Vasquez-Rifo et al. Genome Biology (2019) 20:270
https://doi.org/10.1186/s13059-019-1890-1

Genome Biology

RESEARCH Open Access

The *Pseudomonas aeruginosa* accessory genome elements influence virulence towards *Caenorhabditis elegans*

Alejandro Vasquez-Rifo[1*†], Isana Veksler-Lublinsky[2*†], Zhenyu Cheng[3], Frederick M. Ausubel[4,5] and Victor Ambros[1]

Abstract

Background: Multicellular animals and bacteria frequently engage in predator-prey and host-pathogen interactions, such as the well-studied relationship between *Pseudomonas aeruginosa* and the nematode *Caenorhabditis elegans*. This study investigates the genomic and genetic basis of bacterial-driven variability in *P. aeruginosa* virulence towards *C. elegans* to provide evolutionary insights into host-pathogen relationships.

Results: Natural isolates of *P. aeruginosa* that exhibit diverse genomes display a broad range of virulence towards *C. elegans*. Using gene association and genetic analysis, we identify accessory genome elements that correlate with virulence, including both known and novel virulence determinants. Among the novel genes, we find a viral-like mobile element, the *teg* block, that impairs virulence and whose acquisition is restricted by CRISPR-Cas systems. Further genetic and genomic evidence suggests that spacer-targeted elements preferentially associate with lower virulence while the presence of CRISPR-Cas associates with higher virulence.

Conclusions: Our analysis demonstrates substantial strain variation in *P. aeruginosa* virulence, mediated by specific accessory genome elements that promote increased or decreased virulence. We exemplify that viral-like accessory genome elements that decrease virulence can be restricted by bacterial CRISPR-Cas immune defense systems, and suggest a positive, albeit indirect, role for host CRISPR-Cas systems in virulence maintenance.

Keywords: *C. elegans*, *P. aeruginosa*, Accessory genome, Virulence, CRISPR-Cas

Background

Interactions between environmental bacteria and small invertebrate animals, such as free-living nematodes, are ecologically significant in many terrestrial ecosystems [1]. These interactions comprise many types of ecological relationships that range from reciprocal harm to mutualism. Frequently, animal-bacterial interactions are "predator-prey" relationships, where for example nematodes feed on bacteria. Such predation can in turn drive the evolution of bacterial anti-predator mechanisms, such as the production of noxious toxins, and/or full pathogenic potential where the bacterium can kill and feed on the predator ([2]; reviewed in [3]). One such bacterial species is *Pseudomonas aeruginosa* (*P. aeruginosa*) that is preyed upon by invertebrates, but is also a facultative pathogen of a broad range of hosts including plants, amoeboid protists, insects, mammals, and nematodes [4–7].

The relationship between a facultatively pathogenic bacterium and a predator, such as a free-living nematode, can be bidirectional, with the pathogen either serving as a food source for the predator, or itself thriving on the infected predator. For example, the nematode *Caenorhabditis elegans* (*C. elegans*) [2] can grow from larval stages to the adult by feeding on the pathogenic bacterium *P. aeruginosa*. Interestingly, although *C. elegans* larval development can proceed successfully on *P. aeruginosa*, adults can suffer dramatically reduced lifetimes, depending on the *P. aeruginosa* strain (for example, median adult survival of ~ 2 days on strain PA14 compared to ~ 14 days on *Escherichia coli* strain OP50 that is used as the standard laboratory diet for

* Correspondence: alejandro.vasquezrifo@umassmed.edu; vaksler@post.bgu.ac.il
†Alejandro Vasquez-Rifo and Isana Veksler-Lublinsky contributed equally to this work.
[1]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA
[2]Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

C. elegans). This mutually antagonistic relationship between *C. elegans* and *P. aeruginosa* is well-studied model that serves as a proxy for ecologically coexisting predators of *P. aeruginosa* that are also natural hosts for infection [8].

It is plausible that *C. elegans* and *P. aeruginosa* interact in natural niches, as *P. aeruginosa* is known to inhabit many environments including soils [9–11] and *C. elegans* is often an inhabitant of soil and rotting plant matter [12]. These interactions could be transitory in the wild, due to worm avoidance of *P. aeruginosa* or death of the worms, and thus difficult to catalog, but have been substantiated by a report of natural coexistence of the two species (reviewed in [12]). Nonetheless, independently of their putative co-existence in the wild, *C. elegans* can be used as an experimentally tractable proxy of naturally occurring predator and host of *P. aeruginosa*.

Considering that *P. aeruginosa* is a free-living bacterial species that facultatively engages in pathogenic interactions with invertebrates, and that *C. elegans* is a natural bacterial predator, it seems likely that *P. aeruginosa* strain variation in virulence towards *C. elegans* reflects adaptations of *P. aeruginosa* to its natural niches. In natural settings, virulence may be a character under selection by the frequency with which predators are deterred by virulence mechanisms, and/or by the extent to which the bacterium depends on infection of predator hosts for population growth. Such variability in bacterial virulence should be reflected in the genomic composition of different bacterial isolates, and determining the mechanisms underlying this variability enhances our understanding of the evolution of host-microbe interactions.

In the present work, we addressed the sources and genomic correlates of bacteria-driven variability in the virulence of distinct *P. aeruginosa* strains towards *C. elegans*. A previous study of 20 *P. aeruginosa* natural isolates revealed strain-driven variation in *P. aeruginosa* virulence, highlighting virulence as a complex trait, likely the result of multiple components acting in a combinatorial manner [13]. Extending this previous work, we conducted an in-depth genome-wide comparative survey of a set of 52 *P. aeruginosa* strains. We used comparative genomic approaches to identify correlations between *P. aeruginosa* virulence and the presence/absence of specific accessory genome elements, including bacterial immune defense systems.

Our analysis revealed gene sets in the accessory genome of *P. aeruginosa* (i.e., the set of genes present in some, but not all, of the strains in the species) that correlate either with high or low virulence. Our approach identified known virulence factors, as well as novel factors that can directly modulate bacterial virulence, either positively or negatively, as evidenced through genetic testing. We also identified genes that may indirectly affect virulence. For example, our study revealed a positive role in virulence for certain bacterial immune defense systems which filter horizontal gene transfer (HGT), and hence can impact the composition of the accessory genome. In particular, we found that *P. aeruginosa* strains with active CRISPR-Cas systems have statistically higher levels of virulence towards *C. elegans* and that spacer-targeted genes are among the genes associated with lower virulence. These correlative findings, together with our genetic confirmation of virulence-inhibitory activity of certain accessory genome elements, support an indirect role for CRISPR-Cas systems in contributing to the maintenance and evolution of high virulence against nematodes.

Results

A large *P. aeruginosa* accessory genome underlies substantial strain diversity in gene content

To assess the extent of variation in genetic makeup among a diverse panel of environmental and clinical *P. aeruginosa* strains, we analyzed in silico the genomes of 1488 *P. aeruginosa* strains. The protein-coding genes of the strains were assigned to clusters of homologous genes using the CD-HIT program [14] with a threshold of 70% amino acid similarity. The clustering procedure resulted in the identification of 28,793 distinct gene clusters (i.e., groups of homologous genes). We then examined the distribution and frequency of these 28,793 genes across the 1488 *P. aeruginosa* strains. Five thousand one hundred seventy genes were present in more than 90% of the isolates and were accordingly defined as constituting the *P. aeruginosa* core genome (Fig. 1a). The remaining 23,623 genes constitute the accessory genome of these 1488 *P. aeruginosa* strains. The frequency distribution of the genes is bimodal, with prominent maxima corresponding to the core genome and the set of genes that occur only once in these strains (referred to as "singletons," Fig. 1b). The ratio between the pangenome and the core genome (5.6) agrees with a previously reported ratio: 5.3 [15], confirming that *P. aeruginosa* harbors a large amount of strain-specific variation in protein-coding genes.

To model the phylogenetic relationships between the *P. aeruginosa* isolates, we aligned the core genomes and used the alignments to build a phylogenetic tree (Fig. 1c). The isolation source of the strains, when available, was categorized as clinical or environmental and this designation was mapped to the tree (Fig. 1c). Environmental strains distribute across multiple branches of the tree altogether with the clinical isolates. This pattern is consistent with other studies that showed that both clinical and environmental isolates of *P. aeruginosa* can originate from the same clade [16–19].

In order to experimentally study the effect of bacterial genetic variation on the interaction between *P. aeruginosa* and *C. elegans*, we assembled a collection of 52 representative *P. aeruginosa* strains (Additional file 2: Table S1) included in the in silico collection of 1488. The collection consists of bacterial isolates derived from clinical

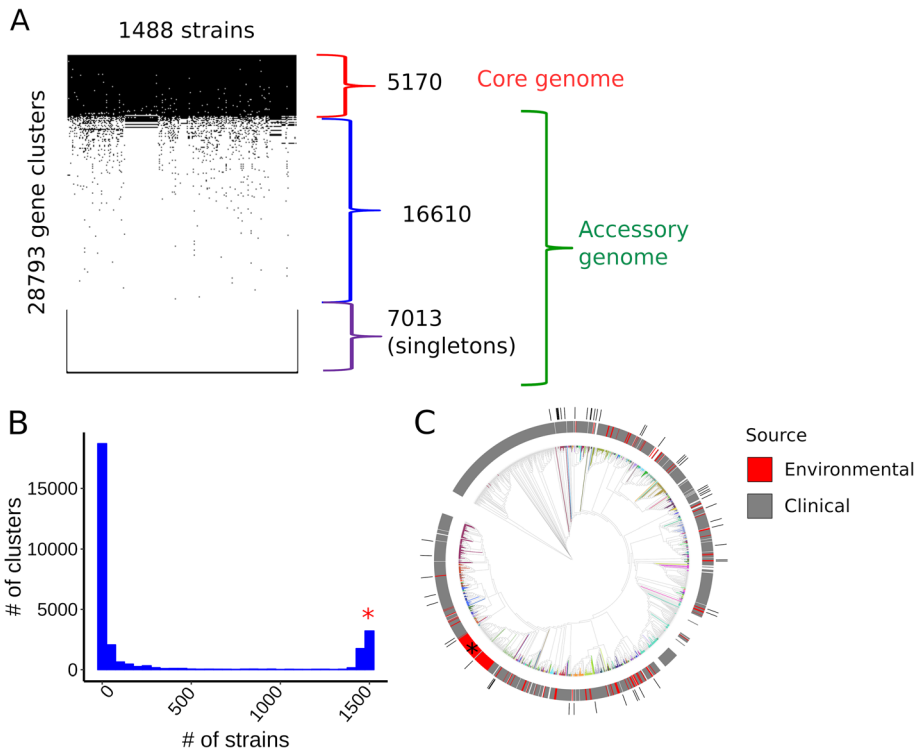

Fig. 1 Pangenomic and phylogenetic features of *Pseudomonas aeruginosa*. **a** Presence/absence matrix of *P. aeruginosa* genes across the strains. Core and accessory genomes are marked. **b** The frequency distribution of the genes among the 1488 strains. The right end of the distribution, marked with red *, corresponds to the core genome, while the left end of the distribution corresponds to singletons and rare accessory genes. **c** Phylogenetic tree of *P. aeruginosa* strains. Phylogenetically related MLST groups are shown in different colors. Isolation source is shown on top of the tree. An apparent clade enriched for environmental strains (indicated by *) is artificially enlarged by the repeated presence of a set of almost identical genomes in the set used to build the phylogeny. The phylogenetic locations of the 52 isolates experimentally tested in this study are indicated in the outer circle (black bars)

(85%, mostly from primary infections) and environmental (15%) settings. The 52 strains distributed widely across *P. aeruginosa* phylogeny (Fig. 1c). The 52-strain cohort have a pangenome of 11,731 genes and an accessory genome of 6537 genes.

Virulence towards the nematode *C. elegans* strongly varies among *P. aeruginosa* strains

To assess phenotypic variation in interactions of *P. aeruginosa* with *C. elegans*, we measured the virulence towards *C. elegans* wildtype worms for the collection of 52 *P. aeruginosa* strains. Young adult *C. elegans* hermaphrodites were exposed to a full lawn of each *P. aeruginosa* strain using so-called slow kill (SK) media [8]. These assay conditions induce bacterial quorum sensing regulation, a system that mediates biofilm, a naturally occurring mode of *P. aeruginosa* growth [20]; minimize the effects of worm behavior on survival [21, 22]; and promote bacterial colonization of the worm gut [8]. Adult lifetime was scored using a semi-automated method [23] to obtain survival curves for worms exposed to each bacterial strain (Fig. 2a). Bacterial strain virulence towards *C. elegans* was measured as the median survival time of worms exposed to each bacterial strain (Fig. 2b). Virulence varied continuously over a fivefold range, spanning from 1.5 to over 10 days (Fig. 2b). Indeed, the median worm survival on *P. aeruginosa* for strain z7, which exhibited the lowest virulence towards *C. elegans*, was greater than that of worms exposed to *E. coli* HB101, a strain commonly used in the laboratory to maintain worm stocks (Fig. 2b). In addition, under SK conditions, the number of viable progeny produced by hermaphrodites exposed to strain z7 was indistinguishable from that of animals exposed to *E. coli* HB101 (Additional file 1: Figure S1A). Altogether, these results show that for our experimental set of 52 *P. aeruginosa* strains, virulence varies continuously over a wide range, from highly virulent strains, which kill *C. elegans* adults within 2 days, to essentially completely avirulent strains that do not detectably impair worm lifespan or reproduction in comparison to their normal laboratory food.

