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 Galan¹, Maria Razzauti¹, Emilie Bard², Maria Bernard³,⁴, Carine Brouat⁵, Nathalie Charbonnel¹, Alexandre Dehne-Garcia¹, Anne Loiseau¹, Caroline Tatard¹, Lucie Tamisier¹, Muriel Vayssier-Taussat⁶, Hélène Vignes⁷, Jean-François Cosson¹,⁶

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.

<table>
<thead>
<tr>
<th>Experimental steps</th>
<th>Sources of errors</th>
<th>Consequences</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample collection</strong></td>
<td>Cross-contamination between individuals [21]</td>
<td>False-positive samples</td>
<td>Rigorous processing (decontamination of the instruments, cleaning of the autopy table, use of sterile bacterial-free consumables, gloves, masks) Negative controls during sampling (e.g., organs of healthy mice during dissection) Use of appropriate storage conditions/buffers. Use of unambiguously identified samples. Double checking of tube labeling during sample collection.</td>
</tr>
<tr>
<td></td>
<td>Collection and storage conditions [21]</td>
<td>False-positive &amp; negative samples</td>
<td></td>
</tr>
<tr>
<td><strong>DNA extraction</strong></td>
<td>Cross-contamination between samples [22]</td>
<td>False-positive samples</td>
<td>Rigorous processing (separation of pre- and post-PCR steps, use of a sterile hood, filter tips and sterile bacterial-free consumables) Negative controls for extraction (extraction without sample)</td>
</tr>
<tr>
<td></td>
<td>Reagent contamination with bacterial DNA [21,23]</td>
<td>False-positive samples</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Small amounts of DNA [21, 24]</td>
<td>False-negative samples</td>
<td>Use of an appropriate DNA extraction protocol. Discarding of samples with a low DNA concentration.</td>
</tr>
<tr>
<td><strong>Target DNA region and primer design</strong></td>
<td>Target DNA region efficiency [19,25]</td>
<td>False-negative due to poor taxonomic identification</td>
<td>Selection of an appropriate target region and design of effective primers for the desired taxonomic resolution Checking of the universality of the primers with reference sequences.</td>
</tr>
<tr>
<td></td>
<td>Primer design [21,26]</td>
<td>False-negative samples due to biases in PCR amplification for some taxa</td>
<td></td>
</tr>
<tr>
<td><strong>Tag/index design and preparation</strong></td>
<td>False-assignments of sequences due to cross-contamination between tags/indices [27,29]</td>
<td>False-positive samples</td>
<td>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</td>
</tr>
<tr>
<td></td>
<td>False-assignments of sequences due to inappropriate tag/index design [29]</td>
<td>False-positive samples</td>
<td>Focusing 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</td>
</tr>
<tr>
<td><strong>PCR amplification</strong></td>
<td>Cross-contamination between PCRs [25]</td>
<td>False-positive samples</td>
<td>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 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</td>
</tr>
<tr>
<td></td>
<td>Reagent contamination with bacterial DNA [21,23]</td>
<td>False-positive samples</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chimeric recombinations by jumping PCR [27,30,31,32,33]</td>
<td>False-negative samples due to artifactual chimeric sequences</td>
<td>Increasing the elongation time. Use of a bioinformatic strategy to remove the chimeric sequences (e.g., Uchime program)</td>
</tr>
<tr>
<td></td>
<td>Poor or biased amplification [45]</td>
<td>False-negative samples</td>
<td>Increasing the amount of template DNA. Optimizing the PCR conditions (reagents and program) Use of technical replicates to validate sample positivity</td>
</tr>
<tr>
<td><strong>Library preparation</strong></td>
<td>Cross-contamination between PCRs/libraries [22]</td>
<td>False-positive samples</td>
<td>Rigorous processing (use of 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) Avoiding PCR library enrichment of pooled samples.</td>
</tr>
<tr>
<td></td>
<td>Chimeric recombinations by jumping PCR [27]</td>
<td>False-negative samples due to inter-individual recombinations</td>
<td></td>
</tr>
<tr>
<td><strong>MiSeq sequencing (Illumina)</strong></td>
<td>Sample sheet errors [21]</td>
<td>False-negative samples</td>
<td>Rigorous processing (use of 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) Avoiding PCR library enrichment of pooled samples.</td>
</tr>
<tr>
<td></td>
<td>Run-to-run carryover (Illumina Technical Support Note No. 770-2013-046)</td>
<td>False-negative samples</td>
<td>Rigorous processing (use of 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) Avoiding PCR library enrichment of pooled samples.</td>
</tr>
<tr>
<td></td>
<td>Poor quality of reads due to flowcell overloading [34]</td>
<td>False-negative due to low quality of sequences</td>
<td>qPCR quantification of the library before sequencing. Decreasing cluster density. Creation of artificial sequence diversity at the flowcell surface (e.g., by adding 5 to 10% PhiX DNA control library)</td>
</tr>
<tr>
<td></td>
<td>Poor quality of reads due to low-diversity libraries (Illumina Technical Support Note No. 770-2013-015)</td>
<td>False-negative due to low quality of sequences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Small number of reads per sample [35,36]</td>
<td>False-negative due to low depth of sequencing</td>
<td>Decreasing the level of multiplexing Discard the sample with a low number of reads.</td>
</tr>
<tr>
<td></td>
<td>Too short overlapping read pairs [18]</td>
<td>False-negative due to low quality of sequences</td>
<td>Increasing paired-end sequence length or decreasing the length of the target sequence</td>
</tr>
<tr>
<td></td>
<td>Mixed clusters on the flowcell [27]</td>
<td>False-positive due to false index-pairing</td>
<td>Use of a single barcode sequence for both the i5 and i7 indices for each sample (when possible, e.g., small number of samples)</td>
</tr>
<tr>
<td><strong>Bioinformatics and taxonomic classification</strong></td>
<td>Poor quality of reads</td>
<td>False-negative samples due to poor taxonomic resolution</td>
<td>Removal of low-quality reads Use of standardized protocols and reproducible workflows</td>
</tr>
<tr>
<td></td>
<td>Errors during processing (sequence trimming, alignment) [18,37,38]</td>
<td>False-positive and negative samples</td>
<td>Selection of an appropriate database for the selected target region and testing of the database for bacteria of particular interest</td>
</tr>
<tr>
<td></td>
<td>Incomplete reference sequence databases [39]</td>
<td>False-negative samples</td>
<td>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)</td>
</tr>
<tr>
<td></td>
<td>Error of taxonomic classification [40]</td>
<td>False-negative samples</td>
<td></td>
</tr>
</tbody>
</table>

