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], Hélène 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

Importance

Several recent public health crises have shown that the surveillance of zoonotic agents in wildlife is important to prevent pandemic risks. Rodents are intermediate hosts for numerous zoonotic bacteria. High-throughput sequencing (HTS) technologies are very useful for the detection and surveillance of zoonotic bacteria, but rigorous experimental processes are required for the use of these cheap and 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 bacterial positivity, prevalence and coinfection, by 16S rRNA amplicon sequencing on the MiSeq platform. This procedure, applied to 711 commensal rodents collected from 24 villages in Senegal, Africa, detected several emerging bacterial genera, some in high prevalence, while never before reported for West Africa. This study constitutes a step towards the use of HTS to improve our understanding of the risk of zoonotic disease transmission posed by wildlife, by providing a new strategy for the

use of HTS platforms to monitor both bacterial diversity and infection dynamics in wildlife. In the future, this approach could be adapted for the monitoring of other microbes such as protists, fungi, and even viruses.

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 of recent developments in 16S rRNA amplicon sequencing to facilitate the multiplexing of large numbers of samples, 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 eliminate the biases generated by HTS methods. We present here an extensive review of these biases and of their consequences, and we propose a trimming strategy for managing them and cleaning the dataset. We also analyzed 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: *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%, depending on the bacterial taxon, rodent species and site considered, and a mean of 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, thus, particularly suitable for continuous pathogen surveillance in the framework of disease monitoring programs

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 human health interest. 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 natures, including non-cultivable bacteria. Such metagenomics 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, with the sequences obtained then compared with those in a reference sequence database [4]. Metagenomics has made a major contribution 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, 90]. However, improvements in the estimation of infectious 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 MiSeq Illumina sequencing platform. The 16S rRNA amplicon sequencing technique is based on the amplification of small fragments of the hypervariable region of the 16S rRNA gene. The sequences of these fragments are then obtained and compared with those in a dedicated database, for taxonomic identification [4,19]. Multiplexed approaches of this kind include short indices (or tags) specific to a PCR product. 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 good characterization of bacterial diversity within each individual samples it provides. However, the 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 necessary mean that the bacterium is effectively 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*. Rodents were screened for bacteria as described by Kozich *et al.* [18], in a protocol based on MiSeq sequencing (Illumina) of the V4 hypervariable region of the 16SrRNA gene. 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 analyzing negative controls (NC) and positive controls (PC) at the various stages of the experiment, and systematically using sample replicates. This strategy delivers realistic and reliable

Table 1. Sources of bias during the experimental and bioinformatic steps of 16S rRNA amplicon sequencing, consequences for data interpretation and solutions for decreasing 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. Use of a bioinformatic strategy to remove the chimeric sequences (e.g., Uchime program)
	Poor or biased amplification [45]	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)

estimates of bacterial prevalence in wildlife populations, and could be used to analyze the co-occurrence of different bacterial species within individuals.

Results & Discussion

Raw sequencing results. The sequencing of 1569 PCR products (from rodent samples and controls, see details in Table S1) in two MiSeq runs generated a total of 23,698,561 raw paired-end sequence reads (251-bp) of the V4 region of the 16SrRNA gene. Overall, 99% of wild rodent PCRs generated more than 3,000 raw reads (mean: 11,908 reads; standard deviation: 6,062). The raw sequence read files in FASTQ format are available for each PCR and each MiSeq run on request to the corresponding author. Using mothur v1.34 [42] 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.

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 S2). 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 genuine 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 [44]. 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. This threshold corresponds to 99% of the distribution of the numbers of sequences between PCR products.

