13	0.72	1177	0.00	16.17
Carboxypeptidase_inhibitor	17	0	0.00	307	0.00	40.24
Cytotoxin	14	1	0.07	718	14.36	16.00
Rapp-40–287	14	1	0.07	478	0.00	47.71
Down syndrome cell adhesion molecule	13	11	0.85	1377	0.00	123.08
Serine carboxypeptidase	12	11	0.92	1399	0.00	55.17
Toll-like	12	2	0.17	1496	0.00	27.55
23–24kDa	10	7	0.70	785	53.70	42.70
JHBP	10	1	0.10	754	0.00	11.90

3. Results

The resulting number of reads per library for both species are displayed on Table 1. We will analyze first the results for *Rhipicephalus zambeziensis* libraries, followed by those for *Ixodes scapularis*.

3.1. Rhipicephalus zambeziensis libraries

The assemblies of the *R. zambeziensis* libraries generated 65,789 predicted peptides from the IL (Supplemental spreadsheet 1) and 143,241 peptides from the PB library (Supplemental spreadsheet 2). Their BUSCO completeness values were 63.9% complete BUSCOs for the PB and 81.4% for the IL library (Supplemental table S2). Using the results from RPSBLAST of the predicted peptide sequences against the TSF v2.0 database, 1256 peptide sequences from the IL assembly were found while the PB library produced 488 peptides (Table 1 and Supplemental spreadsheets 1 and 2). When comparing the number of CDS retrieved by the assembly from each library, we found 25 groups of proteins having 10 or more members identified by the IL library. In all cases, except for the GRP (Glycine rich proteins) group, the PB library produced less hits than the IL library (Table 2). Notably, the IL library identified 343 lipocalins, against 43 identified by the PB library. Forty of the 43 lipocalin predicted sequences from the PB assembly were matched by the Illumina assembly at >90% coverage and >95% identity. One hundred and eleven Kunitz proteins were predicted by the IL library, while only 14 were predicted by the PB library. All 14 Kunitz predicted sequences from the PB assembly were matched by the IL assembly at =>99% coverage and => 99% identity. Eighty-nine metalloproteases were identified by the IL assembly, against 81 from the PB. Fifty-six of the 81 metalloproteases from the PB assembly were matched by IL assembly at =>95% identity and=>90% coverage. The 8.9 kDa family was represented by 55 peptides from the IL and 5 from the PB assembly. All the five 8.9 kDa family transcripts assembled by the PB reads matched IL transcripts at 99 or 100% identity, 3 of which had a coverage of 100%, one of 98% and one of 81% coverage. While the IL library identified 35 cystatins, the PB library identified 3, all of which were matched by the Illumina transcripts at 100% coverage and => 97% identity. Exceptionally, the PB library identified 74 transcripts coding for GRP members, against 37 from the IL assembly.

Overall, the PB assembly identified less transcripts coding for putative secreted proteins than the IL assembly, except for the GRP class. We wondered whether this PB efficiency for the GRP group of proteins could be due to the existence of repeats within their coding sequences. To investigate this hypothesis, the program TRF was run on the IL assembled transcripts identifying the GRP group as having the highest average score of tandem repeats. It is also possible that the PB library was more efficient in assembling transcripts with higher read representation. Indeed, the average TPM score of 26 lipocalins identified by both the PB and IL libraries (at > 95% coverage and > 95% id) was 330, while the 317 lipocalins solely identified by the IL assembly had an average TPM of 22.05. A Mann Whitney test indicated the difference to be highly significant ($P < 0.0001$). Because the GRP class coded for long CDS, we also considered that the PB methodology could be more efficient in assembling long CDS. To verify all these hypotheses, we did a regression analysis of the ratios of CDS transcripts identified by the PB assembly compared to the IL assembly, as a function of the average CDS length, average TRF sum of scores and average number of reads per thousand base pairs per million (RPKM) for each group of secreted sequences that had 10 or more CDS identified by the IL assembly (Table 2). Results showed a significant correlation for all 3 variables (Fig. 1).

An additional observation comparing the PB and IL assemblies revealed that the IL assembly matched 3720 sequences from the UniprotKB database with > 90% coverage and an eval < 1e-15, while the PB assembly matched 2853 in the same conditions.

At this point we considered that the PacBio library would not solve the question of whether the apparently large number of lipocalins and other high numbered families of tick salivary proteins were an artifact of assembly. We then decided to rerun the experiment but this time using *I. scapularis*, which has published genomes and allowing matching of the assembled transcriptomes predictions to the predicted proteome. Furthermore, we saved aliquots of the homogenate used to construct the libraries in order to attempt PCR recovery of the predicted IL-derived lipocalins not detected by the PB library.

