environmental microbiology

Environmental Microbiology (2017) **19**(12), 5010–5024 doi:10.1111/1462-2920.13949

Metagenomic assessment of the interplay between the environment and the genetic diversification of *Acinetobacter*

Marc Garcia-Garcera,[1,2]* **Marie Touchon**,[1,2] **Sylvain Brisse**[1,2] **and Eduardo P.C. Rocha**[1,2]

[1]*Microbial Evolutionary Genomics, Institut Pasteur, 25-28 rue Dr Roux, Paris 75015, France.*

[2]*CNRS, UMR3525, Unité de Génétique des Genomes, 25-28 rue Dr. Roux, Paris 75015, France.*

Summary

Most bacteria have poorly characterized environmental reservoirs and unknown closely related species. This hampers the study of bacterial evolutionary ecology because both the environment and the genetic background of ancestral lineages are unknown. We combined metagenomics, comparative genomics and phylogenomics to overcome this limitation, to identify novel taxa and to propose environments where they can be isolated. We applied this method to characterize the ecological distribution of known and novel lineages of *Acinetobacter* spp. We observed two major environmental transitions at deep phylogenetic levels, splitting the genus into three ecologically differentiated clades. One of these has rapidly shifted towards host-association by acquiring genes involved in bacteria-eukaryote interactions. We show that environmental perturbations affect species distribution in predictable ways: bovines have very diverse communities of *Acinetobacter*, unless they were administered antibiotics, in which case they show highly uniform communities of *Acinetobacter* spp. that resemble those of humans. Our results uncover the diversity of bacterial lineages, overpassing the limitations of classical cultivation methods and highlight the role of the environment in shaping their evolution.

Received 17 January, 2017; revised 8 September, 2017; accepted 26 September, 2017. *For correspondence. E-mail marc.garcia.garcera@gmail.com; Tel. +33 140 613 353; Fax +33 145 688 727.

Introduction

The evolution of many bacterial lineages, like *Acinetobacter*, is driven by rapid change in gene repertoires in response to environmental challenges (Ley *et al*., 2006; Treangen and Rocha, 2011). This is especially apparent in the emergence of nosocomial pathogens, since the acquisition of new genetic tools allows previously inoffensive isolates to become virulent in a context of niche depletion caused by antibiotherapy (Peleg *et al*., 2008; Vallenet *et al*., 2008; Touchon *et al*., 2009; Bialek-Davenet *et al*., 2014). Hence, the historical changes in gene repertoires are associated with environmental adaptations (Gianoulis *et al*., 2009; Martinez, 2009). In theory, the colonization of a new environment is accompanied by adaptive genetic changes that facilitate its colonization (Smillie *et al*., 2011). Despite the theoretical support to the relationship between environment and genetic diversification (Ehrlén and Morris, 2015), its study in microbiology has lagged behind due to two main reasons. First, the environmental reservoirs of bacterial species are not well characterized. This is especially true in species including strains with an antagonistic behaviour towards humans, where most of the knowledge is biased towards clinical isolates even when most lineages are avirulent (Doughari *et al*., 2011). Second, the analysis of the evolution of these lineages is often impaired by lack of known closely related species (Heath *et al*., 2008), which hinders our capacity to understand how the evolutionary history of the lineage is associated with its environmental distribution.

To overcome these limitations, we have studied the environmental distribution of a bacterial lineage using metagenomic data. Metagenomics facilitates the analysis of bacterial biodiversity and the identification of novel taxa because it bypasses the need for microbial cultivation or isolation (Handelsman, 2004; Rodriguez-R and Konstantinidis, 2014). The recent development of these techniques has greatly expanded the nucleotide sequence databanks, which can be queried to identify bacterial taxa. We have used the *Acinetobacter* genus as a model to our study. Its members are thought to colonize a wide variety of environments (Doughari *et al*., 2011), and their phylogenetic relationships have been recently resolved through genome-wide comparative analyses (Sahl *et al*., 2013;

Touchon *et al*., 2014). *Acinetobacter* are intrinsically resistant to many toxics, including antibiotics, which gives them an adaptive advantage in the hospital. One species, *A. baumannii*, is one of the most important nosocomial pathogens, and several other species may be emerging as novel nosocomials (Tjernberg and Ursing, 1989; Bergogne-Bérézin and Joly-Guillou, 1991; Seifert *et al*., 1997).

In this work, we use the Evolutionary Placement Algorithm (EPA) to map metagenomic data in a phylogeny. EPA uses a pre-existing phylogeny and a multiple alignment to place a novel sequence from a single copy gene biomarker in the tree (Berger *et al*., 2011). If a fragment is from a species represented in the tree, the EPA places it next to the corresponding tip. If a fragment is from a species lacking close representatives in the tree, the EPA places it at the internal branch where that species would branch if it were present in the tree. Hence, EPA provides valuable information on the existence of previously unknown taxa, their environmental distribution and their phylogenetic relationship with the known species.

We have used thousands of lineage-specific protein profiles and EPA to obtain a fine-scale classification of metagenomic sequences in a bacterial genus. Our approach identifies known and unknown taxa within the focal lineage, *Acinetobacter* in this case, and associates them with an environment and a position in the phylogenetic tree of the genus. It also serves as a proof of concept for a method that can be applied to other bacterial clades. We test the ability of our method to identify novel lineages in an environment, to partition a clade in relation to habitat preferences and to study the change in community composition following an environmental perturbation. The latter also sheds light on how the evolution of the ability to interact with eukaryotes can facilitate the emergence of virulence in humans, when microbial niches are affected by antibiotic treatments.

Results

Overview of the method and assessment of its quality

We developed a pipeline to identify clade-specific sequences in metagenomic data and place them on a reference phylogenetic tree in four major steps (see Experimental procedures, Supporting Information Fig. S1). First, we built protein profiles for every protein family of the core-genomes of the focal clade, a close outgroup and a distant outgroup. Second, we retrieved, curated and annotated a large set of metagenomes from different environments. We integrated the two datasets by searching in the metagenomic data for proteins (or peptides) matching the profiles of the core-genomes. Third, we used linear discriminant analysis (LDA) and self-organizing maps (SOM) to remove the distantly related sequences from the hits. Finally, the remaining peptides were placed in the focal clade tree by maximum likelihood using EPA. The first three steps ascertain that few non-pertinent sequences are subject to EPA. This makes the procedure much faster than if all peptides were subject to the EPA, because this latter step is very time-consuming. We used this procedure to study *Acinetobacter*, using *Moraxella* and *Psychrobacter* as close outgroups and *Pseudomonas* as a distant outgroup. The choice of these outgroups was based on the phylogenetic distance: the first group is the sister-clade of *Acinetobacter* in our dataset, whereas the *Pseudomonas* is the subsequent one (among clades with several completely sequenced genomes). Although the core-genome of *Acinetobacter* was composed of 923 genes, only 647 of them where shared with the outgroups. To avoid possible misassignments due to the lack of orthologous genes in the outgroups, only the latter were used.

We assessed the quality of the classification (LDA + SOM) by randomly sampling peptides from the genomes of *Acinetobacter* and the close outgroup (Supporting Information Figs S2 and S3). LDA assigned incorrectly only 7.8% of the peptides from core-genomes. However, 16% of peptides from other proteins matched the core-genome profiles, because they were homologous, leading to 39% of erroneous assignations. Expectedly, these matches had lower scores than those from members of core-gene families. To remove them, we restricted our analysis to peptides that had a sufficient score (parameter $S_{A,A,i}$) and that matched the profiles of *Acinetobacter* by at least 9% better (parameter R_i) than those of the outgroups (see Experimental procedures). This effectively removed the matches of paralogs. Nevertheless, 17 core-genes of *Acinetobacter* (out of 647) were consistently misclassified because they had highly similar homologs outside *Acinetobacter* (presumably due to horizontal gene transfer). We discarded the corresponding protein profiles from further analyses. A final number of 630 core genes was used for further analyses. The receiver operational characteristic (ROC) curves of the entire classification procedure showed a remarkably good trade-off between sensitivity and specificity even for small peptides (Supporting Information Fig. S4). Upon validation, the set of parameters used in our subsequent analysis returned less than 0.5% of false positives (Supporting Information Table S5, see Experimental procedures).

We tested the consistency of the method by analysing six novel *Acinetobacter* genomes. These genomes provide an independent validation set because they were not used in the previous analysis (they only became available after the start of the project), and were distant from all the others (average nucleotide identity: ANI < 0.95, see Experimental procedures). To validate the procedure, we selected random parts of proteins (small peptides) from these genomes with sizes representative of the metagenomics datasets and placed them on the reference tree of

Acinetobacter. We also built a new phylogenetic tree of the core genome of the genus including the novel genomes (Supporting Information Fig. S5). For each of the six taxa, we computed the differences between the observed placement of the peptides in the reference phylogenetic tree and the expected one (given by the new core genome tree). The differences were small: 97% of the peptides were placed less than 4% away from the expected position (the percentage is the distance between the positions, divided by the maximal tip to root distance in the tree, Supporting Information Fig. S6). We concluded that our procedure accurately maps metagenomic peptides on the phylogenetic tree of *Acinetobacter*. The EPA has been previously used to obtain a broad taxonomic classification of metagenomic data using a small set of universal marker genes (as done in PhyloSift) (Darling *et al.*, 2014). We evaluated the benefits of using the profiles from the complete core-genome instead of the universal markers. Our approach was better at identifying *Acinetobacter* fragments and at placing them in the tree (Supporting Information Figs S7-S12, see '*Validation of the Core-Genome EPA compared to Phylosift*' in Supporting Information). Hence, the use of a large number of core genes increases the discriminative capacity of EPA when the study is focusing on a given microbial clade.