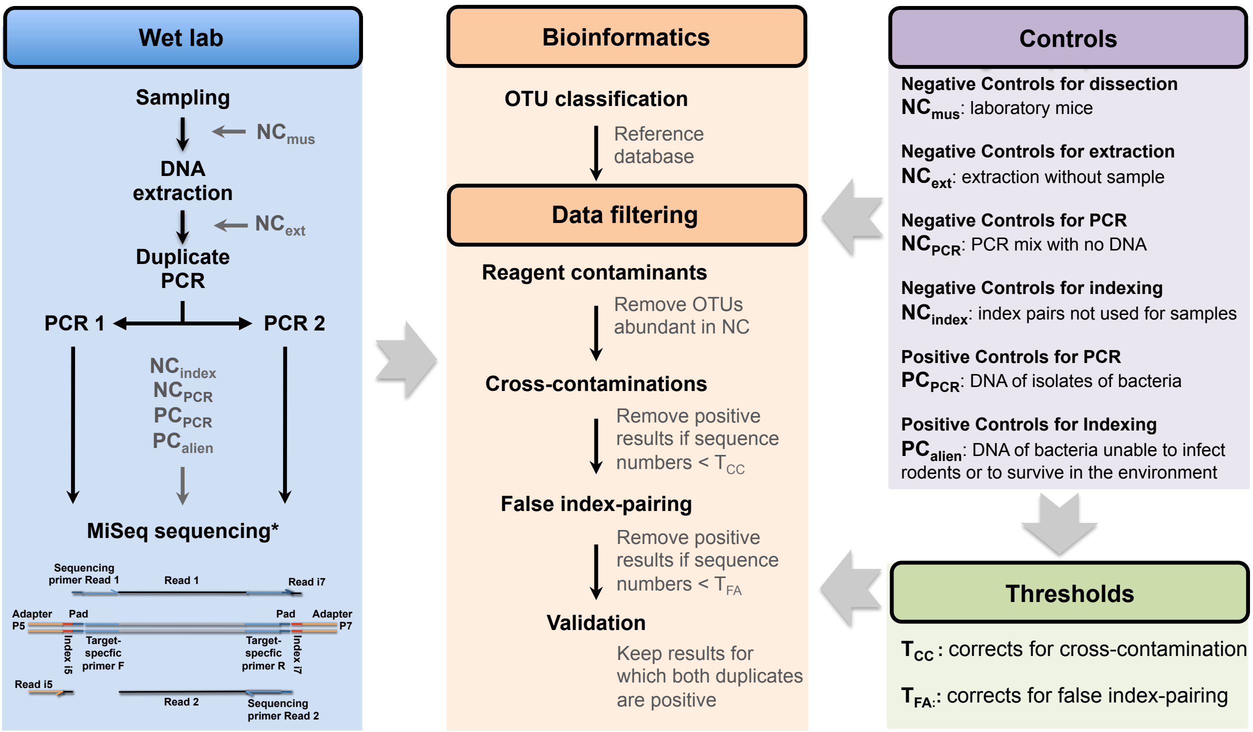
Wet lab
Sampling
NC_{mus}
DNA extraction
NC_{ext}
Duplicate PCR
PCR 1
PCR 2
NC_{index}
NC_{PCR}
PC_{PCR}
PC_{alien}
MiSeq sequencing*
Sequencing primer Read 1
Read 1
Read i7
Adapter P5
Pad
Pad
Adapter P7
Index i5
Target-specfic primer F
Target-specfic primer R
Index i7
Read i5
Read 2
Sequencing primer Read 2
Bioinformatics
OTU classification
Reference database
Data filtering
Reagent contaminants
Remove OTUs abundant in NC
Cross-contaminations
Remove positive results if sequence numbers < T_{CC}
False index-pairing
Remove positive results if sequence numbers < T_{FA}
Validation
Keep results for which both duplicates are positive
Controls
Negative Controls for dissection
NC_{mus}: laboratory mice
Negative Controls for extraction
NC_{ext}: extraction without sample
Negative Controls for PCR
NC_{PCR}: PCR mix with no DNA
Negative Controls for indexing
NC_{index}: index pairs not used for samples
Positive Controls for PCR
PC_{PCR}: DNA of isolates of bacteria
Positive Controls for Indexing
PC_{alien}: DNA of bacteria unable to infect rodents or to survive in the environment
Thresholds
T_{CC} : corrects for cross-contamination
$T_{FA:}$: corrects for false index-pairing

Figure 1. Workflow of the wet laboratory, and for bioinformatics and data filtering procedures, and a list of controls and thresholds included in the process of data filtering for the elimination of false-positive results for 16S rRNA amplicon sequencing. Reagent contaminants were detected by analyzing the sequences in the NC_{ext} and NC_{PCR}; T_{CC}: sequence number threshold for correcting for cross-contamination. T_{CC} values are OTU- and run-dependent and were estimated by analyzing the sequences in the controls, NC_{mus}, NC_{ext}, NC_{PCR} and PC_{index}; T_{FA}: 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}. 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.

Filtering for reagent contaminants. Metagenomics 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. This made it possible to identify the most abundant contaminants, including *Pseudomonas*, *Acinetobacter*, *Herbaspirillum*, *Streptococcus*, *Pelomonas*, *Brevibacterium*, *Brachybacterium*, *Dietzia*, *Brevundimonas*, *Delftia*, *Comamonas*, *Corynebacterium*, and *Geodermatophilus,* which accounted for 29% of the sequences in the dataset (Table S3). The bacterial contaminants detected differed between MiSeq runs: *Pseudomonas*, *Pelomonas* and Herbaspirillum predominated in run 1, whereas *Brevibacterium*, *Brachybacterium* and *Dietzia* predominated in run 2. This difference probably reflects the use of two different PCR kits manufactured at several months apart (Qiagen technical service, pers. com.). Other taxa, such as *Streptococcus,* most originated from the DNA extraction kits used, as they were detected in abundance in the negative controls for extraction (NC_{ext}). These results highlight the importance of carrying out systematic negative controls to filter the taxa concerned, 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 without affecting the depth of sequencing of the samples.