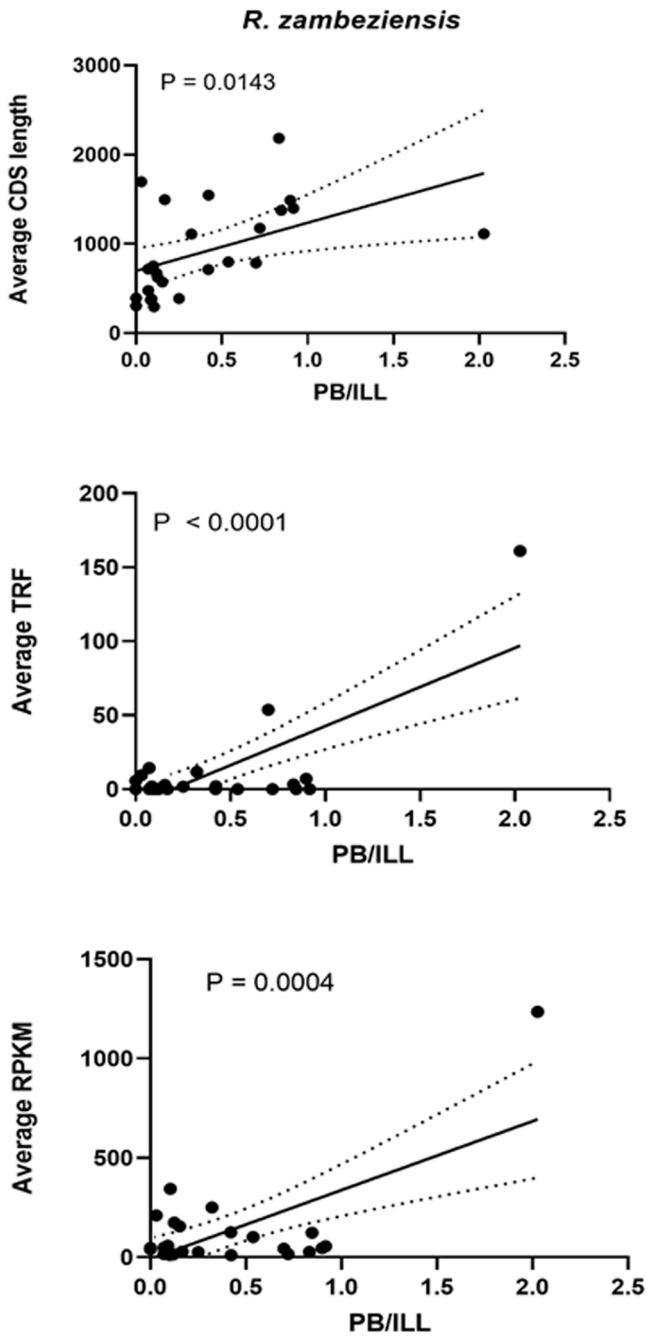

Fig. 1. Correlations between the ratio of the number of salivary transcripts from *Rhipicephalus zambeziensis* coding for secreted proteins in 25 groups of proteins as a function of the groups' RPKM, tandem repeat scores and coding sequences length. The line represents the linear regression of the data, and the dotted lanes represents the 95% confidence interval of the slope. The P values are indicated in each graph.

3.2. *Ixodes scapularis libraries*

A description of the comparison of both libraries following the same format used for the *R. zambeziensis* libraries follows: The assemblies of the *I. scapularis* libraries generated 27,159 predicted peptides from the IL (supplemental spreadsheet 3) and 52,302 peptides from the PB library (supplemental spreadsheet 4). Their BUSCO completeness values were: 89.8% complete for PB and 88.4% for the IL library (Supplemental Table 1). Using the results from RPSBLAST of the predicted peptide sequences against the TSF v2.0 database, 1000 peptide sequences from the IL assembly were found while the PB library produced 959 peptides (Supplemental spreadsheets 3 and 4). When comparing the hits from each library to 26 protein groups that had 10 or more CDS predicted by the IL assembly, the IL library uncovered more CDS than the PB library for 17 groups. Notably, 132 lipocalin CDS's were revealed by the IL assembly, against 82 identified by the PB assembly. Of these 82 PB lipocalins, 50 matched IL lipocalins with at least 90% coverage and 90% id. However, the PB assembly indicated more CDS's than the IL library for 10 of the 26 groups, including the groups Metalloprotease, Serine protease, endothelin converting enzyme and 5′ nucleotidase (Table 3). A regression analysis of the ratios of CDS transcripts identified by the PB assembly compared to the IL assembly, as a function of the average CDS length, average TRF sum of scores and average RPKM for each group of secreted sequences that had 10 or more CDS identified by the IL assembly (Table 3) revealed a significant correlation for the PB/IL ratio as a function of CDS length, but no significance for the other 2 variables (Fig. 2).

3.3. *Validation of lipocalins found only by the assembled IL library*

Fifty lipocalin sequences assembled from the IL library were not matched by transcripts from the PB library (Table 3). These had relatively low TPM's, not exceeding a value of 50. To investigate whether these sequences were artifacts of the assembling methodology, we have selected the 9 most expressed to be validated by PCR. All the target transcripts could be amplified (Fig. 3), and their identities were confirmed by sequencing (Supplemental file 1). This result indicates that the lipocalins identified by the IL method can be trusted. However, it is still possible that some of the non-PCR tested sequences are misassembled.

3.4. *The salivary lipocalins of Ixodes scapularis*

To further investigate the lipocalin diversity in *I. scapularis*, we retrieved all the lipocalins predicted from the three published genome assemblies of this tick, then joined these with the predicted lipocalins found in our PB/IL transcriptome (supplemental spreadsheet 5).The transcriptome sequences, but not the genomic-derived sequences, were clusterized to 95% identity using the program CD-HIT (Li and Godzik, 2006). Accordingly, we obtained 333 lipocalin sequences from the MD genome, 270 from the I6 genome, 31 from the WI genome and 205 from the transcriptome assembly (Supplemental spreadsheet 5). Of these 830 sequences, 694 were unique. These 694 lipocalin sequences were then compared by blastp to each other, and their matches according to different degrees of sequence coverage and identities were tallied. Venn diagrams (Fig. 4) demonstrate the spread of these 694 sequences over the four groups of sequence provenance. Notice that with an identity (id) level of 90% and coverage (cov) of 95% the MD genome indicates 205 unmatched lipocalin sequences (Fig. 4A), decreasing to 135 at an id level of 90% and cov. of 90% (Fig. 4B), 6 when id = 80% and cov = 90% (Fig. 4C) and to 2 when coverage was reduced to 75% (Fig. 4D).