To evaluate the potential contribution of strain isolation source to virulence against *C. elegans*, we compared the set of clinical isolates to the environmental isolates. Strains from clinical settings displayed lower mean virulence when

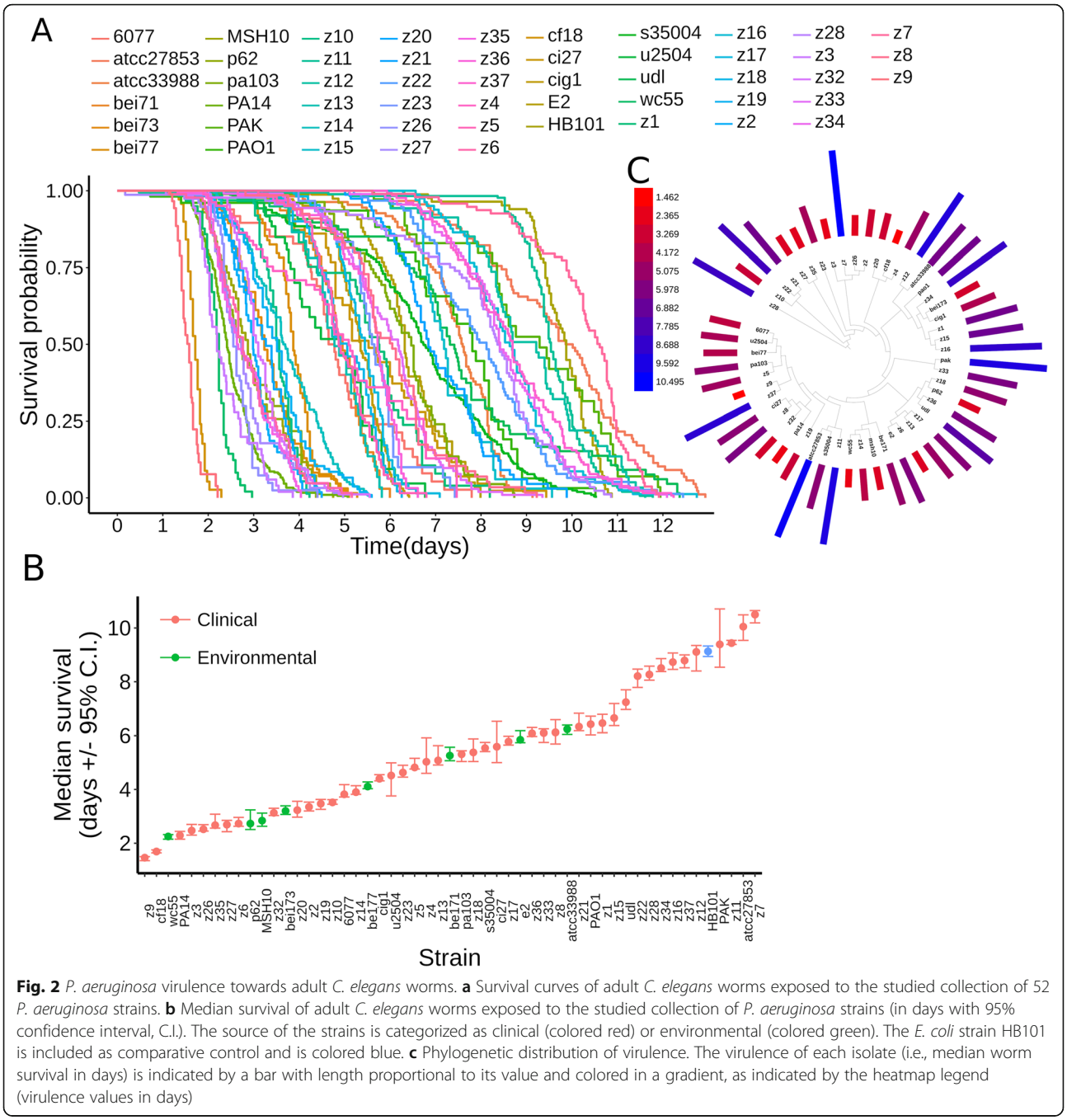

Fig. 2 *P. aeruginosa* virulence towards adult *C. elegans* worms. **a** Survival curves of adult *C. elegans* worms exposed to the studied collection of 52 *P. aeruginosa* strains. **b** Median survival of adult *C. elegans* worms exposed to the studied collection of *P. aeruginosa* strains (in days with 95% confidence interval, C.I.). The source of the strains is categorized as clinical (colored red) or environmental (colored green). The *E. coli* strain HB101 is included as comparative control and is colored blue. **c** Phylogenetic distribution of virulence. The virulence of each isolate (i.e., median worm survival in days) is indicated by a bar with length proportional to its value and colored in a gradient, as indicated by the heatmap legend (virulence values in days)

compared to strains isolated from non-clinical, environmental settings (Welch *t* test, *p* value = 0.047, Additional file 1: Figure S1B). This result suggests that clinical strains isolated from infected humans do not constitute a biased sampling of strains that are relatively more pathogenic to worms than environmental isolates. Rather, it is possible that some clinical strains could harbor variations and adaptations that disfavor virulence towards worms.

Next, we evaluated the distribution of virulence along the *P. aeruginosa* phylogeny. Mapping of virulence onto the phylogenetic tree of the studied isolates showed no phenotypic clustering of virulence towards any particular clade (Fig. 2c). Thus, evolutionarily fluctuations in virulence among isolates occur without any particular affiliation to select phylogenetic clades.

Defects in bacterial growth rates can impair virulence towards *C. elegans*, and such impairments can be detected in vitro (e.g., [24]). Thus, we assessed whether strain-specific virulence against *C. elegans* could primarily reflect the relative growth rate capacity of each strain, as determined by

growth rate in LB media at 25 °C (the temperature of the virulence assays). We found that growth rate in LB medium showed no statistically significant correlation with virulence (Additional file 1: Figure S2, Pearson's correlation, $\rho = -0.3$, p value = 0.08).

P. aeruginosa virulence correlates with the presence of particular accessory genome elements

We employed gene association analysis to test whether virulence of *P. aeruginosa* strains towards *C. elegans* could be associated with the presence or absence of specific bacterial genes. In this analysis, virulence is defined as a quantitative trait for each strain, corresponding to the median lifespan of adult *C. elegans* hermaphrodites when fed each of the strains. The association between genes and virulence was measured using the Mann-Whitney (MW) and linear regression (LR) tests, followed by a gene permutation approach, to control for multiple statistical testing and thus assess the reliability of the p value. Furthermore, genes with significant associations, as determined by the MW and LR tests, were evaluated with two additional metrics that consider phylogeny to resolve confounding effects due to population structure, namely, the "simultaneous" and "subsequent" scores of the treeWAS method described by Collins and Didelot [25] (Additional file 3: Table S2). Gene associations were assessed for the set of 11,731 protein-coding pangenomic genes of the 52 experimental strains and for a set of 83 previously-identified non-coding RNA genes (excluding rRNAs and tRNAs) of *P. aeruginosa.*

The small non-coding RNAs of bacteria fulfill diverse gene regulatory roles and can modulate pathways required for virulence [26, 27]. Interestingly, we noted that most of the non-coding RNA genes we examined are core genome elements (78%, 65/83 genes). We found no statistically significant association between the non-coding RNAs of *P. aeruginosa* and virulence (Additional file 1: Figure S3A, all p value > 0.05 for the MW and LR tests).

Among the 6537 protein-coding accessory genes present in the 52-strain experimental panel, we identified 79 genes significantly associated with virulence, either positively or negatively (Fig. 3, p value < 0.01 for the MW or LR tests). For 35 of these 79 virulence-associated genes (44%), their presence defined a set of strains with higher virulence compared to the strain set where the same genes were absent (Fig. 3a). We refer to them as high virulence-associated genes (or "HVA genes" for short). For the other 44 genes (56%), their presence corresponded to strains with lower virulence (Fig. 3a). We refer to these as low virulence-associated genes (or "LVA genes" for short). Each strain harbors a different subset of the 79 associated genes. For example, strain PA14, a highly virulent strain, has 19 HVA genes and 1 LVA gene (Fig. 3b). On the other side of the spectrum, strain ATCC27853, a poorly virulent isolate, has 5 HVA genes and 41 LVA genes (Fig. 3c). A description of the 79 genes associated with higher or lower virulence is presented in Additional file 3: Table S2. All the LVA genes (44/44 or 100%) were supported by either the simultaneous or subsequent scores (p value < 0.05). Similarly, 30/35 of the HVA genes (86%) were supported by either simultaneous or subsequent scores (p value < 0.05, Additional file 2: Table S1). Altogether, these phylogenetically aware scores suggest that population structure does not confound interpretation of the gene associations observed. This result is also congruent with the absence of phenotypic clustering of virulence in the phylogenetic tree (Fig. 2d).

The 79 virulence-associated genes encompass a variety of functions, although for many of the associated genes, a functional annotation is not available (43% of HVA genes and 64% of the LVA genes are annotated as "hypothetical proteins"). Associated genes could be categorized as follows: (1) Genes with known regulatory roles: Such roles can be ascribed to strain PA14 genes PA14_27700 (HVA gene #13286) and PA14_27690 (HVA gene #15454), which encode a cAMP-dependent protein kinase and RNA polymerase sigma factor, respectively. A second example is the *qsrO* gene (LVA gene #17701), which negatively regulates a highly conserved quorum sensing pathway (Köhler et al., 2014). (2) Genes that encode proteins associated with structural roles: The *pslM* (HVA gene #2628) and *pslK* (HVA gene #2479) genes belong to the psl polysaccharide biosynthetic pathway, a polymer that contributes to biofilm formation [28]. Other examples are the HVA genes #6371, #8276, and #8113, which encode homologs of *wbpZ*, *wbpL*, and *wzz*, respectively. These homologs encode enzymes required for LPS O-antigen synthesis [29], a structural component of the bacterial outer membrane. (3) Mobile genetic elements: Several of the genes associated with low virulence are annotated as integrase (genes #6157, #4439, #10878, #8459), or phage-related (genes #8274, #5222), suggests that these genes are likely to encode components of mobile genetic elements. Further support for the mobility of these elements comes from their targeting by CRISPR spacers (see below).

Among the genes that we found to be associated with high virulence across the 52-strain panel, two HVA genes, PA14_27700 and PA14_27690, have been previously characterized as virulence genes. Previous genetic analysis showed that loss of function mutations in either PA14_27700 (HVA gene #13286) or PA14_27690 (HVA gene #14622) compromised the virulence of strain PA14 against *C. elegans* [24] under the SK assay conditions, the same condition used in the present study. Our examination of the published literature identified a total of 60 previously described *P. aeruginosa* virulence genes (Additional file 4: Table S3) that were identified by genetic analysis of virulence against *C. elegans* for two commonly studied *P. aeruginosa* strains, PA14 and PAO1

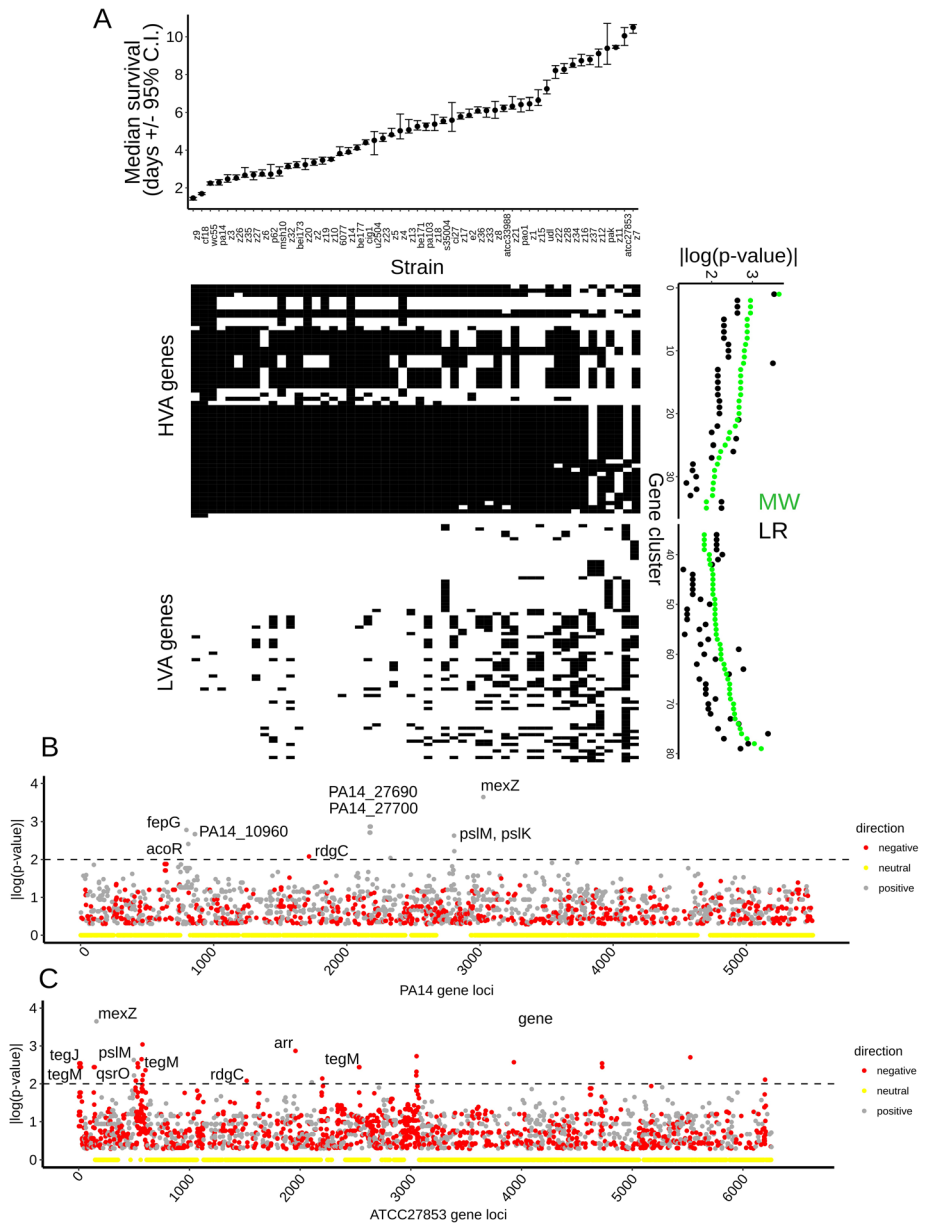

Fig. 3 Association between protein-coding genes of *P. aeruginosa* and bacterial virulence. **a** (Top panel) median survival of adult *C. elegans* worms exposed to a collection of 52 *P. aeruginosa* strains (with 95% confidence interval, C.I.). The strains are ordered from high to low virulence (left to right) and aligned with the matrixes below middle and bottom left panels: Gene presence/absence matrix for HVA genes (middle) and LVA genes (bottom). Gene presence is indicated with black squares and absence with white squares. Genes (rows) are aligned with the corresponding *p* values. Middle and bottom right panels: Association statistics (*p* value of MW and LR tests) for the HVA and LVA genes, shown as $|\log_{10}(\text{pval})|$. **b**, **c** Associated genes present in the strain PA14 (**b**) or ATCC27853 (**c**). Gene loci are plotted against the association statistic (*p* value of MW test), shown as $|\log_{10}(\text{pval})|$. Loci are colored according to the directionality of the gene-virulence association (gray: positively associated; red: negatively associated; yellow: *p* value equals zero). Horizontal dashed lines demarcate a significance threshold ($p < 0.01$)

