Micropathogen community identification in ticks (Acari: Ixodidae) using third-generation sequencing

Jin Luo [a,b], Qiaoyun Ren [a], Wenge Liu [a], Xiangrui Li [b], Hong Yin [a,c], Mingxin Song [d], Bo Zhao [e], Guiquan Guan [a], Jianxun Luo [a], Guangyuan Liu [a,*]

[a] State Key Laboratory of Veterinary Etiological Biology, Key Laboratory of Veterinary Parasitology of Gansu Province, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Science, Xujiaping 1, Lanzhou, Gansu, 730046, PR China
[b] MOE Joint International Research Laboratory of Animal Health and Food Safety, College of Veterinary Medicine, Nanjing Agricultural University, Nanjing, 210095, PR China
[c] Jiangsu Co-Innovation Center for the Prevention and Control of Important Animal Infectious Disease and Zoonoses, Yangzhou University, Yangzhou, 225009, PR China
[d] Heilongjiang Provincial Key Laboratory of Animal-borne Zoonosis/Northeast Agriculture University, Harbin, 150030, PR China
[e] Gansu Agriculture Technology College, Duanjiatan 425, Lanzhou, 730030, PR China

ARTICLE INFO

ABSTRACT

Ticks are important vectors that facilitate the transmission of a broad range of micropathogens to vertebrates, including humans. Because of their role in disease transmission, it has become increasingly important to identify and characterize the micropathogen profiles of tick populations. The objective of the present study was to survey the micropathogens of ticks by third-generation metagenomic sequencing using the PacBio Sequel platform. Approximately 46.481 Gbp of raw micropathogen sequence data were obtained from samples from four different regions of Heilongjiang Province, China. The clean consensus sequences were compared with host sequences and filtered at 90% similarity. Most of the identified genomes represent previously unsequenced strains. The draft genomes contain an average of 397,746 proteins predicted to be associated with micropathogens, over 30% of which do not have an adequate match in public databases. In these data, *Anaplasma phagocytophilum* and *Coxiella burnetii* were detected in all samples, while *Borrelia burgdorferi* was detected only in *Ixodes persulcatus* ticks from G1 samples. Viruses are a key component of micropathogen populations. In the present study, *Simian foamy virus*, *Pustyn virus* and *Crimean-Congo haemorrhagic fever orthonairovirus* were detected in different samples, and more than 10–30% of the viral community in all samples comprised unknown viruses. Deep metagenomic shotgun sequencing has emerged as a powerful tool to investigate the composition and function of complex microbial communities. Thus, our dataset substantially improves the coverage of tick micropathogen genomes in public databases and represents a valuable resource for micropathogen discovery and for studies of tick-borne diseases.

1. Introduction

Micropathogens pose serious threats to the health of livestock, wildlife and even humans. In many cases, the spread of disease is mediated through micropathogens residing within arthropod vectors (Paula et al., 2017). Ticks are important vectors that facilitate the transmission of a broad range of micropathogens to vertebrates, including humans (Eisen et al., 2017). These small arachnids are capable of transmitting micropathogens, including viruses, bacteria, protozoa, and fungi (Jahfari et al., 2017; Krzysztof et al., 2015). Ticks and the

micropathogens that they transmit cause direct damage to animals by reducing animal weight, milk production and leather quality (Luo et al., 2019). However, a wide variety of micropathogens are present in ticks, and the confirmed micropathogens represent only a small percentage of all micropathogens. Because of their role in disease transmission, it has become increasingly important to identify and characterize the micropathogen profiles of tick populations. In addition, the African swine fever incident that was spread by soft ticks in 2018 in China caused substantial losses to the pig industry, attracting the attention of related fields. Some undetected micropathogens are also likely to cause severe

disease in animals or humans. For example, in 2010, a previously undetected bunyavirus severely impacted many people in Henan Province (Xu et al., 2011). Therefore, it is particularly important to predict the micropathogens that may be carried by vectors. Currently, the use of PCR amplicon sequencing is limited to a few predefined targets (Adrian et al., 2020; Latrofa et al., 2020). Thus, deep sequencing is used for micropathogen detection. In 2017, micropathogens were detected in ticks by small RNA sequencing. micropathogens, such as viruses and bacteria, detected in ticks in this study had not been previously detected in ticks, these results provided a good basis for research on important tick-borne pathogens (Luo et al., 2017). However, because the sequences were short, the assembled data were incomplete. As a result, research on the identification of pathogens is largely lacking. In recent years, unbiased next-generation metagenomic sequencing (NGMS) has been used to identify novel and emerging human pathogens circulating in tick vectors. For example, Heartland virus, discovered in Missouri in 2012 via NGMS, is transmitted by the lone star tick (*Amblyomma americanum*) and is a potential cause of febrile illness and death in humans (Laura et al., 2012). Heartland virus has since been detected in mammalian hosts in 13 U.S. states (Riemersma et al., 2015). However, a shortcoming of NGMS is sequence length, which cannot accurately identify similar sequences. In 2015, PacBio launched a new and upgraded third-generation sequencing instrument, the PacBio Sequel sequencing system (Lavezzo et al., 2016; Rhoads et al., 2015; Wagner et al., 2016; Shingo et al., 2018). The associated read length, high throughput, high accuracy and other features of this instrument have brought a new third-generation metagenomic sequencing (TGMS) experience to the research field. In the present study, micropathogens were detected using TGMS to assess population characteristics and the diversity of micropathogens in ticks.

2. Materials and methods

2.1. Tick collection and DNA extraction

Unengorged ticks [*Haemaphysalis japonica* (*H. japonica*) (n = 102), *Ixodes persulcatus* (*I. persulcatus*) (n = 97), *Dermacentor silvarum* (*D. silvarum*) (n = 150) and *Haemaphysalis concinna* (*H. concinna*) (n = 112)] were collected throughout Jiameng forest farm and Sanguliu Yichun city (longitude 128°48′-129°08′ east, latitude 47°41′-48°04′ north), Luobei County, Hegang city (longitude 130°01′-131°34′ east, latitude 47°12′-48°21′ north) and Jiejinkou Tongjiang city (longitude 132°50′25.61″ east, latitude 47°56′3.60″ north), Heilongjiang Province, China, from 15 May to 23 July in 2018. All ticks were collected by the flag method, which is suitable for less populated grasslands and involves dragging a white cloth of approximately 1 square metre on grass to collect unfed ticks. The collected ticks were identified by morphology in the Department of Veterinary Parasitology, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences. Based on the epidemiological characteristics of local dominant ticks, these samples were separated into four different groups. All *H. japonica*, *I. persulcatus*, *D. silvarum* and *H. concinna* ticks were collected from different regions and mixed as groups 1 (G1), 2 (G2), 3 (G3) and 4 (G4). To sterilize the external surface of the ticks to ensure that the micropathogen sequences were internal to the ticks, the collected ticks were immediately placed in phosphate-buffered saline (PBS) and washed twice in a solution containing 0.133 M NaCl, 1.11% sodium dodecyl sulphate (SDS) and 0.0088 M ethylenediaminetetraacetic acid (EDTA) (Luo et al., 2017). These clean ticks were mixed and stored in liquid nitrogen until ground using a mortar and pestle with liquid nitrogen, and genomic DNA (gDNA) was extracted with a QIAamp DNA Mini Kit (QIAGEN, China) following the manufacturer's instructions.

2.2. Library construction and sequencing

Total DNA quality was analysed on a PacBio RS II sequencing platform analyser system and by resolution on a denaturing polyacrylamide gel electrophoresis system. A DNA database library was generated according to the DNA sample preparation instructions. Subsequently, DNA was amplified with Pfx DNA polymerase (Invitrogen, China) using 20 PCR cycles and a PacBio DNA primer set. PCR products were purified, and the recovered DNA was precipitated and quantified with both a Nanodrop Spectrophotometer (Thermo Scientific) and a TBS-380 mini fluorometer (Turner Biosystems) using PicoGreenH dsDNA quantitation reagent (Invitrogen). The sample concentration was adjusted to 10 nM, and a final volume of 10 mL was used for the sequencing reaction. The purified DNA library was used for cluster generation (on the PacBio Cluster Station). Subsequently, DNA was sequenced on a PacBio Sequel machine following the manufacturer's instructions (Nextomics), and the library construction process is shown in Fig. 1. The gDNA concentration was normalized by dilution from a high to a low concentration and was then sequenced using the PacBio platform.

