UNIVERSITE ECOLOGIE
BRETAGNE GEOSCIENCES
LOIRE AGRONOMIE ALIMENTATION

Oniris
École Nationale Vétérinaire, Agroalimentaire et de l'Alimentation — Nantes Atlantique
Santé et alimentation au cœur de la vie

THESE DE DOCTORAT DE

ONIRIS
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 600
Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation
Spécialité : « *Génétique, génomique et bio-informatique* »

Par

« Noémi Pierre CHARRIER »

« Diversité génomique, évolution et adaptation de la tique *Ixodes ricinus* »

Thèse présentée et soutenue à Nantes, le 6 décembre 2018
Unité de recherche : Bioepar
Thèse N° : 2018ONIR119F

Rapporteurs avant soutenance :

Denis Baurain Associate Professor, Université de Liège
Gael Kergoat Directeur de Recherche, INRA

Composition du J ury :

Présidente :	Monique Zagorec	Directrice de recherche, INRA
Examinateurs :	Karen McCoy	Directrice de recherche, CNRS
	Monique Zagorec	Directrice de recherche, INRA
Dir. de thèse :	Claude Rispe	Chargé de recherche, INRA

À ma grand-mère

Résumé

Les tiques sont des acariens hématophages vecteurs de nombreux micro-organismes dont certains sont responsables de maladies humaines ou animales (Borréliose de Lyme par exemple). La tique *Ixodes ricinus*, est largement distribuée en Europe où elle représente le principal vecteur de l'agent responsable de la maladie de Lyme. Trois volets ont été abordés au cours de cette thèse, en réalisant pour chaque point des séquençages à haut-débit de transcriptomes. Dans le premier volet, un catalogue de transcrits a été reconstruit et annoté à partir d'individus provenant de différentes conditions physiologiques (stades de développement, état de gorgement, sexe). Une analyse d'expression différentielle a permis de déterminer quels gènes sont exprimés plus spécifiquement lors du gorgement (protéines cuticulaires notamment, mais également métalloprotéases, etc...). Dans le deuxième volet, la structure génétique d'*I. ricinus* a été explorée à partir de douze populations Européennes. Mes résultats montrent pour la première fois un signal clair de structuration géographique, et d'isolement par la distance, à l'échelle de l'Europe. Dans le troisième volet, j'ai employé une approche phylogénomique sur le groupe des tiques dures : pour cela, j'ai reconstruit les transcriptomes de 27 espèces de tiques (dont neuf espèces séquencées pour ce projet) permettant de proposer un arbre phylogénétique très robuste pour ce groupe.

Mots clés : Transcriptome, *Ixodes ricinus*, RNA-seq, Analyse d'expression différentielle, Génomique des populations, Phylogénomique.

Abstract

Ticks are obligate blood-feeders, able to transmit numerous micro-organisms, including causative agents of human or veterinary diseases (e.g.: Lyme Borreliosis). The tick *Ixodes ricinus* is a widely distributed species in Europe, where it is the principal vector of the Lyme disease agent. Using high-throughput transcriptome sequencing, three lines of research were investigated. First, a large catalogue of transcripts was reconstructed and annotated for ticks in different physiological conditions (feeding or non-feeding, developmental stage, and sex). A differential expression analysis allowed to pinpoint genes associated with blood-feeding at the level of the whole body (genes involved in cuticle production, metalloproteases, etc...). Secondly, I explored the genetic structure of *I. ricinus* at the European scale using transcriptomes from 12 populations. I found a clear signal of phylogeographical strucure probably resulting from an isolation by distance process. Finally, transcriptomes for 27 different species of ticks were reconstructed (including nine species sequenced for this study). This permitted to reconstruct a robust phylogeny for the whole group of hard ticks.

Keywords: Transcriptome, *Ixodes ricinus*, RNA-seq, Differential expression analysis, Population genomics, Phylogenomics

Aknowledgments

Merci d'abord à Claude Rispe, directeur de cette thèse, encadrant et superviseur de ce travail. Merci de m'avoir donné l'opportunité de préparer cette thèse, de m'avoir fait confiance pour travailler sur ce sujet pendant ces trois années. Merci d'avoir continué ma formation de jeune scientifique.

Merci à Gael Kergoat et Denis Baurain d'avoir accepté de rapporter ce travail. Merci à Karen McCoy et Monique Zagorec d'avoir accepté d'examiner ce travail.

Merci aux membres de mon comité de thèse qui en ont suivi le déroulement, merci d'avoir accepté de prendre du temps sur vos emplois du temps et d'avoir évalué mon avancement. Merci donc à Catherine Belloc, Sarah Bonnet, Pierre Peterlongo, Olivier Plantard et Denis Tagu. Un deuxième merci Olivier, pour tes précieuses connaissances sur la biologie des tiques.

Merci à Juliette Bordot et Sylvie Alonso de m'avoir si bien guidé dans le mille-feuille institutionnel ainsi qu'à Erwann Helleu d'avoir compris mes besoins informatiques et d'avoir trouvé des solutions *ad-hoc* en passant par dessus la vétusté du réseau informatique. Merci Claire Bonsergent et Caroline Hervet pour tous les bons moments passés ensemble, pour vos avis et conseils éclairés sur les méthodes de biologie moléculaire. Merci Laurence Malandrin d'avoir permis de confirmer moléculairement l'identité des petites fourmis Nantaises et plus généralement de m'avoir pris au sérieux pendant mon séjour à Bioepar. Merci tout autant à Suzanne Bastian, Albert Agoulon et Maggy Jouglin, Sandie Arnoux. Merci aussi à Aurélie, Racem, Julie, Guillaume, Axelle, Juan, Marie, Romain, Mathilde, George, Niki doctorant.e un jour à Bioepar.

Merci chaleureusement aux membres du groupe TMT (Tique et Maladie à Tique) du REID pour les nombreux échanges bienveillants et très enrichissants. Merci à Pierre de Wit pour ton accueil et tes enseignements. Merci à toute la station de recherche de Tjärnö et à l'Académie des Sciences de Suède d'avoir rendu possible ce séjour profondément épanouissant.
Je tiens aussi à remercier Christophe Douady. Tu m'as laissé faire de nombreux stages pendant ma Licence et mon Master. Je vois dans les travaux préparés pendant cette thèse une continuité avec les chemins parcourus à Lyon. Ces remerciements sont aussi pour Tristan Lefébure et Vincent Lacroix, merci pour vos encouragements et vos enseignements.

Merci au Labo des Savoirs, à tous ses membres, médiateurs et vulgarisateurs, curieux et enthousiastes. Les moments passés avec vous font parti de ceux qui ont empêché que mon gout des sciences du vivant ne soit remplacé par de l'aigreur, de l'indifférence ou pire encore, de l'arrogance. Merci à mes fabuleux, formidables et extra-ordinaires colocataires. Grâce à vous j'ai pu découvrir, et me nourrir de toutes sortes de belles réalisations et d'univers créatifs. Merci donc à Anne-Line, Claire, Lucas, Etienne, Simon et Antoine. Une partie de mes remerciements va au monde du logiciel libre et à tout ceux qui le font vivre.

De très nombreux merci chaleureux et sincères sont encore à délivrer. Je pense à mes parents et ma famille, qui ont encouragé ma curiosité, m'ont trainé en montagne sans me confisquer mes livres. Je pense aussi à Nico et à vous amis de Lycée. Je pense aux nombreuses personnes avec qui j'ai vécu en colocation, vous êtes très nombreux et très importants pour moi. Un immense merci à Jérôme Gippet, ainsi qu'à Julie Toury, Charles Rocabert, Ivaylo Vassilev et Stan Chabert. Les années passées à vos côtés, les fesses sur les bancs de l'université, dans l'herbe du petit parc ou sur le goudron lyonnais m'ont véritablement marquées (humainement et scientifiquement). Nos discussions ont été pour moi un véritable enseignement. Encore Merci.

Et puis si tu ne t'es pas senti concerné par tous ces remerciements et que tu lis ces lignes par curiosité... alors merci pour l'intérêt et la curiosité que tu portes à ce travail.

Contents

1 Dissertation

Contents

1.1 Forewords

This PhD thesis dissertation regroups researches on a group of arthropods, the tick, vectors of disease agents. These researches take advantage of genomic approaches and especially from transriptome sequencing. I used different methodologies in order to study genes' functions via their level of expression, and their genetic diversity. Indeed, variations of the level of expression of these genes allow us to gain clues about specific adaptation linked with biological characteristics, for example which genes are mobilized during the feeding process of the tick. Likewise, the observed genetic diversity gives precious information on the demographic proces in the tick population: what is the overall level of genetic variation and the element structuring this variation (for example, the impact of the geography). Lastly, I explored the evolutionary history of hard ticks, which allowed us to reconstruct the phylogenetical relationships in this group. These results could ultimately help understanding evolution of this group with in particular the evolvability of the host-spectra, as well as the understanding of the maintenance of closely related species with similar ecology.

1.2 Background

1.2.1 Tick and general concern for human and veterinary health

Ticks are terrestrial arthropods, obligate hematophagous ectoparasites, that feed on a variety of tetrapods such as mammals, birds, lizards, snakes or even amphibians [1]. Because of their blood-feeding life-style, they can transmit a wide range of pathogen organisms, from viruses to eukaryotes, which thus represent a concern for human and animal health. In 2004, approximately 10% of the described tick species were considered as vector of pathogens[2] (see Figure 1.1[3] as well as Table 1 in de la Fuente et al, 2008[4]).

1.2.1.1 Ticks concerns for human health

Lyme Borreliosis, or Lyme disease, is the most common tick-borne disease targeting humans in North America and Europe[5]. This disease is caused by a group of spirochaete bacteria, namely *Borrelia burgdorferi* sensu lato (s.l.), which are transmitted by ticks from the *Ixodes* genus. Globally, 85,000 patients annually could be subject to Lyme Borreliosis[6]. Only for the USA, the Center for Disease Control and Prevention (CDC) reports a stable average of 25,000 confirmed cases per year since 2008 (climbing to almost 33,000 with probable cases) (CDC, accessed Oct 10,

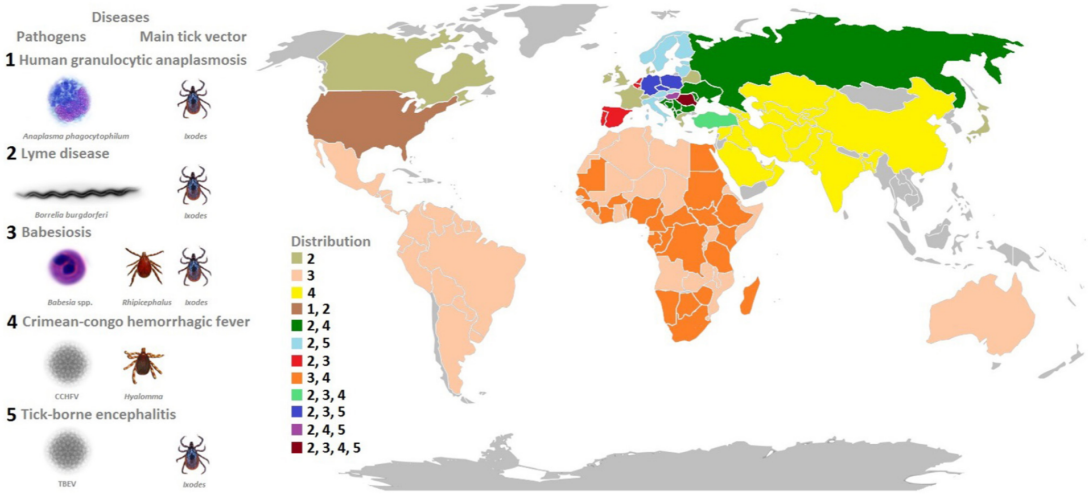

Figure 1.1: Major tick-borne pathogens and their worldwide distribution area (from de la Fuente et al., 2017[3] under CC-BY-4.0)

2018). Tick-borne encephalitis virus is also largely affecting Human[7], as well as multiple spotted fever caused by Rickettsia bacteria[8], and the Crimean-Congo Hemorrhagic fever[9].

1.2.1.2 Ticks concerns for animal health

Ticks have direct consequences on cattle production both by the diseases caused by microorganisms they spread but also by the reported loss of weight induced by tick infestations. In Tanzania, the mortality due to tick-borne disease is estimated at 1.3 millions of cattle per annum[10] and economical losses are estimated to 364 millions of US dollars per annum. An estimation have been done for the dairy industry in Queensland (Australia) and reported an overall cost of 4.1 billions of dollars solely for the 1998 annum[11]. A recent review of the projected economic losses due to vector and vector borne disease in India reported losses superior to 20 billions of US dollars[12].

1.2.1.3 Zoonotic risk

Beside the direct transmission of disease agents to human, ticks can introduce new pathogens by connecting "pathogens reservoirs" and sensible hosts from different environments and/or with different ecological niches. Zoonoses are defined as diseases circulating freely between human and animals. They can be seen as the fate of the livestock farmer life-style associated with our human society[13]. Thereby, predicting zoonotic risk is of main concern, and tick appeared to transmit more pathogens that it was suspected decades ago[14]. This observation resulted

in an integrative approach (under the name of One-health), calling veterinarians and practicians to unify their efforts[15]. Studying tick's biology and correctly appreciating ticks diversity appeared as an important challenge to better understand and effectively control the zoonotic risk induced by their parasitic life-style[16].

1.2.2 Tick's biodiversity

Ticks belong to Arachnida (illustrated in Figure 1.2), a group dated back to 501 Millions years (449-553, 95% confidence interval - see Table2 from Wheat et al., 2013[17]. Ticks are classically described as Acari [18] but the regroupement of parasitiformes and acariformes seems not robust and the reality of the Acari group raised questions[19]. The 900 described extant tick species are classified into three families: i) **Argasidae**, ii) **Ixodidae**, and iii) **Nuttalliellidae**. The **Nuttalliellidae** family is monotypic, with only one species described in 1931 by Bedford in South Africa [20]. *Nutalliella namaqua* is thought to represent a key species to understand the evolution of blood-feeding in ticks[21], given its basal position in the phylogeny of ticks.

The **Argasidae** family or soft ticks contains approximately 200 species[18].They are characterized by leathery integument[18] and do not possess a scutum (a sclerotinized dorsal plate). The particular composition of the cement and lipids deposited on their integument allow them to support relatively low hygrometry[22, 23]. This could explain their presence in relatively dry areas such as tropical and subtropical zones and their presence in arid zones in central Asia and Africa[22, 24]. Mostly nidicolous, they live in microhabitats close to their host achieving multiple short blood meals and between two to eight nymphal stages to become mature adults[22]. Once adult, soft ticks can live as long as 18 years, laying multiple batches of eggs and are able to starve during several years, p.49 in[25]).

The family or hard-ticks regroups ticks characterized by a scutum (a dorsal sclerotinized plate). All the members of this group need a single blood meal to achieve molting between each of their 3 stages (larval to nymphal and nymphal to adult). Adult females need an extra blood meal which can last several days before laying thousands of eggs. There is a morphological character splitting the members in two distinct groups: the relative position of the anal groove in respect of the anal pore. Species that possess this structure posteriorly to the anal pore are called Prostriata (one genus: *Ixodes*) by contrast with the Metastriata (11 genera: *Haemaphysalis* with 167 sp., *Amblyomma* with 130 sp., *Rhypicephalus* with 84sp., *Dermacentor* with 35 sp., *Hyalomma* with 27sp., and 6 others genera with 18sp.). The *Ixodes* genus comprises more than $\sim$34% of the species diversity of the **Ixodidae**. Recently the *Amblyomma* genus has been confirmed to be polyphyletic (see [26, 27]). The genus *Ixodes* appears also to be problematic[28] possibly with two markedly divergent lineages (Australasian and "other- *Ixodes*" following the

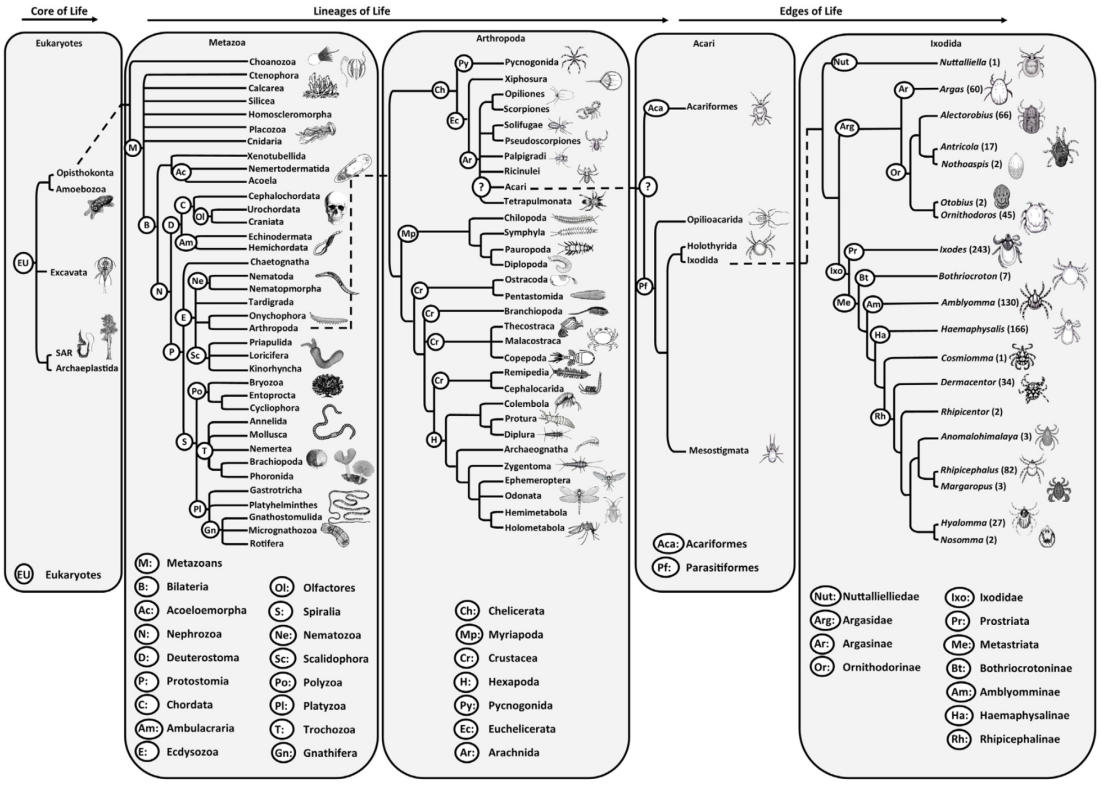

Fig. 1. The lineage specific history of the Ixodida. Indicated are current phylogenetic relationships for Eukaryotes (Adl et al., 2012), Metazoa (Edgecombe et al., 2011), Arthropoda (Giribet and Edgecombe, 2012), Acari (Klompen et al., 2007) and Ixodida (Mans, 2011; Mans et al., 2015). Numbers in brackets indicate current number of species according to Guglielmone et al. (2010) as updated (Mans, 2011).

Figure 1.2: Position of ticks in the current classification of Life. (Taken from Mans et al., 2016[18]. All rights reserved to Elsevier Gmbh.)

terminology of Barker & Murrell[29]). Whether this represents another case of polyphyletic or paraphyletic assemblage is thus still unclear.

Finally, a fourth tick family haas been recently described from fossil evidence: the **Deinocrotonidae**[30]. The fossil record is lacunar and the oldest tick evidence is dated back to 100 Millions years from Burmese amber (see two interesting papers from [31] and [32]).

1.2.2.1 Host-Seeking strategy among hard ticks

Among **Ixodidae**, diverse life-cycle strategies have been described from endophyly to exophyly (respectively living in the nest of their host or host-seeking outside), and from specialist to generalist in terms of host usage. The difference of life-style among hard ticks could cause, or be the result of, different adaptations [33]. For example, the *Ixodes* and *Haemaphysalis* ticks lack eyes while the ticks of the genera *Dermacentor*, *Hyalomma*, and *Rhipicephalus* possess photoreceptors regrouped as

visual organs[33]. The presence or absence of eyes seems to be linked with the preferential strategy of host seeking: through the use of chemoreceptors in the first group and eyes in the second. Indeed, ticks with eyes seems to more frequently use an active strategy rather than ambushing their host. Thus, a mix of this different strategies is generally described. For example, the strategy of host-seeking for *Ixodes ricinus* is called "questing". This strategy is an active waiting for the hosts (see picture Figure 1.3), where ticks climb on a stem of grass depending the environmental conditions (temperature, hygrometry) and sheltering on the litter when the conditions are not favorable to wait. The environmental conditions as well as genetic bases of the questing behavior in *I. ricinus* have been investigated[34]. At the extreme opposite of the hunting strategy (for example *Hyalomma marginatum* is actively host-seeking in open country [35]), there is some ticks living in the nest of their host, (endophylic species) passively waiting for a meal.

Figure 1.3: Questing *I. ricinus*, Carquefou (France), 21-feb-2018. (Canon 1100D - 24 mm)

1.2.2.2 Host-spectra

There is a large range between tick species in terms of effective hosts (host species on which a tick usually feeds on). Some tick species are considered as specialists (meaning that they tend to be associated with a narrow range of hosts) while others can be found on many different species of hosts. For example, three closely related species of hard-ticks (namely *I. vespertillionis*, *I. simplex*, *I. ariadnae*) are specifically found on bats[1, 36], while *I. ricinus* or *I. persulcatus* are found on almost every terrestrial tetrapod in their distribution area[1]. Another example is the bird-associated ticks *I. arboricola*, found on diverse bird species while *I. lividus* is found specifically on a precise bird species: the sand martins (*Riparia riparia*). These observations lead to ask how host-specificity is determined or by which factors the host spectra is influenced. Those questions are not recent, as was pointed by Harry Hoogstraal in 1982[37] who cited questions from Ernst Mayr in 1957: "How, when, and where did host specificity of each parasite group evolve?

How strict is specificity in each case? Why does specificity break down under what circumstances?"[38]. These questions are still an open area of research and need to be addressed towards an evolutionary understanding of the host spectra. This requires a clear view of the evolutionary history of the different ticks lineages and points the need of a solid phylogeny of the hard ticks.
With the capacity of generalist species to feed on diverse hosts, ticks can connect "pathogens reservoirs" and sensible hosts from different environments and/or with different ecological niches. Understanding how plastic is the specificity (and/or host preference) is of importance in the management of zoonotic risks (i.e. the risk of transmission of a pathogen from animals to humans).

1.2.2.3 Vector competency

As described above, host seeking strategy will influence directly which host can be used by the tick to feed on (host-spectra) and therefore which micro-organisms can be carried by the ticks. However, micro-organisms ingested by the tick will not systematically be transmitted to the next host. Indeed, there is multiple environmental filters[39] which should be considered: i) the ability of a micro-organism to be ingested by the tick, ii) the ability by the pathogen to at least survive in the tick body, iii) the ability to reach the salivary glands and to be excreted with salivary products into a new host, and iv) to escape the immunity of the host, at least during the early stage of the introduction. Vector competency is defined as the ability for the tick to ingest, amplify and infect a new host. Some of the environmental filters could be determined by the tick and then subject to phylogenetic inertia.

1.2.2.4 "*ricinus* complex"

The term "*ricinus* complex" was defined to refer to *Ixodes* species sharing similar morphology, life-style and the competency to vectorize *Borrelia burgdorferi* s.l. (the Lyme disease agent)[40]. Previously, many authors referred to these similar species by terms such as the "*Ixodes ricinus* complex", the "*Ixodes persulcatus* complex" and the "*Ixodes ricinus/persulcatus* complex"[40]. It was however later argued that this term should not be used because distant species are also able to transmit *B. burgdorferi* s.l.[41]. Yet, terms such "*ricinus* group" is reported to describe the similarity between these species[42–44] and continue to illustrate the phylogenetical proximity between species of this group. Moreover, hybridization event is for example reported between *I. persulcatus* and *I. ricinus*[45]. In the case of trans-ovarian transmission of pathogen such as *Borrelia myamotoi* [46] (*i.e.* pathogens are transmitted vertically to the next generation of tick), hybridization could permit to extend the range of transmission of a pathogen. A better understanding of the phylogenetical relationships between *Ixodes* species is needed to decipher how are evolving similar features shared by the representatives of the "*ricinus* group/complex" and ultimately measure efficiently the zoonotic risk.

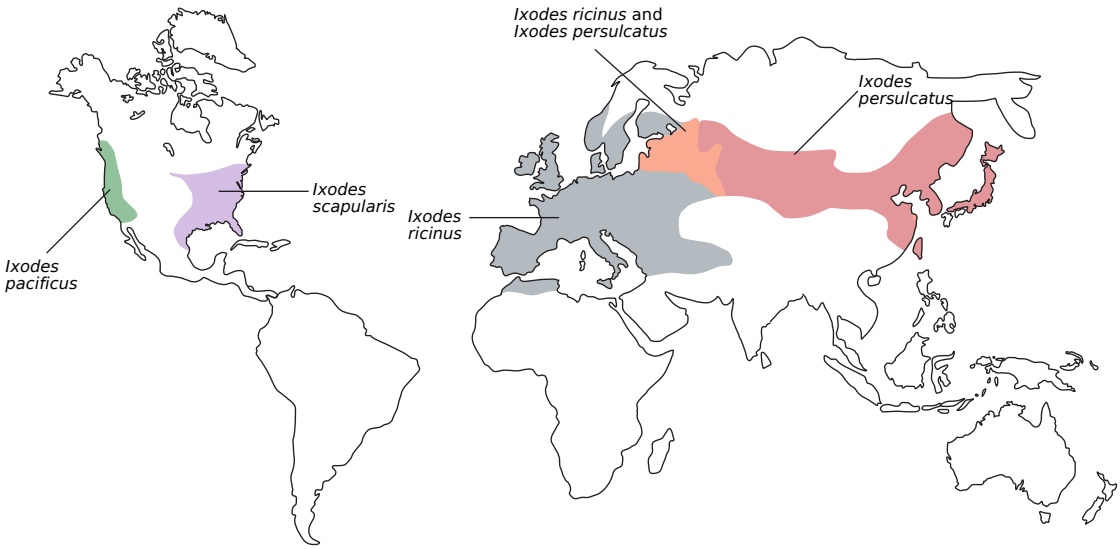

Figure 1.4: Distribution of species from the "*ricinus*" complex, from the European Concerted Action on Lyme Borreliosis

1.2.3 Biology and Ecology of *Ixodes ricinus*

1.2.3.1 Distribution

Ixodes ricinus is observed from North of Africa (Morocco, Tunisia, Algeria) to Scandinavia (south Norway, south Sweden, south of Finland) (see map Figure 1.4). Recent molecular analyses observed a discrepancy in the genetical structure observed between populations from North of Africa compared to those of the rest of Europe[47]. North African populations were elevated at species level by Augustin Estrada-Pena et al. in 2014 with the name of *Ixodes inopinatus*[48]). It appears that this tentative new species might be in sympatry in South of Europe (Spain and Portugal) mostly feeding on lizard. At the Eastern part of Europe, the limit of distribution of *I. ricinus* is borned by *I. persulcatus* (Estonia) and cases of hybridization have been reported [45].

1.2.3.2 Life cycle and host-spectra

Ixodes ricinus needs a minimum of two hosts to become sexually mature and an extra blood meal on an additional host for the female to lay eggs. This life style is therefore described as a three-host life cycle[37] and is illustrated in Figure 1.5. The quantity and quality of the blood, ingested during this third meal will determine the quantity of laid eggs. As illustrated in Figure 1.5, the larva of *I. ricinus* are generally found on small hosts such as micromamals and ground-feeding birds. Roe deer are considered as the principal host but adults can be found on a wide range of hosts (lizards, turtles, birds, human, ...)

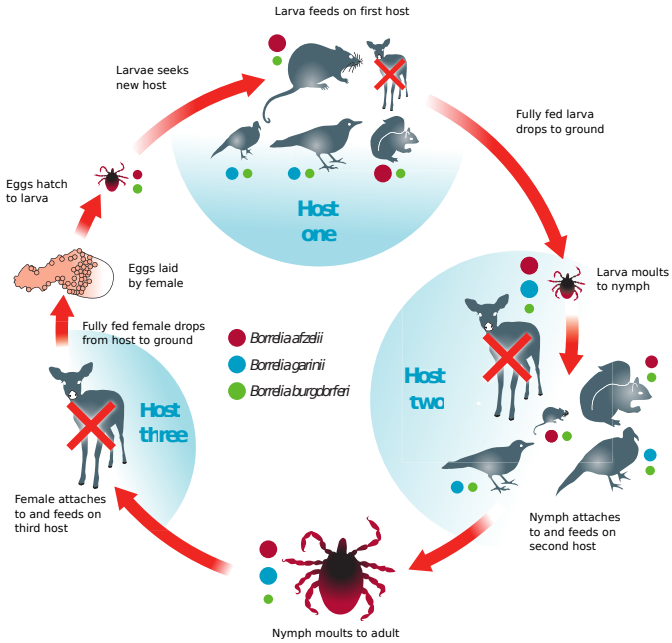

Figure 1.5: Life-cycle of *I. ricinus*, from the European Concerted Action on Lyme Borreliosis

1.2.3.3 Genetic structure

Studying how is genetically structured the population of *I. ricinus* could help understand gene flow and thereby structuration in time space and space. This information could help to predict the spread of zoonotic agents. As it is observed for *I. uriae*[49], potential adaptation to different hosts has been investigated[50–53], particularly at the larvae or nymph. Yet, only a small effect has been observed in some areas of its distribution (especially Southern Europe), suggesting an incipient process of adaptation and eventually of speciation; the authors suggested that it represents the very early stages in the formation of host races[53].
Impact of geography on the genetic structure of *I. ricinus* has been investigated with different sources of material (microsatellites, nuclear and mitochondrial haplotyping, mitogenome reconstruction) under various sampling strategies[42].
While most of the study concluded that no phylo-geographical structure could be found for *I. ricinus*[47, 54–56], Eastern location (Latvia) were found significantly different from Western locations (UK) in terms of haplotype compositions[57]. This result was confirmed by a study which compared the Northern populations to each other[58]. Furthermore, the population from Norway was different from those of UK and north continental Europe. Differences in terms of conclusion between these studies (geographical impact on the genetic structure) could be the consequence of the strategy used to investigate this question. Indeed, studies which concluded that there is no phylogeographical structure relies on very few markers, or few

populations[47, 54–56].

1.2.3.4 morphology and ontogeny

I. ricinus share with the other ecdysozoa clade members the impossibility to grow linearly, constrained by an inelastic exoskeleton. Thus, three stages compose the growing: the larval stage, the nymphal stage, and the adult stage. This number of stages seems conserved in hard ticks but is more variable in soft ticks (from two to eight stages)[22]. To transit from one stage to another, *I. ricinus* tick should achieve molting which is prepared by one long blood meal. The adult female needs an extra bloof meal to lay eggs. It has been reported a full life cycle within 155 days under laboratory conditions[59] but it is generally admitted that the life cycle lasts from 2 to 3 years[60] depending on the host availability and environmental conditions.
Morphologically, ticks share with mites a structure called gnathosoma (or capitulum). This term is used to describe the apical part of the tick body, separated from the rest of the body (or idiosoma). It includes the basis capitula, a pair of palps and the hyposthome. Legs are inserted ventrally on the side of the upper part of the idiosoma. Adults as well as nymphs possess four pairs of legs whereas the larval stages possess only three pairs of legs. Dorsally, a sclerotinized part named scutum is present. Two spiracular plates are present posteriorly to the last pair of legs. These structures are responsible for the regulation of the gaz exchanges and limit the loss of water[61] and they form complex organs able to close and open depending on the environmental conditions[62]. The genital pore is situated ventrally on the upper part of the idiosoma, followed by the anus on the terminal part of the idiosoma. Sexual dimorphism is only present for adults, and sex could not be determined for larval and nymphal stages. Females display porous areas on their basal capituli while males have a scutum covering dorsally most of the idisosoma.

1.2.3.5 Feeding process and salivary glands

Cells from salivary glands will produce many compounds forming a molecular cocktail, generally studied through the lens of the transcription under the term of sialome[63]. During feeding, ticks need first to be firmly attached to the host. After inserting mouthparts into the host skin (mechanical dynamic is described with details in Richter et al., 2013[64]), salivary glands will secrete a “core cement”[65] within 5-15 minutes after attachment[66]. Later on, a second cement (“cortical cement”[65]) is deposited, all along the entire feeding process in the tick *I. scapularis*[67].
The feeding process crystalizes the risk of pathogen transmission in many ways: first, by excreting microorganisms in the host and then by ingesting new microorganisms. As seen above, this feeding process is critical for the tick, especially in

the relation between the host and the parasites. Thereby, gene expressions during feeding in the salivary glands have been subject to previous studies and are the main source of genomic data for the tick *I. ricinus*[63].

1.3 Aims of the thesis

The overall objective of this work was to give insight into the biology of the castor bean tick *Ixodes ricinus* (Linnaeus, 1758). This was achieved by studying *I. ricinus* through the lens of the expressed DNA sequences at three different levels (physiological response: **article I**; population structure: **article II**; and phylogenetic relationships with other tick species: **article III**). Therefore, this thesis is based on the following three parts, which is referred to all along the manuscript by their corresponding roman numerals. More specifically, the aim of each part of this thesis was:

Article I: To explore the transcriptome of *Ixodes ricinus* from a laboratory maintained population with three goals. First, to enrich the catalogue of transcripts publicly available in order to facilitate future functional analysis and genome annotations. Second, to identify the physiological response to feeding at the scale of the whole body. Third, to determine how the laboratory maintenance affected the heterozygosity of the strain with respect to wild individuals.

Article II: To investigate the genetic structure at the European scale of *Ixodes ricinus*, by exploring the transcriptomes of pools of individuals each representing a geographical population. SNP discovery by mapping reads to the high quality reference transcriptome produced in the part 1 of this thesis allowed to generate a substantial number of robust SNPs, which was explored with population genomics tools.

Article III: To consolidate the phylogeny of hard ticks with the power of phylogenomics (in contrast with all studies of tick molecular phylogenies published to date, which have been restrained by the use of a reduced number of markers).

1.4 Methods

High-troughput sequencing of transcripts expressed by the cells of an organism is called RNA-seq. It has been now widely used for many purposes such as the establishment of gene catalogs (including novel discovery of lineage-specific genes), the exploration of expression landscapes (allowing functional inferences), the exploration of the ontogenesis dynamic[68], the survey of intra-specific polymorphisms[69], and the reconstruction of evolutionary history[70]. In **article I**,

transcriptome sequencing allowed us to enrich the catalogue of genes for *I. ricinus* and to explore the physiological response when the tick is feeding at the level of the whole body. I also took advantage of the already available data set of RNA studies based on high-throughput sequencing, to compare the level of polymorphism of different types of tick material, including laboratory maintained strains of ticks, tick cell lines, and ticks from wild populations of ticks. In **article II**, RNA-seq was used to compare polymorphisms between different populations of *I. ricinus* across Europe (population genomics) and investigate the genetic structure. In **article III**, transcriptomes were used to reconstruct a species tree which recapitulates the evolutionary history of hard ticks (phylogenomics).

1.4.1 Genes, transcripts, contigs

Through this work, contig, transcript and gene are employed to designate different types of sequence data. A contig is a sequence of nucleotide virtually reconstructed by an assembly process. It likely originated from the concatenation of the terms contiguous and sequence. A transcript is the product of a gene, one gene could produce many different transcripts during the maturation of the mRNA (Alternative splicing). The term gene is referring to a piece of DNA and by extension the term gene is employed to designate a fragment of DNA responsible for the production of a protein or for a RNA with catalytic activities. Through misuse of language, lengths of contig/transcript/gene are expressed in base pairs (abbreviated 'bp') and follow the International System of Units by the addition of a metrix prefix such as Kb for 1000bp, Mb for 1,000,000bp and so on.

1.4.2 Reconstructing RNA sequences from reads of sequencing

Illumina sequencers produce reads of 70-300 bases. They do not represent the full length transcript and then need to be assembled together in order to retrieve the complete transcript. When a reference genome is available, transcripts can be assembled with the guidance of the genome assembly, otherwise a *de-novo* approach must be used[71]. The published draft genome for *I. ricinus* is still extremely incomplete and simply represents a genome survey [72, 73]. A more complete genome assembly has been published for a closely related species, namely *I. scapularis*[74], but it still has many gaps and partial or missing genes. Therefore, I used a *de-novo* approach to reconstruct the full length transcripts of *I. ricinus*. Briefly, a catalogue of subsequences (also called *k-mer*) is constructed from all the sequences of k-letters found in the set of sequenced reads. Then, these indexed elements are linked with the sequenced reads and form a graph when *k-mers* overlap each other on a length of $k-1$. This de Bruijn graph is then de-convoluted (or linearized) in order to reconstruct the full length transcripts. Through this PhD work, I used the `Trinity` assembler[75], an integrated tool for reference transcriptome reconstruction and analysis[76]. In particular, this platform incorporates the `Trinotate` tool, an

emerging open source alternative to the proprietary solution proposed by `blast2go`.

1.4.3 Preparing and annotating sequences

A raw assembly may produce up to hundreds of thousands contigs which represent complete or fragmented biological transcripts. This large amount of data is not human readable and should be processed before analyzed. The first step is to reduce the redundancy of the assembly by keeping for example the contig which is the most representative of a set of similar contigs. Indeed, because of the splicing process and the polymorphisms between individuals, one gene will generate multiple transcripts. Depending on the goal of the analysis, different levels of compression can be achieved. For the **article II** and **III**, the goal was to obtain as much as possible one contig per gene and so a transcript embedding most of the exons (coding sequences of the gene). This approach is classically done by keeping the longest sequence when two sequences are very similar, using for example the `cd-hit` software[77].
One of the main step achieved in the **article I** was to annotate the different transcripts. Classically, it consists in comparing unknown sequences (reconstructed transcripts) with well annotated sequences. The assumption behind this method is that the similarity between sequences is due to conservation of the functional property of a protein through the molecular evolution mechanisms. This goal was met by following the guidelines of the `Trinotate` pipeline of annotation[76]. Briefly I compared the set of contigs to already annotated database such as Swissprot and Uniref90 and looked for conserved domains of protein (PFAM-A database[78]). Then, I used particular software predicting peptide signals[79] and transmembrane domains[80]. All these layers of information were combined with `Trinotate`. The advantage of this method is to take advantage of the annotation supplied for model organisms and to predict functions associated with the assembled transcripts, by means of semantic convention introduced by the Gene Ontology consortium[81].

1.4.4 Differential gene expression and functional enrichment

RNA-seq methods allow us not only to reconstruct the transcript sequences but also to quantify the level of expression of those transcripts. Indeed, the quantity of sequenced reads per transcript is directly correlated with the number of transcripts extracted from biological tissues. Thereby, gene expression could be compared across physiological conditions in order to identify genes implicated in those conditions[82] - genes differentially expressed.