After filtering for the above reagent contaminants, the seven most relevant pathogenic bacterial genera, *Bartonella*, *Borrelia*, *Ehrlichia*, *Mycoplasma*, *Orientia*, *Rickettsia* and *Streptobacillus,* accounted for 66% of the sequences identified in wild rodent samples. Six different OTUs were obtained for *Mycoplasma* (*Mycoplasma*_OTU_1 to *Mycoplasma*_OTU_6), with one OTU each for the other genera (Table 2). The other 34% of sequences probably corresponded to commensal bacteria (Bacteroidales, Bacteroides, Enterobacteriaceae, *Helicobacter*, *Lactobacillus*), undetected contaminants and rare taxa of unknown function.

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 on request to the corresponding author). 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 data-based 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 numbers of sequences below the thresholds (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. 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 2, Table S4). 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 is essentially due to mixed clusters during the sequencing [27]. We used two kinds of controls to detect incorrect assignments (Figure 1).

First, negative control index pairs (NC_{index}), corresponding to particular index pairs not used to identify our samples, were included to check for cross-contamination between indices or for errors during completion of the Illumina sample sheet. 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, we used “alien” positive controls (PC_{alien}) in the PCR amplification step: $PC_{Mycoplasma\_m}$, corresponding to the DNA of *Mycoplasma mycoides,* which cannot infect rodents, and $PC_{Borrelia\_b}$, corresponding to 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 misassignment of sequences 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 S1).

Table 2. Number of sequences for 12 pathogenic OTUs observed in wild rodents, in negative controls and in 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).

	OTUs	Total	Wild rodents		Negative controls						Positive controls						Thresholds	
			(*n*=711)		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
	M. mycoides***	280151	338	28	0	0	0	0	2	2	34	20	24	18	279753	82767	/	/
	B. burgdorferi***	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
	M. mycoides***	16854	2	1	0	0	0	0	/	/	0	0	0	0	16852	5766	/	/
	B. burgdorferi***	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)

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 of were hundred-fold diluted 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 the 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 another, 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 2, Table S4).

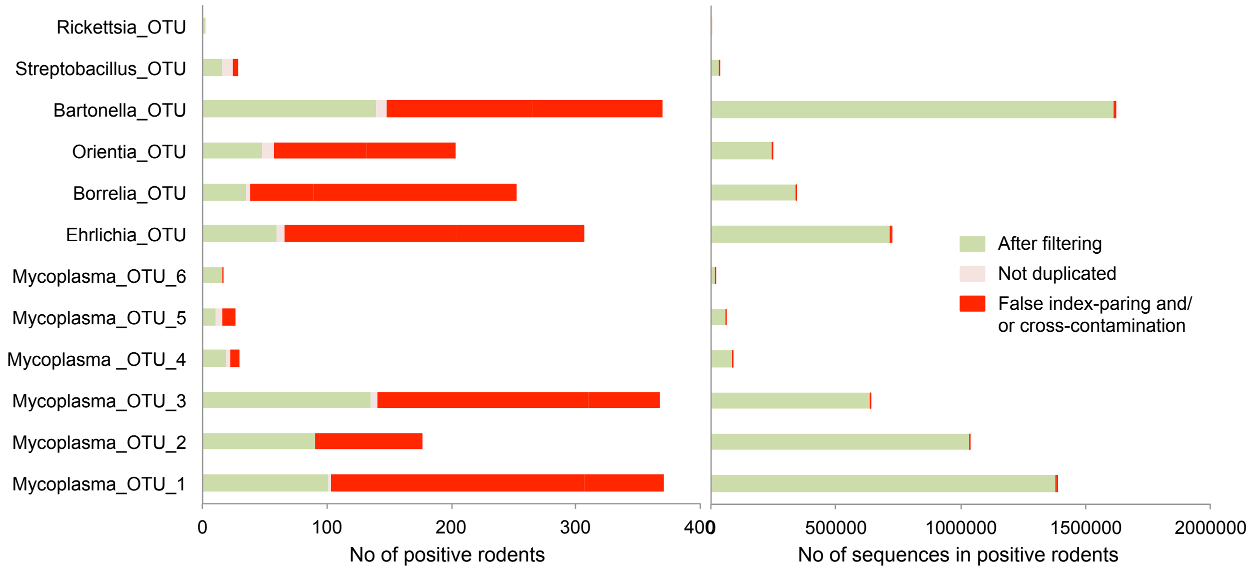

Figure 2. Numbers of positive rodents, and of sequences in positive rodents, removed for each OTU at each step in data filtering. These findings demonstrate that the positive rodents filtered out corresponded to only a very small number of sequences. (A) The histogram shows the number of positive rodents discarded because of likely cross-contamination, false index-pairing and for a negative result in a replicate PCR, and, finally 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 3 and Figure S2). 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.