Abundance of Acinetobacter *in microbial communities*

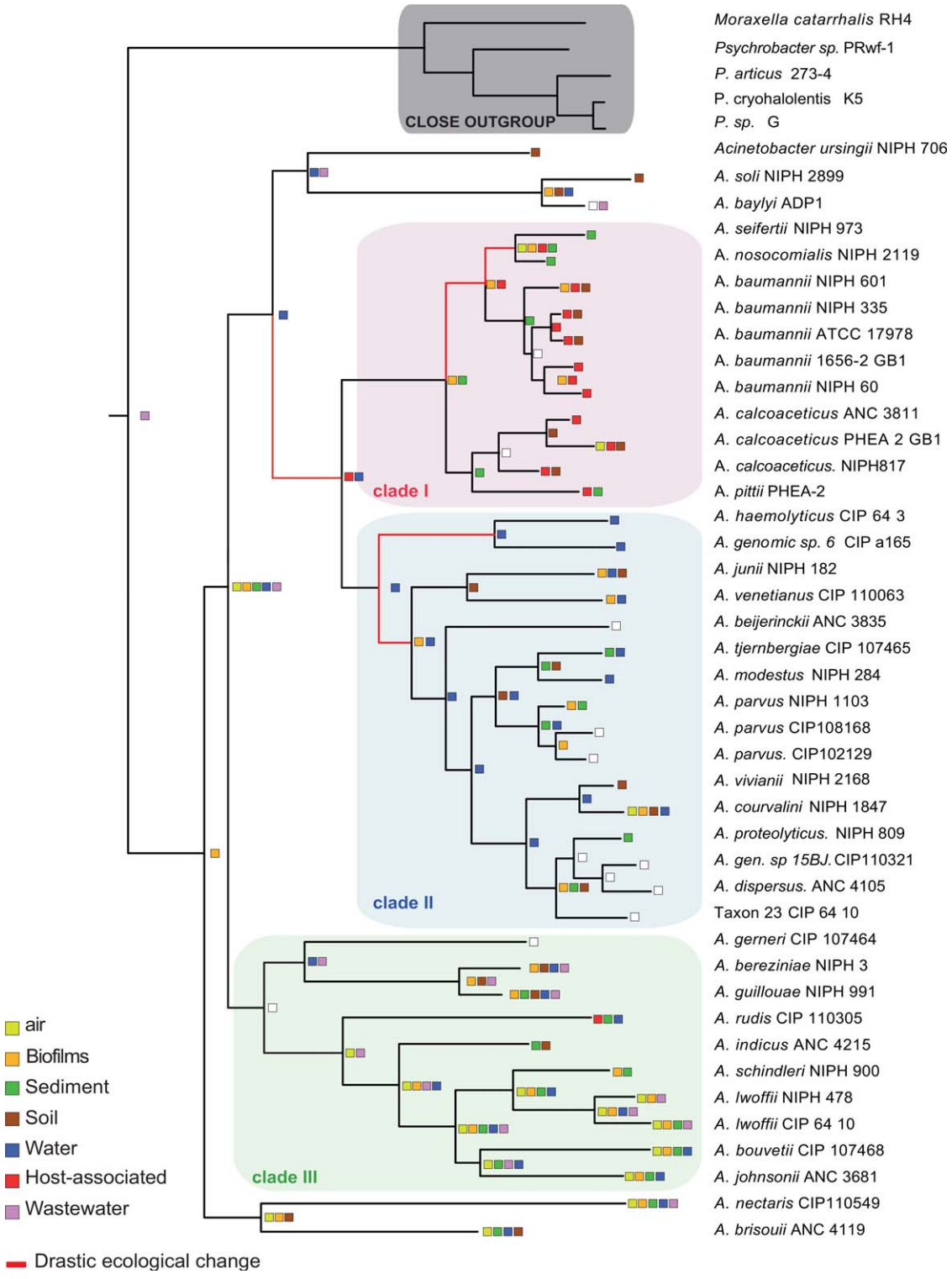
Our method allows to identify the microbial communities whose metagenomes contain sequences from *Acinetobacter*. To evaluate its accuracy, we checked if the six novel *Acinetobacter* genomes used to validate the EPA procedure were placed in branches of the tree over-representing the environments where they were isolated. This was indeed the case for the six taxa (see '*Validation of the EPA using novel Acinetobacter genomes*' in Supporting Information), showing that EPA helps identifying the environments where novel taxa can be found.

We then retrieved 2568 metagenomic datasets from 126 independent locations and classed them in types of environments (see Experimental procedures). We identified *Acinetobacter* in 817 out of 2568 datasets. We placed 274 890 of the peptides of these sets (0.06%) using EPA in the reference tree of *Acinetobacter*. Four environments had particularly high frequencies of *Acinetobacter* spp. (Fig. 1): (i) soil, (ii) host-associated environments (host), (iii) natural aquatic environments (water) and (iv) aquatic environments, rich in organic material, either associated with waste treatment plants (wastewater) or marine sediments. Peptides from *Acinetobacter* were identified at much lower frequencies in the other environments (Supporting Information Fig. S13), and were lacking in extreme environments, notably in hyperthermophilic, hypersaline or mine drainage samples. We found peptides matching the same branches in different environments (Fig. 1). Importantly, we placed 38% of the peptides in the internal branches of the phylogenetic tree, suggesting that many novel taxa of *Acinetobacter* remain to be uncovered.

Distinct environmental distribution in three major clades

The peptides from certain environments were placed much more frequently in certain branches of the phylogenetic tree than in others. In particular, the peptides from the four environments with higher frequency of *Acinetobacter* (aquatic, host-associated, soil and wastewater) were not randomly placed in the tree (all $P < 0.0001$, Bartel's Rank test). To obtain an accurate picture of the distribution of the *Acinetobacter* in terms of phylogeny and environment, we computed the environmental sources over-represented among the peptides placed in each branch of the tree (Fig. 1). As mentioned above (see *Introduction*), the data plotted in the internal branches of the tree does not represent ancestral states. Instead, it represents taxa absent from the tree that branch at the position specified by the EPA. The environments where these taxa are over-represented are plotted at the corresponding internal branches. The EPA revealed three large clades with distinctive over-represented environments (named I to III, Fig. 1). To characterize the differences between these clades, we computed the dissimilarity matrix of the environmental distribution of the peptides placed in each branch of the tree (see Experimental procedures). This revealed that phylogenetically close taxa were more frequently found in similar environments than distantly related taxa (Kruskal–Wallis test between within-clade dissimilarities and between-clade dissimilarities $P < 0.005$). The analysis of the matrix with non-metric multidimensional scaling (NMDS) confirmed that within-clade branches cluster together (Fig. 2), thus revealing the distinctness of the three main *Acinetobacter* clades.

Clade I shows the highest intra-clade ecological divergence among the three clades (ANOVA $P < 0.001$). Members of this clade, including the *Acinetobacter calcoaceticus-A. baumannii (ACB)* complex, were over-represented in soil and human-associated environments. *A. calcoaceticus* and *A. pittii* were very common in soil (the former being the most abundant) and present at low frequency in humans and other hosts. In contrast, *A. baumannii* was very abundant in humans. Members of clade II were frequently found in aquatic environments and rarely associated with hosts. Members of clade III were more frequently found in aquatic environments rich in organic matter, such as wastewater samples and marine sediments. The environments associated with deeper branches of clades II and III were more similar between them than those of branches closer to the tips of the tree of the same clades (Fig. 2). This is revealed by the smaller



Kullback–Leibler (KL) dissimilarities between the sets of deeper branches, which translated into a strong aggregation and a marked overlap in the NMDS representation.

The previous results suggest that we can use our approach to study the evolution of environmental distributions of bacterial taxa. Initially, we found no significant correlation between the patristic distance and the environmental distribution of taxa (Spearman's $\rho = 0.08$, $P = 0.12$). However, when we split the genus in the three clades, we found a highly significant correlation between

Fig. 1. Results of the evolutionary placement analysis.
We computed for each branch the distribution of the environmental categories associated with the fragments placed in the branch. The colour boxes indicate the branches in which the representation of fragments from certain environments was significantly higher than the average abundance for each environment across the tree (one-way Kruskal–Wallis test, *P*-value < 0.001). White boxes represent branches without any significant over-representation. The pale backgrounds represent the three large clades with similar broad environmental distribution. The distribution of dissimilarities between ancestral and descendant branches was calculated. Branches showing the top 95% of the ecological shifts towards their descendants are marked in red (see Experimental procedures). For clarity of the display, the small branches in the tree were slightly extended to allow the inclusion of the colour boxes, and only environments significantly over-represented were kept. A version including all environments can be found in Supporting Information (Fig. S18). The calculations were all done with the original tree (see Supporting Information Fig. S16). [Colour figure can be viewed at wileyonlinelibrary.com]

the two variables (Fig. 3 and Supporting Information Fig. S14). The difference between these two analyses seems to result from an amalgamation effect: the trend present in different groups of data disappears when these groups are combined (Good and Mittal, 1987). Hence, the environmental distribution of bacteria changed abruptly at the origin of each major clade and then changed gradually, at different rates, in each clade. Interestingly, our analysis shows that the rate of habitat diversification was much higher in clade I than in the others.

Host-associated Acinetobacter

Certain taxa were preferentially associated with certain hosts (Fig. 4), usually those where they were first isolated. However, some taxa were found in unexpected hosts. For example, *A. lwoffii* that was previously described in humans (Oh *et al.*, 2014), and its association with clinical samples suggests its possible emergence as an opportunistic pathogen (Tega *et al.*, 2007; Hu *et al.*, 2011; Tayabali *et al.*, 2012), was also found in the gut of *Anopheles gambiae*, where it represented around 20% of the total *Acinetobacter* assignations.

Taxa of clade I were very abundant in human-associated microbiomes, whereas those of clades II and III, and those placed deeper in the tree, were rare (Fig. 4). This suggests that frequent human-association emerged few times in the natural history of the genus and was particularly important in taxa from clade I. To detail the association of *Acinetobacter* with humans, we queried specifically the data of the

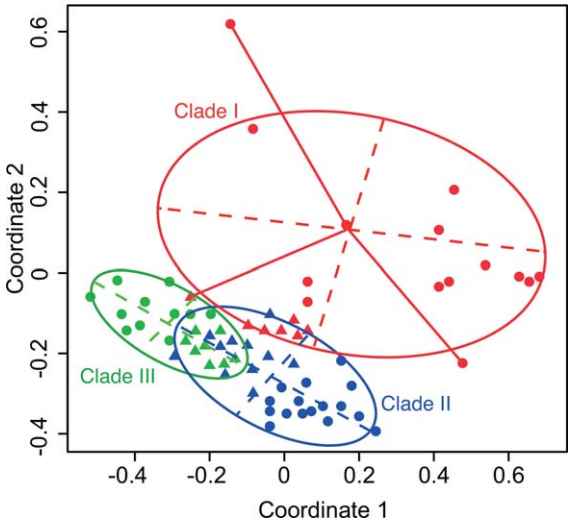


Fig. 2. Non-metric multidimensional scaling analysis of the Kullback–Leibler dissimilarity matrix of the environmental diversity associated to each branch.
Each point represents a branch in the phylogenetic tree of *Acinetobacter*. Colours define the clade of the branch. Branches that do not belong to any of the three clades were not displayed. Terminal branches are represented by circles, and deep branches by triangles. The area defined by the clusters in the *N*-dimensional space was represented as the smallest ellipse covering at least 95% of the variance of the cluster. To calculate this area we used the 'ade4' R package (Thioulouse *et al.*, 1997). [Colour figure can be viewed at wileyonlinelibrary.com]

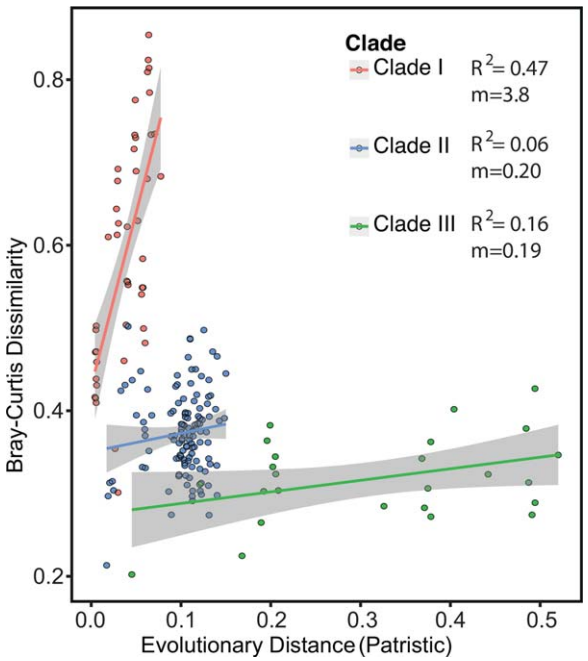


Fig. 3. Scatterplot of the Bray–Curtis dissimilarity (*Y* axis) between the different terminal branches and their phylogenetic distance (*X* axis) inside each clade.
The different clades (I, II and III) are represented by the three different colours. Spearman ρ values of the associations are: 0.61 (clade I), 0.11 (clade II) and 0.31 (clade III), all $P < 0.05$. The fitness (R^2) and slope (m) of the regression line are indicated in the figure, all $P < 0.05$. [Colour figure can be viewed at wileyonlinelibrary.com]

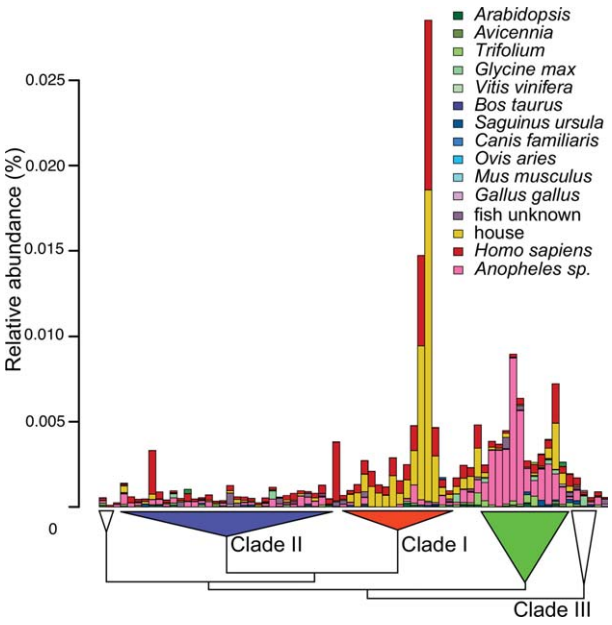


Fig. 4. Relative abundance of peptides from host-associated environments placed on the different branches of the *Acinetobacter* tree. Total abundances have been divided by the total abundance of peptides assigned to each branch, but only host-associated environments are displayed. Colours represent host types. [Colour figure can be viewed at wileyonlinelibrary.com]

Human Microbiome Project (HMP) and the Home Microbiome Project (HoMP) (Consortium, 2012; Lax *et al.*, 2014). We found similar *Acinetobacter* in skin and house-related samples of the same household in HoMP data (Spearman's $\rho = 0.82$, $P < 0.0001$) (Supporting Information Fig. S15). The identification of the natural reservoirs of *A. baumannii* is an important topic of research, given the role of this species as a nosocomial pathogen. *A. baumannii* was over-represented in the metagenomic datasets from host-associated environments (chi-square test, $P < 0.0001$), biofilms ($P < 0.001$, same test) and soil ($P < 0.0001$, same test). This fits previous observations using classical identification methods (Houang *et al.*, 2001; Vangnai and Petchkroh, 2007; Hamouda *et al.*, 2011; Rafei *et al.*, 2015). In the oral samples of the HMP, 86% of the *Acinetobacter* peptides were from *A. baumannii*.

Community response to environmental perturbations

The previous analyses showed that we could detail the evolution of environmental distributions on the tree of the genus. We then enquired on the possibility of identifying differences caused by environmental disturbances. There is now ample evidence that antibiotic treatments shape human microbiomes (Jakobsson *et al.*, 2010; Sommer and Dantas, 2011; Maurice *et al.*, 2013), and it has been suggested that *A. baumannii*'s success as a nosocomial pathogen is largely due to its intrinsic resistance to antibiotics and disinfectants (Fournier *et al.*, 2006; Dijkshoorn *et al.*, 2007; Wisplinghoff *et al.*, 2007; Diancourt *et al.*, 2010; Kempf and Rolain, 2012). We therefore tested if our method was able to identify differences in the *Acinetobacter* present in animals' microbiome treated or not with antibiotics. Comparison of bovine rumen metagenomes of treated and untreated animals with the metagenomes of humans and soil [metagenome references mgm4563763-86; mgm4497370-412 (Chambers *et al.*, 2015)], showed that the human and treated bovine samples had much less genetic diversity than the untreated and soil samples (Fig. 5A). The matrix of KL distances showed that human and treated bovine samples were much more similar than the others in terms of their composition in *Acinetobacter* (average distance to the group centroid of 1.96 and 0.44 respectively). Soil and untreated bovine samples were apart and equidistant from this group (7.5 and 5.97 respectively, ANOVA $P = 0.09$) (Fig. 5B). The resemblance between the composition of *Acinetobacter* in soil and untreated bovines suggests that soil-associated microbiota is acquired during foraging and incorporated into cattle rumen depending on the composition of the individual microbiota (in terms of *Acinetobacter* spp.). On the other hand, the similarity between treated bovines and humans suggests that antibiotic treatments in the former favour over-representation of *Acinetobacter* taxa that are usually identified in humans.

Genetic and functional bases of ecological differentiation

Our method aims at placing taxa in a known phylogenetic tree using the core genome. Yet, if one is interested in analysing genetic determinants associated with environmental transitions, or clades in a tree, one can analyse the pan-genome of the clade. To illuminate the genetic basis of transitions between clades, we searched for the genes associated with clades I to III. For this, we assessed the relative representation of every gene family of the genus pan-genome in the three clades and annotated these families using eggNOG (see Experimental procedures, Supporting Information Table S6). A total of 864 (out of 26 600) gene families were overrepresented in a specific clade ($P < 0.05$ after FDR). The vast majority of them (88%) were over-represented in clade I. Many of those families were associated with metabolism (53%, Chi-square test, $P < 0.0001$), and especially amino-acid metabolism (51% of the metabolism hits, $P < 0.0005$, same test). Some of these genes were found in clusters in the genomes, including some complete operons. For example, the urease operon, involved in colonization and virulence in a number of nosocomial pathogens (Mora and Arioli, 2014), was over-represented in clade I. The remaining families (38%) over-represented in clade I were

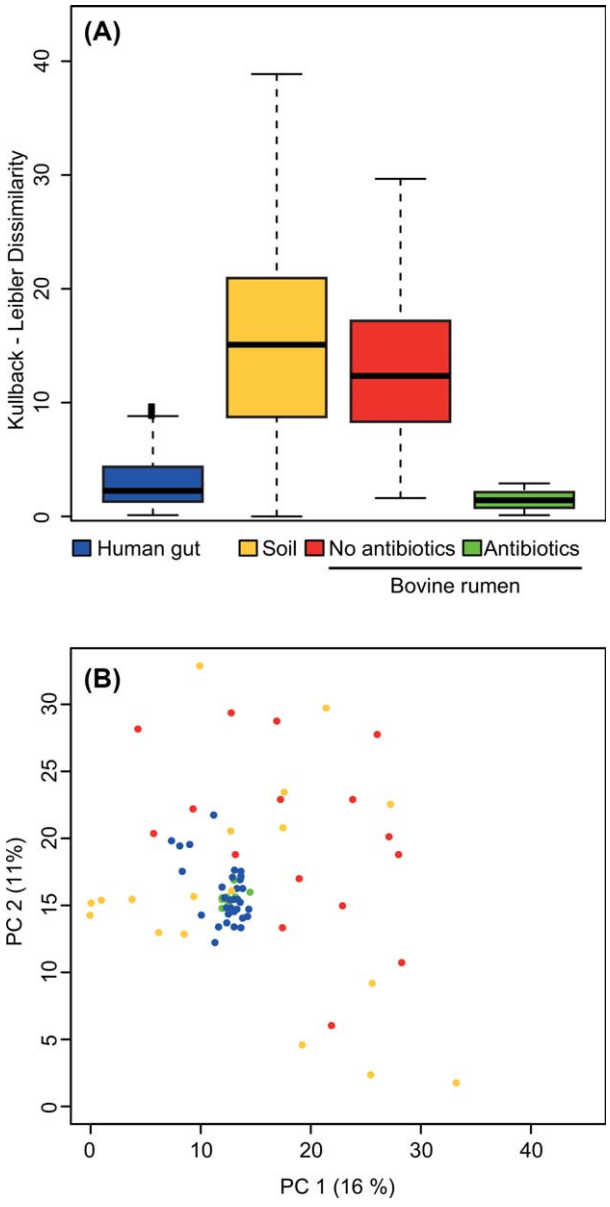


Fig. 5. A. Intragroup divergence variability. Bars represent the complete variance of each dataset. All pairwise comparisons were significantly different ($P < 0.05$, Wilcoxon tests).
B. Principal Coordinate Analysis of the Bray–Curtis dissimilarity associated to taxonomic diversity of four different metagenomic datasets. Each dot represents the projection of the Bray–Curtis dissimilarity into the dissimilarity space. Colours are assigned according to the metagenomic dataset. [Colour figure can be viewed at wileyonlinelibrary.com]

involved in environmental interactions, including siderophore biosynthesis and transport or antibiotic resistance. Only cell wall and envelope biogenesis were over-represented in clade II (Chi-square test $P = 0.00049$). No categories were over-represented in clade III. Hence, most clade-associated genetic traits were acquired by genomes of clade I and may be involved in its evolution towards association with animals.

Discussion

Methodological limitations and implications

Other methods have tried to identify genus-specific sequences in metagenomics data. Methods such as PhymmBL, Kraken or Metaphlan, aim to assess the taxonomic distribution of metagenomic datasets by matching them against a subset of genomic markers (Brady and Salzberg, 2009; Segata *et al*., 2012; Wood and Salzberg, 2014). These methods can process large amounts of information with good accuracy for known taxa, because they look for highly similar hits against either a small set of universal markers (in the case of PhymmBL) or a subset of species-discriminant markers (in the case of Kraken and Metaphlan). However, these methods do not place the sequence in a phylogenetic scenario and do not provide precise information on the evolutionary distance between the environmental taxa and the references. EPA of sequence fragments was pioneered by Phylosift, which uses 37 nearly universal single copy genes to obtain a broad classification of bacteria and archaea (Darling *et al*., 2014). In contrast, our method uses thousands of clade-specific genes and is thus expected to be more accurate at the genus-level, at the cost of having to identify the core genome of the clade, and of the outgroups, and compute the associated protein profiles. This requires a certain degree of expertise from the user in order to produce the required core-genomes, protein alignments and the phylogenetic tree. Nevertheless, there are different user-friendly tools available, such as Roary (Page *et al*., 2015), that produce the necessary core genome data for our pipeline. While the method is reproducible, its accuracy is expected to increase with the number of profiles and with the distance of the closest outgroup. Accordingly, we were able to map ten times more fragments with our method, while fetching seven times fewer false positives, with our method relative to phylosift (see '*Validation of the Core-Genome EPA compared to Phylosift*' in Supporting Information). Hence, these two different ways of using EPA are complementary; phylosift is more adequate to identify large phyla, whereas our approach is more accurate to study the ecological diversification of lineages at the genus level. If the goal of the analysis is to study even narrower taxa, such a clonal complex in a species, then these phylogenetic methods must be replaced by methods focusing on the identification of strain-specific genes.

As the other abovementioned programs, our method assumes that sequences branching in the *Acinetobacter* tree are from *Acinetobacter* genomes. This will produce false positives when genomes from other taxa have recently acquired genes from *Acinetobacter*. We believe that this problem will have little effect on the context of large-scale analyses because core genes, contrary to clade-specific genes, are transferred between distant

species at low rates (Abby *et al*., 2012). Also, the use of a large number of core genes should diminish the effect of a given event of horizontal gene transfer. Accordingly, the results of the LDA + SOM analysis showed a very high accuracy, confirming that sequences from the outgroups were not mistaken as *Acinetobacter*.

In this study, we selected the closest known genera to *Acinetobacter* as the close outgroup. We assessed the performance of our pipeline when the outgroups were more distant, *Moraxella* (closest outgroup) and Enterobacteria (distant outgroup) (*see* Supporting Information). The selection of different outgroups had an impact on the number of fragments kept during the discriminative phase, as 13.4% of the fragments coming from *Moraxella* passed the discriminant analysis and were kept for the evolutionary placement, within which only very few (2.4%) were placed inside the *Acinetobacter* genus. Hence, a few sequences of clades more closely related to the genus than to the closest outgroup considered in the analysis can be mapped erroneously in the genus. Importantly, the placement of *Acinetobacter* fragments was not affected by changing the outgroups.

Our method has the interesting property of identifying where the taxa branch in the known phylogeny and in which environments they are more susceptible to be isolated. This could dramatically accelerate the identification of novel bacterial taxa and their niches. It is interesting to observe that many of the novel lineages of *Acinetobacter* had not been observed before, in spite of previous projects aiming at sequencing all known species in the genus (Touchon *et al*., 2014), and recent efforts to identify novel species (Krizova *et al*., 2015; Maixnerova *et al*., 2015; Nemec *et al*., 2015; Sedo *et al*., 2016). Importantly, these potentially novel lineages were generally found in metagenomic datasets from environments similar to those of their close relatives. Nevertheless, we observed some disagreements between previous literature and our results on the distribution of *Acinetobacter* species in environments. Some of these may result from the coarse-grained classification of metagenomic samples into environmental categories, which is essential to attain sufficient statistical power to test our hypotheses, but may have resulted in some over-simplifications. However, a careful analysis of the most striking discrepancies suggests different reasons. For example, *A. baumannii* has often been described as present in human-associated environments and in biofilms in dry inert surfaces (Peleg *et al*., 2012; McConnell *et al*., 2013). This makes it particularly well-adapted to clinical settings and may explain its success as a nosocomial pathogen. We consistently found *A. baumannii* in household surfaces and other environments classified as biofilms. However, this species was not systematically found in all biofilm-associated datasets, probably due to the characteristics of the sampled environments. We have also identified it in soil samples. While previous association between the *A. baumannii* and the soil has been reported using cultivation-based approaches (Houang *et al*., 2001; Hrenovic *et al*., 2014), the difficulty in reproducing these studies by others has led to suggestions that they resulted from methodological artifacts (Peleg *et al*., 2008). In our work, most *A. baumannii* were identified in the samples from humans and soil. Different metagenomic datasets consistently showed small abundances of *A. baumannii* in soil samples of different geographic locations, such as USA, UK or France (Delmont *et al*., 2012; Fierer *et al*., 2013). These discrepancies between cultivation-based methods and metagenomics are not new. For instance, Benitez-Paez and colleagues described a completely different bacterial composition in oral samples when they compared classical isolation methods to a metagenomic approach (Benítez-Páez *et al*., 2013). These results highlight the interest of complementing classical isolation methods (Browne *et al*., 2016) with cultivation-free metagenomics to study microbial ecology. Indeed, the use of metagenomic approaches should be considered as the preliminary step to characterize the diversity of communities and to identify novel lineages. These approaches will then require comparative genomics analyses to understand the process of diversification of the novel lineages. When these taxa are highly abundant in the sample, genome-resolved metagenomics may allow to recover their genomic content (Alneberg *et al*., 2014; Nielsen *et al*., 2014; Cleary *et al*., 2015; Lu *et al*., 2017). Other alternatives, when the genome assembly cannot be obtained from metagenomes, might include targeted bacteria and targeted cultivation based on initial metabolic characterization.

The diversification of the genus and the emergence of nosocomial lineages

We consistently found *Acinetobacter* in four major types of environments. Several taxa were frequently found in more than one environment, and sporadically in others, suggesting they have the ability to colonize multiple environments transiently or permanently. This may have facilitated their environmental diversification. Closely related taxa tend to inhabit closely related environments. This not only applies to the lineages present in the original dataset used to build the phylogenetic tree, but also the new lineages predicted by EPA. However, there were exceptions to this trend, including some sharp transitions that led to the rapid diversification of clades I and II. In fact, branches close to the origin of clades I and II are among those showing larger differences in their environmental distribution, compared to their close relatives.

We identified the major environmental shifts in the evolutionary history of the genus by comparing the differences

of environmental-associated peptides in immediately ancestral and descendant branches at each node. These differences measure the change in the capacity to inhabit a specific environment by the immediate descendants of that node, probably by the acquisition/loss of a genetic repertoire that allows them to occupy those niches. The transition to clade I is especially interesting because the latter is the most frequently found in the human microbiome. The extensive use of antibiotics by humans started long time after this transition. Yet, the existence of a clade that is often associated with hosts, with an appropriate repertoire of specific host-associated genes, and the selection of more resistant lineages through the last decades, might have facilitated the recent emergence of nosocomial *Acinetobacter* species. Two observations seem to agree with this hypothesis: First, the identification of non-pathogenic lineages from the ACB complex (including *A. calcoaceticus*) in human samples and the rapid diversification between *A. baumannii* and the other members of clade I. Second, the differences observed in the microbiome of bovines exposed to antibiotics (relative to the others), showing a shift towards the *Acinetobacter* spp. typically found in humans.

Clade II and III are markedly different between them but show similar correlations between the environmental dissimilarity and the phylogenetic distance of its members. In contrast, clade I shows a strikingly more rapid ecological diversification. This is reflected in the relative abundances of each species of clade I in the different environments. The difference between the abundances of *A. baumannii* in soil and host-associated environments suggests that this species might be rapidly specializing towards human-associated environments (including humans, human-associated hosts and house-holds), while the other members of the clade remained abundant in soil. Further research might shed light in the association between members of clade I and the human-associated environments and the effect of the recent massive antibiotic usage in humans and other human-associated environments. Our observations about the transitions observed in the branches at the origin of the major clades, together with the sharp diversification observed in clade I support the hypothesis that such environmental disturbance will be accompanied by an important diversification of the lineage.

Experimental procedures

Genome data

We analysed the 133 complete genomes of 29 validly named species and 8 genomic species of *Acinetobacter* analysed in (Touchon *et al.*, 2014) (Supporting Information Table S1). The places of isolation of the bacteria were retrieved from the same reference or from the literature. We also analysed the 2644 complete genomes available in GenBank RefSeq (Supporting Information Table S2, last accessed November 2013). At the end of the study, we retrieved from RefSeq 29 novel genomes of *Acinetobacter* (not identified as *A. baumannii*), that were published after the work of (Touchon *et al.*, 2014) (last accessed October 2015). We only used genomes that mentioned the isolation site of the strain (Supporting Information Table S3). The 16 genomes that lacked annotation were annotated using prodigal v.2.6.2 (Hyatt *et al.*, 2010) (default parameters).

Definition of core-genomes and pan-genomes

Core-genomes were defined as the families of orthologous genes ubiquitous in a given clade. The pan-genome of a clade was defined as the repertoire of gene families present in that clade. Both core-genome and pan-genome reconstructions were performed following the approach from (Touchon *et al.*, 2014). The list of core-genome profiles and a Supporting Information Table S7 listing the gene families are included in Supporting Information. For more information, see Supporting Information.

Identification of protein families over-represented in clades I to III

We used hmmsearch from HMMer v.3.1.2 (Eddy, 2011) to search for the best hit (e-value $< 10^{-5}$) of each protein profile in the eggNOG database v.4.0 (Powell *et al.*, 2013). Protein profiles were annotated using the functional information of the best eggNOG hit. Of them, 40% of the proteins were not assigned to any eggnog category and were discarded from further analyses. We then compared the abundances of the different eggNOG categories in each clade relative to the whole genus' pan-genome. The over-representation of protein families was assessed statistically using the Pearson Chi square test with Benjamini–Hochberg correction for multiple tests (Benjamini and Hochberg, 1995).

Two pan-genome protein families were considered to be in a relation of gene order conservation when the respective genes co-localized (less than 5 CDS apart) in all the genomes where the two families were present.

Metagenomics data

We analysed 2568 metagenomic datasets from 126 independent locations with relevant meta-data retrieved from MG-RAST (Meyer *et al.*, 2008). These sets represent a broad diversity of host-associated and environmental ecosystems. They contain $\sim 6 \times 10^{11}$ metagenomic fragments (6 Terabytes of data). We only retrieved the datasets with multiple samples (to be able to assess the diversity within a sequencing project). We ignored datasets obtained by procedures involving amplification, because they may generate genomic and metagenomic coverage biases, produce chimeric contaminant sequences and over or under-estimate the abundance of certain taxa (Džunková *et al.*, 2014; Marine *et al.*, 2014). We grouped the datasets in 8 major environmental categories and 39 sub-categories (Supporting Information Table S4). We searched for ORFs in the metagenomic data using FragGeneScan v.1.17 using the options to operate with fragmented

data (Rho *et al.*, 2010). We only kept fragments of size higher than 35 amino acids, based on the distribution of false positives and negatives observed during the methodological validation (see *Simulated metagenomic fragments*). Fragments were queried against the core-genome HMM profiles using hmmsearch. Significant hits were kept for the EPA analysis. For more details, see Supporting Information.

Simulated metagenomic fragments

We sampled random fragments from translated genomic data. As the distribution of read sizes usually fits a gamma distribution (Richter *et al.*, 2008), we selected fragments with lengths following this distribution with parameters: $\kappa = X$ and $\lambda = 1$, where X is the average fragment size. We made separate analyses for fragments with $X = \{35, 50, 75, 85, 100, 115\}$ amino acids (see Supporting Information Fig. S3).

Linear discriminant analysis

We queried the simulated metagenomes of *Acinetobacter* and both outgroups with the three sets of protein profiles (one per clade core-genome). The distributions of the nine sets of normalized scores were analysed with LDA. This analysis was performed in R, using an estimation based on a T distribution and assuming equal prior assignation probabilities for all three groups. Based on these results, we calculated for each *Acinetobacter* fragment *i* the ratio (R_i) between the best normalized score obtained with the profiles of *Acinetobacter* profiles ($S_{A,A,i}$) and the best normalized score obtained with the profiles of the close and distant outgroups ($S_{CO,A,i}$; $S_{DO,A,i}$).

We used $S_{A,A,i}$ and R_i to define a metagenomic peptide as *Acinetobacter*, based on the results of LDA. The conditions for a peptide to be classed as *Acinetobacter* were thus: $S_{A,A,I} \geq \text{Max}(S_{CO,A,i}; S_{DO,A,i})$, $S_{A,A,i} > 1.45$ and $R_i > 1.09$.

Phylogenetic reconstruction

The original *Acinetobacter* phylogenetic tree had some short branches (Touchon *et al.*, 2014). EPA cannot assign short reads to these branches accurately. To reduce this problem, we selected a lineage representative at species level, and removed the other strains from that already represented lineage. The pruned protein alignments of the core-genome were back-translated to DNA (each amino acid was replaced by the original codon), as is the best practice in evolutionary analyses [see (Touchon *et al.*, 2014)]. Poorly aligned regions were trimmed with trimAL v.1.4 using the *automated1* algorithm (Capella-Gutiérrez *et al.*, 2009). The final phylogenetic tree was inferred using RAxML v.8.1.2 with the model GTR + I + G (Stamatakis, 2014). We assessed the robustness of the topology with 1000 bootstrap experiments. We used another alignment of 630 orthologous genes common to *Acinetobacter* and the close outgroup to root the tree (using the same method).

Evolutionary placement analysis

We used maximum likelihood to place on the phylogenetic tree the metagenomic peptides preselected with the SOM. To reduce computational time (and inaccurate placement of outgroups), we removed the few remaining very divergent peptides. For this, we computed the minimal pairwise sequence distances ($X_{Fi,Aj}$) between each peptide (F_i) and the reference *Acinetobacter* sequences (A_j) of the *k* genomes corresponding to the protein profile that best hits F_i:

$$X_{F_i,A_j} = \min_{k \in \{1133\}} \left(D_{F_i, A_{j,k}} \right)$$

We also computed the maximal pairwise sequence distances $\left(Y_{A_{i,j}} \right)$ between all the *Acinetobacter* sequences of the core-gene family *j* on the exact region where the peptide F_i matched.

$$Y_{A_{i,j}} = \max_{k,l \in \{1133\},\, k \neq l} \left(D_{A_{i,j,k}, A_{i,j,l}} \right)$$

If $X_{F_i,A_j} > Y_{A_{i,j}}$ then the maximum pairwise distance between the core-genes is smaller than all the matches of the queried fragment with each of the core-genes. These peptides were thus considered to be outgroups and were removed from further analysis. The others were incorporated in the multiple alignments using the *addfragments* algorithm of MAFFT v. 7.153b (Katoh *et al.*, 2002). They were then placed on the phylogenetic tree using the EPA included in RAxML v.8.1.2 (Berger *et al.*, 2011), with the '-f v' option and the same evolutionary model used in the phylogenetic reconstruction. For more information about the validation of EPA, see Supporting Information. The pipeline scripts are available at https://gitlab.pasteur.fr/gem/Core-Genome-EPA.

The number of assignations per metagenome sample was correlated with the size of the metagenomic dataset (Spearman's $\rho = 0.68$, $P < 0.0001$). Hence, we divided the abundance of the hits from a specific dataset by the total number of peptides in the metagenomic dataset.

Community analysis of the Acinetobacter *selected peptides*

We determined the distribution of the peptides placed by the EPA for each environment in each phylogenetic branch of the reference tree of the *Acinetobacter* genus. We then assessed similarities and differences between these environmental distributions across the branches of the tree. For this, we computed Bray–Curtis (BC) and KL dissimilarity matrices using the vegan R-package and the functions defined in Faust and colleagues (Oksanen *et al.*, 2008; Faust *et al.*, 2012). These two measurements are often used to quantify the differences between datasets, either by comparing the compositional dissimilarity between sites (BC, in this case environmental composition between branches) or by comparing the probability distributions derived from the observed frequencies in each dataset (KL) (Gorelick and Bertram, 2010). These two matrices were then independently analysed with NMDS analyses. This method finds a non-parametric monotonic relationship between dissimilarities and ranks them in a smaller set of dimensions that can be represented in a *N*-dimensional space (Kruskal, 1964). We performed a k-means clustering on the KL dissimilarity matrix, using the Calinski–Harabasz index to define the optimal number of clusters (Calinski and Harabasz, 2007). The statistical robustness of clustering was assessed using 100 bootstrap analyses.

We analysed the environmental shifts along the phylogenetic tree. To do so, we computed the KL dissimilarity between the distribution of environmental sources of peptides placed at every ancestral branch and at the two immediately descendent branches. This resulted in two values per node. The distribution of dissimilarities was analysed to highlight the nodes with the most distinct differences between ancestral and descendant branches. When these values were in the top 95% of the distribution, they were marked in red in Fig. 5 in the place of the corresponding descendant branch.

Acknowledgements

This work was supported by the H2020 European Research Council grant [EVOMOBILOME n°281605]. We thank Alexandr Nemec, Pilar Francino, Sandrine Isaac and Pedro Oliveira for comments and criticisms in earlier versions of this manuscript. The authors report no conflicts of interest.

Author Contributions

M.G.G. conceived the project, produced and analysed the data and wrote the paper. M.T. and S.B. provided data, helped interpreting the results and writing the paper. E.P.C.R. conceived and ran the project, helped analysing the data and wrote the paper.

References

Abby, S.S., Tannier, E., Gouy, M., and Daubin, V. (2012) Lateral gene transfer as a support for the tree of life. *Proc Natl Acad Sci USA* **109:** 4962–4967.

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., *et al*. (2014) Binning metagenomic contigs by coverage and composition. *Nat Methods* **11:** 1144–1146.

Benítez-Páez, A., Álvarez, M., Belda-Ferre, P., Rubido, S., Mira, A., and Tomás, I. (2013) Detection of transient bacteraemia following dental extractions by 16S rDNA pyrosequencing: a pilot study. *PLoS One* **8:** e57782.

Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* **57:** 289–300.

Berger, S.A., Krompass, D., and Stamatakis, A. (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol* **60:** 291–302.

Bergogne-Bérézin, E., and Joly-Guillou, M.L. (1991) Hospital infection with Acinetobacter spp.: an increasing problem. *J Hosp Infect* **18:** 250–255.

Bialek-Davenet, S., Criscuolo, A., Ailloud, F., Passet, V., Jones, L., Delannoy-Vieillard, A.-S., *et al*. (2014) Genomic definition of hypervirulent and multidrug-resistant Klebsiella pneumoniae clonal groups. *Emerging Infect Dis* **20:** 1812–1820.

Brady, A., and Salzberg, S.L. (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* **6:** 673–676.

Browne, H.P., Forster, S.C., Anonye, B.O., Kumar, N., Neville, B.A., Stares, M.D., *et al*. (2016) Culturing of "unculturable" human microbiota reveals novel taxa and extensive sporulation. *Nature* **533:** 543–546.

Calinski, T., and Harabasz, J. (2007) A dendrite method for cluster analysis. *Commun Stat* **3:** 1–27.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25:** 1972–1973.

Chambers, L., Yang, Y., Littier, H., Ray, P., Zhang, T., Pruden, A., *et al*. (2015) Metagenomic analysis of antibiotic resistance genes in dairy cow feces following therapeutic administration of third generation cephalosporin. *PLoS One* **10:** e0133764.

Cleary, B., Brito, I.L., Huang, K., Gevers, D., Shea, T., Young, S., and Alm, E.J. (2015) Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat Biotechnol* **33:** 1053–1060.

Consortium, T.H.M.P. (2012) Structure, function and diversity of the healthy human microbiome. *Nature* **486:** 207–214.

Darling, A.E., Jospin, G., Lowe, E., Matsen, F.A., Bik, H.M., and Eisen, J.A. (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2:** e243.

Delmont, T.O., Simonet, P., and Vogel, T.M. (2012) Describing microbial communities and performing global comparisons in the 'omic era. *ISME J* **6:** 1625–1628.

Diancourt, L., Passet, V., Nemec, A., Dijkshoorn, L., and Brisse, S. (2010) The population structure of *Acinetobacter baumannii*: expanding multiresistant clones from an ancestral susceptible genetic pool. *PLoS One* **5:** e10034–e10017.

Dijkshoorn, L., Nemec, A., and Seifert, H. (2007) An increasing threat in hospitals: multidrug-resistant *Acinetobacter baumannii*. *Nat Rev Microbiol* **5:** 939–951.

Doughari, H.J., Ndakidemi, P.A., Human, I.S., and Benade, S. (2011) The ecology, biology and pathogenesis of Acinetobacter spp.: an overview. *Microbes Environ* **26:** 101–112.

Džunková, M., Garcia-Garcerà, M., Martínez-Priego, L., D'Auria, G., Calafell, F., and Moya, A. (2014) Direct sequencing from the minimal number of DNA molecules needed to fill a 454 picotiter plate. *PLoS One* **9:** e97379.

Eddy, S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* **7:** e1002195.

Ehrlén, J., and Morris, W.F. (2015) Predicting changes in the distribution and abundance of species under environmental change. *Ecol Lett* **18:** 303–314.

Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012) Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol* **8:** e1002606.

Fierer, N., Ladau, J., Clemente, J.C., Leff, J.W., Owens, S.M., Pollard, K.S., *et al*. (2013) Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States. *Science* **342:** 621–624.

Fournier, P.-E., Vallenet, D., Barbe, V., Audic, S., Ogata, H., Poirel, L., *et al*. (2006) Comparative genomics of multidrug resistance in *Acinetobacter baumannii*. *PLoS Genet* **2:** e7.

Gianoulis, T.A., Raes, J., Patel, P.V., Bjornson, R., Korbel, J.O., Letunic, I., *et al*. (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci USA* **106:** 1374–1379.

Good, I.J., and Mittal, Y. (1987) The amalgamation and geometry of two-by-two contingency tables. *Ann Stat* **15:** 694–711.

Gorelick, R., and Bertram, S.M. (2010) Multi-way multi-group segregation and diversity indices. *PLoS One* **5:** e10912.

Hamouda, A., Findlay, J., Hassan, A.L., and Amyes, S.G.B. (2011) Epidemiology of *Acinetobacter baumannii* of animal origin. *Int J Antimicrob Agents* **38:** 314–318.

Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68:** 669–685.

Heath, T.A., Zwickl, D.J., Kim, J., and Hillis, D.M. (2008) Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. *Syst Biol* **57:** 160–166.

Houang, E.T.S., Chu, Y.W., Leung, C.M., Chu, K.Y., Berlau, J., Ng, K.C., and Cheng, A.F.B. (2001) Epidemiology and infection control implications of Acinetobacter spp. in Hong Kong. *J Clin Microbiol* **39:** 228–234.

Hrenovic, J., Durn, G., Goic-Barisic, I., and Kovacic, A. (2014) Occurrence of an environmental *Acinetobacter baumannii* strain similar to a clinical isolate in paleosol from Croatia. *Appl Environ Microbiol* **80:** 2860–2866.

Hu, Y., Zhang, W., Liang, H., Liu, L., Peng, G., Pan, Y., *et al.* (2011) Whole-genome sequence of a multidrug-resistant clinical isolate of *Acinetobacter lwoffii*. *J Bacteriol* **193:** 5549–5550.

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11:** 119.

Jakobsson, H.E., Jernberg, C., Andersson, A.F., Sjölund-Karlsson, M., Jansson, J.K., and Engstrand, L. (2010) Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLoS One* **5:** e9836.

Katoh, K., Misawa, K., Kuma, K.-I., and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30:** 3059–3066.

Kempf, M., and Rolain, J.-M. (2012) Emergence of resistance to carbapenems in *Acinetobacter baumannii* in Europe: clinical impact and therapeutic options. *Int J Antimicrob Agents* **39:** 105–114.

Krizova, L., McGinnis, J., Maixnerova, M., Nemec, M., Poirel, L., Mingle, L., *et al.* (2015) Acinetobacter variabilis sp. nov. (formerly DNA group 15 sensu Tjernberg & Ursing), isolated from humans and animals. *Int J Syst Evol Microbiol* **65:** 857–863.

Kruskal, J.B. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29:** 1–27.

Lax, S., Smith, D.P., Hampton-Marcell, J., Owens, S.M., Handley, K.M., Scott, N.M., *et al.* (2014) Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* **345:** 1048–1052.

Ley, R.E., Peterson, D.A., and Gordon, J.I. (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124:** 837–848.

Lu, Y.Y., Chen, T., Fuhrman, J.A., and Sun, F. (2017) COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics* **33:** 791–798.

Maixnerova, M., Sedo, O., Nemec, A., and Krizova, L. (2015) Acinetobacter albensis sp. nov., isolated from natural soil and water ecosystems. *Int J Syst Evol Microbiol* **65:** 3905–3912.

Marine, R., McCarren, C., Vorrasane, V., Nasko, D., Crowgey, E., Polson, S.W., and Wommack, K.E. (2014) Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome* **2:** 3.

Martinez, J.L. (2009) The role of natural environments in the evolution of resistance traits in pathogenic bacteria. *Proc Biol Sci* **276:** 2521–2530.

Maurice, C.F., Haiser, H.J., and Turnbaugh, P.J. (2013) Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell* **152:** 39–50.

McConnell, M.J., Actis, L., and Pachón, J. (2013) *Acinetobacter baumannii*: human infections, factors contributing to pathogenesis and animal models. *FEMS Microbiol Rev* **37:** 130–155.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9:** 386.

Mora, D., and Arioli, S. (2014) Microbial urease in health and disease. *PLoS Pathog* **10:** e1004472.

Nemec, A., Krizova, L., Maixnerova, M., Sedo, O., Brisse, S., and Higgins, P.G. (2015) Acinetobacter seifertii sp. nov., a member of the Acinetobacter calcoaceticus-*Acinetobacter baumannii* complex isolated from human clinical specimens. *Int J Syst Evol Microbiol* **65:** 934–942.

Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., *et al.* (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* **32:** 822–828.

Oh, J., Byrd, A.L., Deming, C., Conlan, S., NISC Comparative Sequencing Program, Kong, H.H., and Segre, J.A. (2014) Biogeography and individuality shape function in the human skin metagenome. *Nature* **514:** 59–64.

Oksanen, J., Kindt, R., Legendre, P., and O'Hara, B. (2008) *Vegan: Community Ecology Package*. R package. Vienna, Austria: R Foundation for Statistical Computing.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., *et al.* (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31:** 3691–3693.

Peleg, A.Y., Seifert, H., and Paterson, D.L. (2008) *Acinetobacter baumannii*: emergence of a Successful Pathogen. *Clin Microbiol Rev* **21:** 538–582.

Peleg, A.Y., de Breij, A., Adams, M.D., Cerqueira, G.M., Mocali, S., Galardini, M., *et al.* (2012) The success of Acinetobacter species; genetic, metabolic and virulence attributes. *PLoS One* **7:** e46984.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., *et al.* (2013) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res* **42:** D231–D239.

Rafei, R., Hamze, M., Pailhoriès, H., Eveillard, M., Marsollier, L., Joly-Guillou, M.-L., *et al.* (2015) Extrahuman epidemiology of *Acinetobacter baumannii* in Lebanon. *Appl Environ Microbiol* **81:** 2359–2367.

Rho, M., Tang, H., and Ye, Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* **38:** e191–e191.

Richter, D.C., Ott, F., Auch, A.F., Schmid, R., and Huson, D.H. (2008) MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One* **3:** e3373.

Rodriguez-R, L.M., and Konstantinidis, K.T. (2014) Bypassing cultivation to identify bacterial species. *Microbe* **9:** 111–118.

Sahl, J.W., Gillece, J.D., Schupp, J.M., Waddell, V.G., Driebe, E.M., Engelthaler, D.M., and Keim, P. (2013) Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to pathogenesis in Acinetobacter. *PLoS One* **8:** e54287.

Sedo, O., Maixnerova, M., Nemec, A., Jezek, P., Vrestiakova, E., and Radolfova-Krizova, L. (2016) Taxonomy of haemolytic and/or proteolytic strains of the genus Acinetobacter with the proposal of Acinetobacter courvalinii sp. nov. (genomic species 14 sensu Bouvet & Jeanjean), Acinetobacter dispersus sp. nov. (genomic species 17), Acinetobacter modestus sp. nov., Acinetobacter proteolyticus sp. nov. and Acinetobacter vivianii sp. nov. *Int J Syst Evol Microbiol* **66:** 1673–1685.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9:** 811–814.

Seifert, H., Dijkshoorn, L., Gerner-Smidt, P., Pelzer, N., Tjernberg, I., and Vaneechoutte, M. (1997) Distribution of Acinetobacter species on human skin: comparison of phenotypic and genotypic identification methods. *J Clin Microbiol* **35:** 2819–2825.

Smillie, C.S., Smith, M.B., Friedman, J., Cordero, O.X., David, L.A., and Alm, E.J. (2011) Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480:** 241–244.

Sommer, M.O., and Dantas, G. (2011) Antibiotics and the resistant microbiome. *Curr Opin Microbiol* **14:** 556–563.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313.

Tayabali, A.F., Nguyen, K.C., Shwed, P.S., Crosthwait, J., Coleman, G., and Seligy, V.L. (2012) Comparison of the virulence potential of Acinetobacter strains from clinical and environmental sources. *PLoS One* **7:** e37024.

Tega, L., Raieta, K., Ottaviani, D., Russo, G.L., Blanco, G., and Carraturo, A. (2007) Catheter-related bacteremia and multidrug-resistant *Acinetobacter lwoffii*. *Emerg Infect Dis* **13:** 355–356.

Thioulouse, J., Chessel, D., Dec, S.D., and Olivier, J.-M. (1997) ADE-4: a multivariate analysis and graphical display software. *Stat Comput* **7:** 75–83.

Tjernberg, I., and Ursing, J. (1989) Clinical strains of Acinetobacter classified by DNA-DNA hybridization. *APMIS* **97:** 595–605.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., *et al.* (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* **5:** e1000344.

Touchon, M., Cury, J., Yoon, E.-J., Krizova, L., Cerqueira, G.C., Murphy, C., *et al.* (2014) The genomic diversification of the whole Acinetobacter genus: origins, mechanisms, and consequences. *Genome Biol Evol* **6:** 2866–2882.

Treangen, T.J., and Rocha, E.P.C. (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* **7:** e1001284.

Vallenet, D., Nordmann, P., Barbe, V., Poirel, L., Mangenot, S., Bataille, E., *et al.* (2008) Comparative analysis of Acinetobacters: three genomes for three lifestyles. *PLoS One* **3:** e1805.

Vangnai, A.S., and Petchkroh, W. (2007) Biodegradation of 4-chloroaniline by bacteria enriched from soil. *FEMS Microbiol Lett* **268:** 209–216.

Wisplinghoff, H., Schmitt, R., Wöhrmann, A., Stefanik, D., and Seifert, H. (2007) Resistance to disinfectants in epidemiologically defined clinical isolates of *Acinetobacter baumannii*. *J Hosp Infect* **66:** 174–181.

Wood, D.E., and Salzberg, S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15:** R46.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Overview of the analysis. Metagenomic data was collected, processed and annotated (red). Genomic data was used to build the core and pan-genomes of *Acinetobacter* and two close outgroups (blue). The core-genome was used to build a phylogenetic tree of the *Acinetobacter* genus. Metagenomic and genomic data were integrated to place the metagenomics fragments on the *Acinetobacter* genus tree using EPA (purple). The latter were then used to analyse the distribution of *Acinetobacter* fragments in the light of their position in the phylogenetic tree and the environment where they were sampled.

Fig. S2. Scatterplot of the scores of fragments matching the protein profiles of the core-genomes of *Acinetobacter* (*X*-axis) and the close outgroup (*Y*-axis). The fragments were sampled from the complete genomes of *Acinetobacter* (blue) and the close outgroup (red). Fragments from genes that were in none of the core-genomes were coloured in grey.

Fig. S3. Scatterplots of the scores of peptides matching the protein profiles of the core genomes of *Acinetobacter* (*X*-axis) and the close outgroup (*Y*-axis). The fragments were sampled from the complete genomes of *Acinetobacter* (in blues) and the close outgroup (reds). Each plot is associated to a different average fragment size.

Fig. S4. Receiver Operating characteristic (ROC) curve illustrating the performance of our binary classifier. The *X* axis shows the false positive rate (FP) and the *Y* axis shows the rate of True Positives (TP). The colour gradient shows the maximum $S_{A,A,i}$ value found at a specific TP/TN rate point.

Fig. S5. Phylogenetic reconstruction of the *Acinetobacter* spp. core genome, including the six new isolates. The new isolates highlighted in red are those used in the validation of the EPA. ANI values for those isolates are included in the figure.

Fig. S6. Distribution of the phylogenetic distances between the placement of a simulated fragment by the EPA, and the true correct position according to the core-genome phylogenetic reconstruction. The distances were divided by the

maximal tip-to-root distance in the tree and are presented as percentages. The different colours represent the genomic origin of the simulated fragments.

Fig. S7. *Acinetobacter sp. MN12*. Distribution of the phylogenetic distances between the placement of a simulated fragment by the EPA, and the true correct position according to the core-genome phylogenetic reconstruction. The distances were divided by the maximal tip-to-root distance in the tree and are presented as percentages. The different colours represent the two methodologies used: Our approach (in red) and Phylosift (in blue). Phylosift consistently displays two separated groups of placements, separated by a small gap. This gap relates to the use of highly conserved markers, which distinguishes perfect matches to sequences originally included in the profile construction and distantly related matches (and therefore internal branches), leading to the granularity observed in the different figures.

Fig. S8. *Acinetobacter sp*. Ver3. Distribution of the phylogenetic distances between the placement of a simulated fragment by the EPA, and the true correct position according to the core-genome phylogenetic reconstruction. The distances were divided by the maximal tip-to-root distance in the tree and are presented as percentages. The different colours represent the two methodologies used: Our approach (in red) and Phylosift (in blue). Phylosift consistently displays two separated groups of placements, separated by a small gap. This gap relates to the use of highly conserved markers, which distinguishes perfect matches to sequences originally included in the profile construction and distantly related matches (and therefore internal branches), leading to the granularity observed in the different figures.

Fig. S9. *Acinetobacter sp*. MDS7A. Distribution of the phylogenetic distances between the placement of a simulated fragment by the EPA, and the true correct position according to the core-genome phylogenetic reconstruction. The distances were divided by the maximal tip-to-root distance in the tree and are presented as percentages. The different colours represent the two methodologies used: Our approach (in red) and Phylosift (in blue). Phylosift consistently displays two separated groups of placements, separated by a small gap. This gap relates to the use of highly conserved markers, which distinguishes perfect matches to sequences originally included in the profile construction and distantly related matches (and therefore internal branches), leading to the granularity observed in the different figures.

Fig. S10. *Acinetobacter* sp. TTH0-4. Distribution of the phylogenetic distances between the placement of a simulated fragment by the EPA, and the true correct position according to the core-genome phylogenetic reconstruction. The distances were divided by the maximal tip-to-root distance in the tree and are presented as percentages. The different colours represent the two methodologies used: Our approach (in red) and Phylosift (in blue). Phylosift consistently displays two separated groups of placements, separated by a small gap. This gap relates to the use of highly conserved markers, which distinguishes perfect matches to sequences originally included in the profile construction and distantly related matches (and therefore internal branches), leading to the granularity observed in the different figures.

Fig. S11. *Acinetobacter* sp. HR7. Distribution of the phylogenetic distances between the placement of a simulated fragment by the EPA, and the true correct position according to the core-genome phylogenetic reconstruction. The distances were divided by the maximal tip-to-root distance in the tree and are presented as percentages. The different colours represent the two methodologies used: Our approach (in red) and Phylosift (in blue). Phylosift consistently displays two separated groups of placements, separated by a small gap. This gap relates to the use of highly conserved markers, which distinguishes perfect matches to sequences originally included in the profile construction and distantly related matches (and therefore internal branches), leading to the granularity observed in the different figures.

Fig. S12. *Acinetobacter* sp. A47. Distribution of the phylogenetic distances between the placement of a simulated fragment by the EPA, and the true correct position according to the core-genome phylogenetic reconstruction. The distances were divided by the maximal tip-to-root distance in the tree and are presented as percentages. The different colours represent the two methodologies used: Our approach (in red) and Phylosift (in blue). Phylosift consistently displays two separated groups of placements, separated by a small gap. This gap relates to the use of highly conserved markers, which distinguishes perfect matches to sequences originally included in the profile construction and distantly related matches (and therefore internal branches), leading to the granularity observed in the different figures.

Fig. S13. Relative abundance of fragments assigned to *Acinetobacter* by EPA (*Y* axis). Each bar represents an environmental category. Distribution of fragments was normalized by the sum of all normalized frequencies. Colours were selected according to the ones assigned in **Fig.** 1. A new version with the figure with the *Y*-axis recalculated to show the total number of sequence per environment, divided by total number of sequences per environment has been included (Supporting Information **Figure S17**).

Fig. S14. Scatterplot of the Bray–Curtis dissimilarity (*Y* axis) between the different terminal branches and their phylogenetic distance (*X*-axis) between all clades.

Fig. S15. Scatterplot of the relative abundance of *Acinetobacter* fragments from skin samples (*Y*-axis) and household-associated samples (*X*-axis) from the Home Microbiome Project. Samples were paired according to the origin of isolation and relationship between each house and their tenants. Given the large distance between the two highest points in the scatterplot and the rest, we have reanalysed the correlation after removing those two points, resulting in a still significant correlation (adjusted rho = 0.973, *P*-value = 1e-08) and no significant difference between the two slopes in the linear regression.

Fig. S16. Phylogenetic reconstruction of the *Acinetobacter* spp. Core genome. The large monophyletic clades highlighted by the three colours correspond to the three environmentally coherent clades represented in **Fig.** 1. The scale of the tree is given in substitutions per site. Only Bootstraps supports below 90% are represented in the corresponding node.

Fig. S17. Relative abundance of fragments assigned to *Acinetobacter* by EPA (*Y* axis). Each bar represents an environmental category. Distribution of fragments was normalized by the sum of all normalized frequencies and the total number of reads per environment. Colours were selected according to the ones assigned in Fig. 1.

Fig. S18. Results of the evolutionary placement analysis. We computed for each branch the distribution of the environmental categories associated with the fragments placed in the branch. The colour boxes indicate the branches in which the representation of fragments from certain environments was significantly higher than the average abundance for each environment across the tree (one-way Kruskal–Wallis test, P-value < 0.001). White boxes represent branches without any significant overrepresentation.

Table S1. List of *Acinetobacter* strains and genomes used in this study

Table S2. List of complete genomes used in the analysis.

Table S3. List of new *Acinetobacter* isolates used to validate our approach.

Table S4. Metagenomic metadata recruited from MG-RAST. All the meta information from each sample was processed and catalogued. The environmental classification was built using a hierarchical classification, from broader to more specific environment type.

Table S5. Percentage of True Positives, False negatives, False positives and True negatives resulted by SOM analysis, in the core and pan genome from *Acinetobacter*, close and distant outgroups. All values have been divided by the total number of events.

Table S6. EggNOG Functional annotation of the differentially enriched protein families in the three environmentally independent clades.

Table S7. List of Core Genome profiles, their presence in both the focal group and the close outgroup and whether they were finally used or not.