1.4.5 Calling and quantifying variants

By the transmission of the genome from one generation to another, coupled with the mutation and the recombination processes, many information could be retrieved from the differences between comparable sequences (polymorphisms). This polymorphisms were precious in the **article I**, **II** and **III** but for different purposes which I will develop below. In **article I**, and **II**, only a particular class of substitutions were considered: the bi-allelic single nucleotide substitutions (SNPs) – positions of interest (or character) in the contig must have two and only two possibilities over four bases (two states of character), and for which no polymorphism was observed on the neighborhood.
In **article I**, SNPs were detected using a graph-based approach[83], allowing to overpass the need for a reference genome or transcriptome. Briefly, SNPs are identified via the bubbles they form in the Transcriptome de Bruijn graph. For the comparison of different populations in the **article II**, I used a more classical approach based on mapping. This approach consists in mapping the reads on a reference transcriptome. In both cases, three caveats have to be avoided: i) false SNPs resulting from sequencing errors, ii) polymorphisms due to alternative splicing event, and iii) polymorphisms due to recent paralogy (duplication event which occurred recently). The error rate associated with sequencing error in Illumina technology is around 1%, independently from the quality of the prediction measured by the phred score. Therefore, unfrequent SNPs (minor allele frequency <1-2 %) were not considered while there are indicative of the recent demographic history.
In the **article III**, sequences were compared between different species and so make use of a different conceptual framework: molecular phylogeny, contrary to population genetic conceptual framework used in **article I** and **II**. Instead of allele frequency or Fixation index measures (F_{ST}), distances between species were estimated using models of substitution.

1.4.6 Predicting Orthologues

One way to access to the evolutionary history of a particular group of organisms is to achieve the reconstruction of a species tree. A species tree represents the most probable scenario of speciation (an event of split between lineages generating two taxonomical units). This most probable scenario is based on the similarity and dissimilarity between a set of taxonomical unit measured from characters. In the case of DNA (or proteins), the different characters are the different sites all along the DNA molecules, and the states of the character are the different bases. In order to correctly infer the most probable scenario of speciation, the alignment should guarantee the homology of positions.
Indeed, a particular gene could have been subject to duplication and therefore two copies in the genome will accumulate differences independently, one copy could be lost. Consequently, reconstructing the species tree should use gene reflecting the

speciation event and not those experiencing duplication events. Genes reflecting the speciation event are called orthologs by contrast with paralogs[84].
Multiple strategies can be applied to predict orthologs[85]. The strategy used in the **article III** is a conservative one-to-one strategy: we makes the hypothesis that if one and only one particular sequence is found in most of the species, this sequence should not have been subject to a duplication event.

1.5 Main results

The work presented in **article I** allowed us to obtain a highly complete transcriptome. Clearly, extensive transcriptomic data for this species already existed, but the choice to focus on a narrow range of tissues or conditions (most often salivary glands during the feeding process) had greatly restrained the diversity of discovered transcripts). By contrast, our transcriptome project was based on a design intended to obtain the broadest possible range of transcripts, including whole bodies of two stages (nymphs and adults), males and females, and both feeding and non-feeding conditions. The collection of transcripts we deposited in the public database (Genbank, TSA division) therefore increased significantly the global catalogue of known transcripts and coding genes for *I. ricinus*.

Based on the annotated GO terms, our analysis of differential expression highlighted the importance of cuticular related products during the slow phase of engorgement. Surprisingly, we did not found a clear enrichment of functions associated with the feeding process in salivary glands. Some of these "missing" terms (such as "metalloproteases") were retrieved however in the semantic analysis of the annotated terms with help of a word cloud representation. Moreover, these results point out the importance of the choice of tissues, physiological states, and time-scale to study the transcriptome of an organism.

The use of transcriptomes, from our data sets and already published data sets, permitted to compare the genetic diversity from different sources of material. We therefore could evaluate heterozygosity for different sources of material (wild strains, cellular lineage, F1). The results of this work were in agreement with our assumption that laboratory maintained strains were subject to inbreeding when compared to natural populations.
In this work, a SNP detection was performed with a "direct from the reads" approach, using the Transcriptome De Bruijn Graph (T-DBG) -ie. without mapping, which is the most common strategy for SNP calling. We thereby provided one of the first (if not the first) real case of use for this approach thanks to the `Kissplice` software.

A different "variant calling" strategy was used to study the polymorphisms in *I. ricinus* at the European scale (**article II**), which is justified by the fact that we had obtained a good and complete reference transcriptome (with reduced redundancy)

and by the volume of data obtained in this second study (which makes the mapping more efficient than the *de-novo* approach). We indeed followed a classical workflow, based on read mapping against a reference transcriptome, the one built upon the work of **article I**. The main result of this population transcriptomics approach was the strong effect of geographical distance on the allelic frequencies of the 12 sampled populations. This result was both observed by F_{ST} estimations and a Principal Coordinates Analysis (PCoA) using the minor allele frequencies. Results from the Mantel tests showed a significant correlation between the geographical distance and different genetic distances. The first corresponded to F_{ST} measure and the second to the Euclidian distance between the population on the PCoA plan formed by the Axis 1 and 2. These results are in favor of an Isolation By Distance effect (IBD), but this interpretation has to be taken with precaution (Diniz-Filho et al., 2013). Indeed, it was also notable that two groups of populations were found when we hierarchically clustered the populations with an UPGMA method based on the F_{ST} distances. It appeared that these two groups correspond respectively to populations from the West (Spain, Ireland, UK) and the East (Germany, Switzerland, Czech Republic, Hungary) of Europe. We finally note the "outgroup" position from Finland.

The phylogenomic approach (**article III**) was in favor of the monophyly of the *Ixodes* genus. Thus, the bayesian inference on the SCO75 and SC50 supermatrices with the CAT-GTR [86] resulted in a split in favor of an origin of the Metastriata from the “other” *Ixodes* lineage (following the terminology of Barker & Murrell, 2004[29]). Yet, the central branch separating these three groups was small enough to enview the possibility of a multifurcation. Our results highlighted the important divergence between the two *Ixodes* clades, namely the Australasian lineage and the "other *Ixodes*".

1.6 Discussion and perspectives

1.6.1 How complete are the transcriptomic ressources? Could they still be enriched?

The original conditions of sequencing successfully enriched the catalogue of transcripts for *I. ricinus*. Indeed, we sequenced RNA from whole body and not from targeted organs together with original materials (males, unfed nymphs) (see Figure 2 of the **article I**). It is likely that we have not reached a complete description of the transcript catalogue, given that many genes remain partial (a significant proportion of genes were predicted to be "partial") and probably many rare transcripts are still entirely missing. This leads us to propose the sequencing of yet unexplored physiological states (*in-natura* questing ticks, molting nymphs or laying females), or specific tissues. Currently, my research group is studying a recently sequenced

transcriptome for brains of *I. ricinus*, which confirms the potential of sequencing the transcriptome of new organs (many genes involved in neural processes are highly specific to the synganglion). We also suggest that tick specific organs, involved at precise time-points of the tick biology could provide a valuable source of information: this could be the case of Haller's organs. Haller described this particular organ situated on the tarsus of the first pair of legs [87, 88]. Initially thought to be an auditory organ, it is now rather considered as a complex sensory-organ helping ticks in the host detection via olfaction and sensing of humidity, temperature and carbon dioxide.
Studying the expressed genes in this particular organ could give us information on the molecular mechanisms implicated in the host-detection (CO2, temperature, etc). This would be all the more interesting and novel that genes involved in sensing are often highly group-specific and fast-evolving (i.e. they are not easily identifiable by simple homology). Another example could be the "Egg-waxing organ", an organ involved in egg production which was firstly described by Géné in 1848. This organ, situated dorsally between the scutum and the gnathosoma permits to cover the eggs with a waterproof substance[89]. This under-studied organ, could provide interesting details on the molecules used to protect eggs from desiccation.

1.6.2 How to accurately evaluate the number of tick transcripts?

The high number of non-redundant transcripts assembled (after compression based on similarity with the `cd-hit-est` tool[77]) raised questions on the total number of different transcripts that a tick can express. We can wonder if these transcripts originated from distinct genes through an important gene duplication history, by an amplification of specific gene families and/or if these transcripts originated from a highly active alternative splicing. These hypotheses are not opposed while duplication and splicing process could act separately[90]. Moreover, a relatively high number of transcripts has no sequence homology with annotated sequences from Uniref90 and Swissprot. A similar observation was pointed by Gibson et al. in 2013[91] when sequencing expressed sequences for the tick *Amblyomma americanum*. Authors conclude that the lack of genomic resources as well as a large amount of tick specific gene could explain these results. Our work tried to fill this gap but more studies should be performed to disentangle the different factors and hypotheses.

1.6.3 Genetic structure

Our results from the **article II** showed that geographical distance was the most important factor to consider for explaining the genetic structure of *I. ricinus*. This is an important result which contrasts with several past works on the differentiation of *I. ricinus* in Europe. Most of the previous investigations concluded indeed that

no phylogeographical structure could be found for *I. ricinus*[47, 54–56]. Significant differentiation between East and West populations were however observed [57, 58]. We argue that the discrepancy between these studies lies in the number of markers and sampling designs (the first group of studies used very few markers, or few populations). Once enough markers are studied, geographic structuration makes no doubt.
Beside the suspected Isolation By Distance effect, we observed two well-marked clusters of populations samples (namely IR, ES, FR-01 and UK on one side and AL, RT, SUI and FR-03) which are coherent with previous East/west differences reported. Yet, the presence of the FR-01 and ES in the East cluster together with UK and IR, seems to indicate that seas are not a barrier for the dispersion of *I. ricinus*. This East/West separation could be explained by migrating birds, carrying ticks from South to North and inversely. Two different flyways are surprisingly fitting perfectly with the geographical repartition of our two clusters (namely the Black sea/Mediterranean flyway and the East Atlantic flyway)[92]. We hypothesized that the differences between east and west populations could originate from the two roads of bird migration which could tend to homogenize genetic diversity in their respective area by long-distance dispersion.
We propose that the case of the Finland population (the most differentiated pool) could be explained either by events of hybridization with the *I. persulcatus*[45], or by a reduced connection and gene flow between Finland and the other European locations.

1.6.4 Ricinus Complex

In the article **III**, we found that five species formed one clade with an extreme closeness, namely *I. acuminatus*, *I. persulcatus*, *I. ricinus*, *I. scapularis*, and *I. ventalloi*. Historically, some of these species were grouped together under the term "*ricinus* complex". This term was first used to encompass their ability to transmit Borrellia, the agent responsible of the lyme disease, their similar ecological life-styles (generalist) and their morphological similarities[40]. However, a study deconstructed this term arguing that some tick species could transmit Borrellia without beeing evolutionary close from the group of closely related species[41]. Furthermore, some species unable to transmit Borrelia were found belonging to the same phylogenetical group than other species from the "*ricinus*-complex". In light of the results presented in the **article III**, we suggested that the expression "*ricinus* complex" should be reintroduced, but limited to the *Ixodes* species sharing close phylogenetic relationship with *I. ricinus* without regards of their ecological caracteristics. Indeed, while *I. acuminatus* is generally found feeding only on micro-mammals[1], or that *I. ventalloi* is not reported to transmit Borrellia, both seem to belong to the complex.It raised questions about the plasticity of these ecological caracteritics.
In order to have a more precise insight into the evolutionary history of this com-

plex, we could take advantage of the methods used in **article II** which allowed to identify SNPs in order to perform a Coalescent approach. Indeed, we are expecting incomplete lineage sorting (ILS), producing incongruance between phylogenetic trees reconstructed with classical substitution models.

These relationships between closely related species of the "*ricinus* group" could also be investigated using other genomic ressources in order to test for introgression (hybridization is reported between *I. ricinus* and *I. persulcatus*). Yet, *I. acuminatus* was found closer from *I. ricinus* while *I. persulcatus* was found sister group of *I. scapularis*. The advent of complete genome sequencing (both for *I. ricinus*, and for other *Ixodes* species) should give tools to explore with more precision the evolutionary history of gene exchanges and speciation within that complex.
Questions could be also be asked for *I. trianguliceps* which are found on the same geographical area and also feeding on micro-mammals[93]. All these elements raised questions about the ability of ticks to maintain distinct lineages. These questions could be linked with the particularly low diversity in terms of species. Indeed, there are 1000 described species in the Ixodida (which originated presumably 300 Millions years ago) compared to the Hymenoptera (153,000 described species, originated 339-229 MA [CI-95%])[94], or Coleoptera (>350,000 described species)[95].

1.6.5 Has there been a large duplication event in an ancestor of ticks?

Van Zee et al. in 2016 proposed that a Large Duplication Event (LDE) could have occurred about 40 million years ago, in an ancestor of *Ixodes scapularis*[96]. This is surprising at first sight, given the very rare occurrence of Whole Genome Duplications events in arthropods. Authors of a recent study argued however that multiple events of that type may have occurred in hexapods[97]. However, they give compelling evidence only in the case of *Bombyx mori* and alternative explanantions could also explain the excess of divergent paralogs they detected in many species. We here propose to further explore the extensive transcriptomic data obtained in our **article III** (with transcriptomes from 27 different species of ticks). A preliminary exploration of this data set (results not shown), following the methodology of Blanc and Wolfe in 2004[98] found no significant deviation from a null model of gene birth-death process. Therefore this does not give support to LDE in tick ancestors.

1.6.6 Genome architecture and Life history traits

How lifestyle impacts or is impacted by the genomic features is an essential question. By their dependence from their host, ticks could experience demographic uncertainties. For example, some very specialized ticks will depend on their host - *I. vespertilionis* is known to feed exclusively on a particular species of bat[1]. These ecological particularities could influence the demographic process taking place in

these species. By influencing the demographic parameters, ecological caracteristics can impact the strength of selection, leading to difficulty for the selection to purge weakly deleterious mutations.
Similarly, by interacting durably and intimately with their host, we could wonder how gene repertoires are evolving to evade host-immunitary system, and how host-spectra is influenced/influencing the gene repertoire. These questions were not reached during my PhD work but now benefited from a phylogenetic backbone. Indeed, in the **article III**, a robust phylogenetical tree was reconstructed, even if there is an uncertainty about the evolutionary history of the Metatrastriata/Prostriata node.
For example, during my PhD work, I supervised Hadrien Jouanne (Master 1 student) who focused on the design of a quantitative metrics which could comprehend the Host-spectra. Using the reconstructed tree as a backbone, other trees could be aggregated with the super-network approach. This resulting super-network could take advantage of the hundreds mitogenomes publicly available as well as classical markers for DNA barcoding (16S and COI). The evolutionary scenario could permit to explore how is evolving host spectra, vector competency as well as host-seeking strategy (from passive with endophylic life-style to active hunter) and possibly to provide clues to the question of Mayr (see page 16 [37, 38]).

Bibliography

1. A. A. Guglielmone et al.: *The hard ticks of the world*. Springer, 2014, pp. 978–94. doi: 10.1007/978-94-007-7497-1.
2. F. Jongejan and G. Uilenberg: The global importance of ticks. *Parasitology* **129**(S1) (2004), S3–S14. doi: 10.1017/S0031182004005967.
3. J. de la Fuente et al.: Tick-Pathogen Interactions and Vector Competence: Identification of Molecular Drivers for Tick-Borne Diseases. *Frontiers in Cellular and Infection Microbiology* **7** (Apr. 2017). doi: 10.3389/fcimb.2017.00114.
4. J. de la Fuente, A. Estrada-Peña, J. M. Venzal, K. M. Kocan, and D. E. Sonenshine: Overview: Ticks as vectors of pathogens that cause disease in humans and animals. *Frontiers in Bioscience* **Volume**(13) (2008), 6938. doi: 10.2741/3200.
5. G. Stanek, G. P. Wormser, J. Gray, and F. Strle: Lyme borreliosis. *The Lancet* **379**(9814) (2012), 461–473. doi: 10.1016/S0140-6736(11)60103-7.
6. Z. Hubálek: Epidemiology of Lyme Borreliosis. *Lyme Borreliosis* (2009), 31–50. doi: 10.1159/000213069.
7. L. Lindquist and O. Vapalahti: Tick-borne encephalitis. *The Lancet* **371**(9627) (2008), 1861–1871. doi: 10.1016/S0140-6736(08)60800-4.
8. P. Parola et al.: Update on Tick-Borne Rickettsioses around the World: a Geographic Approach. *Clinical Microbiology Reviews* **26**(4) (Oct. 2013), 657–702. doi: 10.1128/cmr.00032-13.
9. C. A. Whitehouse: Crimean–Congo hemorrhagic fever. *Antiviral Research* **64**(3) (2004), 145–160. doi: 10.1016/j.antiviral.2004.08.001.
10. F. M. Kivaria: Estimated direct economic costs associated with tick-borne diseases on cattle in Tanzania. *Tropical Animal Health and Production* **38**(4) (May 2006), 291–299. doi: 10.1007/s11250-006-4181-2.
11. N. Jonsson, R. Bock, and W. Jorgensen: Productivity and health effects of anaplasmosis and babesiosis on Bos indicus cattle and their crosses, and the effects of differing intensity of tick control in Australia. *Veterinary Parasitology* **155**(1) (2008), 1–9. doi: 10.1016/j.vetpar.2008.03.022.
12. B. W. Narladkar: Projected economic losses due to vector and vector-borne parasitic diseases in livestock of India and its significance in implementing the concept of integrated practices for vector management. *Veterinary World* **11**(2) (Feb. 2018), 151–160. doi: 10.14202/vetworld.2018.151-160.
13. J. Diamond: *Guns, germs, and steel: the fates of human societies*. NY: WW Norton & Company, 1997.

14. C. D. Paddock: The Science and Fiction of Emerging Rickettsioses. *Annals of the New York Academy of Sciences* **1166**(1) (May 2009), 133–143. doi: 10.1111/j.1749-6632.2009.04529.x.

15. F. Dantas-Torres, B. B. Chomel, and D. Otranto: Ticks and tick-borne diseases: a One Health perspective. *Trends in Parasitology* **28**(10) (Oct. 2012), 437–446. doi: 10.1016/j.pt.2012.07.003.

16. J. M. Medlock et al.: Driving forces for changes in geographical distribution of *Ixodes ricinus* ticks in Europe. *Parasites & Vectors* **6**(1) (2013), 1. doi: 10.1186/1756-3305-6-1.

17. C. W. Wheat and N. Wahlberg: Phylogenomic Insights into the Cambrian Explosion, the Colonization of Land and the Evolution of Flight in Arthropoda. *Systematic Biology* **62**(1) (Nov. 2012), 93–109. doi: 10.1093/sysbio/sys074.

18. B. J. Mans et al.: Ancestral reconstruction of tick lineages. *Ticks and Tick-borne Diseases* **7**(4) (2016). TTP8-STVM Special Issue, 509–535. doi: 10.1016/j.ttbdis.2016.02.002.

19. P. P. Sharma et al.: Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal. *Molecular Biology and Evolution* **31**(11) (2014), 2963–2984. doi: 10.1093/molbev/msu235.

20. G. A. H. Bedford: *Nuttalliella namaqua*, a New Genus and Species of Tick. *Parasitology* **23**(02) (Apr. 1931), 230–232. doi: 10.1017/s0031182000013573.

21. B. J. Mans, D. de Klerk, R. Pienaar, and A. A. Latif: *Nuttalliella namaqua*: A Living Fossil and Closest Relative to the Ancestral Tick Lineage: Implications for the Evolution of Blood-Feeding in Ticks. *PLoS ONE* **6**(8) (Aug. 2011). Ed. by P. L. Oliveira, e23675. doi: 10.1371/journal.pone.0023675.

22. L. Vial: Biological and ecological characteristics of soft ticks (Ixodida: Argasidae) and their impact for predicting tick and associated disease distribution. *Parasite* **16**(3) (Sept. 2009), 191–202. doi: 10.1051/parasite/2009163191.

23. A. D. Lees: Transpiration and the Structure of the Epicuticle in Ticks. *The Journal of Experimental Biology* **23**(3-4) (Apr. 1947), 379.

24. P.-C. Morel: Contribution á la connaissance de la distribution des tiques (Acariens, Ixodidae et Amblyomnisdae) en Ethiopie continentale. PhD thesis. 1969.

25. D. Sonenshine and R. Roe: *Biology of Ticks, 2nd Ed.* Oxford University Press, 2014.

26. T. D. Burger, R. Shao, L. Beati, H. Miller, and S. C. Barker: Phylogenetic analysis of ticks (Acari: Ixodida) using mitochondrial genomes and nuclear rRNA genes indicates that the genus *Amblyomma* is polyphyletic. *Molecular Phylogenetics and Evolution* **64**(1) (2012), 45–55. doi: 10.1016/j.ympev.2012.03.004.

27. T. D. Burger, R. Shao, and S. C. Barker: Phylogenetic analysis of the mitochondrial genomes and nuclear rRNA genes of ticks reveals a deep phylogenetic structure within the genus *Haemaphysalis* and further elucidates the polyphyly of the genus *Amblyomma* with respect to *Amblyomma sphenodonti* and *Amblyomma elaphense*. *Ticks and Tick-borne Diseases* **4**(4) (2013), 265–274. doi: 10.1016/j.ttbdis.2013.02.002.

28. J. Klompen, W. C. Black, J. E. Keirans, and D. E. Norris: Systematics and Biogeography of Hard Ticks, a Total Evidence Approach. *Cladistics* **16**(1) (Mar. 2000), 79–102. doi: 10.1111/j.1096-0031.2000.tb00349.x.

29. S. C. Barker and A. Murrell: Systematics and evolution of ticks with a list of valid genus and species names. *Parasitology* **129 Suppl** (2004), S15–36. doi: 10.1017/S0031182004005207.

30. E. Peñalver et al.: Ticks parasitised feathered dinosaurs as revealed by Cretaceous amber assemblages. *Nature Communications* **8**(1) (Dec. 2017). doi: 10.1038/s41467-017-01550-z.

31. L. Chitimia-Dobler, B. C. De Araujo, B. Ruthensteiner, T. Pfeffer, and J. Dunlop: *Amblyomma birmitum* a new species of hard tick in Burmese amber. *Parasitology* **144**(11) (June 2017), 1441–1448. doi: 10.1017/s0031182017000853.

32. A. Estrada-Peña and J. de la Fuente: The fossil record and the origin of ticks revisited. *Experimental and Applied Acarology* **75**(2) (May 2018), 255–261. doi: 10.1007/s10493-018-0261-z.

33. I. Uspensky: Preliminary Observations on Specific Adaptations of Exophilic Ixodid Ticks to Forests or Open Country Habitats. *Experimental & Applied Acarology* **28**(1) (May 2002), 147–154. doi: 10.1023/A:1025303811856.

34. J. L. Tomkins, J. Aungier, W. Hazel, and L. Gilbert: Towards an Evolutionary Understanding of Questing Behaviour in the Tick *Ixodes ricinus*. *PLoS ONE* **9**(10) (Oct. 2014). Ed. by C. R. E. Lazzari, e110028. doi: 10.1371/journal.pone.0110028.

35. A. Estrada-Peña, M. Martínez Avilés, and M. J. Muñoz Reoyo: A Population Model to Describe the Distribution and Seasonal Dynamics of the Tick *Hyalomma marginatum* in the Mediterranean Basin. *Transboundary and Emerging Diseases* **58**(3) (Jan. 2011), 213–223. doi: 10.1111/j.1865-1682.2010.01198.x.

36. S. Hornok et al.: Contributions to the morphology and phylogeny of the newly discovered bat tick species, *Ixodes ariadnae* in comparison with *I. vespertilionis* and *I. simplex*. *Parasites & vectors* **8**(1) (2015), 47. doi: 10.1186/s13071-015-0665-0.

37. H. Hoogstraal and A. Aeschlimann: Tick-host specificity. *Bulletin de la société entomologique suisse* **55** (1982), 5–32.

38. E. Mayr: Evolutionary aspects of host specificity among parasites of vertebrates. *SYMPOSIUM ON HOST SPECIFICITY AMONG PARASITES OF VERTEBRATES (1st), University of Neuchatel, April 15-18, 1957.* Neuchatel. 1957, 7–14.

39. C. Combes: *The art of being a parasite*. University of Chicago Press, 2005.

40. J. Keirans, G. Needham, and J. Oliver Jr: The *Ixodes ricinus* complex worldwide: diagnosis of the species in the complex, hosts and distribution. *Acarology IX* **2** (1999), 341–347.

41. G. Xu, Q. Q. Fang, J. E. Keirans, and L. A. Durden: Molecular phylogenetic analysis indicate that the *Ixodes ricinus* complex is a paraphyletic group. *Journal of Parasitology* **89**(3) (June 2003), 452–457. doi: 10.1645/0022-3395(2003)089[0452:mpaitt]2.0.co;2.

42. A. Araya-Anchetta, J. D. Busch, G. A. Scoles, and D. M. Wagner: Thirty years of tick population genetics: A comprehensive review. *Infection, Genetics and Evolution* **29** (2015), 164–179. doi: 10.1016/j.meegid.2014.11.008.

43. S. Kovalev, S. Fedorova, and T. Mukhacheva: Molecular features of *Ixodes kazakstani*: first results. *Ticks and Tick-borne Diseases* **9**(3) (2018), 759–761. doi: https://doi.org/10.1016/j.ttbdis.2018.02.019.

44. A. Estrada-Peña, J. M. Venzal, and S. Nava: Redescription, molecular features, and neotype deposition of *Rhipicephalus pusillus* Gil Collado and *Ixodes ventalloi* Gil Collado (Acari, Ixodidae). *Zootaxa* **4442**(2) (July 2018), 262. doi: 10.11646/zootaxa.4442.2.4.

45. S. Kovalev, I. Golovljova, and T. Mukhacheva: Natural hybridization between *Ixodes ricinus* and *Ixodes persulcatus* ticks evidenced by molecular genetics methods. *Ticks and Tick-borne Diseases* **7**(1) (2016), 113–118. doi: 10.1016/j.ttbdis.2015.09.005.

46. G. van Duijvendijk et al.: Larvae of *Ixodes ricinus* transmit *Borrelia afzelii* and *B. miyamotoi* to vertebrate hosts. *Parasites & Vectors* **9**(1) (Feb. 2016). doi: 10.1186/s13071-016-1389-5.

47. R. Noureddine, A. Chauvin, and O. Plantard: Lack of genetic structure among Eurasian populations of the tick *Ixodes ricinus* contrasts with marked divergence from north-African populations. *International Journal for Parasitology* **41**(2) (2011), 183–192. doi: 10.1016/j.ijpara.2010.08.010.

48. A. Estrada-Peña, S. Nava, and T. Petney: Description of all the stages of *Ixodes inopinatus* n. sp. (Acari: Ixodidae). *Ticks and Tick-borne Diseases* **5**(6) (2014), 734–743. doi: 10.1016/j.ttbdis.2014.05.003.

49. K. D. Mccoy, T. Boulinier, C. Tirard, and Y. Michalakis: Host specificity of a generalist parasite: genetic evidence of sympatric host races in the seabird tick *Ixodes uriae*. *Journal of Evolutionary Biology* **14**(3) (May 2001), 395–405. doi: 10.1046/j.1420-9101.2001.00290.x.

50. T. d. Meeûs, L. Béati, C. Delaye, A. Aeschlimann, and F. Renaud: Sex-biased genetic structure in the vector of Lyme Disease, *Ixodes ricinus*. *Evolution* **56**(9) (Sept. 2002), 1802–1807. doi: 10.1111/j.0014-3820.2002.tb00194.x.

51. F. Kempf, T. de Meeûs, C. Arnathau, B. Degeilh, and K. D. McCoy: Assortative Pairing in *Ixodes ricinus* (Acari: Ixodidae), the European Vector of Lyme Borreliosis. *Journal of Medical Entomology* **46**(3) (May 2009), 471–474. doi: 10.1603/033.046.0309.

52. F. Kempf, K. D. McCoy, and T. D. Meeûs: Wahlund effects and sex-biased dispersal in *Ixodes ricinus*, the European vector of Lyme borreliosis: New tools for old data. *Infection, Genetics and Evolution* **10**(7) (2010), 989–997. doi: 10.1016/j.meegid.2010.06.003.

53. F. Kempf et al.: Host races in *Ixodes ricinus*, the European vector of Lyme borreliosis. *Infection, Genetics and Evolution* **11**(8) (2011), 2043–2048. doi: 10.1016/j.meegid.2011.09.016.

54. S. Casati, M. Bernasconi, L. Gern, and J.-C. Piffaretti: Assessment of intraspecific mtDNA variability of European *Ixodes ricinus* sensu stricto (Acari: Ixodidae). *Infection, Genetics and Evolution* **8**(2) (2008), 152–158. doi: 10.1016/j.meegid.2007.11.007.

55. D. Porretta et al.: The integration of multiple independent data reveals an unusual response to Pleistocene climatic changes in the hard tick *Ixodes ricinus*. *Molecular Ecology* **22**(6) (Feb. 2013), 1666–1682. doi: 10.1111/mec.12203.

56. G. Carpi et al.: Mitogenomes reveal diversity of the European Lyme borreliosis vector *Ixodes ricinus* in Italy. *Molecular Phylogenetics and Evolution* **101** (2016), 194–202. doi: 10.1016/j.ympev.2016.05.009.

57. R. E. Dinnis et al.: Multilocus sequence typing using mitochondrial genes (mtMLST) reveals geographic population structure of *Ixodes ricinus* ticks. *Ticks and Tick-borne Diseases* **5**(2) (2014), 152–160. doi: 10.1016/j.ttbdis.2013.10.001.

58. K. H. Røed, K. S. Kvie, G. Hasle, L. Gilbert, and H. P. Leinaas: Phylogenetic Lineages and Postglacial Dispersal Dynamics Characterize the Genetic Structure of the Tick, *Ixodes ricinus*, in Northwest Europe. *PLOS ONE* **11**(12) (Dec. 2016). Ed. by U. G. E. Munderloh, e0167450. doi: 10.1371/journal.pone.0167450.

59. *Fauna of the USSR Arachnida. IV, 2. Ixodid Ticks*. Zoological Institute of the Academy of Science USSR, Moscow, Leningrad, 1950.

60. F. Dantas-Torres and D. Otranto: Seasonal dynamics of *Ixodes ricinus* on ground level and higher vegetation in a preserved wooded area in southern Europe. *Veterinary Parasitology* **192**(1) (2013), 253–258. doi: 10.1016/j.vetpar.2012.09.034.

61. P. J. A. Pugh, P. E. King, and M. R. Fordy: The spiracle of *Ixodes ricinus* (L.) (Ixodidae: Metastigmata: Acarina): a passive diffusion barrier for water vapour. *Zoological Journal of the Linnean Society* **93**(2) (June 1988), 113–131. doi: 10.1111/j.1096-3642.1988.tb01530.x.

62. G. T. Baker: Spiracular Plate of Nymphal and Adult Hard Ticks (Acarina: Ixodidae): Morphology and Cuticular Ultrastructure. *Invertebrate Biology* **116**(4) (1997), 341–347.

63. J. Chmelař et al.: Sialomes and Mialomes: A Systems-Biology View of Tick Tissues and Tick–Host Interactions. *Trends in Parasitology* **32**(3) (2016). Special Issue: Vectors, 242–254. doi: 10.1016/j.pt.2015.10.002.

64. D. Richter, A. Matuschka Franz-Rainer ans Spielman, and L. Mahadevan: How ticks get under your skin: insertion mechanics of the feeding apparatus of *Ixodes ricinus* ticks. *Proceedings of the Royal Society B: Biological Sciences* **280**(1773) (2013), 20131758. doi: 10.1098/rspb.2013.1758.

65. D. Kemp, B. Stone, and K. Binnington: Chapter 4 - Tick Attachment and Feeding: Role of the Mouthparts, Feeding Apparatus, Salivary Gland Secretions and the Host Response. *Physiology of Ticks*. Ed. by F. D. Obenchain and R. Galun. Pergamon, 1982, 119–168. doi: 10.1016/B978-0-08-024937-7.50009-3.

66. J. Suppan, B. Engel, M. Marchetti-Deschmann, and S. Nürnberger: Tick attachment cement - reviewing the mysteries of a biological skin plug system. *Biological Reviews* **93**(2) (Nov. 2017), 1056–1076. doi: 10.1111/brv.12384.

67. T. K. Kim et al.: *Ixodes scapularis* Tick Saliva Proteins Sequentially Secreted Every 24 h during Blood Feeding. *PLOS Neglected Tropical Diseases* **10**(1) (Jan. 2016). Ed. by R. R. Dinglasan, e0004323. doi: 10.1371/journal.pntd.0004323.

68. S. Pantalacci and M. Sémon: Transcriptomics of developing embryos and organs: A raising tool for evo-devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **324**(4) (Nov. 2014), 363–371. doi: 10.1002/jez.b.22595.

69. J. Romiguier et al.: Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**(7526) (Aug. 2014), 261–263. doi: 10.1038/nature13685.

70. B. Misof et al.: Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**(6210) (Nov. 2014), 763–767. doi: 10.1126/science.1257570.

71. V. Cahais et al.: Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. *Molecular Ecology Resources* **12**(5) (Apr. 2012), 834–845. doi: 10.1111/j.1755-0998.2012.03148.x.

72. W. J. Cramaro et al.: Integration of *Ixodes ricinus* genome sequencing with transcriptome and proteome annotation of the naïve midgut. *BMC Genomics* **16**(1) (Oct. 2015). doi: 10.1186/s12864-015-1981-7.

73. W. J. Cramaro, O. E. Hunewald, L. Bell-Sakyi, and C. P. Muller: Genome scaffolding and annotation for the pathogen vector *Ixodes ricinus* by ultra-long single molecule sequencing. *Parasites & Vectors* **10**(1) (Feb. 2017). doi: 10.1186/s13071-017-2008-9.

74. M. Gulia-Nuss et al.: Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. *Nature Communications* **7** (Feb. 2016), 10507. doi: 10.1038/ncomms10507.

75. M. G. Grabherr et al.: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**(7) (May 2011), 644–652. doi: 10.1038/nbt.1883.

76. B. J. Haas et al.: De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**(8) (July 2013), 1494–1512. doi: 10.1038/nprot.2013.084.

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

78. R. D. Finn et al.: Pfam: the protein families database. *Nucleic Acids Research* **42**(D1) (Nov. 2013), D222–D230. doi: 10.1093/nar/gkt1223.

79. T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen: SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**(10) (Oct. 2011), 785–786. doi: 10.1038/nmeth.1701.

80. A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer: Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology* **305**(3) (2001), 567–580. doi: 10.1006/jmbi.2000.4315.

81. M. Ashburner et al.: Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**(1) (May 2000), 25–29. doi: 10.1038/75556.

82. A. Oshlack, M. D. Robinson, and M. D. Young: From RNA-seq reads to differential expression results. *Genome Biology* **11**(12) (2010), 220. doi: 10.1186/gb-2010-11-12-220.

83. H. Lopez-Maestre et al.: SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Research* (July 2016), gkw655. doi: 10.1093/nar/gkw655.

84. E. V. Koonin: Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics* **39**(1) (Dec. 2005), 309–338. doi: 10.1146/annurev.genet.39.073003.114725.

85. T. Gabaldón: Large-scale assignment of orthology: back to phylogenetics? *Genome Biology* **9**(10) (2008), 235. doi: 10.1186/gb-2008-9-10-235.

86. N. Lartillot and H. Philippe: A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution* **21**(6) (June 2004), 1095–109. doi: 10.1093/molbev/msh112.

87. G. Haller: Vorläufige Bemerkungen über das Gehörorgan der Ixodiden. *Zoologischer Anzeiger* **4** (1881), 165–167.

88. G. H. F. Nuttall, W. F. Cooper, and L. E. Robinson: On the Structure of "Haller's Organ" in the Ixodoidea. *Parasitology* **1**(03) (Oct. 1908), 238. doi: 10.1017/s0031182000003486.

89. A. D. Lees and J. W. L. Beament: An egg-waxing organ in ticks. *Quarterly Journal of Microscopical Science* **89**(7) (1948), 291–331.

90. J. Roux and M. Robinson-Rechavi: Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Research* **21**(3) (Dec. 2010), 357–363. doi: 10.1101/gr.113803.110.

91. A. K. Gibson, Z. Smith, C. Fuqua, K. Clay, and J. K. Colbourne: Why so many unknown genes? Partitioning orphans from a representative transcriptome of the lone star tick *Amblyomma americanum*. *BMC Genomics* **14**(1) (2013), 135. doi: 10.1186/1471-2164-14-135.

92. G. C. Boere and D. A. Stroud: The flyway concept: what it is and what it isn't. *Waterbirds around the world* (2006), 40–47.

93. K. Bown et al.: Sympatric *Ixodes trianguliceps* and *Ixodes ricinus* Ticks Feeding on Field Voles (*Microtus agrestis*): Potential for Increased Risk of *Anaplasma phagocytophilum* in the United Kingdom? *Vector-Borne and Zoonotic Diseases* **6**(4) (Dec. 2006), 404–410. doi: 10.1089/vbz.2006.6.404.

94. R. S. Peters et al.: Evolutionary History of the Hymenoptera. *Current Biology* **27**(7) (2017), 1013–1018. doi: 10.1016/j.cub.2017.01.027.

95. T. Hunt et al.: A Comprehensive Phylogeny of Beetles Reveals the Evolutionary Origins of a Superradiation. *Science* **318**(5858) (Dec. 2007), 1913–1916. doi: 10.1126/science.1146954.

96. J. P. Van Zee et al.: Paralog analyses reveal gene duplication events and genes under positive selection in *Ixodes scapularis* and other ixodid ticks. *BMC Genomics* **17**(1) (Mar. 2016). doi: 10.1186/s12864-015-2350-2.

97. Z. Li et al.: Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proceedings of the National Academy of Sciences* (Apr. 2018), 201710791. doi: 10.1073/pnas.1710791115.

98. G. Blanc and K. H. Wolfe: Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *THE PLANT CELL ONLINE* **16**(7) (July 2004), 1667–1678. doi: 10.1105/tpc.021345.

2 Article I: Exploration of the *Ixodes ricinus* transcriptome

Contents

Forewords

This work has been peer-reviewed, accepted and published in the journal Parasites & Vectors journal. This work was first deposited on a preprint server for biology, namely bioRxiv with the doi:10.1101/244830.

This article have the following reference: N. P. Charrier et al.: Whole body transcriptomes and new insights into the biology of the tick *Ixodes ricinus*. *Parasites & Vectors* **11**(1) (June 2018). doi: 10.1186/s13071-018-2932-3

Charrier *et al. Parasites & Vectors*
https://doi.org/10.1186/s13071-018-2932-3

Parasites & Vectors

RESEARCH Open Access

Whole body transcriptomes and new insights into the biology of the tick *Ixodes ricinus*

N. Pierre Charrier[1*], Marjorie Couton[1], Maarten J. Voordouw[2], Olivier Rais[2], Axelle Durand-Hermouet[1], Caroline Hervet[1], Olivier Plantard[1] and Claude Rispe[1]

Abstract

Background: *Ixodes ricinus* is the most important vector of tick-borne diseases in Europe. A better knowledge of its genome and transcriptome is important for developing control strategies. Previous transcriptomic studies of *I. ricinus* have focused on gene expression during the blood meal in specific tissues. To obtain a broader picture of changes in gene expression during the blood meal, our study analysed the transcriptome at the level of the whole body for both nymphal and adult ticks. *Ixodes ricinus* ticks from a highly inbred colony at the University of Neuchâtel were used. We also analysed previously published RNAseq studies to compare the genetic variation between three wild strains and three laboratory strains, including the strain from Neuchâtel.