3.5. *The metalloproteases of I. scapularis*

A similar analysis was performed for the metalloproteases of *I. scapularis* (supplemental spreadsheet 6). We recovered 320 sequences from the predicted proteins of the 3 genomes (172 from the MD, 141 from the I6, and 7 from the WI), plus 152 sequences from our combined transcriptomes (TR) that were identified as metalloproteases by the TSF database. Again, more than 50% of the sequences were unique when compared to each other at 90% coverage and 95% identity (Fig. 4B).

3.6. *The ribosomal proteins and proteins associated with post-translation modification of I. scapularis*

To contrast the results obtained from the lipocalin and

Table 3
Comparison of Illumina (ILL) vs. PacBio (PB) transcriptome assembly from the salivary glands of *Ixodes scapularis* according to the number of coding sequences (CDS) identified for each group of transcripts coding for secreted products (identified by the TSF2.0 database). The ratio of the number of identified CDS by the PB library compared to the ILL library is given in the column PB/ILL. The average length of the CDS for each group as well as the average sum of the Tandem Repeat Finder (TRF) scores and average RPKM's are also shown. Displayed are the results where the ILL assembly identified 10 or more CDS per group.

Group	Number os CDS ILL	Number os CDS PacBio	PB/ILL	Average CDS length	Average TRFs	Average RPKM
Lipocalin	132	82	0.62	809.81	0.39	176.16
Salp15/Ixostatin	123	41	0.33	496.29	3.23	119.66
Kunitz	109	45	0.41	654.03	8.14	187.50
Metalloprotease	74	152	2.05	1434.61	1.36	166.27
10kDa-WC	45	2	0.04	407.84	0.00	78.76
Serine protease	26	54	2.08	1261.12	27.58	36.54
Metalloproteoid	24	7	0.29	705.67	5.25	59.58
Evasin	22	12	0.55	563.86	5.86	271.09
Transposon	21	48	2.29	1776.00	12.48	8.24
Serpin	19	23	1.21	1275.21	0.00	158.11
8.9kDa	17	10	0.59	582.41	8.12	374.00
Cytotoxin	17	11	0.65	879.65	0.00	70.12
GRP	17	20	1.18	2019.59	10.24	1518.88
BTSP	16	10	0.63	1043.25	37.13	269.75
Derf7/JHBP	16	13	0.81	747.69	0.00	97.69
Toll-like	15	23	1.53	1226.20	3.33	82.40
Down syndrome cell adhesion molecule	14	27	1.93	1485.50	11.86	33.57
Ixodegrin	13	3	0.23	557.23	0.00	51.54
Endothelin converting enzyme	13	47	3.62	2271.69	0.00	445.85
ISAC	12	5	0.42	696.50	4.75	320.67
Ficolin/Ixoderin	12	10	0.83	886.08	0.00	82.25
23–24kDa	12	11	0.92	1005.42	13.75	319.42
ML_domain	11	7	0.64	608.36	0.00	111.36
5′nucleotidase	11	23	2.09	1781.45	0.00	88.36
TIL	10	2	0.20	605.00	31.20	143.30
Mucin	10	35	3.50	1331.30	62.80	144.00

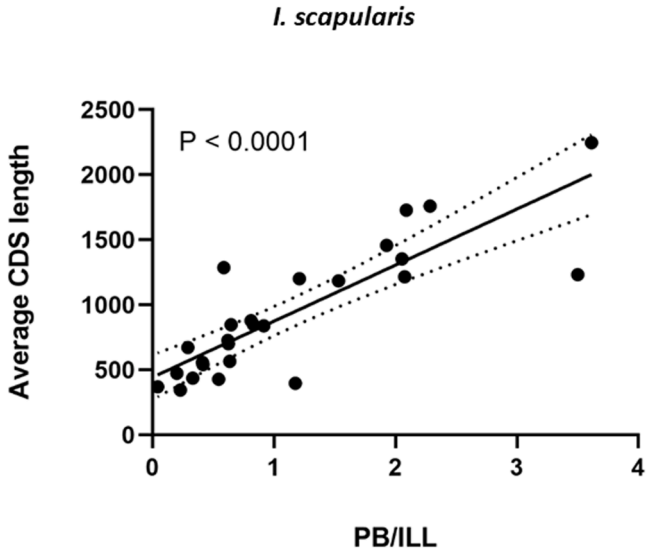

Fig. 2. Correlations between the ratio of the number of salivary transcripts from *Ixodes scapularis* coding for secreted proteins in 25 groups of proteins as a function of the groups' coding sequences length. The line represents the linear regression of the data, and the dotted lines represents the 95% confidence interval of the slope. The P value is .indicated on the figure.