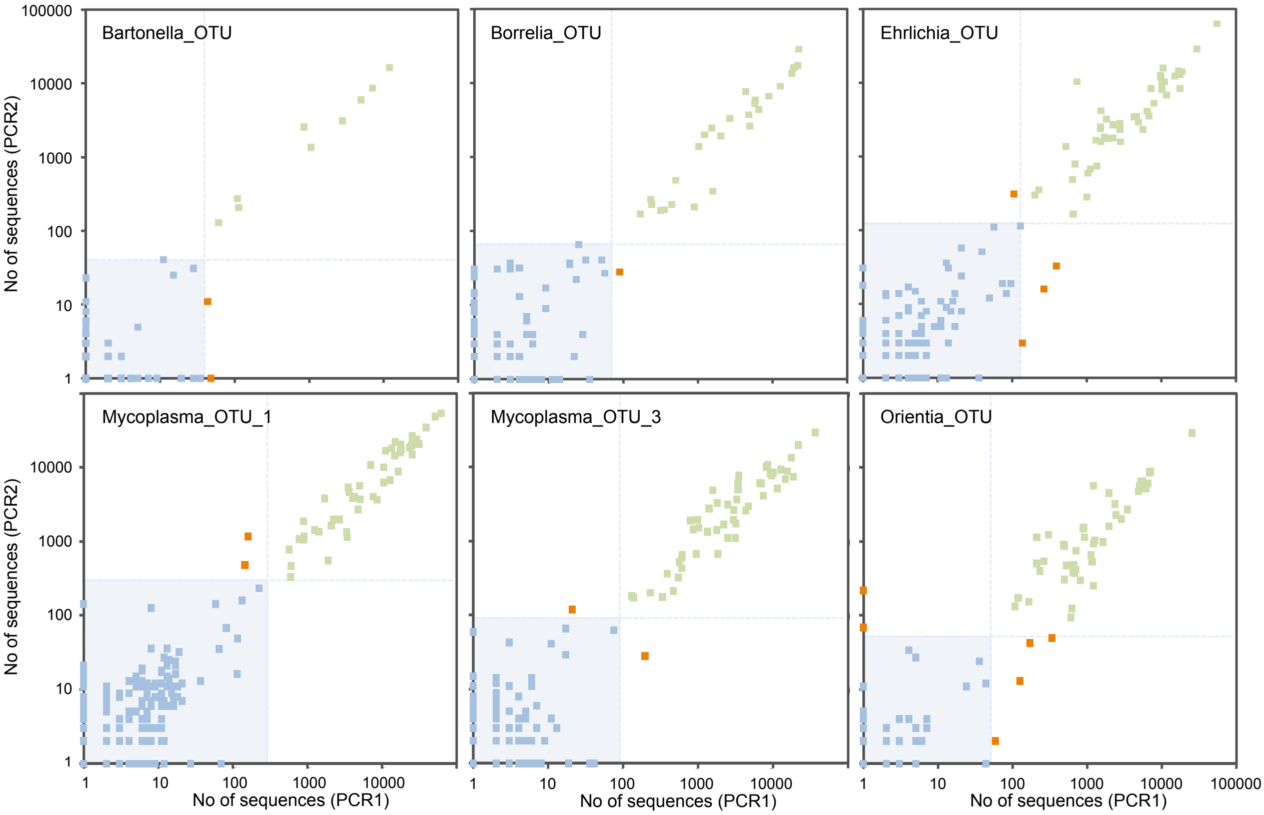

Figure 3. Plots of the number of sequences (log (x+1) scale) from bacterial OTUs in both PCR replicates (PCR1 & PCR2) of the 348 wild rodents analyzed in the first MiSeq run. Note that each rodent was tested with two replicate PCRs. Green points correspond to rodents with two positive results after filtering; 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 the threshold values used for the data filtering: samples below the lines are filtered out. See Figure S2 for plots corresponding to the second MiSeq run.

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 2, Table S4 and Table S5). Note that the number of replicates may be increased, as described in the strategy of Gómez-Díaz *et al* [45].

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 2). 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.

