Large-scale comparative analyses of tick genomes elucidate their genetic diversity and vector capacities

Na Jia[1,27,31], Jinfeng Wang[2,5,31], Wenqiang Shi[1,31], Lifeng Du[2,4,31], Yi Sun[1,31], Wei Zhan[3], Jia-Fu Jiang[1,27], Qian Wang[1,4], Bing Zhang[2], Peifeng Ji[2], Lesley Bell-Sakyi[6], Xiao-Ming Cui[1,27], Ting-Ting Yuan[1], Bao-Gui Jiang[1],Wei-Fei Yang[3],Tommy Tsan-Yuk Lam[7], Qiao-Cheng Chang[8], Shu-Jun Ding[9], Xian-Jun Wang[9], Jin-Guo Zhu[10], Xiang-Dong Ruan[11], Lin Zhao[1,4], Jia-Te Wei[1,4], Run-Ze Ye[1,4],Teng Cheng Que[12], Chun-Hong Du[13], Yu-Hao Zhou[1], Jing Xia Cheng[14], Pei-Fang Dai[14], Wen-Bin Guo[1], Xiao-Hu Han[15], En-Jiong Huang[16], Lian-Feng Li[1], Wei Wei[1], Yu-Chi Gao[3], Jing-Ze Liu[17], Hong-Ze Shao[18], Xin Wang[19], Chong-Cai Wang[20], Tian-Ci Yang[21], Qiu-Bo Huo[22], Wei Li[23], Hai-Ying Chen[24], Shen-En Chen[24], Ling-Guo Zhou[25], Xue-Bing Ni[7], Jun-Hua Tian[26], Yue Sheng[1], Tao Liu[3], Yu-Sheng Pan[1], Luo-Yuan Xia[1], Jie Li[1], Tick Genome and Microbiome Consortium (TIGMIC), Fangqing Zhao[2,5,27,28,29,30*], Wu-Chun Cao[1,4,27,32*].

1. State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, P. R. China
2. Computational Genomics Lab, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, P. R. China
3. Annoroad Gene Technology (Beijing) Company Limited, Beijing 100176, P. R. China
4. Institute of EcoHealth, School of Public Health, Shandong University, 44 Wenhuaxi Street, Jinan 250012, Shandong, P. R. China
5. State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, P. R. China
6. Department of Infection Biology, Institute of Infection and Global Health, University of Liverpool, Liverpool L3 5RF, UK
7. State Key Laboratory of Emerging Infectious Diseases and Centre of Influenza

Research, School of Public Health, The University of Hong Kong, Hong Kong SAR
8. College of Animal Science and Veterinary Medicine, Heilongjiang Bayi Agricultural University, Daqing 163319, Heilongjiang, P. R. China
9. Shandong Center for Disease Control and Prevention, Shandong Provincial Key Laboratory of Communicable Disease Control and Prevention, Jinan 250014, Shandong, P. R. China
10. ManZhouLi Customs District, Manzhouli 021400, Inner Mongolia, P. R. China
11. Academy of Forest Inventory and Planning, State Forestry and Grassland Administration, Beijing 100714, P. R. China
12. Guangxi Zhuang Autonomous Region Terrestrial Wildlife Medical-aid and Monitoring Epidemic Diseases Research Center, Nanjing 530028, Guangxi, P. R. China
13. Yunnan Institute for Endemic Diseases Control and Prevention, Dali 671000, Yunnan, P. R. China
14. Shanxi Provence Center for Disease Control and Prevention, Xian 030012, Shanxi, P. R. China
15. Shenyang Agriculture University, Shenyang 110866, Liaoning, P. R. China
16. Fuzhou International Travel Healthcare Center, Fuzhou 350001, Fujian, P. R. China
17. Key Laboratory of Animal Physiology, Biochemistry and Molecular Biology of Hebei Province, College of Life Sciences, Hebei Normal University, Shijiazhuang 050024, Hebei, P. R. China
18. Animal Husbandry and Veterinary Science Research Institute of Jilin Province, Changchun 130062, Jilin, P. R. China
19. Qingjiangpu District Center for Disease Control and Prevention, Huai'an 223001, Jiangsu, P. R. China
20. Hainan International Travel Healthcare Center, Haikou 570311, Hainan, P. R. China
21. State Key Lab of Mosquito-borne Diseases Hangzhou International Tourism

Healthcare Center Hangzhou Customs of China; Hangzhou 310012, Zhejiang, P. R. China
22. Mudanjiang Forestry Central Hospital, Mudanjiang 157000, Heilongjiang, P. R. China
23. Xinjiang Center for Disease Control and Prevention, Urumqi 830002, Xinjiang, P. R. China
24. The Collaboration Unit for Field Epidemiology of State Key Laboratory for Infectious Disease Prevention and Control, Nanchang Center for Disease Control and Prevention. Nanchang 330038, Jiangxi, P. R. China
25. Shaanxi Natural Reserve and Wildlife Administration Station, Xi'an 710082, Shaanxi, P. R. China
26. Wuhan Center for Disease Control and Prevention, Wuhan 430015, Hubei, P. R. China.
27. Research Unit of Discovery and Tracing of Natural Focus Diseases, Chinese Academy of Medical Sciences, Beijing 100071, P. R. China
28. Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, 650223, Yunan, P. R. China
29. Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou, Zhejiang, P. R. China
30. University of Chinese Academy of Sciences, Beijing 100049, P. R. China
31. These authors contributed equally to this work
32. Lead contact

* Corresponding authors:

Prof. Wu-Chun Cao, State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing 100071, China. Email: caowc@bmi.ac.cn

Prof. Fangqing Zhao, Computational Genomics Lab, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China. Email: zhfq@biols.ac.cn

Summary

Among arthropod vectors, ticks transmit the most diverse human and animal pathogens, leading to an increasing number of new challenges worldwide. Here, we sequenced and assembled the high-quality genomes of six ixodid tick species and further resequenced 678 tick specimens to understand three key aspects of ticks: genetic diversity, population structure and pathogen distribution. We explored the genetic basis common to ticks, including heme and hemoglobin digestion, iron metabolism, and reactive oxygen species, and unveiled for the first time that both genetic structure and pathogen composition in different tick species were mainly shaped by ecological and geographic factors. We further identified species-specific determinants associated with different host ranges, life cycles and distributions. The findings of this study provide an invaluable resource for research and control of ticks and tick-borne diseases.

Ticks (Acari: Ixodida), which are obligate blood-feeding arthropods, are distributed all over the world from tropic to subarctic regions, with the oldest records dating back to the mid-late Cretaceous (Anderson and Magnarelli, 2008; Peñalver et al., 2018). Ticks are the most versatile vectors, capable of transmitting the broadest spectrum of pathogens, including bacteria, protozoa, fungi, nematodes and viruses, to humans, livestock and wildlife. More than 28 tick species are known to cause a variety of human diseases, such as Lyme disease and spotted fever group rickettsiosis (Jongejan and Uilenberg, 2004), even causing deaths due to misdiagnosis and delayed treatment. Persistent and relapsing infections, as well as long-term sequelae caused by tick-borne pathogens, further worsen the quality of human health (Krause et al., 2008; Mac et al., 2019). Furthermore, the global economic burden in animal husbandry due to tick-borne infections is very large. For instance, the most notorious veterinary ectoparasite, *Rhipicephalus microplus*, is estimated to lead to an annual loss of US$ 2.5 billion throughout tropical and subtropical regions (Barker and Walker, 2014).

The threats of tick-borne diseases (TBDs) to human health have unpredictably increased with contemporary urbanization, deforestation, climate change and the rapidly changing interactions between people, animals and their respective habitats. A recent example is the exotic disease vector *Haemaphysalis longicornis*, which has infested multiple states in the United States (Beard et al., 2018) and caused great concern. Even worse, the surging number and geographic expansion of emerging TBDs have caused social anxiety due to unknown health consequences and the lack of approaches to control their transmission. Therefore, fundamental knowledge of tick genomes and genetic diversity is urgently needed, which will undoubtedly open new avenues for research on tick biology, vector-pathogen interactions, disease transmission and control strategies.

The first tick genome sequenced, that of *Ixodes scapularis*, offered a glimpse into the genetic architecture and genomic features of the tick (Gulia-Nuss et al., 2016). However, different tick species adapt to diverse environmental niches, feed on diverse

hosts ranging from reptiles to mammals and birds, and exhibit complex and distinct life cycles. The dominant tick species across China, including *Ixodes persulcatus*, *Haemaphysalis longicornis*, *Dermacentor silvarum*, *Hyalomma asiaticum*, *Rhipicephalus sanguineus*, and *Rhipicephalus microplus*, have their species-specific characteristics. For example, *Hae. longicornis* is a widely distributed tick species indigenous to eastern Asia, whereas *Hy. asiaticum* prefers to live in desert or semidesert environments (Figure 1A). *R. microplus* has a typical one-host cycle, while most others are three-host ticks depending on the number of host animals they attach themselves to during their life cycle (Figure 1B). Therefore, to better understand their genetic complexity and reveal the links between the genomic variation and geographic distribution, ecological adaptation and vector capacity of ticks, we performed large-scale comparative analyses of 684 ixodid tick genomes, representing six dominant tick species across China (Figures S1-S2).

Six high-quality ixodid tick reference genomes

We used larvae of above-mentioned six representative ixodid ticks for *de novo* genome sequencing. We first constructed ≥15 Kb DNA libraries for the PacBio Sequel System and generated 162~303 Gb of subreads with high sequencing depth (approximately 67~95×) (Table S1). Considering the relatively high error rate of PacBio sequencing, we further constructed short-fragment libraries (350 bp) and sequenced them using the Illumina HiSeq X-Ten platform, which generated 106~134 Gb of clean reads (Table 1). We used these high-quality short reads to perform K-mer frequency analyses to estimate the genome sizes (Table S1) and to correct the short indels and substitutions in the PacBio assembly. To further improve the continuity of the assembled tick genomes and anchor the assemblies into chromosomes, we used Hi-C data to order and orient the contigs as well as to correct misjoined sections and merge overlaps (Figure 1C). Finally, we assembled six tick genomes, achieving 8,620~15,174 contigs with scaffold N50 lengths of 533~208,696 Kb and contig N50 lengths of 340~1,800 Kb (Table 1; Table S1). Subsequently, we used Benchmarking Universal Single-Copy Orthologs (BUSCO) and the proportion of properly aligned

Illumina paired-end reads to evaluate the completeness of these assemblies, which further demonstrated their high completeness and accuracy (Table 1).

By combining *de novo* and homology-based approaches, 52.6~64.4% of the repetitive elements were identified from these six assembled tick genomes (Table 1), which is comparable to that from the latest available genome of *I. scapularis* embryonic 6 (ISE6) cell line (~63.5%) (Miller et al., 2018). Among the annotated repeats, LINE and LTR constituted the most abundant known repeat families, representing 8.6~18.3% and 6.5~16.1% of the repetitive sequences, respectively (Table S1). By combining transcriptome-based, homology-based, and *ab initio* approaches, 25,718~29,857 protein-coding genes were predicted from these tick genomes (Table 1). The gene numbers are slightly larger than those predicted in *I. scapularis* and two closely related species, namely, *Centruroides sculpturatus* (bark scorpion) and *Parasteatoda tepidariorum* (common house spider) (Thomas et al., 2018) (Table S1), which could be explained by the high completeness and accuracy of the assembled genomes as well as the pairwise homology searches among these six tick species. The average gene length varied greatly among the six tick species, from the smallest (6,466 bp) in *Hae. longicornis* to the largest (15,067 bp) in *I. persulcatus*, with 3.0~4.8 exons per gene and an average intron length of 2,754~3,760 bp (Table S1), indicating the substantial differences in genetic structure among these ticks.