2.3. Transcriptome sequence analysis

The raw sequencing reads contained low-quality sequences (Table 1). To generate reliable data for analysis, we processed the raw reads as follows. 1) Quality control of the sequencing data: The circular consensus sequencing (CCS) workflow of SMRT Link software (Yan et al., 2020) and the Arrow algorithm (Sontag et al., 2009) were used to obtain high-quality raw CCS reads with the primary parameters of –minPasses 1 –polish -minPredictedAccuracy 0.8. To obtain clean CCS reads, accuracy and length filtering were performed with the parameters accuracy ≥ 99% and length ≥ 500 bp. CCS and host sequence filtering were performed using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to compare the clean CCS reads with the host sequences, which were filtered at a 90% similarity level. In this process, *Rhipicephalus microplus* (https://www.ncbi.nlm.nih.gov/assembly/GCA_002176555.1/) and *Ixodes scapularis* (https://www.ncbi.nlm.nih.gov/assembly/GCF_00020 8615.1/) were used as references. 2) Gene prediction: The genes were predicted from samples of metagenomic sequences using MetageneMark (Wazim et al., 2014), and the primary parameter was -p meta. Redundant sequences were removed from the predicted genes based on a 95% similarity level and 90% coverage using CD-HIT. 3) Species annotation and taxonomic analysis of pathogens: Sequence alignment with the nonredundant (NR) database was performed with DIAMOND and the CAZy database; nonredundant gene sets were compared to the entire NR database, and all gene annotation results were obtained. The primary parameters were –evalue 0.00001 and –sensitive. For species annotation and statistical analysis of the gene annotation results obtained, the LCA algorithm was used to annotate all CCS reads, and finally, species abundance information was calculated based on the CCS annotation results. 4) Functional annotation: To examine the evolutionary genealogy of genes, the non-supervised orthologous groups (eggNOG) database provides a functional annotation of constructed orthologous groups using the Smith-Waterman matching algorithm. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database is now a comprehensive database, at the core of which are the KEGG Pathway and KEGG Orthology (KO) databases.

2.4. Species abundance and community composition

From the phylum to species level, the top 50 species were selected by the maximum ranking method, and a community composition heat map was drawn based on species abundance. The heat map of community composition allowed the proportions of different species in different samples to be easily identified, which is convenient for the discovery of dominant species.

The redundant genes were analysed in the NR database using DIAMOND, and the corresponding gene copies were then combined to calculate the species abundance in each sample. For all samples, from the kingdom to species level (kingdom, phylum, class, order, family, genus and species), the maximum ranking method was used to select the
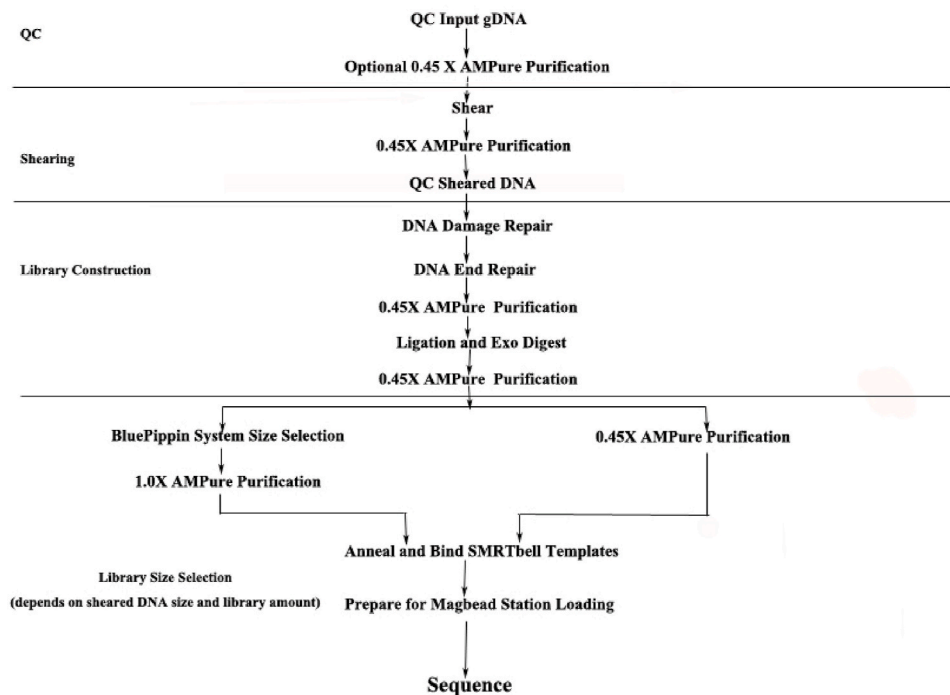
Fig. 1. The genomic DNA library was generated according to the PacBio Sequel sample preparation instructions.

Table 1
Sequencing data statistics for each sample.

Samples	Raw Bases (Gbp)	Raw CCS	Low quality reads	Clean CCS	Host Removal	Bases (Mbp)	N50 Length (bp)
G1	14.876	525,604	142,425	383,179	348,745	494.500	1417
G2	9.844	380,783	185,783	195,000	153,240	368,792	2406
G3	10.798	457,749	246,077	211,672	145,837	359,253	2463
G4	10.963	442,502	225,154	217,348	162,561	405.766	2496
Average	11.620	451,659	199,860	251,799	202,596	407,078	2196

Note: Sample: Sample name. Raw Bases: Raw bases of subreads (Gbp). Raw CCS: Sequence consistency analysis and preliminary quality control of subreads were performed to obtain the number of original CCS data. Clean CCS: The number of CCS results obtained by further performing a series of data quality controls on the original CCS data. Host Removal: The host filtering sequence for Clean CCS is the final sequence set to enter subsequent analyses.

species with an abundance greater than 1% among the top 100 species. Then, a histogram of the relative abundances of species was drawn, which is convenient for viewing the dominant species in each sample.

2.5. Microbial sequence detection using BLAST and PCR confirmation

BLAST searches were conducted with NCBI BLAST 2.2.26 to identify micropathogen sequences in the clean unique reads. The results were then manually analysed to screen for potential viral, bacterial, fungal, and protozoan sequences. To verify the reliability of the data pertaining to microbial community composition in various samples following third-generation sequencing, the presence of previously described micropathogens, including *Simian foamy virus*, *Crimean-Congo haemorrhagic fever orthonairovirus*, *Coxiella burnetii* (*C. burnetii*), *Borrelia burgdorferi* (*B. burgdorferi*) and *Anaplasma phagocytophilum* (*A. phagocytophilum*), was assessed by PCR using micropathogen-specific primers (listed in Table 4). PCR amplification was performed in an automatic DNA thermocycler (Bio-Rad, Hercules, CA, USA), and the PCR products were separated by 1.5% agarose gel electrophoresis to assess the presence of specific bands indicative of micropathogens (data not shown). The DNA fragments generated were recovered, ligated into the pGEM^R-T Easy vector (Invitrogen, Carlsbad, CA, USA) and transformed into competent *Escherichia coli* DH5α cells (Takara Bio Inc., Dalian, China). At least three positive clones were sequenced per sample by GenScript Corporation (Piscataway, NJ, USA).

2.6. Phylogenetic and taxonomic analysis

In the present study, *Simian foamy virus*, *Crimean-Congo haemorrhagic fever orthonairovirus*, *C. burnetii*, *B. burgdorferi* and *A. phagocytophilum* were analysed to assess the diversity of the putative pathogens in the ticks. The different pathogen-specific sequences were amplified and sequenced to construct a phylogenetic tree using the neighbour joining method in MEGA 7 (Livak and Schmittgen, 2001).