Refining bacterial taxonomic identification. We refined the taxonomic identification of the 12 bacterial OTUs through phylogenetic and blast analyses. We were able to identify the bacteria present down to genus level and, in some cases, we could even identify the most likely species (Table 3 and Figure S3). For instance, the sequences of the six *Mycoplasma* OTUs were consistent with three different species — *M. haemomuris* for OTU_1 and 3, *M. coccoides* for OTU_4, 5 and 6, and *M. species novo* [46] for OTU_2 — with high percentages of sequence identity (≥93%) and strong bootstrap support (≥80%). All three of these species belong to the Hemoplasma group, which is known to infect mice, rats and other mammals [47,48], and is thought to cause anemia in humans [49,50]. The *Borrelia* sequences grouped with three different species of the relapsing fever group (*crocidurae*, *duttonii* and *recurrentis*) with a high percentage of identity (100%) and a reasonably high bootstrap value (71%). In West Africa, *B. crocidurae* causes severe borreliosis, a rodent-borne disease transmitted by ticks and lice [51]. The *Ehrlichia* sequences were 100% identical to and clustered with the recently described Candidatus *Ehrlichia khabarensis* isolated from voles and shrews in the Far East of Russia [52]. The *Rickettsia* sequences were 100% identical to the sequence of *R. typhi*, a species of the typhus group responsible for murine typhus [53], but this clade was only weakly differentiated from many other *Rickettsia* species and had only moderate bootstrap support (61%). The most likely species corresponding to the sequences of the *Streptobacillus* OTU was *S. moniliformis*, with a high percentage of identity (100%) and a high bootstrap value (100%). This bacterium is common in rats and mice and causes a form of rat-bite fever, Haverhill fever [54]. The *Orientia* sequences corresponded to *O. chuto*, with a high percentage of identity (100%) and a high

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

		Number of positive wild rodents				
OTUs of interest (genus level)	Closest species* (% identity in GenBank)	*Mastomys erytholeucus* (*n*=219)	*Mastomys natalensis* (*n*=93)	*Mus musculus* (*n*=203)	*Rattus rattus* (*n*=189)	Biology & epidemiology
Bartonella	undetermined	60	68	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 [83].
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 [51].
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* [52]).
Mycoplasma OTU_1	*haemomuris* (96%)	28	41	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 [46]). 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 [50].
Mycoplasma OTU_2	*sp. novo (100%) GenBank AB752303*	0	0	0	90	
Mycoplasma OTU_3	*haemomuris* (93%)	93	23	1	1	
Mycoplasma OTU_4	*coccoides* (96%)	0	0	0	18	
Mycoplasma OTU_5	*coccoides* (95%)	3	7	0	0	
Mycoplasma OTU_6	*coccoides* (97%)	3	14	0	0	
Orientia	*chuto* (100%) *tsutsugamushi* (98%)	0	3	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 [53].
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% [54].

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

bootstrap value (77%). This species was recently isolated from a patient infected in Dubai [55]. Finally, accurate species determination was not possible for *Bartonella*, as the 16S rRNA gene does not resolve the species of this genus well [56]. Indeed, the sequences from the *Bartonella* OTU detected in our rodents corresponded to at least seven different species (*elizabethae, japonica, pachyuromydis, queenslandis, rattaustraliani, tribocorum, vinsonii*) and a putative new species recently identified in Senegalese rodents [57].

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* [51] and *Bartonella* [57] 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. *S. moniliformis* has recently been detected in rodents from South Africa [58] and there have been a few reports of human streptobacillosis in Kenya [59] and Nigeria [60]. *R. typhi* was recently detected in rats from Congo, in Central Africa [61], and human seropositivity for this bacterium has been reported in coastal regions of West Africa [62]. With the exception of one report in Egypt some time ago [63], *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 [89] together with previous reports of possible cases of human ehrlichioses in this region [64]. 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 [65].

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* (73%) 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.