To further elucidate the genetic diversity of these tick species, we compared the chromosome size, abundance of repetitive elements, gene content, GC content, noncoding RNA content and synteny of these six tick genomes (Figure 1D). *D. silvarum* had the largest genome size and the largest chromosome 1 (> 452 Mb), ~100 Mb larger than those of the other species (Table S1). In contrast, the genome size and gene content of *I. persulcatus* were the lowest, while its repetitive elements and noncoding contents were the highest. The GC content was relatively similar across different tick species. Among the six sequenced tick genomes, *I. persulcatus* exhibited very low conserved synteny, which reflects its high genetic divergence from the other tick species. To calculate the evolutionary distances of the six tick species and *I.*

scapularis from arachnids, orthologous protein sequences were obtained from these species and two outgroup species, *C. sculpturatus* and *P. tepidariorum*, and used to construct a maximum likelihood tree. The divergence time was estimated based on the coding sequences of 464 single-copy orthologous genes. As shown in Figure 1E, the phylogenetic analysis divided the ticks into two main clades, with the two ixodids (*I. scapularis* and *I. persulcatus*) closely related to each other and sharing a common ancestor ~200 million years ago (MYA) with the other five ticks. *Hae. longicornis, R. microplus*，*R. sanguineus*, *Hy. asiaticum*, and *D. silvarum* were clustered together and differentiated from a common ancestor about 137.8 MYA. This genome-based phylogeny constitutes mutual confirmation with the morphological evolutionary tree for ticks (Hoogstraal and Aeschlimann, 1982).

Essential genetic basis of tick hematophagy and the related phenotype

The six sequenced genomes provide a unique resource for understanding the genetic basis of tick hematophagy through comparative genomics and transcriptomics analysis. Through protein family (Pfam) domain-based comparison of the six ticks with *I. scapularis* tick (Miller et al., 2018), three other blood-feeding arthropods (*Anopheles gambiae*, *Aedes aegypti*, and *Glossina morsitans*) and two arachnids (*P. tepidariorum* and *C. sculpturatus*), we found that protein families implicated in peptidase activity, transferase activity, transcription regulator activity, transmembrane transporter activity and immunity have notable expansions in ticks (Figure 2A; Table S2). Most of these protein families are relevant to the blood-sucking process. For example, 3~15-fold proliferation of peptidase family M13, ABC-2 family transporter protein, serine protease inhibitor, and glutathione S-transferase occurred in tick genomes (Table S2); these families are involved in hemoglobin digestion, heme transport, blood coagulation, fibrinolysis, detoxification, and oxidative stress (Dickinson and Forman, 2002; Horn et al., 2009; Lara et al., 2015; Rubin, 1996).

Long attachment time to the host (several days to weeks), large volume of blood meal (hundreds of times its unfed weight), and broad meal source range (the blood of almost all terrestrial animals) are unique traits of hematophagous ticks and should be

involved in many physiological processes, including detoxification of xenobiotic factors, host questing, blood meal digestion, nutrient metabolism, and immune response (Figure 2B). The six tick genomes sequenced in this study provided strong evidence that unlike most eukaryotes (Braz et al., 1999; Gulia-Nuss et al., 2016; Perner et al., 2016), blood-dependent ticks have lost most genes encoding heme biosynthesis and degradation, making them strictly dependent on exogenous sources of heme from the host (Table S3). Thus, ticks are likely to have evolved to acquire and transport heme and iron for vitally important physiological processes and at the same time to maintain redox homeostasis, where free heme and iron can catalyze the generation of reactive oxygen species (ROS). To investigate the potential mechanism associated with iron homeostasis, we surveyed the gain and loss of iron metabolism-related genes in tick genomes and found that the transmembrane protease serine 6 family of matripase-2 (TMPRSS6) was significantly expanded (Table S3). In addition, genes associated with antioxidant enzymes, radical scavengers, or heme-mediated activators associated with ROS were mostly conserved across all tick species (Figure 2C, Table S3). This further indicated the importance of maintaining antioxidant systems for ticks, on the one hand to avoid oxidative stress and on the other hand to affect pathogen transmission indirectly by changing its balance with other microbes, as reported in mosquitos (Cirimotich et al., 2011; Kumar et al., 2010; Oliveira et al., 2011). Furthermore, genes related to immune systems and interactions with pathogens were relatively conserved (Figure 2D; Tables S3), which suggests that ticks may have evolved multiple cellular and humoral immunities to achieve success at the tick-host interface and to maintain a balance at the tick-pathogen interface. In addition, we observed the absence of many genes (*Imd, Fadd, Dredd*) in the immune deficiency pathway (Table S3), which is essential for recognition and response to Gram-negative bacteria in *Drosophila* (Palmer and Jiggins, 2015), indicating a different strategy of immunological defense against microbes between ticks and fruit flies.

We further performed comparative transcriptomic analysis between unfed and fed

ticks and found that the differentially expressed genes in various ticks were all enriched in functions of heme and iron ion binding, oxidoreductase activity, and chitin metabolic process (Figure 2E). For example, the upregulated genes in TMPRSS6 family exhibited 3~97 fold change during blood sucking in all ticks. The results further elucidate the common genetic basis for tick blood feeding and highlight the importance of these mechanisms for their parasitic lifestyle. Considering that genes after duplication tend to be nonfunctionalized, neofunctionalized or subfunctionalized (Sandve et al., 2018), we explored their expression changes between unfed and fed ticks and found that duplicated genes in larger gene families exhibited a significantly larger standard deviation of fold change than those in smaller gene families (Spearman's rank correlation test, $p < 0.001$), indicating the diversification of these homologous genes in blood-feeding after gene expansion.

We next explored the genomic features associated with the species-specific traits that are critical for vector control, including evolutionary distance, host range, geographic distribution and life cycle. *I. persulcatus* in the Prostriata clade evolved much earlier and parasitizes a more diverse range of host groups than the other five tick species (Beati and Klompen, 2019; Hoogstraal and Aeschlimann, 1982). A notable expansion of gene families associated with blood meal digestion, detoxification of xenobiotic factors (such as acaricides, poisons, and environmental pollutants), and nutrient metabolism including serine carboxypeptidase, TMPRSS6, cytochrome P450, and alcohol dehydrogenase etc., was found in *I. persulcatus* (Figures 2C-2D; Table S3). These expansions may confer to *I. persulcatus* additional advantages for nutrient acquisition and endogenous/exogenous detoxification during blood feeding. *Hae. longicornis* has the widest geographic distribution (Figure 1A) and was recently detected in the United States (Beard et al., 2018). We discovered the expansion of known gene families implicated in blood feeding by comparative genomic analyses in *Hae. longicornis* (Figures 2C-2D), which may account for its adaptation to colonize diverse habitats and ecological niches.

Another distinguishing trait of ticks is their life cycle. *R. microplus* has a typical

one-host cycle. The expanded chemosensory gene family, e.g., the ionotropic receptors (IRs) (Figures 2F-2G; Table S3), which have been associated with a variety of sensory functions (Eyun et al., 2017), may facilitate the strict parasitization by *R. microplus* of the same host in each developmental stage. In addition, cytochrome P450 genes, encoding a major family of enzymes involved in the detoxification of xenobiotics, were strikingly reduced (Figure 2G; Table S3). The down-regulations genes after blood meal in RNA-seq differential expression analysis were also enriched in P450 gene families of *R. microplus* (Fisher's exact test, $p = 0.03$).Those may be potentially attributed to *R. microplus* one-host life cycle and a lack of selection pressure.

Population structure and genetic diversity of six tick species

Population evolution is particularly challenging for ticks, as their life cycle consists of long off-host periods (months to years) in changing environments and because of their great reproductive potential, with thousands of eggs being laid after repletion. The genetic diversity of ticks is largely unknown due to the lack of genomic data from different habitats. With the advantage of having acquired six high-quality genomes, we resequenced 678 wild-caught specimens of the six tick species across 27 provinces, metropolises or autonomous regions of mainland China, spanning eight ecogeographical faunas and a variety of ecological settings, including coniferous forest, steppe, farmland, desert, shrubland and tropical forest (Figure 1A). Maximum likelihood trees based on full mitochondrial sequences and nuclear single nucleotide variants within single-copy genes were constructed to explore the population structure and genetic diversity among these tick individuals.

Through comparison of the six population structures, we found that different tick species have evolved a common dispersal strategy. An ecogeographical distribution pattern was observed for *I. persulcatus*, *D. silvarum*, *Hy. asiaticum*, and *R. sanguineus* (Figure 3A; Figure S3). *I. persulcatus* was relatively restricted in the boreal coniferous forest and temperate forest; *D. silvarum* detected in Shanxi formed a subdivision; The morphologically indistinguishable *R. sanguineus* could be mainly

subdivided into two clades, one thriving in tropical forest or shrubland and the other in farmland; and *Hy. asiaticum* was distributed in the same ecological fauna but was geographically differentiated between Xinjiang and Inner Mongolia. Although further investigations of diverse ecosystems, different hosts and larger datasets are needed for broader generalization of these results, our findings suggest that the local adaptation to different ecological niches coupled with geographic distance by restriction of active tick movement can explain the observed patterns of population subdivision in ticks.

Hae. longicornis is particularly interesting because it is capable of rapidly invading new areas and explosively proliferating in established ranges (e.g., recent invasion to the USA). A very close genetic distance of the *Hae. longicornis* population was observed in the phylogenetic analysis, although this species had a wide geographic distribution occupying diverse ecosystems (Figures 3A-3C). Population structure models supported the division of *Hae. longicornis* into one major population and one minor population (Figure 3A). The major domestic population lacked clear geographic structuring, which suggested that this species was selected for dispersion rather than local competitiveness, which prevented selection for locally adapted phenotypes. The minor population was mainly from three provinces (Fujian, Shanghai and Jiangsu) along the southern coastline of China (Figure 3B). Compared with the major population, the minor was close to the ancestral root of the phylogenetic tree and shared a high similarity with strains from New Zealand (Guerrero et al., 2019) (Figure 3C). Understanding the contribution of migrating birds to the domestic and overseas movement of *Hae. longicornis* is warranted for further dissection of the dispersion of this vector population.

As a tick with a typical one-host cycle, *R. microplus* has a distinct population structure and gene flow compared with three-host ticks. We found that *R. microplus* can be clustered into three major clades which largely correspond to their geographical subdivisions: Clade 1 includes specimens from Southwest China (Yunnan), Clade 2 from Southeast China (Hainan and Guangdong to Jiangxi and Fujian) and Clade 3 from South Central China (Guizhou and Chongqing to Hubei,

Hunan, Anhui and Zhejiang) (Figure 3B; Figure S4). Comparison of the branches from different provinces showed high F_{ST} values (>0.50), indicating the high genetic differentiation among various *R. microplus* populations in China. Interestingly, phylogenetic analysis based on mitochondrial sequences showed some differences in the tree topology compared with that based on nuclear genome sequences (Figures 3A, 3C), indicating distinct paternal and maternal population structures and migration patterns within this species. We speculated that the host specificity within this species may drive local selection patterns of *R. microplus* and greatly alter its population structure (Araya-Anchetta et al., 2015). We also detected extensive gene gain-and-loss events among three subdivisions of *R. microplus* and found that the discriminated genes were enriched in pathways related to the regulation of epithelial cell proliferation and NF-κB (Figures 3D-3E). The top discriminated genes, such as ubiquitin protein ligase and mucin-6-like protein, indicated some differences of immune response among the three clades.