Results: RNA was extracted from whole tick bodies and the cDNA was sequenced, producing 162,872,698 paired-end reads. Our reference transcriptome contained 179,316 contigs, of which 31% were annotated using Trinotate. Gene expression was compared between ticks that differed by feeding status (unfed *vs* partially fed). We found that blood-feeding in nymphs and female adult ticks increased the expression of cuticle-associated genes. Using a set of 3866 single nucleotide polymorphisms to calculate the heterozygosity, we found that the wild tick populations of *I. ricinus* had much higher levels of heterozygosity than the three laboratory populations.

Conclusion: Using high throughput strand-oriented sequencing for whole ticks in different stages and feeding conditions, we obtained a *de novo* assembly that significantly increased the genomic resources available for *I. ricinus*. Our study illustrates the importance of analysing the transcriptome at the level of the whole body to gain additional insights into how gene expression changes over the life-cycle of an organism. Our comparison of several RNAseq datasets shows the power of transcriptomic data to accurately characterize genetic polymorphism and for comparing different populations or sources of sequencing material.

Keywords: Transcriptomics, RNA-seq, *Ixodes ricinus*, Expression profiling, Polymorphism

Background

Ticks are vectors of numerous pathogenic microorganisms (*Borrelia* spp., *Babesia* spp., tick-borne encephalitis virus, etc.) that cause infectious diseases to both humans and animals [1]. Ticks acquire tick-borne pathogens from infected vertebrate hosts and transmit them to other animals during the blood meal [2]. During blood-feeding, the tick salivary glands secrete a complex cocktail of molecules that allows them to inhibit the different components of the response of the vertebrate host including coagulation, inflammation and immunity [3, 4]. Tick-borne pathogens bind to these tick salivary gland proteins to evade host immunity and enhance their own transmission [5, 6]. Similarly, during acquisition, tick-borne pathogens interact with tick proteins that allow them to persist in the tick midgut [7]. For this reason, anti-tick vaccines have traditionally targeted tick proteins in the salivary glands or midgut in the hope of reducing the efficiency of tick feeding and pathogen transmission [8]. However, a broader knowledge of genes involved in other aspects of the tick life-cycle (e.g. growth and moulting) may lead to alternative control strategies.

* Correspondence: npcharrier@gmail.com
[1]BIOEPAR, INRA, Oniris, Université Bretagne Loire, 44307 Nantes, France
Full list of author information is available at the end of the article

Ixodes ricinus is one of the most abundant and widespread tick species in Europe where it transmits a number of tick-borne diseases including Lyme borreliosis and tick-borne encephalitis [9, 10]. Hard ticks of the genus *Ixodes* have three motile stages: larva, nymph and adult. The immature stages (larvae and nymphs) take a single blood meal, then moult to the next stage; adult females take a blood meal to produce eggs. There is currently much interest in studying the genome of *I. ricinus* and other tick species in the hope of developing vector control strategies. Recent advances in sequencing technology have made it possible to study large catalogues of gene transcripts (the transcriptome) in individual species. By comparing gene expression between different states (e.g. developmental stage, sex, environmental conditions, etc.), these studies can provide insight into gene function. These RNA-sequencing studies can also provide information on genetic variation within and among populations [11]. To date, several studies have investigated gene expression in *I. ricinus* [12–17]. Most of these studies have focused on gene expression during the blood meal in either the nymphal tick or the adult tick, which is when pathogen transmission occurs. The majority of these studies have investigated gene expression in the tick salivary glands and/or the tick midgut because these tissues are critical for pathogen transmission [12–14, 16]. Taken together, these studies have shown that there are thousands of transcripts that are differentially expressed with respect to the duration of the blood meal, the developmental stage (nymph *versus* adult), the specific tissue (salivary glands *versus* midgut), and other conditions [12–19]. Most of these studies have focused on gene expression during the blood meal in either the nymphal tick or the adult tick, which is when pathogen transmission occurs [14].

The purpose of the present study is to explore new transcriptomic data of *I. ricinus* to improve several aspects of the existing knowledge. First, we wanted to enrich the global catalogue of genes for *I. ricinus*. We used whole tick bodies, which is expected to provide a broader description of the transcriptome compared to the previously published tissue-restricted libraries. We used strand-oriented sequencing, which produces contigs in the direction of transcription for the majority of transcripts. This type of sequencing gives a higher accuracy in the process of gene identification, especially for genes without detectable homology. Secondly, our design included different developmental stages and both sexes in order to capture the highest possible transcriptional diversity. The third innovative aspect of our study was the exploration of high throughput transcriptomic sequencing to detect and compare polymorphism levels among different sources of ticks. Transcriptome sequencing can identify single nucleotide polymorphisms (SNPs) on thousands of coding genes (see [14] for an application for tick data). In the present study, we first studied polymorphism in a highly inbred laboratory strain of *I. ricinus*, which was expected to show low heterozygosity. We then compared the results from our study to the results from previously published RNA-Seq studies that used different sources of *I. ricinus* tick material: (i) wild ticks; (ii) F1 offspring obtained from a mating between two wild ticks; and (iii) a tick cell line. Specifically, we compared levels of polymorphism and heterozygosity between four different sources of *I. ricinus* tick material that were expected to differ with respect to their genetic variability. This information provides a large catalogue of polymorphic sites in expressed regions and provides an important database for future population genetic studies.

Methods

Origin of ticks

The *I. ricinus* ticks used in this study came from a laboratory colony reared at the University of Neuchâtel, in Neuchâtel, Switzerland. This colony was initiated in 1978 with a small number of wild ticks collected from a natural population near Neuchâtel and has been maintained as follows. Larval and nymphal ticks are fed on laboratory mice (*Mus musculus*) and adult ticks are fed on rabbits (*Oryctolagus cuniculus*). Completion of the life-cycle of *I. ricinus* (from eggs to eggs) in the laboratory takes about 1 year. Each year, there is at least one cycle of sexual reproduction with a mating population of 40 to 50 adult ticks. There has been no admixture between this tick colony and wild *I. ricinus* ticks for almost 40 years (~40 generations). The ticks from this colony are therefore expected to be pathogen-free and have reduced genetic diversity due to prolonged inbreeding.

Preparation of the biological samples

We obtained ticks in five different biological states (sample size in brackets): (i) unfed nymphs (n = 60); (ii) partially fed nymphs (n = 12); (iii) unfed adult females (n = 8); (iv) partially fed adult females (n = 4); and (v) unfed adult males (n = 12). Sample sizes for cDNA sequencing differed among biological states because the amount of mRNA per individual differed between stages (nymphs *vs* adults) and feeding conditions (unfed *vs* partially fed). Nymphs were fed on mice (*Mus musculus*) and removed after 24 h of attachment while adult females were fed on rabbits (*Oryctolagus cuniculus*) and removed after 48 h of attachment. These feeding durations for both stages (adult ticks feed longer than nymphs) correspond to phase 1 or the slow phase of engorgement following the two phases defined by Lees [20]. During this phase, the nymphs and adults both reach approximately 1/4 of their final engorgement size. Unfed and partially fed ticks were flash-frozen at -80 °C before RNA extractions.

To allow statistical comparisons of gene expression levels among the different conditions, three biological replicates were obtained for each of the five combinations of stage and feeding status (a total of 15 samples were prepared).

Total RNA extraction

Whole tick bodies were ground with a soft plastic pestle in Trizol (Invitrogen, Life Technologies, Carlsbard, CA, USA) on dry ice. RNA was purified as follows: after adding chloroform, the ground material was centrifuged, the aqueous phase was transferred into an RNase-free tube and was topped up with ethanol. RNA was extracted using a NucleoSpin RNA XS column (Macherey-Nagel, Düren, Germany), which included a DNAse treatment. A second DNase treatment (Machery-Nagel) in RNasin (Promega, Madison, USA) was performed to ensure complete degradation of any remaining genomic DNA. The absence of genomic DNA was confirmed by PCR tests that targeted the *18S* ribosomal RNA gene of *I. ricinus*.

Library preparation and sequencing

The quantity and quality of extracted RNA was evaluated with NanoDrop (Thermo Fisher Scientific, Waltham, USA), Qubit (Invitrogen, CA) and Experion machines (Bio-RAD Laboratories Inc., Hercules, USA). All samples had sufficient quantities, concentrations and qualities of RNA to proceed with library preparation. The library preparation kit was NEBNext® Ultra Directional RNA Library Prep Kit, NEB Art. No E7420. Poly-A selection with a magnetic isolation module was used to target mRNAs, followed by strand-specific cDNA synthesis with an insert size of 150–400 bp, PCR amplification and library purification. Individual tags used for the 15 samples allowed multiplex sequencing. Sequencing was done on one lane of an Illumina Hiseq 2500 machine (v4 chemistry).

Quality and assembly of reads

To produce the dataset, the raw paired-end reads (2 × 125 bp) were first cleaned. Adapters were clipped and low-quality regions were filtered using Trimmomatic (release 0.36) [21]; only reads with a minimum of 36 high quality-scored contiguous bases were kept. Summary statistics of the sequence quality were checked for each library by visualizing the FastQC report (release 0.11.5) [22].

Filtering out rRNA reads and *de novo* assembly

To filter out reads corresponding to rRNA gene expression, reads were mapped to a large contig encompassing the *18S*, *28S* and *5.8S* ribosomal genes. This contig was obtained by performing a preliminary *de novo* assembly of publicly available Illumina transcriptomic sequence data for *I. ricinus* as of June 2014 (see Additional file 1: Table S1 for the description of this contig). Reads were mapped to this contig using Bowtie 2 [23]. This contig was expected to be more effective for removing the rRNA reads than the published rRNA sequences of *I. ricinus* because the former contains complete or nearly complete sequences whereas the latter only contains partial sequences. All the reads that did not map to rRNA were assembled with Trinity (release 2014-07-17) [24], using the 'dUTP library preparation' option to take into account strand-oriented sequencing.

Assessment of transcriptome completeness

A common test used to assess the "coverage" (or information completeness) of a given sequence dataset is to analyse random samples of reads, and to create a saturation curve describing the relationship between different metrics (numbers of contigs of a specified length, number of matches to a known set of genes, etc.) and the read sample size. Complete datasets have saturation curves that plateau more quickly compared to incomplete datasets. To determine the coverage or completeness of our dataset, we randomly sampled 1, 2, 5, 10, 20, 50, 80, 100, 140 and 160 million reads from our cleaned libraries, as detailed below. Completeness was also estimated by determining the presence of homologs of conserved arthropod genes using the BUSCO approach [25]. BUSCO v1 uses a reference database of 2675 conserved arthropod genes (BUSCO genes) and searches for potential homologs in the database of interest by running BLAST [26] and HMMER [27]. Conserved BUSCO genes are assigned to four classes of genes: (i) missing; (ii) fragmented; (iii) duplicated; and (iv) complete. To determine if the open reading frames (ORFs) were correctly predicted, we checked the strands of the predicted genes within the contigs matching the BUSCOs. To produce the final assembly, we reduced the potential redundancy resulting from the presence of alternative transcripts in the contigs. We clustered similar sequences using cd-hit-est [28] with 98% of identity, retaining the longest transcript of each cluster. Identity parameters were chosen to cluster nearly identical sequences resulting from alternative splicing. We used relatively stringent parameters for clustering: the local alignment had to comprise more than 50% of the longest alignment and more than 80% of the shortest alignment. To assess the loss of information produced by the clustering, we checked read recruitment in our final set of contigs by mapping with Bowtie 2 [23].

Gene prediction and annotation

The prediction of coding sequences (> 100 amino acids) was performed using TransDecoder, which is part of the Trinity software [24]. The TransDecoder options were set to account for the strand orientation of the sequencing

(i.e. ORFs were searched only on the forward strand). Finally, annotations from comparison with public databases were used to filter multiple ORF predictions by transcripts (see below). Following the pipeline recommendation of Trinotate (release 2.0.2) [29], both contigs and predicted peptides were compared by blastx+ and blastp+ (release 2.2.29) [26] to releases of Swissprot and Uniref90 (available at https://data.broadinstitute.org/Trinity/__deprecated_trinotate_resources/Trinotate_v2.0_RESOURCES/ v2.0 RESOURCES/). Protein domains were identified using HMMER (release 3.0 from March 2010) [27] with PFAM-A [30], signal peptides with SignalP [31], and transmembrane domains with TMHMM [32]. We tagged ribosomal RNAs using RNAmmer [33]. All these layers of annotation were combined by Trinotate to assign gene ontology (GO) information to each contig. In the case of multiple ORF predictions for a contig, if one ORF was similar to a known protein while the others was not, only the former ORF was retained. In other cases (several ORFs with similarity to known proteins, or several ORFs with no similarity to known proteins), the different ORFs were retained. As we were particularly interested in cuticular proteins, all peptides that contained the chitin-binding domain (PF00379) were classified using the CutProtFam-Pred webserver [34].

Comparison of the completeness of our assembly relative to other assembled transcriptomes

In recent years, several research groups have produced RNAseq datasets for *I. ricinus*. For most of these projects, the *de novo* assemblies (or sets of predicted genes derived from these assemblies) have been published in the Transcriptomes Shotgun Assembly (TSA) division of GenBank. Using statistics provided by BUSCO [25], we compared the completeness of our final assembly to that of six different assemblies/gene sets, which were obtained from six different RNAseq projects (see Additional file 1: Table S2; TSA accessions: GADI01, GANP01, GBIH01, GCJO01, GEFM01 and GEGO01). In addition, to determine the relative contribution of each of the different datasets (TSA and or own CDS prediction) to the complete gene collection, we analysed the clustering of all CDSs using cd-hit-est with default parameters [28] -for GCJO01, which corresponded to contigs, we predicted CDSs by using Transdecoder [28].

Differential expression and GO enrichment

In addition to the reconstruction of transcript sequences, RNA sequencing also allows the user to quantify transcript expression by counting the number of sequenced reads that map to a given transcript. Paired reads for each library were pseudo-aligned on the Transcriptome de Bruijn Graph (T-DBG) using Kallisto [35]. We chose this method, based on a k-mer approach, because it is much faster while providing the same accuracy as the best mapping approaches [35]. This method produced raw counts and normalized count statistics (TPM, or transcripts per million reads) for each assembled contig. These counts allowed us to test for differential expression between conditions. To assess feeding-related changes in gene expression, we compared partially fed ticks *versus* unfed ticks. For this, we performed a comparison between two ensembles of libraries, respectively D/E/F/M/N/O and A/B/C/G/H/J/K/L. Differential expression analyses were performed with the R package *DESeq2* [36] using the raw counts from Kallisto, each library been taken as an independent replicate. To describe the relevant biological changes between conditions, we used predictions produced by gene ontology (GO) term annotation, which included: "molecular function", "biological process" and "cellular localization" [37]. GO terms were compared between transcripts that were not differentially expressed between conditions (unbiased transcripts) *versus* transcripts that were differentially expressed between conditions. We defined unbiased transcripts as contigs with no significant change in expression between conditions (fold change less than 2 and adjusted *P*-value higher than 0.05). Enrichment analysis was performed using the *elim* method using Kolmogorov-Smirnov tests developed and implemented by Alexa [38] in the R package *TopGO*. As suggested by this author, multiple testing was taken into consideration by using the false discovery rate (FDR) on the enrichment test *P*-values. The resulting GO enrichments were analyzed using the R package *GOprofiles* [39], which provides visualization tools. GO enrichment comparisons are by definition limited to contigs with assigned GOs, whereas many more contigs can show significant changes in expression between conditions. A substantial number of contigs had no assigned GOs but did have other annotations, such as domains identified through the PFAM analysis. We therefore used a text mining analysis of the PFAM domains to further compare changes in gene expression between the different conditions (unfed/fed, nymph/adult, male/female). This approach allowed us to distinguish the most common terms associated within a text. We therefore extracted the PFAM terms associated with contigs over-expressed in each of the conditions defined above (fed or unfed) and treated them as a single text for each category. To prevent over-representation of transcripts with many PFAM domains, only the first ten PFAM terms with the lowest e-value were retained for a transcript and over-expressed contigs were defined as contigs with a fold change larger than 4 and a significant *P*-value ($P < 0.05$) in the DESeq2 analysis. PFAM descriptions were edited to remove the less informative terms (e.g. "protein", "domain", "motif"). We then used an in-house R script to draw clouds of words with word sizes

proportional to their frequency in the text. This approach is complementary to the GO-enrichment tests, as it helps to visualize major shifts in expression among conditions.

Summary of results by peptide predictions and by contigs

An annotation file with tab-separated values was produced, providing all the information from the Trinotate report, raw counts from Kallisto, log fold changes and *P*-values for differential expression, and BUSCO information. This report contains one line per peptide prediction and was deposited with a DOI on the Zenodo platform (see the section "Availability of data and materials" below).

Polymorphisms

Polymorphism was surveyed in the reads produced through our project (corresponding to an inbred line, here after referred to as NEU) but also for data from five other RNAseq projects of *I. ricinus* publically available in GenBank. Three published datasets used wild tick populations from Sénart in France (SEN) [16], from the Czech Republic (CZ-W) [13], and a mixture of wild tick populations (LUX) that was provided by Charles River Laboratory [17]. The other two datasets were based on F1 full sibs from a cross between wild ticks from the Czech Republic (CZ-F1) [19], and a tick cell line (CL) deposited by the Broad Institute under BioProject accession numbers PRJNA238785-88. More details on those 5 datasets are given in Additional file 1: Table S3. Reads sequences are available at the NCBI Sequence Reads Archive (SRA) and organized by BioProject. After downloading reads from the SRA archives, the reads were cleaned using Trimmomatic (with the same parameters as above). As estimates of polymorphism and heterozygosity depend on sample size, we standardized the sample size for each dataset by randomly sampling 30 million reads from each SRA. The combined datasets were analysed to detect single nucleotide polymorphisms (SNPs). SNPs were predicted with a "direct from the reads" approach, using KisSplice (release 2.4.0) [40], a software that identifies variations by detecting "bubbles" in the De Bruijn graph. SNPs were mapped on our final set of contigs using Blat (version 36) [41]. A report assessing various parameters for each SNP (location, reliability, etc.) was provided by Kiss2refTranscriptome [42]. To minimize false positives, we retained only SNPs that respected the following criteria: (i) SNPs had to be covered by at least ten reads in each dataset; and (ii) SNPs needed to be uniquely mapped (e.g. mapping to a single component). Using this restricted set of SNPs, we used variant counts produced by KisSplice to calculate allele frequencies at each polymorphic site, for each of the 6 RNAseq datasets. We estimated heterozygosity (using the formula He = 2pq, where p and q are the frequency of each variant) for each SNP position, and for each dataset.

Results

Reads and assembly statistics

We obtained a total of 210,229,106 paired strand-oriented reads. After filtering out the poor-quality reads and orphan reads (13,611,318 reads) and after excluding the reads assigned to rRNA (33,745,090 reads), the final set contained 162,872,698 trimmed, good-quality reads (see Table 1). For the fifteen libraries, the mean number of reads was 10,858,180 reads per library (range 7,628,548–15,814,366). Trinity produced an initial assembly set of 427,491 contigs, of which half had a length between 200–300 bp. As these small contigs are expected to be mostly represent gene fragments or untranslated region (UTR) sequences, we discarded them and only considered contigs above 300 bp. After reduction of redundancy [28], the assembly contained a total of 192,050 contigs. For the 15 different tick libraries, 88% and 91% of the reads mapped back to contigs (see Table 1). This good recruitment rate suggests that (i) the Trinity-produced assembly managed to capture most of the information contained by the reads, and (ii) little information was lost after eliminating the smaller class of contigs (< 300 bp). As an internal test of completeness, we analyzed the numbers of contigs of different sizes that resulted from the assemblies of random subsets of reads of increasing sample size. There was no clear saturation when considering all contigs, as the number of contigs was still rising (with only a moderate decrease of the slope) for even the largest read sample sizes (see Additional file 1: Figure S1). However, when only considering contigs above a certain size, the number of contigs clearly tended to plateau; this plateau can already be seen for contig size > 300 bp, whereas the plateau was marked for contigs > 1000 bp. This saturation effect was also shown with the BUSCO approach where the numbers of conserved arthropod genes plateaued at a sample size of 100 million reads (see Additional file 1: Figure S2). Overall, this result suggests that our complete set of reads tended to saturate the information on the mid-size to large transcripts, indicating that the coverage of our *I. ricinus* transcriptome was good (when considering all together the 5 combinations of stage, sex, and feeding conditions in our study).

Taxonomic assignation

We used taxonomic information from the code name of the best hit on the Uniref90 proteins cluster. We found that 6.4% of assembled transcripts (n = 12,368) were assigned to fungi (considering a minimum of 50% of protein identity and an E-value lower than 10e-5). Fungal contamination of individual ticks has been observed in the *I. ricinus* colony at the University of Neuchâtel.

Table 1 Description of the 15 libraries. The 15 libraries refer to the 5 combinations of stage, sex, and feeding condition for *I. ricinus* ticks, each replicated 3 times. The different columns refer to: library identification code, stage, sex, feeding condition, number of cleaned quality reads (Reads), percentage of reads that mapped back to the final set of transcripts with Bowtie 2 (Recruitment) and number of counting events observed by Kallisto divided by the number of paired reads (Kallisto)

Library	Stage	Sex[a]	Condition	Reads	Recruitment (% reads)	Kallisto (% reads)
A	Nymph	Unknown	Unfed	13,839,882	88.1	86.5
B	Nymph	Unknown	Unfed	9,158,226	86.9	85.7
C	Nymph	Unknown	Unfed	13,233,880	88.4	86.5
D	Nymph	Unknown	Fed	12,665,880	89.5	87.8
E	Nymph	Unknown	Fed	8,230,200	91.4	89.9
F	Nymph	Unknown	Fed	7,762,182	89.5	88.0
G	Adult	Male	Unfed	8,191,900	91.5	89.7
H	Adult	Male	Unfed	9,360,492	90.7	89.0
I	Adult	Male	Unfed	11,228,194	91.6	90.7
J	Adult	Female	Unfed	11,604,308	91.6	89.6
K	Adult	Female	Unfed	15,814,366	91.2	88.8
L	Adult	Female	Unfed	10,277,418	90.9	88.7
M	Adult	Female	Fed	12,813,758	90.5	89.1
N	Adult	Female	Fed	11,063,464	94.7	94.5
O	Adult	Female	Fed	7,628,548	90.6	90.2
Total				162,872,698	90.4	88.9

[a]Sex is unknown for nymphs

To facilitate moulting, blood-engorged larvae are placed in tubes with moistened filter paper, which also facilitates the growth of opportunistic fungi, which are the most likely source of the contamination observed in the present study. We counted 887,767 reads that mapped to the fungi-like contigs, which represents only 1.2% of the total number of counting events by Kallisto, suggesting that contamination was modest. Statistics on the abundance of these fungi-like transcripts for each library are presented in Additional file 1: Figure S3, which showed that this contamination was restricted to unfed nymphs. The partially fed nymphs and adults contained blood (and therefore RNA) from mice and rabbits, respectively. After removing the 12,368 fungi-like contigs and 366 mammalian transcripts, the final assembly contained 179,316 contigs (Table 2). A taxonomic assignation based on the best blastx hit on Uniref90 was obtained for 56,773 contigs; of these, 37.5% and 33.6% were assigned to *Ixodes scapularis* and *I. ricinus*, respectively (Fig. 1). Another 17.2% of the contigs had matches to "Acari" (i.e. tick and mite species other than *I. scapularis* and *I. ricinus*) and 2.5% had matches to "Insecta" (n = 1433 contigs). The insect species with the most abundant hits were: pea aphid *Acyrthosiphon pisum* (n = 590 contigs), termite *Zootermopsis* (n = 145), clonal raider ant *Ooceraea biroi* (n = 69), Asian long-horned beetle *Anoplophora glabripennis* (n = 66), and kissing bug *Rhodnius prolixus* (n = 6267). As expected, the mean identity of the hits reflected phylogenetic distance so that identity was highest for hits corresponding to *Ixodes* species, and lower for distant taxonomic groups. However, this expected pattern was reversed at the finest taxonomic level (genus *Ixodes*), where the mean distance to *I. scapularis* hits was lower than the mean distance to *I. ricinus* hits. We explored the distribution of % identities at the amino acid level (see Additional file 1: Figure S4) for best hits to *I. ricinus* or *I. scapularis*. For both species, the highest peak of

Table 2 Statistics of assembly for the final set of contigs: total size of contigs, shortest and largest contigs, mean and median contig size and the N50 contig length are expressed in base pairs

Statistic	Assembly
Number of contigs	179,316
Total size of contigs (bp)	130,913,381
Shortest contig (bp)	301
Longest contig (bp)	20,233
Number of contigs > 0.5 kbp	79,235
Number of contigs > 1 kbp	29,911
Number of contigs > 10 kbp	19
Mean contig size (bp)	730
Median contig size (bp)	461
N50 contig size (bp)	875

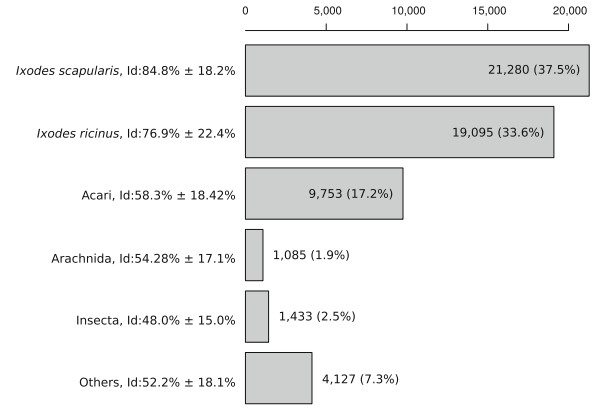

Fig. 1 Taxonomic assignment of contigs using Uniref90 best blast hits. The bars correspond to the number of contigs assigned to different taxonomic groups (the percentage of contigs assigned to each taxonomic group is shown in brackets). Best hits to "Acari", which include ticks and mites are hits to Acari species other than *I. ricinus* and *I. scapularis*. Best hits to "Arachnida", which also include spiders and scorpions, are hits to Arachnida species other than Acari. The mean percentage of identity at the amino acid level and the standard deviation are shown after the taxon name

genes had a very high percentage of identity (> 95%). This peak probably corresponds to the orthologous genes found in both *I. scapularis* and *I. ricinus*. However, the distribution of identities for *I. ricinus* shows a secondary peak of hits with a low identity (~45%). These low identity hits cannot correspond to the same gene, but probably represent distant paralogs of genes, or genes having a similar protein domain. The presence of this secondary peak decreases the mean identity of hits to *I. ricinus* and explains why the mean identity of our dataset is lower for *I. ricinus* than *I. scapularis*.

Annotation

Overall, 56,809 contigs (31.7% of the total) had a significant similarity with known proteins present in Uniref90 or Swissprot (Table 3). The percentage of contigs with a match strongly increased with contig size. For example, the percentage of contigs with a match was as high as 70.8% for contigs longer than 1 Kb. TransDecoder predicted 57,257 peptides, of which 35.6% were predicted as complete. In total, 13,308 GO terms were extracted from 26,702 contigs (14.9% of all contigs). Transmembrane domains and peptide signals were found for 8869 peptides and 3485 peptides, respectively. In addition, 22,082 contigs (12.3%) had a detected PFAM domain. Some contigs had only a GO match or only a PFAM assignation, so these data were complementary.

Completeness

Of the 2675 conserved arthropod genes in the BUSCO database, our assembly contained 2033 complete genes (completeness of 76%) and 250 partial genes (extended completeness of 85.3%). Previously published assemblies or collections of predicted genes (TSA archives) for *I. ricinus* all produced lower percentages of complete conserved genes (Fig. 2a). A comparison of the overlap between our new dataset and the combined TSA datasets found that our assembly contains 295 BUSCO genes (11%) not present in any of these TSA datasets, whereas the combined TSA datasets contained 231 BUSCO genes not present in our assembly (Fig. 2b). We obtained

Table 3 Summary of annotation statistics for the final set of contigs. For each class of contig length the following data are given: number of contigs, number of contigs with a significant hit on Swissprot or Uniref90, number of peptides predicted by TransDecoder and number of contigs for which gene ontology terms could be extracted

Class	No. of contigs	No. of annotated (%)[a]	No of peptides (%)[b]	TGO[c]
301-500	100,081	19,017 (19.00)	16,289 (16.28)	5247 (5.24)
501-1000	49,324	16,604 (33.66)	16,295 (33.04)	6753 (13.69)
1001-1500	13,214	7658 (57.95)	8406 (63.61)	4442 (33.62)
1501-2000	6446	4641 (72.00)	5368 (83.28)	3185 (49.41)
2001-2500	3712	3033 (81.71)	3555 (95.77)	2271 (61.18)
2501-3000	2289	1930 (84.32)	2387 (104.28)	1500 (65.53)
3001-4000	2394	2182 (91.14)	2737 (114.33)	1779 (74.31)
4001,5000	1025	956 (93.27)	1179 (115.02)	814 (79.41)
5001-7500	716	676 (94.41)	882 (123.18)	602 (84.08)
7501-10000	96	93 (96.88)	132 (137.50)	91 (94.79)
10001-100000	19	19 (100.00)	27 (142.11)	18 (94.74)
Total	179,316	56,809 (31.68)	57,257 (31.93)	26,702 (14.89)

[a]Number of contigs with a significant hit on Swissprot or Uniref90
[b]Number of peptides predicted by TransDecoder
[c]Number of contigs for which gene ontology terms could be extracted

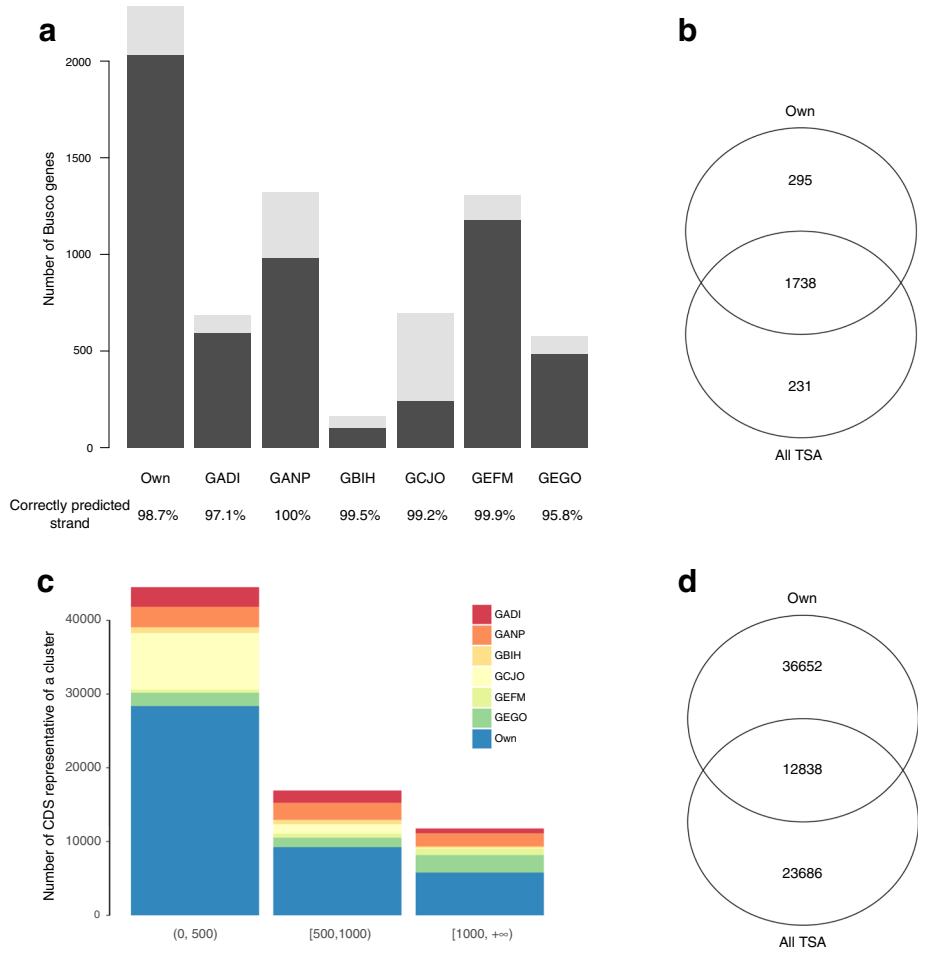

Fig. 2 Completeness of our *I. ricinus* transcriptome compared to Transcriptome Shotgun Assembly (TSA) datasets. Completeness was assessed through a comparison with the BUSCO database and by a direct comparison among CDSs from all datasets (cd-hit-est). **a** Number of BUSCO genes found in our assembly ("Own") and the six TSA projects. The percentage of complete BUSCO genes predicted on the correct strand are shown below each bar. **b** Venn diagram showing the overlap between the number of complete BUSCO genes found in our assembly and in the combined TSA assemblies. **c** Relative contribution of each dataset (different TSA and our assembly) to the combined clustering for three classes of size of contigs. **d** Venn diagram showing how many CDSs are unique and shared between our assembly and published TSA datasets

a similar percentage of BUSCO contigs with correctly predicted strands (98.7%) compared to the combined TSA datasets (mean of 98.6%, minimum of 95.8% for GEGO, maximum of 100% for GANP). We then performed a direct comparison between all the CDSs from our assembly and from the TSA datasets, with a clustering approach. We showed that our assembly permitted to identify 36,652 new CDSs which were not present in any of the published dataset (Fig. 2d). On the other hand, TSAs also contained specific CDSs not found in our study (n = 23,686). For CDSs which were represented both in our assembly and TSA datasets, we found that our assembly often produced the longest and most complete CDS. The relative contribution of each dataset and for different length classes is displayed in Fig. 2c: it shows that our assembly contributed to enriching the collection both for small and large CDSs.

Differential expression

When comparing the 200 most expressed transcripts, the "I" library of unfed adult males showed a strong deviation from all other libraries, including the two other replicate libraries of unfed adult males (libraries "G" and "H") (see Additional file 1: Figure S5). The "I" library also contained a particularly high percentage of rRNA reads (> 60%) suggesting that this sample was of lower quality. For this reason, we decided to exclude the "I" library from the analyses of differential gene expression. Comparing the levels of gene expression among libraries (for all transcripts), we found that libraries clustered by condition for unfed nymphs, fed nymphs, and unfed females (Additional file 1: Figure S6). Such clustering was not observed for the unfed males, whereas only two of the three samples clustered together for the fed females. The number of unbiased transcripts was 39,719. We

found that 11,322 transcripts (6.3%) were differentially up- or downregulated between the unfed condition and the partially fed condition (adjusted *P*-value < 0.05). Using the GO terms associated with DE transcripts between unfed and partially fed ticks, we found a significant enrichment (FDR < 0.05) for 4 molecular function (MF) terms (Additional file 1: Table S4), 41 biological process (BP) terms (Additional file 1: Table S5), and 6 cellular component (CC) terms (Additional file 1: Table S6). We represent top-significantly enriched GO terms (using up and down DE transcripts) between partially fed and unfed ticks without distinction of stage and sex (Fig. 3). The comparison between unfed and partially fed ticks found a significant enrichment of terms associated with cuticle production (Fig. 3); specifically, the expression of cuticle-associated genes was significantly higher in partially fed ticks than unfed ticks. Out of 39 transcripts containing a chitin-binding PFAM domain (PF00379), one contained two PF00379 domains and 22 contained a signal peptide. All these transcripts were classified as members of group 2 using the CutProFam-Pred webserver. Expression levels in each library of cuticle-related transcripts are shown in a heatmap (Fig. 4). Of the three most expressed cuticular transcripts, two transcripts shared a high identity with the proteins described by Andersen & Roepstorff [43]. The c252234_g1_i1 transcript in our study had 97.03% identity with the Ir-ACP10.9 protein (belonging to the CU109 cluster of Uniprot) [43]. This transcript was found to be highly over-expressed in partially fed ticks (588-fold change between unfed and partially fed ticks; *P*-value < 0.0001). Another transcript (c295844_g3_i4) that was highly over-expressed in partially fed adult females had 99.35% of identity with the Ir-ACP16.8 protein (belonging to the CU168 cluster of Uniprot) described in the same study [43]. The next and last two GO terms for Molecular Function showing significant enrichment for partially fed ticks were "extracellular matrix structural constituent" (which may also be related with constituents of the cuticle), and "metalloendopeptidase activity". As we discuss below, previous studies of expression several metalloproteases showed strong fold-changes in expression at different time-points of the blood meal [14]. Text mining of the identified domains in the differentially expressed transcripts provided additional insight on the function associated with each of the conditions (see Fig. 5). For the genes that were over-expressed in partially fed ticks compared to unfed ticks (Fig. 5a), we highlight the abundance of the following terms: reprolysin and metallopeptidase and Kunitz-BPTI (these three terms were often associated in the same transcripts) as well as tick-histamine-binding, immunoglobulin and chitin. For the genes over-expressed in unfed ticks (Fig. 5b), the most common terms were immunoglobulin, zinc-finger and leucine-rich-repeat. Actually, some of the most highly DE genes in unfed ticks had several IgC domains as is the case for a homolog of the gene Turtle. However, this trend was not caused by just one or a few transcripts, but by several genes with a similar structure (probably a gene family).

Polymorphism

Starting from the 368,367 SNPs initially detected by KisSplice using the default parameters, we applied more stringent criteria to increase the specificity: only 8955 SNPs were supported by at least 10 reads in each strain.