3. Results

3.1. Sequencing data statistics

In the present study, we performed metagenomic Sequel sequencing on 4 tick samples. The raw data and clean CCS quality statistics of the samples before and after quality control are shown in Table 1. A total of 46.481 Gbp of raw data were obtained, and the proportions in the raw data were determined, with an average of 451,659 clean CCS reads obtained per sample (data not shown). The identified genes were compiled into a nonredundant catalogue of 282,746 genes, and the average maximum and N50 lengths were 4290 and 2195 bp, respectively. The average sequence length is shown for each sample (Fig. 2).
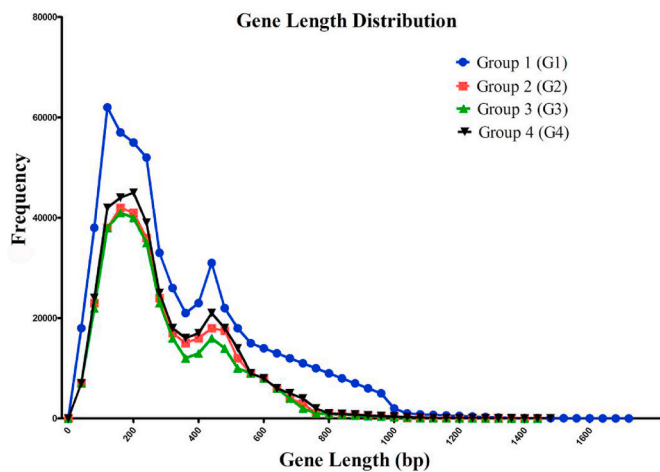
Gene Length Distribution

Fig. 2. Summary of the tag length distribution following the sequencing of genes from the four samples.

3.2. Gene prediction and functional annotation

MetageneMark and FragGeneScan were used to directly predict the genes in the CCS reads to avoid the introduction of error into the assembly (Ismail WM et al., 2014). The results showed that the average gene number was 387,746, the total length of the predicted genes was 44,897,989 and the average max length was 4290 bp for each gene (Table 2).

DIAMOND was used to compare the gene set sequences with the eggNOG database (the expected value of the parameter set was set as 1e-5), and the corresponding functional classification (category) and homology groups (non-supervised orthologous groups, NOGs) were obtained. The results identified nearly 9000 genes as being involved in replication, recombination and repair (Table 3). Genes involved in maintaining nuclear structure were not expressed in almost any of the four samples, with only one gene with a similar function detected in the G2 sample. Of all the genes with known functions, 1376 were common to all samples. There were 503 genes specific to G1, 2226 genes specific to G2, 937 genes specific to G3 and 622 genes specific to G4, while other genes were shared between the two samples (Fig. 3).

Analysis of the CAZy annotations showed that glycosyl transferases had a high match rate in the four samples and that the number of matched genes was more than 4000 copies (data not shown). The KEGG database is also an important tool for studying molecular functions. In this analysis, DIAMOND was used to compare the gene set sequence with the KEGG database (the expected e-value was set as 1e-5), and the corresponding metabolic pathways and KOs were obtained. The results identified a number of human disease-related genes, including those related to neurodegenerative diseases, viral infection-related diseases, parasitic infection-related diseases, bacterial infection-related diseases, immune diseases, endocrine diseases and metabolic diseases. In addition, a large number of pathogen-related genes were detected in different samples (data not shown). Based on the analysis of microbial

Table 2
Gene prediction statistics of each sample.

Samples	Total Number	Total Length (bp)	Max Length (bp)	Min Length (bp)
G 1	513,897	56,853,886	4371	60
G 2	360,245	42,224,994	3726	60
G 3	333,896	36,617,219	4212	60
G 4	382,946	43,895,857	4851	60
Average	397,746	44,897,989	4290	60

Note: MetageneMark was used to directly predict the genes of the CCS reads to avoid the introduction of error into the assembly.

Table 3
Functional classification of the eggNOG annotation results of the four samples.

Functional Category	Description	Samples and Gene Number			
		Group 1 (G1)	Group 2 (G2)	Group 3 (G3)	Group 4 (G4)
A	RNA processing and modification	24	24	15	21
B	Chromatin structure and dynamics	320	443	238	545
C	Energy production and conversion	964	2047	795	618
D	Cell cycle control, cell division, chromosome partitioning	208	377	175	163
E	Amino acid transport and metabolism	898	3288	930	643
F	Nucleotide transport and metabolism	536	975	343	409
G	Carbohydrate transport and metabolism	609	1815	615	450
H	Coenzyme transport and metabolism	691	1195	310	466
I	Lipid transport and metabolism	590	1251	510	444
J	Translation, ribosomal structure and biogenesis	1408	2082	829	949
K	Transcription	617	2047	701	462
L	Replication, recombination and repair	18547	14150	9432	16255
M	Cell wall/membrane/ envelope biogenesis	751	2047	642	435
N	Cell motility	8	324	100	19
O	Posttranslational modification, protein turnover, chaperones	1158	1484	793	832
P	Inorganic ion transport and metabolism	336	1961	670	337
Q	Secondary metabolites biosynthesis, transport and catabolism	208	791	245	211
R	General function prediction only	0	0	0	0
S	Function unknown	42726	30115	21232	26982
T	Signal transduction mechanisms	407	1811	656	385
U	Intracellular trafficking, secretion, and vesicular transport	654	1051	501	538
V	Defence mechanisms	215	588	257	170
W	Extracellular structures	0	5	9	0
Y	Nuclear structure	0	1	0	0
Z	Cytoskeleton	115	69	87	88

genes (16 or 18S rRNA), some of the microbes were common among different ticks, while others were specific to certain ticks (Fig. 4). These genes are most likely involved in pathogen invasion or in protecting the body against various environmental stimuli (Luo et al., 2019).

3.3. Community composition

Ixodes showed a relatively high abundance in the community composition analysis, because all of the raw CCS data of ticks were directly mapped to the genomes of all known species (Fig. 5). Therefore, its relative abundance between different samples was high. *Anaplasma* and *Coxiella* also had a high relative abundance in the four samples, indicating that they are common micropathogens in ticks. Moreover, a

Table 4
Oligonucleotides used as primers for PCR analysis of micropathogens.

| Micropathogens	Primer names	Primer sequence (5′–3′)	References
Simian foamy virus	SFV*pol*	5′-CCTGGATGCAGAGTTGGATC-3′	Reid MJC et al. (2017)
	SFV*pol*874	5′-CACGAATTTCCTGTAAAAAGA-3′	
Crimean-Congo haemorrhagic fever agent	CCHFVF	5′-TGGACACCTTCACAAACTC-3′	Tekin S et al. (2012)
	CCHFV536R	5′-GACAAATTCCCTGCACCA-3′	
Coxiella burnetii	CbUF	5′-AAGGATCCAATTAACCGTTGTAGTT-3′	Qi Y et al. (2018)
	CbUR1042	5′-CGGAATTCTCACTCTTTCCTATGTT-3′	
Borrelia burgdorferi	BBUF	5′- CACGA CTT TCT TCG CCT TAA AGC-3′	Maggi RG et al. (2019)
	BBUR	5′- GTT AAG CTC TTA TTC GCT GAT GGT A-3′	
Anaplasma phagocytophilum	APH*msp*4F	5′-ATGAATTACAGAGAATTGCTTGTAGG-3′	de la Fuente J et al. (2005)
	APH*msp*849R	5′-TTAATTGAAAGCAAATCTTGCTCCTATG-3′	
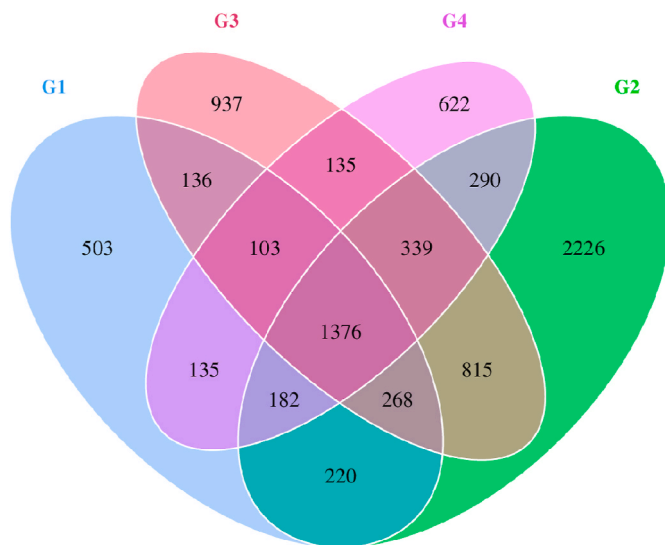
Fig. 3. Venn diagram based on the eggNOG database. Note: The corresponding functional categories and non-supervised orthologous group (NOG) numbers were obtained from the eggNOG database.
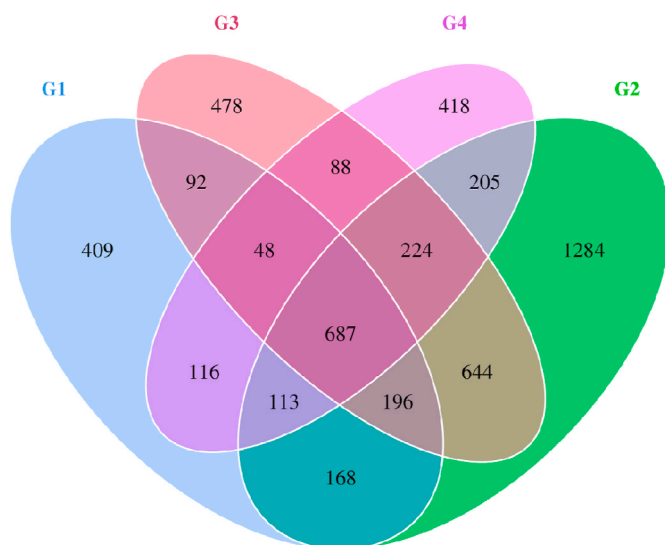