Key drivers of pathogen distribution in ticks

The complex genomic diversity among tick species implies complicated tick-pathogen interactions, which prompted us to further understand the tick-borne pathogen ecology and evolution. We evaluated the impacts of host gene flow on pathogen distribution by metagenomic analysis of the six tick species. Host DNA contamination could be effectively removed by using the six tick genomes obtained. After filtering the host sequences by mapping the sequencing reads to tick genome assemblies, microbial composition analysis and pathogen identification were performed for each of the 678 specimens.

The tick taxonomy is an important factor in defining the potential of a tick to transmit pathogens. Our study for the first time unveiled the landscape of pathogens carried by six tick species collected from a wide range of geographical sources. In general, the relative abundance of certain pathogens was quite different across the six tick species (Figure 4A). *I. persulcatus* and *Hae. longicornis*, traditionally the most important vectors of human and animal diseases (Fang et al., 2015), were found

bearing various bacterial species of *Anaplasma*, *Babesia*, *Borrelia*, *Coxiella*, *Ehrlichia*, and *Rickettsia* (Figure 4A). In contrast, *R. sanguineus* had the lowest abundance of bacterial pathogens. *R. microplus*, which transmits *Babesia* and *Anaplasma* in livestock and wild ruminants, possesses a *Coxiella*-like endosymbiont as the most abundant bacterial taxon (Figure 4A). Notably, *D. silvarum* presented the largest relative abundance of *Rickettsia* (Figure 4A). *Hy. asiaticum* carried the highest relative abundance of *Coxiella burnetii* and *Francisella tularensis* (Figure 4A), the causative agents of Q fever and tularemia, respectively.

The interplay among humans, animals and ecosystems is well acknowledged. However, the driving factors of interactions among the environment, pathogens, vectors and hosts have not yet been clearly addressed for TBDs. Each geographical fauna has specific ecological features and thus favors different forms of animal life. We observed that the bacterial distribution had an overall correlation with the ecogeographical faunal region for a given tick species (Figure 4A). For example, the relative abundances of *Anaplasma* and *Ehrlichia* in *R. sanguineus* were lower in tropical forest and shrubland areas than in farmland faunal regions ($p < 0.05$, Mann–Whitney U test); for *D. silvarum*, nonpathogenic *Anaplasma* was prevalent in North China, and for *R. microplus*, *Rickettsia* was prevalent in Southwest China (Figure 4A) ($p < 0.001$, Mann–Whitney U test). To quantify the microbial divergence across regions, we compared the Bray-Curtis (BC) dissimilarities of the tick microbiota between different geographic faunas and within the same geographic fauna (Figure 4B), and found that the calculated BC dissimilarities varied by geographic distance for each tick species (Figure 4C). We found that the more the geographic fauna or distance diverged, the larger the tick microbiota dissimilarity was, and such a pattern may consequently impact the pathogen distribution. In addition to above key drivers, we also found that different subtypes of *R. microplus* and *Hae. longicornis* exhibited different positive rates of *Rickettsia* ($p < 0.001$ for *R. microplus* and $p < 0.05$ for *Hae. longicornis*, Kruskal-Wallis test) (Figure 4A), further indicating the necessity of determining and monitoring the tick subspecies or subpopulations with more pathogen

load.

We further summarized all the reported human cases with TBDs in China from 1980 to 2020 (Figure 4D; Figure S5; Table S4). During the past 40 years, at least 22 diseases caused by tick-borne bacteria or protozoa have been reported. The northeastern China is a high-risk area where about 15 pathogens, half of which were emerging agents, have caused human infections (Jia et al., 2018; Jia et al., 2013; Jia et al., 2014; Jiang et al., 2018; Jiang et al., 2015; Li et al., 2015). We mapped the abundance and proportion of pathogens of different tick species onto their collection sites (Figure 4D). By overlapping the distributions of TBDs and detected pathogens in ticks, we found that pathogenic *Rickettsia* had both the high prevalence and large abundance in the ticks from the Northeastern China, where spotted fever group rickettsioses were frequently diagnosed. However, the abundance of tick pathogens does not strictly correlate with their transmission rate to human in general. Besides the reason that the identified pathogens in ticks of this study may not be at infectious stage when they were sampled, another possible explanation is that there might be under-reported cases of TBDs due to lack of etiologic diagnosis tests in many endemic areas. It should be noted that although the abundance of *Borrelia* was only 3% of that of *Rickettsia* in the ticks in Northeastern China, it has caused the disease incidence as high as *Rickettsia* (Figures 4E-4F). Taken together, these findings suggest that pathogen abundance may not be the sole factor in determining the risk of human infection, which highlights the necessity of more sensitive approaches to identify the low abundance pathogens in ticks.

In conclusion, the genomes of six representative species generated in this study provide novel insights into tick-specific blood feeding life, tick-pathogen interactions and the development of genetic tools for tick control. The large-scale genomic re-sequencing of 678 wild-caught tick specimens further unveils the high genetic heterogeneity of ticks, reflecting their local adaptation to diverse ecological niches. Based on metagenome profiling and pathogen screening of these tick specimens, we described the landscape of microbial pathogens, including some emerging human

pathogens, carried by six tick species collected from a wide range of geographical sources. The pathogen composition in different tick species is mainly shaped by ecological and geographic factors, and different subpopulations may have diverse tick-borne pathogen profiles. We believe the tick genomes and their associated pathogen profiles generated in this study will undoubtedly benefit the community on global tick and TBD control.

ACKNOWLEDGEMENTS

We thank all the members in the Tick Genome and Microbiome Consortium (TIGMIC) for their help on sample collection. This study is supported by Natural Science Foundation of China (81621005, 81773492, 81760607, 31722031, 31671364) and the State Key Research Development Program of China (2019YFC1200202, 2019YFC1200401, 2018YFC0910400).

AUTHORS CONTRIBUTIONS

W.C.C, F.Z., N.J. designed and supervised research. Y.S., J.F.J., X.M.C., B.G.J., Q.C.C., S.J.D., X.J.W., J.G.Z., X.D.R., T.C.Q., C.H.D., J.X.C., P.F.D., X.H.H., E.J.H., J.Z.L., H.Z.S., X.W., C.C.W., T.C.Y., Q.B.H., W.L., H.Y.C., L.G.Z., J.H.T. collected samples. Q.W, T.T.Y, L.F.L., W.W., L.Y.X., J. L., prepared materials for sequencing. Q.W., L.Z., Y.S., W.B.G., X.B.N. set up the database, W.Z., W.F.Y., Y.C.G, T.L, performed genome sequencing. W.Z., W.F.Y., Y.C.G, T.L, W.S, performed genome assembly and annotation. W.S., L.D., J.W, N.J., F.Z. performed genome analysis and interpretation, J.W. W.S, L.D., N.J., Y.H,Z. R.Z.Y prepared figures and tables. L.B.S. provided the tick cell line and edited the manuscript, N.J., J.W., F.Z. W.C.C. wrote the paper.

DECLARATION OF INTERSTS

The authors declare no competing interests.

Figure Legends

Figure 1 Basic information and genomic comparison of six tick species. (A) Map of sample collections. The size of the circle represents the number of tick samples collected in the area. Geographical fauna were recorded as follows: Northeast China (I); North China (II); Neimenggu-Xinjiang (III); Qinghai-Xizang (IV); Southwest China (V); Central China (VI); South China (VII). Ecological fauna are also shown on the map with different colors. (B) Illustration of ticks with a 3-host life cycle, in which larvae and nymphs feed on blood once before molting, the adults feed once, and then, the fully engorged tick drops from the host and lays thousands of eggs to continue the life cycle. (C) Hi-C interactive heatmap of the genome-wide organization of 11 chromosomes for five ticks. For auxiliary assembly of chromosomes, assemblies were cut into bins of the same length. The effective mapping read pairs between two bins were used as a signal of the strength of the interaction between the two bins. With the numbered chromosomes as the coordinates, the color of each dot represents the log value of the interaction intensity of the corresponding bin pair of the genome, and the interaction intensity increases from yellow to red. Chr represents chromosomes. (D) Comparative genomic analysis of six tick species. From the outer circle to the inner circle, nine types of information, namely, chromosome size, Illumina data coverage, PacBio data coverage, Hi-C data coverage, repeat abundance, gene abundance, GC content, ncRNA, and gene synteny, are labeled successively with the letters a-i. In the synteny analysis, the blue and red lines denote *R. microplus* and *Hae. longicornis*, serving as the reference genome, respectively. (E) Maximum likelihood phylogeny of all sequenced ticks with two species of Arachnoidea as outgroups. The estimated divergence time between clades is labeled on the branch nodes. See also Figures S1-S2.

Figure 2 Genetic basis of tick hematophagy and the related phenotype. (A) Species-specific and shared Pfam family among ticks and other arthropod species. Each cell in the heatmap represents the normalized gene count (across all species on

the left side) of a Pfam family. Only the Pfam that are specific to ticks or common with other blood-feeding arthropod species are shown. Pfams are further grouped according to their functions in biological processes or activities. (B) Unique hematophagous traits of ticks, including detoxification of xenobiotic factors (a), host questing (b), blood meal digestion (c), nutrient metabolism (d), and immune response (e). (C) Gene counts of four gene categories in six tick species: detoxification of xenobiotics (yellow), iron metabolism (deep blue), hemoglobin digestion (green) and oxidative stress (purple). (D) Gene counts of six tick species related to five hematophagous traits of ticks. (E) Gene ontology (GO) enrichment analysis based on the transcriptomic data of unfed and fed ticks. The biological process, cellular component, and molecular function categories are referred to as BP, CC and MF, respectively. From the inner circle to the outer circle, three levels of GO enrichment are displayed with nodes. The sector of the nodes in outermost circle represents the proportion of DE genes in three ticks, namely, *I. persulcatus*, *Hae. longicornis* and *R. microplus*. The sector of the nodes in inner circles represents the absence or presence of DE genes. (F) Gene counts of four different perception pathways to quest preferred hosts in six tick species. (G) Phylogenetic analysis of the IR25a gene (left) and P450 gene group I family (right). The colors of the nodes on the tree represent different tick species. See also Tables S2-S3.

Figure 3 Genetic diversity and population structure of six tick species. (A) Phylogenetic structure of tick populations based on the mitochondrial genome. The subtitle of each tree indicates the species name and the number of specimens. The color of the tree tip represents the ecological fauna type of the sample location. (B) Geographical population structure of *Hae. longicornis* and *R. microplus*. In the top bar plot, each vertical line shows the membership probability of a specimen inherited from each of the inferred ancestral populations (K=5) for *Hae. longicornis*, and specimens are grouped by the sampled province as annotated by the line segment on the top. The bottom plot shows the same information for *R. microplus*. Pie charts on

the map aggregate the same membership probability of ancestral populations for all specimens in each province. Neighboring provinces are connected according to the F_{ST} value between the two provinces. (C) Phylogenetic structures of *Hae. longicornis* (left) and *R. microplus* (right) populations based on their nuclear genomes. The strain previously reported in New Zealand and its close relative were highlighted. (D) Circos plot of genes with elevated copy numbers in the three clades of *R. microplus*. (E) GO enrichment analysis of genes with elevated copy numbers in the three clades of *R. microplus*. The heatmap color represents the adjusted p-value (-log 10). The biological process, cellular component, and molecular function categories are referred to as BP, CC and MF, respectively. See also Figures S3-S4.

