16S rRNA amplicon sequencing for epidemiological surveys of bacteria in wildlife: the importance of cleaning post-sequencing data before estimating positivity, prevalence and co-infection

Maxime Galan[1], Maria Razzauti[1], Emilie Bard[2], Maria Bernard[3,4], Carine Brouat[5], Nathalie Charbonnel[1], Alexandre Dehne-Garcia[1], Anne Loiseau[1], Caroline Tatard[1], Lucie Tamisier[1], Muriel Vayssier-Taussat[6], Helene Vignes[7], Jean-François Cosson[1,6]

1: INRA, CBGP, Montferrier sur Lez, France

2: INRA, EpiA, Clermont-Ferrand, France

3: INRA, Sigenae, France

4: INRA, GABI, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France

5: Ird, CBGP, Montferrier sur Lez, France

6: INRA, Bipar, Maisons-Alfort, France

7: CIRAD, AGAP, Montpellier, France

Corresponding authors: galan@supagro.inra.fr; cosson@supagro.inra.fr

Summary

Human impact on natural habitats is increasing the complexity of human-wildlife interfaces and leading to the emergence of infectious diseases worldwide. Highly successful synanthropic wildlife species, such as rodents, will undoubtedly play an increasingly important role in transmitting zoonotic diseases. We investigated the potential for recent developments in 16S rRNA amplicon sequencing to facilitate the multiplexing of large numbers of samples needed to improve our understanding of the risk of zoonotic disease transmission posed by urban rodents in West Africa. In addition to listing pathogenic bacteria in wild populations, as in other high-throughput sequencing (HTS) studies, our approach can estimate essential parameters for studies of zoonotic risk, such as prevalence and patterns of coinfection within individual hosts. However, the estimation of these parameters requires cleaning of the raw data to mitigate the biases generated by HTS methods. We present here an extensive review of these biases and of their consequences, and we propose a comprehensive trimming strategy for managing these biases. We demonstrated the

application of this strategy using 711 commensal rodents collected from 24 villages in Senegal, including 208 *Mus musculus domesticus,* 189 *Rattus rattus*, 93 *Mastomys natalensis* and 221 *Mastomys erythroleucus*. Seven major genera of pathogenic bacteria were detected in their spleens: *Borrelia*, *Bartonella*, *Mycoplasma, Ehrlichia*, *Rickettsia*, *Streptobacillus* and *Orientia*. The last five of these genera have never before been detected in West African rodents. Bacterial prevalence ranged from 0% to 90% of individuals per site, depending on the bacterial taxon, rodent species and site considered, and 26% of rodents displayed coinfection. The 16S rRNA amplicon sequencing strategy presented here has the advantage over other molecular surveillance tools of dealing with a large spectrum of bacterial pathogens without requiring assumptions about their presence in the samples. This approach is therefore particularly suitable for continuous pathogen surveillance in the context of disease monitoring programs.

Importance

Several recent public health crises have shown that the surveillance of zoonotic agents in wildlife is important to prevent pandemic risks. High-throughput sequencing (HTS) technologies are potentially useful for this surveillance, but rigorous experimental processes are required for the use of these effective tools in such epidemiological contexts. In particular, HTS introduces biases into the raw dataset that might lead to incorrect interpretations. We describe here a procedure for cleaning data before estimating reliable biological parameters, such as positivity, prevalence and coinfection, with 16S rRNA amplicon sequencing on the Illumina MiSeq platform. This procedure, applied to 711 rodents collected in West Africa, detected several zoonotic bacteria, including some at high prevalence despite never before having been reported for West Africa. In the future, this approach could be adapted for the monitoring of other microbes such as protists, fungi, and even viruses.

Introduction

Pathogen monitoring in wildlife is a key method for preventing the emergence of

infectious diseases in humans and domestic animals. More than half the pathogens causing disease in humans originate from animal species [1]. The early identification of zoonotic agents in animal populations is therefore of considerable interest for human health. Wildlife species may also act as a reservoir for pathogens capable of infecting livestock, with significant economic consequences [2]. The monitoring of emerging diseases in natural populations is also important for preserving biodiversity, because pathogens carried by invasive species may cause the decline of endemic species [3]. There is, therefore, a need to develop screening tools for identifying a broad range of pathogens in samples consisting of large numbers of individual hosts or vectors.

High-throughput sequencing (HTS) approaches require no prior assumptions about the bacterial communities present in samples of diverse nature, including non-cultivable bacteria. Such HTS microbial identification approaches are based on the sequencing of all (WGS: whole-genome sequencing) or some (RNAseq or 16S rRNA amplicon sequencing) of the bacterial DNA or RNA in a sample, followed by comparison to a reference sequence database [4]. HTS has made major contributions to the generation of comprehensive inventories of the bacteria, including pathogens, present in humans [5]. Such approaches are now being extended to the characterization of bacteria in wildlife [6-13]. However, improvements in the estimation of infection risks will require more than just the detection of bacterial pathogens. Indeed, we will also need to estimate the prevalence of these pathogens by host taxon and/or environmental features, together with coinfection rates [14,15] and pathogen interactions [16,17].

Razzauti *et al*. [8] recently used 16S rRNA amplicon sequencing with the dual-index sequencing strategy of Kozich *et al*. [18] to detect bacterial pathogens in very large numbers of rodent samples (up to several hundred samples in a single run) on the Illumina MiSeq sequencing platform. The 16S rRNA amplicon sequencing technique is based on the amplification of small fragments of one or two hypervariable regions of the 16S rRNA gene. The sequences of these fragments are then obtained and compared with reference sequences in curated databases for taxonomic identification [4,19]. Multiplexed approaches of this kind include short indices (or tags) linked to the PCR products and specific to a given sample. This makes it possible to assign the sequences generated by the HTS run to a particular sample

following bioinformatic analysis of the dataset [18]. Razzauti *et al.* [8] demonstrated the considerable potential of this approach for determining the prevalence of bacteria within populations and for analyzing bacterial interactions within hosts and vectors, based on the accurate characterization of bacterial diversity within each individual samples it provides. However, various sources of error during the generation and processing of HTS data [20] may make it difficult to determine which samples are really positive or negative for a given bacterium. The detection of one or a few sequences assigned to a given taxon in a sample does not necessarily mean that the bacterium is actually present in that sample. We carried out an extensive literature review, from which we identified several potential sources of error involving all stages of a 16S rRNA amplicon sequencing experiment — from the collection of samples to the bioinformatic analysis — that might lead to false-negative or false-positive screening results (Table 1, [18,19,21-40]). These error sources have now been documented, and recent initiatives have called for the promotion of open sharing of standard operating procedures and best practices in microbiome research [41]. However, no experimental designs minimizing the impact of these sources of error on HTS data interpretation have yet been reported.

We describe here a rigorous experimental design for the direct estimation of biases from the data produced by 16S rRNA amplicon sequencing. We used these bias estimates to control and filter out potential false-positive and false-negative samples during screening for bacterial pathogens. We applied this strategy to 711 commensal rodents collected from 24 villages in Senegal, Western Africa: 208 *Mus musculus domesticus,* 189 *Rattus rattus*, 93 *Mastomys natalensis* and 221 *Mastomys erythroleucus*. Pathogenic bacteria associated with the rodents were analysed using a protocol based on Illumina MiSeq sequencing of the V4 hypervariable region of the 16S rRNA gene [18]. We considered the common pitfalls listed in Table 1 during the various stages of the experiment (see details in the workflow procedure, Figure 1). Biases in assessments of the presence or absence of bacteria in rodents were estimated directly from the dataset, by including and analysing negative controls (NC) and positive controls (PC) at various stages of the experiment (see Box 1), and systematically using sample replicates. This strategy delivers realistic and reliable estimates of bacterial prevalence in wildlife populations, and could be used to analyse the co-occurrence of different bacterial species within individuals.

Table 1. Sources of bias during the experimental and bioinformatic steps of 16S rRNA amplicon sequencing. Consequences for data interpretation and solutions for mitigating these biases.

Experimental steps	Sources of errors	Consequences	Solutions
Sample collection	Cross-contamination between individuals [21]	False-positive samples	Rigorous processing (decontamination of the instruments, cleaning of the autopsy table, use of sterile bacterial-free consumables, gloves, masks)
			Negative controls during sampling (e.g., organs of healthy mice during dissection)
	Collection and storage conditions [21]	False-positive & negative samples	Use of appropriate storage conditions/buffers. Use of unambiguously identified samples. Double checking of tube labeling during sample collection.
DNA extraction	Cross-contamination between samples [22]	False-positive samples	Rigorous processing (separation of pre- and post-PCR steps, use of a sterile hood, filter tips and sterile bacterial-free consumables)
	Reagent contamination with bacterial DNA [21,23]	False-positive samples	Negative controls for extraction (extraction without sample)
	Small amounts of DNA [21, 24]	False-negative samples	Use of an appropriate DNA extraction protocol. Discarding of samples with a low DNA concentration
Target DNA region and primer design	Target DNA region efficacy [19,25]	False-negative due to poor taxonomic identification	Selection of an appropriate target region and design of effective primers for the desired taxonomic resolution
	Primer design [21,26]	False-negative samples due to biases in PCR amplification for some taxa	Checking of the universality of the primers with reference sequences
Tag/Index design and preparation	False-assignments of sequences due to cross-contamination between tags/indices [27,28]	False-positive samples	Rigorous processing (use of sterile hood, filter tips and sterile bacterial-free consumables, brief centrifugation before the opening of index storage tubes, separation of pre- and post-PCR steps)
			Negative controls for tags/indices (empty wells without PCR reagents for particular tags or index combinations)
			Positive controls for alien DNA, i.e. a bacteria strain highly unlikely to infect the samples studied (e.g., a host-specific bacterium unable to persist in the environment) to estimate false assignment rate
	False-assignments of sequences due to inappropriate tag/index design [29]	False-positive samples	Fixing of a minimum number of substitutions between tags or indices. Each nucleotide position in the sets of tags or indices should display about 25% occupation by each base for Illumina sequencing
PCR amplification	Cross-contamination between PCRs [28]	False-positive samples	Rigorous processing (brief centrifugation before opening the index storage tubes, separation of pre- and post-PCR steps)
			Negative controls for PCR (PCR without template) with microtubes left open during sample processing
	Reagent contamination with bacterial DNA [21,23]	False-positive samples	Rigorous processing (use of sterile hood, filter tips and sterile bacterial-free consumables)
			Negative controls for PCRs (PCR without template), with microtubes closed during sample processing
	Chimeric recombinations by jumping PCR [27,30,31,32,33]	False-positive samples due to artifactual chimeric sequences	Increasing the elongation time and decreasing the number of cycles. Use of a bioinformatic strategy to remove the chimeric sequences (e.g., Uchime program)
	Poor or biased amplification [46]	False-negative samples	Increasing the amount of template DNA; Optimizing the PCR conditions (reagents and program)
			Use of technical replicates to validate sample positivity
			Positive controls for PCR (extraction from infected tissue and/or bacterial isolates)
Library preparation	Cross-contamination between PCRs/libraries [22]	False-positive samples	Rigorous processing (use of a sterile hood, filter tips and sterile bacterial-free consumables, electrophoresis and gel excision with clean consumables, separation of pre and post-PCR steps)
			Use of a protocol with an indexing step during target amplification
			Negative controls for indices (changing well positions between library preparation sessions)
	Chimeric recombinations by jumping PCR [27]	False-positive samples due to inter-individual recombinations	Avoiding PCR library enrichment of pooled samples. Positive controls for alien DNA, i.e. a bacterial strain that should not be identified in the sample (e.g. a host-specific bacterium unable to persist in the environment)
MiSeq sequencing (Illumina)	Sample sheet errors [21]	False-positive and negative samples	Negative controls (wells without PCR reagents for a particular index combination)
	Run-to-run carryover (Illumina Technical Support Note No. 770-2013-046)	False-positive samples	Washing of the MiSeq with dilute sodium hypochlorite solution
	Poor quality of reads due to flowcell overloading [34]	False-negative due to low quality of sequences	qPCR quantification of the library before sequencing.
	Poor quality of reads due to low-diversity libraries (Illumina Technical Support Note No. 770-2013-013)		Decreasing cluster density. Creation of artificial sequence diversity at the flowcell surface (e.g., by adding 5 to 10% PhiX DNA control library)
	Small number of reads per sample [35,36]	False-negative due to low depth of sequencing	Decreasing the level of multiplexing
			Discard the sample with a low number of reads
	Too short overlapping read pairs [18]	False-negative due to low quality of sequences	Increasing paired-end sequence length or decreasing the length of the target sequence
	Mixed clusters on the flowcell [27]	False-positive due to false index-pairing	Use of a single barcode sequence for both the i5 and i7 indices for each sample (when possible, e.g. small number of samples)
			Positive controls for alien DNA, i.e., a bacterial strain highly unlikely to be found in the rodents studied (e.g., a host-specific bacterium unable to persist in the environment)
Bioinformatics and taxonomic classification	Poor quality of reads	False-negative samples due to poor taxonomic resolution	Removal of low-quality reads
	Errors during processing (sequence trimming, alignment) [18,37,38]	False-positive and negative samples	Use of standardized protocols and reproducible workflows
	Incomplete reference sequence databases [39]	False-negative samples	Selection of an appropriate database for the selected target region and testing of the database for bacteria of particular interest
	Error of taxonomic classification [40]	False-positive samples	Positive controls for PCRs (extraction from infected tissue and/or bacterial isolates and/or mock communities)
			Checking of taxonomic assignments by other methods (e.g., Blast analyses on different databases)

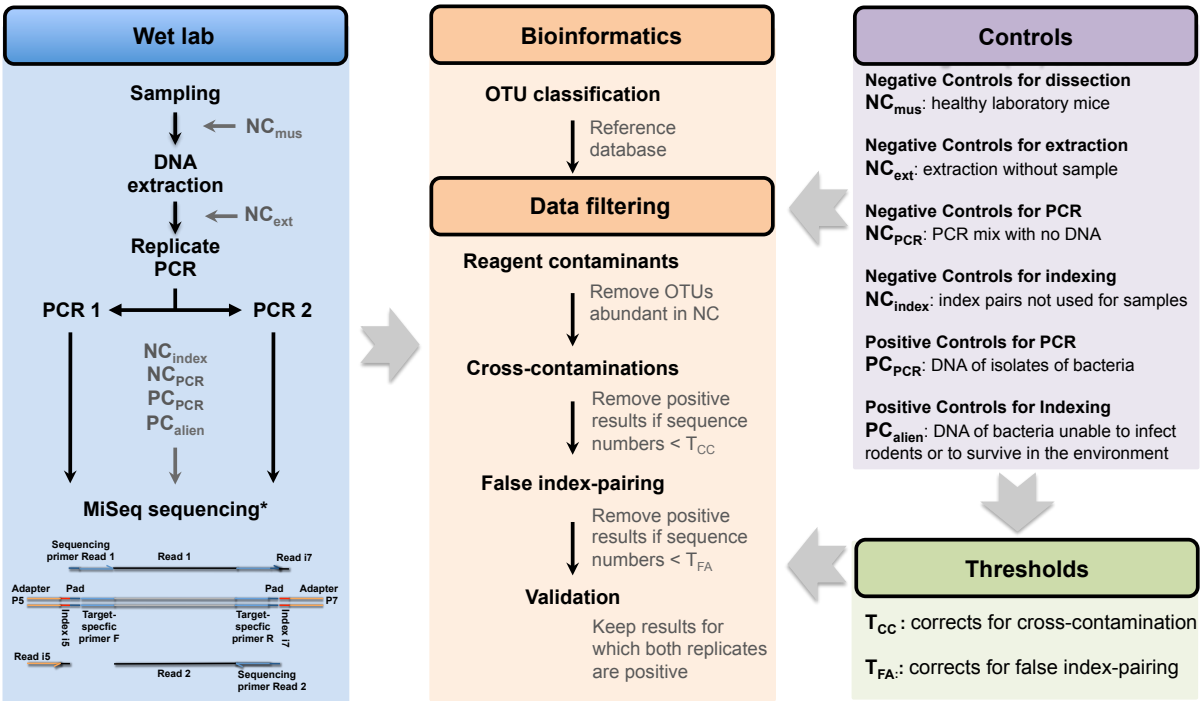


Figure 1. Workflow of the wet laboratory, bioinformatics and data filtering procedures in the process of data filtering for 16S rRNA amplicon sequencing.
Reagent contaminants were detected by analyzing the sequences in the NC_{ext} and NC_{PCR} controls. Sequence number threshold for correcting for cross-contamination (T_{CC}) are OTU- and run-dependent, and were estimated by analyzing the sequences in the NC_{mus}, NC_{ext}, NC_{PCR} and PC_{index} controls. Sequence number threshold for correcting for false index-pairing (T_{FA}) values are OTU- and run-dependent, and were estimated by analyzing the sequences in the NC_{index} and PC_{alien} controls. A result was considered positive if the number of sequences was > T_{CC} and > T_{FA}. Samples were considered positive if a positive result was obtained for both PCR replicates. *see Kozich et al 2013 for details on the sequencing.

Results & Discussion

Raw sequencing results. The sequencing of 1569 PCR products in two MiSeq runs generated a total of 23,698,561 raw paired-end sequence reads (251-bp) of the V4 region of the 16S rRNA gene. Because we made PCR replicates for each rodent sample, and because we included several controls in each sequencing run, we have more PCR products (N=1569) than rodent samples (N=711) (see summary in Table S1 and complete information by sample and run in Table S2). Overall, 99% of PCRs generated more than 3,000 raw reads (mean: 11,908 reads; standard deviation: 6,062). The raw sequence files are available in FASTQ format in the Dryad Digital Repository http://dx.doi.org/10.5061/dryad.m3p7d [42].

Using mothur v1.34 [43] and the MiSeq standard operating procedure (http://www.mothur.org/wiki/MiSeq_SOP), we removed 20.1% of paired-end reads because they were misassembled, 1.5% of sequences because they were misaligned, 2.6% because they were chimeric and 0.2% because they were non-

bacterial. The remaining reads were grouped into operational taxonomic units (OTUs) with a divergence threshold of 3%. Bioinformatics analysis identified 13,296 OTUs, corresponding to a total of 7,960,533 sequences in run 1 and 6,687,060 sequences in run 2.