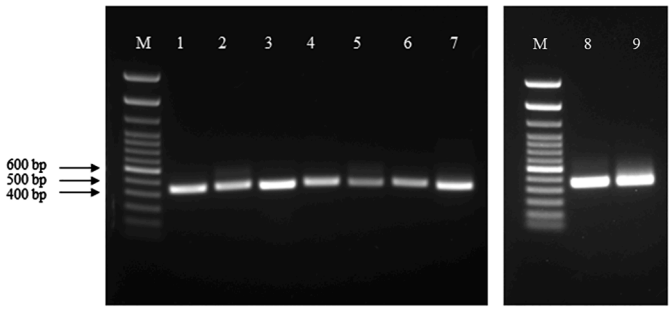

Fig. 3. RT-PCR amplification products of *Ixodes scapularis* lipocalins genes visualized by 1% agarose gel. M: 100 bp DNA marker; 1: transcript 64,033; 2: transcript 34,643; 3: transcript 43,377; 4: transcript 96,024; 5: transcript 108,517; 6: transcript 34,644; 7: transcript 39,780; 8: transcript 48,458; 9: transcript 99,871.

metalloprotease, which are related to secreted salivary protein families, to a set of housekeeping conserved proteins, we retrieved 436 protein sequences that gave matches to PFAM motifs of ribosomal proteins with > 90% coverage (119 from MD, 125 from I6, 82 from WI and 110 from our combined transcriptome) (Supplemental spreadsheet 7 and Fig 4C) and 1885 sequences producing matches of over 90% coverage to the KOG class "Posttranslational modification, protein turnover, chaperones" (supplemental spreadsheet 8) (618 from MD, 658 from I6, 215 from WI and 384 from the TR. Notice that the overall percentage of unique sequences in this group of proteins is much smaller, mainly within the MD and I6 sequences (Fig. 4C and 4D).

A two-way analysis of variance comparing the percentage of unique proteins from the lipocalin and metalloprotease groups found at 90% coverage and 95, 90 and 80% identities with the ribosomal and KOG groups indicated significant differences for the 95 and 90% identities, but not at 80% identity (Table 4). For these analysis we used the MD, I6 and TR data sets, omitting the WI set as it had small sequence representation for the lipocalin and metalloprotease groups of sequences.

To obtain an additional insight into the polymorphism of the lipocalins and metalloproteases, we took advantage of the recently sequenced genomes of *I. scapularis* (Schoville et al., 2023), which publicly disclosed Illumina short reads from 92 ticks collected in diverse areas of the U.S.A. Although the average coverage was relatively small (4.24 fold), the combined data represented a genome coverage of 402 x from geographically diverse organisms. The combined short read archived (SRA) data was mapped to the CDS from the MD genome and the SNPs were determined within each CDS (Supplemental spreadsheet

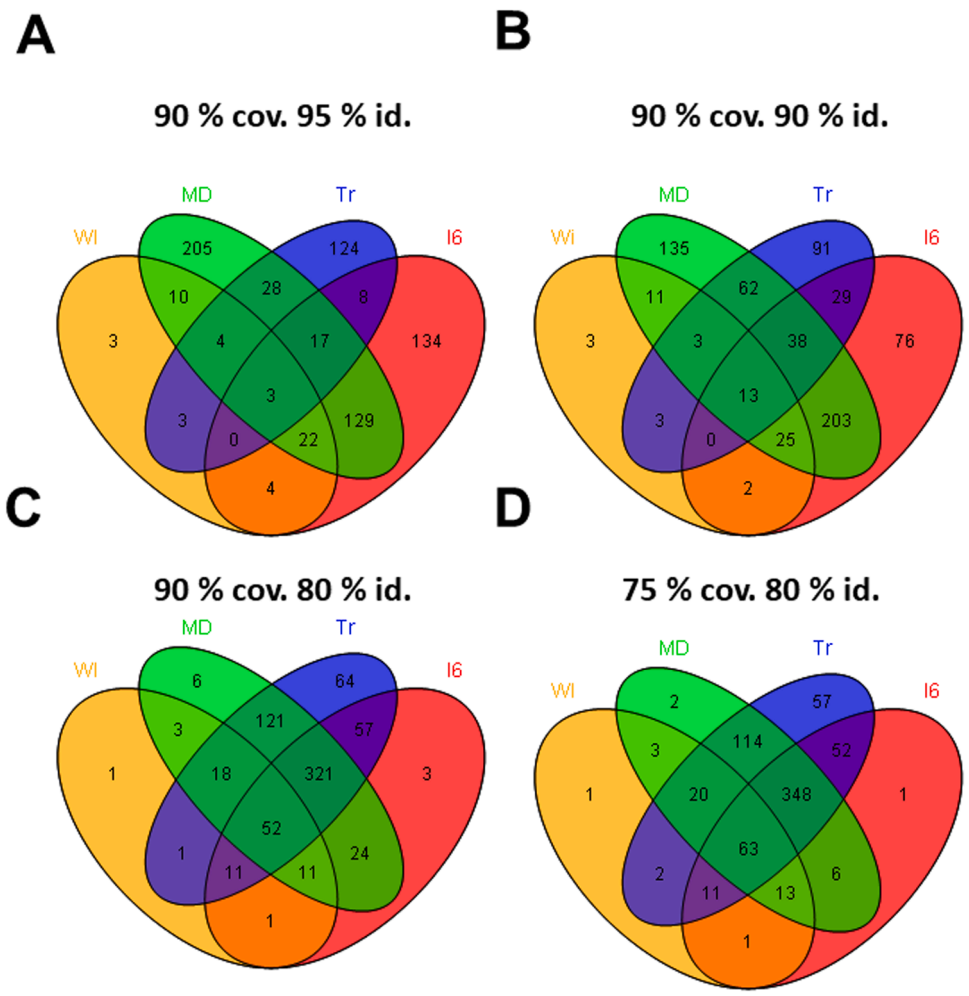

Fig. 4. Venn diagram of 694 unique lipocalin sequences deriving from three *I. scapularis* genome assemblies (MD, I6, and WI) together with those derived from the PacBio and Illumina libraries (Tr), clustered at 90% coverage and 95% identity (A), 90% cov. and 90% id (B), 90% cov. and 80% id. (C), and 75% cov. and 80% id. (D).