Figure 4 Potential pathogen profiling of six tick species. (A) The distribution and abundance of known tick-borne pathogens and their related species in the six tick populations. The relative abundance of the microbes in each sample was estimated by read counts per 100,000 reads. Subtypes of each tick species were classified based on the phylogenetic analysis of the resequenced genomes. Geographic fauna and ecological fauna were selected according to the Chinese fauna classification and were annotated in the corresponding colors. Bacterial species of twelve human pathogenic genera are shown, and each genus name is indicated below the heat map. Human pathogens are annotated in deep gray, and nonhuman pathogens are annotated in light gray. (B) Bray-Curtis dissimilarity between each pair of samples, grouped within the same geographic fauna or between different geographic fauna. (C) Bray-Curtis dissimilarity between each pair of samples varied by geographic distance. (D) Epidemiological distribution of tick-borne disease (TBD) patients and tick pathogens. The cases of human infection were reported between 1980 and 2020. The pies indicate pathogen composition, with the color of circle outline representing tick species. The circle size indicates the relative abundance of all pathogens per 10^5 microbial reads, and the color and area of pies indicate the species and relative abundance of each pathogen, respectively. The Northeastern China is highlighted in

dark gray. (E) The relative abundance (node color) and positive rate of 33 human pathogenic bacteria or protozoa species of the ticks in the Northeastern China. (F) The reported incidences of TBD among the risk population in the Northeastern China. See also Figure S5 and Table S4.

Table 1. Summary of the Assembly and Annotation Information of the Sequenced Tick Genomes

	I. persulcatus	*Hae. longicornis*	*D. silvarum*	*Hy. asiaticum*	*R. sanguineus*	*R. microplus*	*I. scapularis* [a]
Data statistics							
Illumina clean data (Gb)	118.4	115.1	134.1	121.1	105.9	110.1	49.6
Pacbio subreads (Gb)	165.1	303.1	202.7	162.3	183.6	170.6	192.5
Hi-C clean data (Gb)	-	306.0	210.1	201.3	185.5	168.8	-
Assembly statistics							
Contig span (Mb)	1,901.7	2,554.5	2,473.0	1,713.1	2,364.5	2,529.8	2,691.1
Contig N50 (Kb)	532.9	740.0	340.0	555.4	541.9	1,800.7	269.7
Chromosome size (Mb)	-	2,230.7	2,384.8	1,539.3	2,210.2	2,140.8	-
Scaffold N50 (Kb)	532.9	204,922.3	189,477.5	137,335.1	208,696.2	183,350.9	835.7
GC content (%)	46.0	47.4	46.9	46.6	46.8	45.8	46.0
Genome completeness							
Mapping rate (%)	97.5	93.6	98.1	97.9	92.7	97.7	98.7
Coverage rate (%)	98.1	96.7	98.8	99.2	98.1	98.3	96.6
BUSCO (%)	93.2	91.8	91.6	93.3	92.3	90.3	95.0
Annotation statistics							
Repeat content (%)	64.4	59.3	60.2	52.6	61.6	63.1	63.5
Gene numbers	28,641	27,144	26,696	29,644	25,718	29,857	24,501
Mean gene length (bp)	15,067	6,466	12,166	10,574	11,201	8,818	26,459
Mean CDS length (bp)	1,091	892	1,097	960	1,016	1,009	1,348

a. Scaffold N50, GC content, and annotation statistics were calculated using the latest available genome of *I. scapularis* ISE6 cell line (Miller et al., 2018).

SUPPLEMENTAL INFORMATION

Figure S1. Illustrations of the Six *de novo* Sequenced Tick Species, Related to Figure 1.

Figure S2. Life Cycles (A-E) and Parameters (F) under Laboratory Rearing Conditions of Six Tick Species, Related to Figure 1.

Figure S3. Phylogenetic Structures for Tick Populations Based on Mitochondrial (A) and Nuclear (B) genome, Related to Figure 3.

Figure S4. Geographical Population Structures of (A) *Hae. longicornis* and (B) *R. microplus*, Related to Figure 3.

Figure S5. Epidemiological Distribution of the Human Cases Infected with Tick-borne Diseases and of Pathogen Profiles in Six Tick Species, Related to Figure 4.

Table S1. Details of the Assembly and Annotation Information of the Sequenced Tick Genomes and Their Homolog Species, Related to Table 1.

Table S2. Pfam Comparison across Seven Tick Species, Three Other Blood-feeding Arthropods and Two Arachnids, Related to Figure 2A.

Table S3. Comparative Analysis of the Genes Associated with Tick Hematophagy and the Related Phenotype, Related to Figure 2.

Table S4. Epidemiological Distribution of the Human Cases Infected with Tick-borne Diseases from 1980 to 2020, Related to Figure 4D.

Table S5. Detailed Commands, Parameters and Configurations Used in Tick Genome Assembly, Repeat Identification and Gene Annotation, Related to STAR★METHODS.

STAR★METHODS

- **KEY RESOURCE TABLE**
- **RESOURCE AVAILABILITY**
  - **Lead contact**
  - **Materials Availability**
  - **Data and code availability**
- **METHOD DETAILS**
  - **Sample collection**
  - ***De novo* sequencing, assembly and annotation**
  - **Gene family and phylogenetic analysis**
  - **Divergence time estimation**
  - **Gene family analysis and comparison**
  - **Comparative genomics**
  - **Differential transcriptome analysis**
  - **Population structure analysis**
  - **Metagenomic analysis and pathogen detection**
  - **Epidemiological data search strategy**
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

STAR★METHODS

KEY RESOURCE TABLE

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Wu-Chun Cao (caowc@bmi.ac.cn).

Materials Availability

The study did not generate any new reagents.

Data and code availability

The genome assemblies and annotations generated in this study are available at BIGD (https://bigd.big.ac.cn, project accession ID PRJCA002240). We have also submitted the genome assemblies to GenBank (accession ID: JABSTQ000000000-JABSTV000000000) with the project accession ID: PRJNA633311. The raw data of re-sequenced samples are available at BIGD (accession number PRJCA002242). We provided a detailed list of software, commands, parameters and configuration files used in genome data analyses in Table S5.

METHODS DETAILS

Sample collection

From November 2017 to January 2019, ticks were collected from 28 provinces, metropolises or autonomous regions of mainland China. The collection sites were selected according to their ecological environments, including coniferous forest, steppe, farmland, desert, shrubland and tropical forest. Ticks were collected by dragging a standard 1-m^2 flannel flag over vegetation or from domestic or wild animals such as cattle, dogs, sheep, goats, cats, rabbits, camels, deer, and boars. The latitude and longitude of each collection site were recorded. The species, sex and developmental stage of each tick were identified by entomologists. Adult ticks were used for tick genome resequencing to understand their genetic diversity, population

structure and pathogen distribution. Most of the *R. sanguineus* and *R. microplus* ticks were collected from animal hosts. A majority of the *I. persulcatus*, *Hae. longicornis*, *D. silvarum*, and *Hy. asiaticum* specimens were free questing ticks. Live ticks were transported to the laboratory, and dead ticks were directly stored at -80 °C. A total of 678 specimens were used for tick genome resequencing (Figure 1A).

Live adult ticks of *I. persulcatus*, *Hae. longicornis*, *D. silvarum*, *Hy. Asiaticum*, *R. sanguineus*, and *R. microplus* collected from the Heilongjiang (129.22°E, 44.96°N), Shandong (122.32°E, 36.89°N), Shanxi (110.93°E, 38.70°N), Tibet (91.09°E, 30.68°N), Guangxi (109.96°E, 22.41°N) and Guizhou (107.96°E, 26.56°N) provinces, respectively (Figure 1A), were laboratory reared to obtain larvae and then used for *de novo* genome sequencing. Laboratory mice (for *I. persulcatus*), rabbits (for *Hae. longicornis* and *D. silvarum*) and goats (for *Hy. asiaticum*) were used for blood feeding to obtain engorged females. Engorged *R. sanguineus* and *R. microplus* ticks were directly collected from dogs or cattle on site. Engorged female ticks were reared separately under a 12-hour light/12-hour dark photoperiod at 25 °C in desiccators in which a saturated aqueous solution of K_2SO_4 was used to maintain relative humidity. Larvae hatched from a single female were used for the subsequent *de novo* genome sequencing (Illumina, PacBio sequencing and Hi-C experiment), considering their lower contamination of environmental bacteria than those directly collected from natural environments, and their single maternal source which may reduce genetic complexities. In addition, to reduce the genetic heterozygosity of *R. microplus*, the embryo-derived cell line BME/CTVM23 (Alberdi et al., 2012) of *R. microplus* was also subjected to deep sequencing and then used for genome scaffolding.

***De novo* sequencing, assembly and annotation**

Genomic DNA preparation and genome sequencing

Larvae hatched from a single female were used for *de novo* sequencing. Approximately 50-100 larvae of each tick species were collected, thoroughly surface-sterilized (two successive washes of 70% ethanol, 30 s each) and then used for genomic DNA extraction using the DNeasy Blood & Tissue Kit (Qiagen, USA). The integrity of the DNA was determined using an Agilent 4200 Bioanalyzer (Agilent Technologies, Palo Alto, California, Genomic DNA Analysis ScreenTape and

Genomic DNA Reagents). Two high-throughput sequencing platforms, namely, the Illumina HiSeqX-Ten and Pacific Bioscience Sequel, were used to generate sequencing data. First, more than 1 μg of DNA was used to construct short fragmented libraries with an insertion size of 350 bp, which were then sequenced on the Illumina HiSeqX-Ten platform. For each tick species, approximately ~110Gb Illumina sequencing data were generated. Second, 8 mg of DNA was sheared using g-Tubes (Covaris, Woburn, MA) and concentrated with AMPure PB magnetic beads. Each single-molecule real-time (SMRT) bell library was constructed using the Pacific Biosciences SMRTbell Template Prep Kit 1.0. The constructed libraries were size-selected on a BluePippin™ system for molecules ≥15 kb, followed by primer annealing (Sequencing Primer v3) and the binding of SMRTbell templates to polymerases with the Sequel Binding and Internal Control Kit 3.0. Sequencing (Sequel Sequencing Kit 3.0 Bundle, SMRT Cell 1M v3 Tray) was performed on the Pacific Bioscience Sequel platform by Annoroad Gene Technology Beijing Co. Ltd.