Box 1. Guideline for experimental controls to include within high-throughput amplicon sequencing experiments to mitigate false positive results

Recent research has highlighted different biases occurring at different steps of high-throughput sequencing. These biases can be estimated directly from the data by including several controls together with samples in the experiment. We detail below these different controls as well as the rationale for their use.

Negative Controls for sample collection. When possible we advise to include axenic samples during sample collection. The number of sequences observed in these controls are used to estimate cross-contamination rates during sample collection. In our study we used spleens from healthy laboratory mice (**NC_{mus}**), free from rodent pathogens, which were manipulated together with wild samples during the dissections in the field.

Negative Controls for DNA extraction (NC_{ext}). DNA extractions performed without the addition of sample tissue (blanks), which are processed together with the other samples. We advise performing at least one extraction blank for each extraction experiment, although more is better. The numbers of sequences observed in these controls are used to estimate and filter the cross-contaminations during the DNA extractions and to detect for DNA bacterial contaminants in the extraction kit reagents.

Negative Controls for PCR (NC_{PCR}). PCR reactions without any DNA extract included (blank), which are processed together with the other samples. We advise performing at least one PCR blank per PCR microplate, although more is better. The numbers of sequences observed in these controls are used to estimate and filter the cross-contaminations during the PCR preparation and to detect DNA bacterial contaminants in the PCR reagents.

Negative Controls for indexing (NC_{index}). Combinations of barcodes that are not used to identify samples in the sequencing run, but that are searched for during the bioinformatic demultiplexing. In practice, they correspond to empty PCR wells (without reagent and without index). The numbers of sequences recovered for these particular index combinations are used to estimate and filter the cross-contaminations between indexed PCR primers during primer handling or PCR preparation, and to identify errors in the Illumina sample sheet.

Positive Controls for PCR (PC_{PCR}). PCR reactions with DNA of known taxa isolates, which are processed together with the other samples. The sequences obtained for these controls are used to verify the taxonomic assignment and to estimate and filter cross-contaminations.

Positive Controls for Indexing (PC_{alien}). PCR reactions with DNA of taxa isolates that are known to be absent in the samples. They are handled separately from the samples to avoid cross-contaminations with the samples during the wet lab procedures (DNA extractions and PCRs). Sequences from PC_{alien} found in the samples are used to calculate the rate of sample misidentification due to false index-pairing (see text and Kircher et al [27] for details concerning this phenomenon).

In practice, (PC_{PCR}) and (PC_{alien}) could be the same and we advice to use taxa that are phylogenetically distant from the taxa we look for, in order to avoid potential confusion between sequences from alien controls and sequences from the samples.

Taxonomic assignment of sequences. We used the Bayesian classifier (bootstrap cutoff = 80%) implemented in mothur with the Silva SSU Ref database v119 [43] as a reference, for the taxonomic assignment of OTUs. The 50 most abundant OTUs accounted for 89% (min: 15,284 sequences; max: 2,206,731

sequences) of the total sequence dataset (Table S3). The accuracy of taxonomic assignment (to genus level) was assessed with positive controls for PCR, corresponding to DNA extracts from laboratory isolates of *Bartonella taylorii*, *Borrelia burgdorferi* and *Mycoplasma mycoides* ($PC_{Bartonella\_t}$, $PC_{Borrelia\_b}$ and $PC_{Mycoplasma\_m}$, respectively), which were correctly assigned to a single OTU corresponding to the appropriate reference sequences (Table 2). Note that the sequences of $PC_{Mycoplasma\_m}$ were assigned to Entomoplasmataceae rather than Mycoplasmataceae because of a frequent taxonomic error reflected in most databases, including Silva [45]. This problem might also affect other taxa. We therefore recommend systematically carrying out a blast analysis against the sequences of taxa of interest in GenBank to confirm the taxonomic assignment obtained with the 16S databases. Finally, we assumed that the small number of sequences per sample might limit the completeness of bacterial detection [36]. For this reason, we discarded seven rodent samples (2 *M. erythroleucus* and 5 *M. domesticus*) yielding fewer than 500 sequences for at least one of the two PCR replicates (1% of the samples).

	OTUs	Total	Wild rodents (*n*=711)		Negative controls						Positive controls						Thresholds	
					NC_{PCR}		NC_{ext}		NC_{mus}		$PC_{Bartonnela\_t}$		$PC_{Borrelia\_b}$		$PC_{Mycoplasma\_m}$			
		Total no. of sequences	Total no. of sequences	Maximum no. of sequences in one PCR	Total no. of sequences	Maximum no. of sequences in one PCR	Total no. of sequences	Maximum no. of sequences in one PCR	Total no. of sequences	Maximum no. of sequences in one PCR	Total no. of sequences	Maximum no. of sequences in one PCR	Total no. of sequences	Maximum no. of sequences in one PCR	Total no. of sequences	Maximum no. of sequences in one PCR	T_{cc}*	T_{FA}**
Run 1	**Whole dataset**	7960533	7149444	64722	45900	8002	39308	8741	68350	26211	137424	73134	239465	120552	280642	82933	/	/
	***Mycoplasma*_OTU_1**	1410218	1410189	61807	2	1	3	2	9	5	3	3	8	6	4	3	6	282
	***Mycoplasma*_OTU_ 3**	507376	507369	36335	2	1	0	0	0	0	2	2	1	1	2	2	2	101
	***Ehrlichia*_OTU**	649451	649423	63137	4	2	3	2	7	4	1	1	1	1	12	6	6	130
	***Borrelia*_OTU**	345873	345845	28528	4	4	7	4	9	4	1	1	0	0	7	3	4	69
	***Orientia*_OTU**	279965	279957	29503	1	1	4	1	0	0	2	2	0	0	1	1	2	56
	***Bartonella*_OTU**	202127	67973	16145	1	1	1	1	1	1	134124	71163	7	4	20	9	9	40
	$PC_{Mycoplasma_m}$_OTU* **	280151	338	28	0	0	0	0	2	2	34	20	24	18	279753	82767	/	/
	$PC_{Borrelia_b}$_OTU* **	238772	420	43	0	0	0	0	0	0	38	21	238238	119586	76	23	/	/
Run 2	**Whole dataset**	6687060	6525107	42326	61231	9145	53334	7669	/	/	12142	7518	13378	7164	21868	6520	/	/
	***Mycoplasma*_OTU_1**	155486	155486	7703	0	0	0	0	/	/	0	0	0	0	0	0	0	31
	***Mycoplasma*_OTU_2**	1036084	1035890	23588	1	1	192	115	/	/	0	0	0	0	1	1	115	207
	***Mycoplasma*_OTU_ 3**	127591	127590	5072	1	1	0	0	/	/	0	0	0	0	0	0	1	26
	***Mycoplasma* _OTU_4**	85596	85583	20146	0	0	13	13	/	/	0	0	0	0	0	0	13	17
	***Mycoplasma*_OTU_ 5**	56324	56324	10760	0	0	0	0	/	/	0	0	0	0	0	0	0	11
	***Mycoplasma*_OTU_ 6**	13356	13356	1482	0	0	0	0	/	/	0	0	0	0	0	0	0	3
	***Ehrlichia*_OTU**	74017	74017	19651	0	0	0	0	/	/	0	0	0	0	0	0	0	15
	***Borrelia*_OTU**	21636	21636	3085	0	0	0	0	/	/	0	0	0	0	0	0	0	4
	***Orientia*_OTU**	307	307	181	0	0	0	0	/	/	0	0	0	0	0	0	0	0
	***Bartonella*_OTU**	1559028	1547652	14515	1	1	2	2	/	/	11297	6714	2	2	74	59	59	312
	***Streptobacillus*_OTU**	32399	32399	6245	0	0	0	0	/	/	0	0	0	0	0	0	0	6
	***Rickettsia*_OTU**	589	589	329	0	0	0	0	/	/	0	0	0	0	0	0	0	0
	$PC_{Mycoplasma_m}$_OTU* **	16854	2	1	0	0	0	0	/	/	0	0	0	0	16852	5766	/	/
	$PC_{Borrelia_b}$_OTU* **	12197	0	0	0	0	0	0	/	/	0	0	12197	6426	0	0	/	/

*: Threshold T_{cc} is based on the maximum number of sequences observed in a negative or positive control for a particular OTU in each run
**: Threshold T_{FA} is based to the false assignment rate (0.02%) weighted by the total number of sequences of each OTU in each run
***: *Mycoplasma mycoides* and *Borrelia burgdorferi* bacterial isolates added as positive controls for PCR and indexing (i.e., PC_{alien} see Figure 1)

Table 2. Number of sequences for 12 pathogenic OTUs observed in wild rodents, negative controls, and positive controls, together with T_{CC} and T_{FA} threshold values. Data are given for the two MiSeq runs separately. NC_{PCR}: negative controls for PCR; NC_{ext}: negative controls for extraction; NC_{mus}: negative controls for dissection; $PC_{Bartonella\_t}$: positive controls for PCR; $PC_{Borrelia\_b}$ and $PC_{Mycoplsma\_m}$: positive controls for PCR and positive controls for indexing; T_{CC} and T_{FA}: thresholds for positivity for a particular bacterium according to bacterial OTU and MiSeq run (see also Figure 1).

Filtering for reagent contaminants. 16S rRNA amplicon sequencing data may be affected by the contamination of reagents [23]. We therefore filtered the data, using negative controls for extraction (NC_{ext}), corresponding to extraction without the addition of a tissue sample, and negative controls for PCR (NC_{PCR}), corresponding to PCR mixtures to which no DNA was added. We observed between 2,843 and 8,967 sequences in the NC_{ext} and between 5,100 and 9,145 sequences in the NC_{PCR}. Based on their high number of reads in negative controls, we identified 13 contaminant genera, including *Pseudomonas*, *Acinetobacter*, *Herbaspirillum*, *Streptococcus*, *Pelomonas*, *Brevibacterium*, *Brachybacterium*, *Dietzia*, *Brevundimonas*, *Delftia*, *Comamonas*, *Corynebacterium*, and *Geodermatophilus*, some of them having been previously identified in other studies [23]. These contaminants accounted for 29% of the sequences in the dataset (Figure 2). They also differed between MiSeq runs: *Pseudomonas*, *Pelomonas* and Herbaspirillum predominated in run 1, whereas *Brevibacterium*, *Brachybacterium* and *Dietzia* predominated in run 2 (Table S4, Figure S1). This difference probably reflects the use of two different PCR kits manufactured several months apart (Qiagen technical service, pers. com.). The majority of other contaminants, such as *Streptococcus,* most likely originated from the DNA extraction kits used, as they were detected in abundance in the negative controls for extraction (NC_{ext}).

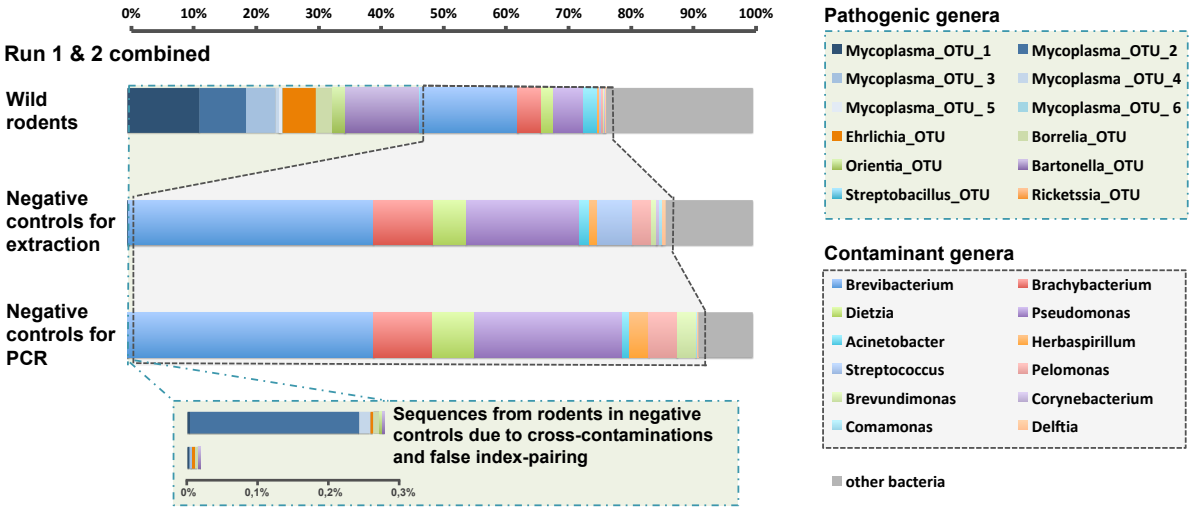


Figure 2. Taxonomic assignment of the V4 16S rRNA sequences in wild rodents, and in negative controls for extraction and PCR. The histograms show the percentage of sequences for the most abundant bacterial genera in the two MiSeq runs combined. Notice the presence of several bacterial genera in the controls, which were likely due to the inherent contamination of laboratory reagents by bacterial DNA (termed 'contaminant genera'). These contaminant genera are also present (to a lesser extent) in the rodent samples. The inserts represent the proportion of sequences from rodent samples, which were incorrectly assigned to the controls. See Figure S1 for separate histograms for both MiSeq runs.

Genera identified as contaminants were then simply removed from the sample dataset. It is important to note, however, that the exclusion of these results does not rule out the possibility that our samples contained true rodent infections (at least for some of them like *Streptococcus* which contains both saprophytic and pathogenic species). However, as mentioned by Razzauti *et al*. [8] distinguishing between those two possibilities seems difficult, if not impossible. Faced with this lack of certainty, it is most prudent to simply remove these taxa from the sample dataset. These results highlight the importance of carrying out systematic negative controls to filter the taxa concerned in order to prevent inappropriate data interpretation, particularly for the *Streptococcus* genus, which contains a number of important pathogenic species. The use of DNA-free reagents would improve the quality of sequencing data and likely increase the depth of sequencing of the samples.

After filtering for the above reagent contaminants, 12 OTUs, belonging to 7 genera for which at least one species or one strain is known to be pathogenic in mammals (therefore referenced as “pathogenic genera”), accounted for 66% of the sequences identified in wild rodent samples for both MiSeq runs combined (Figure 2). These genera are *Bartonella*, *Borrelia*, *Ehrlichia*, *Mycoplasma*, *Orientia*, *Rickettsia* and *Streptobacillus*. Six different OTUs were obtained for *Mycoplasma* (*Mycoplasma*_OTU_1 to *Mycoplasma*_OTU_6), and one OTU each for the other genera (Table 2). Finally, the precise significance of the remaining 34% of sequences was undetermined, potentially corresponding to commensal bacteria (Bacteroidales, Bacteroides, Enterobacteriaceae, Helicobacter, Lactobacillus), unknown pathogens, undetected contaminants, or undetected sequencing errors.

Filtering for false-positive results. Mothur analysis produced a table of abundance, giving the number of sequences for each OTU in each PCR product (data available in the Dryad Digital Repository http://dx.doi.org/10.5061/dryad.m3p7d [42]. The multiple biases during experimental steps and data processing listed in Table 1 made it impossible to infer prevalence and co-occurrence directly from the table of sequence presence/absence in the PCR products. We suggest filtering the data with estimates of the different biases calculated from the multiple controls introduced during the process. This strategy involves calculating sequence number thresholds from our bias estimates. Two different thresholds were set for each of the

12 OTUs and two MiSeq runs. We then discarded positive results associated with sequence counts below the threshold (Figure 1).

Threshold T_{cc}: Filtering for cross-contamination. One source of false positives is cross-contamination between samples processed in parallel (Table 1). Negative controls for dissection (NC_{mus}), consisting of the spleens of healthy laboratory mice manipulated during sessions of wild rodent dissection, and negative controls for extraction (NC_{ext}) and PCR (NC_{PCR}) were used, together with positive controls for PCR ($PC_{Bartonella\_t}$, $PC_{Borrelia\_b}$ and $PC_{Mycoplasma\_m}$), to estimate cross-contamination. For each sequencing run, we calculated the maximal number of sequences for the 12 pathogenic OTUs in the negative and positive controls. These numbers ranged from 0 to 115 sequences, depending on the OTU and the run considered (Table 2), and we used them to establish OTU-specific thresholds (T_{CC}) for each run. For example, in Sequencing Run 2, the highest number of sequences in a control for *Mycoplasma*_OTU_2 was 115 (in a NC_{ext}). Therefore, we established the threshold value at 115 sequences for this OTU in sequencing Run 2. Thus, PCR products with less than 115 sequences for the *Mycoplasma*_OTU_2 in sequencing Run 2 were considered as false-positive for this OTU. The use of these T_{CC} led to 0% to 69% of the positive results being discarded, corresponding to only 0% to 0.14% of the sequences, depending to the OTU considered (Figure 3, Table S5). A PCR product may be positive for several bacteria in cases of coinfection. In such cases, the use of a T_{CC} makes it possible to discard the positive result for one bacterium whilst retaining positive results for other bacteria.

Threshold T_{FA}: Filtering out incorrectly assigned sequences. Another source of false positives is the incorrect assignment of sequences to a PCR product (Table 1). This phenomenon may be due either to cross-contamination between indices during the experiment, or to the generation of mixed clusters during the sequencing [27].

First, the cross-contamination of indexes may happen during the preparation of indexed primer microplates. This cross-contamination was estimated using negative control index pairs (NC_{index}) corresponding to particular index pairs not used to identify the samples. NC_{index} returned very few read numbers (1 to 12), suggesting that there was little or no cross-contamination between indices in our experiment.

Second, the occurrence of mixed clusters during the sequencing of multiplexed samples was reported by Kircher et al [27]. Mixed clusters on the Illumina flowcell

surface are considered by Kircher et al [27] as the predominant source of error of sequence assignment to a PCR product. The impact of this phenomenon on our experiment was estimated using “alien” positive controls (PC_{alien}) sequenced in parallel of the rodent samples: $PC_{Mycoplasma\_m}$, corresponding to the DNA of *Mycoplasma mycoides*, which cannot infect rodents, and $PC_{Borrelia\_b}$, containing the DNA of *Borrelia burgdorferi*, which is not present in Africa. Neither of these bacteria can survive in abiotic environments, so the presence of their sequences in African rodent PCR products indicates a sequence assignment error due to false index-pairing [27]. Using $PC_{Mycoplasma\_m}$, we obtained an estimate of the global false index-pairing rate of 0.14% (i.e. 398 of 280,151 sequences of the *Mycoplasma mycoides* OTU were assigned to samples other than $PC_{Mycoplasma\_m}$). Using $PC_{Borrelia\_b}$, we obtained an estimate of 0.22% (534 of 238,772 sequences of the *Borrelia burgdorferi* OTU were assigned to samples other than $PC_{Borrelia\_b}$). These values are very close to the estimate of 0.3% obtained by Kircher *et al*. [27]. Close examination of the distribution of misassigned sequences within the PCR 96-well microplates showed that all PCR products with misassigned sequences had one index in common with either $PC_{Mycoplasma\_m}$ or $PC_{Borrelia\_b}$ (Figure S2).