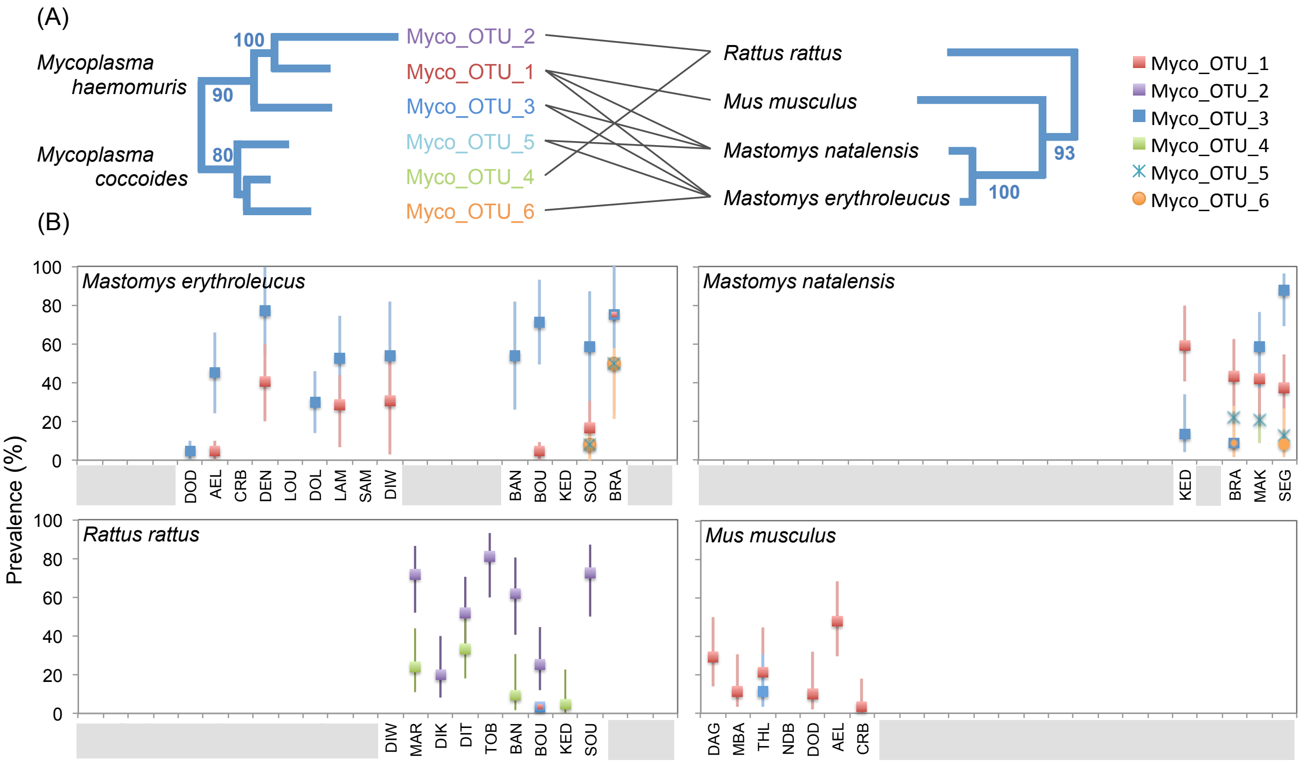

Figure 4. Prevalence of *Mycoplasma* lineages in Senegalese rodents, by site, and phylogenetic associations between *Mycoplasma* lineages and rodent species. (A) Comparison of phylogenetic trees based on the 16S rRNA V4-sequences of *Mycoplasma*, and on the mitochondrial cytochrome *b* gene and the two nuclear gene fragments (IRBP exon 1 and GHR) for rodents (rodent tree redrawn from [91]). 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 [92] by rodent species and site (see [67] 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 4). OTU_2, putatively identified as a recently described lineage isolated from brown rat, *Rattus norvegicus* [46], 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 4). Pathogen prevalence also varied considerably between sites, as shown for the six *Mycoplasma* OTUs (Figure 4). 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 [66,67]

Perspectives

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 that the sensitivity of 16S rRNA amplicon sequencing on the MiSeq platform was equivalent to that of whole RNA sequencing (RNAseq) on the HiSeq platform for detecting bacteria in rodent samples. However, little is known about the comparative sensitivity of HTS approaches relative to qPCR with specific primers, the current gold standard for bacterial detection within biological samples. Additional studies are required to address this question. Moreover, as 16S rRNA amplicon sequencing is based on a short sequence, it does not yield a high enough resolution to distinguish between species in some bacterial genera, such as *Bartonella*. Whole-genome shotgun or RNAseq techniques provide longer sequences, through the production of longer reads or the assembly of contigs, and they might therefore increase the accuracy of species detection [68]. However, these techniques would be harder to adapt for the extensive multiplexing of samples [8]. Other methods could be used to assign sequences to bacterial species for individuals found positive for a bacterial genera following the 16S rRNA screening. For example, positive PCR assays could be carried out with bacterial genus-specific primers, followed by amplicon sequencing, as commonly used in MLSA (multilocus sequence analysis) strategies [69] or high-throughput microfluidic qPCR assays based on bacterial species-specific primers could be used [70]. High-throughput amplicon sequencing approaches could be fine-tuned to amplify several genes for species-level assignment, such as the *gltA* gene used by Gutierrez *et al*. [71] for the *Bartonella* genus, in parallel with the 16S rRNA-V4 region.

This strategy could also easily be adapted for other microbes, such as protists, fungi and even viruses, provided that universal primers are available for their detection (see [72,73] for protists and fungi, and [74] for degenerate virus family-level primers for viruses). Finally, our filtering method could also be translated to any other post-sequencing dataset of indexed or tagged amplicons in the framework of environmental studies (e.g. metabarcoding for diet analysis and biodiversity monitoring [75], the detection of rare somatic mutations [76] or the genotyping of highly polymorphic genes (e.g. MHC or HLA typing, [77,78]).