The copyright holder for this preprint (which was first posted online Feb. 17, 2016; doi: http://dx.doi.org/10.1101/039826) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.
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.

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<sub>ext</sub> and NC<sub>PCR</sub>; T<sub>CC</sub>: sequence number threshold for correcting for cross-contamination. T<sub>CC</sub> values are OTU- and run-dependent and
were estimated by analyzing the sequences in the controls, NC<sub>mus</sub>, NC<sub>ext</sub>, NC<sub>PCR</sub> and PC<sub>PCR</sub>; T<sub>FA</sub>: sequence number threshold for correcting for false index-pairing. T<sub>FA</sub> values are OTU- and run-dependent
and were estimated by analyzing the sequences in the NC<sub>index</sub> and PC<sub>index</sub>. A result was considered
positive if the number of sequences was > T<sub>CC</sub> and > T<sub>FA</sub>. 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<sub>ext</sub>), corresponding to extraction without the addition of a
tissue sample, and negative controls for PCR (NC<sub>PCR</sub>), 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 Tcc: Filtering for cross-contamination.** One source of false positives is cross-contamination between samples processed in parallel (Table 1). Negative controls for dissection (NC\textsubscript{mus}), consisting of the spleens of healthy laboratory mice manipulated during sessions of wild rodent dissection, and negative controls for extraction (NC\textsubscript{ex}) and PCR (NC\textsubscript{PCR}) were used, together with positive controls for PCR (PC\textsubscript{Bartonella\_t}, PC\textsubscript{Borrelia\_b} and PC\textsubscript{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\textsubscript{CC}) for each run. The use of these T\textsubscript{CC} led to 0% to 69% of the positive results being discarded, corresponding to only 0% to 0.14% of the sequences, depending on 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\textsubscript{CC} makes it possible to discard the positive result for one bacterium whilst retaining positive results for other bacteria.

**Threshold T\textsubscript{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\textsubscript{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\textsubscript{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\textsubscript{alien}) in the PCR amplification step: PC\textsubscript{Mycoplasma\_m}, corresponding to the DNA of *Mycoplasma mycoides*, which cannot infect rodents, and PC\textsubscript{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\textsubscript{Mycoplasma\_m}, we obtained an estimate of the global false index-pairing rate of
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_rims: negative controls for dissection; PC_Bartonella_t: positive controls for PCR; PC_Borrelia_b and PC_Mycoplasma_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).