(Fig. 4a, b), both of which are included in our experimental test panel. Upon analysis of these 60 genes, we found that two of the HVA genes associated with virulence in our 52-strain panel (Additional file 3: Table S2), *pslM* (HVA gene #2628) and *pslK* (HVA gene #2479), were not previously identified as virulence genes in PA14 or PAO1, but are contained in the same *psl* operon as the previously identified virulence gene *pslH* (gene #6064), which was shown to be required for full virulence in the PAO1 strain [30].

Other than PA14_27700, PA14_27690, and the *psl* operon genes (*pslM, pslK*), no other genes from the set of 60 previously described virulence factors showed association with virulence in this study (Fig. 4; Additional file

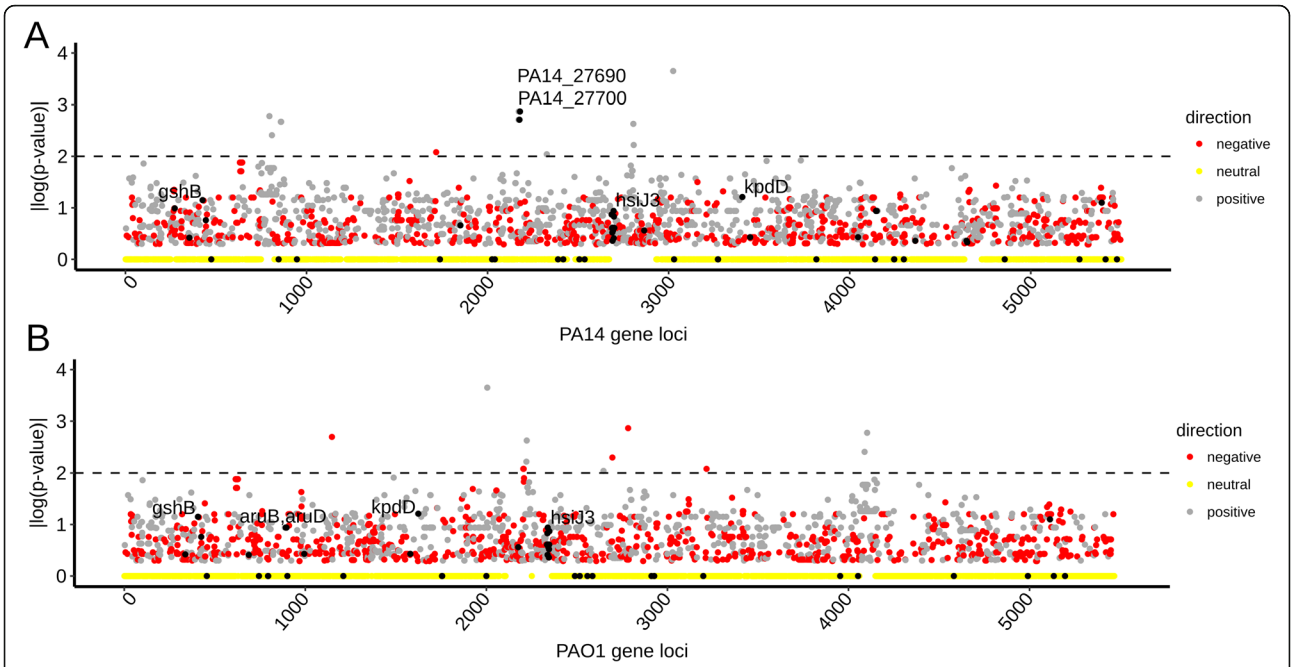

Fig. 4 Distribution and features of previously-identified virulence genes. **a**, **b** Gene association for PA14 (**a**) and PAO1 (**b**) protein-coding gene loci. Gene loci are plotted against the association statistic (*p* value of MW test), shown as |log10(*p* value)|. Previously identified virulence genes are indicated with black dots and the top 5 most associated genes labeled. The top known genes associated with virulence are PA14_27690 and PA14_27700. Loci are also colored according to the directionality of the gene-virulence association (gray: positively associated; red: negatively associated; yellow: *p* value equals zero). Horizontal dashed lines demarcate a significance threshold ($p < 0.01$)

1: Figure S3B). Notably, 51 of the 60 known virulence genes (85%) belong to the core genome of our panel of 52 experimental strains, explaining the null association observed. The remaining previously identified virulence genes that did not emerge as HVA genes in our 52-strain panel may not have a strong enough impact on virulence across our 52 stains for a variety of potential reasons, including strain-specific epistasis from other accessory genome elements.

Genetic tests identify *P. aeruginosa* accessory genome elements that contribute to decreased or increased virulence towards *C. elegans*

The statistical association of particular protein-coding genes with either high virulence (in the case of HVA genes) or low virulence (in the case of LVA genes) across the set of 52 experimental strains tested here could in principle reflect the presence or absence of single genes that are individually necessary and/or sufficient to impact virulence. In such cases, loss-of-function or gain-of-function genetic manipulations of the relevant strains would be expected to measurably impact virulence. However, single gene causality may in some cases be masked by strain-specific epistatic interactions, for example with other accessory genes. It would not be unexpected if some of the HVA and LVA genes that we identified were to function in combination, such that the contribution of each individual gene would not be easily evident from single gene knock out or overexpression tests. It is also possible that a gene with no direct function in virulence could nevertheless show association with virulence because of a physiological or ecological linkage between the function of that gene and the function and/or acquisition of bona fide virulence factors.

The above-expected caveats notwithstanding, we used loss-of-function and gain-of-function approaches to test whether individual HVA genes are necessary and/or sufficient to support high virulence, and conversely, whether LVA genes are necessary and/or sufficient to impose reduced virulence. For most of these genetic tests, we selected strain z8, which exhibits an intermediate level or virulence, contains members of both the HVA and LVA gene sets, and is amenable to genome-editing through use of its endogenous CRISPR-Cas system.

The set of HVA genes included previously validated virulence genes (e.g., PA14_27700, PA14_27690), which we did not re-test here. Instead, we evaluated the potential role in virulence for *mexZ* (gene #14466), which had not been previously tested genetically. We constructed an in-frame deletion of *mexZ* in strain z8 (*ΔmexZ*), but no difference in virulence was found for *ΔmexZ* when compared to the wildtype z8 strain (Additional file 1: Figure S4). The absence of a direct effect on virulence of strain z8 suggests that the association of *mexZ* with virulence among the

panel of 52 strains could be secondary to additional underlying factors. *mexZ* is frequently mutated in clinical isolates, as a part of the bacterial adaptations to acquire antibiotic resistance [31, 32].

We next selected genes associated with low virulence to test their effects by using loss-of-function and gain-of-function approaches. We assigned gene names to the genes selected for study that were not previously named (Fig. 5a and Additional file 5: Table S4). The selected genes belong to three genomic loci: the *ghlO* gene (LVA gene# 25296) is associated with virulence as a single gene (i.e., no additional neighboring genes are associated with virulence); the *qsrO* gene (LVA gene# 17701, [33]) belongs to a four-gene operon (referred to as "*qsr*" operon); and the *tegG* to *tegN* genes (LVA genes # 5222, 5330, 10513, 15466, 21386, 21557, 26140) constitute a block of contiguous genes in bacterial chromosomes (referred to as the "*teg* block" described below).

We constructed strain z8 mutants carrying in-frame deletions of *ghlO*, *qsrO*, and the *teg* gene block (*ΔghlO*, *ΔqsrO*, and *Δteg*, respectively, see also Additional file 6: Table S5) and measured virulence on two *C. elegans* strains: wildtype and *pmk-1(lf)* mutant. The *pmk-1(lf)* mutant has an impaired p38/PMK-1 pathway that compromises the worm's response to *P. aeruginosa* PA14 [34] and z8 strains (Fig. 5b, c). This worm mutant was used as a strain with a genetically "sensitized" background. Deletion of *ghlO* led to marginally reduced survival of wildtype worms (Fig. 5b) but not of *pmk-1(lf)* worms (Fig. 5c). Deletion of *qsrO*, but not of *teg*, led to a significant reduction in the survival of wildtype worms, indicating an increased virulence of the *ΔqsrO z8* bacteria (Fig. 5b). Similarly, deletion of *qsrO*, but not of *teg*, led to a mild but significant reduction in the survival of *pmk-1(lf)* worms (Fig. 5c). These results support a direct negative role for the *qsrO* gene in the regulation of virulence. Interestingly, the *qsrO* gene had been reported previously to have a negative regulatory function on quorum sensing (QS), a key contributor to *P. aeruginosa* virulence [33].

To test if the selected genes associated with low virulence can modulate virulence when their expression is enhanced, we constructed strains containing multi-copy plasmids that encode the *ghlO* gene (p ($ghlO^+$)), the *qsr* operon (p (qsr^+)), and *teg* block genes (p ($tegLM^+$) and p ($tegN^+$)) driven by their native promoters in their respective mutant backgrounds (Additional file 6: Table S5). The virulence of these strains was measured and compared to a strain carrying an empty plasmid control (p (control)). The virulence of strains overexpressing the *qsrO* and *tegN* genes was significantly reduced compared to the control (Fig. 5d, p value $< 10^{-4}$). In contrast, no difference in virulence was observed for strains overexpressing the *ghlO* and *tegLM* genes (Fig. 5d, p value > 0.01). Strains overexpressing *qsrO* or *tegN* also displayed reduced virulence when tested on immunocompromised *pmk-1(lf)* (Fig. 5e, p value < 0.01). This effect of diminished virulence was abolished when the *qsrO* and *tegN* genes in the plasmids were mutated by introduction of an early stop codon (p (qsr^{mut}) and p ($tegN^{mut}$), Fig. 5e, p values > 0.01, see also Additional file 6: Table S5).

These results suggest a direct role for the *qsrO* and *tegN* genes in the negative regulation of virulence. By contrast, our results suggest the associations of *mexZ*, *ghlO*, and *tegL* and *tegM* genes with high virulence may not reflect direct causal roles in virulence per se. Rather, these latter associations may be secondary to additional underlying factors related to physiological or ecological linkages to virulence. In light of these, findings that at least some genes of the accessory genome of *P. aeruginosa* (for example, *qsrO* and *tegN*) can directly modulate virulence imply that processes of selective gene deletion and acquisition (such as horizontal gene transfer, HGT) are critical for the evolution of *P. aeruginosa* virulence in the wild. In summary, the present gene association study identifies 4 previously characterized virulence genes (i.e., PA14_27700, PA14_27690, *pslM*, *pslK*). In addition, we genetically tested 11 LVA genes by deletion approach, and 6 of these LVA genes by an expression approach, identifying direct roles for *qsrO* and *tegN* in reducing virulence. Importantly, *tegN* is evolutionarily gained or lost altogether with a defined set of 8 accompanying neighboring *teg* genes, i.e., in a physically linked "gene block" (see below, and Additional file 3: Table S2). Thus, all *teg* genes show association with virulence by being linked to a bona-fide virulence modifier gene (i.e., *tegN*), even though some may not have direct effects on virulence (e.g., *tegM*). A similar pattern is found in other associated genes that are also found in physically linked gene blocks and are evolutionarily gained or lost as units (e.g., *qsrO*, PA14_27700).

The *teg* block is a mobile genetic element that impinges on virulence

Our gene association analysis revealed that the *teg* genes (i.e., genes *tegG* to *tegN*) are LVA genes. Among the experimental isolate collection, strains where this group of *teg* genes is present had lower virulence compared to those where it is absent (Welch t test, p value = 0.005), as expected from the gene association results. Our finding that *tegN* directly modulates virulence when expressed (Fig. 5d, e) strongly suggests a functional link between the *teg* genes and reduced virulence.