We then estimated the impact of false index-pairing for each PCR product by calculating the maximal number of sequences of “alien” bacteria assigned to PCR products other than the corresponding PC. These numbers varied from 28 to 43, depending on the positive control for run 1 (Table 2) — run 2 was discarded because of the low values of the numbers of sequences, which is likely due to the fact that DNAs of PC were diluted one hundred-fold in run 2 (Table S1). We then estimated a false-assignment rate for each PCR product (R_{fa}), by dividing the above numbers by the total number of sequences from “alien” bacteria in Sequencing Run 1. R_{fa} was estimated for $PC_{Mycoplasma\_m}$ and $PC_{Borrelia\_b}$ separately. R_{fa} reached 0.010% and 0.018% for $PC_{Mycoplasma\_m}$ and $PC_{Borrelia\_b}$, respectively. We adopted a conservative approach, by fixing the R_{fa} value to 0.020%. This number signifies that each PCR product may receive a maximum 0.020% of the total number of sequences of an OTU present in a run due to false index-pairing. Moreover, the number of misassigned sequences for a specific OTU into a PCR product should increase with the total number of sequences of the OTU in the MiSeq run. We therefore defined the second threshold (T_{FA}) as the total number of sequences in the run for an OTU

multiplied by R_{fa}. T_{FA} values varied with the abundance of each OTU in the sequencing run (Table 2). Because the abundance of each OTU varied from one sequencing run to the other, T_{FA} also varied according to the sequencing run. The use of the T_{FA} led to 0% to 87% of positive results being discarded. This corresponded to 0% to 0.71% of the sequences, depending on the OTU (Figure 3, Table S5).

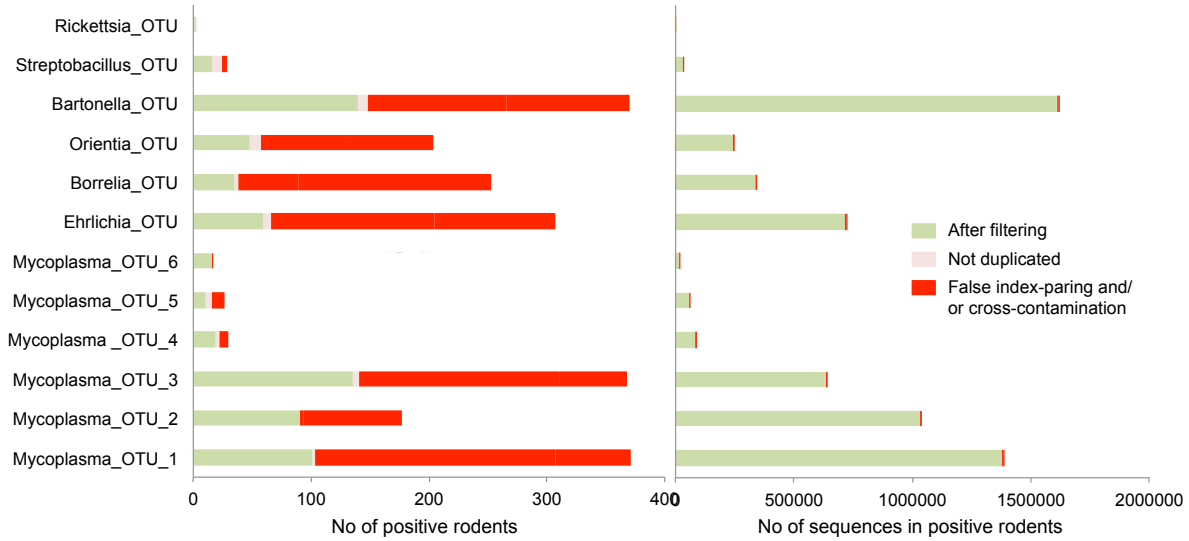


Figure 3. Numbers of positive rodents, and of sequences in positive rodents, removed for each OTU at each step in data filtering. These findings demonstrate that the positive rodents filtered out corresponded to only a very small number of sequences. (A) The histogram shows the number of positive rodents discarded because of likely cross-contamination, false index-pairing, and failure to replicate in both PCRs, as well as the positive results retained at the end of data filtering in green. (B) The histogram shows the number of sequences corresponding to the same class of positive rodents. Note that several positive results may be recorded for the same rodent in cases of co-infection.

Validation with PCR replicates. Random contamination may occur during the preparation of PCR 96-well microplates. These contaminants may affect some of the wells, but not those for the negative controls, leading to the generation of false-positive results. We thus adopted a conservative approach, in which we considered rodents to be positive for a given OTU only if both PCR replicates were considered positive after the filtering steps described above. The relevance of this strategy was supported by the strong correlation between the numbers of sequences for the two PCR replicates for each rodent (R^2>0.90, Figure 4 and Figure S3). At this stage, 673 positive results for 419 rodents were validated for both replicates (note that a rodent may be positive for several bacteria, and may thus be counted several times), whereas only 52 positive results were discarded because the result for the other

replicate was negative. At this final validation step, 0% to 60% of the positive results for a given OTU were discarded, corresponding to only 0% to 7.17% of the sequences (Figure 3, Table S5 and Table S6). Note that the number of replicates may be increased, as described in the strategy of Gómez-Díaz *et al* [46].

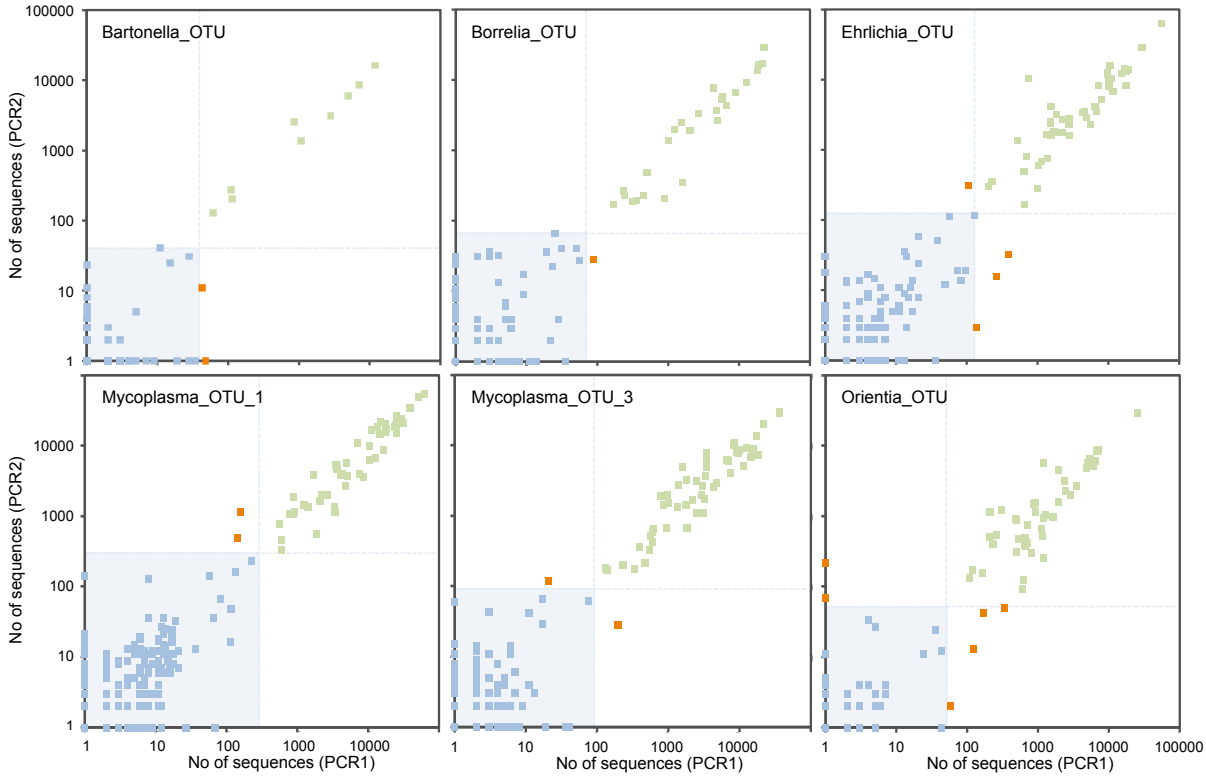


Figure 4. Plots of the number of sequences (log (x+1) scale) from bacterial OTUs in both PCR replicates (PCR1 & PCR2) of the 348 wild rodents analyzed in the first MiSeq run. Note that each rodent was tested with two replicate PCRs. Green points correspond to rodents with two positive results after filtering; red points correspond to rodents with one positive result and one negative result; and blue points correspond to rodents with two negative results. The light blue area and lines correspond to threshold values used for the data filtering: samples below the lines are filtered out. See Figure S3 for plots corresponding to the second MiSeq run.

Post-filtering results. Finally, the proportion of rodents positive for a given OTU filtered out by the complete filtering approach varied from 6% to 86%, depending on the OTU, corresponding to only 1% of the total sequences (Figure 3). Indeed, our filtering strategy mostly excluded rodents with a small number of sequences for the OTU concerned. These rodents were considered to be false-positive.

359 ***Refining bacterial taxonomic identification.*** We refined the taxonomic
360 identification of the 12 bacterial OTUs through phylogenetic and blast analyses. We
361 were able to identify the bacteria present down to genus level and, in some cases,
362 we could even identify the most likely species (Table 3 and Figure S4). For instance,
363 the sequences of the six *Mycoplasma* OTUs were consistent with three different
364 species — *M. haemomuris* for OTU_1 and 3, *M. coccoides* for OTU_4, 5 and 6, and
365 *M. species novo* [47] for OTU_2 — with high percentages of sequence identity
366 (≥93%) and bootstrap values ≥80%. All three of these species belong to the
367 Hemoplasma group, which is known to infect mice, rats and other mammals [48,49],
368 and is thought to cause anemia in humans [50,51]. The *Borrelia* sequences grouped
369 with three different species of the relapsing fever group (*crocidurae*, *duttonii* and
370 *recurrentis*) with a high percentage of identity (100%) and a bootstrap value of 71%.
371 In West Africa, *B. crocidurae* causes severe borreliosis, a rodent-borne disease
372 transmitted by ticks and lice [52]. The *Ehrlichia* sequences were 100% identical to
373 and clustered with the recently described Candidatus *Ehrlichia khabarensis* isolated
374 from voles and shrews in the Far East of Russia [53]. The *Rickettsia* sequences were
375 100% identical to the sequence of *R. typhi*, a species of the typhus group responsible
376 for murine typhus [54], but this clade was only weakly differentiated from many other
377 *Rickettsia* species (bootstrap support of 61%). The most likely species corresponding
378 to the sequences of the *Streptobacillus* OTU was *S. moniliformis*, with a high
379 percentage of identity (100%) and a bootstrap value of 100%. This bacterium is
380 common in rats and mice and causes a form of rat-bite fever, Haverhill fever [55].
381 The *Orientia* sequences corresponded to *O. chuto*, with a high percentage of identity
382 (100%) and a bootstrap value of 77%. This species was recently isolated from a
383 patient infected in Dubai [56]. Finally, accurate species determination was not
384 possible for *Bartonella*, as the 16S rRNA gene does not resolve the species of this
385 genus well [57]. Indeed, the sequences from the *Bartonella* OTU detected in our
386 rodents corresponded to at least seven different species (*elizabethae, japonica,*
387 *pachyuromydis, queenslandis, rattaustraliani, tribocorum, vinsonii*) and a putative
388 new species recently identified in Senegalese rodents [58].

389

390

391 **Table 3. Detection of 12 bacterial OTUs in the four wild rodent species (*n*=704)**
392 **sampled in Senegal;** biology and pathogenicity of the corresponding bacterial genus. *n*= number of rodents
393 analyzed.

OTUs of interest (genus level)	Closest species* (% identity in GenBank)	Number of positive wild rodents				Biology & epidemiology
		Mastomys erytholeucus (*n*=219)	*Mastomys natalensis* (*n*=93)	*Mus musculus* (*n*=203)	*Rattus rattus* (*n*=189)	
Bartonella	undetermined	60	73	1	6	*Bartonella* spp. are intracellular fastidious hemotropic gram-negative organisms identified in a wide range of domestic and wild mammals and transmitted by arthropods. Several rodent-borne *Bartonella* species have emerged as zoonotic agents, and various clinical manifestations are reported, including fever, bacteremia and neurological symptoms [84].
Borrelia	*crocidurae* (100%) *duttonii* (100%) *recurrentis* (100%)	21	0	8	6	*Borrelia* is a genus of spiral gram-negative bacteria of the spirochete phylum. These bacteria are obligate parasites of animals and are responsible for relapsing fever borreliosis, a zoonotic disease transmitted by arthropods (tick and lice). This disease is the most frequent human bacterial disease in Africa. *B. crocidurae* is endemic to West Africa, including Senegal, and *B. duttonii* and *B. recurrentis* have been reported in Central, southern and East Africa [52].
Ehrlichia	*khabarensis* (100%)	40	0	12	8	*The genus Ehrlichia* includes five species of small gram-negative obligate intracellular bacteria. The life cycle includes the reproduction stages taking place in both ixodid ticks, acting as vectors, and vertebrates. *Ehrlichia* spp. can cause a persistent infection in the vertebrate hosts, which thus become reservoirs of infection. A number of new genetic variants of *Ehrlichia* have been recently detected in rodent species (e.g., *Candidatus Ehrlichia khabarensis* [53]).
Mycoplasma OTU_1	*haemomuris* (96%)	28	42	30	1	*Mycoplasma* is a genus including over 100 species of bacteria that lack of a cell wall around their cell membrane. *Mycoplasma coccoides* and *Mycoplasma haemomuris* are blood parasites of wild and laboratory rodents. A new closely related species was recently isolated from brown rats (AB752303 [47]). These species are commonly referred as ‘‘hemoplasmas’’. Hemoplasmas have been detected within the erythrocytes of cats, dogs, pigs, rodents and cattle, in which they may cause anaemia. There have been sporadic reports of similar infections in humans, but these infections have been poorly characterized [51].
Mycoplasma OTU_2	*sp. novo (100%) GenBank AB752303*	0	0	0	90	
Mycoplasma OTU_3	*haemomuris* (93%)	93	40	1	1	
Mycoplasma OTU_4	*coccoides* (96%)	0	0	0	18	
Mycoplasma OTU_5	*coccoides* (95%)	3	8	0	0	
Mycoplasma OTU_6	*coccoides* (97%)	3	13	0	0	
Orientia	*chuto* (100%) *tsutsugamushi* (98%)	0	2	46	0	*Orientia* is a genus of obligate intracellular gram-negative bacteria found in mites and rodents. *Orientia tsutsugamushi* is the agent of scrub typhus in humans. This disease, one of the most underdiagnosed and underreported febrile illnesses requiring hospitalization, has an estimated 10% fatality rate unless treated appropriately. A new species, *Orientia chuto,* was recently characterized in sick patients from the Arabian Peninsula, and new *Orientia* haplotypes have been identified in France and Senegal [9].
Rickettsia	*typhi* (100%)	1	0	0	1	*Rickettsia* is a genus of obligate intracellular gram-negative bacteria found in arthropods and vertebrates. *Rickettsia* spp. are symbiotic species transmitted vertically in invertebrates, and some are pathogenic invertebrates. *Rickettsia* species of the typhus group cause many human diseases, including murine typhus, which is caused by *Rickettsia typhi* and transmitted by fleas [54].
Streptobacillus	*moniliformis* (100%)	10	1	0	5	*Streptobacillus* is a genus of aerobic, gram-negative facultative anaerobe bacteria, which grow in culture as rods in chains. *Streptobacillus moniliformis* is common in rats and mice and is responsible of the Streptobacillosis form of rat-bite fever, the Haverhill fever. This zoonosis begins with high prostrating fevers, rigors (shivering), headache and polyarthralgia (joint pain). Untreated, rat-bite fever has a mortality rate of approximately 10% [55].

*based on phylogenetic analysis, see Figure S3
n: number of rodents screened

394

These findings demonstrate the considerable potential of 16S rRNA amplicon sequencing for the rapid identification of zoonotic agents in wildlife, provided that the post-sequencing data are cleaned beforehand. *Borrelia* [52] and *Bartonella* [58] were the only ones of the seven pathogenic bacterial genera detected here in Senegalese rodents to have been reported as present in rodents from West Africa before. The other bacterial genera identified here have previously been reported to be presented in rodents only in other parts of Africa or on other continents. *Streptobacillus moniliformis* has recently been detected in rodents from South Africa [59] and there have been a few reports of human streptobacillosis in Kenya [60] and Nigeria [61]. *R. typhi* was recently detected in rats from Congo, in Central Africa [62], and human seropositivity for this bacterium has been reported in coastal regions of West Africa [63]. With the exception of one report in Egypt some time ago [64], *Mycoplasma* has never before been reported in African rodents. Several species of *Ehrlichia* (from the *E. canis* group: *E. chaffeensis, E. ruminantium, E. muris, E. ewingii*) have been characterized in West Africa, but only in ticks from cattle [65] together with previous reports of possible cases of human ehrlichioses in this region [66]. Finally, this study reports the first identification of *Orientia* in African rodents [9]. There have already been a few reports of suspected human infection with this bacterium in Congo, Cameroon, Kenya and Tanzania [67].