Table 4
Results from two-way analysis of variance tests comparing the percentual values of unique proteins from the I6, MD and current transcriptome from the lipocalins and metalloproteases groups, compared to the percentual levels of unique ribosomal or KOG groups clusterized at 90% coverage and 95, 90 and 80% identities.

Origin	Lipocalins 90/95	Ribosomal proteins 90/95	Source of variation	% of total variation	P value	P value summary
I6	59.82	1.6				
MD	64.87	0.84	Row Factor	7.894	0.093	ns
Tr	88.57	15.45	Column Factor	91.3	0.0044	**
Origin	Lipocalins 90/90	Ribosomal proteins 90/90	Source of Variation	% of total variation	P value	P value summary
I6	33.93	1.60				
MD	42.72	0.00	Row Factor	15.39	0.1633	ns
Tr	65.00	12.73	Column Factor	81.6	0.0179	*
Origin	Lipocalins 90/80	Ribosomal proteins 90/80	Source of Variation	% of total variation	P value	P value summary
I6	1.34	1.60				
MD	1.90	0.00	Row Factor	63.74	0.2653	ns
Tr	45.71	11.82	Column Factor	13.24	0.3957	ns
Origin	Metalloproteases 90/95	KOG 90/95	Source of Variation	% of total variation	P value	P value summary
I6	54.61	1.27				
MD	60.47	0.56	Row Factor	6.584	.1481	ns
Tr	78.95	9.58	Column Factor	92.27	.0061	**
Origin	Metalloproteases 90/90	KOG 90/90	Source of Variation	% of total variation	P value	P value summary
I6	31.91	1.27				
MD	40.70	0.56	Row Factor	28.96	0.1639	ns
Tr	48.68	9.58	Column Factor	65.37	0.0408	*
Origin	Metalloproteases 90/80	KOG 90/80	Source of Variation	% of total variation	P value	P value summary
I6	2.13	0.69				
MD	3.49	0.56	Row Factor	69.96	.2069	ns
Tr	42.11	4.55	Column Factor	11.79	.3736	ns

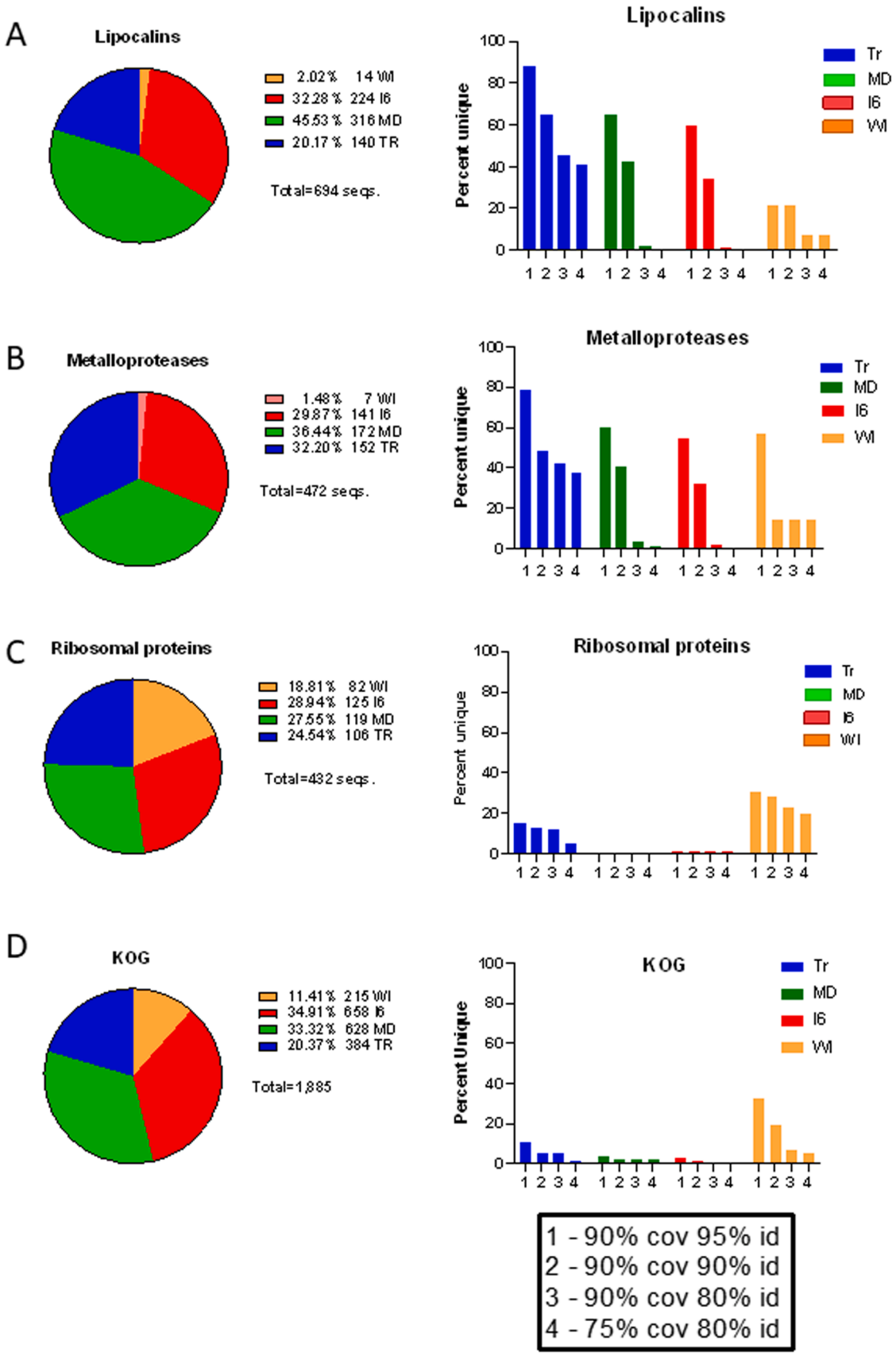

Fig. 5. Pie charts indicating the number of lipocalins (A), metalloproteases (B), ribosomal proteins (C) and Post-translational proteins retrieved from three genome assemblies from *I. scapularis,* plus the ILL + PB transcriptome assemblies. The bar charts indicate the percentage of unique sequences from each assembly under four different thresholds indicated in the bottom of (D).

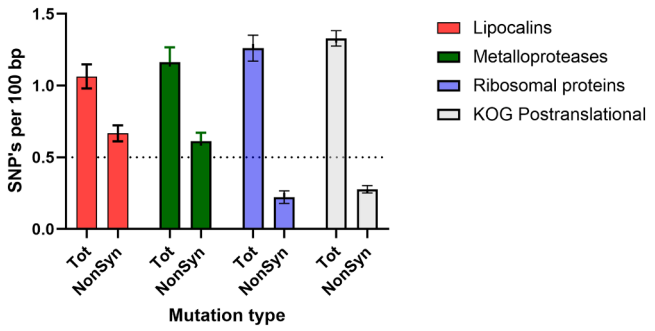

Fig. 6. Bar charts of the number of single nucleotide polymorphisms (SNPs) per 100 nt on coding genes for Lipocalins, Metalloproteases, Ribosomal proteins and KOG Post translational classes. Both the total number of SNPs and the number of SNPs leading to a non-synonymous mutation are shown. The dashed line represents a SNP value of 0.5 SNPs per 100 bp.

9), when the linear coverage of the CDS was larger than 90% and the read depth was larger than 50, Fig. 6 displays the total recovered SNPs as well as the non-synonymous SNPs from the lipocalins, metalloproteases, ribosomal proteins, and the KOG post-translational class of CDS. A Kruskal Wallis test comparing how the median values of the non-synonymous SNPs differ from 0.5 showed non-significant values for the lipocalins and metalloproteases, and significant values for the ribosomal proteins and KOG posttranslational classes, indicating a higher tolerance for non-synonymous mutations in the lipocalin and metalloprotease classes (Fig. 6). These analyses are in accordance with those reported in (Schoville et al., 2023), which found increased allele frequency in some salivary gland genes, including metalloproteinases.

4. Discussion

The primary objective of this work was to assess whether the large number of lipocalins found in IL-based sialotranscriptomes was artifactual. We then compared the assemblies of IL and PB transcriptomes made from the same cDNA aliquot of homogenized salivary glands from the tick *R. zambeziensis*, and verified that the PB assembly was less exhaustive than the IL in the recovery of unique transcripts, but nonetheless it produced transcripts that were not found in the IL assembly. Notably, the PB assembly had better performance than the IL assembly with transcripts having high TPM values, larger frequency of internal repeats, and longer coding sequences (Fig. 1).

We then decided to repeat the work, this time starting from salivary gland homogenates from the tick *I. scapularis*. The PB library had better performance in this instance and contributed to many transcripts not identified in the IL assembly. These were significantly associated with longer coding sequences (Fig. 2). Notice that the BUSCO completeness value for the IL library of *I. scapularis* is similar to the IL library of *R. zambeziensis,* but the BUSCO scores for both PB libraries differ considerably (63.9% for *R. zambeziensis*, 89,8% for *I. scapularis*). This suggests sequencing depth of the *R. zambeziensis* PB library was not optimal and may explain (to some extent) the differences in numbers of detected transcripts between both methods. What solved our initial question, whether the IL lipocalins were a product of artifactual assembly, were the PCR results and their product sequences that validated the predicted assembled sequences (Fig. 3).

Having found support for the existence of a large number of salivary lipocalins in tick sialotranscriptomes, we compared our transcriptome results with the genomic predictions of lipocalins from the three public genomic assemblies of *I. scapularis* in order to have an insight into the possible mechanisms of this diversity. Venn diagrams displaying the degree of sequence identity among the three genomic predictions as well as the transcriptome sequences showed a large degree of unique sequences (or alleles) (Fig. 4 and 5A). The same pattern was verified for the expanded salivary family of metalloproteases (Fig 5B). However, when performing the same analysis using housekeeping proteins (Ribosomal proteins and those belonging to the KOG class "Post-translational modification, protein turnover, chaperones"), it was observed that the number of unique sequences (Fig. 5C and 5D) is significantly smaller (Table 4), suggesting that salivary protein families may occupy a genomic region that favors fast evolution, such as occur with the human MHC locus (Andersson and Mikko, 1995; Carrington, 1999; Trowsdale and Parham, 2004). This was confirmed when the reads from 92 genomes of *I. scapularis* were mapped to the CDS predicted by the MD assembly (Fig. 6). This analysis supports the idea that the lipocalin and metalloprotease genes belong to highly polymorphic loci, thus the number of alleles in a population vastly exceeds the number of genes from a single organism.