To better understand the organization of the *teg* genes and their possible mode of acquisition/loss, we examined features of the *tegN* locus by in silico analysis of three *P. aeruginosa* isolates with complete genomes (strains atcc27853, SCV20265, and PA7790) that allow uninterrupted examination of chromosomal features and synteny around *tegN*. The *teg* locus contains a conserved

A

strain	gene(s)	association	frequency (% of strains)
z8	tegG to tegN	LVA	17
z8	ghlO	LVA	15
z8	qsrO	LVA	40

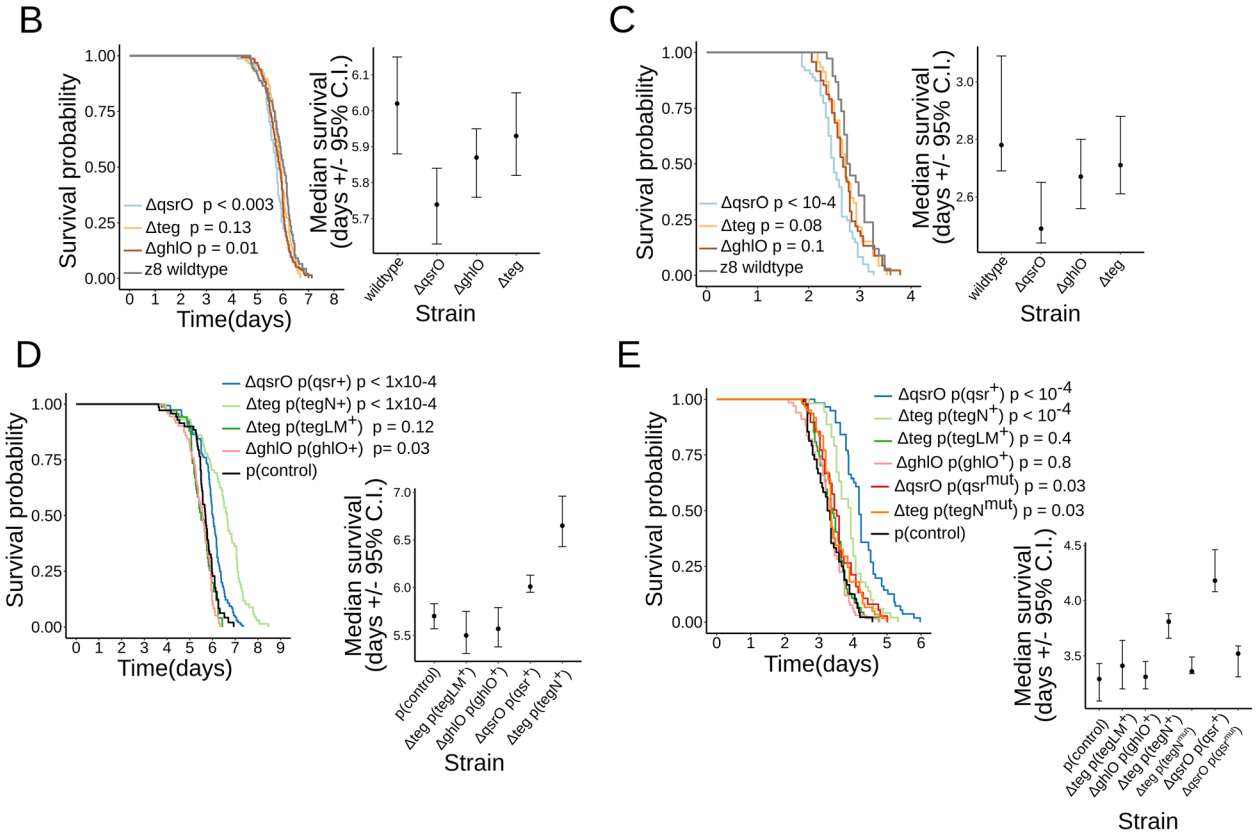

Fig. 5 Bacterial virulence upon loss or expression of genes associated with lower virulence. **a** Summary of the tested LVA genes. Strain, gene nomenclature, and gene frequency are indicated. **b**, **c** Survival curves and median survival (confidence interval, C.I.) of two strains of adult *C. elegans* worms exposed to three strains of *P. aeruginosa* z8 with deletions in genes associated with lower virulence (i.e., Δ*qsrO*; Δ*teg*; Δ*ghlO*). Wildtype worms are analyzed in **b**, *pmk-1(lf)* worms in **c**. Pairwise comparisons of the survival curves between each strain and the z8 wildtype isolate were done using the logrank test. The test *p* values are indicated next to each mutant strain in the legend. **d** Survival curves and median survival (confidence interval, C.I.) of wildtype adult *C. elegans* worms exposed to four strains of *P. aeruginosa* z8 with plasmids expressing genes in gene blocks associated with lower virulence (i.e., Δ*qsrO p (qsr +)*; Δ*teg p (tegN +)*; Δ*teg p (tegLM +)*; Δ*ghlO p (ghlO +)*). Pairwise comparisons of the survival curves between each strain and the z8 wildtype strain with control plasmid (p (control)) were done using the logrank test. The test *p* values are indicated next to the corresponding strain in the legend. **e** Survival curves and median survival (confidence interval, C.I.) of *pmk-1(lf)* adult *C. elegans* worms exposed to six strains of *P. aeruginosa* z8 with plasmids expressing genes associated with lower virulence. Four bacterial strains express wildtype bacterial genes (i.e., Δ*qsrO p (qsr +)*; Δ*teg p (tegN+)*; Δ*teg p (tegLM+)*; Δ*ghlO p (ghlO+)*). Two additional bacterial strains express mutated bacterial genes (i.e., Δ*qsrO p (qsr mut)*; Δ*teg p (tegN mut)*). Pairwise comparisons of the survival curves between each strain and the z8 wildtype strain with control plasmid (p (control)) were done using the logrank test. The test *p* values are indicated next to the corresponding strain in the legend

genomic repeat of ~ 7 kilobases (Fig. 6a). This genomic repeat is found in 2–4 tandem copies in the queried genomes (Fig. 6a). The repeats are not completely identical between strains and display stretches of varying conservation (Fig. 6b). We refer to this tandem genomic repeat unit as the "*teg* block."

The frequency and phylogenetic distribution of the *teg* block in the 52-strain collection suggest that the element is mobile. The block is found in 10 strains, corresponding to 19% of the collection (Additional file 2: Table S1), and it is distributed to multiple clades (Fig. 6c). The simplest hypothesis to account for the phylogenetic pattern of the *teg* block is seven independent acquisitions. A comparison of the genomic neighborhoods surrounding the location of the *teg* block in the 3 complete genomes showed no evident synteny (Fig. 6d), arguing against an ancestrally fixed genomic location, and also supporting the conclusion that the *teg* block is a mobile genetic element. Curiously, two genes (PA2221, PA3080) were commonly shared in 2 distinct pairs of neighborhoods.

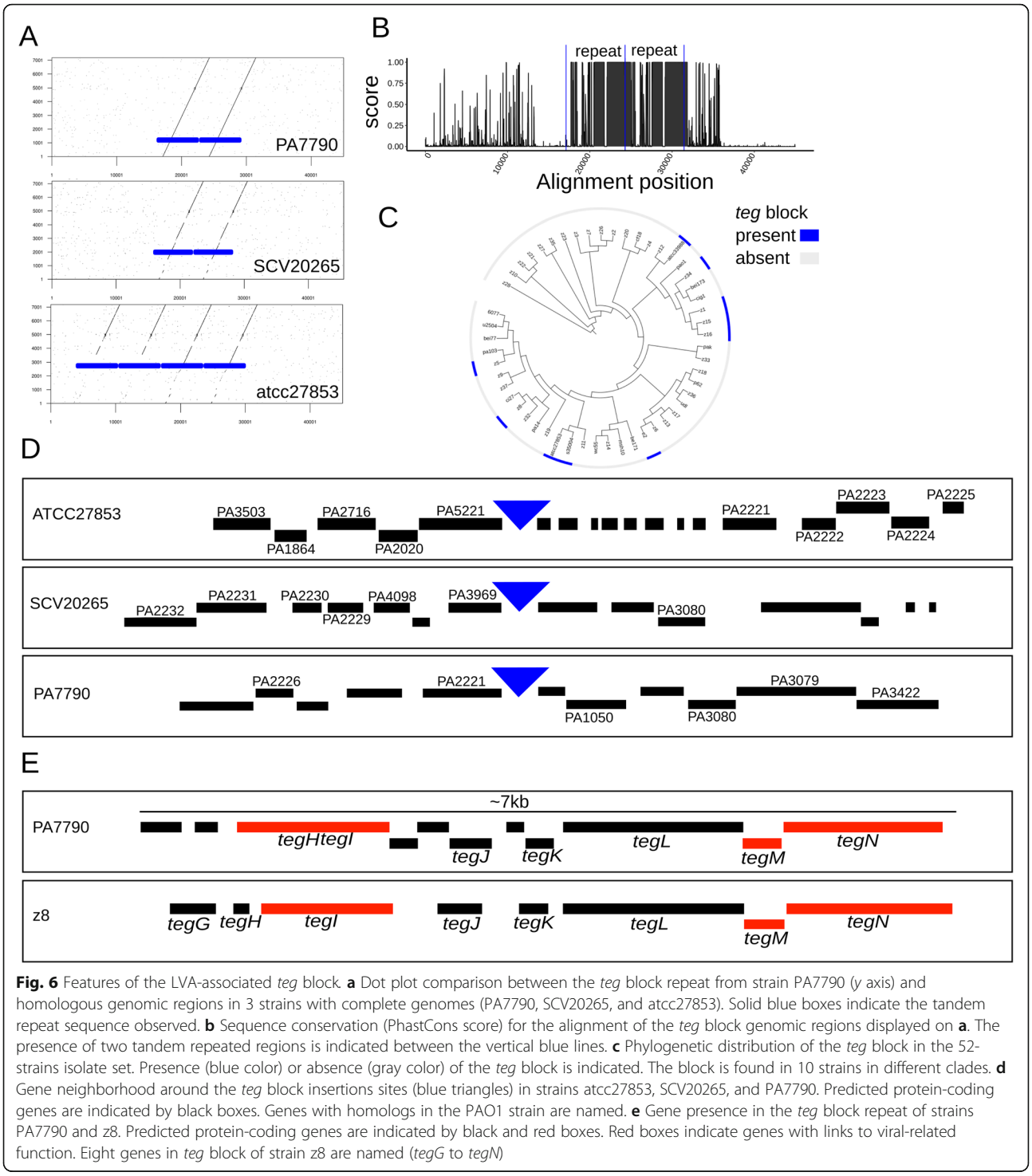

Fig. 6 Features of the LVA-associated *teg* block. **a** Dot plot comparison between the *teg* block repeat from strain PA7790 (*y* axis) and homologous genomic regions in 3 strains with complete genomes (PA7790, SCV20265, and atcc27853). Solid blue boxes indicate the tandem repeat sequence observed. **b** Sequence conservation (PhastCons score) for the alignment of the *teg* block genomic regions displayed on **a**. The presence of two tandem repeated regions is indicated between the vertical blue lines. **c** Phylogenetic distribution of the *teg* block in the 52-strains isolate set. Presence (blue color) or absence (gray color) of the *teg* block is indicated. The block is found in 10 strains in different clades. **d** Gene neighborhood around the *teg* block insertions sites (blue triangles) in strains atcc27853, SCV20265, and PA7790. Predicted protein-coding genes are indicated by black boxes. Genes with homologs in the PAO1 strain are named. **e** Gene presence in the *teg* block repeat of strains PA7790 and z8. Predicted protein-coding genes are indicated by black and red boxes. Red boxes indicate genes with links to viral-related function. Eight genes in *teg* block of strain z8 are named (*tegG* to *tegN*)

The predicted proteins encoded by the *teg* block also support genetic mobility as a potential function. The conserved repeat unit (i.e., *teg* block) has 8 and 11 predicted protein-coding genes in strains PA7790 and z8, respectively, and includes the *tegG* to *tegN* set, named and investigated in strain z8 (Fig. 6e). Five of the predicted *teg* proteins (*tegG*, *tegH*, *tegJ*, *tegK*, *tegL*) have no features or annotations that could help infer their functions. However, three of the *teg* proteins have features and annotations that suggest virus-related functions. The gene *tegI* encodes a viral "replication initiation protein" homologous to *gpII* of phage M13. *tegM* encodes a homologue of viral coat protein g6p of phage Pf3, with a conserved DUF2523 domain (CDD domain accession:

pfam10734). *tegN* encodes a P-loop containing NTPase (CDD domain accession: cl21455), a homologue of *gpI* found in phage M13. These annotations suggest that the *teg* block encodes functions related to DNA replication (*tegI*) and virion assembly (*tegM* and *tegN*) [35, 36], supporting the conclusion that the *teg* block is a virus-related element. The apparent absence of proteins with functionality for chromosomal integration or conjugative transfer may indicate that the *teg* block may rely on proteins from its bacterial host or other mobile genomic elements for these putative functions.

Genomic presence of the *teg* block is restricted by CRISPR-Cas systems

The composition of the *P. aeruginosa* accessory genome is shaped by uptake of genes from other microorganisms via horizontal gene transfer (HGT), frequently involving mobile genetic elements (MGE) such as prophages and ICEs (integrative and conjugative elements). HGT events can be restricted by diverse classes of bacterial defense systems, which protect cells against the acquisition of elements that could confer deleterious phenotypes. Since we observed that the *teg* block, a viral-like element of the *P. aeruginosa* accessory genome, associates and negatively regulates virulence, we investigated if such element would be restricted by the bacteria.

We first explored the possibility that CRISPR-Cas systems could restrict the uptake of the *teg* block. For this purpose, we utilized the existence of an immunity record in the CRISPR spacer loci of *P. aeruginosa* strains. CRISPR repeat spacer sequences identify genes whose restriction by CRISPR-Cas systems of *P. aeruginosa* has been selected for during the recent evolution of the strains examined. Except in rare cases of apparent spacer "self-targeting" [37] (also, see below), CRISPR spacers and their protospacer target genes are predominantly found in different genomes.