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 [79]. Many rodent-borne pathogens cause only mild or undifferentiated disease in healthy people, and these illnesses are often misdiagnosed and underreported [54,80-83]. 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. Rodents were killed by cervical dislocation, as recommended by Mills *et al.* [84] and dissected as described in Herbreteau *et al.* [85]. Rodent species were identified by morphological and/or molecular techniques [67]. Cross-contamination during dissection was prevented by washing the tools used successively in bleach, water and alcohol between rodents. We used the spleen for bacterial detection, because this organ is a crucial site of early exposure to bacteria [86]. Spleens were placed in RNAlater (Sigma) and stored at 4°C for 24 hours and then at -20°C until their use for genetic analyses.

Target DNA region and primer design. We used primers with sequences slightly modified from those of the universal primers of Kozich *et al*. [18] to amplify a 251-bp portion of the V4 region of the 16S rRNA gene (16S-V4F: GTGCCAGCMGCCGCGGTAA; 16S-V4R: GGACTACHVGGGTWTCTAATCC). The ability of these primers to hybridize to the DNA of bacterial zoonotic pathogens was assessed by checking that there were low numbers of mismatched bases over an alignment of 41,113 sequences from 79 zoonotic genera inventoried by Taylor et al [1], extracted from the Silva SSU database v119 [43] (Table S6). The FASTA file is available on request to the corresponding author. We used a slightly modified version of the dual-index method of Kozich *et al*. [18] to multiplex our samples. The V4 primers included different 8-bp indices (i5 in the forward and i7 in the reverse position) and Illumina adapters (i.e. P5 in the forward and P7 in the reverse position) in the 5' position. The combinations of 24 i5-indexed primers and 36 i7-indexed primers made it possible to identify 864 different PCR products loaded onto the same MiSeq flowcell. Each index sequence differed from the others by at least two nucleotides, and each nucleotide position in the sets of indices contained approximately 25% of each base, to prevent problems due to Illumina low-diversity libraries (Table 1).

DNA extraction and PCRs. All laboratory manipulations were conducted with filter tips, under a sterile hood, in a DNA-free room. 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) (Table S1).

Library preparation and MiSeq sequencing. Two MiSeq (Illumina) runs were conducted, including PCR products from wild rodents and the positive and negative controls detailed in Figure 1 and Table S1. The MiSeq platform was chosen because it generates lower error rates than other HTS platforms [87]. The number of PCR products multiplexed was 823 for the first MiSeq run and 746 for the second MiSeq run (Table S1). 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 in 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 (expected size: 375 bp including primers, indexes and adaptors), by excision in a low-melting agarose gel (1.25%) to discard non-specific amplicons and primer dimers. The 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 on request to the corresponding author.

Bioinformatic and taxonomic classification. MiSeq datasets were processed with mothur v1.34 [42] and with the MiSeq standard operating procedure

(SOP) [18]. We used the Silva SSU Reference database v119 [43] and the Silva taxonomy file for taxonomic assignment. 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. 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 [88]. The FASTA files used are available on request to the corresponding author. Phylogenetic trees were generated from the K2P distance with SeaView and species were identified on the basis of the "closest phylogenetic species". We also used our sequences for blast analyses of GenBank, to identify the reference sequences to which they displayed the highest percentage identity.

Acknowledgments

This study was funded by the French National Institute for Agricultural Research (INRA) Meta-omics and microbial ecosystems metaprogram (Patho-ID project: Rodent and tick pathobiomes) and the ANR ENEMI (ANR-11-JSV7-0006). We would like to thank Virginie Dupuy for extracting DNA from bacterial cultures and Julie Sappa from Alex Edelman & Associates for improving the English writing. Analyses were performed on the CBGP HPC computational platform. The funders had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

Authors' contributions

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

Supplementary materials

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).

Table S2. The 50 most abundant OTUs in wild rodents and controls.

Table S3. Bacterial contaminants observed in negative and 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.

Table S4. Proportion of sequences and proportion of positive results removed at each step in data filtering. Note that several positive results may be recorded for the same rodent in cases of co-infection.

Table S5. 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.

Table S6. Number of mismatches between PCR forward and reverse primers and 41,113 bacterial 16S rRNA V4 sequences of 79 zoonotic genera. Data [1] was extracted from the Silva SSU database v119. Numbers of mismatches > 3 correspond to sequences of bad quality from different taxon. The number of mismatches in the 3' side of primers was always <2.

Figure S1. 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 plate 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 plate 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).

Figure S2. 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 duplicate PCRs. Green points correspond to rodents with two positive results after the filtering process; orange 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 S2 for plots corresponding to the second MiSeq run. See Figure 3 for plots corresponding to the first MiSeq run.

Figure S3. 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%. The Fasta files used are available on request to the corresponding author.

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, et al.** 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, et al.** 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, et al. 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, et al. 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, et al.** 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, et al.** 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, et al.** 2011. Metagenomic profile of the bacterial communities associated with Ixodes ricinus ticks. PLoS One 6: e25604.