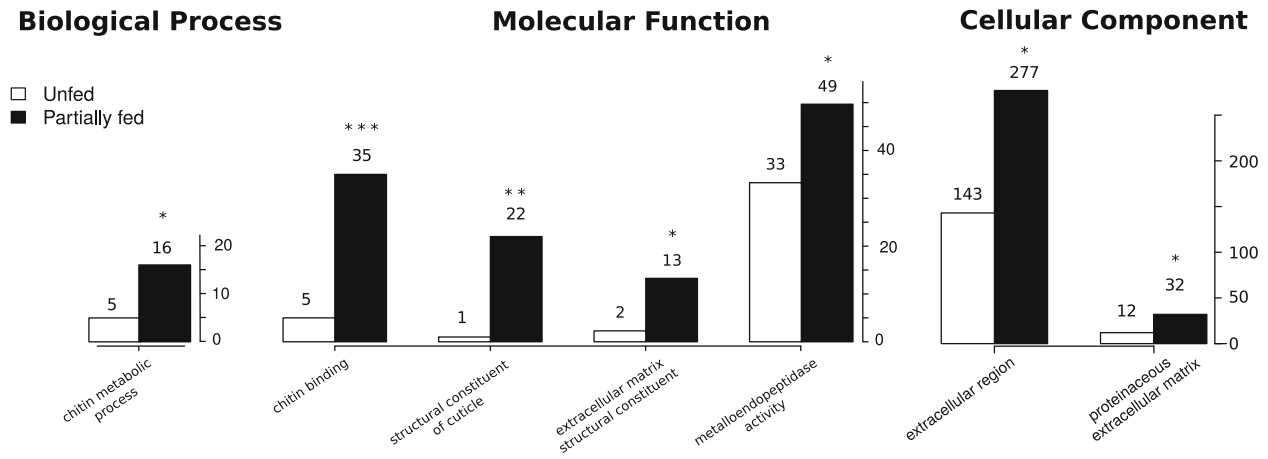

Fig. 3 Abundance of significantly enriched GO terms in partially fed ticks. For comparison, the abundance of the same terms for differentially expressed transcripts in both conditions (unfed and partially fed ticks) are also represented. GO enrichment for the comparison between unfed and partially fed ticks. The X-axis shows the top enriched GO terms for the three categories: Biological Process, Molecular Function and Cellular Component. The Y-axis shows the number of GO terms found in differentially expressed transcripts for unfed and partially fed ticks (see Additional file 1: Tables S4-S9 for detailed GO results). Significance of the Kolmogorov-Smirnoff enrichment test corrected by Benjamini & Hochberg: *$P < 0.05$, ** $P < 0.0005$, *** $P < 0.000005$

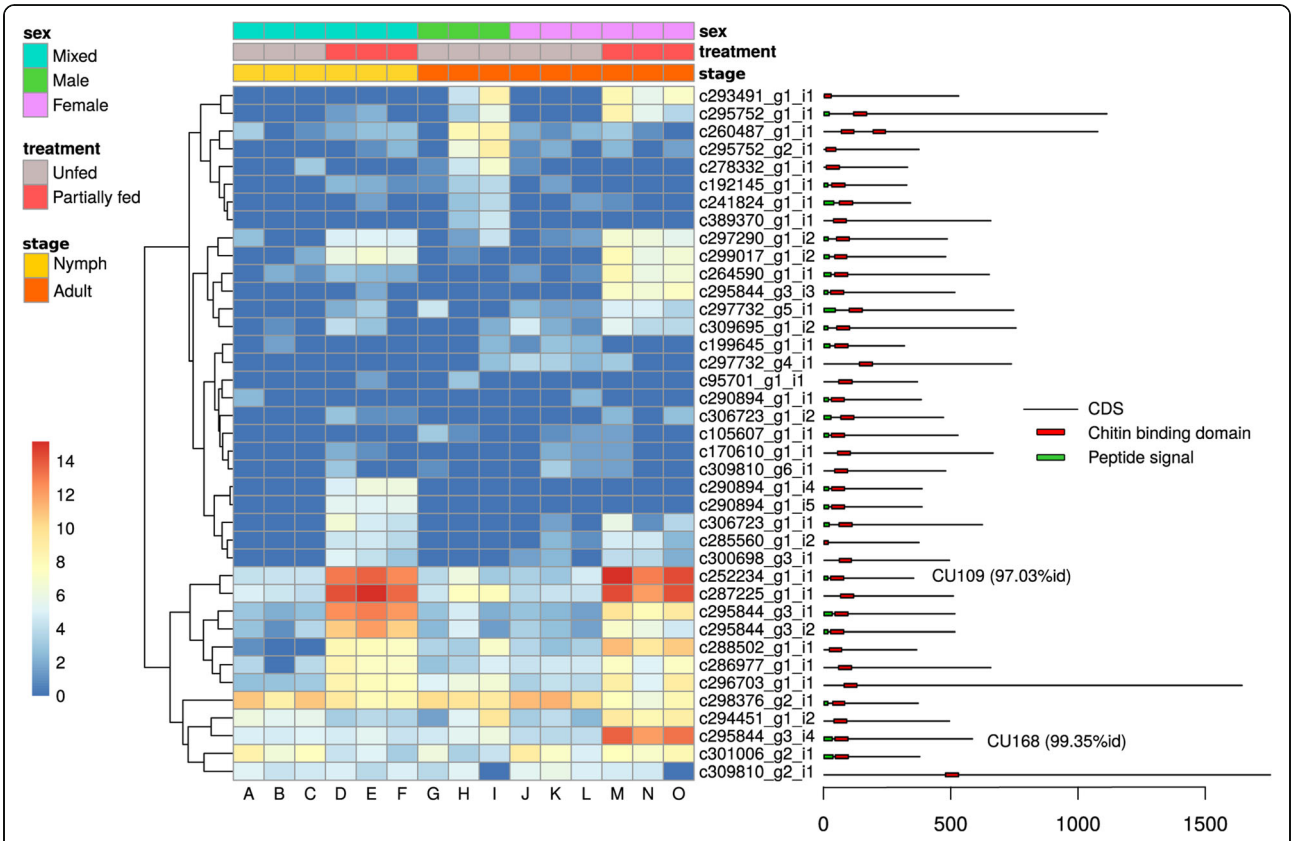

Fig. 4 Condition-dependent expression of genes that contain a chitin-binding domain (PFAM domain PF00379). The first three rows above the heat map indicate the sex, seeding treatment and stage condition of the *I. ricinus* ticks in each of the 15 libraries (columns A to O), according to a color legend detailed on the left. Each row of the heat map corresponds to a different gene (contig). Colors of the cells represent the level of expression (log("raw-count"+1)) of the gene in each library (counted by Kallisto). The legend on the right shows the position of the signal peptide domain and of the chitin-binding domain on the coding sequence of the transcript (CDS). Best-hit UNIPROT names are indicated for genes with more than 90% protein identity

We further filtered out all SNPs not assigned unambiguously to a single component of the T-DBG, resulting in a final set of 3866 SNPs. For this small but very robust set of polymorphic sites, the minor allele frequency (MAF) and heterozygosity were computed. A factorial component analysis (FCA) showed that the three wild strains of *I. ricinus* (SEN, CZ-W and LUX) were grouped very closely together (Fig. 6). These three wild tick strains had very similar levels of heterozygosity across loci on the three plane representations (Fig. 6b: plane of axis 1 *vs* 2, Fig. 6c: plane of axis 1 *vs* 3, and Fig. 6d: plane of axis 2 *vs* 3 of the FCA). By contrast, the laboratory strains (CZ-F1, NEU and CL) were distant from the central cluster of wild strains in the plane of axis 1 *vs* 2 space (Fig. 6b) and were also distant from each other. These laboratory strains formed a separate group supported by the third axis (see plane space representation of corresponding to plane of axis 1 *vs* 3 in Fig. 6c and plane of axis 2 *vs* 3 in Fig. 6d). Densities of the loci's heterozygosities for each strain showed similar patterns (see Additional file 1: Figure S7). Again, the three wild strains showed similar distributions of heterozygosity whereas the laboratory strains (CZ-F1 and NEU) showed very similar profiles: a high proportion of sites with very low heterozygosity (fixed or nearly-fixed SNPs) and a low proportion of sites with intermediate levels of heterozygosity. The CL strain showed an even more striking increase in the proportion of fixed sites. All these results showed that the three laboratory strains have less heterozygosity, differ strongly from wild ticks, and differ strongly from each other.

Discussion

Several recent RNAseq studies on *I. ricinus* have focused on the transcriptomes of specific tissues such as the midgut, salivary glands, haemocytes and ovaries [13–19]. In contrast, we tried to obtain a broader picture of the transcriptome of *I. ricinus* by sequencing and assembling transcripts from whole ticks in different conditions defined by developmental stage, sex and feeding status. First, our *de novo* assembly showed high overall completeness and significantly enlarged the catalogue of

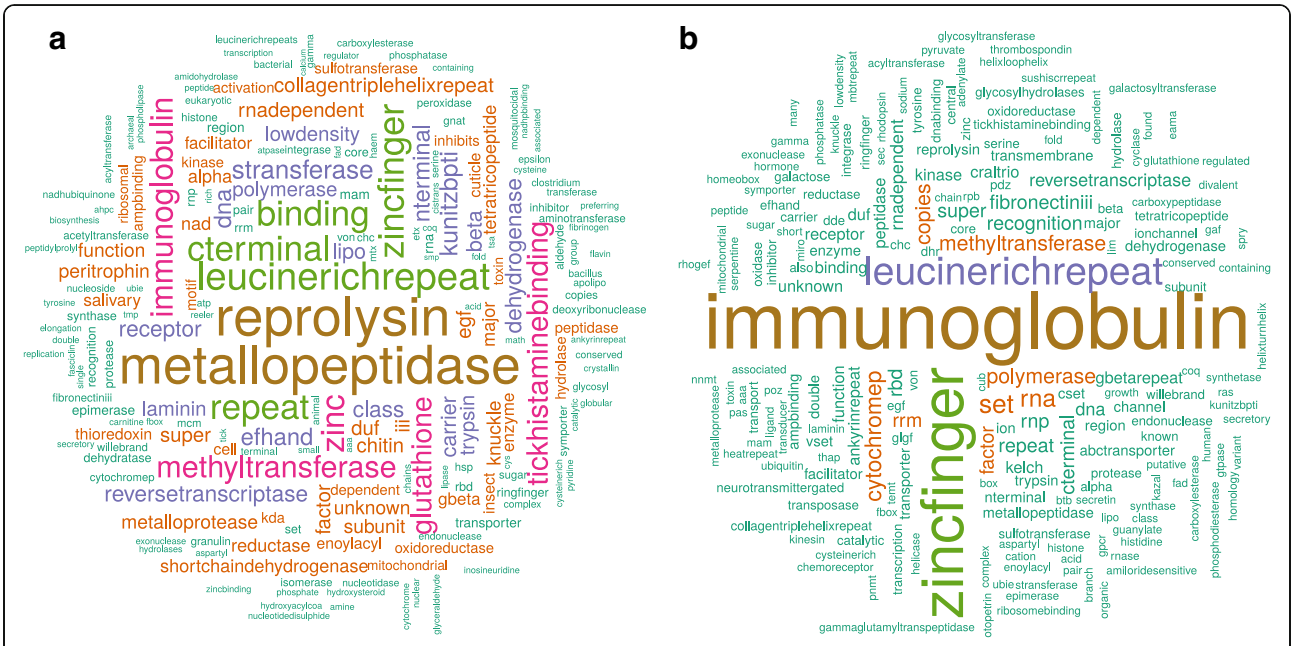

Fig. 5 Relative abundance of terms in the identified PFAM domains of transcripts over-expressed in unfed and partially fed ticks. Over-expressed transcripts were defined as transcripts with a fold-change > 4 and a *P*-value < 0.05 (DESeq2 analysis). Terms were extracted from the PFAM annotations retaining the first 10 domains (with the lowest e-value). Terms or characters that were not informative (e.g. "protein", "domain", "motif", "dots", etc.) were trimmed. Most abundant terms in transcripts upregulated in partially fed ticks (**a**) or upregulated in unfed ticks (**b**)

known coding sequences for *I. ricinus*. Secondly, we detected blood-feeding-induced changes in gene expression at the level of the whole body. Thirdly, our analysis of polymorphism in the transcriptome sequence data of our study and previously published studies allowed us to compare the levels of genetic diversity between outbred wild strains of *I. ricinus versus* inbred laboratory strains and tick cell lines.

Genomic resources for transcribed sequences

Using the BUSCO approach, our assembly showed high completeness with metrics that compared very favorably with previously published *de novo* assemblies (TSA contigs). All of these previous assemblies were based on specific tick tissues (midgut, salivary glands, haemocytes, etc.), which probably contain a smaller repertoire of transcripts which could explain their lower level of completeness. In contrast, we found that 231 BUSCO genes were absent from our assembly but present in the published TSA. By construction, BUSCO genes represent conserved genes found in most arthropods, so this approach does not bring *a priori* much insight for genes specific to ticks (which could be the most significant to understand tick biology). For a more global and more exact comparison, we also compared protein sequences (CDSs) of all published analyses and those predicted in our study. We found again that large numbers of CDSs were predicted only either in our study or in published datasets. Therefore, each analysis had a relatively large level of specificity, a finding thus similar to that of the BUSCO approach. This suggests that it is important to combine the different sources of data to obtain the most complete description of the transcriptome of the species. When analysing shared CDSs in the different analyses (by a clustering method), we found that our predicted CDSs often provided the longest sequence in each cluster, i.e. the most complete gene sequence. Therefore, our predicted collection of CDSs significantly enriched the extant knowledge for *I. ricinus* genes. To understand the differences in completeness between datasets (and their complementary aspects), we stress that tissue-specific transcriptomes (previous studies) will detect transcripts that may be rare at the level of the whole body (our study), whereas our whole-body transcriptome may have helped to identify transcripts present mostly or only in tissues not already covered by previous studies. A second factor is the strong time-dependence of gene expression during the blood meal as shown in previous studies on *I. ricinus* [14] and *I. scapularis* [44]. Our experimental design had only a single time point for nymphs and adult ticks, which might have resulted in missing some of the transcripts. We also suggest that the sequencing strategy used for previously published assemblies (which was often a combination of 454 and Illumina technologies), may have produced sub-optimal results. For example, the error-rich 454 sequences may have caused numerous indels in the final contigs, causing difficulties for ORF prediction [45]. Finally, we suggest that the choice of a highly

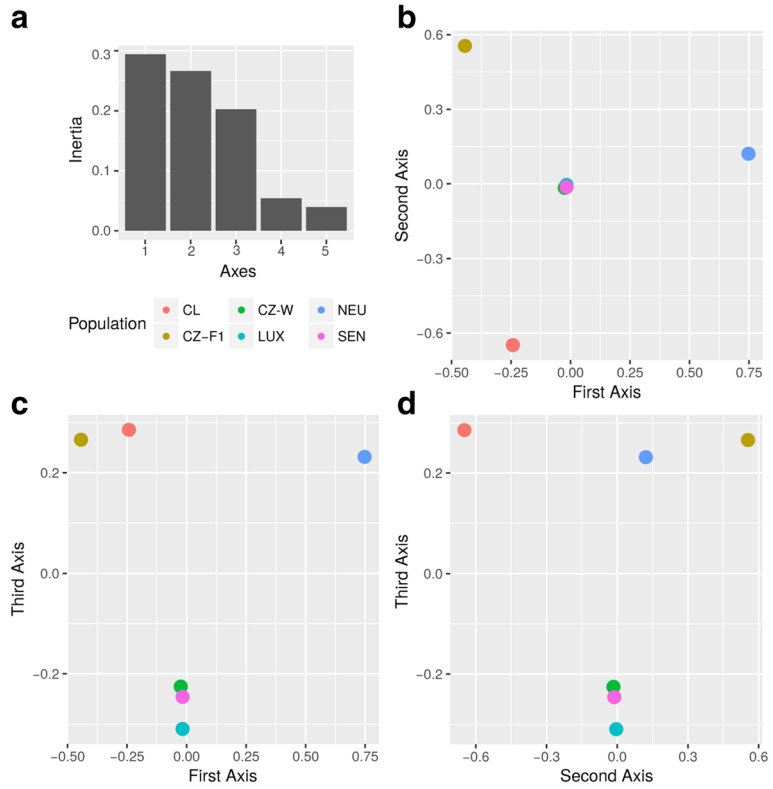

Fig. 6 Results of the factorial component analysis (FCA). The FCA explored how six sources of material (from different RNAseq datasets) of *I. ricinus* differ based on the heterozygosity of single nucleotide polymorphism (SNP) loci. The different sources are: wild ticks from the Czech Republic (CZ-W); wild ticks from a commercial strain (LUX); wild ticks from Sénart, France (SEN); a cell line (CL); a first generation of full sibs (CZ-F1); and a laboratory strain (NEU). The variables correspond to heterozygosity at the 3866 selected SNPs. **a** The eigenvalue (percentages of variation or inertia) explained by each eigenvector (axes). **b** The 6 sources on the plane formed by the first two axes. **c** The plane formed by axes 1 and 3. **d** The plane formed by axes 2 and 3

inbred line in our study must have facilitated the *de novo* assembly and produced more continuous and complete contigs (and then, CDSs). Indeed, polymorphism may create complexity in the resolution of the De Bruijn graph resulting in more fragmented results.

Overall, one third of the assembled transcripts showed similarity with known proteins and, in particular 70.8% of the contigs larger than 1 kb were annotated. The majority of the best hits (71.1%) were matches to *Ixodes* tick species, as expected given that a complete genome sequence is available for *I. scapularis* [46]. A smaller fraction had best matches to other groups, primarily arthropods (19.6%). These matches could be genes that are not found in other tick genomic resources due to their relative incompleteness or to the true absence of homologous genes in other tick species. A small number (n = 366) of contigs were assigned to genes of mice and rabbit (on which the nymphs and adults had fed), which indicates that host RNA was ingested during the blood meal. As expected, the mean identity at the amino acid level reflected phylogenetic distance, with one exception. We found a higher mean amino acid identity of our transcripts to *I. scapularis* than to *I. ricinus*. One explanation for this counter-intuitive result is that the available genomic resources are more complete for *I. scapularis* than *I. ricinus*. A complete genome has been published for *I. scapularis* [46], but only a genome survey is available for *I. ricinus* [17, 47]. For *I. ricinus*, most protein sequences in the databases are derived from *de novo* assembled transcriptomes that are still rather incomplete. Thus, one explanation for the relatively low mean identity of the hits to *I. ricinus* could be the absence of the same gene in the data banks (matches would often correspond to paralogous gene copies or shared protein domains).

Gene expression in response to blood-feeding

Differential expression (DE) analyses were focused on the comparison between two feeding conditions, the unfed and partially fed ticks. For this, we had to cope with two potential difficulties, (i) a relatively high rRNA contamination in four of the libraries, and (ii) an imperfect clustering of libraries with regard to sex, stage and feeding condition. These two aspects may be related, since

one of the libraries (I, for males) both had a high rRNA content and appeared as a clear outlier, which lead us to discard it from DE analyses. Replicates of a same feeding condition × stage × sex condition are expected to cluster together, and we observed indeed this pattern for unfed nymphs, fed nymphs and unfed adult females. But the two remaining male samples did not cluster together, and there was also an incomplete grouping for the fed females' libraries. Because it was not clear how to determine which grouping was the most legitimate, we preferred not to censor further libraries. Rather we expect that the inclusion of all "unfed" libraries and all the "fed" libraries together should buffer possible adverse effects of the imperfect clustering, but also lead to sort out the transcripts with the strongest effects in terms of expression change associated with feeding condition (this represents therefore a conservative approach).

DE analysis showed that 11,322 transcripts were either up- or downregulated during blood-feeding. The enrichment tests for unfed and partially fed conditions compared the GO terms between the differentially expressed transcripts and 39,719 unbiased transcripts. The three most important enriched GO terms were related to production of the tick cuticle. Over the course of the blood meal (3 to 8 days depending on the stage), hard ticks undergo dramatic changes in body size. When adult females reach repletion, their body weight increases by a factor of 100 [48]; which means that ticks must completely remodel their cuticle during repletion. Previous studies on adult female *I. ricinus* ticks observed an increase in cuticle thickness from 30 μm to 105 μm during the slow phase (first phase) of engorgement, followed by a decrease to 45 μm during the rapid phase (second phase) of engorgement [20, 43, 48, 49]. These studies support our finding that *I. ricinus* ticks increase their production of cuticular proteins during the blood meal, in order to greatly expand their body size. The next significant GO term (for Molecular Function) associated with upregulated transcripts in the fed condition was GO:0004222 (metalloendopeptidase activity) (Additional file 1: Table S4).

Several studies of the transcriptome of *I. ricinus* during the blood meal have shown the prominent role of several functional groups [12–16, 18, 19], including metalloproteases [12–14] and proteases inhibitors, such as the Kunitz-BPTI group [12, 14, 16]. Unexpectedly, with the exception of the single term found for metalloproteases, these groups of genes were not identified in the GO-enrichment tests made for the comparison between unfed and partially fed ticks. We therefore found only limited overlap between our work and previous studies in terms of GO functions of upregulated genes. We now discuss the possible origins of these differences among our work and previous studies. One reason could be the relatively limited fraction of genes with a GO assignation (15%), possibly resulting in insufficient statistical power in the present work. A second reason is the fact that the previous studies investigated specific tissues (e.g. salivary glands) whereas our study concerned whole ticks. If expression of secreted BP Kunitz proteins is mostly restricted to the tick salivary glands, then any change in gene expression would be diluted when considering the whole tick body where other metabolic processes dominate, such as those related to cuticle production. Our work therefore gives an indication of change of expression only at the very global level of the whole body, and cannot pretend to reach the same level of description of fine-scale changes between tissues, or between precise time-points of the blood meal. A third reason is that GO terms associated with BP Kunitz proteins, metalloproteases, reprolysin, etc., were not exclusively upregulated in the partially fed ticks. In fact, several of the genes that were considered as unbiased or genes that were strongly upregulated in unfed ticks had the same GO terms. The same observation applies to other terms that have been associated with blood-feeding in previous studies, or to other terms (e.g. zinc finger). The text mining analysis of the PFAM domains also found a high frequency of terms like reprolysin and Kunitz domains in the upregulated transcripts of the partially fed ticks. This result is in agreement with previous publications [12–16, 18, 19]. However, we note that these domains are also common in the upregulated transcripts in unfed ticks. This observation suggests that ticks contain multigenic families that share these common domains and that different genes in these families are expressed at different points in the tick life-cycle. This observation also suggests a subtle transcriptional landscape, where shifts in gene expression should be studied at the level of the gene rather than at the level of blocks of genes sharing the same GO or domain assignation.

Genetic diversity

We therefore established a catalogue of robust SNPs which could be useful for future population genetics analyses, as it could be used to design a genotyping study of multiple individuals, allowing a shift in scale compared to traditional population genetics approaches based on small numbers of markers. To our knowledge, only one similar analysis with RNAseq data has been conducted to date for *I. ricinus* [14], but the numbers of identified SNPs between this study and our work are difficult to compare given the different strategies (*de novo* SNP identification from the reads *versus* mapping to a set of CDSs) and because the reference sets used for localization of the SNPs are not the same. We however stress that the *de novo* (direct from the reads) approach for RNAseq datasets has been successfully applied in recent studies for different organisms [50]. Even though

the different datasets analysed in this work comprise pools of individuals or materials of very different origins, our approach illustrates the power of these analyses to determine the genetic characteristics of the different materials. Our study of single nucleotide polymorphisms found indeed substantial differences in heterozygosity between the different sources of *I. ricinus* used in RNA-seq studies. The factorial component analysis indicated that three datasets corresponding to wild ticks (SEN, CZ-W and LUX) clustered together, whereas each of the three laboratory strains (NEU, CZ-F1 and CL) differed markedly from that cluster and from each other. The clustering of the wild tick populations indicates that they have similar levels of heterozygosity (and should not be interpreted as the absence of genetic diversity between the three wild tick populations). In contrast, the laboratory lines show reduced levels of heterozygosity and frequent allele fixation at each of the SNP sites (Additional file 1: Figure S7). Reduction in heterozygosity was expected in the CZ-F1 strain because it was derived from a single mating of two wild ticks. Reduced heterozygosity was also expected in the NEU strain because it was derived from a laboratory colony that has a long history of inbreeding and small effective population size. The material from the tick cell line (CL) had an even more extreme reduction of heterozygosity, which is a typical feature of cell lines [51]. The laboratory strains differ from each other because of genetic drift, which results in the random fixation of alleles in each laboratory strain. Our study illustrates that sequencing the transcriptome of tick populations allows the genotyping of thousands of SNPs at hundreds of genes, which can greatly extend the traditional populations genetic approaches based on much fewer genetic markers [14, 42, 50].

Conclusions

Our *de novo* assembly showed high overall completeness and significantly enlarged the catalogue of known coding sequences for *I. ricinus*. Our study investigated the transcriptome of whole ticks that differed with respect to their developmental stage, sex and feeding condition. This approach allowed us to detect changes in gene expression at the level of the whole body instead of specific tissues. We found that blood-feeding induced a strong upregulation of transcripts associated with cuticle production. Finally, our analysis of polymorphism in the transcriptome sequence data of our study and previously published studies allowed us to identify 3866 robust SNPs from expressed-regions, and to compare the levels of genetic diversity between outbred wild strains of *I. ricinus versus* inbred laboratory strains and tick cell lines.

Additional file

Additional file 1: Table S1. Annotation of a rRNA containing contig from a preliminary *de novo* assembly. **Table S2.** Assembly statistics for Transcriptome Shotgun Assembly datasets (*de novo* assemblies corresponding to published papers, or still unpublished). **Table S3.** Details of published RNAseq studies for *I. ricinus*. **Table S4.** GO Enrichment for partially fed ticks (Molecular Function). **Table S5.** GO Enrichment for partially fed ticks (Biological Process). **Table S6.** GO Enrichment for partially fed ticks (Cellular Component). **Figure S1.** Reads sub-sampling and assembly statistics. **Figure S2.** Reads sub-sampling and assembly completeness. **Figure S3.** Quantification of Fungi-like reads in the different libraries. **Figure S4.** Distribution of the contig's identity with Acari and *Ixodes* species. **Figure S5.** Expression of the 200 most expressed genes among all libraries. **Figure S6.** Heatmap showing the hierarchical clustering of the 15 libraries based on expression counts. **Figure S7.** Distribution of estimated heterozygosity for six populations, using SNPs discovered by KisSplice. (PDF 380 kb)

Abbreviations

BP: Biological Process; CC: Cellular Component; DE: Differential Expression; FCA: Factorial Component Analysis; GO: Gene Ontology; MAF: Minor Allele Frequency; MF: Molecular Function; ORFs: Open Reading Frame; SNPs: Single Nucleotide Polymorphisms; SRA: Sequence Read Archive; T-DBG: Transcriptome De Bruijn Graph; TSA: Transcriptome Sequence Assembly

Acknowledgements

We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrénées (Bioinfo Genotoul) for providing support and computing resources. Part of the computations (multiple transcriptome assembly for subsamples of reads) were performed at the Bordeaux Bioinformatics Center (CbiB). We thank two anonymous reviewers and the editors for their constructive criticism and comments which helped improve this manuscript.

Availability of data and materials

Reads from sequencing the 15 libraries were deposited in the SRA section of NCBI under the BioProject ID: PRJNA395009. The contigs are available under accession GFVZ00000000 (NCBI, TSA). The annotation table was deposited on the Zenodo repository with a DOI (10.5281/zenodo.1137702), available on https://zenodo.org/record/1137702.

Authors' contributions

CR and OP conceived and designed the experiments. OR and MJV reared the ticks and provided the raw material for this study. AD and CH performed the molecular work (RNA extraction). NPC, MC and CR analyzed the data. NPC and CR wrote the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The use of vertebrate animals (mice and rabbits) to maintain the *I. ricinus* colony at the University of Neuchâtel was performed following the Swiss legislation on animal experimentation. The commission that is part of the 'Service de la Consommation et des Affaires Vétérinaires (SCAV)' of the canton of Vaud, Switzerland evaluated and approved the ethics of this part of the study. The SCAV of the canton of Neuchâtel, Switzerland issued the animal experimentation permit (NE05/2014)

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]BIOEPAR, INRA, Oniris, Université Bretagne Loire, 44307 Nantes, France. [2]Laboratoire d'Ecologie et Evolution des parasites, Institut de Biologie, Université de Neuchâtel, Rue Emile-Argand 11, CH-2000 Neuchâtel, Switzerland.

Received: 11 January 2018 Accepted: 4 June 2018
Published online: 26 June 2018

References

1. Jongejan F, Uilenberg G. The global importance of ticks. Parasitology. 2004; 129(Suppl.):S3–S14.
2. de la Fuente J, Antunes S, Bonnet S, Cabezas-Cruz A, Domingos AG, Estrada-Peña A, et al. Tick-pathogen interactions and vector competence: identification of molecular drivers for tick-borne diseases. Front Cell Infect Microbiol. 2017;7:114.
3. Kazimírová M, Štibrániová I. Tick salivary compounds: their role in modulation of host defences and pathogen transmission. Front Cell Infect Microbiol. 2013;3:43.
4. Brossard M, Wikel S. Tick immunobiology. Parasitology. 2004;129(S1):161–76.
5. Hovius JW, van Dam AP, Fikrig E. Tick-host-pathogen interactions in Lyme borreliosis. Trends Parasitol. 2007;23:434–8.
6. Ramamoorthi N, Narasimhan S, Pal U, Bao F, Yang XF, Fish D, et al. The Lyme disease agent exploits a tick protein to infect the mammalian host. Nature. 2005;436:573–7.
7. Kung F, Anguita J, Pal U. *Borrelia burgdorferi* and tick proteins supporting pathogen persistence in the vector. Future Microbiol. 2013;8:41–56.
8. Schuijt TJ, Hovius JW, van der Poll T, van Dam AP, Fikrig E. Lyme borreliosis vaccination: the facts, the challenge, the future. Trends Parasitol. 2011; 27:40–7.
9. Rizzoli A, Hauffe HC, Carpi G, Vourc'h GI, Neteler M, Rosa R. Lyme borreliosis in Europe. Eurosurveillance. 2011;16:19906.
10. Süss J. Tick-borne encephalitis 2010: epidemiology, risk areas, and virus strains in Europe and Asia - an overview. Ticks Tick Borne Dis. 2011;2:2–15.
11. Milano I, Babbucci M, Panitz F, Ogden R, Nielsen RO, Taylor MI, et al. Novel tools for conservation genomics: comparing two high-throughput approaches for SNP discovery in the transcriptome of the European hake. PloS One. 2011;6:e28008.
12. Chmelař J, Anderson JM, Mu J, Jochim RC, Valenzuela JG, Kopecký J. Insight into the sialome of the castor bean tick, *Ixodes ricinus*. BMC Genomics. 2008; 9:233.
13. Schwarz A, von Reumont BM, Erhart J, Chagas AC, Ribeiro JM, Kotsyfakis M. *De novo Ixodes ricinus* salivary gland transcriptome analysis using two next-generation sequencing methodologies. FASEB J. 2013;27:4745–56.
14. Kotsyfakis M, Schwarz A, Erhart J, Ribeiro JM. Tissue-and time-dependent transcription in *Ixodes ricinus* salivary glands and midguts when blood-feeding on the vertebrate host. Sci Rep. 2015;5:9103.
15. Kotsyfakis M, Kopáček P, Franta Z, Pedra JHF, Ribeiro JMC. Deep sequencing analysis of the *Ixodes ricinus* haemocytome. PLOS Negl Trop Dis. 2015;9: 1–22.
16. Liu XY, de la Fuente J, Cote M, Galindo RC, Moutailler S, Vayssier-Taussat M, et al. IrSPI, a tick serine protease inhibitor involved in tick feeding and *Bartonella henselae* infection. PLoS Negl Trop Dis. 2014;8:e2993.
17. Cramaro WJ, Revets D, Hunewald OE, Sinner R, Reye AL, Muller CP. Integration of *Ixodes ricinus* genome sequencing with transcriptome and proteome annotation of the naïve midgut. BMC Genomics. 2015;16:871.
18. Schwarz A, Tenzer S, Hackenberg M, Erhart J, Gerhold-Ay A, Mazur J, et al. A systems level analysis reveals transcriptomic and proteomic complexity in *Ixodes ricinus* midgut and salivary glands during early attachment and feeding. Mol Cell Prot. 2014;13:2725–35.
19. Perner J, Provazník J, Schrenková J, Urbanová V, Ribeiro JM, Kopáček P. RNA-seq analyses of the midgut from blood-and serum-fed *Ixodes ricinus* ticks. Sci Rep. 2016;6:36695.
20. Lees A. The role of cuticle growth in the feeding process of ticks. Proc Zool Soc London. 1952;121:759–72.
21. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
22. Andrews S. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 10 Oct 2015.
23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
24. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
25. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
26. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
27. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7: e1002195.
28. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
29. Trinotate: Transcriptome Functional Annotation and Analysis. https://trinotate.github.io/. Accessed 1 Feb 2016
30. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2013;42(D1):D222–30.
31. Petersen TN, Brunak S, Von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785–6.
32. Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305:567–80.
33. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35:3100–8.
34. Ioannidou ZS, Theodoropoulou MC, Papandreou NC, Willis JH, Hamodrakas SJ. CutProtFam-Pred: detection and classification of putative structural cuticular proteins from sequence alone, based on profile hidden Markov models. Insect Biochem Mol Biol. 2014;52:51–9.
35. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
36. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
37. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32(Database issue):D258–61.
38. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. R package version 2.18.0: 2010. http://master.bioconductor.org/packages/devel/bioc/citations/topGO/citation.html.
39. Sanchez A, Ocana J, Salicru M. goProfiles: an R Package for the Statistical Analysis of Functional Profiles. R package version 1.28.0; 2010.
40. Sacomoto GA, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot MF, et al. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. BMC Bioinformatics. 2012;13(Suppl. 6):S5.
41. Kent WJ. BLAT - the BLAST-like alignment tool. Genome Res. 2002;12:656–64.
42. Lopez-Maestre H, Brinza L, Marchet C, Kielbassa J, Bastien S, Boutigny M, et al. SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. Nucleic Acids Res. 2016;44:e148–3.
43. Andersen SO, Roepstorff P. The extensible alloscutal cuticle of the tick, *Ixodes ricinus*. Insect Biochem Mol Biol. 2005;35:1181–8.
44. Kim TK, Tirloni L, Pinto AFM, Moresco J, Yates JRIII, da Silva Vaz I Jr, et al. *Ixodes scapularis* tick saliva proteins sequentially secreted every 24 h during blood-feeding. PLOS Negl Trop Dis. 2016;10:1–35.
45. Francis WR, Christianson LM, Kiko R, Powers ML, Shaner NC, Haddock SH. A comparison across non-model animals suggests an optimal sequencing depth for *de novo* transcriptome assembly. BMC Genomics. 2013;14:167.
46. Gulia-Nuss M, Nuss AB, Meyer JM, Sonenshine DE, Roe RM, Waterhouse RM, et al. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nat Commun. 2016;7:10507.
47. Cramaro WJ, Hunewald OE, Sakyi-Lesley L, Muller CP. Genome scaffolding and annotation for the pathogen vector *Ixodes ricinus* by ultra-long single molecule sequencing. Parasit Vectors. 2017;10:71.
48. Flynn PC, Kaufman WR. Female ixodid ticks grow endocuticle during the rapid phase of engorgement. Exp Appli Acarol. 2011;53:167–78.
49. Dillinger S, Kesel A. Changes in the structure of the cuticle of *Ixodes ricinus* L., 1758 (Acari, Ixodidae) during feeding. Arthropod Struct Dev. 2002;31:95–101.
50. De Wit P, Pespeni MH, Palumbi SR. SNP genotyping and population genomics from expressed sequences - current advances and future possibilities. Mol Ecol. 2015;24:2310–23.
51. Tischfield JA. Loss of heterozygosity or: how I learned to stop worrying and love mitotic recombination. Am J Hum Genet. 1997;61:995–9.

Supplemental material of "Whole body transcriptomes and new insights into the biology of the tick *Ixodes ricinus*"

N. Pierre Charrier[1], Marjorie Couton[1], Maarten J. Voordouw[2], Olivier Rais[2], Axelle Durand-Hermouet[1], Caroline Hervet[1], Olivier Plantard[1] and Claude Rispe[1]

[1]BIOEPAR, INRA, Oniris, Université Bretagne Loire, 44307, Nantes, France
[2]Laboratoire d'Ecologie et Evolution des parasites, Institut de Biologie, Université de Neuchâtel, Rue Emile-Argand 11, CH-2000, Neuchâtel, Switzerland

May 29, 2018

List of Figures

List of Tables

Table S1: Annotation of a rRNA containing contig from a preliminary *de novo* assembly. Annotation of this contig with Rfam (release 12.1 2016-04-26) confirmed the presence of three rRNA units 18S (SSU: positions 3102-7008), 5.8S (positions 2244-2396) and 28S (LSU: positions 3102-7008).

ID	accession	start	end	bits score	E-value	strand
LSU_rRNA_eukarya	RF02543	3102	7008	2943.5	0	+
SSU_rRNA_eukarya	RF01960	1	1816	1773.3	3.7e-34	+
SSU_rRNA_bacteria	RF00177	1	1821	359.3	3.1e-13	+
SSU_rRNA_microsporidia	RF02542	1	1816	874.4	5e-266	+
5_8S_rRNA	RF00002	2244	2396	126.9	1.5e-32	+
SSU_rRNA_archaea	RF01959	1	1819	437.0	7.5e-14	+
LSU_rRNA_bacteria	RF02541	2991	6751	1056.8	0	+
LSU_rRNA_archaea	RF02540	3032	6757	1271.4	0	+

Table S2: Assembly statistics for Transcriptome Shotgun Assembly data sets (*de novo* assemblies corresponding to published papers, or still unpublished).

Name	GADI	GANP	GBIH	GCJO	GEFM	GEGO
BioProject	PRJNA177622	PRJNA217984	PRJNA183509	PRJNA270980	PRJNA311553	PRJNA312361
Date	July 2015	March 2015	September 2014	April 2015	February 2016	April 2016
Study	Schwarz, 2013[13]	Schwarz, 2014[18]; Kotsyfakis, 2015[14]	Kotsyfakis, 2015[15]	Cramaro, 2015[17]	Perner, 2016[19]	Unpublished
Number of CDS	8,685	16,002	2,854	(25,962)[1]	7,215	7,692
Number of contigs	8,685	16,002	2,854	59,924	7,215	7,692
total size of contigs	5,465,358	14,483,137	2,128,460	25,311,446	9,540,547	9,280,828
Shortest contig	201	201	201	200	201	201
Longest contig	5,352	16,869	6,909	9,464	17,475	19,044
Number of contigs > 500	4,343	10,145	1,599	12,771	5,933	5,231
Number of contigs > 1k	1,269	4,668	523	2,664	3,724	3,317
Number of contig > 10k	0	7	0	0	10	6
mean contig size	629	905	746	422	1,322	1,207
median contig size	501	645	549	333	1,026	846
N50 contig length	741	1,212	882	437	1,683	1,818

1 - GCJO consists in a collection of contigs, number in perenthesis is the number of CDSs predicted by TransDecoder.
References are numbered accordingly to the list of cited references in the main manuscript.

Table S3: Details of published RNAseq studies for *I. ricinus*. First column, data set acronym (Strain). Second column, type of material: wild ticks, F1 between wild ticks, Cell line, (Origin). Third column, NCBI BioProject accession number. Column 4, accession number (TSA). Column 5, references. Column 6, sequencing technology used. Column 7, total amount of bases sequenced in Gigabases. Column 8, tissues used (SG, salivary glands, H, haemocytome, MG, midgut). Column 9, stages (Ny, nymphs, Ad, adults). Column 10, condition (U, unfed, Pfed, for partially fed ,F, fully fed).

Strain	Origin	BioProject	TSA	Study	Technology	Gb.	Tissues	Stage	Condition
CZ-W	Wild	PRJNA183509	GBIH	Kotsyfakis, 2015 [15]	Illumina	29.9	SG, H	Ny, Ad	Pfed and F
		PRJNA177622	GADI	Schwarz, 2013 [13]	Illumina, 454	8.87	SG, H	Ny, Ad	Pfed and F
CZ-F1	F1	PRJNA312361	GEGO	Unpublished	Illumina	24	SG	Fe	N.A.
		PRJNA311553	GEFM	Perner, 2016 [19]	Illumina	56	MG	Fe	Pfed and F
LUX	Wild	PRJNA270980	GCJO	Cramaro, 2015[17]	Ion torrent	4.3	MG	Ad	U
CL	Cell line	PRJNA238785			Illumina	1.9	CL	-	-
		PRJNA238786			Illumina	1.5	CL	-	-
		PRJNA238787			Illumina	2.4	CL	-	-
		PRJNA238788			Illumina	2.7	CL	-	-
SEN	Wild	PRJNA237360		Liu, 2014 [16]	Illumina	17.7	SG	Ad	Pfed

References are numbered accordingly to the list of cited references in the main manuscript.