We identified the set of all CRISPR spacers present in 1488 strains and searched for their targets in the *P. aeruginosa* pangenome. In this manner, we identified 688 genes that are targeted by spacers (Additional file 7: Table S6). The vast majority (670 out of 688, corresponding to 97%) of the identified spacer-targeted genes are not found on the same genomes as the spacers that target them and thus reflect genes whose integration into the genome of a given strain was successfully blocked by CRISPR-Cas during the evolution of that strain. We next determined the relationship of the spacer-targeted genes to virulence. At the single gene level, the vast majority of the spacer-targeted genes (678) showed no statistically significant correlation with virulence (Fig. 7a). Nonetheless, a set of 9 genes was associated with low virulence (i.e., LVA genes, Fig. 7a, p value < 0.01 by M-W test). In contrast, only one spacer-targeted gene (cluster #18193) showed significant association with high virulence.

Among the LVA spacer-targeted gene set, 5 out of 9 genes were found to be genes in the *teg* block (Fig. 7b). Thus, the spacer-encoded immunity record shows repeated restriction of the *teg* block by CRISPR-Cas systems, consistent with it being detrimental to bacteria. Additional spacer-targeted genes included mostly genes of unknown function, although some annotations related them to mobile elements (i.e., integrase for gene #6157, "phage capsid" for gene #8274) as expected.

Considering that the spacer-encoded record of restricted genes is finite and reflects recent restriction events, we evaluated the *teg* block presence or absence in relationship to the genomic presence or absence of CRISPR-Cas systems in the isolates. Significantly, the "*teg* block" is found predominantly among strains with inactive/absent CRISPR-Cas systems (9/10 strains, Fig. 7c, Welch t-test, p value = 0.038). Altogether, these results show that the *teg* block, a virulence-inhibiting viral-like accessory genome element, is restricted by CRISPR-Cas systems, as indicated by the pangenomic presence of spacers targeting it, and its predominant presence in strains without active CRISPR-Cas systems.

Active CRISPR-Cas systems positively but indirectly correlate with *P. aeruginosa* virulence

Extending our analysis beyond the *teg* block, we analyzed the overall statistical features of the spacer-targeted genes. The statistical distribution of the gene association statistic (p value of the LR test) revealed that the set of spacer-targeted genes, associates preferentially with lower virulence, when compared to not spacer-targeted genes (Fig. 8a, two-sample K-S test, p value 7×10^{-12}). Furthermore, the statistical distribution of spacer-targeted genes separated by their affiliation to higher or lower virulence also differs significantly (Fig. 8b, two sample K-S test, p value 2.2×10^{-16}), and this difference in the distributions remains upon removal of the *teg* loci from the comparison (two sample K-S test, p value 2.2×10^{-16}). Altogether, these results suggest that spacer-targeted genes are enriched in their association with lower virulence, and this enrichment is driven by a plethora of gene associations, in addition to those of the *teg* genes. Moreover, we anticipate that association studies using larger isolate collections should allow better resolution of the individual gene association scores, and may assist in identification of additional spacer-targeted LVA genes.

Since we observed that elements of the *P. aeruginosa* accessory genome can negatively associate with virulence, and such elements can be restricted by bacterial CRISPR-Cas systems, we used gene association analysis to test for the association of virulence against *C. elegans* with the presence or absence of restriction-modification

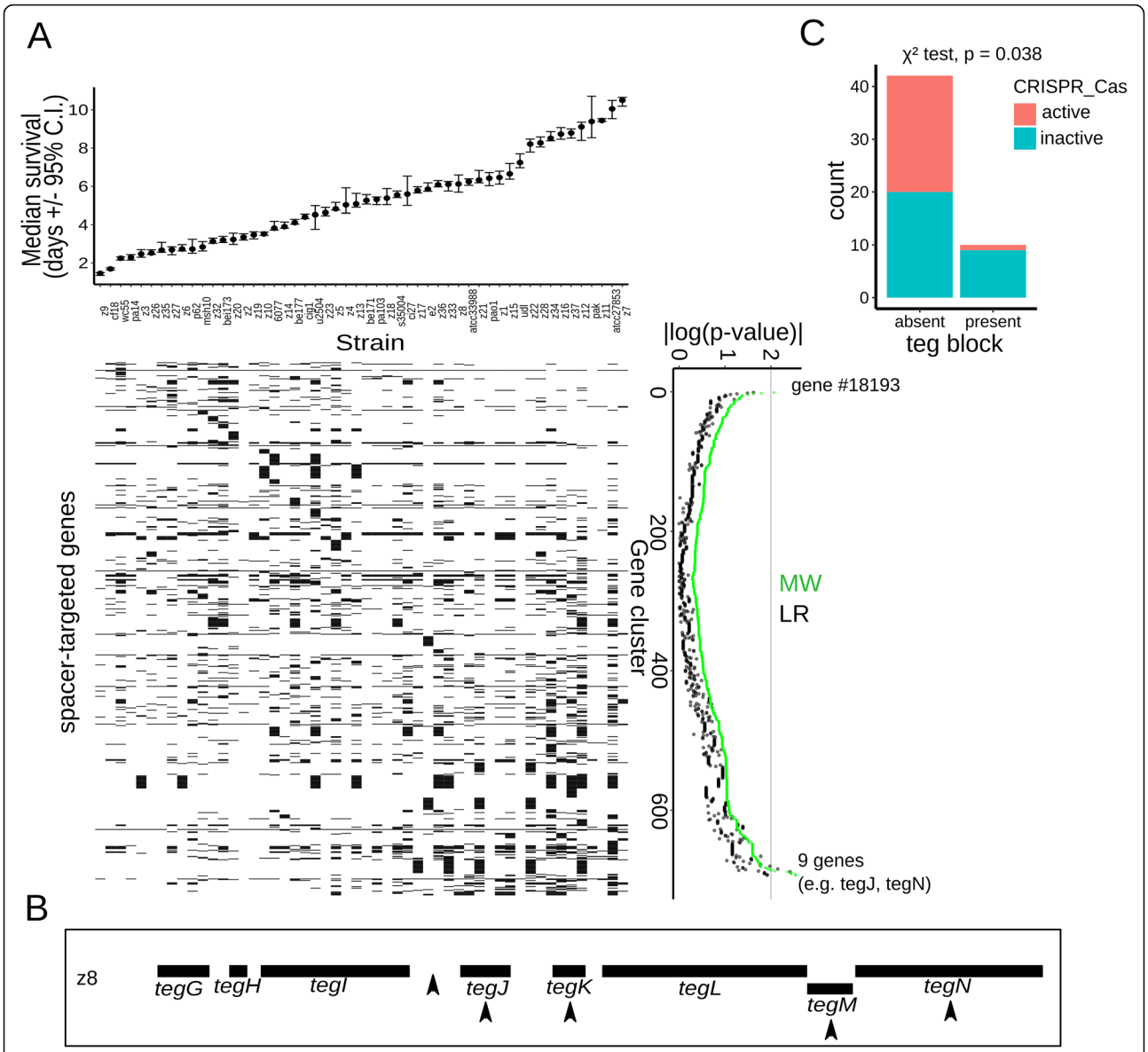

Fig. 7 Restriction of the *teg* block by spacers and correlation with CRISPR-Cas systems. **a** Top panel: median survival of adult *C. elegans* worms exposed to the studied collection of *P. aeruginosa* strains (with 95% confidence interval, C.I.). The strains are ordered from high to low virulence (left to right) and aligned with the matrix below. Bottom left panel: gene presence/absence matrix for genes targeted by CRISPR spacers. Gene presence is indicated with black squares and absence with white squares. Genes (rows) are aligned with the corresponding *p* values. Bottom right panel: Association statistics (*p* value of MW and LR tests) for the CRISPR-targeted genes, shown as $|\log_{10}(\text{pval})|$. Rows are ordered from association with high virulence to association with low virulence. **b** Schematic of the *teg* block in strain z8. Black boxes indicate *teg* genes and arrowheads spacers that target the element. **c** Number of strains (i.e., count) where the *teg* block is present or absent in relationship with the status of the host CRISPR-Cas system (active in red color, inactive in cyan color). The *p* value of a chi-square test is indicated

(RM) systems, CRISPR-Cas systems, and a recently identified cohort of ten novel defense systems [38]. These kinds of defense systems are widely distributed in bacteria and display innate (RM systems) or adaptive immune characteristics (CRISPR-Cas systems). We first analyzed adaptive immune systems on the premise that these systems may be able to selectively filter out deleterious genetic elements.

Type I CRISPR-Cas systems (Cas proteins and spacer arrays) are present in 71% of the 52 strains (37/52 strains; Additional file 2: Table S1) and belong to three different subtypes, which can be absent/present independently of each other: type I-F (73%), type I-E (35%), and I-C (21%). This distribution of CRISPR-Cas systems is consistent and similar to previous surveys of *P. aeruginosa* CRISPR-Cas systems [15]. In addition to the genomic presence of

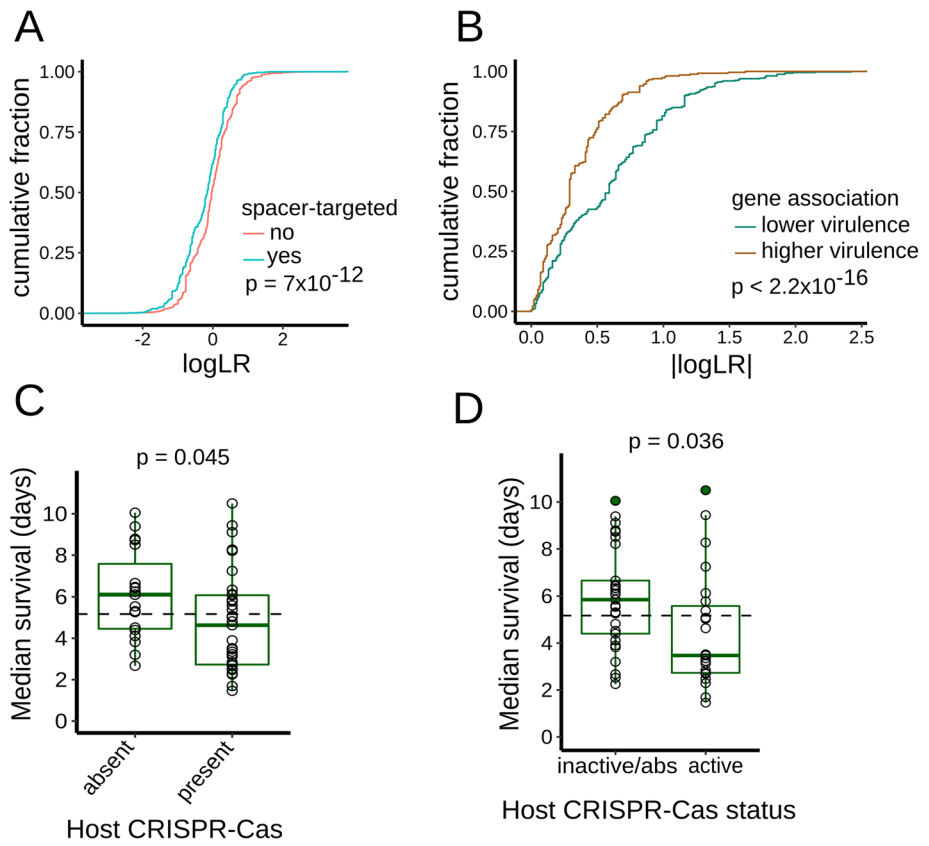

Fig. 8 Relationship between virulence and spacer-targeted genes and CRISPR-Cas systems of *P. aeruginosa*. **a** Cumulative distributions for the association statistic (log *p* value of LR test). Genes in the accessory genome are partitioned according to whether they are targeted by spacers (in cyan color) or not (in red color). The *p* value of two-sample K-S test is indicated. **b** Cumulative distributions for the association statistic (log *p* value of LR test). Spacer-targeted genes are partitioned according to whether they are associated with higher (in brown color) or lower (in green color) virulence. The *p* value of two-sample K-S test is indicated. **c**, **d** Box-and-whisker plots of worm median survival in relationship with CRISPR-Cas presence/absence and activity status. **c** Strains are partitioned according to the presence/absence of host CRISPR-Cas systems (I-E, I-F). **d** Strains are displayed according to their CRISPR-Cas status in two categories: active or inactive-absent (inactive/absent). The median virulence of the complete set of strains displayed on each graph is indicated with the dashed horizontal line. *p* values are indicated for the Welch *t* test comparison of virulence between the two groups represented (**c**, **d**)

CRISPR-Cas loci, we also investigated if the identified CRISPR-Cas systems were predicted to be active or inactive based on the presence/absence of known anti-CRISPR genes. Anti-CRISPR proteins are virus-encoded and can inhibit CRISPR-Cas systems, blocking their immune function (reviewed in [39]). We identified a set of 22 anti-CRISPR gene families in 31% of the 52 *P. aeruginosa* genomes and cataloged each strain's CRISPR-Cas status as (1) "active" if it has at least one CRISPR-Cas system with no known cognate anti-CRISPR gene present in genome or (2) having an "inactive/absent" system if CRISPR-Cas is absent or where cognate anti-CRISPR gene(s) are found concomitantly with CRISPR-Cas (Additional file 2: Table S1). We compared the above anti-CRISPR approach for identifying strains with inactive CRISPR/Cas to an alternative criterion: the presence in the same bacterial genome, of a CRISPR-Cas spacer with its DNA target, a condition referred to as spacer "self-targeting" [37] and obtained similar results (see the "Methods" section).