To further improve the continuity of the assembled genomes, approximately 100 ~ 200 larvae of five tick species were used for chromosome conformation capture (Hi-C) experiments (*I. persulcatus* was not included due to its limited sample size). Cells/tissues were crosslinked using 40 ml of 2% formaldehyde solution at room temperature for 15 min. A total of 4.324 ml of 2.5 M glycine was added to quench the crosslinking reaction. The supernatant was removed, and the tissues were ground with liquid nitrogen and resuspended in 25 ml of extraction buffer I containing 0.4 M sucrose, 10 mM Tris-HCl (pH 8.0), 10 mM $MgCl_2$, 5 mM β-mercaptoethanol, 0.1 mM phenylmethylsulfonyl fluoride (PMSF), and 13 protease inhibitors (Sigma) and then filtered through Miracloth (Calbiochem). The filtrate was centrifuged at 4000 rpm and 4 °C for 20 min. The pellet was resuspended in 1 ml of extraction buffer II (0.25 M sucrose, 10 mM Tris-HCl (pH 8), 10 mM $MgCl_2$, 1% Triton X-100, 5 mM β-mercaptoethanol, 0.1 mM PMSF, and 13 protease inhibitors) and centrifuged at 14,000 rpm and 4 °C for 10 min. The pellet was resuspended in 300 ml of extraction buffer III (1.7 M sucrose, 10 mM Tris-HCl (pH 8), 0.15% Triton X-100, 2 mM $MgCl_2$, 5 mM β-mercaptoethanol, 0.1 mM PMSF, and 1 μl of protease inhibitor), loaded on top of an equal amount of clean extraction buffer III and then centrifuged at 14,000 rpm for 10 min. The supernatant was discarded, and the pellet was washed twice by resuspending in 500 μl of ice-cold 1× CutSmart buffer and then centrifuged

for 5 min at 2,500 ×g. The nuclei were washed with 0.5 ml of restriction enzyme buffer and transferred to a safe-lock tube. Next, the chromatin was solubilized with dilute SDS and incubated at 65 °C for 10 min. After quenching the SDS with Triton X-100, overnight digestion was performed with a 4-cutter restriction enzyme (400 units of MboI) at 37 °C on a rocking platform. The next step was Hi-C specific, including marking of the DNA ends with biotin-14-dCTP and performing blunt-end ligation of crosslinked fragments. The proximal chromatin DNA was religated using the ligation enzyme. The nuclear complexes were reverse-crosslinked by incubating with proteinase K at 65 °C. DNA was purified by phenol-chloroform extraction, and biotin-C was removed from nonligated fragment ends using T4 DNA polymerase. Fragments were sheared to 100-500 bp by sonication. The fragment ends were repaired using a mixture of T4 DNA polymerase, T4 polynucleotide kinase and Klenow DNA polymerase. Biotin-labeled Hi-C samples were specifically enriched using streptavidin magnetic beads. A-tails were added to the fragment ends by Klenow (exo-), and then the Illumina paired-end sequencing adapter was added via a ligation mix. Finally, the Hi-C libraries were amplified by 10-12 cycles of PCR and sequenced on an Illumina HiSeqX-Ten (HiSeq X Ten Reagent Kit v2.5).

Genome size estimation

Before *de novo* assembly, we estimated the genome size of each tick species. For each tick species, we built an Illumina short-read library using the DNA material from the same source as the PacBio sequencing library, and ~110 Gb Illumina sequencing data were generated. Based on the Illumina data, Jellyfish (v2.1.3) (Marçais and Kingsford, 2011) was employed to calculate the frequency of each K-mer (k=21). Then, the genome size was estimated using a previously described method based on K-mer distribution (Liu et al., 2013).

Genome assembly and quality assessment

PacBio reads were first assembled using four *de novo* assemblers: Canu (Koren et al., 2017), Falcon (Chin et al., 2016), SMARTdenovo (Istace et al., 2017) and wtdbg (Ruan and Li, 2020). The best assembly was selected according to the optimal continuity and completeness, and the final version of the genome assembly was polished by Arrow and error-corrected by Pilon (Walker et al., 2014) using Illumina reads. The completeness of the final assembly was evaluated using two criteria: (1) BUSCO (v3.0, arthropoda_odb9) (Simão et al., 2015) based on the evolutionarily

informed expectations of gene content from near-universal single-copy orthologs; (2) mapping rate and coverage of Illumina reads on the assembled genomes.

Scaffolding was performed using Hi-C-based proximity-guided assembly for five tick species, excluding *I. persulcatus*. Hi-C reads were first aligned to the draft genome using the bowtie2.2.3algorithm (Langmead and Salzberg, 2012). According to the Hi-C protocol and the fill-in strategy, unmapped reads were mainly composed of chimeric fragments spanning the ligation junction. HiC-Pro (V2.7.8) was used to identify ligation sites and align back to the genome using the 5' fraction of the read (Servant et al., 2015). The assembly package Lachesis (Burton et al., 2013) was used to perform clustering, ordering and orienting. Based on the agglomerative hierarchical clustering algorithm, we clustered the scaffolds into 11 chromosome groups based on the karyotypes of chromosomes from a previous report (Qin et al., 1997). Contigs from the polished and corrected assembly were anchored to chromosome groups with a length ratio of 80% ~ 95%.

Additional assembly procedures for Hae. longicornis

The initial genome size of *Hae. longicornis* was estimated to be 5.4 G based on the Illumina sequencing data of 100 larvae, which was much larger than those of the other five tick species. Considering its nontypical K-mer Poisson distribution, we assume that the elevated genome size could be attributed to the heterozygosity of the larvae used for *de novo* sequencing. Therefore, we resequenced additional *Hae. longicornis* specimens from three provinces (Beijing, Shandong and Zhejiang), with one male and one female from each province. The genome sizes of three males and two females were approximately 2.4-2.8 Gb. Interestingly, the genome size of the female from Shandong was approximately 3.6 Gb. The larger genome size of this female may be related to the additional chromosomes in the parthenogenetic lineage, which was supported by the detected genetic markers of the parthenogenetic lineage in the female sample (Chen et al., 2014).

The overestimated genome size of *Hae. longicornis* indicated its high genome heterozygosity in the PacBio library. Therefore, additional assembly procedures were adopted beyond the conventional pipeline to improve the assembly quality. First, before assembly, we used the Illumina reads of a single female sample to correct the PacBio reads using LorDEC version 0.8 (Salmela and Rivals, 2014). Second, we filtered a subset of the PacBio reads that showed a low LorDEC correction ratio

(<25%, i.e., proportion of PacBio reads covered by Illumina reads). After filtering, the corrected PacBio reads were fed into the assembler. Third, we obtained a core genome by removing the genome sequences from 7 redundant homologous chromosomes of the core female genome by using redundans (Pryszcz and Gabaldón, 2016) (with parameters including an identity of 80% and overlap of 50%). Finally, contigs of the core male *Hae. longicornis* assembly were anchored in 11 chromosomes using the Hi-C data.

Repeat annotation

Repetitive sequences and transposable elements (TEs) in each tick genome were identified using a combination of *de novo* and homology-based approaches. Briefly, RepeatMasker (open-4.0.6) (Chen, 2004) and RepeatProteinMask (v.4.0.6) were used to identify and classify different TEs by aligning genome sequences against Repbase version 23.12 (Jurka et al., 2005) with default parameters. To identify tandem repeats, TRF v4.0.6 (Benson, 1999) was used with the following parameters: Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50, MaxPerid = 500, -d, -h.

Genome annotation

Gene annotation was accomplished by integrating evidence or predictions from transcriptome-, homology- and *ab initio*-based approaches. In the transcriptome-based approach, RNA was extracted from six tick species. In brief, ticks were quickly washed in RNase-free water twice and homogenized in RLT solution under liquid nitrogen. The homogenate was then incubated at 55 °C for 10 min with proteinase K (Qiagen, USA) and centrifuged for 30 s at full speed. The homogenized lysate was used for further RNA extraction using the RNeasy Mini Kit (Qiagen, USA). RNA quality was assessed using an Agilent Bioanalyzer 2200 (Agilent Technologies, Inc.). RNA-seq libraries were generated by using RiBo-Zero Gold rRNA Removal Reagents (Human/Mouse/Rat) (Illumina). Paired-end (150 bp) sequencing of the RNA library was performed on an Illumina HiSeq 4000 platform. RNA-seq reads generated from each tick species were assembled by Trinity (v2.4.0, https://github.com/trinityrnaseq/trinityrnaseq) with default parameters (Haas et al., 2013). The assembled transcripts were aligned to each assembled genome and were used to predict gene structure by PASA (v2.3.3 http://wfleabase.org/release1/PASA_gene_annotation.html) (Haas et al., 2008). The protein sequences of homologous species, including *I. scapularis*

(https://www.vectorbase.org/), *C. sculpturatus* (https://i5k.nal.usda.gov/content/data-downloads) and *P. tepidariorum* (https://i5k.nal.usda.gov/content/data-downloads), were retrieved from public databases. In addition, as the six tick species sequenced are closely related species, the genes of all five species annotated only by PASA were also added to the homologous gene dataset. Homologous protein sequences were aligned to the tick genome assemblies using TBLASTN v2.2.28+ (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.28/) with e-value=1e-5 (Camacho et al., 2009), and the gene structure was predicted by GeneWise v 2.2.0 (Birney et al., 2004). *Ab initio* gene prediction was performed using Augustus v3.3 (Stanke et al., 2004), GlimmerHMM v 3.0.4 (Majoros et al., 2004), SNAP (Korf, 2004), and GeneMark v3.51 (Besemer and Borodovsky, 2005). Based on the above evidence, we used EvidenceModeler (EVM) v1.1.1(Haas et al., 2008) to integrate the gene models predicted by the above approaches into a nonredundant and more complete gene set. Finally, the functions of the protein-coding genes were predicted by searching against multiple gene annotation databases, including SwissProt (http://www.ebi.ac.uk/interpro/search/sequence-search), NT (https://www.ncbi.nlm.nih.gov/nucleotide/), NR (https://www.ncbi.nlm.nih.gov/protein/), Pfam (http://xfam.org/), Eggnog (http://eggnogdb.embl.de/), GO (http://geneontology.org/page/go-database), and KEGG (http://www.genome.jp/kegg/).

Noncoding RNA annotation

Four types of noncoding RNAs (ncRNAs), namely, microRNAs (miRNAs), transfer RNAs (tRNAs), ribosomal RNAs (sRNAs) and small nuclear RNAs (snRNAs), were identified. The tRNA genes were identified using tRNAscan-SE v1.3.1 (Lowe and Eddy, 1997) with default parameters. The rRNA fragments were predicted by aligning human rRNA sequences to the assembled genome sequences by BLASTN with the parameter e-value <1e-5. The miRNA and snRNA genes were searched using BLAST against the Rfam v13.0 database using INFERNAL v1.0 with a family-specific "gathering" cutoff of Rfam (Griffiths-Jones et al., 2005).

Collinearity analysis

Collinear segments were detected between assembled genomes using JCVI software (v0.8.4, https://github.com/tanghaibao/jcvi) (Tang et al., 2015) with default parameters.

Gene family and phylogenetic analysis

To infer tick evolutionary history, a maximum likelihood phylogenetic tree was built based on the protein sequences of nine species, including the six tick species sequenced in this study, *I. scapularis* and two outgroup species (*C. sculpturatus* and *P. tepidariorum*) (Thomas et al., 2018) . First, single-copy genes within the nine species were identified, and all-to-all BLAST was performed for all protein sequences (E-value <10-10 and identity >30%). Gene families (i.e., ortholog or paralog groups) were identified using OrthoMCL (Li et al., 2003) with the parameters -I=1.5. Single-copy gene families (n=464) were used for subsequent phylogenetic analysis. The protein sequences of these single-copy genes were aligned using MUSCLE (v3.6) (Edgar, 2004) and then used to construct a maximum likelihood tree by PhyML (v3.0) (Guindon et al., 2010).

Divergence time estimation

The divergence time within the nodes of the phylogenetic tree was estimated by the MCMCTREE program of PAML (v4.4) (Yang, 2007) with parameters RootAge=500, model=4, alpha=0, clock=3, sample frequency=2, burn-in=20000, nsample=100000, and finetune="0.00876 0.03724 0.06828 0.00789 0.44485". The divergence time was corrected using calibration points from the TimeTree website (http://timetree.org/) (Kumar et al., 2017).