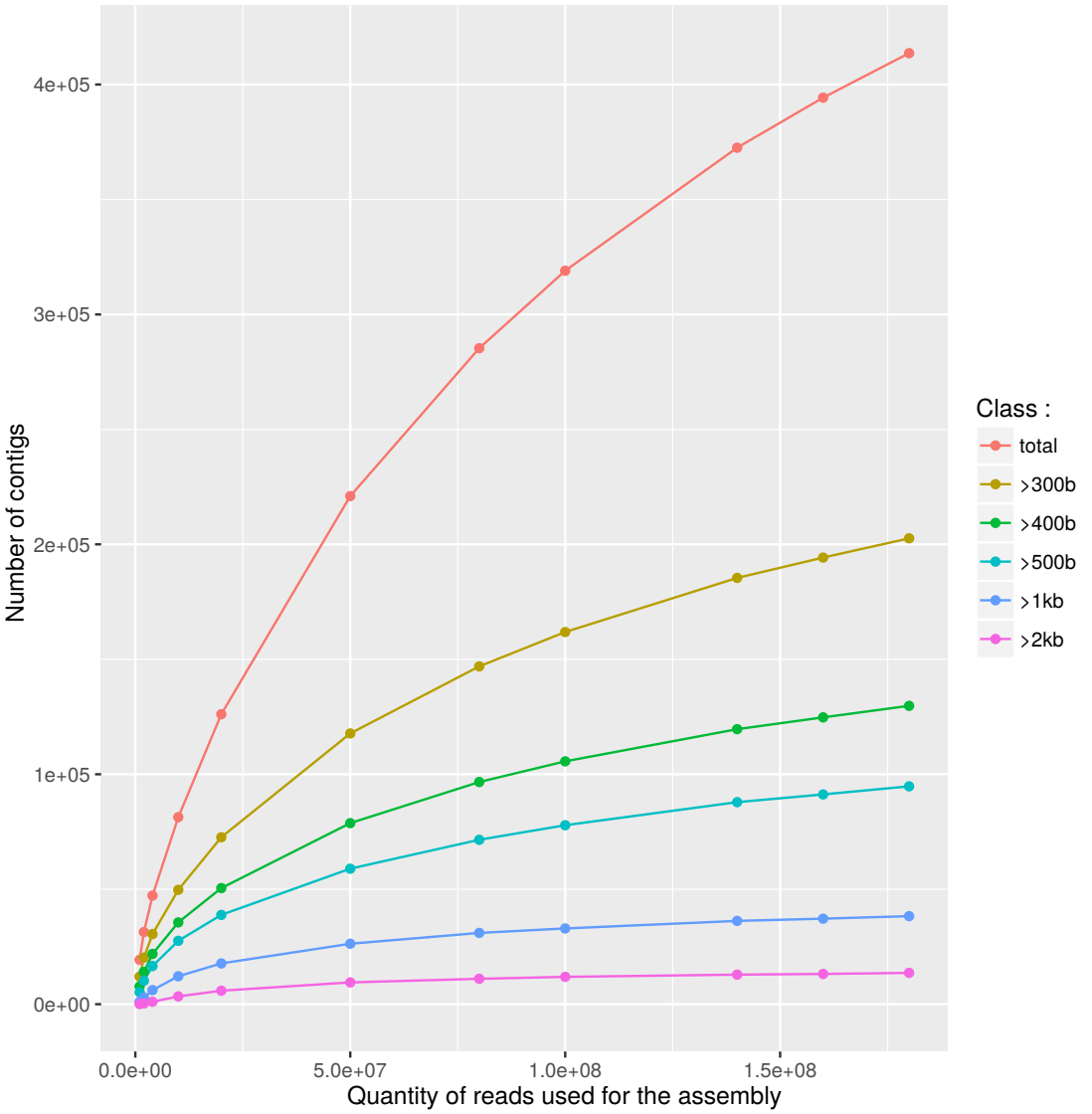

Figure S1: Reads sub-sampling and assembly statistics. Number of contigs obtained after a *de novo* assembly with `Trinity`, for different sample size (number of reads in abscissus). Numbers of transcripts: total (red), transcripts > 300bp (yellow), > 400bp (green), > 500bp (light blue), >1kb (marine blue), and >2kb (purple).

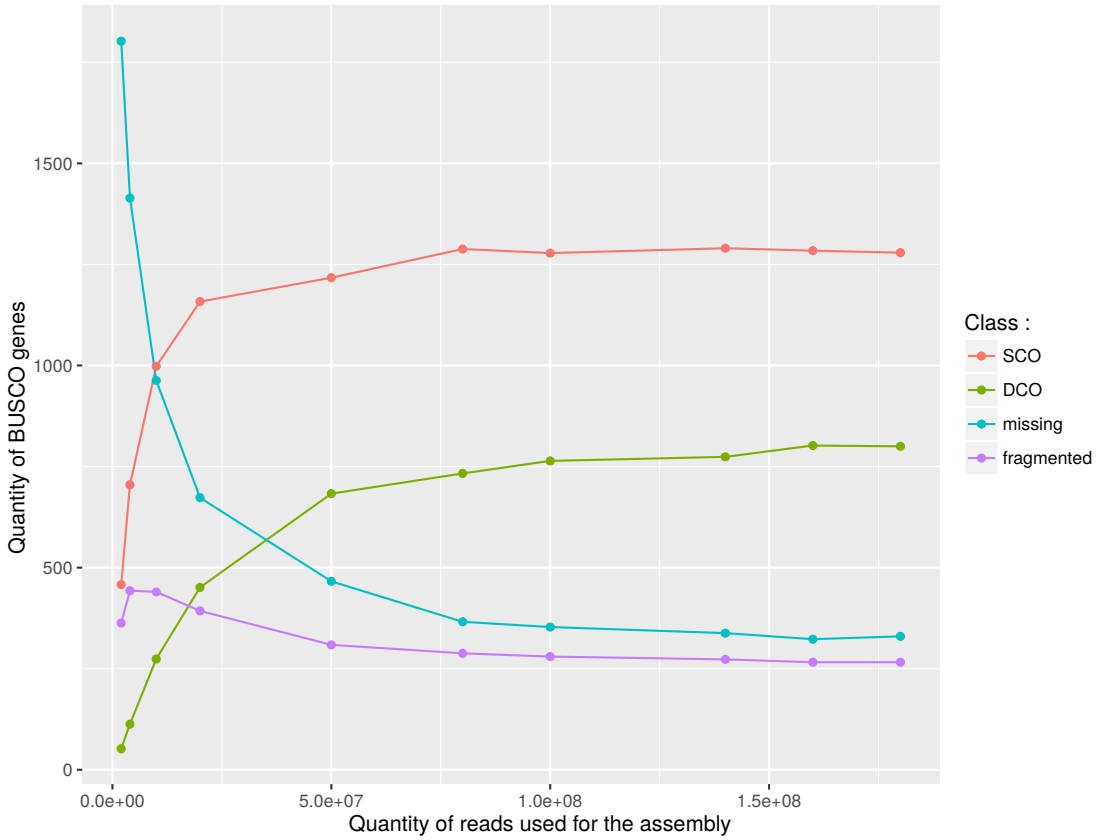

Figure S2: Reads sub-sampling and assembly completeness. Number of BUSCO genes (y-axis) found as a function of sample sizes of reads used for the assembly (number of reads sampled in abscissus). Abbreviations: SCO, Single Copy Orthoguous BUSCO genes; DCO, Duplicated Copy Orthologs; missing, number of absent BUSCO genes; fragmented: BUSCO genes found partially.

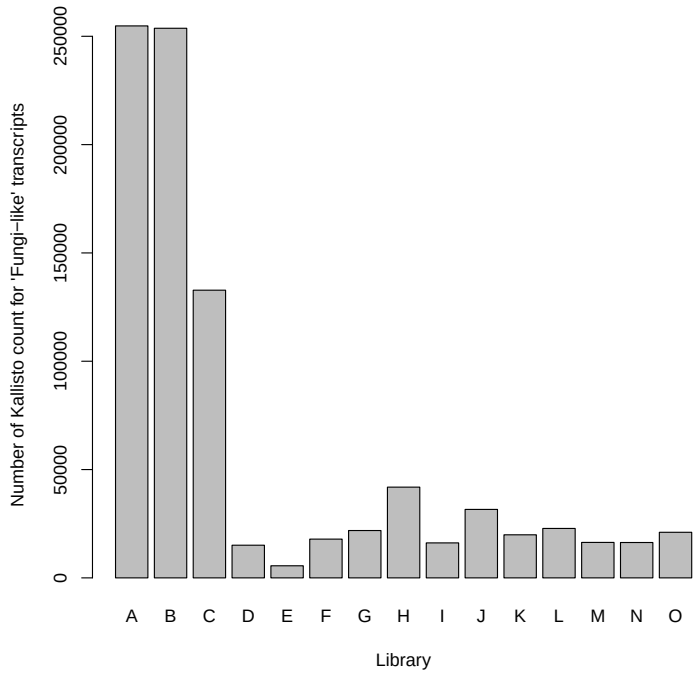

Figure S3: Quantification of Fungi-like reads in the different libraries. Total of expression counts (with Kallisto) for all contigs showing similarity with Fungi, based on Uniref90 annotation (first hit to Fungi, proteic identity $> 50\%$ and E-value lower than $10e-5$). The total of expression counts across all the libraries for Fungi-like transcripts was 887,767, for total of 72,392,431 `kallisto` counts (1.23%).

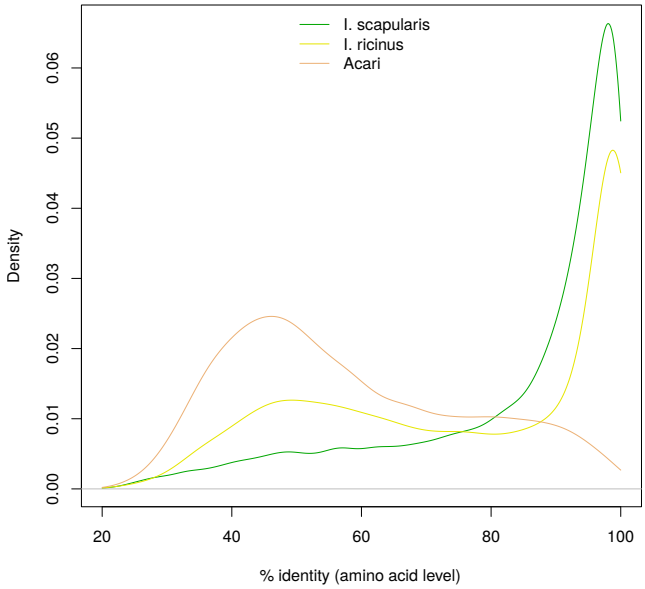

Figure S4: Distribution of the contig's identity with Acari and Ixodes species. Distribution of the best hit identity (at the amino acid level) on Uniref90 for two *Ixodes* species and the "Acari" taxon (other species of Acari).

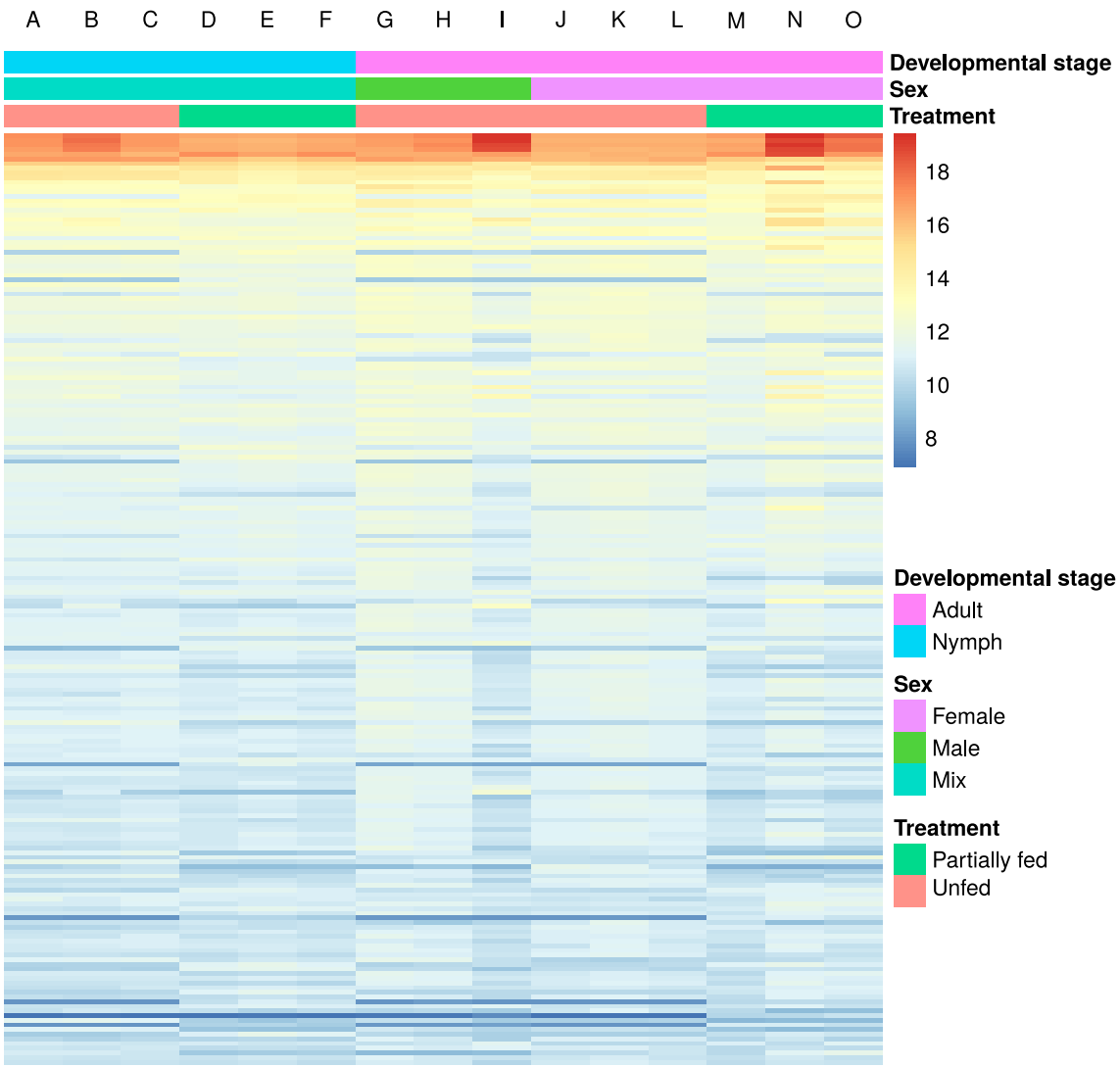

Figure S5: Expression of the 200 most expressed genes among all libraries represented by a color scale (red for highest expression, blue for lowest expression) and measured by a log2-transformed Kallisto count.

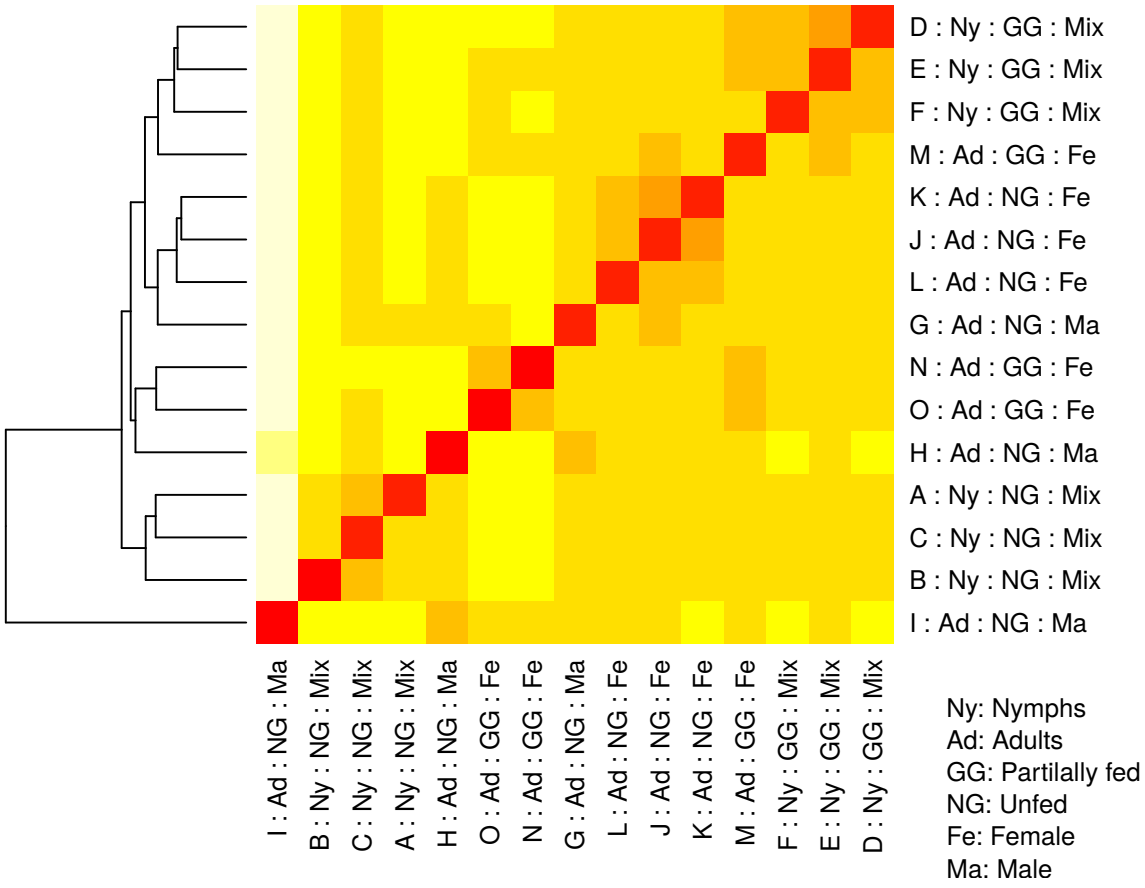

Figure S6: Heatmap showing the hierarchical clustering of the 15 libraries based on expression counts.

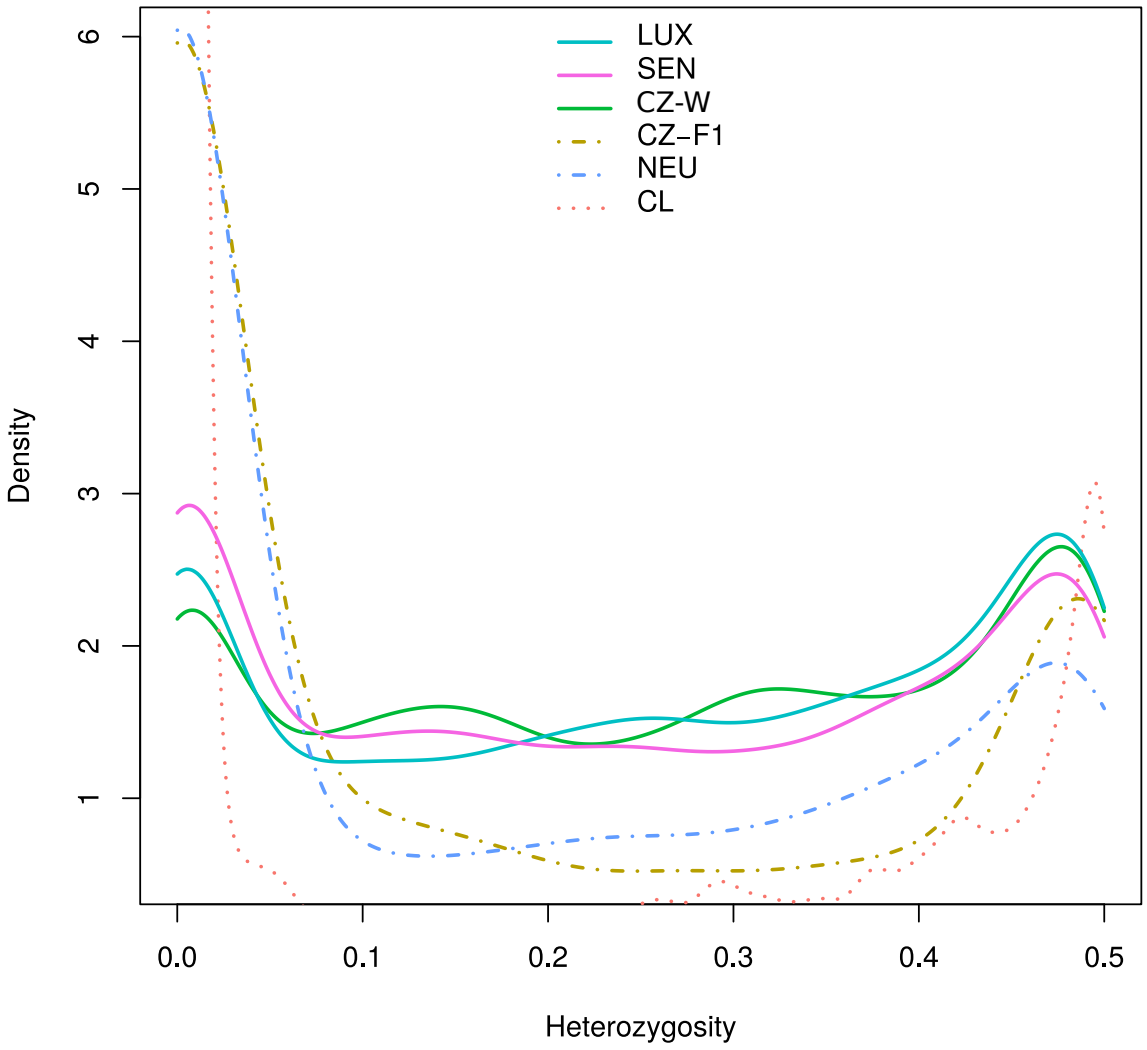

Figure S7: Distribution of estimated heterozygosity for six populations, using SNPs discovered by `KisSplice`. Heterozygosity was calculated on well-covered loci for each of the six data set (n=3,866), estimating variant frequency ($He = 2pq$) for each locus and each data set.

Table S4: GO Enrichment for partially fed ticks (Molecular Function). The elimKS column represents the significativity of the the Elim-Kolmogorov Smirnoff test implemented in the topGO R package, maintained by Adrian Alexa and Jorg Rahnenfuhrer. This tests for enriched GO terms in genes significantly over-expressed in partially fed ticks in comparison with unbiased genes.

GO.ID	Term	elimKS	adjustedFDR
GO:0008061	chitin binding	1.8e-10	9.0e-09
GO:0042302	structural constituent of cuticle	2.6e-06	6.5e-05
GO:0005201	extracellular matrix structural constituent	0.0011	0.018
GO:0004222	metalloendopeptidase activity	0.0021	0.026

Table S5: GO Enrichment for partially fed ticks (Biological Process). The elimKS column represents the significativity of the the Elim-Kolmogorov Smirnoff test implemented in the topGO R package, maintained by Adrian Alexa and Jorg Rahnenfuhrer. This tests for enriched GO terms in genes significantly over-expressed in partially fed ticks in comparison with unbiased genes.

GO.ID	Term	elimKS	adjustedFDR
GO:0006030	chitin metabolic process	1.1e-05	0.00055
GO:0006457	protein folding	0.0019	0.0432
GO:0006508	proteolysis	0.0032	0.0432
GO:0071214	cellular response to abiotic stimulus	0.0045	0.0432
GO:0048608	reproductive structure development	0.0059	0.0432
GO:0040014	regulation of multicellular organism growth	0.0071	0.0432
GO:0006729	tetrahydrobiopterin biosynthetic process	0.0077	0.0432
GO:0006091	generation of precursor metabolites and energy	0.0106	0.0432
GO:0072593	reactive oxygen species metabolic process	0.0111	0.0432
GO:0045333	cellular respiration	0.0111	0.0432
GO:0030707	ovarian follicle cell development	0.0113	0.0432
GO:0009056	catabolic process	0.0126	0.0432
GO:0044248	cellular catabolic process	0.0127	0.0432
GO:0072001	renal system development	0.0139	0.0432
GO:0002064	epithelial cell development	0.0165	0.0432

GO:0031669	cellular response to nutrient levels	0.0184	0.0432
GO:0001666	response to hypoxia	0.0201	0.0432
GO:0036293	response to decreased oxygen levels	0.0201	0.0432
GO:0070482	response to oxygen levels	0.0201	0.0432
GO:0001655	urogenital system development	0.0206	0.0432
GO:0043648	dicarboxylic acid metabolic process	0.0207	0.0432
GO:0016192	vesicle-mediated transport	0.0211	0.0432
GO:0015980	energy derivation by oxidation of organic compounds	0.0214	0.0432
GO:0002065	columnar/cuboidal epithelial cell differentiation	0.0216	0.0432
GO:0002066	columnar/cuboidal epithelial cell cell development	0.0216	0.0432
GO:0051641	cellular localization	0.0231	0.0444
GO:0042743	hydrogen peroxide metabolic process	0.0277	0.0478
GO:0042744	hydrogen peroxide catabolic process	0.0277	0.0478
GO:0019220	regulation of phosphate metabolic process	0.0287	0.0478
GO:0051174	regulation of phosphorus metabolic process	0.0287	0.0478
GO:0061326	renal tubule development	0.0338	0.0491
GO:0061333	renal tubule morphogenesis	0.0338	0.0491
GO:0006099	tricarboxylic acid cycle	0.0355	0.0491
GO:0008406	gonad development	0.0368	0.0491
GO:0045137	development of primary sexual characteristics	0.0368	0.0491
GO:0000281	mitotic cytokinesis	0.0382	0.0491
GO:0061640	cytoskeleton-dependent cytokinesis	0.0382	0.0491
GO:0044743	intracellular protein transmembrane import into intracellular organelle	0.0399	0.0491
GO:0065002	intracellular protein transmembrane transport	0.0399	0.0491
GO:0071806	protein transmembrane transport	0.0399	0.0491
GO:0006536	glutamate metabolic process	0.0403	0.0491

Table S6: GO Enrichment for partially fed ticks (Cellular Component).The elimKS column represents the significativity of the the Elim-Kolmogorov Smirnoff test implemented in the topGO R package, maintained by Adrian Alexa and Jorg Rahnenfuhrer. This tests for enriched GO terms in genes significantly over-expressed in partially fed ticks in comparison with unbiased genes.

GO.ID	Term	elimKS	adjustedFDR

GO:0005578	proteinaceous extracellular matrix	2.6e-05	0.0013
GO:0005576	extracellular region	0.00031	0.0077
GO:0042470	melanosome	0.00279	0.0331
GO:0005788	endoplasmic reticulum lumen	0.00281	0.0331
GO:0005581	collagen trimer	0.00331	0.0331
GO:0044421	extracellular region part	0.00483	0.0402

3 Article II: Investigation of the population structure of *Ixodes ricinus* at the European scale with transcriptomes

Contents

3.1 Forewords

This work have benefited from a grant from the Royal Swedish Academy of Sciences in order to visit Pierre de Wit during three months.

It allowed me to begin manipulate RNA-seq data for population genomics studies with a mapping-based approach. The presented work, is rather incomplete and should be consolidated even if results seems converging to the same answer.

Proposed title: A Pool-RNAseq approach demonstrates the genetic structuration of *Ixodes ricinus* by geographical distance.

Authors: N. Pierre CHARRIER[1*], Pierre DEWIT[2], Axelle HERMOUET[1], Caroline HERVET[1], Olivier PLANTARD[1], Claude RISPE[1]
[1]: BIOEPAR, INRA, Oniris, Université Bretagne Loire, F-44307 Nantes, France
[2]: University of Gothenburg, Department of Marine Sciences, Sven Lovèn Centre for Marine Sciences, Tjärnö, Sweden
[*]: Corresponding author – npcharrier@gmail.com

Samplers: - Elsa Quillery (Sweden)
- Ionut Pavel (Romania)
- Martin Pfeffer, Anna Obiegala (Germany)
- Franck Boué (France, Nancy)
- Albert Agoulon (France, Carquefou)
- Annetta Zindl (Ireland)
- Maarten Vordouw (Switzerland)
- Sandor Hornok (Hungary)
- Ana Garcia-Perez (Spain)
- Heikki Henttonen (Finland)
- Michalis Kotsyfakis (Czeh Republic)
- Swaid Abdullah (Bristol, UK)

Target journals: BMC genomics, Genome Biology, Molecular ecology

Keywords: Population genomic, SNPs, RNA-Seq, Transcriptome, Mantel, IBD

3.2 Introduction

Ticks are fascinating organisms, which as blood-feeding parasites represent a concern in animal and Human health, being the potential vectors of many pathogens. The tick, *Ixodes ricinus* has been a particular focus of interest because of its large distribution and abundance in Europe, and because it is the major vector of diverse pathogens, for example the Lyme borreliosis spirochaetes (*Borrelia spp.*) [1]) and the tick-borne encephalitis virus [2]. Found from the Mediterranean coast to the Scandinavia, *I. ricinus* is already widespread in Europe but global change could even increase its current geographical repartition [3–5]. *I. ricinus* is a generalist blood-feeder, found on all sorts of terrestrial tetrapods and occasionally on Human. As a vector, this capacity to connect reservoir hosts and incidental hosts implies risk of zoonoses [6]. Thus, understanding how this species is genetically structured is important in particular to appreciate the spread of tick-born pathogens across Europe which has been the subject of several studies.

Different factors influencing the genetic structure have been explored, such as geographic variation (see below), or potential adaptation to different hosts, particularly at the larvae or nymph stage [7–10]. These works indeed suggested that at least in some areas of its distribution (especially Southern Europe), *I. ricinus* ticks found on different hosts are genetically different, suggesting an incipient process of adaptation and eventually of speciation. Because of the weak pattern of differentiation however, the authors suggested it represented the very early stages in the formation of host races [10]. Of note, a stronger pattern of genetic differentiation associated with feeding on different host species has been demonstrated in another tick species, *I. uriae*, which can be found on different marine bird species [11].
Contrasted results were obtained when studies investigated the geographical structuration of *I. ricinus* [12]. Using five partial mitochondrial genes for 26 individual from 22 locations across Europe, no phylogeographical structure could be found [13]. In the same way, no differentiation was observed at the European scale using two mitochondrial genes and four nuclear genes [14], even though a clear differentiation between north Africa and European populations was observed. This north-African population was recently elevated as a new species, namely *I. inopinatus* [15]. The same conclusions were drawn from a study exploring the response to Pleistocene climatic changes using two mitochondrial and two nuclear genes for 210 individuals from 22 locations across Europe [16]. While sequencing full mitogenomes for two populations from Italy and one from Slovakia, authors reveals a diversity of lineages in this three populations without revealing population structure between sites [17]. By contrast, differentiation between two geographically distant populations was observed using multilocus sequence typing of six mitochondrial genes (mtMLST) – Latvia and United Kingdom (UK) were different enough in terms of haplotype composition to reject the null hypothesis that there is no difference between populations [18]. A recent study focusing on the North-European population of ticks also found significant differences between UK, Norway and Baltic region using two fragments of mitochondrial genes for up to 442 individuals from 22 sites [19].
These different results could be explained by the choice of genetic markers and sampling strategies, some studies focusing on few distant locations [17, 18], North European [18, 19] or South European region [14, 16, 17]. Furthermore when detected, genetic differentiation seems to be moderate except in the most variable genetic regions such as the control region in the mitochondrial genome [19]. To solve the question of geographic structuration in Europe, and assessing more precisely the level of differentiation at different scales of distance, there is therefore a need of both a high number of genetic markers and of adequate sampling strategy (covering multiple sampling points all over Europe). To our knowledge, no such studies have been performed yet, and in particular no study used high-throughput sequencing strategies in order to explore the genetic diversity of *I. ricinus* at the European scale.
One possibility could be the use of restriction-site-associated DNA sequencing (RADseq). This approach was used in two studies, respectively for *Amblyomma*

americanum [20] and *Ixodes scapularis* [21]. This method potentially provides many single-nucleotide polymorphism sites (SNPs) but does not easily allow to localize sequenced sites in the absence of a reference genome. Here, we preferred an approach based on transcriptomes, which by definition targets transcribed sequences, for potentially thousands of coding genes thus representing a large sample and largely distributed number of markers. Furthermore, this approach will permit to enrich the already well-stocked RNA-seq based studies on *I. ricinus* and thus contributes to enlarge the genomic ressources for that species. While detecting, extracting and analyzing SNPs from RNA-seq represents challenges which must be carefully addressed [22], population transcriptomics approaches have been successfully used in several recent works, for example to identify genetic markers associated with growth in the rainbow truit [23] or to explore the determinants of genetic diversity across the tree of life [24]. Using transcriptomes obtained for twelve different populations of ticks, we investigated the genetic strucure of *I. ricinus* populations across Europe. Our study revealed clear patterns of geographic structuration at the level of this continent.

3.3 Material and method

3.3.1 Tick collection

Sampling was done at 12 locations in Europe, either in autumn 2016 or in the spring of 2017. Questing ticks of the nymph stage were sampled by flagging vegetation at each location (details Table 3.1, illustrated Figure 3.1). After sampling, live ticks were washed with a 10% solution of concentrated bleach, rinsed twice in a pure water solution and then placed in a solution of RNAlater. Each sample was constituted of (ideally) 50 nymphs. These samples were then immediately shipped to our laboratory (BIOEPAR, Nantes), where they were put into Trizol (Invitrogen, Life Technologies, Carlsbad, CA, USA) and frozen.

3.3.2 RNA extraction

Whole bodies of nymphs were grinded with a soft plastic pestle in Trizol (Invitrogen, Life Technologies, Carlsbard, CA, USA) on dry ice. Then RNA was purified by adding chloroform, and proceeding to centrifugation. The aqueous phase was then placed into a NucleoSpin RNA XS column (Macherey-Nagel, Düren, Germany), including a DNase treatment. RNA was stored into RNasin (Promega, Madison, USA). RNA samples were then sent in dry ice to a sequencing platform (GeT-PlaGe, Toulouse). Dosages and quality were assessed with (Nanodrop Thermo Fisher Scientific, Waltham, USA), Qubit (Invitrogen, CA, USA), an Experion machines (Bio-RAD Laboratories Inc., Hercules, USA) and Fragment Analyzer (Agilent, Santa Clara, USA).

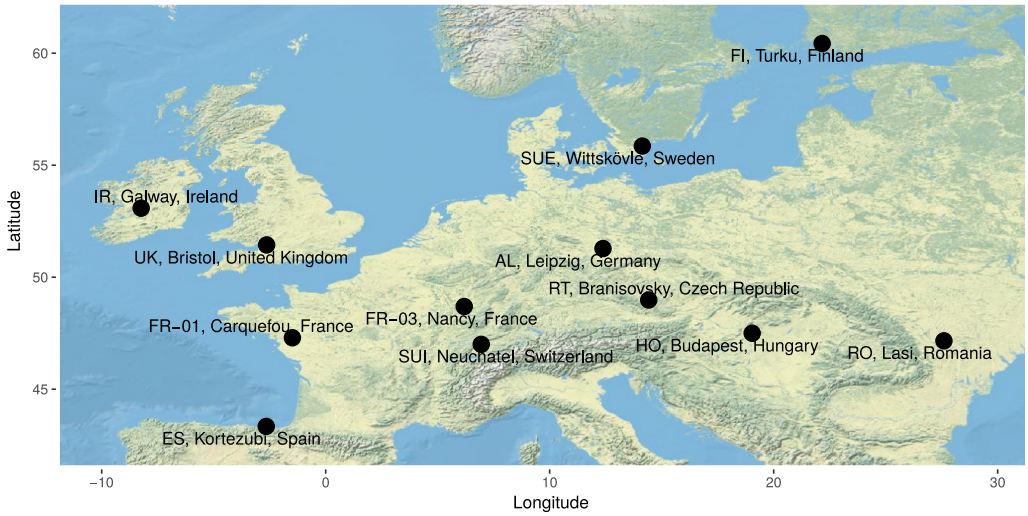

Figure 3.1: Geographical map of the twelve sequenced populations of *I. ricinus*, see details in Table 3.1

Table 3.1: Details on the sampled populations. The code represents the name used to designate the population through this study. Locality is the name of the place associated with the country name where the nymphs were flagged. Latitude and Longitude are given as well as the number of nymphs used to prepare the libraries (see map Figure 3.1)

Code	Locality	Country	lat	lon	Pool (n nymphs)
AL	Leipzig	Germany	51.28	12.38	50
ES	Kortezubi	Spain	43.34	-2.65	50
FI	Turku	Finland	60.43	22.16	50
FR-01	Carquefou	France	47.29	-1.49	50
FR-03	Nancy	France	48.69	6.18	50
HO	Budapest	Hungary	47.50	19.04	50
IR	Galway	Ireland	53.08	-8.24	47
RO	Lasi	Romania	47.16	27.60	50
RT	Branisovsky	Czech Republic	48.98	14.42	30
SUE	Wittskövle	Sweden	55.85	14.14	50
SUI	Neuchatel	Switzerland	47.00	6.95	50
UK	Bristol	United Kingdom	51.44	-2.64	50

3.3.3 Library preparation and sequencing

All of the twelve samples had sufficient quantities, concentrations and qualities of RNA to proceed with library preparation (but the relative low quantity of RNA in three samples, FR-01, SUE and UK led us to use three additional cycles of PCR, see below). The library preparation kit was NEBNext Ultra Directional RNA Library Prep Kit, NEB Art. No E7420. Poly-A selection with a magnetic isolation module, was used to target mRNAs, followed by strand-specific cDNA synthesis with an insert size of 150–400 bp, PCR amplification and library purification. Individual tags used for the 12 samples allowed multiplex sequencing. Sequencing was done on one lane of an Illumina Hiseq 3000 machine.