Estimating prevalence and coinfection. After data filtering, we were able to estimate the prevalence in rodent populations and to assess coinfection in individual rodents, for the 12 bacterial OTUs. Bacterial prevalence varied considerably between rodent species (Table 3). *Bartonella* was highly prevalent in the two multimammate rats *M. natalensis* (79%) and *M. erythroleucus* (27%); *Orientia* was prevalent in the house mouse *M. musculus* (22%) and *Ehrlichia* occurred frequently in only one on the two multimammate rats *M. erythroleucus* (18%). By contrast, the prevalence of *Streptobacillus* and *Rickettsia* was low in all rodent species (<5%). Coinfection was common, as 184 rodents (26%) were found to be coinfected with bacteria from two (19%), three (5%), four (2%) or five (0.1%) different bacterial pathogens.

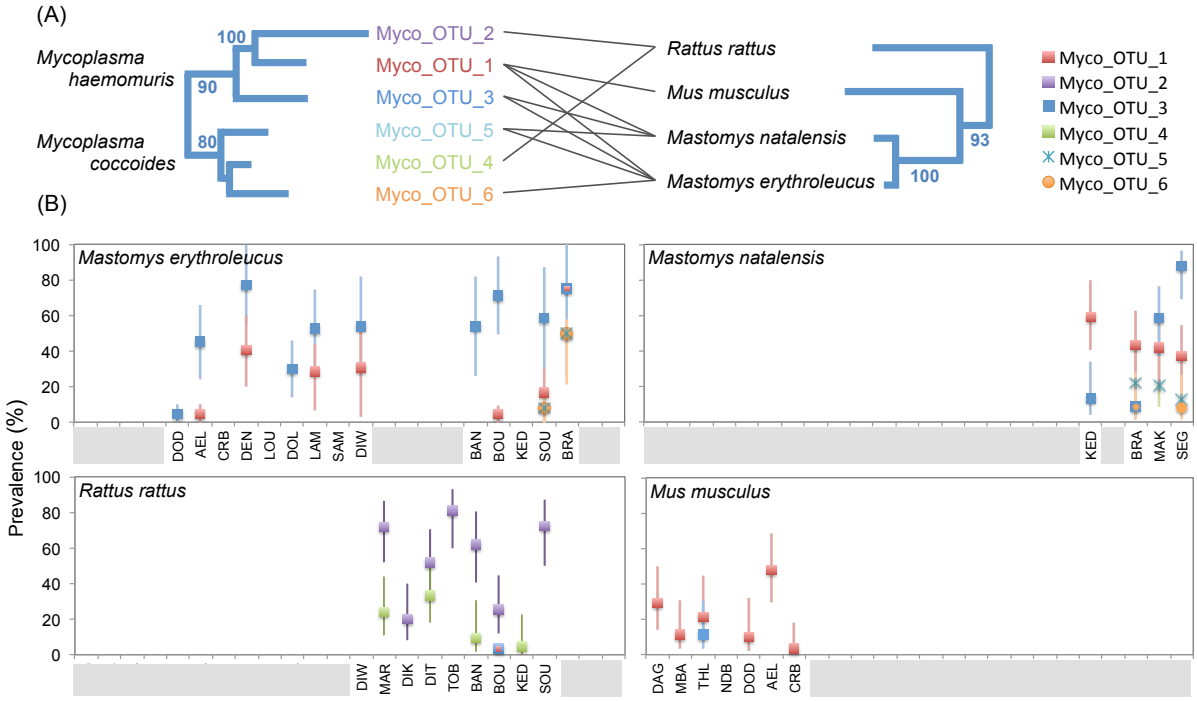


Figure 5. Prevalence of *Mycoplasma* lineages in Senegalese rodents, by site, and phylogenetic associations between *Mycoplasma* lineages and rodent species. (A) Comparison of phylogenetic trees based on the 16S rRNA V4-sequences of *Mycoplasma*, and on the mitochondrial cytochrome *b* gene and the two nuclear gene fragments (IRBP exon 1 and GHR) for rodents (rodent tree redrawn from [93]). Lines link the *Mycoplasma* lineages detected in the various rodent species (for a minimum site prevalence exceeding 10%). The numbers next to branches are bootstrap values (only shown if >70%). (B) Plots of OTU prevalence with 95% confidence intervals calculated by Sterne's exact method [94] by rodent species and site (see [69] for more information about site codes and their geographic locations). The gray bars in the X-legend indicate sites from which the rodent species concerned is absent.

Interestingly, several *Mycoplasma* OTUs appeared to be specific to a rodent genus or species (Table 3; Figure 5, Panel A). OTU_2, putatively identified as a recently described lineage isolated from brown rat, *Rattus norvegicus* [47], was specifically associated with *R. rattus* in this study. Of the OTUs related to *M. coccoides*, OTU_4 was found exclusively in *R. rattus,* whereas OTUs_5 and 6 seemed to be specific to the two multimammate rats (*M. erytholeucus and M. natalensis*). Comparative phylogenies of *Mycoplasma* OTUs and rodents showed that *R. rattus*, which is phylogenetically more distantly related to the other three rodents, contained a *Mycoplasma* community different from that in the *Mus-Mastomys* rodent clade (Figure 5, Panel A). Pathogen prevalence also varied considerably between sites, as shown for the six *Mycoplasma* OTUs (Figure 5, Panel B). This suggests that the infection risks for animals and humans vary greatly according to environmental characteristics and/or biotic features potentially related to recent changes in the distribution of rodent species in Senegal [68,69]

Perspectives

Recommendation for future experiments. Our experiments demonstrated the need to include many different kind of controls, at different steps, in order to avoid data misinterpretation. In particular, alien positive controls are important for establishing threshold values for OTUs positivity. These alien positive controls should include taxa distant enough from potential pathogens in order to avoid potential confusion between sequences of alien controls and sequences that result from actual infection of rodent samples. Ideally, one should choose alien positive controls from bacterial genera which are not able to infect the study's host species. In our study, the use of *Mycoplasma* and *Borrelia* species as alien positive controls was not ideal because both genera are potential rodent pathogens. Thankfully, the species used as alien controls could be easily distinguished from the species found in rodents on the basis of the phylogenetic analyses of the V4 sequences. However, based on our experience, we recommend using bacterial genera phylogenetically distant from pathogenic genera as alien controls when possible.

The inclusion of negative controls of DNA extraction in studies based on massive sequencing of 16S rRNA amplicons had long been overlooked, until the publication of Salter in 2014 [23] demonstrated the high pollution of laboratory reagents by bacterial DNA. Most studies published prior to this reported no extraction controls in their protocols. Here, we have performed one negative control for extraction per DNA extraction microplate; with each run consisting of four DNA extraction microplates, and each control having been analyzed in two replicate, we have a total of 8 negative controls for extraction per run which are analyzed twice. Based on our experience, we recommend performing at least this number of extraction controls per run. Further increases in the number of extraction controls per microplate would further improve the efficiency of data filtering and so the quality of the data produced.

The protocol of PCR amplification is also of importance for insuring data quality. In our study, we built separate amplicon libraries for each sample separately, and used very long PCR elongation times (5min) in order to mitigate the formation of chimeric reads [18] (also called jumping PCR). High numbers of PCR cycles are also known to increase chimera formation, yet as mentioned by Schnell et al [70], this parameter is mainly only critical when bulk amplification of pools of tagged/indexed amplicons is

performed (e.g. when using the Illumina TrueSeq library preparation kit). As we used separate amplicon libraries for each sample, we believe that the relatively high number of PCR cycles we used (40 cycles) had minimal impact on chimera formation, and this protocol ensures the absence of chimeric sequences between samples. We had chosen to maximize the number of cycles to enhance our ability to detect pathogenic bacteria, which are sometimes in low quantity in animal samples. Fine-tuning the balance between these parameters deserves further study.

Moreover, in our study we targeted the spleen to detect bacterial infections based on the fact that this organ is known to filter microbial cells in mammals. However we lack the data to be certain that the spleen is the best organ for giving a global picture of bacterial infection in rodents (and more broadly, vertebrates). We are currently conducting new experiments to address this issue.

Finally, in our experiments, about a third of OTU sequences were attributed neither to contamination nor to (known) pathogenic genera. We currently have no precise idea of the significance of the presence of these OTUs in the rodent spleens. Part of these OTUs could be linked to further undetected biases during data generation; in spite of all the precautions we have implemented here, other biases may still elude detection. Such biases could explain the very high numbers of rare OTUs (11,947 OTUs < 100 reads), which together represent more than 88% of the total number of OTUs but less than 1% of the total number of sequences (both runs combined).

Additionally, the presence of an OTU in a rodent spleen does not necessarily imply that the OTU is pathogenic. We know little about the microbiome of healthy vertebrates organs, yet the sharp increase of microbiome studies over the last few years has led to the discovery that microbiota communities appear to be specific to each part of the vertebrate's body, including internal tissues and blood [71] The OTUs detected in rodent’s spleen could thus simply be part of the healthy microbiome of the organ. These issues deserve better documentation. Our results thus pave the way for future research on unknown bacterial pathogens and the microbiome of healthy organs in vertebrates.

Improving HTS for epidemiological surveillance. The screening strategy described here has the considerable advantage of being non-specific, making it possible to detect unanticipated or novel bacteria. Razzauti *et al*. [8] recently showed

514 that the sensitivity of 16S rRNA amplicon sequencing on the MiSeq platform was
515 equivalent to that of whole RNA sequencing (RNAseq) on the HiSeq platform for
516 detecting bacteria in rodent samples. However, little is known about the comparative
517 sensitivity of HTS approaches relative to qPCR with specific primers, the current gold
518 standard for bacterial detection within biological samples. Additional studies are
519 required to address this question. Moreover, as 16S rRNA amplicon sequencing is
520 based on a short sequence, it does not yield a high enough resolution to distinguish
521 between species in some bacterial genera, such as *Bartonella*, nor to distinguishing
522 between pathogenic and non-pathogenic strains within the same bacterial species.
523 To get this information, we thus need to follow up the 16S rRNA amplicon
524 sequencing with complementary laboratory work. Whole-genome shotgun or RNAseq
525 techniques provide longer sequences, through the production of longer reads or the
526 assembly of contigs, and they might therefore increase the accuracy of species
527 detection [72]. However, these techniques would be harder to adapt for the extensive
528 multiplexing of samples [8]. Other methods could be used to assign sequences to
529 bacterial species or strains for samples found positive for a bacterial genus following
530 the 16S rRNA screening. For example, positive PCR assays could be carried out with
531 bacterial genus-specific primers, followed by amplicon sequencing, as commonly
532 used in MLSA (multilocus sequence analysis) strategies [73] or high-throughput
533 microfluidic qPCR assays based on bacterial species-specific primers could be used
534 [74]. High-throughput amplicon sequencing approaches could be fine-tuned to
535 amplify several genes for species-level assignment, such as the *gltA* gene used by
536 Gutierrez *et al*. [75] for the *Bartonella* genus, in parallel with the 16S rRNA-V4 region.

537 This strategy could also easily be adapted for other microbes, such as protists, fungi
538 and even viruses, provided that universal primers are available for their detection
539 (see [76,77] for protists and fungi, and [78] for degenerate virus family-level primers
540 for viruses). Finally, our filtering method could also be translated to any other post-
541 sequencing dataset of indexed or tagged amplicons in the framework of
542 environmental studies (e.g. metabarcoding for diet analysis and biodiversity
543 monitoring [79], the detection of rare somatic mutations [80] or the genotyping of
544 highly polymorphic genes (e.g. MHC or HLA typing, [81,82]).

545

Monitoring the risk of zoonotic diseases. Highly successful synanthropic wildlife species, such as the rodents studied here, will probably play an increasingly important role in the transmission of zoonotic diseases [83]. Many rodent-borne pathogens cause only mild or undifferentiated disease in healthy people, and these illnesses are often misdiagnosed and underreported [55,84-87]. The information about pathogen circulation and transmission risks in West Africa provided by this study is important in terms of human health policy. We show that rodents carry seven major pathogenic bacterial genera: *Borrelia*, *Bartonella*, *Mycoplasma, Ehrlichia*, *Rickettsia*, *Streptobacillus* and *Orientia.* The last five of these genera have never before been reported in West African rodents. The data generated with our HTS approach could also be used to assess zoonotic risks and to formulate appropriate public health strategies involving the focusing of continued pathogen surveillance and disease monitoring programs on specific geographic areas or rodent species likely to be involved in zoonotic pathogen circulation, for example.

Materials & Methods

Ethics statement. Animals were treated in accordance with European Union guidelines and legislation (Directive 86/609/EEC). The CBGP laboratory received approval (no. B 34-169-003) from the Departmental Direction of Population Protection (DDPP, Hérault, France), for the sampling of rodents and the storage and use of their tissues. None of the rodent species investigated in this study has protected status (see UICN and CITES lists).

Sample collection. We sampled rodents in 24 villages of the Sahelian and Sudanian climatic and biogeographical zones in Senegal (see Dalecky et al. [69] for details on the geographic location and other information on the villages). Rodents were sampled by live trapping according to the standardised protocol described by Dalecky et al. [69]. Briefly, traps were set within homes (one single-capture wire-mesh trap and one Sherman folding box trap per room) during one to five consecutive days. Each captured rodent was collected alive and transported to the field laboratory. There, rodents were killed by cervical dislocation, as recommended by Mills *et al.* [88] and dissected as described in Herbreteau *et al.* [89]. Rodent

577 species were identified by morphological and/or molecular techniques [69]. The
578 information concerning the rodent collection (sample ID, locality and species) is
579 provided in the Table S2. Cross-contamination during dissection was prevented by
580 washing the tools used successively in bleach, water and alcohol between rodents.
581 We used the spleen for bacterial detection, because this organ is a crucial site of
582 early exposure to bacteria [90]. Spleens were placed in RNAlater (Sigma) and stored
583 at 4°C for 24 hours and then at -20°C until their use for genetic analyses.

584 ***Target DNA region and primer design.*** We used primers with sequences
585 slightly modified from those of the universal primers of Kozich *et al*. [18] to amplify a
586 251-bp portion of the V4 region of the 16S rRNA gene (16S-V4F:
587 GTGCCAGCMGCCGCGGTAA; 16S-V4R: GGACTACHVGGGTWTCTAATCC). The
588 ability of these primers to hybridize to the DNA of bacterial zoonotic pathogens was
589 assessed by checking that there were low numbers of mismatched bases over an
590 alignment of 41,113 sequences from 79 zoonotic genera inventoried by Taylor et al
591 [1], extracted from the Silva SSU database v119 [44]. The FASTA file is available in
592 the Dryad Digital Repository http://dx.doi.org/10.5061/dryad.m3p7d [42].

593 We used a slightly modified version of the dual-index method of Kozich *et al*. [18] to
594 multiplex our samples. The V4 primers included different 8-bp indices (called i5 index
595 in the forward and i7 index in the reverse) and Illumina adapters (called P5 adapter in
596 the forward and P7 adapter in the reverse) in the 5' position. The combinations of 24
597 i5-indexed primers and 36 i7-indexed primers made it possible to identify 864
598 different PCR products loaded onto the same MiSeq flowcell. Each index sequence
599 differed from the others by at least two nucleotides, and each nucleotide position in
600 the sets of indices contained approximately 25% of each base, to prevent problems
601 due to Illumina low-diversity libraries (Table 1).

602 ***DNA extraction and PCRs.*** All pre-PCR laboratory manipulations were
603 conducted with filter tips under a sterile hood in a DNA-free room, i.e. room dedicated
604 to the preparation of PCR mix and equipped with hoods that are kept free of DNA by
605 UV irradiation and bleach treatment. DNA from bacterial isolates (corresponding to
606 DNA extracts from laboratory isolates of *Bartonella taylorii*, *Borrelia burgdorferi* and
607 *Mycoplasma mycoides*) was extracted in another laboratory, and PCRs from these
608 isolates were performed after the amplifications of the DNA from rodents to avoid

cross-contamination between samples and bacterial isolates. DNA was extracted with the DNeasy 96 Tissue Kit (Qiagen) with final elution in 200 µl of elution buffer. One extraction blank (NC_{ext}), corresponding to an extraction without sample tissue, was systematically added to each of the eight DNA extraction microplates. DNA was quantified with a NanoDrop 8000 spectrophotometer (Thermo Scientific), to confirm the presence of a minimum of 10 ng/µl of DNA in each sample. DNA amplification was performed in 5 µL of Multiplex PCR Kit (Qiagen) Master Mix, with 4 µL of combined i5 and i7 primers (3.5µM) and 2 µL of genomic DNA. PCR began with an initial denaturation at 95°C for 15 minutes, followed by 40 cycles of denaturation at 95°C for 20 s, annealing at 55°C for 15 s and extension at 72°C for 5 minutes, followed by a final extension step at 72°C for 10 minutes. PCR products (3 µL) were verified by electrophoresis in a 1.5% agarose gel. One PCR blank (NC_{PCR}), corresponding to the PCR mix with no DNA, was systematically added to each of the 18 PCR microplates. DNA was amplified in replicate for all wild rodent samples (*n*=711) (summary Table S1 and details by sample Table S2).

Library preparation and MiSeq sequencing. Two Illumina MiSeq runs were conducted. Run 1 included the PCR products (two or three replicates per sample) from wild rodents collected in north Senegal (148 *Mastomys erythroleucus* and 207 *Mus musculus*) plus the positive controls and the negative controls. Run 2 included the PCR products (two replicates per samples) from wild rodents collected in south Senegal (73 *Mastomys erythroleucus*, 93 *Mastomys natalensis* and 190 *Rattus rattus*) plus the positive controls and the negative controls. Full details on the composition of runs are given in Table S2. The MiSeq platform was chosen because it generates lower error rates than other HTS platforms [91]. The number of PCR products multiplexed was 823 for the first MiSeq run and 746 for the second MiSeq run (Table S2). Additional PCR products from other projects were added to give a total of 864 PCR products per run. PCR products were pooled by volume for each 96-well PCR microplate: 4 µL for rodents and controls, and 1.5 µL for bacterial isolates. Mixes were checked by electrophoresis on 1.5% agarose gels before their use to generate a ‘‘super-pool’’ of 864 PCR products for each MiSeq run. We subjected 100 µL of each ‘‘super-pool’’ to size selection for the full-length amplicon (V4 hypervariable region expected median size: 375 bp including primers, indexes and adaptors and 251bp excluding primers, indexes and adaptors), by excision from

a low-melting agarose gel (1.25%) to discard non-specific amplicons and primer dimers. A PCR Clean-up Gel Extraction kit (Macherey-Nagel) was used to purify the excised bands. DNA was quantified by using the KAPA library quantification kit (KAPA Biosystems) on the final library before loading on a MiSeq (Illumina) flow cell (expected cluster density: 700-800 K/mm^2) with a 500-cycle Reagent Kit v2 (Illumina). We performed runs of 2 x 251 bp paired-end sequencing, which yielded high-quality sequencing through the reading of each nucleotide of the V4 fragments twice after the assembly of reads 1 and reads 2. The raw sequence reads (.fastq format) are available in the Dryad Digital Repository http://dx.doi.org/10.5061/dryad.m3p7d [42].