Gene family analysis and comparison

The expansion and contraction of gene families were determined by comparing the cluster size differences between the ancestor and each of the six investigated tick species and *I. scapularis* using the CAFE program (http://sourceforge.net/projects/cafehahnlab/) (De Bie et al., 2006). CAFE used a random birth-and-death model to infer gene family size across the tree. To calculate the probability of the transitions in each gene family size from parent to child nodes in the tree, a probabilistic graphical model was introduced. According to the conditional

likelihoods, we calculated the possible p-value in each lineage. A p-value of 0.05 was used to identify significantly expanded/contracted families.

Comparative genomics

Pfam analysis

We searched the potential Pfam domains from 12 species of three groups, including six ticks sequenced in this study, *I. scapularis* (Miller et al., 2018), other blood-feeding arthropods *A. aegypti* (Matthews et al., 2018), *A. gambiae* (Holt et al., 2002), and *Glossina* (International Glossina Genome Initiative, 2014), and a non-blood-feeding outgroup *C. sculpturatus* and *P. tepidariorum* (Thomas et al., 2018). Briefly, amino acid sequences of each species were scanned using all profiles from Pfam database version 31 (El-Gebali et al., 2019) by hmmscan version hmmer-3.1b1. The scanned results were filtered with an e-value cutoff of 1e-3, and overlapping/redundant hmm matches were removed. Genes assigned to Pfam were counted within each species. To identify Pfams that differed between the three groups, we used a fold change >2 of the group median value as the selection criteria. Two sets of Pfams were identified using the two-fold criteria: (1) Pfams that were abundant in ticks compared with other blood-feeding arthropods and the outgroup; (2) The Pfams showed similar abundances (fold change ≤2) among ticks and other blood-feeding arthropods but were more abundant in these organisms than in the outgroup.

Orthology analysis

We performed orthology analysis for our six genomes and *I. scapularis* (Miller et al., 2018) genome. First, the protein sequences of gene families with various functions, including iron metabolism, carbohydrate metabolism, amino acid metabolism, chemosensory functions, gustatory functions, immune functions, heme and hemoglobin digestion, detoxification of xenobiotic factors, opsin-related functions, lipid metabolism, oxidative stress, purine metabolism, and mechanosensation, were retrieved and divided into subgroups according to their specific functions (Anderson et al., 2008; Antunes et al., 2012; Bohbot et al., 2014; Cabezas-Cruz et al., 2017; Della Noce et al., 2019; Eyun et al., 2017; Galay et al., 2013; Graça-Souza et al., 2002; Graça-Souza et al., 2006; Gulia-Nuss et al., 2016; Hajdušek et al., 2013; Hajdusek et al., 2016; Hajdusek et al., 2009; Horn et al., 2009; International Glossina Genome Initiative, 2014; Iovinella et al., 2016; Josek et al., 2018; Liu et al., 2012; Liu et al., 2011; Merino et al., 2011; Pal et al., 2004; Perner et al., 2016; Salem et al.,

2014; Sanders et al., 2003; Sonenshine and Macaluso, 2017; Sultana et al., 2010; Weisheit et al., 2015; Whiten et al., 2017; Winzerling and Pham, 2006). The gene families in each subgroup are shown in Table S3. Second, using collected sequences in each subgroup as query sequences, a second step of BLASTp were performed to search ortholog protein sequences in our assembled genomes (e-value <1e-5, identity ≥50%, match percentage of shorter sequence between query and subject ≥25%). For each subgroup, multiple-sequence alignments were performed using MUSCLE v3.8.31 with the default parameters, and PhyML v3.3.20190321 was employed to construct a phylogenetic tree. Based on the phylogenetic tree, genes with high reliability and the gene number for each subgroup were determined after filtering distantly related genes.

Differential transcriptome analysis

Unfed and fed ticks of *I. persulcatus* (3 vs 6 ticks were pooled as unfed vs fed group, respectively), *Hae. longicornis* (10 vs 10) and *R. microplus* (14 vs 24) were used for RNA extraction and transcriptome sequencing. The high-quality transcriptomic data were aligned to the reference genome using HISAT2 v2.1.0 (Kim et al., 2019). The read count of each gene was calculated for each sample by HTSeq v0.6.0 (Anders et al., 2015), and fragments per kilobase per million mapped reads (FPKM) values were then determined. DE genes were analyzed using EdgeR(v3.28.1) (Robinson et al., 2010) with false discovery rate (FDR) ≤0.05 and |log2(fold change) |≥1. The dispersion parameter of DE model was estimated using the estimateCommonDisp() function in the EdgeR package. Enriched GO terms (http://geneontology.org/) of the DE genes were identified using Fisher's exact test in the topGO package (Alexa and Rahnenfuhrer, 2007) (FDR <0.05). Enriched pathways were tested based on the KEGG database (Kyoto Encyclopedia of Genes and Genomes, http://www.kegg.jp/) using clusterProfiler (Yu et al., 2012) (FDR <0.05).

Population structure analysis

Genomic DNA extraction and library preparation for resequencing

All 678 adult ticks collected from the wild were thoroughly surface-sterilized, and genomic DNA for resequencing was isolated using the AllPrep DNA/RNA Mini Kit

(Qiagen, USA). The DNA concentration was measured using the Qubit dsDNA HS Assay Kit in a Qubit® 2.0 fluorometer (Life Technologies, CA, USA). Sequencing libraries were constructed using the NEBNext® UltraTM DNA Library Prep Kit for Illumina (NEB, USA) following the manufacturer's recommendations, and index barcodes were added to attribute sequences to each sample. The library preparations were sequenced on an Illumina NovaSeq platform (NovaSeq 6000 SP Reagent Kit), and paired-end reads were generated.

Variant calling and population structure models

Illumina reads of 678 tick samples were aligned to the corresponding reference genome using BWA (version 0.7.17-r1188) (Li and Durbin, 2009). Variants were called following the recommended GATK 4.0 pipeline (Van der Auwera et al., 2013). Variant sites with quality scores ≥ 30 were kept for subsequent analysis. Based on the called variants, we generated the full mitochondrial sequence of each specimen and built maximum likelihood trees by MEGA7 (Kumar et al., 2016) using the GTR+F+I substitution model. The tree was rooted using mitochondrial sequence of *Ornithodoros hermsi* (NC_039832.1) as outgroup. For variant calling on the nuclear genome, we selected variants with sufficient reads (8≤ read depth ≤12, genotype rate>70%), as the mean genome read coverage was ~8×. To build the phylogenetic tree of the nuclear genome, we used SNPs (minor allele frequency ≥ 5%) in 464 single-copy genes that are supposed to be conserved across tick species. An external dataset from New Zealand (SRR9226159) (Guerrero et al., 2019) was added to the phylogenetic analysis of *Hae. longicornis* and processed using the same pipeline as that used for the six tick genomes sequenced in this study.

Geographical population structure was analyzed using fastSTRUCTURE (Raj et al., 2014) using SNPs in the mitochondria. For each tick species, fastSTRUCTURE was run for K (number of ancestral populations) from 2 to 10 with fivefold cross-validation. The fastSTRUCTURE model selected the best value of K=2 for *Hae. longicornis* and K=3 for *R. microplus* by maximizing the marginal likelihood of the fastSTRUCTURE model. However, to enable fair comparison between the two species, we chose a more detailed population structure (K=5), as shown in Fig. 3B. The population structure was plotted using Pophelper (2.3.0) package (Francis, 2017) and CLUMPAK (http://clumpak.tau.ac.il) (Kopelman et al., 2015). To measure population differentiation, we calculated the F_{ST} between all pairs of populations in

each province for *Hae. longicornis* and *R. microplus* based on the SNPs within their mitochondria. First, the numerator and denominator of the Hudson F_{ST} estimator were calculated for each SNP. Then, across all SNPs, the ratio of the average numerator and denominator was calculated as the final F_{ST} estimator between two populations (Bhatia et al., 2013). F_{ST} calculations were conducted using the python scikit-allel package (version 1.2.1, https://github.com/cggh/scikit-allel) (Alistair Miles and Harding, 2016).

Copy number variation detection in the genomes of R. microplus

We found that *R. microplus* can be clustered into three major clades. First, genes with read counts > 2 in at least half of the samples were selected to calculate the copy number changes in the three clades of *R. microplus*. Second, the read counts of the genes were normalized to gene length. In each sample, the normalized gene read count was divided by the median of all genes to calculate the fold change (cf) of the copy number (CN). Third, the cfCN of the gene was compared with each sample median cfCN by the function of t.test (paired=T) in R to calculate the significance in each clade. The p-values were adjusted for multiple testing correction using Benjamini-Hochberg correction as a function of p.adjust (method = "BH") in R. In each clade, genes with adjusted p-values < 0.001 and median cfCN ≥ 2 were referred to as increased CN genes. According to the gene annotation results, GO enrichment analysis was limited to the 4-level GO terms and implemented by a hypergeometric test with the phyper() function in R. The enrichment p-value was adjusted by the p.adjust function (method = "BH") in R.

Metagenomic analysis and pathogen detection

Tick sequences were filtered by SAMtools (version 0.9.24) (Li et al., 2009) after mapping the reads of 678 specimens to tick genomes by BWA (version: 0.7.17), and all unmapped reads were retained for subsequent analysis. Taxonomic classification was performed by aligning the filtered reads to the NR database using DIAMOND (version 0.9.24, parameters: -f 102 -top 10) (Buchfink et al., 2015). To estimate the relative abundances of different bacterial species, we extracted all taxonomic IDs according to the NCBI taxdump (ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdmp.zip) (*Rickettsia*: TaxID780, *Anaplasma*: TaxID768, *Ehrlichia*: TaxID 943, *Borrelia*: TaxID 138, *Babesia*: TaxID

5864, *Theileria*: TaxID 5873, *Francisella*: TaxID262, *Bartonella*: TaxID 773, *Coxiella*: TaxID 776, *Hepatozoon*: TaxID 75741, *Toxoplasma*: TaxID 5810, *Candidatus Neoehrlichia*: TaxID 467749). Sequence similarity (>70%) were used as the threshold to screen the alignment results. The classification of species pathogenic to human or not to was based on currently available literatures. After normalizing all classified sequences to 100,000 microbial reads, the relative abundance of each pathogen was estimated by calculating the sequences classified to this species. We also adopted a widely-used tool, Metaphlan2 (Segata et al., 2012), for metagenomic taxonomic profiling, but only a very limited number of pathogens could be found in different tick species. Therefore, we used the results of NR-blast-based method for downstream analyses.

Epidemiological data search strategy

We searched PubMed and ISI (Web of Science) for articles published in English, and WanFang database, China National knowledge Infrastructure, and Chinese Scientific Journal Database of articles published in Chinese between Jan 1, 1980 and April 30, 2020. We used the following search terms: “tick-borne disease”, “tick-borne zoonosis”, “tick-borne zoonotic disease”, “tick-associated agent”, “tick-associated microbe”, and “China”. The articles about tick-borne viral diseases were excluded. We did a secondary manual search of the references cited in these articles to find relevant articles. We investigated all the articles related to detection, identification, or case reports of tick-borne microbes in human beings. Each case was geo-referenced to a Chinese map in the prefecture-level with ArcGIS 10.2 (Johnston et al., 2004) (ESRI, Redlands, CA, USA) according to the patient’s living location or visiting hospital.

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification and analysis procedures of genome, transcriptome and metagenome data were provided in the relevant sections of Method Details. To test the correlation between gene family size and standard deviation of gene expression (fold change), Spearman’s rank correlation coefficient was calculated. Fisher’s exact test was used to test the enrichment of down-regulated genes in P450 families. Mann–Whitney U test was used to compare the prevalence of pathogen in different faunal or geographical regions. Kruskal-Wallis test was used to compare the positive rate of *Rickettsia* in different subtypes of ticks. All these tests were performed in R environment and p

value below 0.05 was considered statistically significant. For all analyses, the meaning and value of n and the measures of center, dispersion, and precision used can be found in the relevant main text or in Method Details.