Next, we analyzed the CRISPR-Cas systems in relationship to virulence. We first considered separately the subtypes I-F, I-E, I-C, and their combinations (Additional file 1: Figure S5A). Strains with type I-C CRISPR-Cas systems showed lower virulence compared to that of all other strains (Welch *t* test, *p* value = 0.03). The distinct association observed for I-C systems coincides with the fact that *P. aeruginosa* type I-C CRISPR-Cas systems have been exclusively found inside pKLC102-like ICEs [15]. Defense systems inside ICEs, such as type I-C CRISPR-Cas systems, likely fulfill a primary role in the ICE's lifecycle and may provide minor to negligible immune protection to the bacterial host. Based on the distinct I-C association with virulence and their ownership by ICEs, we did not consider I-C systems part of *P. aeruginosa* complement of immune systems, and so in subsequent analysis, we considered only subtypes I-E and I-F as comprising the bacterial cell's CRISPR-Cas systems.

Interestingly, we found that the presence of a host CRISPR-Cas system (i.e., either subtypes I-E or I-F), significantly associates with higher virulence (Fig. 8c, Welch t test, $p = 0.045$). To investigate if this association is related to the immune function of CRISPR-Cas systems, we considered the status of activity of the host CRISPR-Cas systems. Notably, the presence of active CRISPR-Cas systems (by the criterion of absence of anti-CRISPR genes) also statistically correlates with increased virulence (Fig. 8d, two-sided Welch t test, $p = 0.036$). Moreover, upon inclusion of strains with spacer self-targeting to the "inactive" strain set, the statistical association between active CRISPR-Cas and higher virulence is maintained (one-sided Welch t test, $p = 0.038$). To further investigate the relationship between CRISPR-Cas and virulence, we applied an alternative analysis. The survival curves for the strain collection were pooled, forming two groups based on the presence or absence of CRISPR-Cas in the isolates. The survival curves between these two groups differ significantly (Additional file 1: Figure S5B, K-M method, logrank test, p value $< 2 \times 10^{-16}$), and the strain group with CRISPR-Cas systems has a lower median survival (4.2 days, 95% C.I. 4.0–4.4 days) compared to the group without this defense system (median survival of 6.5 days, 95% C.I. 6.3–6.6 days).

The association of active CRISPR-Cas systems with high virulence suggested a positive role for this immune system in the maintenance of virulence. Thus, we explored whether or not CRISPR-Cas could have a direct role in virulence. First, we constructed a deletion of the entire six Cas genes of strain PA14 (strain PA14ΔCas) to abolish CRISPR-Cas activity, but we observed no significant difference in virulence between the PA14ΔCas and wildtype PA14 (Additional file 1: Figure S5C). In addition, we tested if the Cas proteins have the ability to modulate virulence when expressed from a plasmid in strain PAO1 that lacks CRISPR-Cas. The PAO1 strain expressing CRISPR/Cas from a plasmid (strain PAO1 p (Cas^+) displayed no significant difference in virulence compared to PAO1 expressing a plasmid control (p (control)) (Additional file 1: Figure S5D). In summary, these results indicate that CRISPR-Cas is neither necessary nor sufficient to directly modulate bacterial virulence, at least under the assayed laboratory conditions.

We next proceeded to analyze known and presumed innate immune systems of *P. aeruginosa*: RM systems [40] and the cohort of ten novel defense systems [38], respectively. We identified RM systems based on annotations from the REBASE database [40] (Additional file 2: Table S1). We observed a weak association between the total number of RM systems and virulence (Additional file 1: Figure S6A, spearman rank correlation, rho: 0.25) that does not reach significance ($p = 0.08$). Similarly, the relationship between each separate RM system type and virulence shows weak association for the types I and II, while the association for type III and IV RM systems cannot be reliably assessed (Additional file 1: Figure S6). None of the abovementioned correlations reached statistical significance (all p values ≥ 0.08).

Next, we evaluated the presence of ten novel defense systems [38] by homology of the system's diagnostic proteins to genes in our strain collection (Additional file 2: Table S1). We found no statistically significant association with virulence for any of the novel immune systems (Additional file 1: Figure S7). Similarly, we observed no association between the overall number of novel defense systems per strain and virulence (spearman rank correlation, rho: 0.03, $p = 0.81$, Additional file 8: Figure S7). These results show that the presence or absence of the recently identified immune systems bears no apparent relationship with strain virulence. Interestingly, we noted that the gabija system of strain PA14 (genes PA14_60070 and PA14_60080) and strain CF18 (genes #2421 and ID #Q002_01766) are found inside ICEs: PAPI-1 [41] for PA14, and an unnamed ICE (predicted with ICEfinder [42]) for CF18. Altogether, these observations highlight that ICEs can harbor multiple defense systems, as previously exemplified with type I-C CRISPR-Cas systems.

To summarize this section, we found that RM and novel defense systems have a weak or no significant relationship with virulence. In contrast, the presence and activity of CRISPR-Cas systems associates with higher virulence. The statistical association between active CRISPR-Cas systems and *P. aeruginosa* virulence suggests that CRISPR-Cas activity may indirectly affect virulence-related phenotypes, most likely by regulating acquisition and/or retention of accessory genome virulence factors and other elements that impinge on virulence. A verified instance of such CRISPR-Cas-mediated restriction process is exemplified by the *teg* block. Moreover, the statistical distribution of the gene association statistic for the spacer-targeted genes suggest the possibility that additional restricted LVA genes may be identified in more powerful association studies.

Discussion and conclusions

In the present study, we investigated bacterial-driven variation in the interactions between *C. elegans* and *P. aeruginosa*. Fifty-two *P. aeruginosa* wild isolate strains were found to cover a wide virulence range, spanning from highly virulent strains, which induce a worm median survival of 1.5 days (~ 11% of their lifespan under standard conditions at 25 °C) to strains with almost no virulence, which induce worm lifetimes similar to those observed with non-pathogenic *E. coli* HB101, and which do not affect progeny production.

We posit that bacterial strain variation in virulence towards *C. elegans* reflects adaptations of *P. aeruginosa* to

its natural niches. In natural settings, virulence may be a character under selection by the frequency with which predators are deterred by virulence mechanisms, and/or by the extent to which the bacterium depends on infection of predator hosts for population growth. It should be noted that because *P. aeruginosa* is a multi-host pathogen of many species, including insects and single-celled eukaryotes, as well as nematodes, we cannot say with any certainty whether any of the *P. aeruginosa* strains chosen for this study have undergone selection in the wild through direct interaction with *C. elegans*. We observed that among our 52-strain panel, environmental strain isolates exhibited on average greater virulence against *C. elegans* than did clinical isolates (Additional file 1: Figure S1B), consistent with previous findings [43]. This suggests that some of the strain variation in virulence against *C. elegans* could be influenced by adaptations of *P. aeruginosa* to its pathogenic association with humans, and that such adaptations may not necessarily confer pathogenic benefit against *C. elegans*. The virulence of clinical isolates could reflect genetic and genomic makeup of the bacterium that is favorable in the context of human immune responses and/or therapeutic antibiotics. Indeed, among the genes associated with virulence, we observed several genes involved with antibiotic resistance, such as *mexZ*, a negative regulator of the *mexXY* bacterial efflux pump [31, 32] and *arr*, which functions to induce biofilms in response to aminoglycoside exposure [44].

The variation in virulence among *P. aeruginosa* strains parallels the substantial genomic diversity of this bacterial species. *P. aeruginosa* strains contain relatively large genomes for a prokaryote (5–7 Mb; 5000–7000 genes) with a sizable contribution of accessory genome elements (Fig. 1). Our data show that strain variation in *P. aeruginosa* virulence is mediated by specific accessory genome elements (Figs. 3 and 4), in combination with the core genome, including previously described *P. aeruginosa* virulence-related factors (Fig. 4). Notably, we find particular accessory genome elements that contribute to increased virulence, and others that promote decreased virulence (Figs. 3 and 5). The existence of genes whose functions lead to the negative regulation of virulence (for example, *qsrO* and *tegN*) suggests (1) strain adaptations to niches where capping virulence is advantageous, either for environmental reasons (e.g., infrequent bacterial predators or hosts for bacteria to feed on) or for clinical reasons (e.g., evasion of immune surveillance at lower virulence), and (2) detrimental effects of MGEs (e.g., *teg* block) that are chromosome integrated and likely engage into parasitic relationship with its bacterial host.

The results of our genetic analysis of HVA and LVA genes indicate a direct role for a subset of these genes in modulating virulence, whereas for other HVA and LVA genes our genetic results do not support a direct role. A direct role in virulence for genes PA14_27700, PA14_27680, *pslK*, and *pslM* was expected based on previous findings (Fig. 4), and hence, their identification as HVA genes supports our comparative genomics approach. For 11 LVA genes that we tested genetically, the results suggest a direct contribution for *qsrO* and *tegN* to virulence (Fig. 5). On the other hand, genetic ablation (for t*egG* to *tegN* and *ghlO*) or ectopic expression of *mexZ*, *tegL*, *tegM*, *ghlO* (Fig. 5, Additional file 1: Figure S4), or the *Cas* genes (Additional file 1: Figure S5) did not measurably alter virulence. Importantly, associated genes can be evolutionarily gained or lost as multigene units—physical blocks with defined sets of accompanying neighboring genes. Genes in such blocks all show association with virulence by being linked to a bona-fide virulence modifier gene, even though some may not have direct effects on virulence. This situation is exemplified by the *teg* block that comprises 8 LVA genes (Fig. 6), including one that affects virulence (i.e., *tegN*) and others that do not (i.e., *tegL, tegM*).

What could account for why certain genes would not exhibit essential virulence functions in genetic tests, despite being correlated with virulence in gene association analysis? One possibility could be statistical false discoveries. However, we assessed the reliability of our statistical analysis in two ways: by using permutation-based testing to filter out false discoveries and by employing phylogenetically aware scoring approaches to control for any confounding effect mediated by population structure.

It is also possible that some of the genes that tested negatively in the genetic tests actually do function in some contexts as bona fide virulence factors, but their effects could be masked by epistasis in the genomic background of the particular strains in which we conducted our loss-of-function and gain-of-function tests. The possibility of such strain-specific epistasis could be investigated by conducting parallel genetic tests for the full cohort of relevant strains.

This study shows that genome-wide association (GWAS) analysis of a panel of genomically diverse strains of a bacterial species can identify previously unrecognized accessory genome elements influencing a phenotype of interest, in this case virulence of *P. aeruginosa* against the invertebrate bacterivore *C. elegans*. What sorts of genetic bases for virulence variation might have been missed in our study? First, some of the accessory genome genes that scored below statistical cutoffs in our study might emerge as high-confidence candidate virulence modulators from studies of larger and/or more diverse panels of bacterial strains. It should also be noted that our gene-association analysis scored for the presence or absence of intact (accessory genome) genes. We did not attempt to test for association of virulence with amino acid coding mutations or with non-coding sequence polymorphisms that could alter *cis*-regulatory regulation of direct virulence modulators. Such higher

resolution (GWAS) analysis could be the basis for future inquiries.

Our analysis of the *teg* block illustrates that LVA genes can reside within MGEs that decrease virulence (Fig. 5) and that are restricted by host CRISPR-Cas systems (Figs. 6 and 7). The *teg* block is likely not the only MGE with a negative association to virulence, because the cohort of spacer-targeted genes shows an overall enriched association with lower virulence (Fig. 8a, b). We thus suggest that additional MGEs, detrimental for virulence and CRISPR-Cas restricted, could be unveiled utilizing more powerful association studies with enlarged isolate collections.

We observe a positive correlation between the virulence of *P. aeruginosa* strains against *C. elegans* and the presence of CRISPR-Cas bacterial immunity (Fig. 8c, d), even though our genetic tests with CRISPR-Cas loss-of-function mutants or ectopic expression indicate that CRISPR-Cas activity is neither necessary nor sufficient for increased virulence (Additional file 1: Figure S5C-D). This suggests that bacterial adaptive immunity and anti-predator virulence may be somehow indirectly coupled via the effects of physiological, ecological, and/or evolutionary factors.

Although there are undoubtedly numerous potential underlying causes for a linkage between CRISPR-Cas and virulence, two broad classes of potential scenarios are suggested. One scenario is based on possibility that the evolution of accessory genomes is highly influenced by bacterial restriction systems, such as CRISPR-Cas that function to limit horizontal gene transfer (HGT) and thereby help shape the makeup of the accessory genome. Our finding that accessory genome elements can modulate virulence supports the supposition that bacterial immune systems could indirectly contribute to the maintenance or evolvability of virulence towards invertebrate predators such as *C. elegans.* This scenario is further supported by our findings that *P. aeruginosa* genes associated with low virulence include detrimental viral-like mobile genetic elements and are more enriched for targeting by CRISPR-Cas spacers that are those associated with higher virulence. A second scenario, not mutually exclusive with the first, is based on the fact that bacterial restriction systems such as CRISPR-Cas are themselves often part of the accessory genome, as evidenced in the case of *P. aeruginosa* by the fact that some strains contain one or more CRISPR-Cas loci, while other strains contain none. Apparently, CRISPR-Cas adaptive immunity is selected for or against, depending on particular environmental conditions. Therefore, high virulence and the capacity to restrict HGT could be linked by the co-occurrence of environmental factors that simultaneously select for both features. For example, in certain *P. aeruginosa* natural habitats, abundant predation by invertebrates such as *C. elegans* may commonly co-occur with pressure from an abundance of phages. Conversely, clinical settings may frequently present conditions that simultaneously disfavor high virulence and restriction of HGT. Testing of these hypotheses will benefit from further studies.