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

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

16. **Vayssier-Taussat M, Albina E, Citti C, Cosson JF, Jacques MA, et al.** 2014. Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. Front Cell Infect Microbiol 4: 29.

17. **Vayssier-Taussat M, Kazimirova M, Hubalek Z, Hornok S, Farkas R, et al.** 2015. Emerging horizons for tick-borne pathogens: from the 'one pathogen-one disease' vision to the pathobiome paradigm. Future Microbiol 10: 2033-2043.

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

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

20. **Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, et al.** 2012. Experimental and analytical tools for studying the human microbiome. Nat Rev Genet 13: 47-58.

21. **Sinha R, Chen J, Amir A, Vogtmann E, Shi J, et al.** 2015. Collecting Fecal Samples for Microbiome Analyses in Epidemiology Studies. Cancer Epidemiol Biomarkers Prev.

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

23. **Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, et al.** 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12: 87.

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

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

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

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

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

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

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

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

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

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

34. **Whiteford N, Skelly T, Curtis C, Ritchie ME, Lohr A, et al.** 2009. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics 25: 2194-2199.

35. **Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, et al.** 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6: 1621-1624.

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

37. **Kilianski A, Haas JL, Corriveau EJ, Liem AT, Willis KL, et al.** 2015. Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer. Gigascience 4: 12.

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

39. **Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, et al.** 2014. The SILVA and "All-species Living Tree Project. LTP)" taxonomic frameworks. Nucleic Acids Res 42: D643-648.

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

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

42. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al.** 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537-7541.

43. **Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al.** 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41: D590-596.

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

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

46. **Sashida H, Sasaoka F, Suzuki J, Fujihara M, Nagai K, et** al. 2013. Two clusters among Mycoplasma haemomuris strains, defined by the 16S-23S rRNA intergenic transcribed spacer sequences. J Vet Med Sci 75: 643-648.

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

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

49. **Steer JA, Tasker S, Barker EN, Jensen J, Mitchell J, et al.** 2011. A novel hemotropic Mycoplasma. hemoplasma. in a patient with hemolytic anemia and pyrexia. Clin Infect Dis 53: e147-151.

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

51. **Trape JF, Diatta G, Arnathau C, Bitam I, Sarih M, et al.** 2013. The epidemiology and geographic distribution of relapsing fever borreliosis in West and North Africa, with a review of the Ornithodoros erraticus complex. Acari: Ixodida). PLoS One 8: e78473.

52. **Rar VA, Pukhovskaya NM, Ryabchikova EI, Vysochina NP, Bakhmetyeva SV, et al.** 2015. Molecular-genetic and ultrastructural characteristics of 'Candidatus Ehrlichia khabarensis', a new member of the Ehrlichia genus. Ticks Tick Borne Dis 6: 658-667.

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

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

55. **Izzard L, Fuller A, Blacksell SD, Paris DH, Richards AL, et al.** 2010. Isolation of a novel Orientia species. O. chuto sp. nov.. from a patient infected in Dubai. J Clin Microbiol 48: 4404-4409.

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

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

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

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

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

61. **Laudisoit A, Falay D, Amundala N, Akaibe D, de Bellocq JG, et al. 2014**. High prevalence of Rickettsia typhi and Bartonella species in rats and fleas, Kisangani, Democratic Republic of the Congo. Am J Trop Med Hyg 90: 463-468.

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

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

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

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

66. **Konecny A, Estoup A, Duplantier JM, Bryja J, Ba K, et al.** 2013. Invasion genetics of the introduced black rat. Rattus rattus. in Senegal, West Africa. Mol Ecol 22: 286-300.

67. **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

68. **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.

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

70. **Michelet L, Delannoy S, Devillers E, Umhang G, Aspan A, et al.** 2014. High-throughput screening of tick-borne pathogens in Europe. Front Cell Infect Microbiol 4: 103.

71. **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.

72. **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.

73. **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.

74. **Anthony SJ, Epstein JH, Murray KA, Navarrete-Macias I, Zambrana-Torrelio CM, et al.** 2013. A strategy to estimate unknown viral diversity in mammals. MBio 4: e00598-00513.

75. **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.

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

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

78. **Lange V, Bohme I, Hofmann J, Lang K, Sauter J, et al.** 2014. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics 15: 63.

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

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

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

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

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

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

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

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

87. **D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, et al.** 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics 17: 55.

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

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

90. **Qiu Y, Nakao R, Ohnuma A, Kawamori F, Sugimoto C.** 2014. Microbial population analysis of the salivary £*%^M:P@ Phylogeny and biogeography of African Murinae based on mitochondrial and nuclear gene sequences, with a new tribal classification of the subfamily. BMC Evol Biol 8:199

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