References

Alberdi, M.P., Nijhof, A.M., Jongejan, F., and Bell-Sakyi, L. (2012). Tick cell culture isolation and growth of *Rickettsia raoultii* from Dutch Dermacentor reticulatus ticks. Ticks Tick Borne Dis. *3*, 349-354.

Alexa, A., and Rahnenfuhrer, J. (2007). Gene set enrichment analysis with topGO. https://bioc.ism.ac.jp/packages/2.0/bioc/vignettes/topGO/inst/doc/topGO.pdf

Alistair Miles, and Harding, N. (2016). scikit-allel: A Python package for exploring and analysing genetic variation data. http://github.com/cggh/scikit-allel.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Anderson, J.F., and Magnarelli, L.A. (2008). Biology of ticks. Infect. Dis. Clin. North Am. *22*, 195-215, v.

Anderson, J.M., Sonenshine, D.E., and Valenzuela, J.G. (2008). Exploring the mialome of ticks: an annotated catalogue of midgut transcripts from the hard tick, *Dermacentor variabilis* (Acari: Ixodidae). BMC Genomics *9*, 552.

Antunes, S., Galindo, R.C., Almazán, C., Rudenko, N., Golovchenko, M., Grubhoffer, L., Shkap, V., do Rosário, V., de la Fuente, J., and Domingos, A. (2012). Functional genomics studies of *Rhipicephalus (Boophilus)* annulatus ticks in response to infection with the cattle protozoan parasite, *Babesia bigemina*. Int. J. Parasitol. *42*, 187-195.

Araya-Anchetta, A., Busch, J.D., Scoles, G.A., and Wagner, D.M. (2015). Thirty years of tick population genetics: a comprehensive review. Infect. Genet. Evol. *29*, 164-179.

Barker, S.C., and Walker, A.R. (2014). Ticks of Australia. The species that infest domestic animals and humans. Zootaxa, 1-144.

Beard, C.B., Occi, J., Bonilla, D.L., Egizi, A.M., Fonseca, D.M., Mertins, J.W., Backenson, B.P., Bajwa, W.I., Barbarin, A.M., Bertone, M.A., *et al.* (2018). Multistate infestation with the exotic disease-vector tick *Haemaphysalis longicornis* - United States, August 2017-September 2018. MMWR Morb. Mortal. Wkly. Rep. *67*, 1310-1313.

Beati, L., and Klompen, H. (2019). Phylogeography of Ticks (Acari: Ixodida). Annu. Rev. Entomol. *64*, 379-397.

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

519 Besemer, J., and Borodovsky, M. (2005). GeneMark: web software for gene finding in
520 prokaryotes, eukaryotes and viruses. Nucleic Acids Res. *33*, W451-454.
521 Bhatia, G., Patterson, N., Sankararaman, S., and Price, A.L. (2013). Estimating and
522 interpreting F_{ST}: the impact of rare variants. Genome Res. *23*, 1514-1521.
523 Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome
524 Res. *14*, 988-995.
525 Bohbot, J.D., Sparks, J.T., and Dickens, J.C. (2014). The maxillary palp of *Aedes*
526 *aegypti*, a model of multisensory integration. Insect Biochem. Mol. Biol. *48*, 29-39.
527 Braz, G.R., Coelho, H.S., Masuda, H., and Oliveira, P.L. (1999). A missing metabolic
528 pathway in the cattle tick *Boophilus microplus*. Curr. Biol. *9*, 703-706.
529 Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment
530 using DIAMOND. Nat. Methods *12*, 59-60.
531 Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O., and Shendure, J.
532 (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on
533 chromatin interactions. Nat. Biotechnol. *31*, 1119-1125.
534 Cabezas-Cruz, A., Alberdi, P., Valdés, J.J., Villar, M., and de la Fuente, J. (2017).
535 *Anaplasma phagocytophilum* infection subverts carbohydrate metabolic pathways in
536 the tick vector, *Ixodes scapularis*. Front. Cell. Infect. Microbiol. *7*, 23.
537 Camacho, C., Coulouris, G., Avagyan, V,. Ma N., Papadopoulos, J., Bealer, K., and
538 Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics,
539 *10*: 421.
540 Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic
541 sequences. Curr. Protoc. Bioinformatics *Chapter 4*, Unit 4.10.
542 Chen, X., Xu, S., Yu, Z., Guo, L., Yang, S., Liu, L., Yang, X., and Liu, J. (2014).
543 Multiple lines of evidence on the genetic relatedness of the parthenogenetic and
544 bisexual *Haemaphysalis longicornis* (Acari: Ixodidae). Infect. Genet. Evol. *21*,
545 308-314.
546 Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A.,
547 Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., *et al.* (2016).
548 Phased diploid genome assembly with single-molecule real-time sequencing. Nat.
549 Methods *13*, 1050-1054.
550 Cirimotich, C.M., Dong, Y., Clayton, A.M., Sandiford, S.L., Souza-Neto, J.A.,
551 Mulenga, M., and Dimopoulos, G. (2011). Natural microbe-mediated refractoriness to

Plasmodium infection in *Anopheles gambiae*. Science *332*, 855-858.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics *22*, 1269-1271.

Della Noce, B., Carvalho Uhl, M.V., Machado, J., Waltero, C.F., de Abreu, L.A., da Silva, R.M., da Fonseca, R.N., de Barros, C.M., Sabadin, G., Konnai, S., *et al.* (2019). Carbohydrate metabolic compensation coupled to high tolerance to oxidative stress in ticks. Sci. Rep. *9*, 4753.

Dickinson, D.A., and Forman, H.J. (2002). Glutathione in defense and signaling: lessons from a small thiol. Ann. N. Y. Acad. Sci. *973*, 488-504.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792-1797.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., *et al.* (2019). The Pfam protein families database in 2019. Nucleic Acids Res. *47*, D427-D432.

Eyun, S.I., Soh, H.Y., Posavi, M., Munro, J.B., Hughes, D.S.T., Murali, S.C., Qu, J., Dugan, S., Lee, S.L., Chao, H., *et al.* (2017). Evolutionary history of chemosensory-related gene families across the Arthropoda. Mol. Biol. Evol. *34*, 1838-1862.

Fang, L.Q., Liu, K., Li, X.L., Liang, S., Yang, Y., Yao, H.W., Sun, R.X., Sun, Y., Chen, W.J., Zuo, S.Q., *et al.* (2015). Emerging tick-borne infections in mainland China: an increasing public health threat. Lancet Infect. Dis. *15*, 1467-1479.

Francis, R.M. (2017). POPHELPER: An R package and web app to analyse and visualise population structure. Mol. Ecol. Resour. *17*, 27-32.

Galay, R.L., Aung, K.M., Umemiya-Shirafuji, R., Maeda, H., Matsuo, T., Kawaguchi, H., Miyoshi, N., Suzuki, H., Xuan, X., Mochizuki, M., *et al.* (2013). Multiple ferritins are vital to successful blood feeding and reproduction of the hard tick *Haemaphysalis longicornis*. J. Exp. Biol. *216*, 1905-1915.

Graça-Souza, A.V., Arruda, M.A., de Freitas, M.S., Barja-Fidalgo, C., and Oliveira, P.L. (2002). Neutrophil activation by heme: implications for inflammatory processes. Blood *99*, 4160-4165.

Graça-Souza, A.V., Maya-Monteiro, C., Paiva-Silva, G.O., Braz, G.R., Paes, M.C., Sorgine, M.H., Oliveira, M.F., and Oliveira, P.L. (2006). Adaptations against heme

toxicity in blood-feeding arthropods. Insect Biochem. Mol. Biol. *36*, 322-335.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. *33*, D121-124.

Guerrero, F.D., Bendele, K.G., Ghaffari, N., Guhlin, J., Gedye, K.R., Lawrence, K.E., Dearden, P.K., Harrop, T.W.R., Heath, A.C.G., Lun, Y., *et al.* (2019). The Pacific Biosciences *de novo* assembled genome dataset from a parthenogenetic New Zealand wild population of the longhorned tick, *Haemaphysalis longicornis* Neumann, 1901. Data Brief *27*, 104602.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307-321.

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., *et al.* (2016). Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nat. Commun. *7*, 10507.

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

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. *9*, R7.

Hajdušek, O., Síma, R., Ayllón, N., Jalovecká, M., Perner, J., de la Fuente, J., and Kopáček, P. (2013). Interaction of the tick immune system with transmitted pathogens. Front. Cell. Infect. Microbiol. *3*, 26.

Hajdusek, O., Sima, R., Perner, J., Loosova, G., Harcubova, A., and Kopacek, P. (2016). Tick iron and heme metabolism - New target for an anti-tick intervention. Ticks Tick Borne Dis. *7*, 565-572.

Hajdusek, O., Sojka, D., Kopacek, P., Buresova, V., Franta, Z., Sauman, I., Winzerling, J., and Grubhoffer, L. (2009). Knockdown of proteins involved in iron metabolism limits tick reproduction and development. Proc. Natl. Acad. Sci. USA *106*,

1033-1038.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., *et al.* (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. Science *298*, 129-149.

Hoogstraal, H., and Aeschlimann, A. (1982). Tick-Host Specificity. Mitt. Schweiz Entomol. Ges. *55*.

Horn, M., Nussbaumerová, M., Sanda, M., Kovárová, Z., Srba, J., Franta, Z., Sojka, D., Bogyo, M., Caffrey, C.R., Kopácek, P., *et al.* (2009). Hemoglobin digestion in blood-feeding ticks: mapping a multipeptidase pathway by functional proteomics. Chem. Biol. *16*, 1053-1063.

International Glossina Genome Initiative (2014). Genome sequence of the tsetse fly (*Glossina morsitans*): vector of African trypanosomiasis. Science *344*, 380-386.

Iovinella, I., Ban, L., Song, L., Pelosi, P., and Dani, F.R. (2016). Proteomic analysis of castor bean tick *Ixodes ricinus*: a focus on chemosensory organs. Insect Biochem. Mol. Biol. *78*, 58-68.

Istace, B., Friedrich, A., d'Agata, L., Faye, S., Payen, E., Beluche, O., Caradec, C., Davidas, S., Cruaud, C., Liti, G., *et al.* (2017). *de novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. Gigascience *6*, 1-13.

Jia, N., Zheng, Y.C., Jiang, J.F., Jiang, R.R., Jiang, B.G., Wei, R., Liu, H.B., Huo, Q.B., Sun, Y., Chu, Y.L., *et al.* (2018). Human Babesiosis Caused by a *Babesia crassa*-Like Pathogen: A Case Series. Clin. Infect. Dis. *67*, 1110-1119.

Jia, N., Zheng, Y.C., Jiang, J.F., Ma, L., and Cao, W.C. (2013). Human infection with *Candidatus* Rickettsia tarasevichiae. N. Engl. J. Med. 369, 1178-1180.

Jia, N., Zheng, Y.C., Ma, L., Huo, Q.B., Ni, X.B., Jiang, B.G., Chu, Y.L., Jiang, R.R., Jiang, J.F., and Cao, W.C. (2014). Human infections with *Rickettsia raoultii*, China. Emerg. Infect. Dis. 20, 866-868.