3.3.4 Read cleaning and mapping

Raw reads were cleaned using `Trimmomatic` (v-0.32) with a minimum of 36 remaining bases, a minimum average quality of 15 over 4 bases [25]. Only pair-end reads were conserved. The length of raw reads was 140 bp, and the average length of cleaned reads was XXX. Reads were then mapped on a reference transcriptome using `bwa mem` (version: 0.7.12-r1039) [26]. This transcriptome of *I. ricinus* was assembled from 15 libraries from different conditions [27]. The NCBI accession (TSA section) of this assembly is GFVZ.
Bam files were treated using the `samtools suite` (version: 1.3.1). Duplicates were removed using `Picard MarkDuplicate` (version: 2.1.1). Following the recommendations of `PoPoolation2` pipelines [28], variants were called using `samtools mpileup`, synchronized and cleaned using a phred-quality threshold of 20. For any detected indel (covered by at least 5% of the mapped reads) five bases in each direction were discarded. We considered only bi-allelic variants covered at least by 20 reads in each library.
To help us adjusting our filtering parameters and distinguishing true SNPs and sequencing errors, we computed the Transition-Transversion ratio (Ti/Tv ratio) as well as the ratio of the number of non-synonymous variants over the synonymous variants (NS/S ratio). All analyses were achieved by writing R and Perl scripts (available on https://github.com/npchar/PopulationTranscriptomics). These two ratios are expected to reach stability when criteria are conservative (high specificity but low sensitivity). The inclusion of false positive SNPs in a populational data set is is expected to increase the NS/S ratio and decrease the Ti/Tv ratio, as compared to robust SNPs (which have typically an excess of synonymous over non-synonymous variants, and an excess of transitions over transversions). To determine the minimum allele frequency (MAF) at the level of the whole data set (all 12 populations combined) that would provide an optimal compromise between the total number of SNPs analyzed and the robustness of these variants, we examined changes in the NS/S and Ti/Tv ratios (averages among all sites) with MAF. We computed the two ratios for MAF values ranging between 0 and 6%,

with steps of 0.1% (between 0 and 1%) or of 0.2% (between 1 and 6%).

3.3.5 Measure of genetic distance

Minor allele frequencies were computed for each population for the final set of bi-allelic variants (estimated from allele counts reported by the sync file of `Popoolation2` [28]). Then, we computed the matrix of euclidean distance between populations using the vectors of allele frequencies. A principal coordinate analysis was performed using the `ade4 R` package. We first estimated the Fixation index (F_{ST}), a common statistic used to study population structure firstly by running the `PoPoolation2` [28] internal estimators. This script provided Fst per locus for every pair of populations.

Because we were interested in measuring the distance between populations and not screened loci, we needed to compute an F_{ST} measure which recapitulates the genetic distance across loci. Two strategies have been documented i.) averaging F_{ST} values across loci (noted $\hat{F}_{ST}^{U}$), and ii.) dividing the sum of nominator's estimator across loci by the sum of denominator's estimator across loci (see [29, 30], noted $\hat{F}_{ST}^{W}$). Averaging F_{ST} over loci is known to underestimate the difference between two populations by reducing the effect of rare variants. We chose to have a possible comparison by also estimating the second global $\hat{F}_{ST}$. However, the second strategy supposed to have access to the parameters per loci which was not possible from the `PoPoolation2` output.
Different FST estimators have been proposed and we choose to take advantage of their differences[29, 31]. Considering π_s as the expected divergence between a pair of alleles from two different populations and π_b as the expected divergence within populations and following the development of Charlesworth [32] found in Jackson et al. (expression 4, 2014) [30], we define as follow:

$$\pi_s = p_a * (1 - p_a) + p_b * (1 - p_b)$$

$$\pi_b = p_a * (1 - p_b) + p_b * (1 - p_a)$$

With respectively p_a and p_b the frequency of one allele in the population A and B. For a particular locus, according to Jackson et al. (2014)[30], the Weir and Cocherman (1984) [33] F_{ST} estimator can be expressed as follow:

$$\hat{F}_{ST}^{Jackson} = \frac{\pi_b - \pi_s}{\pi_b}$$

$$\hat{F}_{ST}^{U.Jackson} = \frac{1}{S} \sum_{i=1}^{S} \hat{F}_{ST}^{Jackson}(i)$$

$$\hat{F}_{ST}^{W.Jackson} = \frac{\sum_{i=1}^{S}(\pi_b^{(i)} - \pi_s^{(i)})}{\sum_{i=1}^{S}\pi_b^{(i)}}$$

In Bhatia et al. (2013), the development take into account the size of the sample (n_a and n_b for sample size in pop a and pop b) used also by Chen et al., (2016):

$$\hat{F}_{ST}^{Bhatia} = \frac{(p_a - p_b)^2 - \frac{p_a*(1-p_a)}{n_a-1} - \frac{p_b*(1-p_b)}{n_b-1}}{p_a*(1-p_b) + p_b*(1-p_a)}$$

$$\hat{F}_{ST}^{U.Bhatia} = \frac{1}{S}\sum_{i=1}^{S}\hat{F}_{ST}^{Bhatia}(i)$$

$$\hat{F}_{ST}^{W.Bhatia} = \frac{\sum_{i=1}^{S}((p_a^{(i)} - p_b^{(i)})^2 - \frac{p_a^{(i)}*(1-p_a^{(i)})}{n_a-1} - \frac{p_b^{(i)}*(1-p_b^{(i)})}{n_b-1})}{\sum_{i=1}^{S}(p_a^{(i)}*(1-p_b^{(i)}) + p_b^{(i)}*(1-p_a^{(i)}))}$$

Additionally we used the internal $\hat{F}_{ST}$ computation implemented in `PoPoolation2` which was averaged across all variants.

Matrices of pairwise F_{ST} values between locations were used to build a dendrograms using an UPGMA method. The resulting dendrogram from the five F_{ST} estimators were combined into a supernetwork in order to evaluate the robustness of our clusterings. We used a quartet based method conserving the branch lengths implemented in `SuperQ` [34] with the Gurobi solver.

3.4 Results

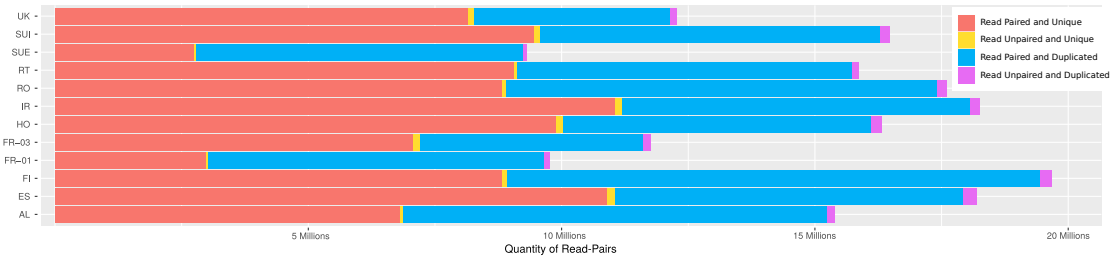

Figure 3.2: Quantity of reads per library uniquely mapped to the reference transcriptome. Categories of mapped reads are indicated as follow: Unique pairs of reads, Unique reads of a pair, Duplicated pairs of reads, and Duplicated reads of a pair.

Table 3.2: Sequencing and mapping informations for the 12 libraries. The two columns indicate the number of read pairs obtained by library and the number of not duplicated pairs mapping without ambiguity.

	Number of read pairs	Read Pairs Unique
AL	26,951,602	6,808,944
ES	30,585,788	10,892,032
FI	32,147,572	8,827,466
FR-01	17,387,041	2,992,495
FR-03	21,434,909	7,073,497
HO	26,630,241	9,901,267
IR	28,556,585	11,058,093
RO	31,191,860	8,815,596
RT	26,348,778	9,056,867
SUE	17,966,680	2,759,404
SUI	28,563,411	9,456,629
UK	20,394,332	8,157,883
sum	308,158,799	95,800,173
mean	25,679,899.92	7,983,347.75

3.4.1 Sequencing and mapping

Sequencing resulted in average of 25.7M pairs of reads per library (with a minimum of 17.4 M for FR-01 and a maximum of 32.1M for FI). We obtained on average 15.1M pairs of reads which mapped without ambiguity on one and only one location in our set of transcripts (minimum of 9.3M for SUE and a maximum of 19.7M for FI). From this uniquely mapping reads, we considered only unique pairs of reads in order to avoid optical and PCR duplicates (Figure 3.2). In average, 8.0M pairs of reads were found unique, mapped without ambiguity on one transcript and for both reads of the pair (minimum of 2.9M for SUE and a maximum of 11.1M for IR, see Table 3.2).

3.4.2 Selected variants

Our exploration of the relative influence of the percentage of counting events on the Ti/Tv ratio and the N/S ratio resulted in an estimation of a Ti/Tv at 2.26 and a N/S at 0.33 (Figure 3.3). We chose a criterion of 3% which was respecting the plateau of the N/S and the Ti/Tv in function of the minimum counting events. While loosing rare variants and then decreasing the sensibility, we theoretically improved our specificity. It resulted in a set of 155,957 SNPs from 5786 transcripts with an average density of 1 SNP every 20bp (see Figure S1).

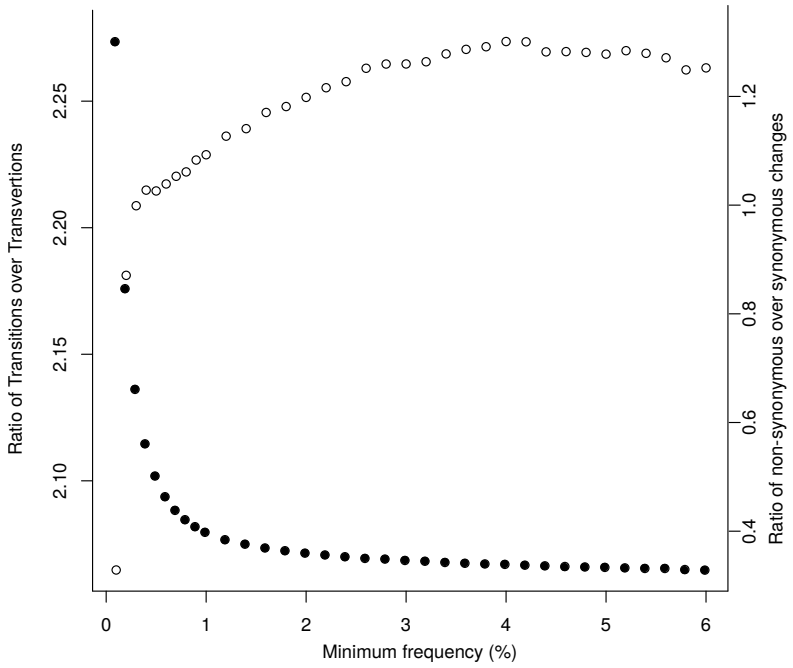

Figure 3.3: Ti/Tv ratio (circle) and number of Non-synonymous over Synonymous variants (black dot) in function of the minimum frequency of the minor allele to consider a variant across all libraries.

3.4.3 Genetic distance from Fixation index and geographical distance

We observed an increase of the genetic distance measured by the FST statistics and the geographical distance (Figure 3.4-a). This correlation was significant whithout regard of the F_{ST} estimator; the variance of the genetic distance explained at 35% by the variance of the geographical distance with a F(1,64)=36.46, p-value=$8.705e-08$ (values are indicated for the $\hat{F}_{ST}^{W.Jackson}$). This positive correlation was significant (Figure S2) when F_{ST} was divided by (1-F_{ST}) and the geographical distance was naturally log-transformed (R^2=0.31, F(1,64)=30.29, pvalue=$7.0e-07$) following recommendation of Diniz-Filho et al., 2013 [35].
We observed a strong correlation between the two matrices (F_{ST} and geographical distance) with a r Mantel's statistics of 0.60. Using 999 permutations, we rejected the null hypothesis (Ho: no correlation between the two matrices) with a risk alpha inferior of 0.002 (Figure 3.4-b). The link between geographical distance and the genetic distance was maximum for the 400-800km class of distance (Figure 3.4-c). No correlation between geographical distance and genetic distance was observed after the class 1200-1600km.

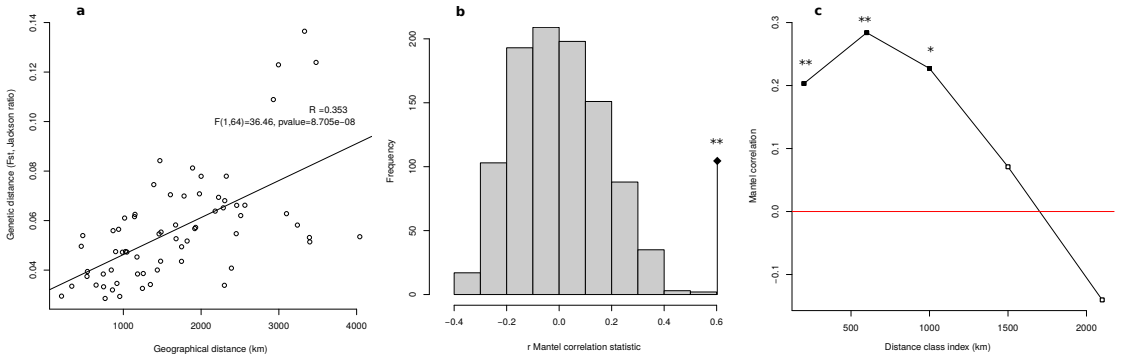

Figure 3.4: Isolation by distance from F_{ST} matrix. Panel a represents the F_{ST} measure between two populations in function of the geographical distance. Panel b represents the distribution of the r Mantel correlation statistics for 999 permutations from the F_{ST} matrix and the geographical distance. The observed r value is represented by a square at the tip of a line. The Mantel correlogram represents the r value for a Mantel test when considering five clases of distance. Significativity: * <0.05, ** = <0.005.

3.4.4 Principal Coordinate Analysis

The PCoA analysis based on the minor allele frequency (Figure 3.5), revealed that the plan formed by the first two axes was a picture of the sampling location (Figure 3.5-b). Indeed, correlation between the geographical distance and a measure of the

genetic distance was even higher using the coordinate of the populations on the plane formed by the first and second axis of the PCoA (rMantel=0.74, simulated p-value=0.001 based on 999 permutations). When not considering the Finland population (FI), the positions on the first PCoA axis were correlated with the longitudinal position of the sample (R^2=0.70, F(1,9)=24.71, pvalue=$7.70e-04$). The second axis isolated the RO and HO from the rest of the other locations (Figure 3.5-b and 3.5-c) while the third axis isolated the FR-01 location (Figure 3.5-c and 3.5-d).

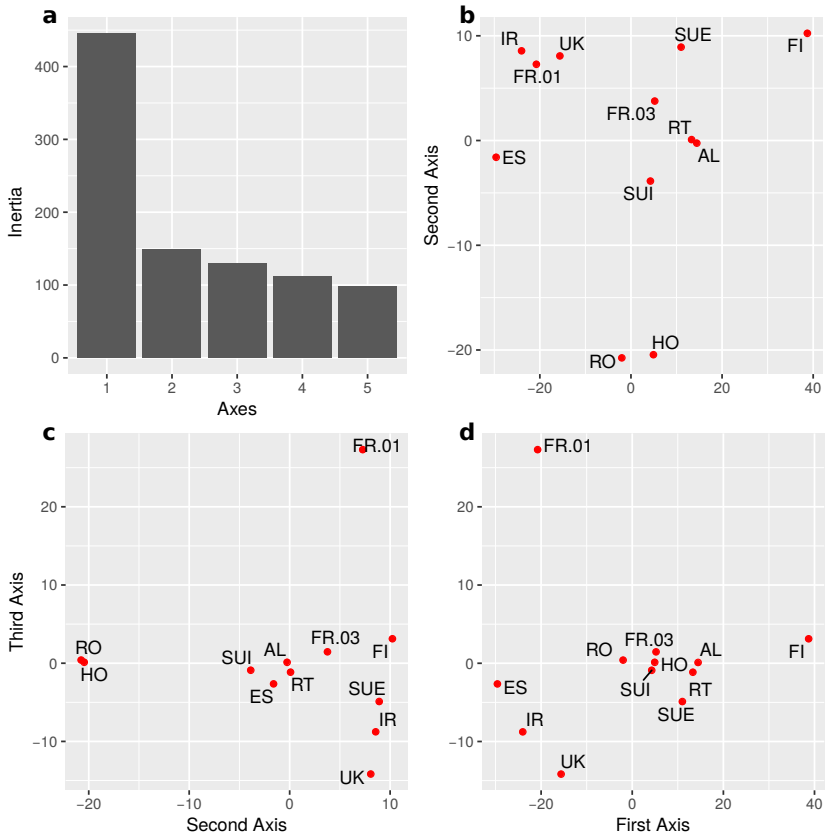

Figure 3.5: PcoA based on minor allele frequencies

3.4.5 Dendrograms from genetic distance

Dendrograms reconstructed by the F_{ST} pairwise measures using a UPGMA method were consistent regardless of the method used to compute the Fixation index (see Figure S3). A representation of the differences between the clustering of these fiveF_{ST} measures were achieved using a quartet method which produced a supernetwork (Figure 3.6). In all the cases, RT, SUI, AL and FR-03 were clustered all together, as well as RO and HO and ES, IR and UK. In most of the cases, FR-01

was sister group of the ES-IR-UK group (Figure S3) while FI was highly mobile and generally distant from the other locations.

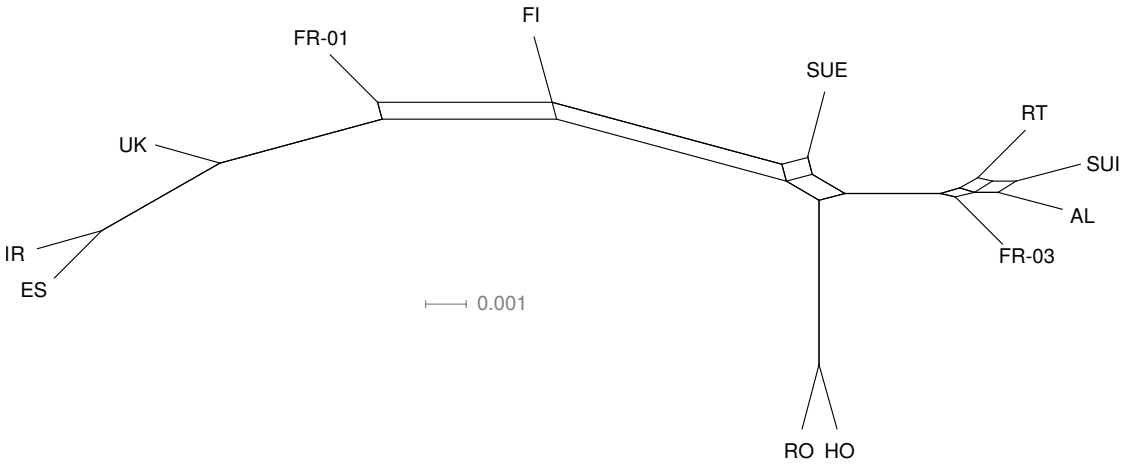

Figure 3.6: Supernetwork constructed by a quartet decomposition of dendrograms based on five different F_{ST} measures.

3.5 Discussion

Our survey of transcriptomic data from twelve wild populations of *I. ricinus* allowed us to identify 155,957 SNPs localized on 5786 transcripts. Such amount of genetic variability may seem surprising. We stress that we chose very conservative parameters to retain polymorphic sites: our criteria were indeed the unique mapping for pairs of reads, a minimum allele frequency across all populations of 3%, and a coverage of 20 reads per library (i.e. population). Furthermore, in order to deal with problematic alternative transcripts, we used a finely compressed set of transcripts as our reference transcriptome. Alternative transcripts could generate indels which can contribute to overestimate genetic variability (by erroneous mapping on flanking region). These indels were removed, as well as the flanking regions (5 bases from each side of each indel). Another factor that could lead to overestimating polymorphism could be the existence of recent paralogs. A *de-novo* assembly (as our reference transcriptome) may for example contain consensus sequences of two paralogs (which cannot be easily separated if the sequence divergence is low). This could be the case here given our choice to reduce transcript redundancy for highly similar sequences (but note that we used a 98% identity cutoff, so only very recent paralogs with low divergence could be affected). In this case, reads from the two gene copies should map to the same gene, and divergence among gene copies will be confused with sequence polymorphism. However, should this phenomenon account for an excess polymorphism, it

would be concentrated on given number of genes (corresponding to recent paralogs). Here, on the contrary, we observe a rather evenly distributed level of polymorphism.

The observed level of polymorphism reaches 1 SNP in average every 20 bases, which is high compared to densities reported for human 1-10 SNPs per 10kb[36]. But this result is not far away from level of polymorphisms observed for *Aedes aegypti* (1 SNPs per 83 bp) and coherent with the previous even higher estimate for the black legged tick (1 SNPs every 14bp) [37]. Given the filtering and strict criteria we used, we therefore believe that our estimation of polymorphisms level (for transcribed regions) is unlikely to be strongly affected by sequence error biases or compression of recent paralogous copies.

Using the Euclidian distance between the geographically distant populations, we observed a strong influence of the geographical distance on the genetic variability. This was seen both in the ordination analysis (PCoA) based on the MAF as well as with the clustering of populations based on their genetic distance (F_{ST}). This correlation between geographical distance and genetic distance was confirmed by a Mantel test. So far few studies reported a geographical factor of variation of the genetic diversity in *I. ricinus*, and even several works concluded that there was no phylogeographical structure[13, 14, 16, 17]. These studies defended that long term dispersion by host eroded geographical structure, thus making of *I. ricinus* a single panmictic population in Europe. Our results are thus in stark contrast with these findings. The absence of phylo-geographical structure in several previous studies could be explained by differences in sampling strategy as well as by the molecular markers used. Indeed, while Carpi et al. in 2016 used high throughput sequencing methods in order to obtain information from complete mitogenomes, only three locations were investigated[17]. At the opposite, 22 locations across Europe were investigated by Porretta et al. in 2013, but with only 4 markers[16].

However two recent studies showed a clear difference between East and West populations[18, 19]. This result is in accordance with our work, and we therefore argue that geographic structuration in this species is clear, but can only be identified by using a sufficient number of markers and adequate sampling (i.e. by using several distant populations). Though our clustering analysis, we could identify two groups of populations: i) the first composed of SUI, AL, RT and FR-03 (Eastern cluster), and ii) the second composed of ES, IR, UK (Western cluster). Theses two clusters are coherent with structures found previously in a sense that the Latvia populations was found significantly different from the UK based population[18] and that the German populations were found to be quite different from those sampled in the UK[19]. Indeed, the FR-01 and ES populations are genetically closer from UK and IR than from FR-03 or SUI. This is unexpected because, these populations are separated by seas which could represent a genetic barrier as observed between Norway and UK populations in Roed et

al, 2016[19]. However, we must consider the fact that birds may serve as hosts to *I. ricinus* and could connect distant populations separated by seas. In that perspective differences between east and west populations could originate from the two different flyways (namely the Black sea/Mediterranean flyway and the East Atlantic flyway)[38]. Indeed, these two markedly different roads of bird migration could tend to homogenize genetic diversity in their respective area by long-distance dispersion. An alternative hypothesis is that the difference in terms of climate between the oceanic versus continental regions could play an important selective pressure as it is observed for the transmission of the tick-borne encephalitis virus[39].

The FI pool of individuals did not cluster with other populations. This situation could be explained by geographical isolation of this population, and a reduced gene flow from either the Western or Eastern populations. Alternatively, we must consider the possibility of hybridization events with a different species, *I. persulcatus*, whose westernmost populations occur in Northern Finland and in Latvia. Indeed, in Latvia, these two species are in sympatry and hybrids have been reported[40]. Presence of hybrids could explain the particular position of FI pool in F_{ST} dendrograms. This hypothesis could be tested by sequencing pools of nymphs for *I. persulcatus* individuals from Finland and Latvia, to determine if the finnish populations of *I. ricinus* contain alleles that come from *I. persulcatus*. A similar scenario could occur in Portugal and Spain were *I. inopinatus* can be found in sympatry with *I. ricinus*[15].

For the present study, we could not obtain populations from Southern Europe. These populations could however provide an interesting insight on the possible recolonization process described by Porretta et al. in 2013[16]. Future studies including a denser sampling with more diverse localities would help obtaining a more accurate description of the geographical structuration of this species. Our strategy based on pools of individuals permitted to obtain hundreds of thousands variable sites for an affordable price and then to study the genetic structure of *I. ricinus* at the European scale. But because of the pooled nature of our design, we were unable to obtain individual-based metrics, for example to check the Hardy-Weinberg equilibrium, which permitted at Kempf et al. to suggest that *I. ricinus* could have sex-specific dispersion ability and/or host usage[10]. The pool-RNA-seq approach used in our present study could be applied to test other hypotheses on other factors of structuration: in particular, this could allow to give a more complete answer to the question of incipient host-race formations (as proposed by Kempf et al, 2011[10] and summarized in McCoy et al, 2013[41]). Pool of adults could be sampled on different hosts (birds, roe deer, etc) following a South to North gradient to test both the a post-glacial recolonization process and the incipient host race formation, as hypothesized by Elsa Léger during her PhD work (see McCoy et al, 2013[41]).

3.6 Supplementary materials

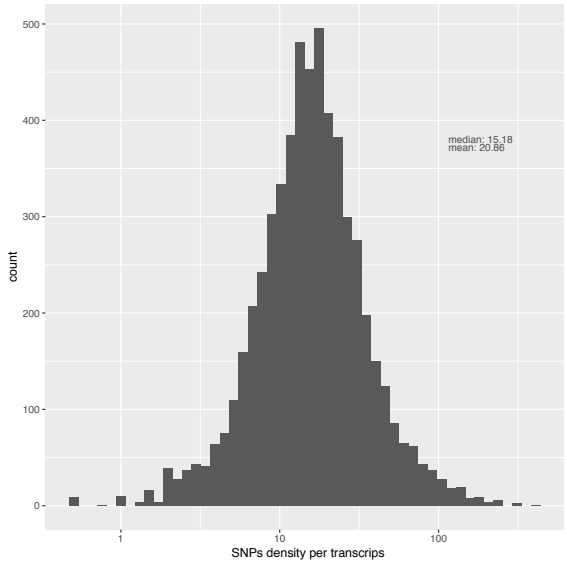

Figure S1: Histogram of the SNPs density per transcripts (on a log10 scale).

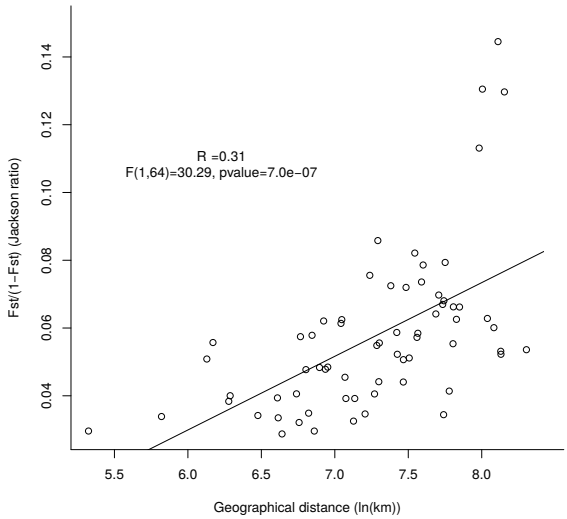

Figure S2: Genetic and geographic distance. F_{ST}/(1-F_{ST}) function of ln(km) see Diniz-Filho, 2013 ; Rousset, 1997

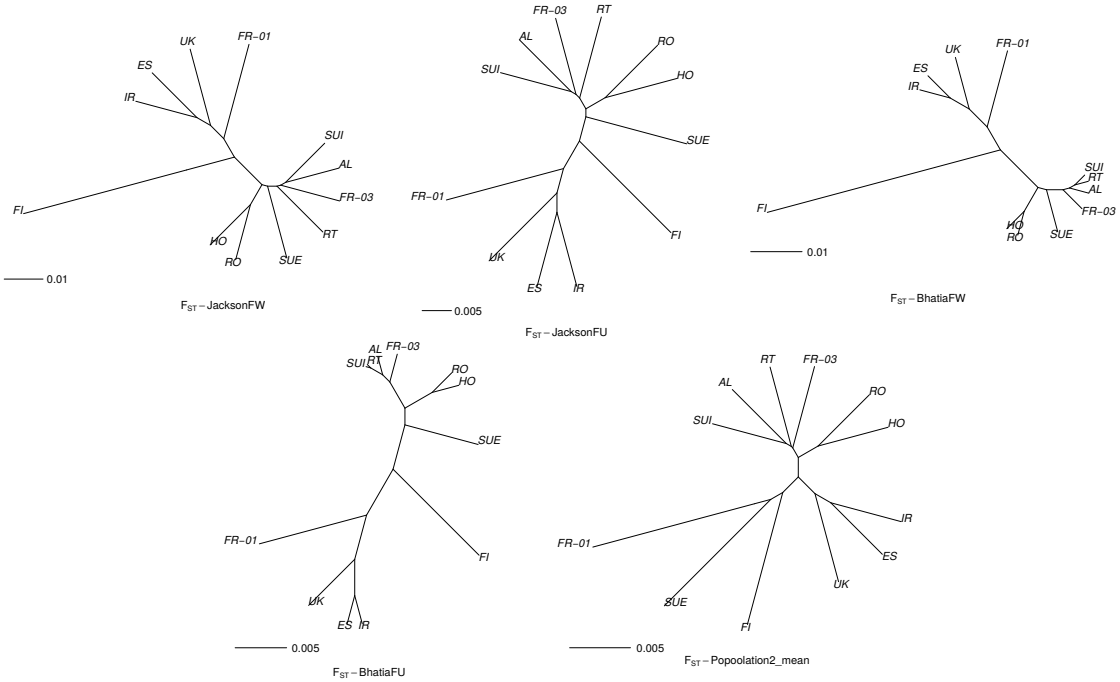

Figure S3: Dendrogram based on the UPGMA clustering for the five different F_{ST} estimators.

Bibliography

1. A. Rizzoli et al.: Lyme borreliosis in Europe. *Eurosurveillance* **16**(27) (2011), 19906. doi: 10.2807/ese.16.27.19906-en.
2. J. Süss: Tick-borne encephalitis 2010: Epidemiology, risk areas, and virus strains in Europe and Asia—An overview. *Ticks and Tick-borne Diseases* **2**(1) (2011), 2–15. doi: 10.1016/j.ttbdis.2010.10.007.
3. T. G. Jaenson and E. Lindgren: The range of *Ixodes ricinus* and the risk of contracting Lyme borreliosis will increase northwards when the vegetation period becomes longer. *Ticks and Tick-borne Diseases* **2**(1) (2011), 44–49. doi: 10.1016/j.ttbdis.2010.10.006.
4. J. M. Medlock et al.: Driving forces for changes in geographical distribution of *Ixodes ricinus* ticks in Europe. *Parasites & Vectors* **6**(1) (2013), 1. doi: 10.1186/1756-3305-6-1.
5. D. Porretta et al.: Effects of global changes on the climatic niche of the tick *Ixodes ricinus* inferred by species distribution modelling. *Parasites & Vectors* **6**(1) (2013), 271. doi: 10.1186/1756-3305-6-271.
6. A. Rizzoli et al.: *Ixodes ricinus* and Its Transmitted Pathogens in Urban and Peri-Urban Areas in Europe: New Hazards and Relevance for Public Health. *Frontiers in Public Health* **2** (Dec. 2014). doi: 10.3389/fpubh.2014.00251.
7. T. d. Meeûs, L. Béati, C. Delaye, A. Aeschlimann, and F. Renaud: Sex-biased genetic structure in the vector of Lyme Disease, *Ixodes ricinus*. *Evolution* **56**(9) (Sept. 2002), 1802–1807. doi: 10.1111/j.0014-3820.2002.tb00194.x.
8. F. Kempf, T. de Meeûs, C. Arnathau, B. Degeilh, and K. D. McCoy: Assortative Pairing in *Ixodes ricinus* (Acari: Ixodidae), the European Vector of Lyme Borreliosis. *Journal of Medical Entomology* **46**(3) (May 2009), 471–474. doi: 10.1603/033.046.0309.
9. F. Kempf, K. D. McCoy, and T. D. Meeûs: Wahlund effects and sex-biased dispersal in *Ixodes ricinus*, the European vector of Lyme borreliosis: New tools for old data. *Infection, Genetics and Evolution* **10**(7) (2010), 989–997. doi: 10.1016/j.meegid.2010.06.003.
10. F. Kempf et al.: Host races in *Ixodes ricinus*, the European vector of Lyme borreliosis. *Infection, Genetics and Evolution* **11**(8) (2011), 2043–2048. doi: 10.1016/j.meegid.2011.09.016.

11. K. D. Mccoy, T. Boulinier, C. Tirard, and Y. Michalakis: Host specificity of a generalist parasite: genetic evidence of sympatric host races in the seabird tick *Ixodes uriae*. *Journal of Evolutionary Biology* **14**(3) (May 2001), 395–405. doi: 10.1046/j.1420-9101.2001.00290.x.

12. A. Araya-Anchetta, J. D. Busch, G. A. Scoles, and D. M. Wagner: Thirty years of tick population genetics: A comprehensive review. *Infection, Genetics and Evolution* **29** (2015), 164–179. doi: 10.1016/j.meegid.2014.11.008.

13. S. Casati, M. Bernasconi, L. Gern, and J.-C. Piffaretti: Assessment of intraspecific mtDNA variability of European *Ixodes ricinus* sensu stricto (Acari: Ixodidae). *Infection, Genetics and Evolution* **8**(2) (2008), 152–158. doi: 10.1016/j.meegid.2007.11.007.

14. R. Noureddine, A. Chauvin, and O. Plantard: Lack of genetic structure among Eurasian populations of the tick *Ixodes ricinus* contrasts with marked divergence from north-African populations. *International Journal for Parasitology* **41**(2) (2011), 183–192. doi: 10.1016/j.ijpara.2010.08.010.

15. A. Estrada-Peña, S. Nava, and T. Petney: Description of all the stages of *Ixodes inopinatus* n. sp. (Acari: Ixodidae). *Ticks and Tick-borne Diseases* **5**(6) (2014), 734–743. doi: 10.1016/j.ttbdis.2014.05.003.

16. D. Porretta et al.: The integration of multiple independent data reveals an unusual response to Pleistocene climatic changes in the hard tick *Ixodes ricinus*. *Molecular Ecology* **22**(6) (Feb. 2013), 1666–1682. doi: 10.1111/mec.12203.

17. G. Carpi et al.: Mitogenomes reveal diversity of the European Lyme borreliosis vector *Ixodes ricinus* in Italy. *Molecular Phylogenetics and Evolution* **101** (2016), 194–202. doi: 10.1016/j.ympev.2016.05.009.

18. R. E. Dinnis et al.: Multilocus sequence typing using mitochondrial genes (mtMLST) reveals geographic population structure of *Ixodes ricinus* ticks. *Ticks and Tick-borne Diseases* **5**(2) (2014), 152–160. doi: 10.1016/j.ttbdis.2013.10.001.

19. K. H. Røed, K. S. Kvie, G. Hasle, L. Gilbert, and H. P. Leinaas: Phylogenetic Lineages and Postglacial Dispersal Dynamics Characterize the Genetic Structure of the Tick, *Ixodes ricinus*, in Northwest Europe. *PLOS ONE* **11**(12) (Dec. 2016). Ed. by U. G. E. Munderloh, e0167450. doi: 10.1371/journal.pone.0167450.

20. J. D. Monzón, E. G. Atkinson, B. M. Henn, and J. L. Benach: Population and Evolutionary Genomics of *Amblyomma americanum*, an Expanding Arthropod Disease Vector. *Genome Biology and Evolution* **8**(5) (Apr. 2016), 1351–1360. doi: 10.1093/gbe/evw080.

21. M. Gulia-Nuss et al.: Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. *Nature Communications* **7** (Feb. 2016), 10507. doi: 10.1038/ncomms10507.

22. P. De Wit, M. H. Pespeni, and S. R. Palumbi: SNP genotyping and population genomics from expressed sequences - current advances and future possibilities. *Molecular Ecology* **24**(10) (Apr. 2015), 2310–2323. doi: 10.1111/mec.13165.

23. M. Salem et al.: RNA-Seq Identifies SNP Markers for Growth Traits in Rainbow Trout. *PLoS ONE* **7**(5) (May 2012). Ed. by Z. Liu, e36264. doi: 10.1371/journal.pone.0036264.

24. J. Romiguier et al.: Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**(7526) (Aug. 2014), 261–263. doi: 10.1038/nature13685.

25. A. M. Bolger, M. Lohse, and B. Usadel: Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15) (Apr. 2014), 2114–2120. doi: 10.1093/bioinformatics/btu170.

26. H. Li and R. Durbin: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14) (May 2009), 1754–1760. doi: 10.1093/bioinformatics/btp324.

27. N. P. Charrier et al.: Whole body transcriptomes and new insights into the biology of the tick *Ixodes ricinus*. *Parasites & Vectors* **11**(1) (June 2018). doi: 10.1186/s13071-018-2932-3.

28. R. Kofler, R. V. Pandey, and C. Schlotterer: PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27**(24) (Oct. 2011), 3435–3436. doi: 10.1093/bioinformatics/btr589.

29. G. Bhatia, N. Patterson, S. Sankararaman, and A. L. Price: Estimating and interpreting FST: The impact of rare variants. *Genome Research* **23**(9) (July 2013), 1514–1521. doi: 10.1101/gr.154831.113.

30. B. C. Jackson, J. L. Campos, and K. Zeng: The effects of purifying selection on patterns of genetic differentiation between *Drosophila melanogaster* populations. *Heredity* **114**(2) (Sept. 2014), 163–174. doi: 10.1038/hdy.2014.80.

31. V. Hivert, R. Leblois, E. J. Petit, M. Gautier, and R. Vitalis: Measuring genetic differentiation from Pool-seq data. *Genetics* **210**(1) (2018), 315–330. doi: 10.1534/genetics.118.300900.

32. B. Charlesworth: Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution* **15**(5) (May 1998), 538–543. doi: 10.1093/oxfordjournals.molbev.a025953.

33. B. S. Weir and C. C. Cockerham: Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**(6) (Nov. 1984), 1358. doi: 10.2307/2408641.

34. S. Grünewald, A. Spillner, S. Bastkowski, A. Bögershausen, and V. Moulton: SuperQ: computing supernetworks from quartets. *IEEE/ACM Trans Comput Biol Bioinform* **10**(1) (2013), 151–60. doi: 10.1109/TCBB.2013.8.

35. J. A. Diniz-Filho et al.: Mantel test in population genetics. *Genetics and molecular biology* **36**(4) (2013), 475–485. doi: 10.1590/S1415-47572013000400002.

36. Z. Zhao, Y.-X. Fu, D. Hewett-Emmett, and E. Boerwinkle: Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene* **312** (2003), 207–213. doi: 10.1016/S0378-1119(03)00670-X.

37. J. V. Zee et al.: High SNP density in the blacklegged tick, *Ixodes scapularis*, the principal vector of Lyme disease spirochetes. *Ticks and Tick-borne Diseases* **4**(1) (2013), 63–71. doi: 10.1016/j.ttbdis.2012.07.005.

38. G. C. Boere and D. A. Stroud: The flyway concept: what it is and what it isn't. *Waterbirds around the world* (2006), 40–47.

39. S. E. Randolph, P. M. F. Green R. M., and D. J. Rogers: Seasonal synchrony: the key to tick-borne encephalitis foci identified by satellite data. *Parasitology* **121**(1) (2000), 15–23.

40. S. Kovalev, I. Golovljova, and T. Mukhacheva: Natural hybridization between *Ixodes ricinus* and *Ixodes persulcatus* ticks evidenced by molecular genetics methods. *Ticks and Tick-borne Diseases* **7**(1) (2016), 113–118. doi: 10.1016/j.ttbdis.2015.09.005.