<table>
<thead>
<tr>
<th>OTUs</th>
<th>Total sequences</th>
<th>Wild rodents (n=111)</th>
<th>Negative controls</th>
<th>Positive controls</th>
<th>Thresholds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total no. of sequences in the PCR</td>
<td>NC_PCR</td>
<td>NC_ext</td>
<td>NC_rims</td>
<td>PC_Bartonella_t</td>
</tr>
<tr>
<td>Whole dataset</td>
<td>756537</td>
<td>714944</td>
<td>64722</td>
<td>45900</td>
<td>8002</td>
</tr>
<tr>
<td>Mycoplasma_OTU_1</td>
<td>1410218</td>
<td>1410199</td>
<td>61607</td>
<td>2 1</td>
<td>1 3</td>
</tr>
<tr>
<td>Mycoplasma_OTU_3</td>
<td>507376</td>
<td>507369</td>
<td>36335</td>
<td>2 1</td>
<td>0 0</td>
</tr>
<tr>
<td>Orientia_OTU</td>
<td>694851</td>
<td>694823</td>
<td>65137</td>
<td>4 2</td>
<td>3 2</td>
</tr>
<tr>
<td>Borrelia_OTU</td>
<td>346783</td>
<td>346845</td>
<td>28528</td>
<td>4 4</td>
<td>7 4</td>
</tr>
<tr>
<td>Bartonella_OTU</td>
<td>279865</td>
<td>279857</td>
<td>26503</td>
<td>1 1</td>
<td>4 1</td>
</tr>
<tr>
<td>M. mycoides***</td>
<td>280151</td>
<td>338 28</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>B. burgdorferi***</td>
<td>238772</td>
<td>420 43</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td></td>
<td>6687080</td>
<td>6520107</td>
<td>42226</td>
<td>61231</td>
<td>9145</td>
</tr>
</tbody>
</table>

* 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 on 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_OTU 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\textsubscript{Mycoplasma\textunderscore m} and PC\textsubscript{Borrelia\textunderscore b} separately. \( R_{fa} \) reached 0.010\% and 0.018\% for PC\textsubscript{Mycoplasma\textunderscore m} and PC\textsubscript{Borrelia\textunderscore 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.

<table>
<thead>
<tr>
<th>OTUs of interest (genus level)</th>
<th>Closest species* (% identity in GenBank)</th>
<th>Number of positive wild rodents</th>
<th>Biology &amp; epidemiology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bartonella</strong></td>
<td>undetermined</td>
<td>60 68 1 6</td>
<td>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.</td>
</tr>
<tr>
<td><strong>Borrelia</strong></td>
<td>crocidurae (100%) duttonii (100%) recurrentis (100%)</td>
<td>21 0 8 6</td>
<td>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.</td>
</tr>
<tr>
<td><strong>Ehrlichia</strong></td>
<td>khabarensis (100%)</td>
<td>40 0 12 8</td>
<td>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).</td>
</tr>
<tr>
<td><strong>Mycoplasma</strong></td>
<td>OTU_1 haemomuris (96%)</td>
<td>28 41 30 1</td>
<td>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].</td>
</tr>
<tr>
<td>Mycoplasma OTU_2</td>
<td>sp. novo (100%) GenBank AB752303</td>
<td>0 0 0 90</td>
<td></td>
</tr>
<tr>
<td>Mycoplasma OTU_3</td>
<td>haemomuris (93%)</td>
<td>93 23 1 1</td>
<td></td>
</tr>
<tr>
<td>Mycoplasma OTU_4</td>
<td>coccoides (96%)</td>
<td>0 0 0 18</td>
<td></td>
</tr>
<tr>
<td>Mycoplasma OTU_5</td>
<td>coccoides (95%)</td>
<td>3 7 0 0</td>
<td></td>
</tr>
<tr>
<td>Mycoplasma OTU_6</td>
<td>coccoides (97%)</td>
<td>3 14 0 0</td>
<td></td>
</tr>
<tr>
<td><strong>Orientia</strong></td>
<td>chuto (100%) tsutsugamushi (98%)</td>
<td>0 3 46 0</td>
<td>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].</td>
</tr>
<tr>
<td><strong>Rickettsia</strong></td>
<td>typhi (100%)</td>
<td>1 0 0 1</td>
<td>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].</td>
</tr>
<tr>
<td><strong>Streptobacillus</strong></td>
<td>moniliformis (100%)</td>
<td>10 1 0 5</td>
<td>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].</td>
</tr>
</tbody>
</table>

*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.

![Diagram of Mycoplasma lineages](image)

**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. erythroleucus 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: GTGCCAGCMGCGCGGTAA; 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 (NCext), 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 (NCPCR), 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/mm²) 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
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, NCmus; Negative Controls for extraction, NCext; Negative Controls for PCR, NCPCR; Negative Controls for indexing, NCindex; Positive Controls for PCR, PCPCR; Positive Controls for Indexing, PCindex. 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\textsubscript{Borrelia\_b} (in blue) and PC\textsubscript{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\textsubscript{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\textsubscript{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
Borrelia 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