Jiang, B.G., Jia, N., Jiang, J.F., Zheng, Y.C., Chu, Y.L., Jiang, R.R., Wang, Y.W., Liu, H.B., Wei, R., Zhang, W.H., et al. (2018). *Borrelia miyamotoi* Infections in Humans and Ticks, Northeastern China. Emerg. Infect. Dis. 24, 236-241.

Jiang, J.F., Zheng, Y.C., Jiang, R.R., Li, H., Huo, Q.B., Jiang, B.G., Sun, Y., Jia, N., Wang, Y.W., Ma, L., et al. (2015). Epidemiological, clinical, and laboratory characteristics of 48 cases of "*Babesia venatorum*" infection in China: a descriptive

study. Lancet Infect. Dis. 15, 196-203.
Johnston, K., Ver Hoef, J., Krivoruchko, K., and Lucas, N. (2004). Using ArcGIS geostatistical analyst. In, p. 300.
Jongejan, F., and Uilenberg, G. (2004). The global importance of ticks. Parasitology, S3-14.
Josek, T., Walden, K.K.O., Allan, B.F., Alleyne, M., and Robertson, H.M. (2018). A foreleg transcriptome for *Ixodes scapularis* ticks: Candidates for chemoreceptors and binding proteins that might be expressed in the sensory Haller's organ. Ticks Tick Borne Dis. *9*, 1317-1327.
Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462-467.
Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907-915.
Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., and Mayrose, I. (2015). CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. *15*: 1179-1191.
Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res. *27*, 722-736.
Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59.
Krause, P.J., Gewurz, B.E., Hill, D., Marty, F.M., Vannier, E., Foppa, I.M., Furman, R.R., Neuhaus, E., Skowron, G., Gupta, S., *et al.* (2008). Persistent and relapsing babesiosis in immunocompromised patients. Clin. Infect. Dis. *46*, 370-376.
Kumar, S., Molina-Cruz, A., Gupta, L., Rodrigues, J., and Barillas-Mury, C. (2010). A peroxidase/dual oxidase system modulates midgut epithelial immunity in *Anopheles gambiae*. Science *327*, 1644-1648.
Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. *33*, 1870-1874.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357-359.
Kumar S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: a resource for

timelines, Timetrees and divergence times. Mol. Biol. Evol. *34*, 1812-1819.

Lara, F.A., Pohl, P.C., Gandara, A.C., Ferreira Jda, S., Nascimento-Silva, M.C., Bechara, G.H., Sorgine, M.H., Almeida, I.C., Vaz Ida, S., Jr., and Oliveira, P.L. (2015). ATP binding cassette transporter mediates both heme and pesticide detoxification in tick midgut cells. PloS one *10*, e0134779.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, H., Zheng, Y.C., Ma, L., Jia, N., Jiang, B.G., Jiang, R.R., Huo, Q.B., Wang, Y.W., Liu, H.B., Chu, Y.L., et al. (2015). Human infection with a novel tick-borne *Anaplasma* species in China: a surveillance study. Lancet Infect. Dis. 15, 663-670.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178-2189.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., and Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Quant. Biol.

Liu, L., Dai, J., Zhao, Y.O., Narasimhan, S., Yang, Y., Zhang, L., and Fikrig, E. (2012). *Ixodes scapularis* JAK-STAT pathway regulates tick antimicrobial peptides, thereby controlling the agent of human granulocytic anaplasmosis. J. Infect. Dis. *206*, 1233-1241.

Liu, L., Narasimhan, S., Dai, J., Zhang, L., Cheng, G., and Fikrig, E. (2011). *Ixodes scapularis* salivary gland protein P11 facilitates migration of *Anaplasma phagocytophilum* from the tick gut to salivary glands. EMBO Rep. *12*, 1196-1203.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955-964.

Mac, S., Bahia, S., Simbulan, F., Pullenayegum, E.M., Evans, G.A., Patel, S.N., and Sander, B. (2019). Long-term sequelae and health-related quality-of-life associated with Lyme disease: A systematic review. Clin. Infect. Dis. doi: 10.1093/cid/ciz1158

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics *20*, 2878-2879.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel

717 counting of occurrences of *k*-mers. Bioinformatics *27*, 764-770.
718 Matthews, B.J., Dudchenko, O., Kingan, S.B., Koren, S., Antoshechkin, I., Crawford,
719 J.E., Glassford, W.J., Herre, M., Redmond, S.N., Rose, N.H., *et al.* (2018). Improved
720 reference genome of *Aedes aegypti* informs arbovirus vector control. Nature *563*,
721 501-507.
722 Merino, O., Almazán, C., Canales, M., Villar, M., Moreno-Cid, J.A., Galindo, R.C.,
723 and de la Fuente, J. (2011). Targeting the tick protective antigen subolesin reduces
724 vector infestations and pathogen infection by *Anaplasma marginale* and *Babesia*
725 *bigemina*. Vaccine *29*, 8575-8579.
726 Miller, J.R., Koren, S., Dilley, K.A., Harkins, D.M., Stockwell, T.B., Shabman, R.S.,
727 and Sutton, G.G. (2018). A draft genome sequence for the *Ixodes scapularis* cell line,
728 ISE6. F1000Res. *7*, 297.
729 Oliveira, J.H., Gonçalves, R.L., Lara, F.A., Dias, F.A., Gandara, A.C., Menna-Barreto,
730 R.F., Edwards, M.C., Laurindo, F.R., Silva-Neto, M.A., Sorgine, M.H., *et al.* (2011).
731 Blood meal-derived heme decreases ROS levels in the midgut of *Aedes aegypti* and
732 allows proliferation of intestinal microbiota. PLoS Pathog *7*, e1001320.
733 Pal, U., Li, X., Wang, T., Montgomery, R.R., Ramamoorthi, N., Desilva, A.M., Bao, F.,
734 Yang, X., Pypaert, M., Pradhan, D., *et al.* (2004). TROSPA, an *Ixodes scapularis*
735 receptor for Borrelia burgdorferi. Cell *119*, 457-468.
736 Palmer, W.J., and Jiggins, F.M. (2015). Comparative genomics reveals the origins and
737 diversity of arthropod immune systems. Mol. Biol. Evol. *32*, 2111-2129.
738 Peñalver, E., Arillo, A., Delclòs, X., Peris, D., Grimaldi, D.A., Anderson, S.R.,
739 Nascimbene, P.C., and Fuente, R.P. (2018). Publisher Correction: Ticks parasitised
740 feathered dinosaurs as revealed by Cretaceous amber assemblages. Nat. Commun. *9*,
741 472.
742 Perner, J., Sobotka, R., Sima, R., Konvickova, J., Sojka, D., Oliveira, P.L., Hajdusek,
743 O., and Kopacek, P. (2016). Acquisition of exogenous haem is essential for tick
744 reproduction. Elife *5*. doi: 10.7554/eLife.12318.
745 Pryszcz, L.P., and Gabaldón, T. (2016). Redundans: an assembly pipeline for highly
746 heterozygous genomes. Nucleic Acids Res. *44*, e113.
747 Qin, Z., Zhou, H., and Yangchun (1997). Progress in the research of karyotypes of
748 chromosomes of tick. Acta Arachnol. Sin. *6*, 74-80.
749 Raj, A., Stephens, M., and Pritchard, J.K. (2014). fastSTRUCTURE: variational

750 inference of population structure in large SNP data sets. Genetics *197*, 573-589.
751 Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor
752 package for differential expression analysis of digital gene expression data.
753 Bioinformatics *26*, 139-140.
754 Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat.
755 Methods *17*, 155-158.
756 Rubin, H. (1996). Serine protease inhibitors (SERPINS): where mechanism meets
757 medicine. Nat. Med. *2*, 632-633.
758 Salem, H., Bauer, E., Strauss, A.S., Vogel, H., Marz, M., and Kaltenpoth, M. (2014).
759 Vitamin supplementation by gut symbionts ensures metabolic homeostasis in an insect
760 host. Proc. Biol. Sci. *281*, 20141838.
761 Salmela, L., and Rivals, E. (2014). LoRDEC: accurate and efficient long read error
762 correction. Bioinformatics *30*, 3506-3514.
763 Sanders, H.R., Evans, A.M., Ross, L.S., and Gill, S.S. (2003). Blood meal induces
764 global changes in midgut gene expression in the disease vector, *Aedes aegypti*. Insect
765 Biochem. Mol. Biol. *33*, 1105-1122.
766 Sandve, S.R., Rohlfs, R.V., and Hvidsten, T.R. (2018). Subfunctionalization versus
767 neofunctionalization after whole-genome duplication. Nat. Genet. *50*, 908-909.
768 Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower,
769 C. (2012). Metagenomic microbial community profiling using unique clade-specific
770 marker genes. Nat. Methods 9, 811-814.
771 Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E.,
772 Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for
773 Hi-C data processing. Genome Biol. *16*, 259.
774 Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M.
775 (2015). BUSCO: assessing genome assembly and annotation completeness with
776 single-copy orthologs. Bioinformatics *31*, 3210-3212.
777 Sonenshine, D.E., and Macaluso, K.R. (2017). Microbial invasion vs. tick immune
778 regulation. Front. Cell. Infect. Microbiol. *7*, 390.
779 Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a
780 web server for gene finding in eukaryotes. Nucleic Acids Res. *32*, W309-312.
781 Sultana, H., Neelakanta, G., Kantor, F.S., Malawista, S.E., Fish, D., Montgomery,
782 R.R., and Fikrig, E. (2010). *Anaplasma phagocytophilum* induces actin

783 phosphorylation to selectively regulate gene transcription in *Ixodes scapularis* ticks. J.
784 Exp. Med. *207*, 1727-1743.
785 Tang, H., Krishnakumar, V., and Li, J. (2015). Jcvi: Jcvi Utility Libraries. Zenodo.
786 https://zenodo.org/record/31631/export/xd
787 Thomas, G.W.C., Dohmen, E., Hughes, D.S.T., Murali, S.C., Poelchau, M., Glastad,
788 K., Anstead, C.A., Ayoub, N.A., Batterham, P., Bellair, M., *et al.* (2018). The genomic
789 basis of arthropod diversity. bioRxiv, 382945. http://dx.doi.org/10.1101/382945.
790 Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G.,
791 Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., *et al.* (2013).
792 From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best
793 practices pipeline. Curr. Protoc. Bioinformatics *43*, 11.10.11-11.10.33.
794 Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo,
795 C.A., Zeng, Q., Wortman, J., Young, S.K., *et al.* (2014). Pilon: an integrated tool for
796 comprehensive microbial variant detection and genome assembly improvement. PloS
797 one *9*, e112963.
798 Weisheit, S., Villar, M., Tykalová, H., Popara, M., Loecherbach, J., Watson, M.,
799 Růžek, D., Grubhoffer, L., de la Fuente, J., Fazakerley, J.K., *et al.* (2015). *Ixodes*
800 *scapularis* and *Ixodes ricinus* tick cell lines respond to infection with tick-borne
801 encephalitis virus: transcriptomic and proteomic analysis. Parasit. Vectors *8*, 599.
802 Whiten, S.R., Eggleston, H., and Adelman, Z.N. (2017). Ironing out the details:
803 exploring the role of iron and heme in blood-sucking arthropods. Front. Physiol. *8*,
804 1134.
805 Winzerling, J.J., and Pham, D.Q. (2006). Iron metabolism in insect disease vectors:
806 mining the *Anopheles gambiae* translated protein database. Insect Biochem. Mol. Biol.
807 *36*, 310-321.
808 Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol.
809 Evol. *24*, 1586-1591.
810 Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an *R* package for
811 comparing biological themes among gene clusters. Omics *16*, 284-287.