41. K. D. McCoy, E. Léger, and M. Dietrich: Host specialization in ticks and transmission of tick-borne diseases: a review. *Frontiers in Cellular and Infection Microbiology* **3** (2013), 57. doi: 10.3389/fcimb.2013.00057.

4 Article III: Reconstruction of the Hard-ticks phylogeny using transcriptomes

Contents

4.1 Forewords

This work benefited from comments and critics after a presentation upon the Alphy meeting (Montpellier, february 2018), as well as during the 5th YNHM meeting (Paris, march 2018). The present manuscript have been subject to correction by some of the co-authors.

Proposed title: Transcriptomics help resolve phylogenetic relationships in ticks.

Authors: N. Pierre CHARRIER[1*], Axelle HERMOUET[1], Caroline HERVET[1], Olivier LAMBERT[1,2], Albert AGOULON[1], Stephen C. BARKER[3], Dieter HEYLEN[4]+, Céline TOTY[5], Karen D. MCCOY[5], Olivier PLANTARD[1], Claude RISPE[1]
[1]: BIOEPAR, INRA, Oniris, Université Bretagne Loire, F-44307 Nantes, France
[2]: Centre Vétérinaire de la Faune Sauvage et des Ecosystèmes des Pays de la Loire, Oniris, Nantes, France
[3]: Department of Parasitology, School of Chemistry & Molecular Biosciences, The University of Queensland, Brisbane, Qld, 4072, Australia
[4]: Evolutionary Ecology Group, Department of Biology, University of Antwerp, Belgium.
[5]: Laboratoire MIVEGEC (Maladies Infectieuses & Vecteurs: Ecologie, Génétique, Evolution & Contrôle), Université de Montpellier – CNRS – IRD, Montpellier, France
*: Corresponding author – npcharrier@gmail.com

Traget journals: BMC genomic, Genome Biology

Keywords : Phylogenomics, RNA-Seq, Transcriptome, Supermatrix, Supertree, Ticks

4.2 Abstract

Ticks represent a group of 900 described species classified in hard ticks, soft ticks, plus a monospecific family. Among hard ticks, the *Ixodes* genus represents a world-widely distributed group of species able to transmit pathogens such as *Borrelia burgdorferi s.l.*, the causative agent of the Lyme desease. Whether or not this genus is monophyletic and what kind of phylogenetic relationships has *Ixodes* genus with the other hard ticks genera is subject to debate. To better understand evolutionary patterns in this genus (for example how genes related to blood-feeding evolved, and how fast did they change over evolutionary time), an accurate phylogeny of the whole group is needed. In this study we rely on nine new transcriptomic data sets from high throughput sequencing, together with 18 other transcriptomes from public databases. After reconstructing transcriptomes for each species (27 in total), we predicted their coding sequences and performed sequence comparisons among them in order to identify Single copy orthologs (SCO). Using Maximum-likelihood and Bayesian frameworks, we combined a supertree and a supermatrix approach to obtain a reliable scenario of the hard tick phylogeny. If major nodes of the tree were well resolved, parts of the phylogenetic tree still remain difficult to resolve. Overall, our results confirmed previous work on tick phylogeny and bring new insight, in particular for the debated "ricinus complex" species. This work provides a solid framework to study the evolutionary history of ticks, and will facilitate further analyses of phylogeny and gene evolutionary patterns.

4.3 Introduction

Ticks are blood-feeding arthropods which parasitize many terrestrial vertebrates species including mammals, birds, lizards, snakes. Ticks are a concern in human and animal health because they may transmit pathogens, including many bacteria, viruses, protozoans, and nematodes. One prime example is *Borrelia burgdorferi s.l.* the causative agent of Lyme Borreliosis, which is carried by different ticks of the *Ixodes* genus. Understanding phylogenetic relationships among tick species is an important objective to better apprehend biological differences among genera or species, including traits related with the risk of pathogen transmission. The aim of this study was to help resolve the phylogeny of hard ticks, by using a much larger number of markers than in preceding studies, while covering a broad sample of tick species (with a focus on the *Ixodes* genus). Ticks possess large genomes [1], with for example 2 Gb in *I. scapularis* [2] and as much as 5 Gb in *Rhipicephalus microplus*, and associated with a high repetitive content [3]. Perhaps, for this reason, only a single whole genome in the whole tick group has been published and annotated (*I. scapularis*, [2]). Therefore, whole-genome based phylogenetic reconstructions are not yet feasible. An alternative approach consists in sequencing transcriptomes with high throughput methods (RNAseq) and reconstructing gene collections for each species after *de novo* assembly. With careful treatment (for exemple accounting for the high redundancy created by the frequent presence of different contigs for the same gene[4]), this approach allows one to obtain data for several hundred to several thousand orthologous genes in different species, providing a basis for robust phylogenetic construction [5, 6]. This approach has been used to decipher deep-branching events (e.g. in Bivalva [7], in Arachnida [8], or in Myriapoda [9]), to infer phylogeny and timing of diversification of the major lineages of Hymenoptera [10], to resolve the position of poorly known arachnid orders (Ricinulei: Arachnida) [11], but also to resolve phylogenetic relationships between more closely related species [12].

There are at least 900 recorded species of ticks, separated into three families: i) soft ticks or *Argasidae*, ii) hard ticks or *Ixodidae*, and iii) a monotypic family named *Nuttalliallidae*. Based on morphology, hard ticks are separated in two groups: Metastriata and Prostriata. One important character supporting these two groups is the position of the anal groove contouring the anal pore anteriorly for the Prostriata while running posteriorly for the Metastriata. Prostriata or (*Ixodinae*) are composed of a single genus, namely *Ixodes*; it represents the richest genus of hard ticks with 244 described species [13]. Metastriata are classically divided into five subfamilies and 10 genera [13]: i) *Amblyomminae* (*Amblyomma*, 130sp), ii) *Bothriocrotinae* (*Bothriocroton*, 7sp), iii) *Haemaphysalinae* (*Haemaphysalis*, 166sp), iv) *Hyalomminae* (*Hyalomma* 27sp; *Nosoma* 2sp), and v) *Rhipicephalinae* (*Dermacentor*, 34sp; *Rhipicentor*, 2sp; *Anomalohimalaya*: 3sp; *Rhipicephalus* 83sp; *Margaropus* 3).

Significant advances in the classification of ticks were achieved thanks to molecular sequencing and confirmed morphology based systematics [14, 15]. But questions have been raised in particular about the monophyly of *Ixodinae* [15, 16]. The 244 described species of the *Ixodes* genus are organized in 14 sub-genera [17]. While recognized as coherent from a systematic point of view [18], these sub-genera are still debated [19] and some species are not included in them (for exemple *I. ventalloi*) or changed of sub-genera (for exemple *I. frontalis* [17]). The relation between the Australasian *Ixodes* lineage and the "other *Ixodes*" lineage (following the terminology of Barker & Murrel, [20])has raised questions about the monophyly of the *Ixodes* genus [21, 22]. Still today this point remains ambiguous: one scenario is that *Ixodes* are monophyletic and form the Prostriata (the Australasian *Ixodes* species would be more closely related to "other *Ixodes*" species) [15, 16, 22–24], while the alternative scenario would result in a paraphyletic *Ixodes* genus (Australasian species grouping with the Metastriata) [22, 25, 26]. Some of these phylogenetic uncertainties could be resolved by using large-scale data sets and a much larger number of molecular markers than the first molecular studies.

Recent years have seen a rapid accumulation of transcriptome studies for ticks [24]. However, to our knowledge, sequencing has essentially been performed species by species, with the objective to explore mechanisms implicated in the feeding process, in blood digestion or in interactions with pathogens. In contrast, there has not yet been an RNAseq based phylogeny for ticks as a group. We also point that the large majority of RNAseq data are for the Metastriata, whereas there are only three species in the Prostriata (*I. scapularis*, *I. ricinus*, *I. persulcatus*) with large-scale RNAseq data. This is a concern if we want to obtain a broader and robust phylogenetic scenario for the evolution of *Ixodes* species. This is especially true if we take into account the fact that most species feeding on humans belong to the *Ixodes* genus, which argues for a better knowledge of the genetic distances and relationships among the different species of the family.

In order to clarify the phylogenetic relationships between hard ticks, we combined RNAseq for nine new species of the *Ixodes* genus with previously published RNASeq data sets for other tick species (our study thus comprised 27 species overall). We thereby obtained thousands of coding genes in each species and hundreds of orthologuous genes among species. This RNAseq based study allowed us to propose a robust phylogenetic framework for this group.

4.4 Material and methods

4.4.1 Taxon sampling and transcriptome sequencing

Nine tick species from the *Ixodes* genus (*I. acuminatus* Neumann 1901, *I. arboricola* Schule & Schlottke 1930, *I. canisuga* Johnston 1849, *I. frontalis* Panzer 1798, *I. holocyclus* Neumann 1899, *I. hexagonus* Leach 1815, *I. uriae* White 1852, *I. ventalloi* Gil Collado 1936, and *I. vespertilionis* Koch 1844) were collected in the field during 2016, either directly on their animal hosts or in the habitat of their animal hosts (Table 4.1). These nine species, plus three other *Ixodes* species for which we analysed published data sets (see below) represent six described sub-genera, plus one ungrouped species (*I. ventalloi*). Morphological identifications were conducted using the books of Hillyard [27], Perez-Eid [28], Barker & Walker [29], and the key of ornitophylic ticks [30]. Our aim was to obtain a reference transcriptome and a catalogue of transcripts as broad as possible for each species. For that purpose, whenever it was possible, we tried to collect a diversity of morphs and conditions (nymphs, adult females, males, fed or unfed ticks). A laboratory stock of *I. arboricola* that originate from woodlands around Antwerp (Belgium) was established in 2008, and has been maintained for 8 years by allowing ticks to feed on great tits [31].
Ticks from distant sampling locations (*I. arboricola*, *I. holocyclus*) were sent to our laboratory in RNAlater in cold conditions (4°C), whereas the other species were kept alive at 4°C until RNA extraction.

For eight of the nine *Ixodes* species (all species except *I. uriae*), whole tick bodies were ground in Trizol (Invitrogen, Life Technologies, Carlsbad, CA, USA). RNA was then purified, keeping the aqueous phase after a centrifugation with chloroform. RNA was extracted using NucleoSpin RNA XS column (Macherey-Nagel, Düren, Germany) including a DNase and was stored in RNAsin (Pro- mega, Madison, USA). RNA concentration and quality of each sample were determined using Nanodrop (Thermo Fisher Scientific, Waltham, USA), Qubit (Invitrogen, CA) and Agilent Bioanalyzer. For species with multiple conditions (different stages, different feeding status), RNA extraction was performed independently for each condition; these different samples were used to produce an equi-molar mix and a single final sample per species. After mRNA isolation using polyA Magnetic Isolation Module, cDNA libraries were prepared using the TruSeq stranded mRNA preparation kit and cDNA was sequenced on one lane of an Illumina HiSeq 2500 machine, producing strand oriented read pairs (2x125bp).

For *I. uriae*, ticks were sampled on Hornoya (Norway, 70°22'N, 31°10'E) in a colony of common murres (*Uria aalge*) in 2016. Ticks were kept alive at 4°C until RNA extraction. Just before RNA extraction, ticks were frozen at -80°C during 24H. Extraction was performed using the Qiagen RNeasy mini kit. Adults were

Table 4.1: Information on the nine collected species of *Ixodes*. The subgenus level is based on either Clifford et al. 1973[17] or Camicas et al. 1998 [32]. Localizations correspond to the name of the country followed by the name of the locality. Composition lists the life stage and physiological state of ticks used to generate the library: FU-female unengorged; FE-female engorged; M-male; NU-nymph unengorged; NE-nymph engorged.

Name	Sub-genus	Host	Localization	Composition
I. acuminatus	*Ixodes* Latreille, 1795	Wood mouse (*Apodemus silvaticus*)	France, Ille et Vilaine, Pleine-Fougères	10 FU, 10M, 49NU, 45NE
I. arboricola	*Pholeoixodes* Shulze, 1942	Great tit (*Parus major*)	Belgium, Antwerp	6FU, 3FE, 5M, 17NE
I. canisuga	*Pholeoixodes* Shulze, 1942	Red fox (*Canis vulpes*)[2]	France, Maine et Loire, Vendée, Grotte de Chanzelles, Saint Georges, Saint-Michel Le Cloud	6FU, 3FE, 5M, 7NU
I. frontalis	*Trichotoixodes*[1] Reznik, 1961	Black bird (*Turdus-merula*)	France, Loire-Atlantique, Nantes	5FE, 9NE
I. hexagonus	*Pholeoixodes* Shulze, 1942	European hedgehog (*Erinaceus europeus*)	France, Loire-Atlantique, Nantes	2FU, 1FE, 15NU, 5NE
I. holocyclus	*Sternalixodes* Schulze, 1935	Questing ticks	Australia, Queensland	10 FU
I. uriae	*Ceratixodes* Neumann, 1902	Common murre (*Uria aalge*)	Norway, Hornoya	FU, M, NU, NE[3]
I. ventalloi	ungrouped	European rabbit (*Oryctolagus cuniculus*)	France, Morbihan, Ile d'Houat	4FU, 3FE
I. vespertilionis	*Eschatocephalus* von Frauenfeld, 1853	Lesser horseshoe bat (*Rhinolophus hipposideros*)[2]	France, Maine et Loire, Vendée, Grotte de Chanzelles, Saint Georges, Saint-Michel Le Cloud	4FU

[1]: the only discrepancy in subgenus name for the listed species is for *Ixodes frontalis* that was considered as belonging to the *Ixodes* subgenus by Clifford et al. 1973 [17] although these authors consider that "It should be noted that species of the frontalis group do provide many exceptions in the definitions of the subgenus *Ixodes*").
[2]: inferred host. Ticks were collected on the soil or on the walls of caves when the mammal species was known to be present.
[3] : For *I. uriae*, different libraries were made for each condition.

extracted by pools of five and nymphs by pools of ten individuals (2 pools of unfed nymphs, one pool of partially engorged nymphs). A DNase treatment was applied following the protocol of Invitrogen TURBO™ DNase kit company inofrmations. Quality was assessed with the same methods employed for the eight other field-collected species. In contrast with the other species, the different conditions of *I. uriae* (unfed nymphs, unfed adult females, males, partially fed nymphs) were not mixed, a different library was generated for each of them). After mRNA isolation using polyA Magnetic Isolation Module, libraries were prepared using the TruSeq stranded mRNA preparation kit (libraries were obtained independently for each life stage or feeding status), and sequencing was done on one lane of a Hiseq3000 at the Get-Plage platform (Toulouse) producing strand oriented read pairs (2x150 bp).

We also included in our analysis previously published sequences for three other *Ixodes* species (*I. ricinus*, *I. persulcatus* and *I. scapularis*) as well as 15 non-*Ixodes* ticks species. Although different types of sequences have been published, for homogeneity purposes we chose to analyse only RNAseq projects performed with the Illumina technology. This type of data was indeed the most common and yielded a high output for the highest number of species. The non-*Ixodes* species were chosen to reflect the diversity of genera in Metastriata ticks (12 species of five different genera being included) and also included outgroup species belonging to soft ticks (three species within the *Ornithodroros* genus). For these species, raw reads were downloaded from the Sequence Read Archive section of NCBI (see details in the supplementary table S1). The next steps (quality checks and de novo assembly) were applied to 26 of the species studied here, whereas for *I. ricinus*, we used an assembly previously obtained by our group with the same methods [33].

4.4.2 Quality check

Newly obtained sequences as well as reads obtained from NCBI (SRA division) were cleaned using the same automated approach. For all species, a first round of cleaning was done using `Trimmomatic-0.36`: this step removed adapters and trimmed sequences by quality, with a minimum length of trimmed sequences of 36 bp [34]. The resulting set of reads was checked with `FastQC` [35]. If necessary, adapter removal was subject to a second round of read cleaning by `Trimmomatic` until no sign of adapters could be detected from the FastQC report. Cleaned reads from the 9 newly sequenced species will be deposited into the SRA section of the NCBI under BioProject n° (pending).

4.4.3 *De-novo* transcriptome assembly

Resulting reads for each of the 27 species were *de-novo* assembled using a De Bruijn graph approach with `Trinity` (v2.2.0) [36]. For the data sets with strand-oriented sequencing, the corresponding option was taken into account in the assembly process. Coding sequences were retrieved using `Transdecoder.longOrfs` (3.0.1) [36]. To reduce redundancy resulting from alternative splicing, we proceeded in two steps. First, coding sequences were clusterized by `CD-HIT` (v4.6) with a 95% identity threshold at the amino acid level [37] and only the corresponding contigs were kept. Among this reduced data set, we extracted the longest contig by component of the Transcriptome-DeBruijn Graph (T-DBG) in order to reduce redundancy resulting from alternative splicing. Several statistics were computed at each step of the pipeline to detect errors and to establish quality measures of our assembly. We measured completeness using `BUSCO` (v2 with Arthropod database v1, n=2675 conserved elements) [38], as well as general assembly statistics (using a home-made R script available at http://www.github.com/npchar/Transcriptome-Assembly-Statistics/).

4.4.4 Orthologues predictions and gene matrix construction

Orthologous genes are defined as sequences from different species that are homologous and have diverged after speciation events. This excludes genes produced by ancient duplication events which could lead to erroneous inferences of the species tree. We used `SiLiX` 1.2.9 [39] to infer homology relationships from the result of an all-versus-all `blastp+` (v 2.3.0) [40] using 50 threads from the BIRD Computational Resources Center (Nantes, FR). We followed the guidelines of the HOGENOM database [41] and chose to define orthologues as genes present in a family only once per species (single copy orthologues, or SCO). The `SiLiX` approach needs 4 user-defined parameters to consider a blast hit as significant for defining homology and producing a family (minimum identity, overlap percentage, minimum length and partial overlap). To find the optimal combination of `SiLiX` parameters, we conducted 1050 different clusterings simulation varying the different parameters. We then searched which parameters optimized the number of SCO genes present in 100% of our species [42]. Based on the presence/absence of a gene in our 27 species, we defined 3 SCO matrices with different gene occupancy levels (100%, 75%, 50%) [7].

4.4.5 SCO alignments, saturation assessment and concatenation

For each matrix (differing by gene occupancy level), SCO genes were first aligned at the protein level using `clustalW` (v 2.1) [43]. For DNA-level analyses, protein sequence alignments were reported on DNA sequences using a home-made perl

script. Alignments were trimmed using `Gblocks` (v091b) [44], preserving the codon structure. In order to assess the quality of DNA alignments, we investigated potential saturation produced by hidden substitutions for every SCO alignment: i) we calculated the pairwise p-distance and a distance corrected by the TN93 model [45] and ii) we then tested if the p-distance reached or not a plateau when represented against the TN93 corrected distance. Sequences that showed a deviation from a linear relationship between the two distances ($R^2 < 0.7$) were discarded. For ML and Bayesian approaches, the resulting alignments were then concatenated to produce the 3 super-alignments (SCO present at least in 100%, 75%, or 50% of the 27 species). Finally, for the Quartet method only, we conducted analyses for two other data sets: i) DNA alignments containing only the first and second positions (considering that the third position is the most sensitive to saturation effects), and ii) amino acid alignments. The alignments used were the nucleotide alignments produced as described above, which were either translated or in which third positions were deleted.

4.4.6 Species tree inference

Species relationships were inferred with three different approaches: Maximum Likelihood inferences were made on the three concatenated supermatrices with `IQ-TREE` [46], with the best estimated partition sites [47] and under the best model of substitution per partition [48] as well as 1000 ultrafast bootstrap estimations [49]. Bayesian Inferences were made using a CAT-GTR model implemented in `Phylobayes` [50] removing constant sites, for the three supermatrices. The convergence of the Bayesian sampler was checked by comparing two independent runs for every supermatrices. We used `tracecomp` and `pbcomp` [51] to assess the convergence of the 8 parameters of the trace and the bipartition list (Effective size $\gg 100$ and relative difference of estimators $\ll 0.1$). Super-networks were constructed using a mixture of gene trees, based on a quartet approach as implemented in the `SuperQ` software [52]. For the super-network approach, **SCO** gene trees were independently inferred with `IQ-TREE` [46] under the best model of substitution [48] with 100 classic non-parametric bootstraps. The best model was chosen according to the Bayesian Information Criterion (BIC) associated with the different models tested [48]. The resulting trees were divided in quartets (subtree of 4 taxons) and assembled together with `SuperQ` [52] to produce a planar supernetwork with the `Gurobi` Optimizer. An advantage of this method is to scale input trees in order to keep information of branch lengths [52].

4.5 Results

4.5.1 Sequencing statistics

We produced new transcriptomes for nine species of *Ixodes* for this study. For *I. acuminatus*, *I. arboricola*, *I. canisuga*, *I. frontalis*, *I. hexagonus*, *I. holocyclus*, *I. ventalloi*, and *I. vespertilionis*, we obtained between 29.2 and 43.9M cleaned paired-end (PE) reads. We also generated 225.8M cleaned PE reads for *I. uriae*. For published data sets also used in this paper (concerning 18 species of ticks), the numbers of cleaned reads ranged between 6.8M (*Rhipicephalus appendiculatus*) and 204.8M (*Ornithodoros erraticus*). Since read length varied among these projects (between 76 and 300), this factor also contributed to the variation in output in terms of base-pairs sequenced. A detailed table of the Sequencing Reads Archive files used in this study is provided in Supplementary Table S1.

4.5.2 Assembly and gene prediction

Raw assemblies produced between 23,992 (*I. hexagonus*) and 517,072 contigs (*I. scapularis*), with an average of 30,425 contigs per species. We predicted between 4613 peptides for *I. hexagonus* and 289,763 peptides for *I. scapularis*, with an average of 96,268 predicted peptides per species. The compression step (clusterization of predicted peptides with cd-hit and selection of the longest ORF by component) reduced on average the number of predicted peptides by 61.3% (Table 4.2). Completeness of these reduced assemblies was assessed with the BUSCO approach (totalizing single copy and duplicated complete genes). This metric ranged between 4.86% for *I. hexagonus* and 77.53% for *I. uriae* with an average completeness of 51.56% (Suppl. Fig. S1).

While a tendency appeared when fitting a linear model explaining the variability of completeness by the number of cleaned read pairs, log10 transformed ($F_{(1,25)} = 4.23$, $p = 0.05$), this relationship became statistically significant when the length of the read was added as an explanatory variable ($F_{(2,24)} = 4.21$, $p = 0.027$). This multiple regression explained almost 20% of the variability of the completeness (adjusted $R^2 = 0.198$) (Suppl. Fig. S2) We note that the number of complete BUSCO genes tends to reach a plateau above 50M reads, indicating a saturation of the sequence information for larger data sets. For smaller data sets, completeness varied strongly among projects (possibly related to the diversity of conditions used, but also to variation in RNA quality). For example, we assume that a quality issue (of RNA, or during library preparation) explains the very low completeness of the *I. hexagonus* data set.

4.5.3 Orthologous identification

For 1050 different clustering scenarios generated by `SiLiX`, we observed an optimum of the number of SCO found in 100% of 27 species (SCO100) for an identity of 75%

Table 4.2: Sequencing and assembly statistics per species. Sp, Species names (species with * were added by this study). Depth is the number of reads used for the assembly. Third column is the number of predicted peptides after redundancy reduction by cd-hit (95% identity). BUSCO is the number of BUSCO genes found complete in the assembly. Last column is the percentage of completeness.

Sp	Depth	Peptides	BUSCO	percentBusco
A. americanum	56,543,545	27,189	922	34.47
A. maculatum	24,145,506	11,125	568	21.23
A. sculptum	47,874,567	22,578	982	36.71
D. andersoni	28,724,901	27,26	1,763	65.91
D. variabilis	35,862,432	27,066	1,806	67.51
Ha. flava	39,952,891	57,981	1,890	70.65
Hy. excavatum	65,365,211	14,505	661	24.71
*I. acuminatus**	14,621,31	20,250	764	28.56
*I. arboricola**	20,908,888	22,179	1,735	64.86
*I. canisuga**	19,483,281	15,283	881	32.93
*I. frontalis**	21,015,354	18,187	1,069	39.96
I. hexagonus	21,956,106	3,215	130	4.86
*I. holocyclus**	18,632,047	15,520	1,321	49.38
I. persulcatus	52,974,333	41,620	1,891	70.69
*I. ricinus**	81,436,349	71,841	1,936	72.37
I. scapularis	178,730,952	63,018	1,973	73.76
*I. uriae**	112,892,945	49,056	2,074	77.53
*I. ventalloi**	19,296,195	16,563	997	37.27
*I. vespertilionis**	16,817,412	21,090	1,133	42.36
O. erraticus	102,396,846	43,106	2,056	76.86
O. moubata	55,303,643	12,839	1,845	68.97
O. rostratus	26,091,718	25,454	1,830	68.41
R. annulatus	15,500,635	28,256	1,418	53.01
R. appendiculatus	6,091,347	35,728	1,689	63.14
R. microplus	31,366,350	23,471	926	34.62
R. pulchellus	27,645,802	46,076	1,219	45.57
R. zambeziensis	21,463,386	61,008	1,759	65.76

and an overlap of 75% (see Suppl. Fig. S3). Other parameters (partial overlap and minimum length) did not have an impact on the number of SCO100. These parameters were set to a partial overlap with a minimal length of 200 amino acid length and an overlap superior to 90%. For these parameters, `SiLiX` predicted 30 SCO shared by the 27 species, 502 SCO present at least in 75% of species, and 952 SCO present at least in 50% of species (Figure 4.1).

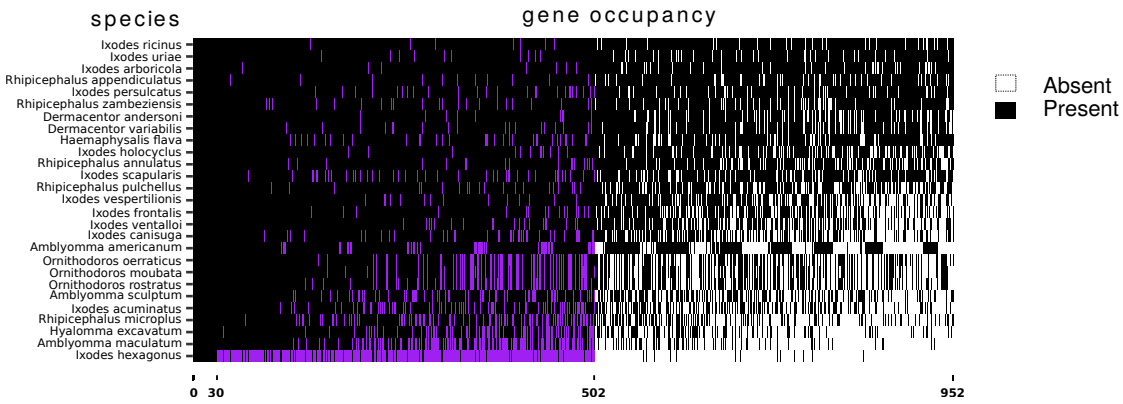

Figure 4.1: Single Copy Orthologues occupancy matrix, generated after Silix clustering. Each row represents a species (from top to bottom, species with a decreasing mean % of occupancy), whereas each column is a SCO (presence of the gene is indicated by a black-filled cell). Different levels of shades indicate occupancy level: at left (columns 1 to 30), SCO present in 100% of the species, in the center (columns 31 to 502) SCO present in less 100% but more than 75% of the species, at right (columns 503 to 952), SCO present in less than 75% but more than 50% of the species.

4.5.4 Alignment

Two genes from the SCO50 matrix showed signs of saturation (deviation from a linear relationship between p-distance and corrected distance by the TN93 model of substitution) and were discarded. Cleaned alignments will be soon deposited on Dryad/Zenodo, under a DOI number. The total numbers of nucleotide sites obtained for the three supermatrices (SCO100, SCO75, and SCO50) were respectively 20,094 (0% missing sites – 11,417 constant sites), 397,632 (11.7% missing sites – 221,141 constant sites), and 824,406 (25.5% missing sites – 460,607 constant sites). Finally, we calculated the percentage of missing data (in terms of sites) per species for the SCO50 supermatrices which ranged between 4.42% for *I. uriae* and 92.18% for *I. hexagonus*.

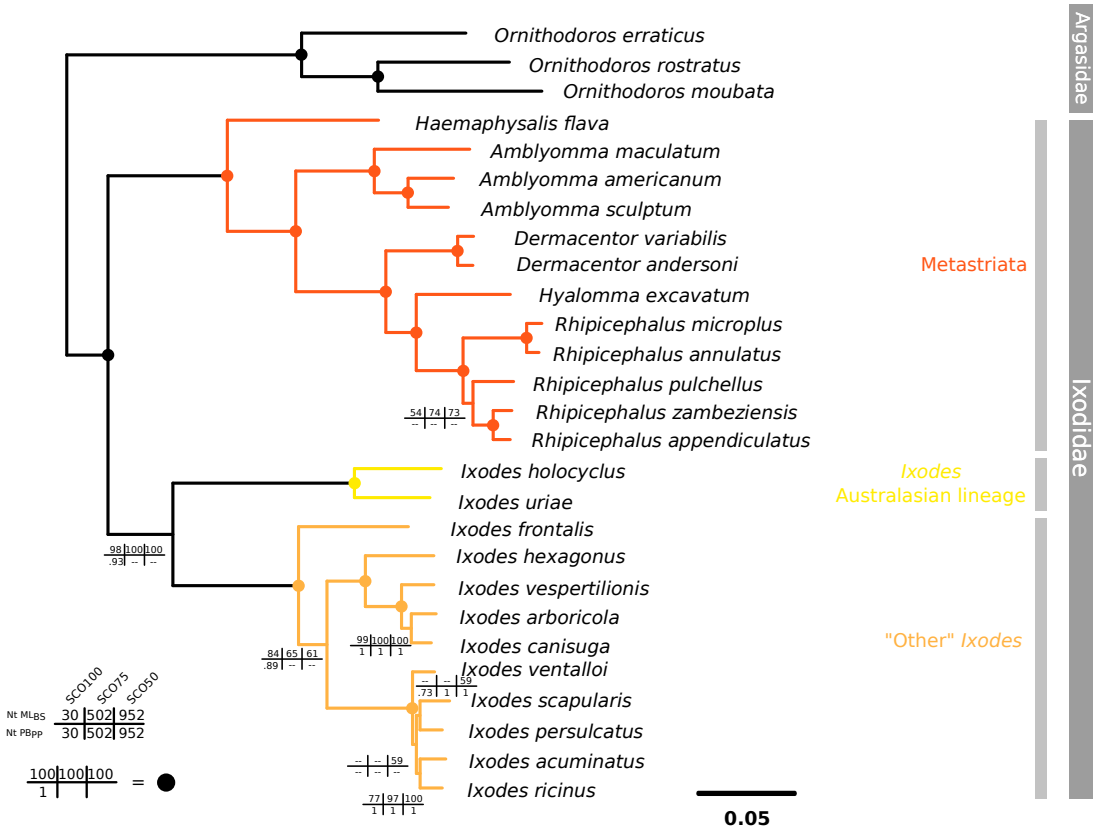

Figure 4.2: Phylogenetic tree based on the SCO50 supermatrix, mid-point rooted. Node with dots represent maximum support values (100% in all methods), otherwise bootstrap values and posterior probabilities were represented. These values are indicated for each of the SCO matrixes (occupancy level of 100%, 75%, or 50%): bootstrap values with the ML method on the top line, and posterior probabilities in the Bayesian approach on the bottom line.

4.5.5 Species tree inference

For both ML and Bayesian methods, we found the following topology (Figure 4.2): Metastriata appeared to form a robust clade (100% bootstrap or posterior probability support, with the three supermatrices). There was also a strong support for a monophyletic *Ixodes* genus in the ML method (all three matrices) and with the Bayesian approach (with the SCO100 matrix). But unexpectedly, the Bayesian approach gave different topologies for SCO75 and SCO50 matrices (based on more sites overall, but with some incompleteness), since in the case *Ixodes uriae* and *I. holocyclus* species grouped with the Metastriata (but with a low support value). With the Bayesian approach, branch length separating the three lineages (Metastriata, Australasian *Ixodes* and "other" *Ixodes*) was very short with 0.019 and 0.010 respectively for SCO75 and SCO50. The number of iteration, effective size and relative difference between chains for the different parameters used to check the convergence of the bayesian approach are available in Suppl. Table S2. With all methods, there was a clear separation of the *Ixodes* genus into two subclades, the Australasian lineage on one side (represented by *I. holocyclus* and *I. uriae*) and a second clade comprising all other species. Distances between these two clades (Australasian species versus other *Ixodes*) were only slightly below distances between a Metastriata and any *Ixodes* species (Table 4.3).

Within Metastriata, *Haemaphylasis flava* was found to be the most basal species followed by the *Amblyomma* species (this was supported in all analyses). We also found that *Hyalomma excavatum* grouped with the *Rhipicephalinae* subfamily (*Dermacentor* and *Rhipicephalus*). While *Rhipicephalus* species formed a robust clade, the position of *Rhipicephalus pulchellus* was not robust among methods, or among data sets.

Among the non-Australasian *Ixodes* lineage, the bird-associated species *I. frontalis* appeared to be basal (in all analyses). Then, the remaining species made two well-supported clades. The first one comprised four *Ixodes* species (*I. hexagonus*, *I. vespertilionis*, *I. arboricola*, *I. canisuga*), while the second comprised the five remaining species (*I. ventalloi*, *I. scapularis*, *I. persulcatus*, *I. acuminatus* and *I. ricinus*). These five species were particularly close to each other and we could only determine a robust grouping between *I. acuminatus* and *I. ricinus*.

Supernetworks were reconstructed for three different gene occupancy levels (100%, 75%, and 50%) and for three data sets: (i) nucleotide level alignments including all codon positions, (ii) nucleotide level alignments including only the first and second codon positions, and (iii) amino-acid level alignments. Supernetworks from the SCO75 matrix based on the nucleotide alignments with all codon positions was represented (Figure 4.3), as well as the nine different combinations (Suppl. Fig. S4). Supernetworks obtained with the three types of alignments showed the same topology (Figure 4.3, Suppl. Fig. S4). Essentially, supernetworks gave the

Table 4.3: Patristic distance across lineages. The upper-right corner indicate the mean patristic distance between species of different groups while the bottom-left indicate the standard deviation (when more than 2 species were present). Patristic distance are computed from the ML consensus tree of the SCO50 data set.

Lineage	*Amblyomma*	*Dermacentor*	*Haemaphysalis*	*Hyalomma*	*Ixodes*	Aust.*Ixodes*	*Ornithodoros*	*Rhipicephalus*
Amblyomma		0.174	0.196	0.193	0.348	0.348	0.426	0.200
Dermacentor	0.005		0.204	0.109	0.356	0.356	0.434	0.116
Haemaphysalis	0.006	>0.001		0.223	0.307	0.307	0.386	0.230
Hyalomma	0.006	>0.001	n.a		0.375	0.375	0.453	0.103
"other-*Ixodes*"	0.007	0.006	0.006	0.006		0.270	0.415	0.382
Australasian *Ixodes*	0.006	0.003	0.004	0.004	0.006		0.415	0.382
Ornithodoros	0.018	0.018	0.020	0.020	0.017	0.018		0.460
Ripicephalus	0.009	0.008	0.008	0.008	0.009	0.008	0.018	

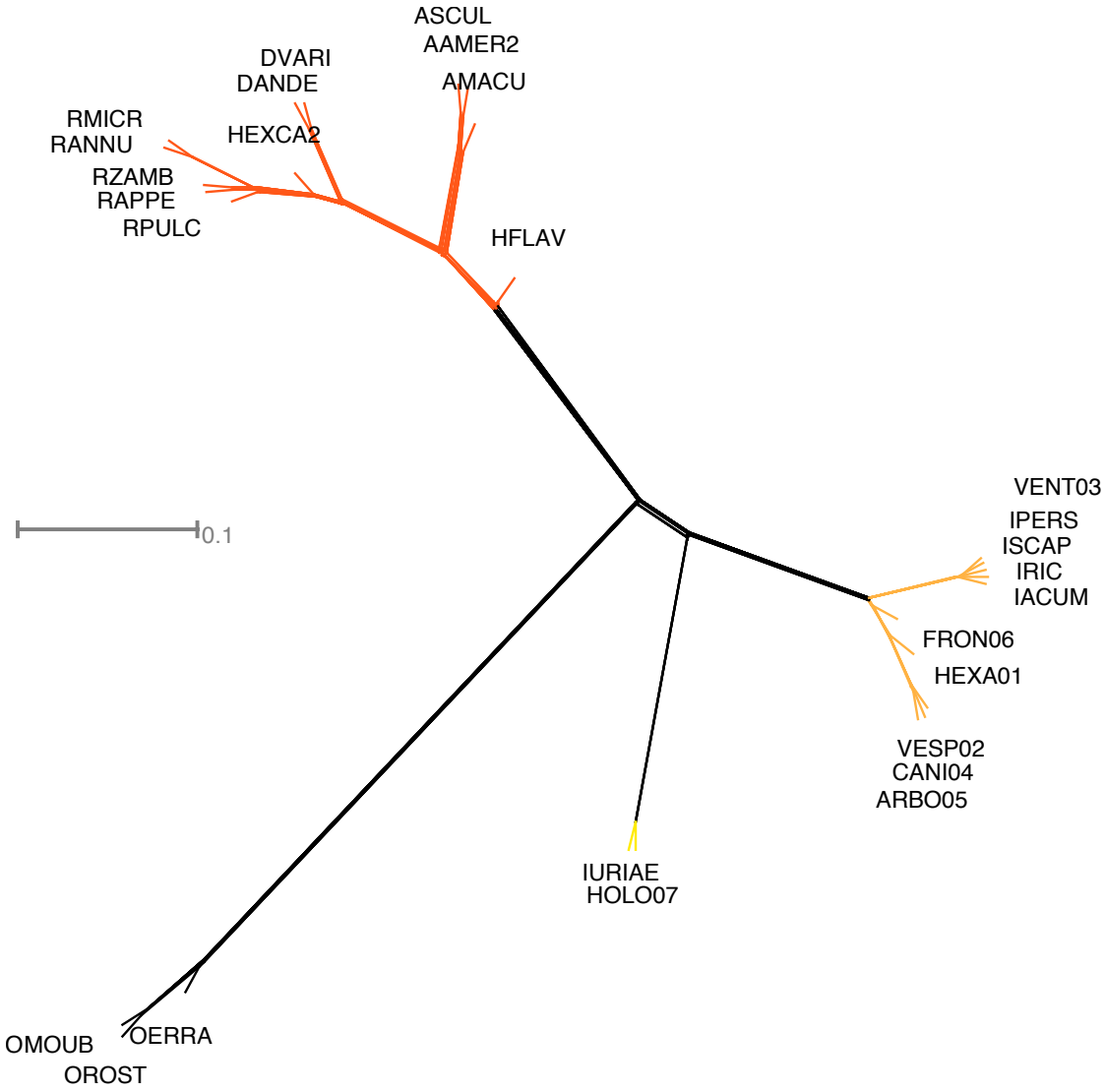

Figure 4.3: Supernetwork based on genes from the SCO75 matrix (occupancy level of 75% among 27 tick species). Supernetwork was built from nucleotidic alignments with all codon positions (see Suppl. Fig. S4 for the different supernetworks reconstructions).

same topology as the ML approach (and generally also as the Bayesian approach). We observed some level of reticulation, mostly for the data set with third positions removed, and for the amino acid data set. By contrast, the nucleotidic alignments resulted in supernetworks with less evidence of reticulation (Suppl. Fig. S4).