Bioinformatic and taxonomic classification. MiSeq datasets were processed with mothur v1.34 [43] and with the MiSeq standard operating procedure (SOP) [18]. Briefly, the MiSeq SOP (http://www.mothur.org/wiki/MiSeq_SOP) allowed us to: 1) construct contigs of paired-end read 1 and read 2 using the make.contig command; 2) remove the reads with poor quality of assembly (> 275 bp); 3) align unique sequences on the SILVA SSU Reference alignment v119 [44]; 4) remove the misaligned, non-specific (eukaryotic) and chimeric reads (uchime program); 5) regroup the reads into Operational Taxonomic Units (OTUs) with a 3% divergence threshold; and 6) classify the OTUs using the Bayesian classifier included in mothur (bootstrap cutoff = 80%) and the Silva taxonomic file. At the end of the process, we obtained a table giving the number of reads for each OTU in line and each PCR product in column. For each OTU, the taxonomic classification (up to genus level) was provided. The abundance table generated by mothur for each PCR product and each OTU was filtered as described in the Results section. The most abundant sequence for each OTU in each sample was extracted from the sequence dataset with a custom-written Perl script (available in the Dryad Digital Repository http://dx.doi.org/10.5061/dryad.m3p7d [42]). The most abundant sequences for the 12 OTUs are available from GenBank (Accession Number KU697337 to KU697350). The sequences were aligned with reference sequences from bacteria of the same genus available from the SILVA SSU Ref NR database v119, using SeaView v4 [92]. We used a neighbor-joining method (bioNJ) to produce phylogenetic trees with a Kimura 2-Parameter model using SeaView software, and species were identified on the basis of the “closest phylogenetic species”. We also used our sequences for blast

675 analyses of GenBank (blastn against nucleotide collection (nr/nt) performed in
676 january 2016) to identify the reference sequences to which they displayed the highest
677 percentage identity. The raw abundance table, the mothur command lines, the
678 mothur output files, the Perl script and the FASTA files used for the phylogenetic
679 analyses are available in the Dryad Digital Repository
680 http://dx.doi.org/10.5061/dryad.m3p7d [42].

681

682 Acknowledgments

683 This study was funded by the French National Institute for Agricultural Research
684 (INRA) Meta-omics and microbial ecosystems metaprogram (Patho-ID project:
685 Rodent and tick pathobiomes), the ANR ENEMI (ANR-11-JSV7-0006) and supported
686 by the COST Action TD1303 (EurNegVec). We would like to thank Virginie Dupuy for
687 extracting DNA from bacterial cultures as well as Julie Sappa from Alex Edelman &
688 Associates, Jessie L Abbate and Petra Villette for improving the English writing.
689 Analyses were performed on the CBGP HPC computational platform. The funders
690 had no role in study design, data collection and analysis, the decision to publish, or
691 preparation of the manuscript.

692

693 Authors' contributions

694 The study was conceived and designed by MG and JFC. MG, AL, CT, LT, HV and
695 MR carried out the molecular biology procedures and validated the MiSeq data. MG,
696 EB, MB and ADG contributed to the development of bioinformatics methods and
697 validated taxonomic assignments. JFC and MTV coordinated the Patho-ID project
698 and CB and NC coordinated the ENEMI project. MG, JFC, LT, CB and NC analyzed
699 the data. MG and JFC wrote the manuscript. CB, NC, MR and MVT helped to draft
700 and to improve the manuscript. All the authors have read and approved the final
701 manuscript.

702

703

704

References

1. **Taylor LH, Latham SM, Woolhouse ME.** 2001. Risk factors for human disease emergence. Philos Trans R Soc Lond B Biol Sci 356: 983-989.

2. **King DA, Peckham C, Waage JK, Brownlie J, Woolhouse ME.** 2006. Epidemiology. Infectious diseases: preparing for the future. Science 313: 1392-1393.

3. **Grogan LF, Berger L, Rose K, Grillo V, Cashins SD, Skerratt LF. 2014.** Surveillance for emerging biodiversity diseases of wildlife. PLoS Pathog 10: e1004015.

4. **Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J. 2009.** Metagenomic pyrosequencing and microbial identification. Clin Chem 55: 856-866.

5. **Hugon P, Dufour JC, Colson P, Fournier PE, Sallah K, Raoult D.** 2015. A comprehensive repertoire of prokaryotic species identified in human beings. Lancet Infect Dis 15: 1211-1219.

6. **Rynkiewicz EC, Hemmerich C, Rusch DB, Fuqua C, Clay K. 2015.** Concordance of bacterial communities of two tick species and blood of their shared rodent host. Mol Ecol 24: 2566-2579.

7. **Gofton AW, Doggett S, Ratchford A, Oskam CL, Paparini A, Ryan U, Irwin P.** 2015. Bacterial Profiling Reveals Novel "Ca. Neoehrlichia", Ehrlichia, and Anaplasma Species in Australian Human-Biting Ticks. PLoS One 10: e0145449.

8. **Razzauti M, Galan M, Bernard M, Maman S, Klopp C, Charbonnel N, Vayssier-Taussat M, Eloit M, Cosson JF.** 2015. A Comparison between Transcriptome Sequencing and 16S Metagenomics for Detection of Bacterial Pathogens in Wildlife. PLoS Negl Trop Dis 9: e0003929.

9. **Cosson JF, Galan M, Bard E, Razzauti M, Bernard M, Morand S, Brouat C, Dalecky A, Bâ K, Charbonnel N, Vayssier-Taussat M.** 2015. Detection of Orientia sp. DNA in rodents from Asia, West Africa and Europe. Parasit Vectors 8: 172.

10. **Vayssier-Taussat M, Moutailler S, Michelet L, Devillers E, Bonnet S, Cheval J, Hebert C, Eloit M.** 2013. Next generation sequencing uncovers unexpected bacterial pathogens in ticks in western Europe. PLoS One 8: e81439.

11. **Williams-Newkirk AJ, Rowe LA, Mixson-Hayden TR, Dasch GA.** 2014. Characterization of the bacterial communities of life stages of free living lone star ticks. Amblyomma americanum). PLoS One 9: e102130.

12. **Williams-Newkirk AJ, Rowe LA, Mixson-Hayden TR, Dasch GA.** 2012. Presence, genetic variability, and potential significance of "Candidatus Midichloria mitochondrii" in the lone star tick Amblyomma americanum. Exp Appl Acarol 58: 291-300.

13. **Carpi G, Cagnacci F, Wittekindt NE, Zhao F, Qi J, Tomsho LP, Drautz DI, Rizzoli A, Schuster SC.** 2011. Metagenomic profile of the bacterial communities associated with Ixodes ricinus ticks. PLoS One 6: e25604.

748 14. **Vaumourin E, Vourc'h G, Gasqui P, Vayssier-Taussat M.** 2015. The
749 importance of multiparasitism: examining the consequences of co-infections
750 for human and animal health. Parasit Vectors 8: 545.

751 15. **Tollenaere C, Susi H, Laine AL.** 2016. Evolutionary and Epidemiological
752 Implications of Multiple Infection in Plants. Trends Plant Sci 21: 80-90.

753 16. **Vayssier-Taussat M, Albina E, Citti C, Cosson JF, Jacques MA, Lebrun MH,**
754 **Le Loir Y, M Ogliastro M, Petit MA, Roumagnac P, Candresse T.** 2014.
755 Shifting the paradigm from pathogens to pathobiome: new concepts in the
756 light of meta-omics. Front Cell Infect Microbiol 4: 29.

757 17**. Vayssier-Taussat M, Kazimirova M, Hubalek Z, Hornok S, Farkas R, Cosson**
758 **JF, Bonnet S, Vourch G, Gasqui P, Mihalca AD, Plantard O, Silaghi C,**
759 **Cutler S, Rizzoli A** 2015. Emerging horizons for tick-borne pathogens: from
760 the 'one pathogen-one disease' vision to the pathobiome paradigm. Future
761 Microbiol 10: 2033-2043.

762 18. **Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD.** 2013.
763 Development of a dual-index sequencing strategy and curation pipeline for
764 analyzing amplicon sequence data on the MiSeq Illumina sequencing
765 platform. Appl Environ Microbiol 79: 5112-5120.

766 19. **Claesson MJ, Wang Q, O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP,**
767 **O'Toole PW.** 2010. Comparison of two next-generation sequencing
768 technologies for resolving highly complex microbiota composition using
769 tandem variable 16S rRNA gene regions. Nucleic Acids Res 38: e200.

770 20. **Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D,**
771 **Knight R.** 2012. Experimental and analytical tools for studying the human
772 microbiome. Nat Rev Genet 13: 47-58.

773 21. **Sinha R, Abnet CC, White O, Knight R, Huttenhower C.** 2015. The microbiome
774 quality control project: baseline study design and future directions. Genome
775 Biology16: 276

776 22. **Kircher M, Heyn P, Kelso J. 2011.** Addressing challenges in the production and
777 analysis of illumina sequencing data. BMC Genomics 12: 382.

778 23. **Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P,**
779 **Parkhill J, Loman NJ, Walker AW.** 2014. Reagent and laboratory
780 contamination can critically impact sequence-based microbiome analyses.
781 BMC Biol 12: 87.

782 24. **Horvath A, Peto Z, Urban E, Vagvolgyi C, Somogyvari F.** 2013. A novel,
783 multiplex, real-time PCR-based approach for the detection of the commonly
784 occurring pathogenic fungi and bacteria. BMC Microbiol 13: 300.

785 25. **Caro-Quintero A, Ochman H. 2015.** Assessing the Unseen Bacterial Diversity in
786 Microbial Communities. Genome Biol Evol 7: 3416-3425.

787 26. **Walters W, Hyde ER, Berg-Lyons D, Ackermann G, Humphrey G, Parada A,**
788 **Gilbert JA, Jansson JK, Caporaso JG, Fuhrman JA, Apprill A, Knight B.**
789 2015. Improved bacterial 16S rRNA gene. V4 and V4-5. and fungal internal
790 transcribed spacer marker gene primers for microbial community surveys.
791 mSystems 1:e00009-15

792 27. **Kircher M, Sawyer S, Meyer M**. 2012. Double indexing overcomes inaccuracies
793 in multiplex sequencing on the Illumina platform. Nucleic Acids Res 40: e3.

794 28. **Esling P, Lejzerowicz F, Pawlowski J.** 2015. Accurate multiplexing and filtering
795 for high-throughput amplicon-sequencing. Nucleic Acids Res 43: 2513-2524.

796 29. **Bystrykh LV. 2012.** Generalized DNA barcode design based on Hamming
797 codes. PLoS One 7: e36852.

798 30. **Meyerhans A, Vartanian JP, Wain-Hobson S.** 1990. DNA recombination during
799 PCR. Nucleic Acids Res 18: 1687-1691.

800 31. **Paabo S, Irwin DM, Wilson AC.** 1990. DNA damage promotes jumping between
801 templates during enzymatic amplification. J Biol Chem 265: 4718-4721.

802 32. **Odelberg SJ, Weiss RB, Hata A, White R.** 1995. Template-switching during
803 DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic Acids Res
804 23: 2049-2057.

805 33. **Lahr DJ, Katz LA**. 2009. Reducing the impact of PCR-mediated recombination in
806 molecular evolution and environmental studies using a new-generation high-
807 fidelity DNA polymerase. Biotechniques 47: 857-866.

808 34. **Whiteford N, Skelly T, Curtis C, Ritchie ME, Lohr A, Zaranek AW, Abnizova I,**
809 **Brown C.** 2009. Swift: primary data analysis for the Illumina Solexa
810 sequencing platform. Bioinformatics 25: 2194-2199.

811 35. **Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N,**
812 **Owens SM, Berg-Lyons D, Betley J, Fraser L, Bauer M, Gormley N,**
813 **Gilbert JA, Smith G, Knight R.** 2012. Ultra-high-throughput microbial
814 community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:
815 1621-1624.

816 36. **Smith DP, Peay KG.** 2014. Sequence depth, not PCR replication, improves
817 ecological inference from next generation DNA sequencing. PLoS One 9:
818 e90234.

819 37. **Kilianski A, Haas JL, Corriveau EJ, Liem AT, Willis KL, Kadavy DR,**
820 **Rosenzweig CN, Minot SS.** 2015. Bacterial and viral identification and
821 differentiation by amplicon sequencing on the MinION nanopore sequencer.
822 Gigascience 4: 12.

823 38. **Callahan B, Proctor D, Relman D, Fukuyama J, Holmes S.** 2016.
824 Reproducible Research Workflow in R for the Analysis of Personalized Human
825 Microbiome Data. Pac Symp Biocomput 21: 183-194.

826 39. **Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T,**
827 **Peplies J, Ludwig W, Gloeckner FO.** 2014. The SILVA and "All-species
828 Living Tree Project. LTP)" taxonomic frameworks. Nucleic Acids Res 42:
829 D643-648.

830 40. **Westcott SL, Schloss PD.** 2015. De novo clustering methods outperform
831 reference-based methods for assigning 16S rRNA gene sequences to
832 operational taxonomic units. PeerJ 3: e1487.

833 41. **Sinha R, Abnet CC, White O, Knight R, Huttenhower C.** 2015. The microbiome
834 quality control project: baseline study design and future directions. Genome
835 Biol 16: 276.

836 42. **Galan M, Razzauti M, Bard E, Bernard M, Brouat C, Charbonnel N, Dehne-**
837 **Garcia A, Loiseau A, Tatard C, Tamisier L, Vayssier-Taussat M, Vignes H,**
838 **Cosson JF.** 2016. Data from: 16S rRNA amplicon sequencing for
839 epidemiological surveys of bacteria in wildlife: the importance of cleaning post-
840 sequencing data before estimating positivity, prevalence and co-infection.
841 *http://dx.doi.org/10.5061/dryad.m3p7d.*

842 43. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB,**
843 **Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B,**
844 **Thallinger GG, Van Horn DJ, Weber CF.** 2009. Introducing mothur: open-
845 source, platform-independent, community-supported software for describing
846 and comparing microbial communities. Appl Environ Microbiol 75: 7537-7541.

847 44. **Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J,**
848 **Gloeckner FO.** 2013. The SILVA ribosomal RNA gene database project:
849 improved data processing and web-based tools. Nucleic Acids Res 41: D590-
850 596.

851 45. **Gasparich GE, Whitcomb RF, Dodge D, French FE, Glass J, Williamson DL.**
852 2004. The genus Spiroplasma and its non-helical descendants: phylogenetic
853 classification, correlation with phenotype and roots of the Mycoplasma
854 mycoides clade. Int J Syst Evol Microbiol 54: 893-918.

855 46. **Gomez-Diaz E, Doherty PF, Jr., Duneau D, McCoy KD.** 2010. Cryptic vector
856 divergence masks vector-specific patterns of infection: an example from the
857 marine cycle of Lyme borreliosis. Evol Appl 3: 391-401.

858 47. **Sashida H, Sasaoka F, Suzuki J, Fujihara M, Nagai K, Kobayashi S,**
859 **Furuhama K, Harasawa R.** 2013. Two clusters among Mycoplasma
860 haemomuris strains, defined by the 16S-23S rRNA intergenic transcribed
861 spacer sequences. J Vet Med Sci 75: 643-648.

862 48. **Neimark H, Johansson KE, Rikihisa Y, Tully JG.** 2001. Proposal to transfer
863 some members of the genera Haemobartonella and Eperythrozoon to the
864 genus Mycoplasma with descriptions of 'Candidatus Mycoplasma haemofelis',
865 'Candidatus Mycoplasma haemomuris', 'Candidatus Mycoplasma haemosuis'
866 and 'Candidatus Mycoplasma wenyonii'. Int J Syst Evol Microbiol 51: 891-899.

867 49. **Neimark H, Peters W, Robinson BL, Stewart LB.** 2005. Phylogenetic analysis
868 and description of Eperythrozoon coccoides, proposal to transfer to the genus
869 Mycoplasma as Mycoplasma coccoides comb. nov. and Request for an
870 Opinion. Int J Syst Evol Microbiol 55: 1385-1391.

871 50. **Steer JA, Tasker S, Barker EN, Jensen J, Mitchell J, Stocki T, Chalker VL,**
872 **Hamon M.** 2011. A novel hemotropic Mycoplasma. hemoplasma. in a patient
873 with hemolytic anemia and pyrexia. Clin Infect Dis 53: e147-151.

874 51. **Pitcher DG, Nicholas RA.** 2005. Mycoplasma host specificity: fact or fiction? Vet
875 J 170: 300-306.

876 52. **Trape JF, Diatta G, Arnathau C, Bitam I, Sarih M, Belghyti D, Bouattour A,**
877 **Elguero E, Vial L, Mané Y, Baldé C, Pugnolle F, Chauvancy G, Mahé G,**
878 **Granjon L, Duplantier JM, Durand P, Renaud F.** 2013. The epidemiology
879 and geographic distribution of relapsing fever borreliosis in West and North
880 Africa, with a review of the Ornithodoros erraticus complex. Acari: Ixodida).
881 PLoS One 8: e78473.

882 53. **Rar VA, Pukhovskaya NM, Ryabchikova EI, Vysochina NP, Bakhmetyeva**
883 **SV, Zdanovskaia NI, Ivanov LI, Tikunova NV.** 2015. Molecular-genetic and
884 ultrastructural characteristics of 'Candidatus Ehrlichia khabarensis', a new
885 member of the Ehrlichia genus. Ticks Tick Borne Dis 6: 658-667.

886 54. **Perlman SJ, Hunter MS, Zchori-Fein E.** 2006. The emerging diversity of
887 Rickettsia. Proc Biol Sci 273: 2097-2106.

888 55. **Elliott SP. 2007.** Rat bite fever and Streptobacillus moniliformis. Clin Microbiol
889 Rev 20: 13-22.