Unlike CRIPSR-Cas, we did not observe a similar association of virulence with other restriction systems, including restriction/modification (RM) and a set of recently identified restriction systems of less well-characterized mechanisms [38]. These other systems, particularly the RM systems, differ from CRISPR-Cas fundamentally in that they are not adaptive immune systems, and hence, they would tend to limit uptake of foreign DNA elements regardless of whether those elements confer positive or negative phenotypes. CRISPR-Cas systems are much more discriminatory: Restriction of an element by CRIPSR-Cas requires programming the spacer array with a sequence from the targeted element, enabling selection for targeting of deleterious elements, and selection against targeting of advantageous elements. Thus, the association that we observe between virulence and CRISPR-Cas may reflect such selection for restriction of uptake of elements that are particularly deleterious in the context of anti-predator virulence.

Methods

C. elegans worm strains

The *C. elegans* N2 strain was used as wildtype strain. In addition, strain KU25: *pmk-1(ku25)*, referred to as *pmk-1(lf)*, was used for some virulence assays. All nematode strains were maintained using standard methods on NGM plates [45] and fed with *E. coli* HB101.

Bacterial strains

The *P. aeruginosa* strains were routinely grown on LB media at 37 °C without antibiotics, unless otherwise noted. A list of the 52 bacterial isolates established as our experimental panel is listed in Additional file 2: Table S1. The collection was assembled using strains procured from numerous distinct sources, and although we strove to obtain a diverse collection of both environmental and clinical stains, there was limited control over the collection composition with regard to specific features. The description and genotypes of bacterial strains constructed in the present study are listed in Additional file 6: Table S5. For a portion of the strains in the collection, we found that genetic manipulation is limited, because a considerable fraction of the isolates exhibit strong restriction to uptaked DNA or high levels of resistance to antibiotics.

Worm survival assays (virulence assays)

Worm survival assays (virulence assays) were performed using slow killing (SK) conditions [8]. Briefly, an aliquot of an overnight liquid LB culture of each *P. aeruginosa* strain was plated on SK agar plates. The bacterial lawn was spread to cover the entire surface of the agar, to prevent worms from easily escaping the bacterial lawn. The plates were incubated at 37 °C for 24 h and then at

25 °C for 24 h, to allow growth of the lawn and the induction of pathogenic activity [8]. Prior to use, FUDR (100 ng/μL) was added to the plates to a final concentration in the agar medium of 300 μM. A synchronous population of young adult (YA) hermaphrodite N2 worms was prepared by standard hypochlorite treatment, followed by culture of larvae from L1 stage to YA stage on NGM agar seeded with *E. coli* HB101. The young adult (YA) worms were then transferred to the SK plates to initiate their exposure to *P. aeruginosa* lawns. The time-course of death of the worms on each plate was determined with the aid of a "lifespan machine" [23], an automated system based on a modified flatbed scanner. A minimum of 3 plates of worms were scanned per isolate, total median $n = 84$ (Additional file 2: Table S1). Image analysis was optimized to fit the *P. aeruginosa* slow killing conditions as described previously [46]. The collected survival information was manually curated and analyzed using R (i.e., survminer package) with the Kaplan-Meier (K-M) method. K-M was used to estimate median survival and its confidence interval. The K-M based estimate of the "median survival" of worms exposed to a particular bacterial isolate corresponds to our measure of bacterial virulence. The semiparametric Cox proportional hazards model is not applicable to the obtained survival information, as the proportional-hazards (PH) assumption does not hold (R "survival" package, proportional hazards test, global p value = 0; p value < 0.05 for 15 strains).

In the alternative analysis of the survival data to study the relationship of virulence to CRISPR-Cas, the survival data (i.e., individual worm lifespans) of all strains with host CRISPR-Cas systems was aggregated into a first group (n = 2656), and the survival data for strains without host CRISPR-Cas systems was aggregated into a second group (n = 1549). The aggregated data was analyzed using R (i.e., survminer package) with the Kaplan-Meier (K-M) method.

To assess the accuracy of the above semi-automated method for determination of survival curves, the survival curves generated by the lifespan machine were compared to manually obtained survival curves for four strains of varied virulence and no appreciable difference was observed between lifespans determined automatically compared to manually (Additional file 1: Figure S8). Virulence assays that involved the use of plasmid-carrying bacterial strains were performed on SK plates supplemented with 20 μM gentamicin.

Generation of mutant and transgenic *P. aeruginosa* strains

Generation of PA14 strains

A PA14*Δcas* in-frame deletion mutant was constructed using a method described previously [47] that employed a sequence that contained regions immediately flanking the coding sequence of the *cas* genes. This fragment was generated by a standard 3-step PCR protocol using Phusion DNA polymerase (New England Biolabs) and then cloned into the *Xba*I and *Hind*III sites of pEX18A [48], resulting in plasmid pEX18-*CIF*. pEX18-*CIF* was used to introduce the deleted region into the wildtype PA14 strain (RRID:WB-STRAIN:PA14) by homologous recombination. *Escherichia coli* strain SM10 pir was used for triparental mating. The deletion of the Cas genes was confirmed by PCR. For the expression of Cas genes in PAO1, the *P. aeruginosa* PA14 *cas* genes were cloned into the *Hind*III and *Xba*I sites of pUCP19 [49], creating plasmids pUCP-*cas* (referred to as p (Cas+)). The resulting plasmid was transformed into *P. aeruginosa* PAO1 by electroporation to generate the strain PAO1 p (Cas^+).

Generation of z8 strains

Gene deletions in the z8 strain were obtained using the endogenous type I-F CRISPR-Cas present in this strain. The gentamicin selectable plasmid pAB01 was modified to introduce a spacer targeting the gene of interest and also a homologous recombination (HR) template with arms flanking the genomic region to be deleted (600–800 bp homology arms). The corresponding plasmid so obtained is referred to as "editing plasmid." The cloning of spacer sequences was performed with the restriction enzyme ligation method. The pAB01 plasmid (pHERD30T backbone with the I-F repeat-spacer-repeat sequence: 5′-GTT CAC TGC CGT GTA GGC AGC TAA GAA AGT CTT CAG TTC TCT GGA AGC TCA AAG AAG ACG TTC ACT GCC GTG TAG GCA GCT AAG AAA-3′ incorporated into MCS) was digested with BbsI enzyme. An insert fragment with the gene-targeting spacer (32 nt) was assembled by annealing of two complementary oligos, extended to seal the BbsI site (e.g., insert spacer targeting *teg* block: 5′-aag aaa GGG GGA TGC GTT CTC GAC ACG AGT AAC CAT Cggt-3′ and 5′-gtg aac CGA TGG TTA CTC GTG TCG AGA ACG CAT CCC Cct t-3′).

Cloning of HR sequences was performed with the Gibson assembly method. The HR arms were PCR amplified from bacterial genomic DNA and incorporated into the NheI site of the pAB01 vector. The sequence of plasmid pAVR85 (used for *teg* gene block editing) is provided in Additional file 10: Table S9 as an example.

The z8 bacterial cells were washed twice with 300 mM sucrose and subjected to electroporation (800 ng of editing plasmid, 2 mm gap width cuvettes, 200 Ω, 25 μF, 2500 V using a Gene Pulser XCell machine (Bio-Rad)). All steps were performed at room temperature. Transformants were selected on LB plates with gentamicin 50 μg/mL. Transformant colonies were re-streaked in LB Gentamicin plates and genotyped by PCR. After obtaining the desired genomic modification, the editing plasmid was cured by passage of the strain in liquid LB culture without antibiotic. Plasmid pHERD30T (gentamicin selectable) was

used for the expression of genes associated with virulence; gene(s) of interest (with surrounding regulatory sequences) were cloned using Gibson assembly.

Bacterial growth rates

A random subset of 33 strains that span the virulence range was used to determine bacterial growth rates. Overnight cultures of each strain (20 μl, O.D. = 1.5–2) were inoculated into 180 μl of LB medium in 96-well plates. The optical densities at 650 nm were measured using the SpectraMax 340 microplate reader (Molecular Devices, CA, USA) every 15 min for 33 h. The experiment was performed at 25 °C, the same temperature used for the worm assays, and the plates were shaken for 5 s before the measurements by the plate reader to allow aeration. The Softmax Pro 6.2.1 (Molecular devices, CA, USA) software was used to analyze the data. Specific growth rates (μ) were calculated based on the exponential phase of the growth curves. The μ values were calculated using the following formula: $OD = N\ e^{\mu t}$ where OD is the measured optical density, N the initial optical density, and t the time.

Genomic analysis of *P. aeruginosa* strains

A full list of *P. aeruginosa* species, consisting of 1734 strains, was downloaded from RefSeq database [50] (on December 2016). In addition, the corresponding annotation files that include (1) genomic sequences, (2) nucleotide and (3) protein sequences for coding genes, and (4) feature tables were downloaded from the RefSeq database as well. Next, several filtration steps were applied to remove strains that (1) had no proper 16S rRNA annotations (missing sequence, or sequence that is shorter than 1000 nts, or sequence that showed less than 80% identity to PA14 16S rRNA) and (2) contained more than 100 core genes with multiple members or were missing more than 15% of the core genes. The second filter was applied after one round of clustering with CD-HIT [14] and identification of core genes (see details below). This process resulted in a final set of 1488 strains (Additional file 8: Table S7).

Clustering analysis of *P. aeruginosa* coding sequences

The protein sequences of 1488 strains (obtained from the RefSeq database ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/) were clustered using CD-HIT (v4.6.5), with the following settings -c 0.70 -n 5 -g 1 -p 1. The procedure yielded 23,793 clusters of homologous genes. The output of the clustering analysis was post-processed to generate a statistical report that lists for each cluster (i.e., each homologous gene) the representative sequence, its function, the total number of occurrences of the gene across the full set of 1488 strains, and the number of strains that contain at least one copy of the gene. A presence/absence matrix for each gene across 1488 strains was generated. In addition to the full matrix, a presence/absence matrix for the collection of 52 experimentally studied strains was extracted. Gene clusters that had no representatives in these 52 strains were removed, resulting in a matrix with 11,731 genes (Additional file 9: Table S8).

Phylogenetic analysis

Core-genes across the 1488 strains were defined as genes present in more than 90% of the strains in a single copy only (resulted in 3494 core-genes). For each cluster representing a core gene, the following steps were applied: the corresponding DNA sequences were aligned using MAFFT default parameters (version 7.273) [51]; gblocks (ver 0.91b) [52] was applied on the alignment to remove poorly aligned positions (with parameters -t = d -b5 = a); an in-house code was used to remove all the invariant positions (excluding gaps); the alignments were padded with gaps for strains in which the core gene was missing. All the alignments were then concatenated to a final alignment of 523,361 nucleotides. The program FastTree [53], version 2.1, with settings: -gtr, was then used to generate the phylogenetic tree of the 1488 strains. The recombination-aware approach, ClonalFrameML [54], was used to reconstruct the phylogenetic tree with corrected branch lengths. The input to the method was the tree generated by the FastTree program and the multi-fasta alignments of 3494 core-genes. The interactive Tree of Life web-based tool [55] was used for visualization of the resulting phylogenetic tree. Information about MLST, source (clinical/environmental), and strains that are part of the experimental collection was incorporated into the tree view. A phylogenetic tree of the 52 experimentally studied strains was extracted from the ClonalFrameML phylogenetic tree of the 1488 strains using the "ape" package in R.

Statistical test for association of genetic elements (coding/non-coding genes) with virulence

The Mann-Whitney (MW) ranking test and linear-regression (LR) analysis were applied to every gene to test the association of the presence/absence pattern with virulence. Genes were considered associated if both tests yielded a p value lower than 0.05, and at least one of the tests yielded a p value smaller than 0.01. Among the virulence-associated genes, genes with negative slope (based on linear regression) were associated with low survival/high virulence (referred to as high-virulence associated or HVA), while genes with positive slope were associated with high survival/low virulence (referred to as low virulence associated or LVA). All the p values are shown in log10 scale as absolute values. The control for multiple hypothesis testing was performed using a permutation test as described below.

Permutation test to control for multiple hypothesis testing

Ten thousand permutations of the virulence values and their assignment to strains were generated (i.e., median worm survival values), and the MW and LR association tests were repeated for each permutation. Then, for each gene, the number of times that it received a better p value using the shuffled virulence data compared to the original one was recorded, separately for MW and LR. The above count was divided by 10,000 to obtain the permutation corrected p value for the MW and LR tests. The MW and LR p values were considered significant if their corresponding corrected p value was lower than 0.05.

Assessment of confounding effects due to population structure

The phylogenetic method reported by Collins and Didelot [25], known as treeWas, was used to address the potential influence of population structure in the statistical association between accessory genes and virulence. The method was applied on the input consisting of (1) 11,731 gene clusters presence/absence matrix, (2) median survival vector, and (3) ClonalFrameML phylogenetic tree of the 52 strains. The method returns as output three types of scores and their corresponding p values for every gene cluster: (1) "Terminal Score" which measures sample-wide association between genotype (gene presence) and phenotype (median survival), without relying on the phylogenetic tree; (2) "Simultaneous Score" which measures the degree of simultaneous change in the phenotype and genotype across branches of the phylogeny; and (3) "Subsequent Score" which measures the proportion of the tree in which genotype and phenotype co-exist. The computed scores were considered significant if their p values < 0.05 (Additional file 3: Table S2).

Collection of known non-coding RNA (ncRNA) in *P. aeruginosa*

The collection of ncRNAs (excluding rRNAs and tRNAs) in *P. aeruginosa* was constructed using two resources: RFAM 12.2 [56] and RefSeq annotations [50]. First, 75 non-coding RNA families were extracted from RFAM, with a total of 1363 sequences across *P. aeruginosa* strains. To get the representative sequences (there could be more than one) for each family, the sequences of each family were clustered using CD-HIT-est (with 80% identity). This analysis resulted in 115 sequences (representing 75 different ncRNA families). Second, using RefSeq annotations of the 1488 strains, 2549 ncRNA sequences were extracted. Altogether, our collection comprised of 83 ncRNA families, represented by 123 sequences. Finally, the collection of the 123 sequences was blasted against the 1488 genomic sequences, and a presence/absence matrix for each of the sequences in all the strains was generated. Rows that represent sequence members from the same family were collapsed, resulting in matrix with 83 rows.