Fig. 4. Venn diagram based on the KEGG Orthology database. Note: DIAMOND was used to compare gene sequences with the KEGG database, and the corresponding metabolic pathway information and KEGG orthology results of the genes were obtained from the KEGG database.

relatively high abundance of *Pseudomonas* was only observed in the G2 sample, indicating that this bacterium was sensitive to *I. persulcatus* ticks.

3.4. Species abundance detection and annotation

The above results showed that bacteria play an important role in the composition of the microbial populations in ticks. Analysis of the relative abundance of the same species in different samples indicated that *A. phagocytophilum* was an important pathogen in these libraries and accounted for 40, 22, 38 and 50% of the G1, G2, G3 and G4 community compositions, respectively. Importantly, *B. burgdorferi* was detected in G1 from *I. persulcatus* ticks, and *C. burnetii* was detected in all samples; both of these species induce severe diseases in humans (Fig. 6).

Moreover, a number of viruses that infect humans and animals were identified, such as *Simian foamy virus*, *Pustyn virus* and *Crimean-Congo haemorrhagic fever orthonairovirus* from G2, G3 and G4, respectively. (Fig. 7). However, in addition to these known micropathogens, there are still many important undetected micropathogens that remain to be identified (Fig. 8).

In G1, *Bole tick virus*, *Wuhan tick virus* and *Tjuloc virus* had a high abundance of 18% (Fig. 9A). In G2, *Lymphocystis disease virus* comprised 33% of the community (Fig. 9B). In G3, *Blacklegged tick phlebovirus* and *Pustyn virus* had high abundances of 20 and 19%, respectively (Fig. 9C). In G4 (Fig. 9D), *Hubei diptera virus and Tacheng tick virus* had high abundances of 21 and 19%, respectively, while other viruses, such as *Crimean-Congo haemorrhagic fever orthonairovirus*, *Ambidensovirus CaaDV1*, *Lambdina fiscellaria nucleopolyhedrovirus* and *Culex tritaenio-rhynchus totivirus*, accounted for 10% of the viral composition in the communities. Interactive pie charts were generated using KRONA for the species annotation results, where the order from the inside to the outside represents the different classification levels, and the sectorial areas represent the relative proportions of different species. The results indicated that the levels of different pathogens in the samples were suggestive of how likely the pathogens were to cause harm to the host.

3.5. Bacterial infection of ticks identified by PCR

To assess the presence of micropathogens in different tick samples, PCR was performed on the samples collected. We assessed the presence of some major micropathogens, including *Simian foamy virus*, *Crimean-Congo haemorrhagic fever orthonairovirus*, *Coxiella burnetii*, *Borrelia burgdorferi* and *Anaplasma phagocytophilum*, using the specific primers listed in Table 4. The results were negative for *Simian foamy virus*, while the other pathogens tested positive in different samples, and the results were consistent with the data generated from third-generation metagenomic sequencing.

3.6. Phylogenetic distribution of novel lineages

The 16S rRNA gene analysis indicated high level of conservation between *C. burnetii* (identified in the four samples analysed). The same
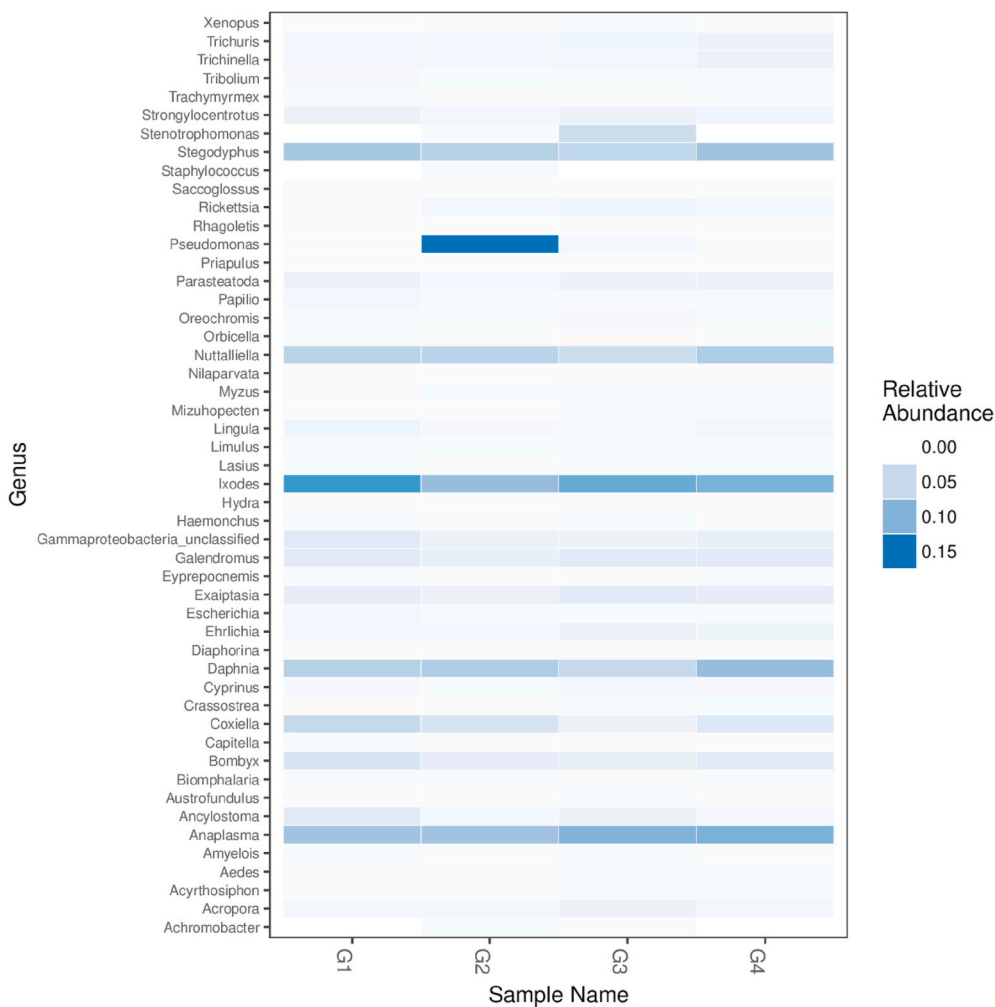
Fig. 5. Heat maps of community composition based on genera. Note: *X*-axis, template name; *Y*-axis, genus. The darker the blue colour is, the higher the enrichment of the genus in the sample.
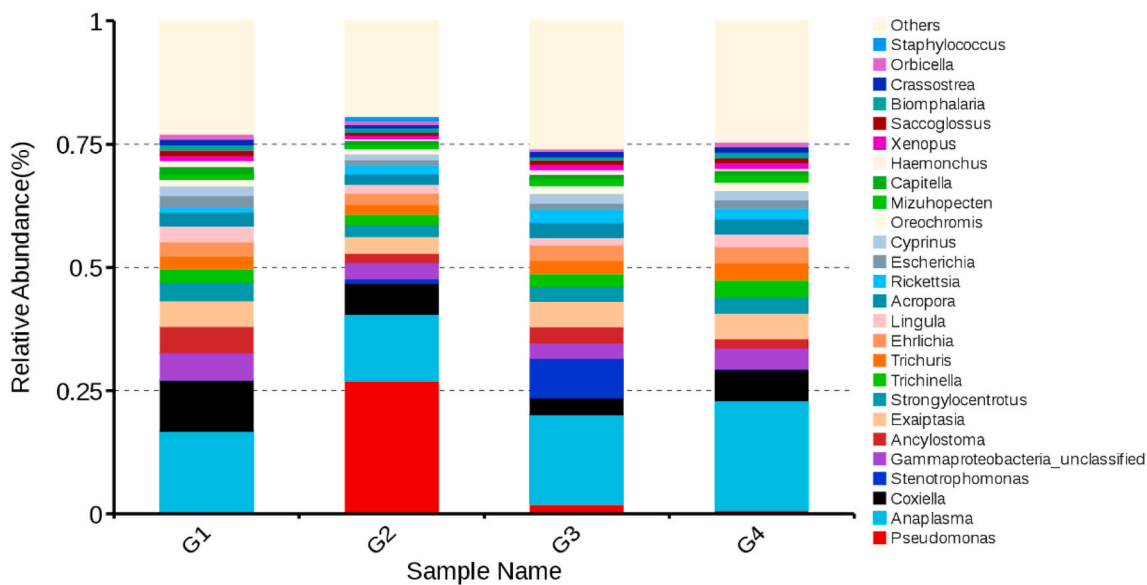