890 56. **Izzard L, Fuller A, Blacksell SD, Paris DH, Richards AL, Aukkanit N, Nguyen**
891 **C, Jiang J, Fenwick S, Day NPJ, Graves S, Stenos J.** 2010. Isolation of a
892 novel Orientia species. O. chuto sp. nov.. from a patient infected in Dubai. J
893 Clin Microbiol 48: 4404-4409.

894 57. **Buffet JP, Kosoy M, Vayssier-Taussat M**. 2013. Natural history of Bartonella-
895 infecting rodents in light of new knowledge on genomics, diversity and
896 evolution. Future Microbiol 8: 1117-1128.

897 58. **Mediannikov O, Aubadie M, Bassene H, Diatta G, Granjon L, Fenollar F**. 2014.
898 Three new *Bartonella* species from rodents in Senegal. Int J Infec Dis
899 21S:335.

900 59. **Julius R, Bastos A, Brettschneider H, Chimimba C.** 2012. Dynamics of
901 Rodent-borne zoonotic diseases and their reservoir hosts: invasive *Rattus* in
902 South Africa. Proc 25th Vertebrate Pest Conference, Monterey, California,
903 USA.

904 60. **Bhatt KM, Mirza NB. 1992**. Rat bite fever: a case report of a Kenyan. East Afr
905 Med J 69: 542-543.

906 61. **Gray HH. 1967.** Squirrel bite fever. Trans R Soc Trop Med Hyg 61:857.

907 62. **Laudisoit A, Falay D, Amundala N, Akaibe D, de Bellocq JG, Van Houtte N,**
908 **Breno M, Verheyen E, Wilschut L, Parola P, Raoult D, Socolovshi C.**
909 2014. High prevalence of Rickettsia typhi and Bartonella species in rats and
910 fleas, Kisangani, Democratic Republic of the Congo. Am J Trop Med Hyg 90:
911 463-468.

912 63. **Dupont HT, Brouqui P, Faugere B, Raoult D.** 1995. Prevalence of antibodies to
913 Coxiella burnetti, Rickettsia conorii, and Rickettsia typhi in seven African
914 countries. Clin Infect Dis 21: 1126-1133.

915 64. **Ammar AM, Sabry MZ, Kirchhoff H.** 1980. Distribution of mycoplasmas in field
916 and laboratory rodents in Egypt. Z Versuchstierkd 22: 216-223.

917 65. **Parola P, Inokuma H, Camicas JL, Brouqui P, Raoult D.** 2001. Detection and
918 identification of spotted fever group Rickettsiae and Ehrlichiae in African ticks.
919 Emerg Infect Dis 7: 1014-1017.

920 66. **Ndip LM, Labruna M, Ndip RN, Walker DH, McBride JW.** 2009. Molecular and
921 clinical evidence of Ehrlichia chaffeensis infection in Cameroonian patients
922 with undifferentiated febrile illness. Ann Trop Med Parasitol 103: 719-725.

923 67. **Kelly DJ, Foley DH, Richards AL.** 2015. A Spatiotemporal Database to Track
924 Human Scrub Typhus Using the VectorMap Application. PLoS Negl Trop Dis
925 9: e0004161.

926 68. **Konecny A, Estoup A, Duplantier JM, Bryja J, Ba K, Galan M, Tatard C, Cosson JF.** 2013. Invasion genetics of the introduced black rat. Rattus rattus. in Senegal, West Africa. Mol Ecol 22: 286-300.

929 69. **Dalecky A, Ba K, Piry S, Lippens C, Diagne CA, Kane M, Sow A, Diallo M, Niang Y, Konecny A, Sarr N, Artige E, Charbonnel N, Granjon L, Duplantier JM, Brouat C.** 2015. Range expansion of the invasive house mouse *Mus musculus domesticus* in Senegal, West Africa: a synthesis of trapping data over three decades, 1983–2014. Mammal Review 45: 176-190

934 70 **Schnell IB, Bohmann K, Gilbert MTP.** 2015. Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. Mol Ecol Resources 15: 1289–1303

937 71 **Lloyd-Price J, Abu-Ali G, Huttenhower C.** (2016) The healthy human microbiome. Genome Medicine, 8:51.

939 72. **Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL.** 2016. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun 469: 967-977.

942 73. **Glaeser SP, Kampfer P.** 2015. Multilocus sequence analysis. MLSA. in prokaryotic taxonomy. Syst Appl Microbiol 38: 237-245.

944 74. **Michelet L, Delannoy S, Devillers E, Umhang G, Aspan A, Juremalm M, Chirico J, van der Wal FJ, Sprong H, Pihl TPB, Klitgaard K, Bødker R, Fach P, Moutailler S**. 2014. High-throughput screening of tick-borne pathogens in Europe. Front Cell Infect Microbiol 4: 103.

948 75. **Gutierrez R, Morick D, Cohen C, Hawlena H, Harrus S.** 2014. The effect of ecological and temporal factors on the composition of Bartonella infection in rodents and their fleas. ISME J 8: 1598-1608.

951 76. **Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM.** 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS One 4: e6372.

955 77. **Mueller RC, Gallegos-Graves LV, Kuske CR.** 2016. A new fungal large subunit ribosomal RNA primer for high-throughput sequencing surveys. FEMS Microbiol Ecol 92.

958 78. **Anthony SJ, Epstein JH, Murray KA, Navarrete-Macias I, Zambrana-Torrelio CM, Solovyov A, Ojeda-Flores R, Arrigo NC, Islam A, Ali Khan S, Hosseini P, Bogich TL, Olival KJ, Sanchez-Leon MD, Karesh WB, Goldstein T, Luby SP, Morse SS, Mazet JAK, Daszak P, Lipkin WI.** 2013. A strategy to estimate unknown viral diversity in mammals. MBio 4: e00598-00513.

964 79. **Galan M, Pages M, Cosson JF.** 2012. Next-generation sequencing for rodent barcoding: species identification from fresh, degraded and environmental samples. PLoS One 7: e48374.

967 80. **Robasky K, Lewis NE, Church GM.** 2014. The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet 15: 56-62.

969 81. **Galan M, Guivier E, Caraux G, Charbonnel N, Cosson JF**. 2010. A 454
970 multiplex sequencing method for rapid and reliable genotyping of highly
971 polymorphic genes in large-scale studies. BMC Genomics 11: 296.

972 82. **Lange V, Bohme I, Hofmann J, Lang K, Sauter J, Albrecht V, Andreas JM,**
973 **Baier DM, Nething J, Ehninger U, Schwarzelt C, Pingel J, Ehninger G,**
974 **Schmidt AH.** 2014. Cost-efficient high-throughput HLA typing by MiSeq
975 amplicon sequencing. BMC Genomics 15: 63.

976 83. **McFarlane R, Sleigh A, McMichael T.** 2012. Synanthropy of wild mammals as a
977 determinant of emerging infectious diseases in the Asian-Australasian region.
978 Ecohealth 9: 24-35.

979 84. **Meerburg BG, Singleton GR, Kijlstra A.** 2009. Rodent-borne diseases and their
980 risks for public health. Crit Rev Microbiol 35: 221-270.

981 85. **Civen R, Ngo V.** 2008. Murine typhus: an unrecognized suburban vectorborne
982 disease. Clin Infect Dis 46: 913-918.

983 86. **Watt G, Parola P.** 2003. Scrub typhus and tropical rickettsioses. Curr Opin Infect
984 Dis 16: 429-436.

985 87. **Vayssier-Taussat M, Moutailler S, Féménia F, Raymond P, Croce O, La**
986 **Scola B, Fournier PE, Raoult D.** 2016. Identification of new zoonotic
987 *Bartonella* species responsible for bacteremia in humans bitten by ticks.
988 Emerg Inf Dis 22:

989 88. **Mills JN, Childs J, Ksiazek TG, Peters CJ, Velleca WM.** 1995. Methods for
990 trapping and sampling small mammals for virologic testing. CDC, Atlanta.

991 89. **Herbreteau V, Rerkamnuaychoke W, Jittapalapong S, Chaval Y, Cosson JF,**
992 **Morand S.** 2011. Field and laboratory protocols for rodent studies. Kasetsart
993 University Press 46 p. http://www.ceropath.org/research/protocols

994 90. **Mebius RE, Kraal G.** 2005. Structure and function of the spleen. Nat Rev
995 Immunol 5: 606-616.

996 91. **D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya**
997 **M, Podar M, Quince C, Hall N.** 2016. A comprehensive benchmarking study
998 of protocols and sequencing platforms for 16S rRNA community profiling.
999 BMC Genomics 17: 55.

1000 92. **Gouy M, Guindon S, Gascuel O.** 2010. SeaView version 4: A multiplatform
1001 graphical user interface for sequence alignment and phylogenetic tree
1002 building. Mol Biol Evol 27: 221-224.

1003 93. **Lecompte E, Aplin K, Denys C, Catzeflis F, Chades M, Chevret P.** 2008
1004 Phylogeny and biogeography of African Murinae based on mitochondrial and
1005 nuclear gene sequences, with a new tribal classification of the subfamily. BMC
1006 Evol Biol 8:199.

1007 94. **Reiczigel J. 2003**. Confidence intervals for the binomial parameter: some new
1008 considerations. Statistics in Medicine 22:611-621.

1009

1010

Supplemental material

Figure S1. Taxonomic assignment of the V4 16S rRNA sequences in wild rodents and in negative controls for extraction and of PCR. The histograms show the percentage of sequences for the most abundant bacterial genera in the MiSeq run 1 and run 2. Notice the presence of several bacterial genera in the controls, which were likely due to the inherent contamination of laboratory reagents by bacterial DNA and which are thereafter called contaminant genera. These contaminant genera are also present (in lower percentage) in the rodent samples. The different in bacterial contaminant composition between run 1 and run 2 reflects the use of different kits manufactured at several months apart (Qiagen technical service, pers. com.). The differences in the pathogenic bacteria proportions and compositions between run 1 and run 2 reflects the different origins of the samples (A) run 1: *Mastomys erythroleucus* (n=148) and *Mus musculus* (n=207) from the north Senegal ; (B) run 2: *Mastomys erythroleucus* (n=73), *Mastomys natalensis* (n=93) et *Rattus rattus* (n=190) from the south Senegal).

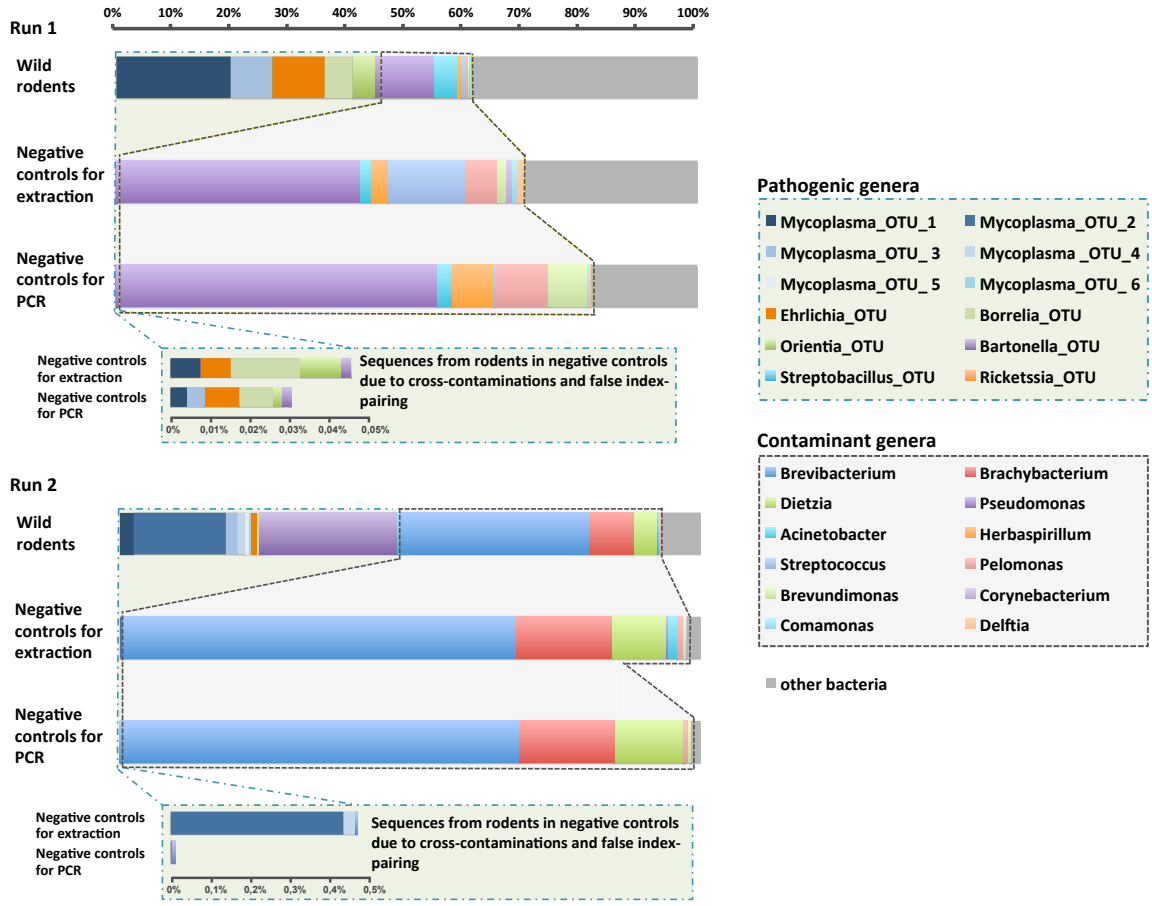
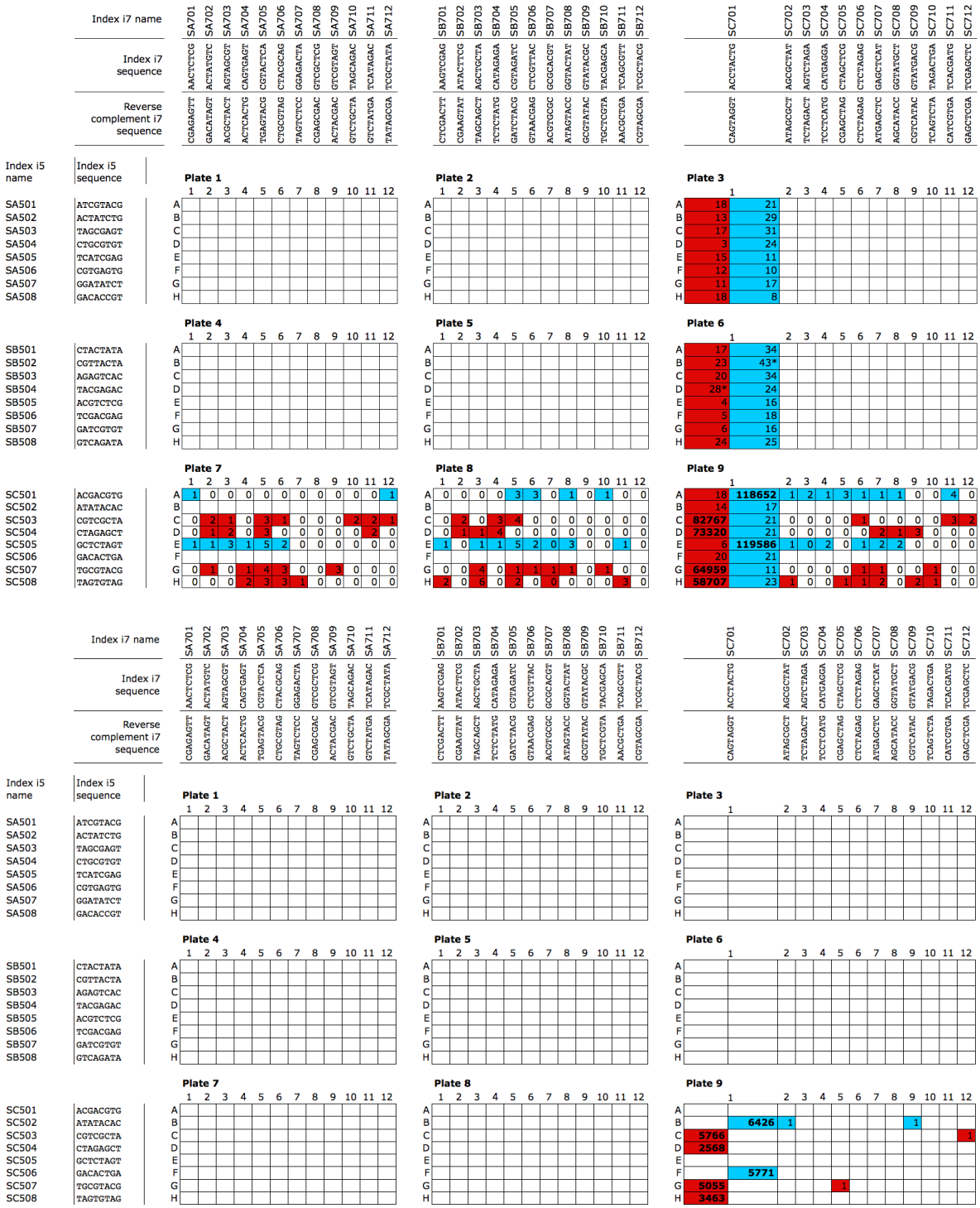
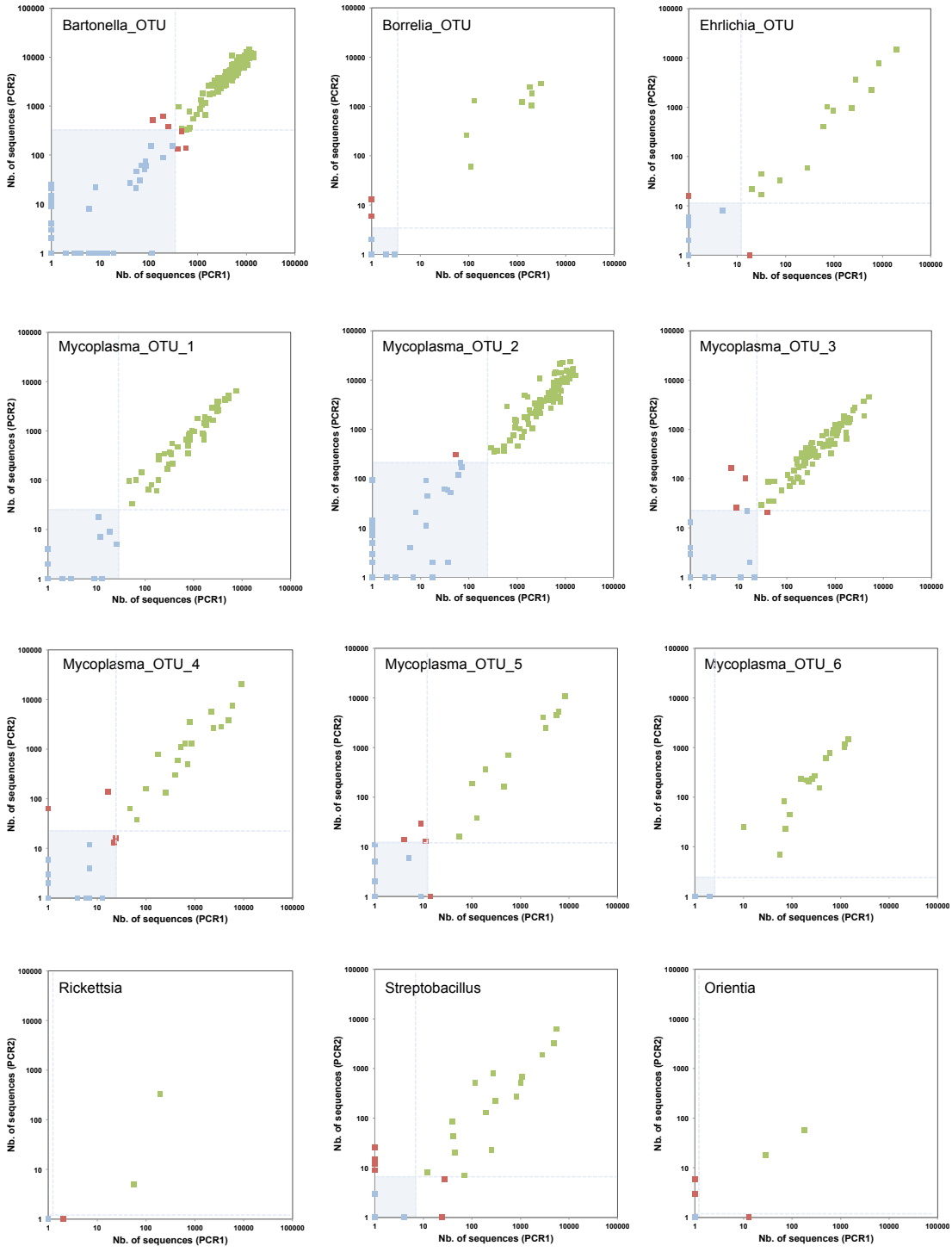