4.6 Discussion

Our study aimed to obtain a better phylogenetic resolution of hard tick, with a focus on the Prostriata (the *Ixodes* genus), by using RNAseq based phylogeny to provide a large number of common markers. To this end we produced transcriptome sequences for 9 species of *Ixodes*, therefore signficantly enriching the taxonomic spectrum of RNAseq projects for this genus (note that while we were conducting our project, an independant project for one of the species, *I. holocyclus* was also published [53]; for the eight other species, our data represent the first high throughput sequence projects). We also took advantage of previously published RNAseq data sets, principally for the Metastriata. For the first time, we used these data together in a phylogenetic framework and were able to obtain a robust phylogeny of the group.
The resulting topology was consistent with previous studies but supported by several hundred nuclear genes. The Maximum Likelihood approach on the three supermatrices with different levels of missing data were consistent with each other, especially for deepest nodes. These trees all supported the monophyly of the *Ixodes* genus. On this point however, the Bayesian approach did not produce a consistent answer. Indeed, for the SCO75 and SCO50 supermatrices, the "other" *Ixodes* lineage was a sister group to the rest of hard ticks (Metastriata). But this result was associated with very short internal branch (separating Metastriata, Australasian and "other-*Ixodes*" lineages) leading us to interpret this result as an unresolved relationship with the CAT-GTR model. All the supernetworks were essentially identical, despite the fact that they were built with different types of alignments. All three approaches (ML, Bayesian, supernetworks) resulted in two highly divergent groups of *Ixodes* (the Australasian lineage and the other species).
Our study did not include mitochondrial genes, which are often used in phylogenetic studies, either as individual markers, or as entire mitogenomes [25, 54, 55]. As pointed by Mans et al [24], the polyphyletic nature of the *Ixodes* genus is generally found with nuclear genes while phylogenetic reconstruction based on mithochondrial marked are in favor of monophyly. Despite the use of nuclear markers only, our results agree with results from mitochondrial gene concatenation [25, 54, 55]. It has been found by comparing the structure of mitogenomes that the Australasian *Ixodes* lineage and Metastriata share the presence of two large non coding regions [56]. This could be taken as evidence of a common ancestry, but this character is not necessarily synapomorphic and could haved occured independently in Metastriata and Australasian *Ixodes*. Our study therefore supports the mono-

phyly of *Ixodes* (including the Australasian lineages), although other approaches (based on mitogenomes) indicated another scenario (but with a weak support). This precise point cannot therefore be fully resolved for the present time, as this might require even larger data sets (based on a larger taxonomic sampling and/or using complete genomes). Whichever of these two scenarios is finally validated, our present work shows that three groups, the Metastriata (non-*Ixodes* hard tick species), the Australasian *Ixodes* lineage, and the "other" *Ixodes* species diverged from each other within a very short amount of time (i.e. this branching is close to be a trifurcation). These observations suggest that Australasian *Ixodes* species should be erected as a distinct genus.

Inside the "Other" *Ixodes* lineage, our results highlight the extreme closeness of five species (*I. ventalloi*, *I. ricinus*, *I. acuminatus*, *I. persulcatus* and *I. scapularis*). Initially the expression "ricinus complex" was used to describe three of these five species (namely *I. ricinus*, *I. persulcatus*, *I. scapularis*). This term was introduced to describe species sharing similar morphology, life-styles and competency for transmitting Lyme spirochaetes (see [57]). Because of the presence of closely related species unable to transmit such pathogen [58], this term was progressively abandoned [59]. However, the term "ricinus group" is still used [60–62] to describe the fact that many *Ixodes* species are sharing interesting traits and are closely related. We argue that this term could be reintroduced to describe this sub-clade of five species within the genus, as well as their close parents. The presence of *Ixodes ventalloi* within that subclade (then close to *I. ricinus*) is coherent with the recent work describing two *I. ventalloi* genotypes [63] as well as the redescription of a neotype [61]. The relative position of the different species in this complex seems to indicate that the host spectrum is a fast evolving trait. Indeed *I. ventalloi* and *I. acuminatus* are considered more specialist (being respectively associated with rabbits and small rodents) whereas the three other species of that subclade, *I. ricinus*, *I. persulcatus*, and *I. scapularis*, are considered as generalists.

Recently, phylogenetic reconstruction from whole mitochondrial genomes confirmed that the *Amblyomma* genus is polyphyletic with some highly divergent species [54]. Although, the sampled *Amblyomma* species in this study did not contain *A. sphaenodonti* and *A. elaphense* [55] – we found that *Amblyomma sensu stricto* formed a well-supported clade in all the results. Which genera from *Amblyomma* or *Haemaphysalis* is basal among Metastriata remains ambiguous: classically *Amblyomma* is placed basally among Metastriata [14, 15, 20, 24, 64, 65] but recent studies placed the *Amblyomma sensu stricto* genus less basally than the *Haemaphysalinae* together with the *Bothriocrotinae* [22, 25, 54, 55]. Our results find that the most basal species were *Haemaphysalinae* instead of the clade formed by *Amblyomma sensu stricto* species. We also found a strong support for the *Rhipicephalinae* subfamily; that is formed by the five *Rhipicephalus* species, the *Hyalomma* species and the three *Dermacentor* species.

Among the new data sets produced for this study, one of them (*I. hexagonus*) yielded a comparatively low number of transcripts and of conserved genes (genes used in the SCO matrixes). Many questions were raised about missing data in phylogenetic reconstruction at the present time of high throughput sequencing era [66]. Despite the relatively poor result for the *I. hexagonus* data set (which we can not fully explain), the phylogenetic position of this species was very robust. The position of *I. hexagonus* was coherent with a recent study of the *Pholeoixodes* subgenus [19]. Indeed, *I. hexagonus* feeds preferentially on *canidae* (as well as *I. canisuga*) and was found to be phylogenetically close to the *Eschatocephalus* subgenus represented here by *I. vespertilionis*, a bat tick [19]. Because we found the same conclusion as the previous study [19], we argue that the lack of assembled transcripts for this species was not a problem in this precise case. Another relatively incomplete transcriptome was that of *Hy. excavatum* (despite a high sequencing depth). This transcriptome was therefore less complete than that species with similar quantity of paired-end reads used in the de-novo assembly process. The specific origin of the tissue (salivary gland only [67]) could explain the low diversity and completeness of transcripts of this data set. Again, our methods allowed however a robust placement of this species with respect to *Rhipicephalinae*.

The method used to detect orthologues was free from topology-based assumptions, defined as a pairwise based one-to-one strategy of orthology prediction [68, 69]. We assume that this strategy is conservative and decreased the occupancy of the SCO matrix but, in our case, still showed overall good results. We may have expected that the different data sets (with all codon positions, first and second positions, or aminoacids) would provide different topologies at different depths of the phylogenetic tree. Interestingly, supernetwork topologies were identical overall although the level of reticulation was higher for aminoacid alignments than for nucleotide alignments with all codon positions.
Unfortunately, no RNAseq data is yet available for the *Nuttalliallidae* species, which could provide an interesting taxon to decipher the relative position of the different tick lineages and provide clues about the evolution of haematophagy [24]. Indeed, this particular species was seen as the "closest living relative to the ancestral tick lineage" [70]. Molecular analyses support 4 scenarios [24], as member of the hard tick group by the concatenation of 10 mitochondrial genes [25], poorly supported as sister group of the soft tick [55], as basal to hard and soft ticks [26, 71] and even placed between the Holothyrida and the Mesostigmata [25]. The inclusion of RNAseq data for some representant of the *Bothriocrotinae* (such data is not yet available) could also help to better reconstruct the evolutionary tree of the hard ticks. Indeed, this subfamily is expected to be basal in the metastriate diversification [24].

By sequencing RNAseq data for 9 new *Ixodes* species, we are bringing precious resources, with new gene catalogues for this group, comprising both genes shared among ticks and species-specific genes. In future studies, this data could be used

to study the evolution of gene families [24, 59, 72], as a backbone to explore the impact of tick life-styles on gene evolution.

4.7 Supplementary materials

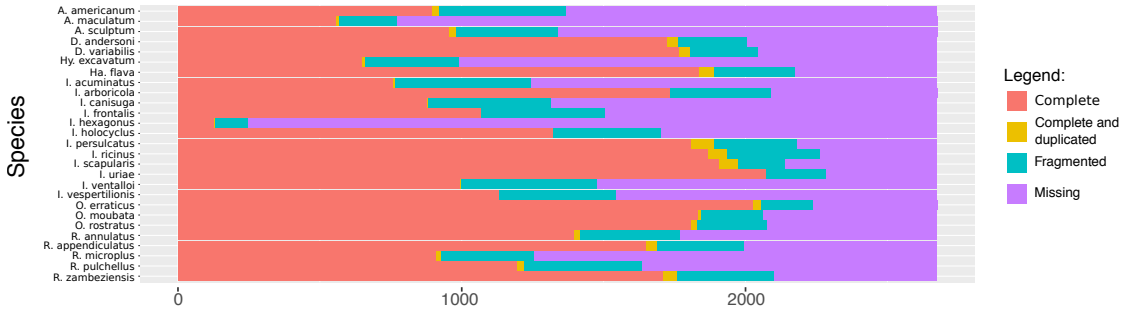

Figure S1: Completeness summarized for the 27 assemblies. Species are represented in row and different color represent the proportion of BUSCO genes (total of 2675 genes) classified into four categories: i) found complete in one copy (Complete), ii) found complete in more than one copy (Complete and Duplicated), iii) Found but incomplete or fragmented (Fragmented), and iv) Not found (Missing).

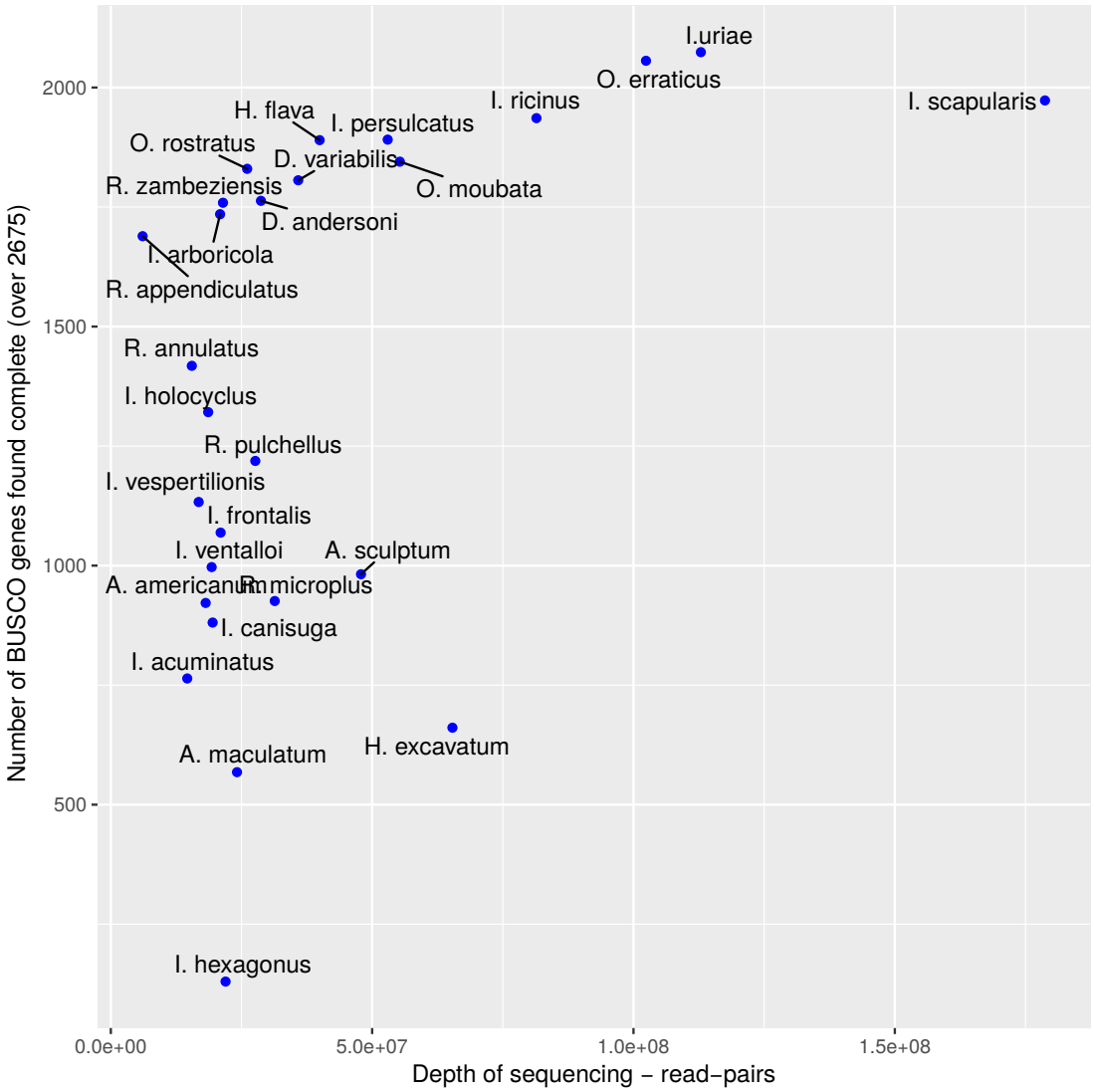

Figure S2: Completeness as a function of depth of sequencing. Y-axis, number of complete BUSCO genes (Complete or Duplicated), x-axis, number of pair of reads used for de-novo assembly.

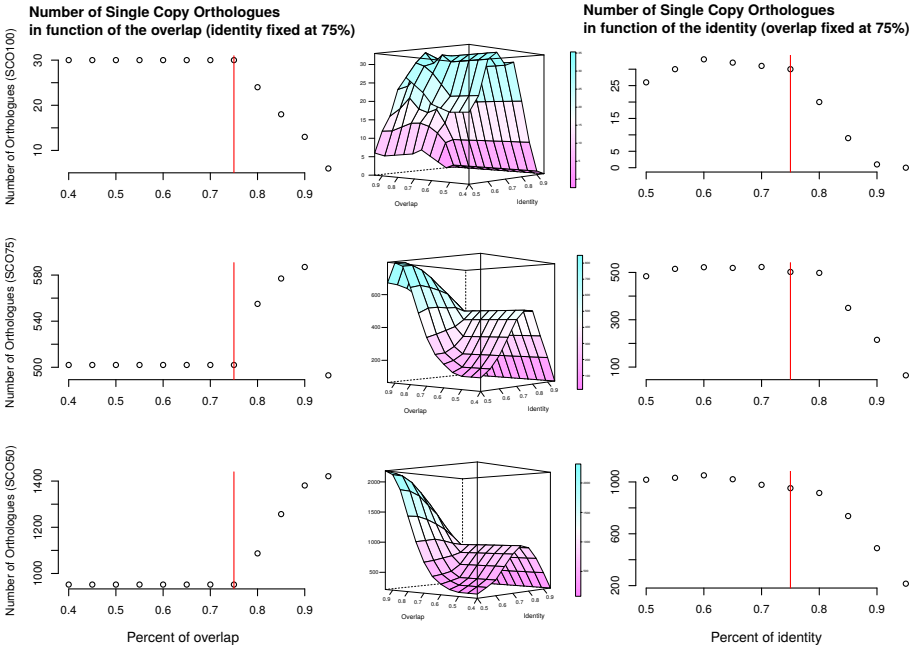

Figure S3: Number of SCO detected by `SiLiX` for different overlap and identity parameters. Each line represents the the Number of SCO detected at three threshold of occupancy (respectively for 100%, 75% and 50%). The first column represents the number of SCO as a function of the overlap for an identity fixed at 75%. The second column represent the number of SCO as a function of the identity and the overlap. The third column represents the number of SCO as a function of the identity for an overlap fixed at 75%. A red line represents the chosen parameters for the final SCO detection (identity of 75%, overlap of 75%).

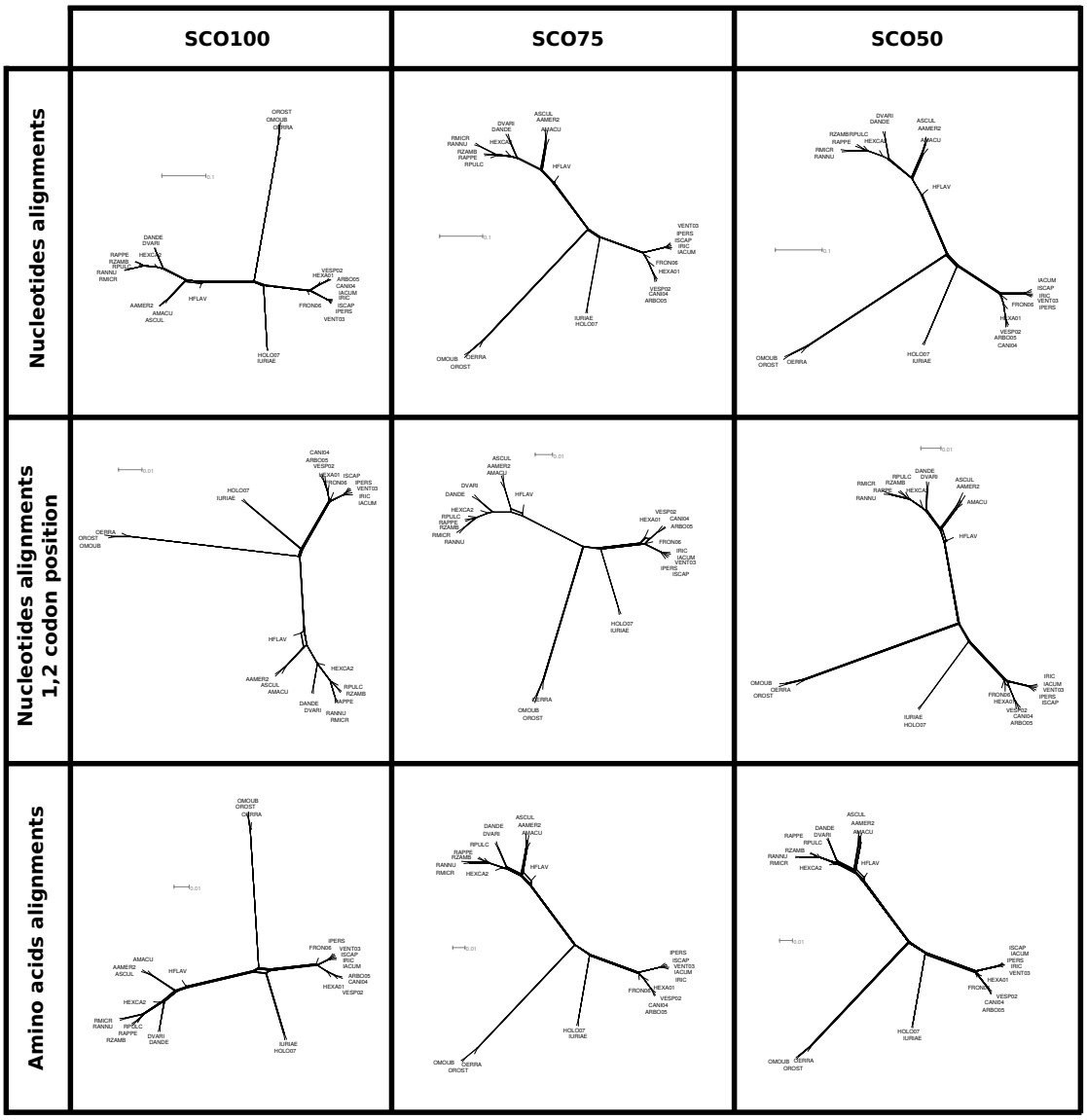

Figure S4: Supernetworks from the different source and type of data set. In columns, the three supermatrixes used to reconstruct the gene trees with different gene occupancy thresholds (SCO100, 75 and 50); in rows, the type of alignments used (nucleotides alignments, third codon positions removed, and amino-acids alignments).

Bibliography

1. N. S. Geraci, J. S. Johnston, J. P. Robinson, S. K. Wikel, and C. A. Hill: Variation in genome size of argasid and ixodid ticks. *Insect Biochemistry and Molecular Biology* **37**(5) (2007), 399–408. doi: 10.1016/j.ibmb.2006.12.007.
2. M. Gulia-Nuss et al.: Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. *Nature Communications* **7** (Feb. 2016), 10507. doi: 10.1038/ncomms10507.
3. A. J. Ullmann, C. M. R. Lima, F. D. Guerrero, J. Piesman, and W. C. Black: Genome size and organization in the blacklegged tick, *Ixodes scapularis* and the Southern cattle tick, *Boophilus microplus*. *Insect Molecular Biology* **14**(2) (Apr. 2005), 217–222. doi: 10.1111/j.1365-2583.2005.00551.x.
4. Y. Yang and S. A. Smith: Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* **14**(1) (2013), 328. doi: 10.1186/1471-2164-14-328.
5. F. Delsuc, H. Brinkmann, and H. Philippe: Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* **6**(5) (May 2005), 361–375. doi: 10.1038/nrg1603.
6. O. Jeffroy, H. Brinkmann, F. Delsuc, and H. Philippe: Phylogenomics: the beginning of incongruence? *Trends in Genetics* **22**(4) (2006), 225–231. doi: 10.1016/j.tig.2006.02.003.
7. V. L. González et al.: A phylogenetic backbone for Bivalvia: an RNA-seq approach. *Proceedings of the Royal Society B* **282**(1801) (2015), 20142332. doi: 10.1098/rspb.2014.2332.
8. P. P. Sharma et al.: Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal. *Molecular Biology and Evolution* **31**(11) (2014), 2963–2984. doi: 10.1093/molbev/msu235.
9. R. Fernández, G. D. Edgecombe, and G. Giribet: Phylogenomics illuminates the backbone of the Myriapoda Tree of Life and reconciles morphological and molecular phylogenies. *Scientific Reports* **8**(1) (Jan. 2018). doi: 10.1038/s41598-017-18562-w.
10. R. S. Peters et al.: Evolutionary History of the Hymenoptera. *Current Biology* **27**(7) (2017), 1013–1018. doi: 10.1016/j.cub.2017.01.027.
11. R. Fernández and G. Giribet: Unnoticed in the tropics: phylogenomic resolution of the poorly known arachnid order Ricinulei (Arachnida). *Royal Society open science* **2**(6) (2015), 150065. doi: 10.1098/rsos.150065.

12. G.-H. Lin et al.: Transcriptome sequencing and phylogenomic resolution within Spalacidae (Rodentia). *BMC Genomics* **15** (Jan. 2014), 32. doi: 10.1186/1471-2164-15-32.

13. A. A. Guglielmone et al.: *The hard ticks of the world*. Springer, 2014, pp. 978–94. doi: 10.1007/978-94-007-7497-1.

14. W. C. Black and J. Piesman: Phylogeny of hard- and soft-tick taxa (Acari: Ixodida) based on mitochondrial 16S rDNA sequences. *Proceedings of the National Academy of Sciences* **91**(21) (Oct. 1994), 10034–10038. doi: 10.1073/pnas.91.21.10034.

15. S. C. Barker and A. Murrell: Phylogeny, evolution and historical zoogeography of ticks: a review of recent progress. *Experimental and Applied Acarology* **28**(1-4) (2002), 55–68. doi: 10.1007/978-94-017-3526-1_3.

16. H. Hutcheson et al.: Current progress in tick molecular systematics. *3rd International Conference On Ticks and Tick-Borne Pathogens: Into the 21st Century, Proceedings*. 2000, 11–19.

17. C. M. Clifford, D. E. Sonenshine, J. E. Keirans, and G. M. Kohls: Systematics of the subfamily Ixodinae (Acarina: Ixodidae). 1. the subgenera of *Ixodes*. *Annals of the Entomological Society of America* **66**(3) (1973), 489–500. doi: 10.1093/aesa/66.3.489.

18. A. Estrada-Peña, A. D. Mihalca, and T. N. Petney: *Ticks of Europe and North Africa: A Guide to Species Identification*. Springer, 2018. doi: 10.1007/978-3-319-63760-0.

19. S. Hornok et al.: Contributions to the phylogeny of *Ixodes* (*Pholeoixodes*) *canisuga*, *I.* (*Ph.*) *kaiseri*, *I.* (*Ph.*) *hexagonus* and a simple pictorial key for the identification of their females. *Parasites & vectors* **10**(1) (Nov. 2017), 545. doi: 10.1186/s13071-017-2424-x.

20. S. C. Barker and A. Murrell: Systematics and evolution of ticks with a list of valid genus and species names. *Parasitology* **129 Suppl** (2004), S15–36. doi: 10.1017/S0031182004005207.

21. J. Klompen: Phylogenetic relationships in the family Ixodidae with emphasis on the genus *Ixodes* (Parasitiformes: Ixodidae). *Acarology IX Symposia*. Ohio State University Ohio, USA. 1999, 349–354.

22. J. Klompen, W. C. Black, J. E. Keirans, and D. E. Norris: Systematics and Biogeography of Hard Ticks, a Total Evidence Approach. *Cladistics* **16**(1) (Mar. 2000), 79–102. doi: 10.1111/j.1096-0031.2000.tb00349.x.

23. J. S. H. Klompen, W. C. Black, J. E. Keirans, and J. H. Oliver: Evolution of Ticks. *Annual Review of Entomology* **41**(1) (Jan. 1996), 141–161. doi: 10.1146/annurev.en.41.010196.001041.

24. B. J. Mans et al.: Ancestral reconstruction of tick lineages. *Ticks and Tick-borne Diseases* **7**(4) (2016). TTP8-STVM Special Issue, 509–535. doi: 10.1016/j.ttbdis.2016.02.002.

25. T. D. Burger, R. Shao, M. B. Labruna, and S. C. Barker: Molecular phylogeny of soft ticks (Ixodida: Argasidae) inferred from mitochondrial genome and nuclear rRNA sequences. *Ticks and Tick-borne Diseases* **5**(2) (2014), 195–207. doi: 10.1016/j.ttbdis.2013.10.009.

26. B. J. Mans, D. de Klerk, R. Pienaar, M. H. de Castro, and A. A. Latif: Next-generation sequencing as means to retrieve tick systematic markers, with the focus on *Nuttalliella namaqua* (Ixodoidea: Nuttalliellidae). *Ticks and Tick-borne Diseases* **6**(4) (2015), 450–462. doi: 10.1016/j.ttbdis.2015.03.013.

27. P. D. Hillyard et al.: *Ticks of north-west Europe.* Field Studies Council, 1996.

28. C. Pérez-Eid: *Les tiques: identification, biologie, importance médicale et vétérinaire.* Lavoisier, 2007.

29. S. C. Barker and A. R. Walker: Ticks of Australia. The species that infest domestic animals and humans. *Zootaxa* **3816**(1) (2014), 1–144. doi: 10.11646/zootaxa.3816.1.1.

30. D. Heylen, E. D. Coninck, F. Jansen, and M. Madder: Differential diagnosis of three common *Ixodes spp.* ticks infesting songbirds of Western Europe: *Ixodes arboricola*, *I. frontalis* and *I. ricinus*. *Ticks and Tick-borne Diseases* **5**(6) (2014), 693–700. doi: 10.1016/j.ttbdis.2014.05.006.

31. D. Heylen and E. Matthysen: Contrasting detachment strategies in two congeneric ticks (Ixodidae) parasitizing the same songbird. *Parasitology* **137**(4) (2010), 661–667. doi: 10.1017/S0031182009991582.

32. J.-L. Camicas, J.-P. Hervy, F. Adam, P. C. Morel, et al.: *The ticks of the world (Acarida, Ixodida): nomenclature, described stages, hosts, distribution.* Éditions de l'ORSTOM, 1998.

33. N. P. Charrier et al.: Whole body transcriptomes and new insights into the biology of the tick *Ixodes ricinus*. *Parasites & Vectors* **11**(1) (June 2018). doi: 10.1186/s13071-018-2932-3.

34. A. M. Bolger, M. Lohse, and B. Usadel: Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15) (2014), 2114–2120. doi: 10.1093/bioinformatics/btu170.

35. S. Andrews: *FastQC A Quality Control tool for High Throughput Sequence Data.* url: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

36. B. J. Haas et al.: De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**(8) (2013), 1494–1512. doi: 10.1038/nprot.2013.084.

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

38. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov: BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**(19) (June 2015), 3210–3212. doi: 10.1093/bioinformatics/btv351.

39. V. Miele, S. Penel, and L. Duret: Ultra-fast sequence clustering from similarity networks with `SiLiX`. *BMC bioinformatics* **12**(1) (2011), 116. doi: 10.1186/1471-2105-12-116.

40. C. Camacho et al.: BLAST+: architecture and applications. *BMC Bioinformatics* **10**(1) (2009), 421. doi: 10.1186/1471-2105-10-421.

41. S. Penel et al.: Databases of homologous gene families for comparative genomics. *BMC bioinformatics*. **10**. 6. BioMed Central. 2009, S3. doi: 10.1186/1471-2105-10-S6-S3.

42. T. Lefébure et al.: Less effective selection leads to larger genomes. *Genome Research* **27**(6) (Apr. 2017), 1016–1028. doi: 10.1101/gr.212589.116.

43. J. D. Thompson, D. G. Higgins, and T. J. Gibson: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**(22) (1994), 4673–4680. doi: 10.1093/nar/22.22.4673.

44. J. Castresana: Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution* **17**(4) (Apr. 2000), 540–552. doi: 10.1093/oxfordjournals.molbev.a026334.

45. K. Tamura and M. Nei: Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution* **10**(3) (1993), 512–526. doi: 10.1093/oxfordjournals.molbev.a040023.

46. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh: IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**(1) (Nov. 2014), 268–274. doi: 10.1093/molbev/msu300.

47. O. Chernomor, A. von Haeseler, and B. Q. Minh: Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic biology* **65**(6) (Nov. 2016), 997–1008. doi: 10.1093/sysbio/syw037.

48. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin: ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* **14**(6) (June 2017), 587–589. doi: 10.1038/nmeth.4285.

49. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh: UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular biology and evolution* **35**(2) (Feb. 2018), 518–522. doi: 10.1093/molbev/msx281.

50. N. Lartillot and H. Philippe: A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution* **21**(6) (June 2004), 1095–109. doi: 10.1093/molbev/msh112.

51. N. Lartillot, T. Lepage, and S. Blanquart: PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**(17) (Sept. 2009), 2286–8. doi: 10.1093/bioinformatics/btp368.

52. S. Grünewald, A. Spillner, S. Bastkowski, A. Bögershausen, and V. Moulton: SuperQ: computing supernetworks from quartets. *IEEE/ACM Trans Comput Biol Bioinform* **10**(1) (2013), 151–60. doi: 10.1109/TCBB.2013.8.

53. M. Rodriguez-Valle et al.: Transcriptome and toxin family analysis of the paralysis tick, *Ixodes holocyclus*. *International Journal for Parasitology* **48**(1) (2018), 71–82. doi: 10.1016/j.ijpara.2017.07.007.

54. T. D. Burger, R. Shao, L. Beati, H. Miller, and S. C. Barker: Phylogenetic analysis of ticks (Acari: Ixodida) using mitochondrial genomes and nuclear rRNA genes indicates that the genus *Amblyomma* is polyphyletic. *Molecular Phylogenetics and Evolution* **64**(1) (2012), 45–55. doi: 10.1016/j.ympev.2012.03.004.

55. T. D. Burger, R. Shao, and S. C. Barker: Phylogenetic analysis of the mitochondrial genomes and nuclear rRNA genes of ticks reveals a deep phylogenetic structure within the genus *Haemaphysalis* and further elucidates the polyphyly of the genus *Amblyomma* with respect to *Amblyomma sphenodonti* and *Amblyomma elaphense*. *Ticks and Tick-borne Diseases* **4**(4) (2013), 265–274. doi: 10.1016/j.ttbdis.2013.02.002.

56. Mitochondrial genomes of parasitic arthropods: implications for studies of population genetics and evolution. *Parasitology* **134**(02) (Feb. 2007), 153. doi: 10.1017/s0031182006001429.

57. J. Keirans, G. Needham, and J. Oliver Jr: The *Ixodes ricinus* complex worldwide: diagnosis of the species in the complex, hosts and distribution. *Acarology IX* **2** (1999), 341–347.

58. G. Xu, Q. Q. Fang, J. E. Keirans, and L. A. Durden: Molecular phylogenetic analysis indicate that the *Ixodes ricinus* complex is a paraphyletic group. *Journal of Parasitology* **89**(3) (June 2003), 452–457. doi: 10.1645/0022-3395(2003)089[0452:mpaitt]2.0.co;2.

59. V. Daix et al.: Ixodes ticks belonging to the *Ixodes ricinus* complex encode a family of anticomplement proteins. *Insect Molecular Biology* **16**(2) (Apr. 2007), 155–166. doi: 10.1111/j.1365-2583.2006.00710.x.

60. I. A. Akimov and I. V. Nebogatkin: Ixodid Ticks (Acari, Ixodidae) in Urban Landscapes. A review. *Vestnik Zoologii* **50**(2) (Apr. 2016), 155–162. doi: 10.1515/vzoo-2016-0018.

61. A. Estrada-Peña, J. M. Venzal, and S. Nava: Redescription, molecular features, and neotype deposition of *Rhipicephalus pusillus* Gil Collado and *Ixodes ventalloi* Gil Collado (Acari, Ixodidae). *Zootaxa* **4442**(2) (July 2018), 262. doi: 10.11646/zootaxa.4442.2.4.

62. S. Kovalev, S. Fedorova, and T. Mukhacheva: Molecular features of *Ixodes kazakstani*: first results. *Ticks and Tick-borne Diseases* **9**(3) (2018), 759–761. doi: https://doi.org/10.1016/j.ttbdis.2018.02.019.

63. M. S. Latrofa et al.: *Ixodes ventalloi*: morphological and molecular support for species integrity. *Parasitology Research* **116**(1) (Jan. 2017), 251–258. doi: 10.1007/s00436-016-5286-9.

64. H. Hoogstraal and A. Aeschlimann: Tick-host specificity. *Bulletin de la société entomologique suisse* **55** (1982), 5–32.

65. H. Hoogstraal: Argasid and nuttalliellid ticks as parasites and vectors. *Adv Parasitol* **24** (1985), 135–238.

66. B. Roure, D. Baurain, and H. Philippe: Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular biology and evolution* **30**(1) (Jan. 2013), 197–214. doi: 10.1093/molbev/mss208.

67. J. M. Ribeiro, M. Slovák, and I. M. Francischetti: An insight into the sialome of *Hyalomma excavatum*. *Ticks and Tick-borne Diseases* **8**(2) (2017), 201–207. doi: 10.1016/j.ttbdis.2016.08.011.

68. T. Gabaldón: Large-scale assignment of orthology: back to phylogenetics? *Genome Biology* **9**(10) (2008), 235. doi: 10.1186/gb-2008-9-10-235.

69. Y. Yang and S. A. Smith: Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Molecular Biology and Evolution* **31**(11) (Aug. 2014), 3081–3092. doi: 10.1093/molbev/msu245.

70. B. J. Mans, D. de Klerk, R. Pienaar, and A. A. Latif: *Nuttalliella namaqua*: A Living Fossil and Closest Relative to the Ancestral Tick Lineage: Implications for the Evolution of Blood-Feeding in Ticks. *PLoS ONE* **6**(8) (Aug. 2011). Ed. by P. L. Oliveira, e23675. doi: 10.1371/journal.pone.0023675.

71. D.-S. Chen et al.: The Complete Mitochondrial Genomes of Six Species of *Tetranychus* Provide Insights into the Phylogeny and Evolution of Spider Mites. *PLoS ONE* **9**(10) (Oct. 2014). Ed. by H. Escriva, e110625. doi: 10.1371/journal.pone.0110625.

72. A. Schwarz, A. Cabezas-Cruz, J. Kopecký, and J. J. Valdés: Understanding the evolutionary structural variability and target specificity of tick salivary Kunitz peptides using next generation transcriptome data. *BMC Evolutionary Biology* **14**(1) (2014), 4. doi: 10.1186/1471-2148-14-4.

Titre : Diversité génomique, évolution et adaptation de la tique *Ixodes ricinus*.

Mots clés : Transcriptome, *Ixodes ricinus*, RNA-seq, Analyse d'expression différentielle, Génomique des populations, Phylogénomique.

Résumé : Les tiques sont des acariens hématophages vecteurs de nombreux micro-organismes dont certains sont responsables de maladies humaines ou animales (Borréliose de Lyme par exemple). La tique *Ixodes ricinus*, est largement distribuée en Europe où elle représente le principal vecteur de l'agent responsable de la maladie de Lyme. Trois volets ont été abordés au cours de cette thèse, en réalisant pour chaque point des séquençages à haut-débit de transcriptomes. Dans le premier volet, un catalogue de transcrits a été reconstruit et annoté à partir d'individus provenant de différentes conditions physiologiques (stades de développement, état de gorgement, sexe). Une analyse d'expression différentielle a permis de déterminer quels gènes sont exprimés plus spécifiquement lors du gorgement (protéines cuticulaires notamment, mais également métallo-protéases, etc...). Dans le deuxième volet, la structure génétique d'*I. ricinus* a été explorée à partir de douze populations Européennes. Mes résultats montrent pour la première fois un signal clair de structuration géographique, et d'isolement par la distance, sà l'échelle de l'Europe. Dans le troisième volet, j'ai employé une approche phylogénomique sur le groupe des tiques dures : pour cela, j'ai reconstruit les transcriptomes de 27 espèces de tiques (dont neuf espèces séquencées pour ce projet) permettant de proposer un arbre phylogénétique très robuste pour ce groupe.

Title : Genomic diversity, evolution and adaptation of the tick *Ixodes ricinus*.

Keywords : Transcritome, *Ixodes ricinus*, RNA-seq, Differential expression analysis, Population genomics, Phylogenomics

Abstract : Ticks are obligate blood-feeders, able to transmit numerous micro-organisms, including causative agents of human or veterinary diseases (e.g.: Lyme Borreliosis). The tick *Ixodes ricinus* is a widely distributed species in Europe, where it is the principal vector of the Lyme disease agent. Using high-throughput transcriptome sequencing, three lines of research were investigated. First, a large catalogue of transcripts was reconstructed and annotated for ticks in different physiological conditions (feeding or non-feeding, developmental stage, and sex). A differential expression analysis allowed to pinpoint genes associated with blood-feeding at the level of the whole body (genes involved in cuticle production, metalloprotease, etc...).Secondly, I explored the genetic structure of *I. ricinus* at the European scale using transcriptomes from 12 populations. I found a clear signal of phylogeographical strucure probably resulting from an isolation by distance process. Finally, transcriptomes for 27 different species of ticks were reconstructed (including nine species sequenced for this study). This permitted to reconstruct a robust phylogeny for the whole group of hard ticks.