An intriguing finding was the relatively poor yield of lipocalins and metalloproteases in the WI genome. While the WI genome is known to be partial, having covered 57% of the genome (Gulia-Nuss et al., 2016), we recovered 82 ribosomal proteins, which is 68% of the 119 found in the MD genome or 65.6% of the 125 found in the I6 genome. Similarly, we recovered 118 sequences matching the KOG class "Posttranslational modification, protein turnover, chaperones" which is 33% of the 355 seqs recovered from the MD genome, and 35% of the 340 seqs recovered from the I6 genome. However, only 14 lipocalins were found in the WI genome, which represents only 4% of the 316 lipocalins found in the MD genome, or 6% of the 224 lipocalins found in the I6 genome. The equivalent values for the seven metalloproteases found in the WI genome are again 4% of the metalloproteases found in the MD genome and 5% of the metalloproteases found in the I6 genome. The lipocalin and metalloprotease-coding genes in the WI genome is nearly 10 x depleted when compared to the MD and I6 genomes. Why are these salivary proteins so depleted in the WI genome? The *I. scapularis* Wikel strain used for extraction of the DNA sequenced for the genome assembly originated from approximately 30 pairs of male and female ticks that were bred in the laboratory for 12 generations without additions of field collected material (Gulia-Nuss et al., 2016). This colony fed on naïve mammals and thus there was no immune pressure from the host regarding the tick salivary antigenic material, possibly leading to a loss of genes or haplotypes related to a highly polymorphic locus. This is also supported by the genomic BUSCO results, which showed 98.9, 96.5 and 86.8% complete BUSCOs for the MD, I6 and WI genomes and 30.5, 35.3 and 1.6% duplicated BUSCOs for the same genomes, respectively, indicating a loss on the WI genome for duplicated genes.

5. Conclusion

Differences currently observed between transcriptomes, genomes or sequencing approaches are multifactorial, including differences in the technologies and analysis methods used, genetic and expression differences between strains of the same species or possible elevated mutation rates in highly expressed protein families. Taking all this in consideration, this analysis supports the idea that the lipocalin and metalloprotease genes belong to highly polymorphic loci, thus the number of alleles in a population vastly exceeds the number of genes from a single organism. The existence of robust tick omic data was crucial and has supported the results of this work.

Funding

JMCR and MGG were supported by the Intramural Research Program of the National Institute of Allergy and Infectious Diseases (Vector-Borne Diseases: Biology of Vector Host Relationship, Z01 AI000810-18). BM was supported in part by the National Research Foundation of South Africa (Grant Number: 137966).

CRediT authorship contribution statement

Melina Garcia Guizzo: Conceptualization, Validation, Formal analysis, Investigation, Writing – review & editing, Visualization. **Ben Mans:** Conceptualization, Formal analysis, Resources, Writing – review & editing, Funding acquisition. **Ronel Pienaar:** Investigation, Writing – review & editing. **Jose M.C. Ribeiro:** Conceptualization, Software, Formal analysis, Resources, Writing – original draft, Visualization, Funding acquisition.

Data availability

I have shared links on supplemental file \"Supplemental-spreadsheet-links\"

Acknowledgements

This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ttbdis.2023.102209.

References

Andersen, J.F., Ribeiro, J.M., 2017. Salivary kratagonists: scavengers of host physiological effectors during blood feeding. Arthropod Vector: Controller of Disease Transmission, Volume 2. Elsevier, pp. 51–63.

Andersson, L., Mikko, S., 1995. Generation of MHC class II diversity by intra-and intergenic recombination. Immunol. Rev. 143, 5–12.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27 (2), 573–580.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A.J., Poux, S., Bougueleret, L., Xenarios, I., 2016. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. Plant Bioinformatics. Springer, pp. 23–54.

Braunger, K., Ahn, J., Jore, M.M., Johnson, S., Tang, T.T., Pedersen, D.V., Andersen, G. R., Lea, S.M., 2022. Structure and function of a family of tick-derived complement inhibitors targeting properdin. Nat. Commun. 13 (1), 1–12.

Byrne, A., Cole, C., Volden, R., Vollmers, C., 2019. Realizing the potential of full-length transcriptome sequencing. Philos. Trans. R. Soc. B 374 (1786), 20190097.

Carrington, M., 1999. Recombination within the human MHC. Immunol. Rev. 167 (1), 245–256.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., 2021. Twelve years of SAMtools and BCFtools. Gigascience 10 (2) giab008.

De, S., Kingan, S.B., Kitsou, C., Portik, D.M., Foor, S.D., Frederick, J.C., Rana, V.S., Paulat, N.S., Ray, D.A., Wang, Y., 2023. A high-quality Ixodes scapularis genome advances tick science. Nat. Genet. 1–11.

Francischetti, I.M., Sa-Nunes, A., Mans, B.J., Santos, I.M., Ribeiro, J.M., 2009. The role of saliva in tick feeding. Front. Biosci. 14, 2051.

Giraldo-Calderón, G.I., Harb, O.S., Kelly, S.A., Rund, S.S., Roos, D.S., McDowell, M.A., 2021. VectorBase. org updates: bioinformatic resources for invertebrate vectors of human pathogens and related organisms. Curr. Opin. Insect Sci.

Gulia-Nuss, M., Nuss, A.B., Meyer, J.M., Sonenshine, D.E., Roe, R.M., Waterhouse, R.M., Sattelle, D.B., De La Fuente, J., Ribeiro, J.M., Megy, K., 2016. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nat. Commun. 7 (1), 1–13.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8 (8), 1494–1512.

Heyne, H., Elliott, E., Bezuidenhout, J.D., 1987. Rearing and infection techniques for *Amblyomma* species to be used in heartwater transmission experiments. Onderstepoort J. Vet. Res. 54 (1), 461–471.

Jia, N., Wang, J., Shi, W., Du, L., Sun, Y., Zhan, W., Jiang, J.-F., Wang, Q., Zhang, B., Ji, P., 2020. Large-scale comparative analyses of tick genomes elucidate their genetic diversity and vector capacities. Cell 182 (5), 1328–1340 e1313.

Krueger, F., 2016. TrimGalore: a wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. TrimGalore (accessed on 27 August 2019).

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9 (4), 357–359.

Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12 (1), 1–16.

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein Or nucleotide sequences. Bioinformatics (Oxford, England) 22 (13), 1658–1659.

Mans, B.J., Andersen, J.F., Ribeiro, J.M., 2022. A deeper insight into the tick salivary protein families under the light of Alphafold2 and Dali: introducing the TickSialoFam 2.0 database. Int. J. Mol. Sci. 23 (24), 15613.

Mans, B.J., De Castro, M.H., Pienaar, R., De Klerk, D., Gaven, P., Genu, S., Latif, A.A., 2016. Ancestral reconstruction of tick lineages. Ticks Tick Borne Dis. 7 (4), 509–535.

Miller, J.R., Koren, S., Dilley, K.A., Harkins, D.M., Stockwell, T.B., Shabman, R.S., Sutton, G.G., 2018. A draft genome sequence for the *Ixodes scapularis* cell line, ISE6. F1000Res. 7.

Nielsen, H., 2017. Predicting Secretory Proteins With SignalP, Protein function Prediction. Springer, pp. 59–73.

Nunn, M.A., Sharma, A., Paesen, G.C., Adamson, S., Lissina, O., Willis, A.C., Nuttall, P.A., 2005. Complement inhibitor of C5 activation from the soft tick *Ornithodoros moubata*. J. Immunol. 174 (4), 2084–2091.

Oliver, J.D., Lynn, G.E., Burkhardt, N.Y., Price, L.D., Nelson, C.M., Kurtti, T.J., Munderloh, U.G., 2016. Infection of immature *Ixodes scapularis* (Acari: ixodidae) by membrane feeding. J. Med. Entomol. 53 (2), 409–415.

Rhoads, A., Au, K.F., 2015. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13 (5), 278–289.

Ribeiro, J., 1995. Insect Saliva: Function, Biochemistry, and Physiology, Regulatory Mechanisms in Insect Feeding. Springer, pp. 74–97.

Ribeiro, J., Mans, B.J., 2020. TickSialoFam (TSFam): a database that helps to classify tick salivary proteins, a review on tick salivary protein function and evolution, with considerations on the tick sialome switching phenomenon. Front. Cell Infect. Microbiol. 374.

Ribeiro, J.M., Bayona-Vásquez, N.J., Budachetri, K., Kumar, D., Frederick, J.C., Tahir, F., Faircloth, B.C., Glenn, T.C., Karim, S., 2022. A draft of the genome of the Gulf Coast tick. Amblyomma Maculatum, 102090. Ticks and tick-borne diseases.

Ribeiro, J.M., Mans, B.J., Arcà, B., 2010. An insight into the sialome of blood-feeding Nematocera. Insect Biochem. Mol. Biol. 40 (11), 767–784.

Santiago, P.B., de Araujo, C.N., Charneau, S., Praca, Y.R., Bastos, I.M.D., Ribeiro, J.M.C., Santana, J.M., 2020. The pharmacopea within triatomine salivary glands. Trends Parasitol. 36 (3), 250–265. https://doi.org/10.1016/j.pt.2019.12.014.

Schoville, S., Burke, R., Dong, D.-y., Ginsberg, H., Maestas, L., Paskewitz, S., Tsao, J., 2023. Genome resequencing reveals population divergence and local adaptation of blacklegged ticks in the United States.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics (Oxford, England) 31 (19), 3210–3212.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I., 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19 (6), 1117–1123.

Tirloni, L., Lu, S., Calvo, E., Sabadin, G., Di Maggio, L.S., Suzuki, M., Nardone, G., da Silva Vaz Jr., I., Ribeiro, J.M.C., 2020. Integrated analysis of sialotranscriptome and sialoproteome of the brown dog tick *Rhipicephalus sanguineus* (s.l.): insights into gene expression during blood feeding. J. Proteomics 229, 103899. https://doi.org/10.1016/j.jprot.2020.103899.

Trowsdale, J., Parham, P., 2004. Mini-review: defense strategies and immunity-related genes. Eur. J. Immunol. 34 (1), 7–17.