Collection of previously identified virulence genes in *P. aeruginosa*

A list of virulence genes, in either PA14 or PAO1, was downloaded from [57]. The list was filtered to contain only genes that were reported to contribute to *P. aeruginosa* virulence towards *C. elegans*, resulting in 56 genes. Another four genes were added based on the publication [30]. The homologous gene clusters that contained the above genes were marked as virulence genes. The full list of 60 virulence genes is found in Additional file 4: Table S3.

Analysis of CRISPR-Cas systems

Identification of CRISPR-Cas systems

The presence of CRISPR-Cas systems in the genomes of our *P. aeruginosa* collection was determined by identifying the gene clusters that encode for Cas proteins.

Identification of anti-CRISPR genes

The most up to date collection of anti-CRISPR genes was downloaded from [58], consisting of 41 sequences (https://tinyurl.com/anti-CRISPR). Annotations (e.g., CRISPR-Cas subtype inhibited) for each sequence were maintained. The representative sequences of the clusters of homologous genes (see CD-HIT clustering above) were blasted against the anti-CRISPR sequences using blastp [59] and *e*-value threshold of e^{-10}. A coverage of more than 35% of the anti-CRISPR sequence was considered a hit.

Determination of active/inactive systems

The annotation on the type of CRISPR-Cas system(s) that is inhibited by each anti-CRISPR protein was used to define CRISPR-Cas activity. The type(s) of CRISPR-Cas systems of every strain were matched to the type(s) inhibited by the anti-CRISPR genes present in the same genome. Strains where all present CRISPR-Cas system(s) are inhibited by type-matching anti-CRISPR proteins were considered inactive.

A second approach to determine active/inactive systems was compared to the method above. The presence in the same genome of a CRISPR-Cas locus and one or more self-targeting spacers is considered to reflect an inactive effector status of that CRISPR-Cas locus, because genome cleavage by an active CRISPR-Cas system is expected to be lethal to the bacterial cell [60, 61]. In our collection, we found 11 strains with CRISPR-Cas and at least one self-targeting spacer with a full match to its genomic target (Additional file 2: Table S1). Most of these strains (9 out of 11, corresponding to 82% of them) were included in the set of inactive strains by the anti-CRISPR approach. The determination of CRISPR-Cas

"inactivity" with the two approaches is highly similar (McNemar's chi-squared test, p value = 1).

CRISPR spacer arrays collection

The collection of CRISPR spacer sequences across all 1488 strains was generated by applying the CRISPR Recognition Tool (CRT1.2-CLI.jar) [62] on genomic sequences, with default parameters. Since the tool works only with single fasta records, the genomic sequences (contigs and scaffolds) of each strain were merged before the application of the tool, and then, the results were mapped back to the original sequences using an in-house code. A total of 35,340 spacer sequences were identified (some sequences were present more than once in the collection) with 94% of spacer sequences in the length range of 32–34 nucleotides.

Targets of CRISPR spacers on P. aeruginosa pangenome

The program blastn [59], with default parameters, was used to identify matches for the full spacer's collection against the DNA sequences of all protein coding genes. Blast hits in which the alignment of the spacer query started after position 2 or had less than 95% identity were filtered out. The homologous gene clusters that contained the targeted genes were marked as CRISPR targets. The above set of targets and spacers was further filtered, and spacers where its target is located in the same genome were tagged as "self-targeting" spacers. In order to use self-targeting spacers to estimate CRISPR-Cas "inactivity," an additional criterion was included: the target (protospacer) should be conductive to CRISPR-Cas cutting of the bacterial DNA, i.e., a full spacer-target alignment with PAM presence should exist. A strain was considered CRISPR-Cas "inactive" by the presence of a CRISPR-Cas locus and at least one spacer satisfying the above criterion.

Analysis of restriction modification (RM) systems

Sequences of RM systems and their type classification were downloaded from REBASE (The Restriction Enzyme Database) [40]. The representative sequences of the clusters of homologous genes (see CD-HIT clustering above) were blasted against the RM sequences using blastp and e-value threshold of e^{-10}. Several filtration steps were ten applied before marking a gene cluster as an RM gene. Gene clusters were excluded if (1) the coverage of the RM sequence by the representative sequence was less than 35%, (2) if the gene cluster represents a core gene, and (3) the function associated with the gene cluster is not diagnostic to an RM system (e.g., permease, topoisomerase). Two hundred twenty-seven gene clusters passed the criteria.

Next, the RM genes of every strain were extracted and re-ordered based on their genomic location. Using the location of the genes, "gene blocks" were determined as groups of genes separated by less than 8 intervening genes.

For every gene, the best matching RM component from REBASE was used to assign an RM type (either type I, II, III, or IV) and identity the RM component (methylase, nuclease, specificity factor, etc.). Every gene with a match to a type IV RM was established as a type IV system.

Next, all other RM systems (types I to III) were defined based on the presence of methylase genes. A gene singleton (i.e., not belonging to any gene block) matching a type II methylase was established as type II RM system. RM systems inside gene blocks were assigned based on the following criteria: (a) 1 or 2 methylases must be present per RM system and (b) all gene components of a given RM system, congruently match a single type of RM system. To assess the quality of our RM data, we compared our predictions to REBASE data. Seven strains from our collection have their genomes annotated in the REBASE website. Four strains have the exact same number of RM systems, while the RM count of the 3 remaining strains differ by one RM. No statistical difference exists between our method and REBASE with regard to the RM count of strains (chi-square test, $p = 0.18$).

Analysis of novel defense systems

Protein accession numbers belonging to ten novel defense systems were downloaded from [38] and were filtered to keep only *P. aeruginosa* proteins. Each protein sequence was annotated with system type and specific system component. The protein sequences were then extracted from RefSeq. The representative sequences of the clusters of homologous genes (see CD-HIT clustering above) were blasted against the protein sequences using blastp [59] and an e value threshold of e^{-10}. A filtration step was applied before marking a gene cluster as a defense system gene. Gene clusters were excluded if (1) the coverage of the defense system sequence by the representative sequence was less than 35%. Next, the candidate genes for novel defense systems of every strain were extracted and re-ordered based on their genomic location. Using the location of the genes, "gene blocks" were determined as groups of genes separated by less than 8 intervening genes. All novel defense systems were defined based on the presence of a set of 2 or more genes uniformly matching a variant of the novel systems as reported by [38].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-019-1890-1.

Additional file 1. Supplemental Figures S1 to S8.

Additional file 2: Table S1. Description and features of the experimentally studied collection of 52 *P. aeruginosa* strains. The 52

strains experimentally studied strains are listed, altogether with all the features derived from this study.

Additional file 3: Table S2. Genes significantly associated with virulence. Description of the 79 genes that comprise the HVA and LVA sets.

Additional file 4: Table S3. Known virulence genes in the interactions between *P. aeruginosa* and *C. elegans* under SK condition.

Additional file 5: Table S4. Nomenclature for the experimentally studied bacterial genes. A set of genes associated with virulence are termed for the *P. aeruginosa* strains z8 and PAO1. Genes that constitute a gene block frequently found in multiple tandem copies in various strains are termed teg(G to N), for 'tandem element gene'. The region encompassing from tegG to tegN is referred to as 'teg gene block'. The Refseq gene 'NT41_RS12090' is termed ghlO (glycosyl hydrolase like ORF) as it exhibits similarity to domain Cdd:cd06549 (E-value: 0.02, CDD database) . The PAO1 genes: PA2228, vqsM, qsrO, and PA225, constitute a putative operon [33] that is referred to as 'qsr' operon.

Additional file 6: Table S5. Bacterial strains generated in the present study. Strains generated in the present study are described with a strain name (AVPae #) and genotype (in both full and short formats).

Additional file 7: Table S6. Gene targeted by CRISPR spacers.

Additional file 8: Table S7. Description of in silico studied set of 1448 *P. aeruginosa* strains.

Additional file 9: Table S8. Gene clustering analysis for the in silico studied *P. aeruginosa* strains. Shown are only gene clusters that contain sequences from the studied 52 strains.

Additional file 10: Table S9. Sequence of plasmid pAVR85.

Additional file 11. Review history.

Acknowledgements
We would like to acknowledge members of the Ambros and Mello laboratories for feedback about this research project. We would also like to thank Deborah McEwan for assistance with the lifespan machine, Zeynep Mirza for contributing the growth rate measurements, Joseph Bondy-Denomy for sharing plasmid reagents, and Veronica Kos for help with strain requests. Some of the investigated bacterial strains were obtained from International Health Management Inc. Regarding *C. elegans*, some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440).

Review history
The review history is available as Additional file 11.

Peer review information
Andrew Cosgrove was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
AV-R, IV-L, and VA designed the study, performed and analyzed the experiments, and wrote the manuscript. ZC constructed *P. aeruginosa* strains; ZC and FA reviewed the results and the manuscript. All authors read and approved the final manuscript.

Funding
This research was supported by funding from NIH grants R01GM088365 and R01GM034028 (V.A.), R01AI085581 and P30DK040561 (F.M.A.), and the Pew Charitable Trusts (A.V-R.).

Availability of data and materials
The analyzed bacterial genomes are publicly available in NCBI website and downloaded from RefSeq [50]. All datasets generated in the paper are included in the supplemental tables. The in-house developed source code used in this paper is available on Github at https://github.com/isana18/PaeG-WAS [63] under the MIT license and Zenodo at https://doi.org/10.5281/zenodo.3534092 [64].

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA. [2]Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel. [3]Department of Microbiology & Immunology, Dalhousie University, Halifax, Nova Scotia, Canada. [4]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA. [5]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.

Received: 31 July 2019 Accepted: 15 November 2019
Published online: 10 December 2019

References
1. Ferris H. Contribution of nematodes to the structure and function of the soil food web. J Nematol. 2010;42:63–7.
2. Weitere M, Bergfeld T, Rice SA, Matz C, Kjelleberg S. Grazing resistance of *Pseudomonas aeruginosa* biofilms depends on type of protective mechanism, developmental stage and protozoan feeding mode. Environ Microbiol. 2005;7:1593–601.
3. Jousset A. Ecological and evolutive implications of bacterial defences against predators. Environ Microbiol. 2012;14:1830–43.
4. Mahajan-Miklos S, Tan MW, Rahme LG, Ausubel FM. Molecular mechanisms of bacterial virulence elucidated using a *Pseudomonas aeruginosa* - *Caenorhabditis elegans* pathogenesis model. Cell. 1999;96:47–56.
5. Pukatzki S, Kessin RH, Mekalanos JJ. The human pathogen *Pseudomonas aeruginosa* utilizes conserved virulence pathways to infect the social amoeba Dictyostelium discoideum. Proc Natl Acad Sci U S A. 2002;99:3159–64.
6. Rahme LG, Stevens EJ, Wolfort SF, Shao J, Tompkins RG, Ausubel FM. Common virulence factors for bacterial pathogenicity in plants and animals. Science. 1995;268:1899–902.
7. Rahme LG, Tan MW, Le L, Wong SM, Tompkins RG, Calderwood SB, et al. Use of model plant hosts to identify *Pseudomonas aeruginosa* virulence factors. Proc Natl Acad Sci U S A. 1997;94:13245–50.
8. Tan MW, Mahajan-Miklos S, Ausubel FM. Killing of *Caenorhabditis elegans* by *Pseudomonas aeruginosa* used to model mammalian bacterial pathogenesis. Proc Natl Acad Sci U S A. 1999;96:715–20.
9. Deredjian A, Colinon C, Hien E, Brothier E, Youenou B, Cournoyer B, et al. Low occurrence of *Pseudomonas aeruginosa* in agricultural soils with and without organic amendment. Front Cell Infect Microbiol. 2014;4:53.
10. Kaszab E, Szoboszlay S, Dobolyi C, Háhn J, Pék N, Kriszt B. Antibiotic resistance profiles and virulence markers of *Pseudomonas aeruginosa* strains isolated from composts. Bioresour Technol. 2011;102:1543–8.
11. Rutherford V, Yom K, Ozer EA, Pura O, Hughes A, Murphy KR, et al. Environmental reservoirs for exoS+ and exoU+ strains of *Pseudomonas aeruginosa*. Environ Microbiol Rep. 2018;10:485–92.
12. Schulenburg H, Félix M-A. The natural biotic environment of *Caenorhabditis elegans*. Genetics. 2017;206:55–86.
13. Lee DG, Urbach JM, Wu G, Liberati NT, Feinbaum RL, Miyata S, et al. Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. Genome Biol. 2006;7:R90.
14. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinforma Oxf Engl. 2012;28:3150–2.
15. van Belkum A, Soriaga LB, LaFave MC, Akella S, Veyrieras J-B, Barbu EM, et al. Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant *Pseudomonas aeruginosa*. mBio. 2015;6:e01796–15.
16. Kidd TJ, Ritchie SR, Ramsay KA, Grimwood K, Bell SC, Rainey PB. *Pseudomonas aeruginosa* exhibits frequent recombination, but only a limited association between genotype and ecological setting. PLoS One. 2012;7:e44199.
17. Pirnay J-P, Matthijs S, Colak H, Chablain P, Bilocq F, Van Eldere J, et al. Global *Pseudomonas aeruginosa* biodiversity as reflected in a Belgian river. Environ Microbiol. 2005;7:969–80.

18. Pirnay J-P, Bilocq F, Pot B, Cornelis P, Zizi M, Van Eldere J, et al. *Pseudomonas aeruginosa* population structure revisited. PLoS One. 2009;4: e7740.
19. Selezska K, Kazmierczak M, Müsken M, