MICROBIAL GENOMICS

RESEARCH ARTICLE

Mangas *et al.*, *Microbial Genomics* 2019;5

DOI 10.1099/mgen.0.000309

Pangenome of *Acinetobacter baumannii* uncovers two groups of genomes, one of them with genes involved in CRISPR/Cas defence systems associated with the absence of plasmids and exclusive genes for biofilm formation

Eugenio L. Mangas[1], Alejandro Rubio[1], Rocío Álvarez-Marín[2], Gema Labrador-Herrera[2], Jerónimo Pachón[2,3], María Eugenia Pachón-Ibáñez[2,3], Federico Divina[4] and Antonio J. Pérez-Pulido[1,*]

Abstract

Acinetobacter baumannii is an opportunistic bacterium that causes hospital-acquired infections with a high mortality and morbidity, since there are strains resistant to virtually any kind of antibiotic. The chase to find novel strategies to fight against this microbe can be favoured by knowledge of the complete catalogue of genes of the species, and their relationship with the specific characteristics of different isolates. In this work, we performed a genomics analysis of almost 2500 strains. Two different groups of genomes were found based on the number of shared genes. One of these groups rarely has plasmids, and bears clustered regularly interspaced short palindromic repeat (CRISPR) sequences, in addition to CRISPR-associated genes (*cas* genes) or restriction-modification system genes. This fact strongly supports the lack of plasmids. Furthermore, the scarce plasmids in this group also bear CRISPR sequences, and specifically contain genes involved in prokaryotic toxin–antitoxin systems that could either act as the still little known CRISPR type IV system or be the precursors of other novel CRISPR/Cas systems. In addition, a limited set of strains present a new *cas9-like* gene, which may complement the other *cas* genes in inhibiting the entrance of new plasmids into the bacteria. Finally, this group has exclusive genes involved in biofilm formation, which would connect CRISPR systems to the biogenesis of these bacterial resistance structures.

DATA SUMMARY

All supporting data, code and protocols used during this study have been provided in the supplementary material (available with the online version of this article) and in the GitHub repository (https://github.com/upobioinfo/aba).

INTRODUCTION

Multidrug-resistant *Acinetobacter baumannii* causes hospital-acquired endemic and outbreak infections with high mortality and morbidity rates. Recently, the World Health Organization has classified carbapenem-resistant *A. baumannii* as priority one for the development of new antibiotics in the fight against the species [1]. This pathogen can colonize the human body or even persist in the hospital environment [2, 3], and opportunistically causes pneumonia, urinary tract infections and occasionally bacteraemia [4]. It has also been isolated from other animals [5], although some authors propose that these isolates are really other species of the same genus [6]. *A. baumannii* has a wide collection of genes, and many of them are known to contribute to its virulence, such as several outer-membrane proteins and siderophores [7, 8]. Some genes can be transferred by plasmids, which allows the bacteria to adapt to the environment [9]. The knowledge of the whole set of genes of a particular strain, in addition

Received 19 August 2019; Accepted 03 October 2019; Published 18 October 2019

Author affiliations: [1]Centro Andaluz de Biología del Desarrollo (CABD UPO-CSIC-JA), Facultad de Ciencias Experimentales (Área de Genética), Universidad Pablo de Olavide, 41013, Seville, Spain; [2]Institute of Biomedicine of Seville (IBiS), University Hospital Virgen del Rocío/CSIC/University of Seville, Seville, Spain; [3]Department of Medicine, University of Seville, Seville, Spain; [4]School of Engineering, Pablo de Olavide University, Ctra. Utrera s/n, Seville, Spain.

***Correspondence:** Antonio J. Pérez-Pulido, ajperez@upo.es

Keywords: *Acinetobacter baumannii*; CRISPR; bacterial genomics; toxin–antitoxin; plasmids; biofilm.

Abbreviations: ANI, average nucleotide identity; CRISPR, clustered regularly interspaced short palindromic repeat; GO, gene ontology; NCBI, National Center for Biotechnology Information.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Three supplementary tables and four supplementary figures are available with the online version of this article.

to the virulence information, could be important to predict the strain's expected pathogenicity. This is now possible due to the vast availability of genomics data stored in molecular databases [10].

Currently, thousands of sequences of complete genomes of *A. baumannii* are available from different genomics projects. Therefore, we can compare such genomes and associate genes with characteristics of the strains, e.g. the isolation source or the disease that they caused. The set of shared genes in this species has been calculated in numerous projects with a limited number of strains, and it amounts to around 2000 genes [2, 11–14]. But it is estimated that the number of genes in the core genome (genes shared by all the strains) is around 2200, with 20000 in the pangenome (genes coming from all the strains) [15, 16]. However, a study with thousands of genomes has never been carried out; the previous ones were usually performed with a restricted number of *A. baumannii* isolates, and might not reflect the complete scenario.

We studied a set of genes from almost 2500 genomes of *A. baumannii*, and discovered that a few of them could be erroneously classified in this species and others have a very low quality that does not justify their use in genomics studies. From the remaining genomes, we validated the previous estimations for core genome and pangenome, and found that a small group of strains share a lower number of genes due to the lack or low prevalence of plasmids. Remarkably, many strains in this group have a high number of clustered regularly interspaced short palindromic repeats (CRISPRs), although only half of strains have the genes of CRISPR/Cas systems, which are an essential prerequisite for the proper functioning of these systems. In *A. baumannii*, only the CRISPR type I–F system has been found so far, which is characterized by a cluster with the genes *cas1*, *cas2_3*, *cas8f*, *cas5*, *cas7* and *cas6f* [17]. However, we found one strain with a putative CRISPR type IV system, which has been mainly described in plasmids [18]. The group of strains with a lower number of plasmids also contains a high number of genes from toxin–antitoxin systems, which have been proposed to be precursors of modules of the CRISPR type IV system, but are still not very well known [18, 19]. Finally, another set of strains in this group has a homologue of the *cas9* gene, one of the most currently used genes in genetic engineering [20], with only the endonuclease domain conserved. All of this suggests there are new CRISPR elements in *A. baumannii*, with the potential for use in the future in the study of virulence relationships [21].

Impact Statement

Thousands of complete bacterial genomes are currently available in public databases. We have performed an *in silico* strategy for analysing the pangenome of almost 2500 strains of *Acinetobacter baumannii* and found that the genomes are divided into two groups with regard to the mean number of shared genes. The group sharing a lower number of genes rarely has plasmids, half of the strains from this group have genes from clustered regularly interspaced short palindromic repeat (CRISPR)/Cas systems, and the majority of strains always have CRISPR arrays. In addition, this group has specific genes from toxin–antitoxin systems, which have already been proposed as ancestral elements of CRISPR/Cas systems. All of this constitutes a demonstration that CRISPR/Cas systems prevent the acquisition of new genes in *A. baumannii*. Finally, this group of strains, despite the presence of a low number of shared genes, has specific genes involved in biofilms that would link CRISPR/Cas systems with this bacterial resistance structure, and could constitute new elements involved in both prokaryotic immunity and virulence.

METHODS

Genome collection

Assembled sequences of *A. baumannii* stored in the National Center for Biotechnology Information (NCBI) genome database as of January 2018 were collected, including 2467 genomes (98 completed and 2369 draft genomes). Metadata were also collected for each genome both from GenBank database and BioSample entries, which included, among others (Table S1), information on the assembly about the collection date, country, isolation source, host disease or host disease outcome [22]. The sequences of genomes of the other species from the genus *Acinetobacter* were also collected from the same source, including 490 additional genomes coming from 51 different species. Three of them were considered as part of the *Acinetobacter* ACB complex (*Acinetobacter pittii*, *Acinetobacter nosocomialis* and *Acinetobacter calcoaceticus*), where *A. baumannii* is usually grouped [23]. The remaining species were tagged as 'other'.