Figure S2. Numbers of sequences of the positive controls for indexing $PC_{Borrelia\_b}$ (in blue) and $PC_{Mycoplasma\_m}$ (in red) in the various PCR products, with a dual-indexing design, for MiSeq runs 1 (a) and 2 (b). The two PCRs for $PC_{Borrelia\_b}$ were performed with 96-well microplate 9, positions A1 and E1 for run 1 and B1 and F1 for run 2, and the four PCRs for $PC_{Mycoplasma\_m}$ were performed with 96-well microplate 9, positions C1, D1, G1 and H1 for the two runs. The numbers of sequences for the other wells correspond to indexing mistakes due to false index-pairing due to mixed clusters during the sequencing (see Table 1).



1014

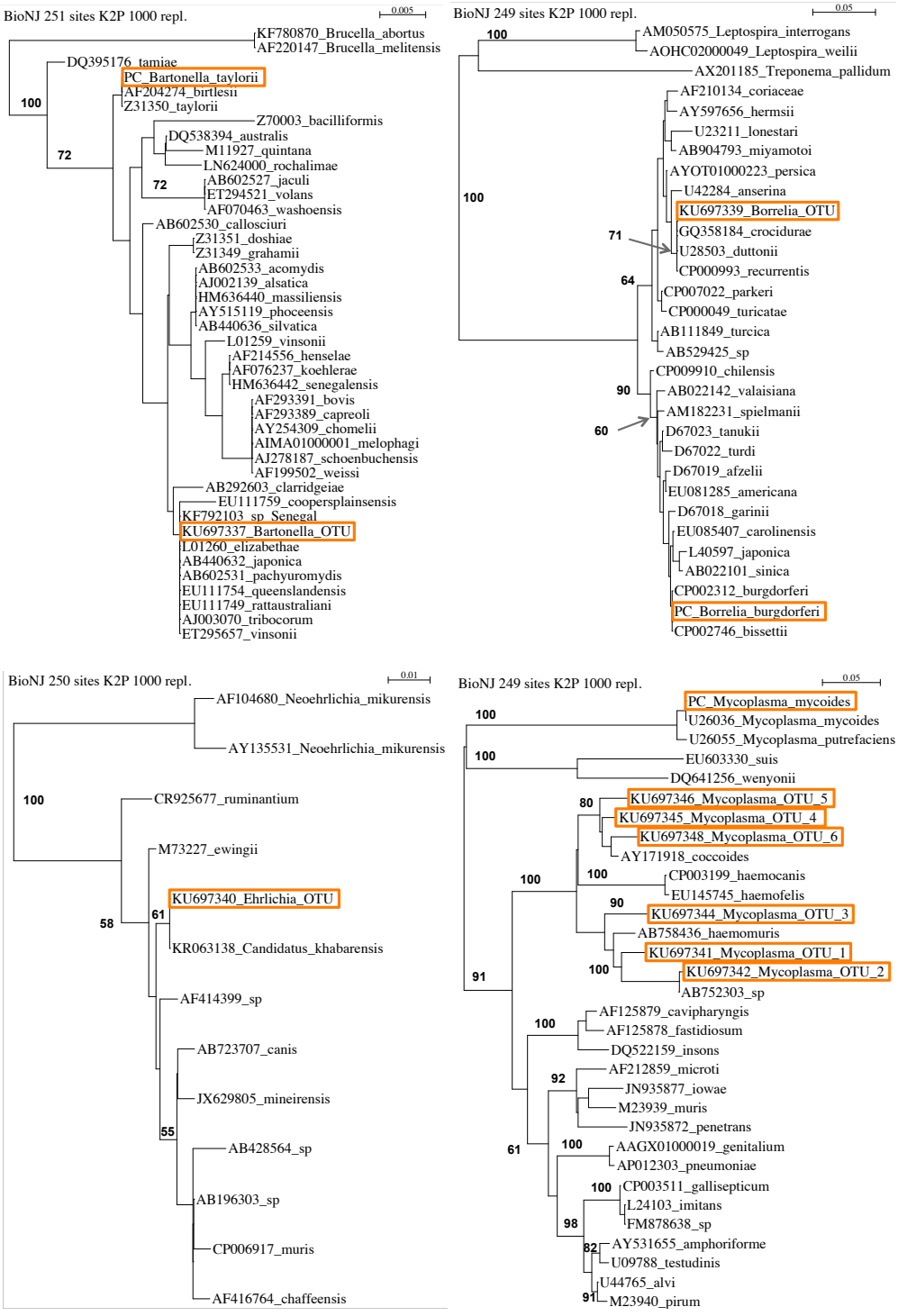
Figure S3. Plots of the number of sequences (log (x+1) scale) from bacterial OTUs in both PCR replicates (PCR1 & PCR2) for the 356 wild rodents analyzed in the second MiSeq run. Note that each rodent was tested with two replicate PCRs. Green points correspond to rodents with two positive results after the filtering process; red points correspond to rodents with one positive result and one negative result; and blue points correspond to rodents with two negative results. The light blue area and lines correspond to the threshold values used for the data filtering: samples below the lines are filtered out. See Figure 4 for plots corresponding to the first MiSeq run.



1015

1016

Figure S4. Phylogenetic trees of the 16S rRNA V4 sequences for 12 pathogenic bacterial OTUs detected in wild rodents from Senegal. Sequences boxed with an orange line were retrieved from African rodents and/or corresponds to positive controls (PC) for *Borellia burgdorferi*, *Mycoplasma mycoides* and *Bartonella taylorii*. The other sequences were extracted from the SILVA database and GenBank. Trees include all lineages collected for *Rickettsia*, *Bartonella*, *Ehrlichia* and *Orientia*, but only lineages of the Spotted Fever Group for *Borrelia*, and lineages of the pneumonia group for *Mycoplasma*. The numbers indicated are the bootstrap values >55%. Fasta files used have been deposited in the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.m3p7d.



1017

1018

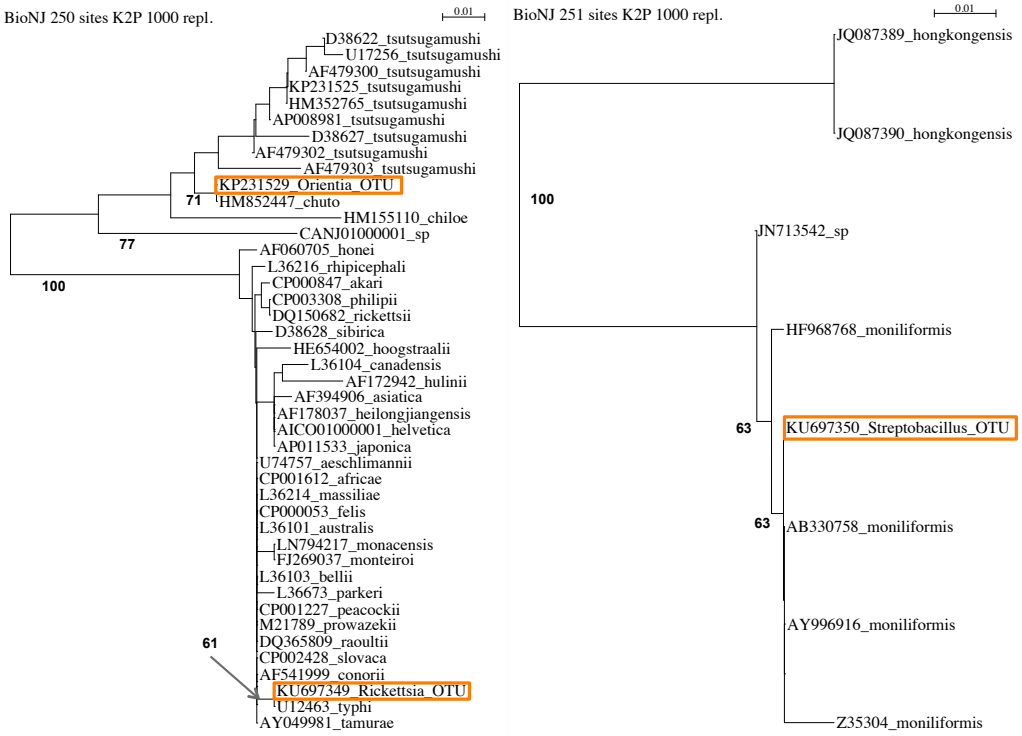


Table S3. The 50 most abundant OTUs in wild rodents and controls. The twelve pathogenic OTUs from wild rodents are in bold and italic. The two OTUs from PC_{alien} ($PC_{Borrelia\_b}$ & $PC_{Mycoplasma\_m}$) are highlighted in grey. A blank space was added at the end of the table to distinguish the first 50 most abundant OTUs and the *Mycoplasma*_OTU_6 and *Rickettsia*_OTU ranked in position 57 and 574 respectively.

						Run 1						Run 2				
OTU name	Phylum	Class	Order	Family	Genus	Total number of sequences (Run1 & Run 2)	Wild rodents	All Negative Controls (NC)	NC_{Bor} & NC_{PCR}	All Positive Controls (PC)	PC_{Calien} ($PC_{Borrelia\_b}$ & $PC_{Mycoplasma\_m}$)	Wild rodents	All Negative Controls (NC)	NC_{Bor} & NC_{PCR}	All Positive Controls (PC) (dilution 1/100th)	PC_{Calien} ($PC_{Borrelia\_b}$ & $PC_{Mycoplasma\_m}$) (dilution 1/100th)
Otu000001	Actinobacteria(100)	Actinobacteria(100)	Micrococcales(100)	Brevibacteriaceae(100)	*Brevibacterium(100)*	2 206 731	310	0	0	0	0	2 123 547	78 360	78 360	4 514	4 397
***Bartonella*_OTU**	**Proteobacteria(100)**	**Alphaproteobacteria(100)**	**Rhizobiales(100)**	**Bartonellaceae(100)**	***Bartonella(100)***	**1 761 155**	**67 973**	**3**	**2**	**134 151**	**27**	**1 547 662**	**3**	**3**	**11 373**	**76**
***Mycoplasma*_OTU_1**	**Tenericutes(100)**	**Mollicutes(100)**	**Mycoplasmatales(100)**	**Mycoplasmataceae(100)**	***Mycoplasma(100)***	**1 565 704**	**1 410 189**	**14**	**5**	**15**	**12**	**155 486**	**0**	**0**	**0**	**0**
***Mycoplasma*_OTU_2**	**Tenericutes(100)**	**Mollicutes(100)**	**Mycoplasmatales(100)**	**Mycoplasmataceae(100)**	***Mycoplasma(100)***	**1 036 084**	**0**	**0**	**0**	**0**	**0**	**1 035 890**	**193**	**193**	**1**	**1**
***Ehrlichia*_OTU**	**Proteobacteria(100)**	**Alphaproteobacteria(100)**	**Rickettsiales(100)**	**Anaplasmataceae(100)**	***Ehrlichia(100)***	**723 468**	**649 423**	**14**	**7**	**14**	**13**	**74 017**	**0**	**0**	**0**	**0**
Otu000004	Proteobacteria(100)	Gammaproteobacteria(100)	Pseudomonadales(100)	Pseudomonadaceae(100)	*Pseudomonas(100)*	700 806	645 274	50 018	41 700	469	465	4 864	178	178	3	3
***Mycoplasma*_OTU_3**	**Tenericutes(100)**	**Mollicutes(100)**	**Mycoplasmatales(100)**	**Mycoplasmataceae(100)**	***Mycoplasma(100)***	**634 967**	**507 369**	**2**	**2**	**5**	**3**	**127 590**	**1**	**1**	**0**	**0**
Otu000003	Actinobacteria(100)	Actinobacteria(100)	Micrococcales(100)	Dermabacteraceae(100)	*Brachybacterium(100)*	526 788	3 161	97	0	0	0	503 804	18 976	18 976	750	670
***Borrelia*_OTU**	**Spirochaetae(100)**	**Spirochaetes(100)**	**Spirochaetales(100)**	**Spirochaetaceae(100)**	***Borrelia(100)***	**367 509**	**345 845**	**20**	**11**	**5**	**7**	**21 636**	**0**	**0**	**0**	**0**
$PC_{Mycoplasma_m}$_OTU	Tenericutes(100)	Mollicutes(100)	Entomoplasmatales(100)	Entomoplasmataceae(100)	*Incertae_Sedis(100)*	297 005	338	2	0	279 811	279 777	2	0	0	16 852	16 852
*Orientia*_OTU	Proteobacteria(100)	Alphaproteobacteria(100)	Rickettsiales(100)	Rickettsiaceae(100)	*Orientia(100)*	280 272	279 957	5	5	3	1	307	0	0	0	0
Otu000007	Actinobacteria(100)	Actinobacteria(100)	Corynebacteriales(100)	Dietziaceae(100)	*Dietzia(100)*	275 814	809	34	0	0	0	262 593	12 032	12 032	346	327
Otu000009	Proteobacteria(100)	Gammaproteobacteria(100)	Pseudomonadales(100)	Moraxellaceae(100)	*Acinetobacter(100)*	262 495	248 538	1 556	1 366	6	6	11 678	711	711	6	6
Otu000010	Firmicutes(100)	Bacilli(100)	Lactobacillales(100)	Lactobacillaceae(100)	*Lactobacillus(100)*	258 007	244 861	995	1	5	5	12 146	0	0	0	0
$PC_{Borrelia_b}$_OTU	Spirochaetae(100)	Spirochaetes(100)	Spirochaetales(100)	Spirochaetaceae(100)	*Borrelia(100)*	250 969	420	0	0	238 352	238 314	0	0	0	12 197	12 197
Otu000011	Firmicutes(100)	Bacilli(100)	Lactobacillales(100)	Lactobacillaceae(100)	*Lactobacillus(100)*	232 870	219 167	600	4	0	0	13 103	0	0	0	0
Otu000012	Firmicutes(100)	Bacilli(100)	Lactobacillales(100)	Lactobacillaceae(100)	*Lactobacillus(100)*	150 241	136 796	2 467	1	3	3	10 975	0	0	0	0
Otu000017	Proteobacteria(100)	Gammaproteobacteria(100)	Enterobacteriales(100)	Enterobacteriaceae(100)	*unclassified(70)*	91 877	43 644	114	114	827	827	47 261	30	30	1	1
***Mycoplasma*_OTU_4**	**Tenericutes(100)**	**Mollicutes(100)**	**Mycoplasmatales(100)**	**Mycoplasmataceae(100)**	***Mycoplasma(100)***	**85 596**	**0**	**0**	**0**	**0**	**0**	**85 583**	**13**	**13**	**0**	**0**
Otu000018	Proteobacteria(100)	Gammaproteobacteria(100)	Pasteurellales(100)	Pasteurellaceae(100)	*unclassified(83)*	84 403	78 632	1 894	1 770	3	2	3 872	0	0	2	2
Otu000024	Firmicutes(100)	Bacilli(100)	Lactobacillales(100)	Leuconostocaceae(100)	*Weissella(100)*	82 171	77 033	50	4	3	3	5 085	0	0	0	0
Otu000028	Proteobacteria(100)	Betaproteobacteria(100)	Burkholderiales(100)	Oxalobacteraceae(100)	*Herbaspirillum(69)*	74 189	65 973	5 435	5 284	37	37	2 696	46	46	2	2
Otu000016	Proteobacteria(100)	Betaproteobacteria(100)	Burkholderiales(100)	Comamonadaceae(100)	*Pelomonas(75)*	73 751	50 487	8 966	7 973	64	64	12 874	1 350	1 350	10	10
Otu000023	Proteobacteria(100)	Epsilonproteobacteria(100)	Campylobacterales(100)	Helicobacteraceae(100)	*Helicobacter(97)*	72 142	71 317	0	0	1	1	824	0	0	0	0
Otu000019	Proteobacteria(100)	Gammaproteobacteria(100)	Enterobacteriales(100)	Enterobacteriaceae(100)	*unclassified(98)*	69 999	69 010	4	1	0	0	985	0	0	0	0
Otu000021	Proteobacteria(100)	Gammaproteobacteria(100)	Enterobacteriales(100)	Enterobacteriaceae(100)	*Yersinia(95)*	62 633	57 413	4 658	4 416	74	74	488	0	0	0	0
Otu000032	Bacteroidetes(100)	Bacteroidia(100)	Bacteroidales(100)	Bacteroidaceae(100)	*Bacteroides(100)*	56 644	11 247	1	1	0	0	45 396	0	0	0	0
***Mycoplasma*_OTU_5**	**Tenericutes(100)**	**Mollicutes(100)**	**Mycoplasmatales(100)**	**Mycoplasmataceae(100)**	***Mycoplasma(100)***	**56 324**	**0**	**0**	**0**	**0**	**0**	**56 324**	**0**	**0**	**0**	**0**
Otu000030	Bacteroidetes(100)	Bacteroidia(100)	Bacteroidales(100)	Bacteroidaceae(100)	*Bacteroides(100)*	50 179	46 668	18	0	1	0	3 492	0	0	0	0
Otu000025	Firmicutes(100)	Bacilli(100)	Lactobacillales(100)	Lactobacillaceae(100)	*Lactobacillus(100)*	46 332	41 591	266	1	0	0	4 475	0	0	0	0
Otu000027	Firmicutes(100)	Bacilli(100)	Bacillales(100)	Bacillaceae(100)	*Geobacillus(97)*	42 581	41 128	1 179	5	0	0	271	0	0	3	3
Otu000036	Firmicutes(100)	Bacilli(100)	Lactobacillales(100)	Streptococcaceae(100)	*Streptococcus(100)*	39 271	31 561	6 137	5 309	0	0	1 568	2	2	3	2
Otu000026	Proteobacteria(100)	Gammaproteobacteria(100)	Pseudomonadales(100)	Moraxellaceae(100)	*Acinetobacter(100)*	38 173	33 294	797	580	15	15	3 443	67	67	557	557
Otu000022	Firmicutes(100)	Bacilli(100)	Bacillales(100)	Staphylococcaceae(100)	*Staphylococcus(67)*	35 526	22 065	923	515	0	0	11 996	527	527	15	10
Otu000038	Actinobacteria(100)	Actinobacteria(100)	Bifidobacteriales(100)	Bifidobacteriaceae(100)	*Bifidobacterium(100)*	35 027	31 814	212	1	0	0	3 001	0	0	0	0
Otu000142	Proteobacteria(100)	Gammaproteobacteria(100)	Aeromonadales(100)	Aeromonadaceae(100)	*Aeromonas(100)*	34 429	34 331	0	0	0	0	98	0	0	0	0
Otu000061	Proteobacteria(100)	Deltaproteobacteria(100)	Myxococcales(100)	Myxococcaceae(96)	*unclassified(96)*	32 821	32 669	42	0	1	0	109	0	0	0	0
***Streptobacillus*_OTU**	**Fusobacteria(100)**	**Fusobacteriia(100)**	**Fusobacteriales(100)**	**Leptotrichiaceae(100)**	***Streptobacillus(100)***	**32 399**	**0**	**0**	**0**	**0**	**0**	**32 399**	**0**	**0**	**0**	**0**
Otu000039	Proteobacteria(100)	Alphaproteobacteria(100)	Caulobacterales(100)	Caulobacteraceae(100)	*Brevundimonas(100)*	23 719	15 160	3 778	3 672	0	0	4 306	473	473	2	2
Otu000125	Bacteroidetes(100)	Bacteroidia(100)	Bacteroidales(100)	Prevotellaceae(100)	*Alloprevotella(100)*	23 197	22 943	43	0	0	0	211	0	0	0	0
Otu000049	Deinococcus-Thermus(100)	Deinococci(100)	Thermales(100)	Thermaceae(100)	*Meiothermus(100)*	23 074	21 780	964	1	0	0	330	0	0	0	0
Otu000051	Proteobacteria(100)	Epsilonproteobacteria(100)	Campylobacterales(100)	Helicobacteraceae(100)	*Helicobacter(99)*	22 844	4 985	0	0	0	0	17 859	0	0	0	0
Otu000070	Bacteroidetes(100)	Bacteroidia(100)	Bacteroidales(100)	Prevotellaceae(100)	*Prevotella(99)*	22 649	19 711	890	2	0	0	2 048	0	0	0	0
Otu000045	Proteobacteria(100)	Gammaproteobacteria(100)	Enterobacteriales(100)	Enterobacteriaceae(100)	*Proteus(99)*	22 344	87	0	0	0	0	22 257	0	0	0	0
Otu000033	Proteobacteria(100)	Alphaproteobacteria(100)	Sphingomonadales(100)	Sphingomonadaceae(99)	*Sphingomonas(59)*	19 819	15 822	519	246	13	13	3 330	135	135	0	0
Otu000058	Proteobacteria(100)	Gammaproteobacteria(100)	Xanthomonadales(100)	Xanthomonadaceae(100)	*Luteimonas(99)*	16 794	15 899	722	2	0	0	173	0	0	0	0
Otu000076	Bacteroidetes(100)	Bacteroidia(100)	Bacteroidales(100)	S24-7(100)	*unclassified(100)*	16 544	15 459	206	1	0	0	879	0	0	0	0
Otu000040	Firmicutes(100)	Bacilli(100)	Bacillales(100)	Planococcaceae(98)	*unclassified(58)*	16 005	12 278	1 819	1	0	0	1 901	7	7	0	0
Otu000050	Bacteroidetes(100)	Bacteroidia(100)	Bacteroidales(100)	Bacteroidaceae(100)	*Bacteroides(100)*	15 883	14 752	0	0	0	0	1 131	0	0	0	0
Otu000043	Firmicutes(100)	Clostridia(100)	Clostridiales(100)	Ruminococcaceae(100)	*unclassified(99)*	15 284	14 210	127	0	0	0	947	0	0	0	0
Mycoplasma*_OTU_6	**Tenericutes(100)**	**Mollicutes(100)**	**Mycoplasmatales(100)**	**Mycoplasmataceae(100)**	***Mycoplasma(100)***	**13 356**	**0**	**0**	**0**	**0**	**0**	**13 356**	**0**	**0**	**0**	**0**
Rickettsia*_OTU*	**Proteobacteria(100)**	**Alphaproteobacteria(100)**	**Rickettsiales(100)**	**Rickettsiaceae(100)**	***Rickettsia(100)***	**593**	**4**	**0**	**0**	**0**	**0**	**589**	**0**	**0**	**0**	**0**
Other OTUs	/	/	/	/	/	1 668 136	1 376 047	57 967	12 205	3 650	438	228 260	1 461	1 461	751	138
					Total number of sequences:	14 647 593	7 149 444	153 558	85 208	657 531	520 107	6 525 107	114 565	114 565	47 388	35 246

* *Mycoplasma*_OTU_6 is ranked in position 57
** *Rickettsia*_OTU is ranked in position 574

1025

Table S1. Numbers of samples and numbers of PCRs for wild rodents and controls. Negative Controls for dissection, NC_{mus} ; Negative Controls for extraction, NC_{ext} ; Negative Controls for PCR, NC_{PCR} ; Negative Controls for indexing, NC_{index} ; Positive Controls for PCR, PC_{PCR} ; Positive Controls for Indexing, PC_{alien}. See also Figure 1 for more details concerning negative controls (NC) and positive controls (PC). See also Figure 1 and Box 1.

MiSeq run	Types of samples	Number of samples	Number of PCRs*
Run 1	Wild rodents	355	790
	PC_{PCR}: *Bartonella taylorii* (no dilution)	1	2
	PC_{PCR}/PC_{alien}: *Borrelia burgdorferi* (no dilution)	1	2
	PC_{PCR}/PC_{alien}: *Mycoplasma mycoides* (no dilution)	1	4
	NC_{mus}	4	8
	NC_{ext}	4	8
	NC_{PCR}	/	9
Run 2	Wild rodents	356	712
	PC_{PCR}: *Bartonella taylorii* (dilution: 1/100th)	1	2
	PC_{PCR}/PC_{alien}: *Borrelia burgdorferi* (dilution: 1/100th)	1	2
	PC_{PCR}/PC_{alien}: *Mycoplasma mycoides* (dilution: 1/100th)	1	4
	NC_{ext}	4	8
	NC_{PCR}	/	9
	NC_{index}	/	9
	Total:	729	1569

*PCR was performed in replicate for rodent samples and controls

1026

1027

1028

Table S4. Bacterial contaminants observed in negative ad positive controls. They were identified as contaminants on the basis of negative controls for extraction and PCR. Taxa in bold correspond to the sequences of DNA extracted from laboratory isolates.

Run name	Negative and positive controls (no. of PCR replicates)	Number of sequences				Taxon (frequency)
		Total	Mean	Min.	Max.	
Run 1	*Bartonella taylorii* (*n*=2), no dilution	137424	68712	64290	73134	***Bartonella* (0.975)**, *Propionibacterium* (0.023), other bacteria (0.002)
	Borrelia burgdorferi (*n*=2), no dilution	239465	119733	118913	120552	***Borrelia* (0.995)**, other bacteria (0.005)
	Mycoplasma mycoides (*n*=4), no dilution	280642	70161	58896	82933	**Entomoplasmataceae° (0.997)**, other bacteria (0.003)
	NC_{ext} (*n*=8)	39308	4914	2843	8967	*Pseudomonas** (0.42), *Streptococcus** (0.134), *Pelomonas** (0.054), *Haemophilus* (0.042), *Yersinia* (0.029), *Herbaspirillum** (0.028), *Granulicatella* (0.02), *Acinetobacter** (0.019), *Actinomyces* (0.017), *Brevundimonas** (0.016), *Veillonella* (0.013), *Staphylococcus* (0.013), *Delftia** (0.013), Comamonadaceae* (0.012), Pasteurellaceae (0.012), *Porphyromonas* (0.011), *Corynebacterium** (0.011), *Gemella* (0.01), *other bacteria (0.126)*
	NC_{mus} (*n*=8)	68350	8544	32*	26211	*Pseudomonas** (0.121), *Lactobacillus* (0.063), Bacillales* (0.037), Planococcaceae (0.033), *Microvirga* (0.031), Bacteroidales (0.028), Thermomicrobia (0.027), Lachnospiraceae (0.027), Nonomuraea (0.026), *Geodermatophilus** (0.023), *Sphingobacterium* (0.022), *Prevotella* (0.022), *Blautia* (0.019), *Pseudonocardia* (0.017), Geodermatophilaceae* (0.017), *Geobacillus* (0.017), *Meiothermus* (0.014), *Defluviimonas* (0.013), *Streptococcus** (0.013), *Pelomonas** (0.012), *Luteimonas* (0.01), other bacteria (0.408)
	NC_{PCR} (*n*=9)	45900	5100	3144	8002	*Pseudomonas** (0.552), *Pelomonas** (0.092), *Herbaspirillum** (0.072), *Brevundimonas** (0.067), *Yersinia* (0.065), *Acinetobacter** (0.026), other bacteria (0.125)
Run 2	*Bartonella taylorii* (*n*=2), dilution: 1/100th	12142	6071	4624	7518	***Bartonella* (0.928)**, *Propionibacterium* (0.042), *Brevibacterium*^ (0.013), other bacteria (0.017)
	Borrelia burgdorferi (*n*=2), dilution: 1/100th	13378	6689	6214	7164	***Borrelia* (0.912)**, *Acinetobacter** (0.046), *Brevibacterium*^ (0.036), other bacteria (0.006)
	Mycoplasma mycoides (*n*=4), dilution: 1/100th	21868	5467	4104	6520	**Entomoplasmataceae° (0.771)**, *Brevibacterium*^ (0.179), *Brachybacterium*^ (0.028), *Dietzia*^ (0.014), other bacteria (0.007)
	NC_{ext} (*n*=8)	53334	6667	5275	7669	*Brevibacterium*^ (0.679), *Brachybacterium*^ (0.166), *Dietzia*^ (0.093), *Acinetobacter** (0.015), *Pelomonas** (0.011), other bacteria (0.036)
	NC_{index} (*n*=9)	52	6	1	12	NA
	NC_{PCR} (*n*=8)	61231	7654	5855	9145	*Brevibacterium*^ (0.689), *Brachybacterium*^ (0.165), *Dietzia*^ (0.117), other bacteria (0.029)

° sequences of *Mycoplasma mycoides* were identified as Entomoplasmataceae due to a frequent taxonomic error present in most databases [44]
* taxa identified as reagent contaminants by Salter *et al*. [23]
^ taxa identified as PCR kit contaminants (Qiagen, personal communication)

1029

1030

1031

1032

1033

Table S5. Proportion of sequences and proportion of positive results removed at each step in data filtering. Note that severa positive results may be recorded for the same rodent in cases of co-infection.

	OTUs of interest	Sequences*					Positive results				
		No. before filtering	% removed from previous step			% removed (total)	No. before filtering	% removed from previous step			% removed (total)
			T_{CC}	T_{FA}	PCR Replicates			T_{CC}	T_{FA}	PCR Replicates	
Run 1	Mycoplasma_OTU_1	1226193	0,01%	0,36%	0,14%	0,51%	297	22%	78%	4%	83%
	Mycoplasma_OTU_3	507237	0,02%	0,27%	0,06%	0,35%	265	20%	75%	4%	80%
	Ehrlichia_OTU	644244	0,04%	0,34%	0,17%	0,55%	283	36%	72%	8%	83%
	Borrelia_OTU	319305	0,14%	0,34%	0,03%	0,50%	238	69%	62%	4%	89%
	Orientia_OTU	242299	0,04%	0,25%	0,40%	0,69%	199	36%	59%	12%	77%
	Bartonella_OTU	67921	0,07%	0,71%	0,14%	0,91%	124	32%	87%	18%	93%
Run 2	Mycoplasma_OTU_1	155486	0,00%	0,10%	0,00%	0,10%	74	0%	31%	0%	31%
	Mycoplasma_OTU_2	1035890	0,10%	0,05%	0,03%	0,18%	177	47%	3%	1%	49%
	Mycoplasma_OTU_3	127590	0,00%	0,13%	0,26%	0,40%	103	6%	10%	5%	19%
	Mycoplasma_OTU_4	85583	0,08%	0,04%	0,29%	0,41%	30	27%	0%	14%	37%
	Mycoplasma_OTU_5	56324	0,00%	0,12%	0,17%	0,29%	26	0%	38%	31%	58%
	Mycoplasma_OTU_6	13356	0,00%	0,01%	0,00%	0,01%	17	0%	6%	0%	6%
	Ehrlichia_OTU	74017	0,00%	0,05%	0,05%	0,09%	24	0%	38%	13%	46%
	Borrelia_OTU	21636	0,00%	0,05%	0,09%	0,13%	15	0%	33%	20%	47%
	Orientia_OTU	307	0,00%	0,00%	7,17%	7,17%	5	0%	0%	60%	60%
	Bartonella_OTU	1547652	0,01%	0,22%	0,19%	0,42%	246	26%	24%	4%	47%
	Streptobacillus_OTU	32399	0,00%	0,06%	0,46%	0,52%	29	0%	17%	33%	45%
	Rickettsia_OTU	589	0,00%	0,00%	0,34%	0,34%	3	0%	0%	33%	33%

*:sum of sequences in both duplicates

T_{CC} based on the maximum number of sequences observed in a control for each OTU in each run

T_{FA} based on the false assignment rate (0.02%) weighted by the total number of sequences for each OTU in each run

1034

1035

Table S6. Proportion of positive results for both PCR products at each step in data filtering. Note that several positive results may be recorded for the same rodent in cases of co-infection.

	OTUs of interest	% of rodents positive for both PCR replicates		
		Before filtering	T_{CC}	T_{FA}
Run 1	Mycoplasma_OTU_1	68%	64%	96%
	Mycoplasma_OTU_3	49%	46%	96%
	Ehrlichia_OTU	56%	56%	92%
	Borrelia_OTU	38%	53%	96%
	Orientia_OTU	43%	54%	88%
	Bartonella_OTU	19%	20%	82%
Run 2	Mycoplasma_OTU_1	76%	76%	100%
	Mycoplasma_OTU_2	59%	96%	99%
	Mycoplasma_OTU_3	86%	92%	95%
	Mycoplasma_OTU_4	77%	91%	82%
	Mycoplasma_OTU_5	62%	62%	69%
	Mycoplasma_OTU_6	94%	94%	100%
	Ehrlichia_OTU	58%	58%	87%
	Borrelia_OTU	53%	53%	80%
	Orientia_OTU	40%	40%	40%
	Bartonella_OTU	66%	83%	96%
	Streptobacillus_OTU	59%	59%	67%
	Rickettsia_OTU	67%	67%	67%

T_{CC} based on the maximum number of sequences observed in a control for each OTU in each run

T_{FA} based on the false assignment rate (0.02%) weighted by the total number of sequences for each OTU in each run

1036

1037

Table S7. Number of mismatches between PCR forward and reverse primers and 41,113 bacterial 16S rRNA V4 sequences of 79 zoonotic genera. Bacterial genera were selected according to the inventory of Taylor et al [1] and sequences were extracted from the Silva SSU database v119. Numbers of mismatches > 3 correspond to sequences of bad quality from diverse taxa. The number of primer mismatches in the 10 bases of the 3' side was ≤ 2 for 99.93% of the reference sequences.

Forward primer		Reverse primer	
No. of mismatches	No. of sequences	No. of mismatches	No. of sequences
0	40063	0	39901
1	841	1	967
2	101	2	132
3	42	3	43
4	8	4	24
5	8	5	8
6	6	6	4
7	3	7	4
8	2	8	4
9	1	9	1
10	4	10	1
11	0	11	3
12	0	12	1
13	0	13	0
14	0	14	1
15	0	15	0
16	0	16	0
17	0	17	0
18	0	18	0
19	0	19	0
NA*	34	20	0
		21	0
		22	0
		NA*	19

* Partial sequences for the primer region

1038