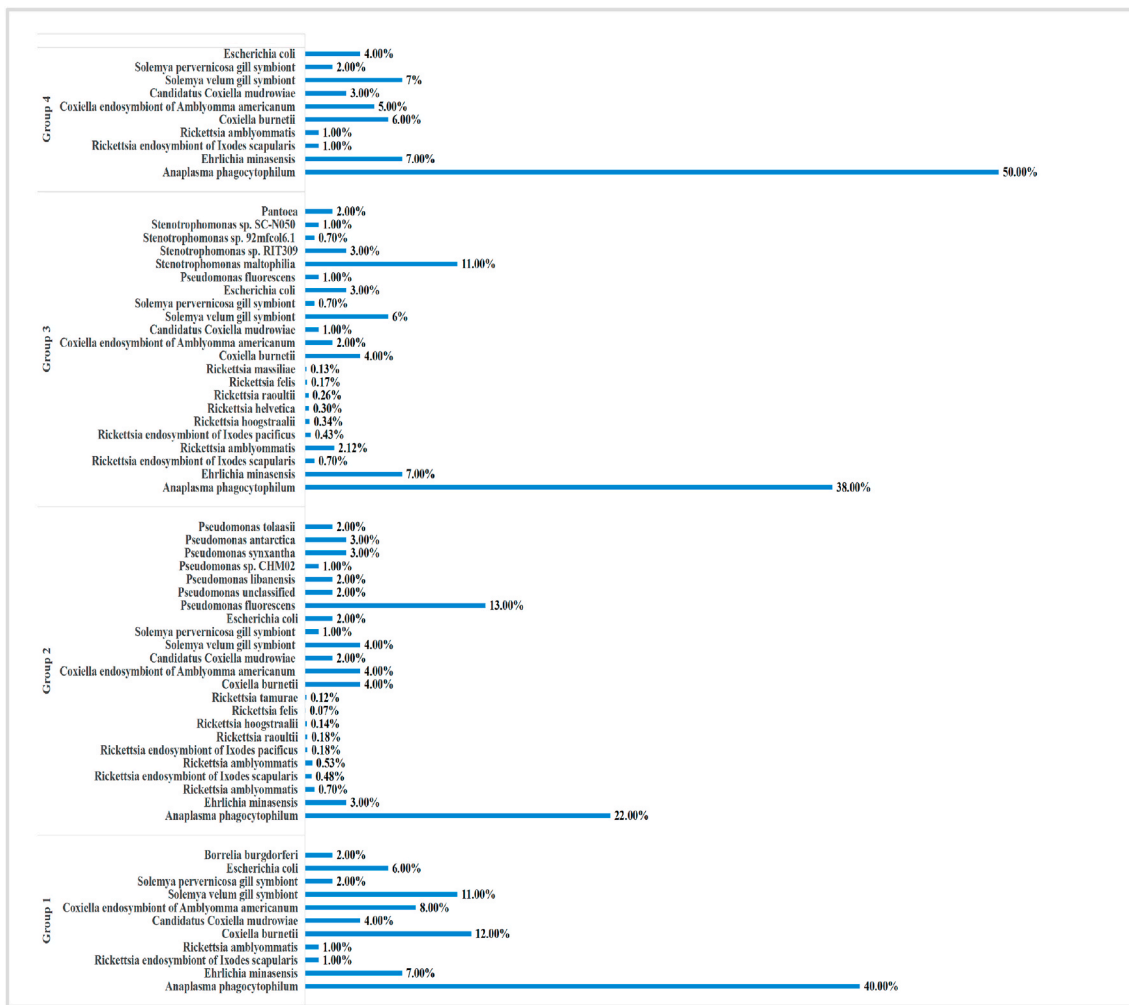**Fig. 6.** Bar chart of the relative abundances of genera.

Fig. 7. Population distribution of micropathogens in different samples. Note: "%" represents the proportion of micropathogen in the total community from each sample.
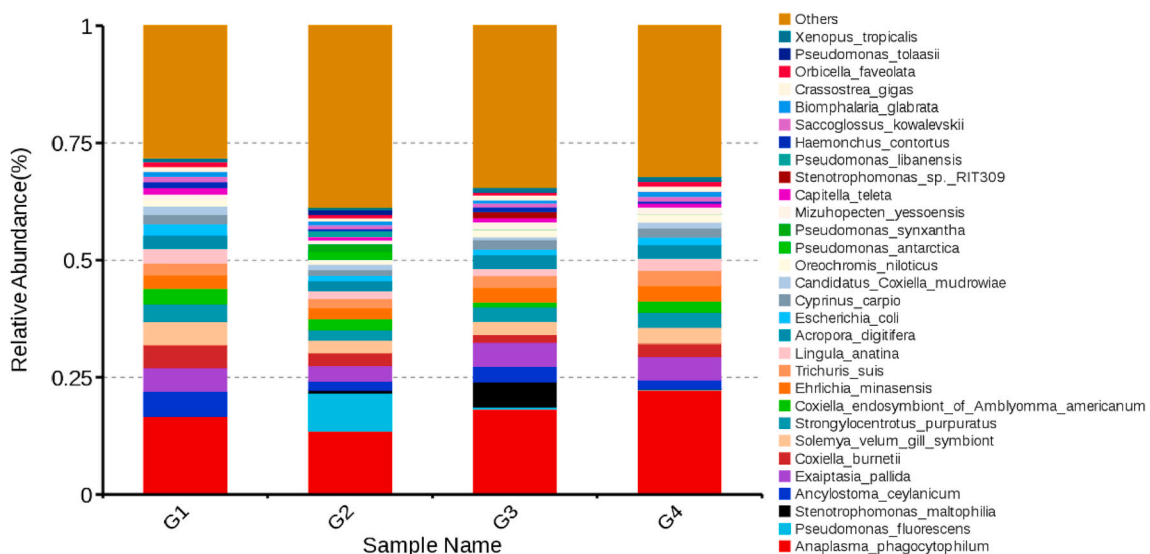

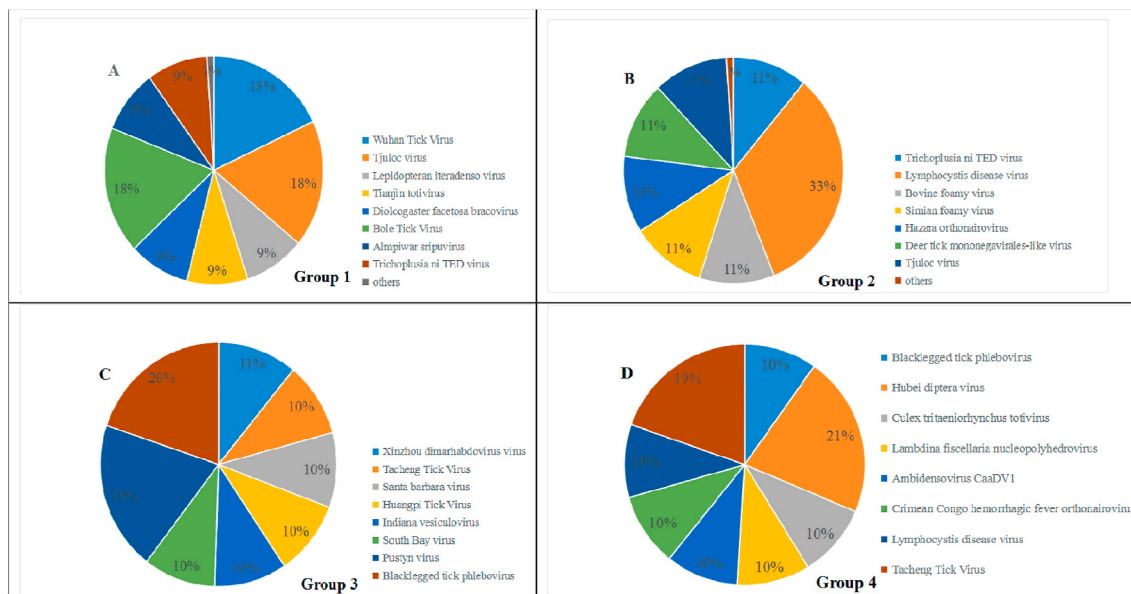**Fig. 8.** Bar chart of the relative abundances of bacteria.

Fig. 9. Pie chart of the distribution of viral abundances in different samples. Note: "%" represents the proportion of the virus in the total community from each sample, and the different colours represent different viruses. A: The primary micropathogens analysed in G1. B: The primary micropathogens analysed in G2. C: The main micropathogens analysed in G3. D: The main micropathogens analysed in G4.

phenomenon was observed for *Borrelia burgdorferi*, but *Borrelia* was highly conserved at the genus level (Fig. 10A), while *Coxiella* remained highly conserved at the species level (Fig. 10B). The MSP4 gene of *A. phagocytophilum* was obviously divided into three main genotypes (Fig. 10C). In addition, *Simian foamy virus* and *Crimean-Congo haemorrhagic fever orthonairovirus* show a more complex taxonomy, and the sequences are less conserved for the same genes (Fig. 10D and 10E).

4. Discussion

The vast woodland resources in northeast China are an important base for livestock breeding and provide an excellent environment for tick survival. Therefore, understanding ticks and tick-borne diseases is important for animal and human health. Thus, the future identification of complex communities of micropathogens via metagenomic sequencing may facilitate the development of effective control measures against ticks and tick-borne diseases. In the present study, we surveyed field-collected ticks from northeast China for tick-borne micropathogens. These findings will improve our understanding of the factors that affect the transmission of micropathogens by tick vectors and also be beneficial for the analysis of the relationship between ticks and their resident microbial populations.

A total of 47 Gbp of read data were generated. After generating the raw sequence data, some insertion tags, low-quality tags, poly-A tags and small tags were removed. Then, the length distributions of clean CCSs and common/specific tags were summarized for the four samples. In addition, the length distribution results indicate that the sequencing length was continuous (Fig. 2), which is a unique feature for TGMS and indirectly indicates the reliability of the data and integrity of the gDNA for further analysis.

Gene prediction is an important index in micropathogen sequence analysis. Gene composition is of great significance to the source of sequences, functional analysis of genes, and species and population compositions of micropathogens (Figs. 3 and 4). The results showed that disease-related pathogens comprised an important part of all detected genes, including genes from viruses, bacteria and parasites, and some expressed genes were related to immunity and metabolism (data not shown). This result strongly suggests that micropathogens are involved most possible infectious diseases transmitted by ticks and indicate that the livestock industry and human health will likely be severely harmed if

prevention efforts are not strengthened. These results also provide a good foundation for screening disease-related pathogens and carrying out research on gene-related vaccines. This is the first time that a TGMS method has been used to assess the potential threat of ticks, and the results provide important guidance for the prevention and control of ticks in the future.

To further analyse the structure of micropathogen communities in ticks, all sequences were mapped to the genomes of different species. The microbial population in ticks is a complex system that includes viruses, bacteria and parasites and fungi.

At the genus level, a high abundance of *Ixodes* sequences was identified in the four different samples (Fig. 5) that obviously originated from the samples themselves but also demonstrated the reliability of our sequencing quality. Moreover, *Anaplasma* and *Coxiella* also had high abundances in these samples, indicating that these pathogens were common micropathogens in ticks from Heilongjiang Province, China, and revealed the broad spectrum of these micropathogens in ticks. These results serve as a reminder that attention must be paid to the prevention of tick bites in daily activities due to a high risk of infection-related tick-borne diseases, which can result in unnecessary health-related and economic losses.

The clean CCS reads mapped to the genomes of all micropathogens, revealing the presence of many micropathogens and demonstrating that these micropathogens infect not only animals but also humans. Among the identified micropathogens, *A. phagocytophilum*, *C. burnetii* and *B. burgdorferi* cause serious diseases in humans, although *B. burgdorferi* was only detected in *I. persulcatus* ticks in the G2 sample. *B. burgdorferi* causes severe clinical symptoms and even death in infected animals. Although *B. burgdorferi* was only detected in *I. persulcatus*, as it is a dominant tick in Heilongjiang Province, this pathogen is likely responsible for the prevalence of Lyme disease in the region and is potentially harmful to human health and animal husbandry. *A. phagocytophilum* (formerly *Ehrlichia phagocytophilum*) (Dumler et al., 2001) is a gram-negative bacterium that is unusual in its tropism to neutrophils. This pathogen causes anaplasmosis in sheep and cattle, also known as tick-borne fever and pasture fever, and causes the zoonotic disease human granulocytic anaplasmosis (Annetta et al., 2017). *C. burnetii* is one of the most infectious organisms (Li et al., 2005; David et al., 2017), and the disease caused by this bacterium occurs in two stages: an acute stage in which patients present with headaches, chills, and respiratory
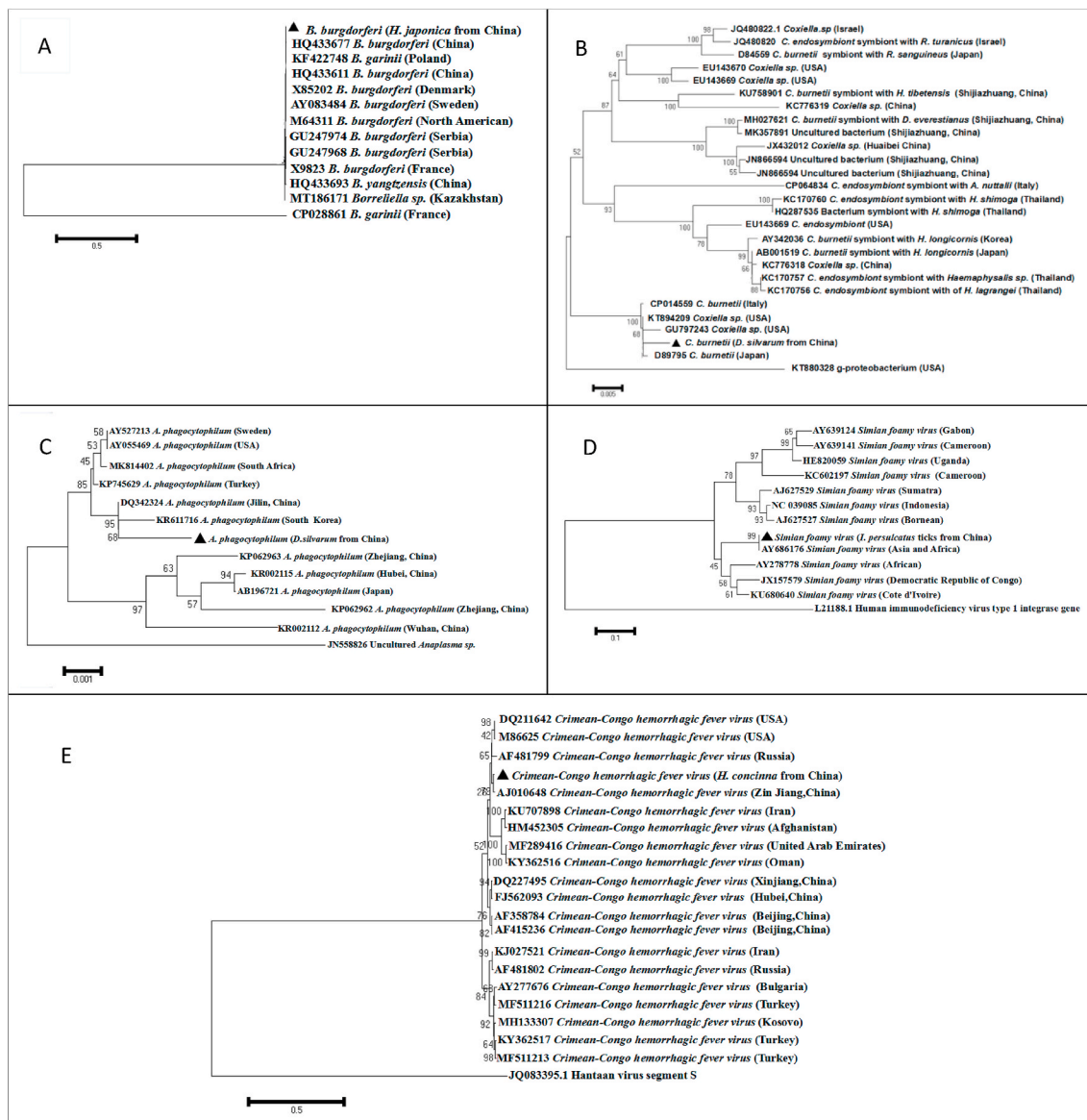
Fig. 10. Phylogenetic analysis of the isolated bacteria/viruses. Reference oligonucleotide sequences were selected by BLAST searches of the NCBI nt database. (A) Subtrees of the experimental sequences from the *Borrelia burgdorferi* 16S rRNA gene. (B) Subtrees of the experimental sequences from the *Coxiella burnetii* 16S rRNA gene. (C) Subtrees of the experimental sequences from the *Anaplasma phagocytophilum* MSP4 gene. (D) Subtrees of the experimental sequences from the *Simian foamy virus* pathogen. (E) Subtrees of the experimental sequences from the *Crimean-Congo haemorrhagic fever orthonairovirus* segment-S gene.

symptoms and an insidious chronic stage. These micropathogens were detected in ticks in different geographical environments from Heilongjiang Province by TGMS, suggesting the need for further attention and continued public health monitoring. In addition, disease prevention in the livestock breeding industry should be strengthened to achieve healthy breeding.

According to the results presented in Fig. 6, approximately 30% of the sequences in each group of samples could not be mapped to any genome. Because the species range of micropathogens is wide, there is no database containing all micropathogens. TGMS has also been able to detect only a small number of micropathogens, while many remain undetected. Therefore, it is necessary to continue to identify and study these undetected micropathogens. Furthermore, these unknown sequences pose a challenge to our study of the microbial community composition of ticks and of the prevention and control of tick-borne diseases that may harm human health and animal husbandry. Because little is known about ticks, especially tick-borne pathogens, substantial effort needs to be devoted to studies on tick-borne pathogens to provide

better and more accurate information for enhanced prevention and control of diseases caused by these vectors. Among some of the known sequences, many virus sequences were detected, such as *Simian foamy virus*, *Bole tick virus*, and *Pustyn virus* sequences (Fig. 7). *Tjuloc virus* was present in the G1 and G2 samples and accounted for 8 and 11% of the total community, respectively. *Blacklegged tick phlebovirus* was identified in the G3 and G4 samples and accounted for 20 and 10% of the total community, respectively. In particular, viruses that are named for the place in which they were found, such as *Wuhan tick virus*, *Hubei diptera virus*, *Tacheng tick virus* and *Indiana vesiculovirus*, were identified, probably because of factors related to animal trade or migration, which introduces micropathogens from other regions and even other countries to Heilongjiang Province. These findings are highly significant for the investigation and research of foreign diseases. The above results are also shown in Figs. 8 and 9.

The pathways, enzymes, and protein families that were observed to be overrepresented in the tick groups are potentially relevant to tick-related processes. Glycoside transferases play an important role as

antibiotic glycosyltransferases and constitute a category of compounds that are widely used in the clinic for their antibacterial and anticancer activities (Huang et al., 2018). With the development of sequencing techniques, large numbers of glycosyltransferases have been identified in various microbial genomes, of which a few are known for their glycosylation specificity and efficiency (Dirk et al., 2002; Blanchard et al., 2002; Luzhetskyy et al., 2008). The use of efficient glycosyl-transferases from among these unexplored glycosyltransferase sequences is a promising strategy for the biosynthesis of glycoside compounds. Based on the CAZy database annotation of the DIAMOND gene sequences, glycoside transferases were identified as the most abundant, followed by glycoside hydrolases, and the fold difference between the two was as high as 25. Therefore, these two enzymes play an important role in glycoside metabolism. This finding also indicates that glycoside metabolism is of great importance for micropathogens infecting hosts and for protecting the body from injury.

The distinct geographical environment and climatic conditions in Heilongjiang Province may influence the diversity of resident species. To characterize the potential microbial population characteristics of different tick species at the regional scale, we used PCR to detect some major micropathogens, such as *Simian foamy virus, Crimean-Congo hae-morrhagic fever orthonairovirus*, *C. burnetii, B. burgdorferi* and *A. phagocytophilum* in *H. japonica, I. persulcatus, D. silvarum* and *H. concinna* tick samples from four regions in Heilongjiang. Approximately 45.62% of the four tick species specimens analysed were positive for *A. phagocytophilum* and *C. burnetii. B. burgdorferi* was only detected in the G1 sample, and *Crimean-Congo haemorrhagic fever orthonairovirus* was only detected in the G1 and G3 samples. However, we were unable to detect any genes belonging to *Simian foamy virus*. Furthermore, we did not detect any genes from the causative agents responsible for Crimean-Congo haemorrhagic fever in the wild *I. persulcatus* and *H. concinna* ticks. In contrast, *A. phagocytophilum* and *C. burnetii* were clearly detected in different tick species isolated from Heilongjiang Province. This result suggests that these micropathogens might have increased potential to lead to epidemic outbursts, especially where their distribution is high. However, further experiments are necessary to verify that some of these microbial agents can be transmitted in the wild.

Based on the 16S rRNA gene, *C. burnetii* and *B. burgdorferi* are highly conserved pathogens and show only slight differences in sequence among various strains, indicating the potential for the detection prevention control of these pathogens. In addition, the genotype MSP4 from *A. phagocytophilum* is obviously divided into three major types, of which type I is relatively limited in distribution, while types II and III are widely distributed worldwide and are also the primary pathogens causing zoonotic diseases. Although virus sequences are simple, they vary based on the environment or host, allowing them to adapt in different environments, as has been observed for *Simian foamy virus* and *Crimean-Congo haemorrhagic fever orthonairovirus*.

5. Conclusion

In the present study, for the first, time TGMS methodologies were used for micropathogen discovery in wild-caught tick vectors. The results suggest that metagenomic sequencing can facilitate the identification of numerous micropathogens in the microbiomes of wild-caught ticks. This approach could be used to not only monitor microbial communities in infectious insect vectors but also as an ideal tool for the surveillance of novel emerging bacterial and viral diseases. Finally, a more thorough understanding of the ecological factors associated with the prevalence and persistence of micropathogen lineages associated with vectors will ultimately aid in the prediction and prevention of the spread of disease. Finally, the technology used to analyse the micropathogens in ticks by TGMS will ultimately help to identify potentially novel micropathogens and predict and prevent the spread of disease.

Author contributions

G.L. and J.L. designed the experiments. Q.R. and J.L. performed the experiments. J.L. and H.Y. analysed the data. J.L., H.S., and X.L. wrote the manuscript. G.G., B.Z., G.L., X.Q., Y.T., and M.S. collected experimental materials. All authors read and approved the final version of the manuscript.

Funding

Declaration of competing interest

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

Acknowledgements

References

de la Fuente, J., Massung, R.F., Wong, S.J., Chu, F.K., Lutz, H., Meli, M., von Loewenich, F.D., Grzeszczuk, A., Torina, A., Caracappa, S., Mangold, A.J., Naranjo, V., Stuen, S., Kocan, K.M., 2005. Sequence analysis of the msp4 gene of *Anaplasma phagocytophilum* strains. J. Clin. Microbiol. 43, 1309–1317.

Dumler, J.S., Barbet, A.F., Bekker, C.P., Dasch, G.A., Palmer, G.H., Ray, S.C., Rikihisa, Y., Rurangirwa, F.R., 2001. Reorganization of genera in the families Rickettsiaceae and Anaplasmataceae in the order Rickettsiales: unification of some species of Ehrlichia with Anaplasma, Cowdria with Ehrlichia and Ehrlichia with Neorickettsia, descriptions of six new species combinations and designation of *Ehrlichia equi* and 'HGE agent' as subjective synonyms of *Ehrlichia phagocytophila*. Int. J. Syst. Evol. Microbiol. 51, 2145–2165.

Eisen, R.J., Kugeler, K.J., Eisen, L., Beard, C.B., Paddock, C.D., 2017. Tick-borne zoonoses in the United States: persistent and emerging threats to human health. ILAR J. 58, 319–335.

Huang, F.C., Giri, A., Daniilidis, M., Sun, G., Härtl, K., Hoffmann, T., Schwab, W., 2018. Structural and functional analysis of UGT92G6 suggests an evolutionary Link between mono- and disaccharide glycoside-forming transferases. Plant Cell Physiol. 59, 857–870.

Ismail, W.M., Ye, Y., Tang, H., 2014. Gene finding in metatranscriptomic sequences. BMC Bioinf. 15 (Suppl. 9), S8.

Latrofa, M.S., Iatta, R., Toniolo, F., Furlanello, T., Ravagnan, S., Capelli, G., Schunack, B., Chomel, B., Zatelli, A., Mendoza-Roldan, J., Dantas-Torres, F., Otranto, D., 2020. A molecular survey of vector-borne pathogens and haemoplasmas in owned cats across Italy. Parasites Vectors 13, 116.

Lavezzo, E., Barzon, L., Toppo, S., Palù, G., 2016. Third generation sequencing technologies applied to diagnostic microbiology: benefits and challenges in applications and data analysis. Expert Rev. Mol. Diagn 16, 1011–1023.

Li, Q.F., Niu, D.S., Wen, B., Chen, M.L., Qiu, L., Zhang, J.B., 2005. Protective immunity against Q fever induced with a recombinant P1 antigen fused with HspB of *Coxiella burnetii*. Ann. N. Y. Acad. Sci. 1063, 130–142.

Livak, K.J., Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods 25, 402–408.

Luo, J., Liu, M.X., Ren, Q.Y., Chen, Z., Tian, Z.C., Hao, J.W., Wu, F., Liu, X.C., Luo, J.X., Yin, H., Wang, H., Liu, G.Y., 2017. Micropathogen community analysis in *Hyalomma rufipes* via high-throughput sequencing of small RNAs. Front Cell Infect Microbiol 7, 374.

Luo, J., Ren, Q.Y., Chen, Z., Liu, W.G., Qu, Z.Q., Xiao, R.H., Chen, R.G., Lin, H., Wu, Z.G., Luo, J.X., Yin, H., Wang, H., Liu, G.Y., 2019. Comparative analysis of microRNA profiles between wild and cultured *Haemaphysalis longicornis* (Acari, Ixodidae) ticks. Parasite 26, 18.

Luzhetskyy, A., Bechthold, A., 2008. Features and applications of bacterial glycosyltransferases: current state and prospects. Appl. Microbiol. Biotechnol. 80, 945–952.

Maggi, R.G., Toliver, M., Richardson, T., Mather, T., Breitschwerdt, E.B., 2019. Regional prevalences of *Borrelia burgdorferi*, *Borrelia bissettiae*, and *Bartonella henselae* in Ixodes affinis, *Ixodes pacificus* and *Ixodes scapularis* in the USA. Ticks Tick Borne Dis 10, 360–364.

Qi, Y., Yin, Q., Shao, Y., Li, S., Chen, H., Shen, W., Rao, J., Li, J., Li, X., Sun, Y., Lin, Y., Deng, Y., Zeng, W., Zheng, S., Liu, S., Li, Y., 2018. Rapid and visual detection of *Coxiella burnetii* using recombinase polymerase amplification combined with lateral flow strips. BioMed Res. Int. 2018, 6417354.

Reid, M.J.C., Switzer, W.M., Schillaci, M.A., Klegarth, A.R., Campbell, E., Ragonnet-Cronin, M., Joanisse, I., Caminiti, K., Lowenberger, C.A., Galdikas, B.M.F., Hollocher, H., Sandstrom, P.A., Brooks, J.I., 2017. Bayesian inference reveals ancient origin of *simian foamy virus* in orangutans. Infect. Genet. Evol. 51, 54–66.

Rhoads, A., Au, K.F., 2015. PacBio sequencing and its applications. Dev. Reprod. Biol. 13, 278–289.

Riemersma, K.K., Komar, N., 2015. Heartland virus neutralizing antibodies in vertebrate wildlife, United States, 2009-2014. Emerg. Infect. Dis. 21, 1830–1833.

Sontag, M.K., Wright, D., Beebe, J., Accurso, F.J., Sagel, S.D., 2009. A new cystic fibrosis newborn screening algorithm: IRT/IRT1 upward arrow/DNA. J. Pediatr. 155, 618–622.

Tekin, S., Bursali, A., Mutluay, N., Keskin, A., Dundar, E., 2012. *Crimean-Congo hemorrhagic fever virus* in various ixodid tick species from a highly endemic area. Vet. Parasitol. 186, 546–552.

Wagner, J., Coupland, P., Browne, H.P., Lawley, T.D., Francis, S.C., Parkhill, J., 2016. Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. BMC Microbiol. 16, 274.

Xu, B., Liu, L., Huang, X., Ma, H., Zhang, Y., Du, Y., Wang, P., Tang, X., Wang, H., Kang, K., Zhang, S., Zhao, G., Wu, W., Yang, Y., Chen, H., Mu, F., Chen, W., 2011. Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. PLoS Pathog. 7, e1002369.

Yang, Y., Li, X., Zhang, Y., Liu, J., Hu, X., Nie, T., Yang, X., Wang, X., Li, C., You, X., 2020. Characterization of a hypervirulent multidrug-resistant ST23 Klebsiella pneumoniae carrying a bla CTX-M-24 IncFII plasmid and a pK2044-like plasmid. J Glob Antimicrob Resist 22, 674–679.