The metadata were filtered and only those with a high representativeness were used. We joined isolates from homogeneous sources to have the most numerous groups: blood (blood cultures), catheters, inert surfaces, non-human hosts (cat, chicken, dog, goose, parrot, stork and a plant), bone/joints (bone and joints samples), perianal (mainly perianal swabs), respiratory airways, skin and soft tissues, urinary tract, and others (the remaining infrequent sources). To create these general groups, we used both isolation source and host disease categories.

Genome and gene annotation

To homogenize the genome annotation, the sequences of all genomes were analysed using the same protocol. The protein-encoding genes were predicted using Prokka version 1.13 [24], and the predicted protein sequences were functionally annotated using Sma3s v2 and UniProt taxonomic division bacteria 2019_01 as a reference database [25, 55].

To annotate specific genes, GO (gene ontology) terms and UniProt keywords assigned by Sma3s were used. To find genes associated with CRISPR systems, protein sequences coming

from UniProt bacteria 2019_05 that had the word CRISPR (mainly with the GO term GO:0043571; maintenance of CRISPR repeat elements) were collected. Then, homologues to these proteins were searched for in all the *A. baumannii* strains, with a minimal identity threshold of 30 and 80% of query coverage. To complete the dataset of *cas* genes, CRISPR/Cas domains from the Conserved Domain Database (CDD) were searched for in the proteins of the pangenome. A total of 134 PSSM (position-specific scoring matrix) from the CDD [26] were searched using RPS-BLAST from the BLAST 2.2.31+ package [27], and 1×10^{-5} as the *E* value threshold. Finally, when a cluster of *cas* genes was found in a strain, and 1–3 unknown genes were inside the cluster, these new genes were also included.

Phylogenetic analysis

To create the 16S rRNA phylogenetic tree, the 16S rRNA gene sequences were searched by the software ssu-align v0.1.1 and Infernal v1.1 with default parameters [28]. Only genomes that showed a complete 16S rRNA gene (with the expected length of this gene in *Acinetobacter*) were taken into account, and when several 16S rRNA genes were found, the longest one was used. Then, the most probable 16S rRNA gene for each strain was used to do a multiple alignment using MAFFT v7.271 [29], the auto mode and default options, because of the high expected sequence similarity. Gap regions appearing in more than 10% of the sequences were removed using trimAL v1.2 [30]. The phylogeny was finally created with the alignment and the tool DNAPARS from the suite PHYLIP v3.697, which uses the parsimony method. A bootstrap of 1000 was calculated to evaluate the phylogeny confidence.

The core phylogeny was created with the amino acid sequences of genes present in the strains of all *Acinetobacter* genera (184 strains of non-*A. baumannii* genomes were removed due to low quality). The core proteins of *A. baumannii* were searched in the remaining species by the standalone version 2.2.31+ of BLAST [27], using the reference core sequences of *A. baumannii* as query sequences, and 70% both for identity and query coverage as thresholds. This resulted in 14 sequences, mainly annotated in metabolism, translation or ribosome biosynthesis: *rplB*, *spot*, *rpmC*, *cgtA*, *hemF*, *aroK*, *proB*, *lolD*, *gmk*, *lysA*, *purM*, *recG*, *truB* and *rpoA*. The sequences were joined and aligned by MAFFT v7.271 using the L-INS-I option. Gap regions appearing in more than 10% of the sequences were also removed using trimAL v1.2. The phylogeny was reconstructed by RAxML v8.2.9 with the model PROTGAMMAWAG and bootstrap of 1000 [31]. Both trees were viewed by R ggtree v1.10.5.

To calculate the average nucleotide identity (ANI), pyani v0.2.4 with method ANIb and BLASTN was used [32]. All strains were compared against the reference strain ATCC 17978, and the ANI percentages were collected.

Core genome and pangenome assessment

To assess the core genome and pangenome, Roary version 3.11.2 was used with an identity threshold of 90% and the -s parameter for not separating paralogs at this identity threshold [33]. This process creates groups of genes that assume the same gene coming from different strains, and each group is represented by a reference sequence. Only protein-encoding genes were considered. To be more exhaustive and to try to discover all the corresponding genes in all the strains, the reference genes were functionally annotated by Sma3s, and proteins with the same gene name were collapsed. In this way, we have a high confidence in the presence/absence of every gene in the pangenome. Finally, when a group contains genes from all the selected strains it is considered to be a core gene (we considered that a gene belonged to the core genome when it appeared in 99% of the strains), and the remaining genes are considered as accessory genes. The annotations of genes with the same gene name were combined to improve the functional information for these genes, since many of them are fragments coming from the same gene that has repeats that are hard to assemble.

The list of presence/absence of genes was used to reconstruct a phylogenetic profile that groups the strains by the co-occurrence of shared genes. The tree was created by Roary and FastTree 2.1.8 [34], and the results were viewed by the Phandango web application [35] and exported in SVG format.

Functional enrichment analysis

To discover the functional enrichment of a group of genes, such as the core genome or the different groups from the accessory genome, TopGO R package version 2.30.1 was used [36], which uses GO terms from a specific ontology (biological process was selected). The GO terms used were those annotated by Sma3s. The figures were created using the ggplot2 R library in a customized script.

UniProt keywords were included as an additional annotation source to enrich the two groups of strains sharing a different number of genes. In this case, the enrichment protocol of the module 3 from Sma3s was used [37], which is based on the hypergeometric distribution. We analysed all functional annotations from genes present in at least in 30% of the strains of a group and in less than 5% of the other group. Annotation terms with a *P* value higher than 0.01 were selected and terms appearing in only one gene were not considered. The two groups were chosen by the minimum point in the distribution of the mean shared genes.

CRISPR array discovery

To collect the plasmids bearing CRISPR arrays, the term 'CRISPR Arrays' was searched for in the COMMENT field of their GenBank entry, and the plasmid was annotated as containing CRISPR-repeats when it appeared. To discover CRISPR repeats in the complete genomes, CRISPRCas-Finder 1.4 was used with default parameters [38]. The number of CRISPR arrays with an evidence-level ≥1 in each genome was counted and assigned to the genome.

Plasmid discovery

When a genome sequence is closed, the plasmids coming from this genome are known and can be used; but when the genome sequence is a draft, it is necessary to predict the plasmids. To do that, plasmids in the NCBI RefSeq nucleotide database were searched by the term: 'Acinetobacter baumannii[Organism]', taking only records from the genetic compartment 'Plasmid' that had at least one of the two terms 'complete sequence' or 'complete genome' in their Description field. Thus, partial sequences and those lower than 1 kb in length were discarded. In this way, 638 plasmids were downloaded. Then, homologues to these plasmids in the *A. baumannii* strains were searched by BLASTN with both an identity and plasmid coverage threshold of 95%. Only sequences from the strains with a minimal coverage of 95% were considered to be more restricted. The candidate plasmids in the different strains were later clustered by psi-CD-HIT, an implementation of CD-HIT for long sequences [39], with an identity threshold of 90% to remove redundant plasmids, and the remaining sequences were clustered by mob_cluster from the Mob-Suite toolkit v1.4.9.1 [40], which uses default parameter to cluster plasmid sequences. Finally, all the clustered plasmids were considered the same plasmid in the different strains. The final number of non-redundant plasmids was 84, although only 48 of them were found in the analysed genomes.

RESULTS

Genome collection and phylogenetic analysis

The first step in the analysis of the set of genes present in *A. baumannii* was collecting 2467 available genome sequences from public databases, together with descriptive information, including their isolation source. To estimate the quality of the genomes and measure the evolutionary relationship among them, a 16S rRNA-based phylogeny was reconstructed. This included the 16S rRNA gene sequences from all the strains, although for 328 of them no complete 16S rRNA gene was found and, thus, we discarded these strains from later analysis. The phylogeny also included the 16S rRNA sequences of 490 strains from other species of the genus *Acinetobacter*, which allowed us to assess the validity of the *A. baumannii* taxonomic assignments (Fig. 1). The result showed that 14 strains could belong to species other than *A. baumannii* from the same genus. As a consequence, only 2125 strains were finally kept for further analysis (Table 1). The highest number of discarded genomes belonged to the group of blood and perianal isolation sources. It suggests that some strains causing bloodstream infections or collected from surveillance studies in the perianal region either are wrongly classified as *A. baumannii* or have incomplete genomes. It is also remarkable that none of the non-human host strains, coming from

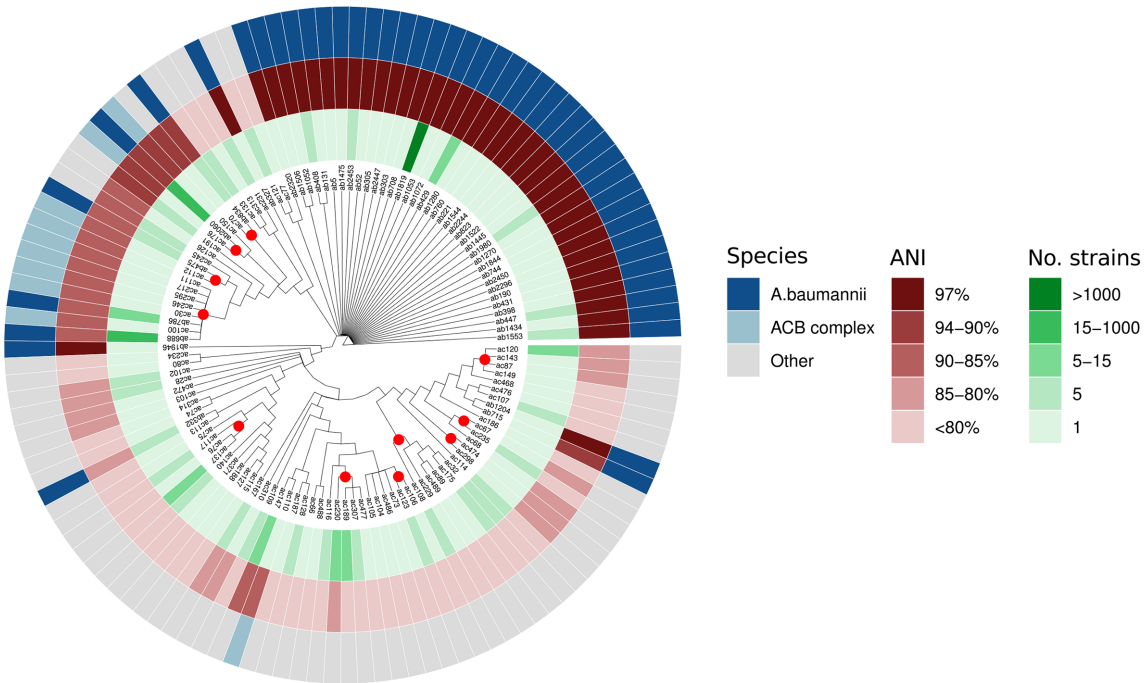

Fig. 1. 16S rRNA-based phylogenetic tree. The three rings around the tree represent: the species (2139 from *A. baumannii*, 166 from the ACB complex, 3 similar related *Acinetobacter* species that are usually grouped in this complex together with *A. baumannii* and 140 from other species of the genus); the percentage similarity with the reference *A. baumannii* ATCC 17978 measured as the ANI; and the number of strains in the terminal node (strains with exactly the same 16S rRNA sequence were collapsed in one node). A bootstrap of 1000 repeats was used, and nodes with a bootstrap lower than 70% are marked with a red circle . A total of 328 strains from *A. baumannii* and 184 from the other species were not used in this phylogeny due to the absence of a complete 16S rRNA gene. Only strains of the *A. baumannii* clade in the upper-left corner were retained in the study.

Table 1. Number of genomes from the different isolation sources

Frequencies are shown before and after the 16S rRNA analysis. The last column shows the frequencies after the posterior low-quality test, where 13 additional strains were discarded. Some sources have been homogenized, and the 'others' class includes both low frequency and non-reported sources.

	Number of strains		
Isolation source classes	**Initially**	**After 16S rRNA**	**Finally**
Respiratory airways	814	777	773
Blood	285	166	163
Skin and soft tissues	155	139	139
Urinary tract	130	126	126
Perianal	151	83	83
Inert surfaces	35	21	21
Non-human hosts	17	17	17
Catheters	15	15	14
Bone/joints	16	9	9
Others	849	772	767
Total strains	2467	2125	2112

animals (mainly birds) and plants (Table S1, available with the online version of this article), were discarded after phylogenetic analysis.

Average number of shared genes among strains

Protein-encoding genes were predicted, and orthologous genes were clustered using a restricted strategy. Two genes from different strains were considered orthologues when they had the same predicted gene name or shared ≥90%amino acid identity. The core genome and pangenome were found to be 2221 and 19272 genes, respectively (Table S2). Functional annotation of the pangenome showed that 42% of the genes were not assigned any function. However, only 3% of such genes were in the core genome (Table S3). The different genomes show a similar number of genes, with a mean around 3600 (Fig. 2a). However, the average number of shared genes by strain shows a bimodal distribution with two mean values, where the second group of strains shares about 150 more genes (Fig. 2b). Both distributions show outliers with extremely low values, which would represent genomes of low quality, and they were discarded from later analysis. In fact, when correlating the two distributions, extreme values suggest both low-quality genomes and misclassification of the species name, sometimes matching to strains previously discarded after the 16S rRNA-based phylogeny (Fig. 2c). So, strains removed because they previously brought other *Acinetobacter* species into the phylogeny showed a number of shared genes under that expected based on their total number of genes, and many of the strains not showing the 16S rRNA gene have a low total number of genes, suggesting a low quality of the genome assembly. It is notable that three strains isolated from blood cultures were discarded at this stage (Table 1).

When analysing the total number of genes and the number of shared genes, another meaningful fact is that the non-human host strain correlated with the majority of other *A. baumannii* strains, highlighting again the isolation of this species in animals. Alternatively, strains isolated from inert surfaces, mainly hospital equipment and furniture, showed a low number of shared genes, or did not show 16S rRNA genes, and some of them were removed from the final dataset.

Functional enrichment of the core genome and pangenome

Once the most contradictory genomes were discarded, the total number of *A. baumannii* strains was 2112. These strains have more than 19000 different genes, though the majority, almost 16000 genes, are present only in less than 20% of the strains. The core genome is highly enriched in genes involved in metabolism, biosynthesis and protein translation functions (Fig. 3). Specifically, more than 450 genes are involved in organonitrogen compound metabolic processes, and almost 400 in small molecule metabolic processeses. However, the accessory genome (genes present in only some strains) is enriched mainly in regulation functions, with more than 90 genes involved in transcription processes. Expression regulation is mainly highlighted in genes present in more than 20% of strains and supports the previous idea that differences between *A. baumannii* isolates are strongly supported by a heterogeneous gene expression of the core genes [15, 41]. A remarkable function that appears in the limited accessory genome is 'maintenance of DNA repeat elements', which is mainly related to CRISPR/Cas systems. Finally, genes only appearing in 1% of strains (21 or less) are related to both DNA modification and integration, characteristics of mobile sequences as transposons or sequences with phage origin. Although, it is noteworthy that the latter group also highlights the function cell motility, which would suggest motile strains in an essentially non-motile genus [2].

The core genome is an important reference in the analysis of phylogenetic relationships between strains. Thus, a new phylogeny was reconstructed with 14 genes present in the entire genus *Acinetobacter*. The tree obtained supports the relationships already found with the 16S rRNA-based phylogeny (Fig. S2). The tree brings together neither strains coming from the same source nor the two groups of strains sharing a different mean number of shared genes, although a small number of the strains sharing a low number of genes appear near to the node of other *Acinetobacter* ssp. However, the analysis of gene presence/absence from the pangenome suggests that both groups have specific genes, which rarely appear in the contrary group (Fig. 4). So, from now on, we name 'group 1' as the group sharing a lower number of genes, and 'group 2' as the group sharing a higher number of them.

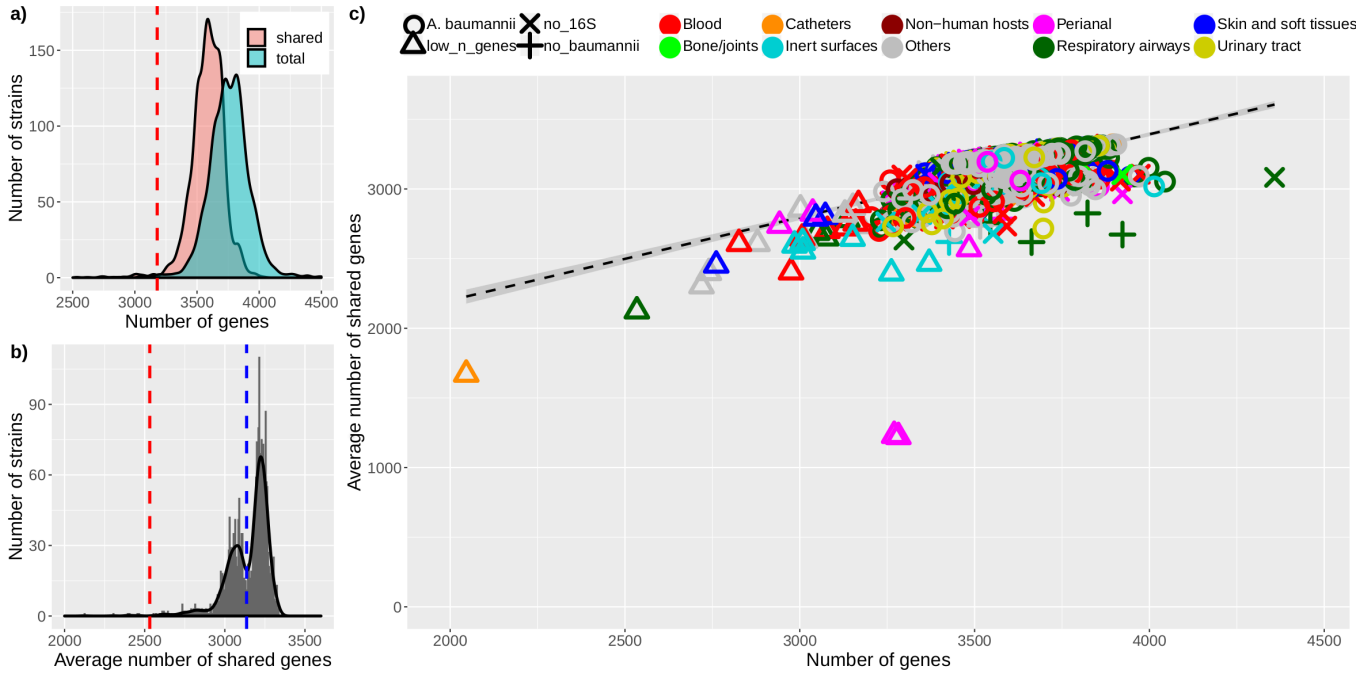

Fig. 2. Distribution of the number of genes and mean number of shared genes between strains. (a) Distribution of the number of genes for all the *A. baumannii* strains: total (total number of genes) and shared (genes shared with at least another different strain). The red dashed line marks the threshold below which strains were removed due to low quality (at 2.5 SD). (b) Distribution of the average number of shared genes for each strain against the other strains. The red dashed line marks the threshold below which strains were removed due to low quality (at 2.5 SD of the first peak), and the blue line marks the minimum in the distribution that highlights the separation between the two clearly independent groups of strains. The Hartigans' dip test for multimodality suggests non-unimodality with *P* value=2.2×10^{-16}. (c) Scatter plot comparing the two previous distributions. Every point represents a strain, the colour shows the isolation source of this strain. The shape shows whether the strain was discarded due to low quality (triangle), the 16S rRNA gene was not found (x), the 16S rRNA phylogeny suggested that it was not *A. baumannii* (+) or the strain was retained as *A. baumannii* in the analysis (circle). An updated figure that only shows the strains remaining in the analysis can be reviewed in Fig. S1.

Strains with CRISPR systems usually lack plasmids

Groups 1 and 2 have 717 and 1395 strains, respectively. In addition, there are genes specifically found in each group, which seem to appear uniquely in the respective group (Fig. 4). Remarkably, genes in group 2 were enriched in plasmid annotation and also in exonuclease activity (Fig. 5), which could help plasmid maintenance [42]. Altogether, this result suggests that the higher number of shared genes in these strains could be due to the presence of plasmids bearing genes, which are not present in the other group. In fact, group 1 was mainly enriched in genes involved in maintenance of CRISPR repeat elements, which could, hence, be involved in avoiding plasmid entry into the bacteria. In addition, many genes in group 1 have signal peptides and encode membrane lipoproteins. Remarkably, four of these genes encode proteins bearing the spore coat protein U domain (*P23_0613*, *XAC1424*) or fimbrial biogenesis outer membrane usher proteins (*yehB/FimD*, *J532_1634/FimC*). All of these genes could be involved in biofilm formation; therefore, they might be relevant in both surviving and virulence [43, 44]. They appear in more than 70% of strains of group 1 and less than 2% of group 2, what suggests its relevance in the first group.

To confirm the inverse relationship between plasmids and CRISPR genes, both plasmids and CRISPR-associated genes (*cas*), in addition to CRISPR arrays, were searched for in the genomes of the different groups. As expected, the number of plasmids in group 1 was lower than in group 2, and group 1 presented a much higher number of both *cas* genes and CRISPR arrays (Table 2, Fig. S3).

Only a small number of strains in group 1 have a high number of plasmids. But in these strains, plasmids bear CRISPR arrays, different to the case of group 2 (Fig. 6a). CRISPR arrays are usually in the bacteria chromosome, though they can also appear in plasmids [18]. In fact, 12 strains in group 1 have plasmids with annotated CRISPR repeats. To measure the abundance of CRISPR systems in group 1 with regard to group 2, CRISPR arrays were predicted in the complete *A. baumannii* genomes. Nearly all group 1 strains are characterized by a high number of CRISPR arrays, and half of them bear *cas* genes. However, only three-quarters of strains belonging to group 2 have CRISPR arrays. They appear in a much lower number, and do not match with *cas* genes in the same strains (Table 2, Fig. 6b). All of this suggests that to avoid plasmid entry, the bacterium needs to have not only *cas* genes

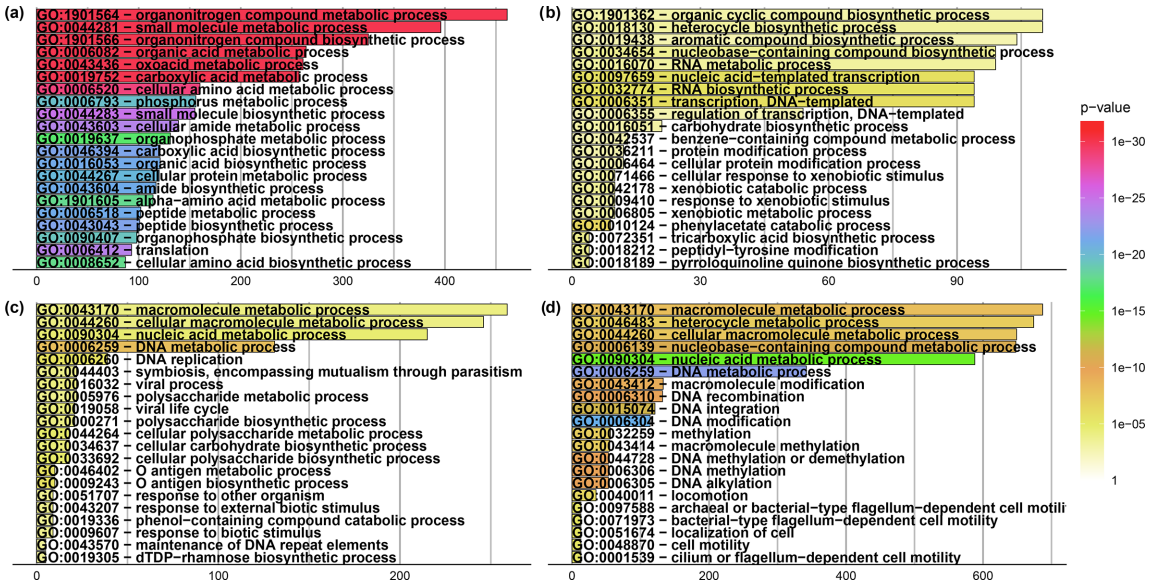

Fig. 3. Functional enrichment of the different pangenome groups. (a) Core genome (>99% of strains; 2221 genes). (b) Accessory genome of frequent genes (20–99% of strains; 1341 genes). (c) Accessory genome of limited genes (1–20% of strains; 3074 genes). (d) Accessory genome of rare genes including singletons (<1% of strains; 12635 genes). Biological process ontology from GO was used. The *x*-axes show the number of genes with a GO term, and the colour shows the *P* value of the GO term. Note that the most significant terms appear in the core genome (coloured in red).

but also CRISPR arrays. In addition, strains in group 1 with CRISPR arrays but without *cas* genes might suggest either they previously had *cas* genes or that they have unknown CRISPR systems or even that they could bear another bacterial immunity system. To test this hypothesis, genes in the strains of group 1 that had no *cas* genes were analysed. From this study, 11 genes emerged, which were present in 82% of the strains of group 1 and only in 3% of group 2 strains. More specifically, the genes were *hsdS*, *mshD*, *yjhX*, *pepO*, *moeA_2*, *kstR* and five unknown genes (annotated as BIT33_04875,

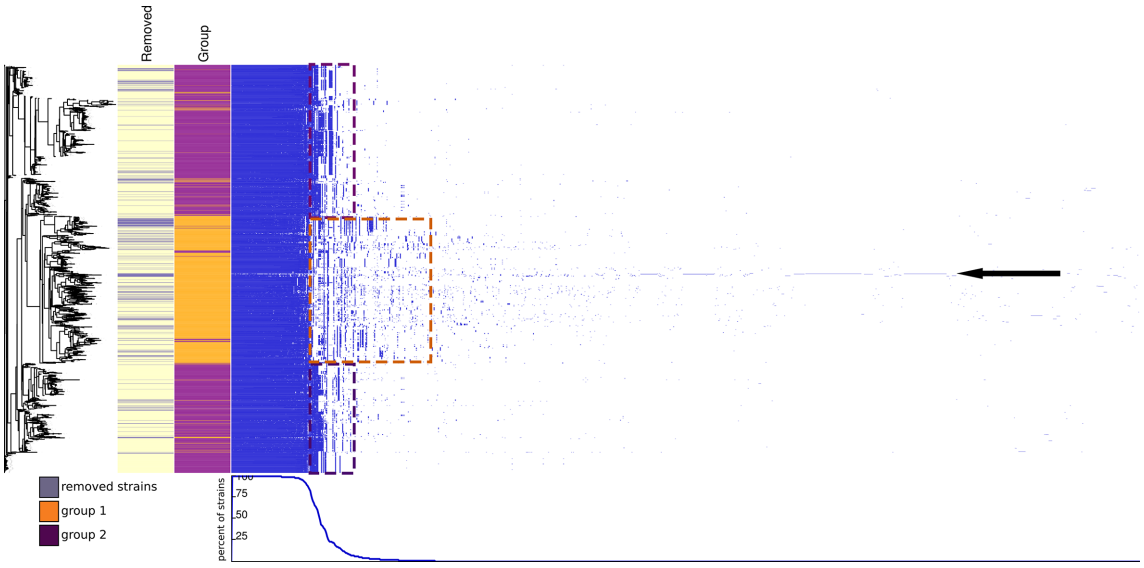

Fig. 4. Phylogenetic profile with a presence/absence matrix of all genes in the pangenome. The presence of a gene is indicated by the colour blue, and at the bottom of the figure the percentage of strains that have each gene is shown. The columns show: in mauve, the strains removed in previous steps; in orange, the strains belonging to group 1; and in purple, the strains belong to group 2. Note that the members of the groups 1 and 2 share specific genes with their group that do not appear in the other group, which are highlighted with dashed boxes. The arrow marks a strain with many specific genes (mostly fragments) that is in the group of discarded strains.

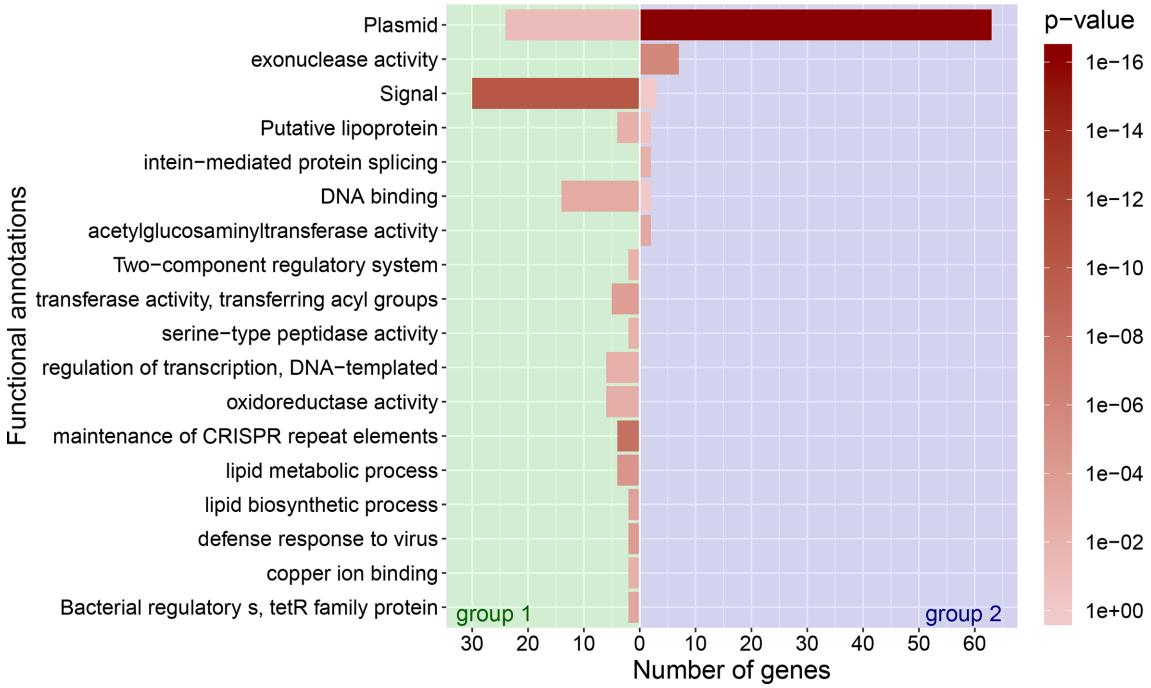

Fig. 5. Functional enrichment of the strains from groups 1 and 2. GO terms (lowercase), and UniProt keywords and descriptions (uppercase) were used as the annotation sources. The *x*-axis shows the number of genes with a term, and the colour shows the *P* value of the term. Note that the most significant terms are *signal* and *maintenance of CRISPR repeat element* in group 1, and *plasmid* and *exonuclease activity* in group 2.

NT90_15810, BAV2244, ACIAD1781 and unknown126). Some of these genes are hypothetical proteins, but others, such as *hdsS*, are involved in a type I restriction-modification system, and together with the peptidase *pepO*, and the transcriptional regulator *kstR* from the TetR family involved in antibiotic resistance, could help in blocking plasmid entrance.

One of the *cas* genes found in the strains of group 1 was a *cas9-like* gene that encodes the endonuclease HNH domain of the Cas9 protein, but not the other expected domains such as the RuvC domain and the PAM-interacting domain, both necessary for the proper functioning of this protein. This *cas9-like* gene usually appears adjacent to several other genes (*tehB*, *rhpA*, *hsdM*, *vsr*, *hepA* and uncharacterized genes named as *unknown3592*, *unknown2528*, *unknown3372*, *unknown2565*, *unknown2538*, *p3ABAYE0073* and *BUC_0564*). These genes have functions putatively related to CRISPR systems such as helicase (*rhpA* and the protein containing a DEAD/DEAH-box helicase domain, *hepA*). For example, *p3ABAYE0073* is a WYL domain-containing protein, and CRISPR systems have been predicted to be transcriptionally regulated by multiple ligand-binding proteins containing this kind of domain, which would sense modified nucleotides during foreign DNA entrance [45]. Strains bearing the *cas9-like* gene not only belong to the group 1 but also have a very low number of plasmids and a high number of CRISPR arrays, with a mean of 3.47±2.70 arrays per strain, which only drops to 2.10±1.74 when considering strains with *cas9-like* as the only *cas* gene (Fig. 6c).

Table 2. Number of plasmids and CRISPR genes in both group 1 and group 2

The number of plasmids, *cas* genes and CRISPR arrays is shown, and also the proportion of strains with at least one of them in the group (in brackets). The remaining strains bearing *cas* genes have different variants of the type I–F.

Group	No. of strains	Predicted plasmids	*cas* genes	CRISPR arrays	Type I–F	Cas9-like
Group 1	717	264 (25%)	1553 (45%)	2628 (95%)	121 (16%)	61 (8%)
Group 2	1395	2091 (73%)	110 (2%)	1348 (73%)	14 (1%)	1 (0%)

Another remarkable result was that plasmids in group 1 are especially enriched in toxin–antitoxin systems (Fig. S4). In particular, 9 genes involved in these systems and present in plasmids appear in 182 strains from group 1, and only 27 from group 2: *yoeB*, *vapC*, *yafQ_1*, *relB*, *dinJ*, *brnA*, *txe*, *yefM* and *brnT*. Many of these systems have already been found in *A. baumannii* [46]. They are composed of two proteins coming from the same operon, where the first gene encodes an antitoxin that neutralizes a regulatory toxin encoded by the second gene. In the analysed genomes, these genes seem to be more frequent in strains isolated from blood cultures, but not in those from the perianal region. A gene coming

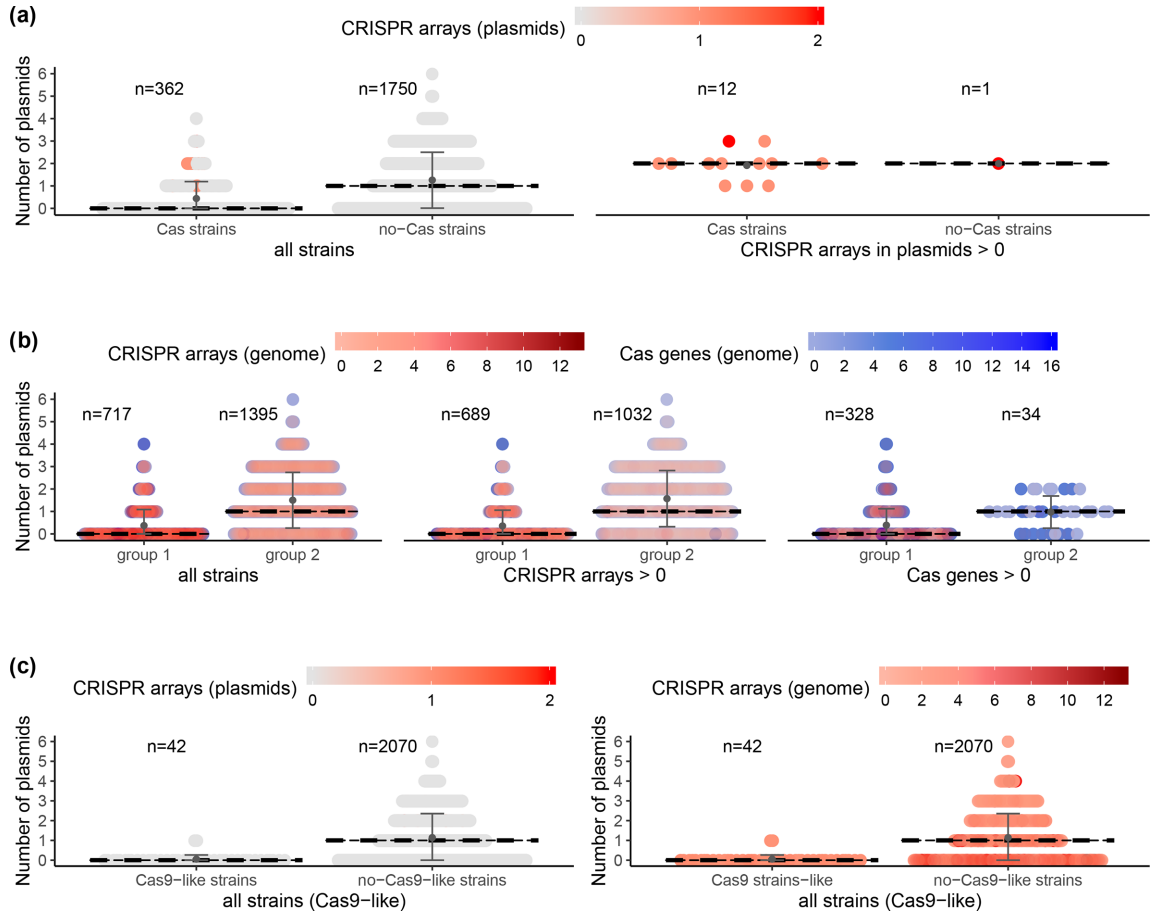

Fig. 6. Distribution of the number of plasmids between different groups together with the number of *cas* genes and CRISPR arrays. Each point represents one strain, and the number of plasmids for each strain is shown on the *y*-axes. The colour of the point represents the number of CRISPR arrays (red colour) or *cas* genes (blue colour). The dashed line shows the median of the plasmid number distribution, and the mean (dark grey dot) and SD (error bars) are also shown. Finally, the number of strains is shown for each group (*n*). (a) Strains carrying *cas* genes (Cas strains) versus strains without any *cas* gene (no-Cas strains), followed by the same two groups but now only showing strains with at least one plasmid carrying CRISPR arrays. (b) Strains of group 1 versus group 2, followed by the same two groups but now only showing strains with at least one CRISPR array in its genome. Note that the overlapping of CRISPR arrays and *cas* genes makes a purple colour. (c) Strains with *cas9-like* as the only *cas* gene versus the remaining strains.

from these systems, but in this case with a high frequency in both groups, is *relE*. This gene is a ribosome-dependent mRNA endoribonuclease that inhibits translation during amino acid starvation [47], and it is present in more than 70% of infective strains (isolated from blood cultures and respiratory airways), but in less than 50% of the remaining isolation sources. Microbial toxin–antitoxin modules seem to have been important contributors to the evolution of CRISPR/Cas systems [19]. For instance, Cas2 is one of the key proteins in the first phase of CRISPR systems, and it derives from the toxins of the VapD family of mRNA interferases [48]. In fact, the most common unknown toxin mechanisms are those that indiscriminately cleave mRNAs inside the ribosome, resulting in microbial dormancy or cell death [49]. The interferases belong to several unrelated protein families including HEPN, RelE and VapD [50]. Specifically, four proteins have the HEPN domain in our described pangenome. All of these are uncharacterized proteins predicted to be in plasmids, and again appearing in group 1 (51 versus 3 in group 2; see Table S3). Furthermore, one of the analysed strains has a cluster with genes *csf3*, *csf4* (also known as *dinG*) and *cas6e*, together with CRISPR arrays at both ends, which could constitute a CRISPR type IV variant.

DISCUSSION

The current genomics era is enabling the achievement of large-scale genomics analysis. In this paper, we present what is believed to be the largest analysis of *A. baumannii* genomes conducted to date. The proposed core genome and pangenome agree with other smaller analyses [11, 15], and specially with previous forecasts expecting an exponential pangenome growth when a great number of genomes were

available [16]. Although several strains were discarded due to poor quality or putative misclassification, none of those came from non-human hosts. This fact suggests that strains from non-human hosts may be true *A. baumannii*, despite several authors suggesting that this species is not present in any other host but humans [6].

The analysis of shared genes between strains gave a bimodal distribution. This kind of distribution used to be related to a population skew or a technical mistake. For example, Gweon *et al.* in 2017 proved that bacterial genome sizes have a bimodal distribution due to a high sequencing effort towards a certain group of species [51]. However, we found two significant *A. baumannii* groups, which share specific genes within each group that are not present in the other group. In fact, one group has a high number of plasmids (and their genes) and the other has CRISPR/Cas genes, which could block plasmid entry. This is something already proposed and expected in CRISPR studies [19], but the presence of CRISPR/Cas does not seem to be enough to repel the plasmids out of the bacterium. However, half of the strains in group 1 specifically have genes that are involved in other bacterial immunity systems, such as the type I restriction-modification system. Type I restriction endonucleases are components of prokaryotic DNA restriction-modification mechanisms that protect the organism against invading foreign DNA [52]; therefore, they could complement the CRISPR/Cas systems. Another defence system found in several prokaryotes is the DND system, which labels DNA by phosphothiolation and destroys unmodified DNA [53]. Remarkably, only three of the analysed strains present two genes from this system, namely *dndD* and *dndC*, and they belong to group 1.

However, group 1 also has plasmids, but they seem to have CRISPR array sequences, which contain guide RNA, essential for the correct functioning of this system. Although these repeated sequences are usual in the bacterial chromosome, they can also be transported in mobile sequences. Species of the family *Vibrionaceae* are an example of this, where CRISPR systems identified were present predominantly within mobile genetic elements, including transposon-like elements and plasmids [54]. However, it would also suggest the presence of still unknown CRISPR/Cas systems, in addition to other complementary immunity systems. These could be related to toxin–antitoxin systems, which are abundant in prokaryotes [47]. The toxin–antitoxin systems are involved in restricting the growth of competing bacteria, the response to starvation and other stresses and, therefore, have been linked to virulence [55]. In *Clostridium difficile*, the systems have been recently related to a new functional antisense RNA system, which is part of a type I toxin–antitoxin system adjacent to CRISPR arrays [56]. In fact, microbial toxin–antitoxin modules seem to be important contributors to the evolution of CRISPR/Cas systems, especially in the type IV system, which usually appear in plasmid sequences [19]. This observation could explain why a great number of the *A. baumannii* strains are characterized by a great number of CRISPR arrays in both their chromosome and plasmids, but only half of them have *cas* genes. In addition, the presence of what is believed to be the first described CRISPR/Cas type IV system in *A. baumannii*, and the abundance of toxin–antitoxin systems found in the strains of the group with a lower number of plasmids in this study, would support the idea that this latter system could contribute to the appearance of defensive CRISPR/Cas systems. It is important to note that we found a new *cas9-like* gene in the *A. baumannii* strains of the group that seemed to block plasmid entrance. Although it seems to conserve only the central endonuclease domain, the adjacent genes that it has, as well as the number of CRISPR arrays present in these strains, and the low number of plasmids that they possess, would suggest that it could be part of a still unknown functional CRISPR system.

Finally, we found genes involved in biofilm formation that appeared almost exclusively in the group enriched in CRISPR systems. The loss of function of proteins involved in CRISPR systems, such as the endonuclease Cas3, seems to affect to biofilm formation [57, 58], then CRISPR systems could help in the survival of the bacterium on inert surfaces, which is meaningful for the resistance of *A. baumannii* in hospitals [2].

In summary, we showed that strains of *A. baumannii* are divided into two groups that share a different number of genes, which could be maintained by CRISPR/Cas systems, some of them still not described. The group with these defence systems seems to have specific biofilm genes, and would avoid the entrance of plasmids and probably foreign genes, including resistance elements. These findings could be useful insights to help in the fight against this bacterium.

Funding information
The project has been supported by Plan Nacional de I+D+i 2013-2016 and Instituto de Salud Carlos III, Subdirección General de Redes y Centros de Investigación Cooperativa, Ministerio de Ciencia, Innovación y Universidades, Spanish Network for Research in Infectious Diseases (REIPI RD16/0016/0009), and co- financed by European Development Regional Fund 'A way to achieve Europe', Operative program Intelligent Growth 2014-2020. E.L.M. is supported by 'Programa de Empleo Joven' (FEDER/Junta de Andalucía, Fase I, II convocatoria), R.A.-M. by a Juan Rodes grant (JR17/00025), and G.L.-H. by a i-PFIS grant (IFI15/00128) from Instituto de Salud Carlos III, Subdirección General de Redes y Centros de Investigación Cooperativa, Ministerio de Ciencia, Innovación y Universidades, Spain.

Acknowledgements
We would like to thank C3UPO for the HPC support.

Author contributions
E.L.M.: performed the majority of analyses and created ***in silico*** protocols and scripts. A.R.: performed the phylogenies and helped in the development of different protocols. R.A.-M. (0000-0002-7101-2514): provided clinical knowledge of the bacterium and helped in the initial classification and result discussion. G.L.-H.: provided clinical knowledge of the bacterium and helped in the initial classification and result discussion. J.P. (0000-0002-8166-5308): provided clinical knowledge of the bacterium and helped in the initial classification and result discussion. M.E.P.-I. (0000-0001-7969-8162): provided clinical knowledge of the bacterium and helped in the initial classification and result discussion. F.D. (0000-0002-0964-9506): supported the training of E.L.M. and helped in protocol design. A.J.P.-P. (0000-0003-3343-2822): conceived the study, carried out the design and coordination, performed some analysis, and wrote the manuscript. All authors read and approved the final manuscript.

Conflicts of interest
The authors declare that there are no conflicts interest.

Data bibliography
The genome assemblies of all the genomes in this project were downloaded from the NCBI genome database, and identifiers are listed in Table S1.

References

1. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M *et al*. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis* 2018;18:318–327.
2. Peleg AY, Seifert H, Paterson DL. *Acinetobacter baumannii*: emergence of a successful pathogen. *Clin Microbiol Rev* 2008;21:538–582.
3. Harding CM, Hennon SW, Feldman MF. Uncovering the mechanisms of *Acinetobacter baumannii* virulence. *Nat Rev Microbiol* 2018;16:91–102.
4. Cisneros JM, Reyes MJ, Pachón J, Becerril B, Caballero FJ *et al*. Bacteremia due to *Acinetobacter baumannii*: epidemiology, clinical findings, and prognostic features. *Clin Infect Dis* 1996;22:1026–1032.
5. Eveillard M, Kempf M, Belmonte O, Pailhoriès H, Joly-Guillou M-L. Reservoirs of *Acinetobacter baumannii* outside the hospital and potential involvement in emerging human community-acquired infections. *Int J Infect Dis* 2013;17:e802–e805.
6. Towner KJ. Acinetobacter: an old friend, but a new enemy. *J Hosp Infect* 2009;73:355–363.
7. McConnell MJ, Actis L, Pachón J. *Acinetobacter baumannii*: human infections, factors contributing to pathogenesis and animal models. *FEMS Microbiol Rev* 2013;37:130–155.
8. Gaddy JA, Arivett BA, McConnell MJ, López-Rojas R, Pachón J *et al*. Role of acinetobactin-mediated iron acquisition functions in the interaction of *Acinetobacter baumannii* strain ATCC 19606T with human lung epithelial cells, *Galleria mellonella* caterpillars, and mice. *Infect Immun* 2012;80:1015–1024.
9. Salto IP, Torres Tejerizo G, Wibberg D, Pühler A, Schlüter A *et al*. Comparative genomic analysis of *Acinetobacter* spp. plasmids originating from clinical settings and environmental habitats. *Sci Rep* 2018;8:7783.
10. Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V *et al*. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res* 2016;44:D73–D80.
11. Adams MD, Goglin K, Molyneaux N, Hujer KM, Lavender H *et al*. Comparative genome sequence analysis of multidrug-resistant *Acinetobacter baumannii*. *J Bacteriol* 2008;190:8053–8064.
12. Farrugia DN, Elbourne LDH, Hassan KA, Eijkelkamp BA, Tetu SG *et al*. The complete genome and phenome of a community-acquired *Acinetobacter baumannii*. *PLoS One* 2013;8:e58628.
13. Chan AP, Sutton G, DePew J, Krishnakumar R, Choi Y *et al*. A novel method of consensus pan-chromosome assembly and large-scale comparative analysis reveal the highly flexible pan-genome of *Acinetobacter baumannii*. *Genome Biol* 2015;16:143.
14. Sahl JW, Johnson JK, Harris AD, Phillippy AM, Hsiao WW *et al*. Genomic comparison of multi-drug resistant invasive and colonizing *Acinetobacter baumannii* isolated from diverse human body sites reveals genomic plasticity. *BMC Genomics* 2011;12:291.
15. Imperi F, Antunes LCS, Blom J, Villa L, Iacono M *et al*. The genomics of *Acinetobacter baumannii*: insights into genome plasticity, antimicrobial resistance and pathogenicity. *IUBMB Life* 2011;63:1068–1074.
16. Antunes LCS, Visca P, Towner KJ. *Acinetobacter baumannii*: evolution of a global pathogen. *Pathog Dis* 2014;71:292–301.
17. Karah N, Samuelsen Ø, Zarrilli R, Sahl JW, Wai SN *et al*. CRISPR-Cas subtype I-Fb in *Acinetobacter baumannii*: evolution and utilization for strain subtyping. *PLoS One* 2015;10:e0118205.
18. Koonin EV, Makarova KS. Mobile genetic elements and evolution of CRISPR-Cas systems: all the way there and back. *Genome Biol Evol* 2017;9:2812–2825.
19. Koonin EV, Makarova KS. Origins and evolution of CRISPR-Cas systems. *Philos Trans R Soc Lond B Biol Sci* 2019;374:20180087.
20. Barrangou R, Doudna JA. Applications of CRISPR technologies in research and beyond. *Nat Biotechnol* 2016;34:933–941.
21. Louwen R, Staals RHJ, Endtz HP, van Baarlen P, van der Oost J. The role of CRISPR-Cas systems in virulence of pathogenic bacteria. *Microbiol Mol Biol Rev* 2014;78:74–88.
22. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E *et al*. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res* 2012;40:D57–D63.
23. Nemec A, Krizova L, Maixnerova M, van der Reijden TJK, Deschaght P *et al*. Genotypic and phenotypic characterization of the *Acinetobacter calcoaceticus-Acinetobacter baumannii* complex with the proposal of *Acinetobacter pittii* sp. nov. (formerly Acinetobacter genomic species 3) and *Acinetobacter nosocomialis* sp. nov. (formerly Acinetobacter genomic species 13TU). *Res Microbiol* 2011;162:393–404.
24. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
25. Casimiro-Soriguer CS, Muñoz-Mérida A, Pérez-Pulido AJ, Pulido P. Sma3s: a universal tool for easy functional annotation of proteomes and transcriptomes. *Proteomics* 2017;17:1700071.
26. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ *et al*. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* 2017;45:D200–D203.
27. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
28. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;29:2933–2935.
29. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;30:772–780.
30. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972–1973.
31. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;30:1312–1313.
32. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* 2016;8:12–24.
33. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
34. Price MN, Dehal PS, Arkin AP. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
35. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM *et al*. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* 2018;34:292–293.
36. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 2006;22:1600–1607.
37. Muñoz-Mérida A, Viguera E, Claros MG, Trelles O, Pérez-Pulido AJ. Sma3s: a three-step modular annotator for large sequence datasets. *DNA Res* 2014;21:341–353.
38. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J *et al*. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res* 2018;46:W246–W251.
39. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28:3150–3152.
40. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genomics* 2018;4:mgen.0.000206.

41. Sahl JW, Gillece JD, Schupp JM, Waddell VG, Driebe EM *et al*. Evolution of a pathogen: a comparative genomics analysis identifies a genetic pathway to pathogenesis in *Acinetobacter*. *PLoS One* 2013;8:e54287.
42. Bassett CL, Kushner SR. Exonucleases I, III, and V are required for stability of ColE1-related plasmids in *Escherichia coli*. *J Bacteriol* 1984;157:661–664.
43. Tomaras AP, Dorsey CW, Edelmann RE, Actis LA. Attachment to and biofilm formation on abiotic surfaces by *Acinetobacter baumannii*: involvement of a novel chaperone-usher pili assembly system. *Microbiology* 2003;149:3473–3484.
44. Doughty S, Sloan J, Bennett-Wood V, Robertson M, Robins-Browne RM *et al*. Identification of a novel fimbrial gene cluster related to long polar fimbriae in locus of enterocyte effacement-negative strains of enterohemorrhagic *Escherichia coli*. *Infect Immun* 2002;70:6761–6769.
45. Makarova KS, Anantharaman V, Grishin NV, Koonin EV, Aravind L. Carf and WYL domains: ligand-binding regulators of prokaryotic defense systems. *Front Genet* 2014;5:102.
46. Jurenaite M, Markuckas A, Suziedeliene E. Identification and characterization of type II toxin-antitoxin systems in the opportunistic pathogen *Acinetobacter baumannii*. *J Bacteriol* 2013;195:3165–3172.
47. Pandey DP, Gerdes K. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Res* 2005;33:966–976.
48. Makarova KS, Aravind L, Wolf YI, Koonin EV. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct* 2011;6:38.
49. Cook GM, Robson JR, Frampton RA, McKenzie J, Przybilski R *et al*. Ribonucleases in bacterial toxin-antitoxin systems. *Biochim Biophys Acta* 2013;1829:523–531.
50. Makarova KS, Wolf YI, Koonin EV. Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biol Direct* 2009;4:19.
51. Gweon HS, Bailey MJ, Read DS. Assessment of the bimodality in the distribution of bacterial genome sizes. *ISME J* 2017;11:821–824.
52. Sistla S, Rao DN. S-Adenosyl-L-methionine-dependent restriction enzymes. *Crit Rev Biochem Mol Biol* 2004;39:1–19.
53. Makarova KS, Wolf YI, Koonin EV. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res* 2013;41:4360–4377.
54. McDonald ND, Regmi A, Morreale DP, Borowski JD, Boyd EF. CRISPR-Cas systems are present predominantly on mobile genetic elements in *Vibrio* species. *BMC Genomics* 2019;20.
55. Lobato-Márquez D, Díaz-Orejas R, García-Del Portillo F. Toxin-antitoxins and bacterial virulence. *FEMS Microbiol Rev* 2016;40:592–609.
56. Maikova A, Peltier J, Boudry P, Hajnsdorf E, Kint N *et al*. Discovery of new type I toxin-antitoxin systems adjacent to CRISPR arrays in *Clostridium difficile*. *Nucleic Acids Res* 2018;46:4733–4751.
57. Tong Z, Du Y, Ling J, Huang L, Ma J. Relevance of the clustered regularly interspaced short palindromic repeats of *Enterococcus faecalis* strains isolated from retreatment root canals on periapical lesions, resistance to irrigants and biofilms. *Exp Ther Med* 2017;14:5491–5496.
58. Tang B, Gong T, Zhou X, Lu M, Zeng J *et al*. Deletion of Cas3 gene in *Streptococcus mutans* affects biofilm formation and increases fluoride sensitivity. *Arch Oral Biol* 2019;99:190–197.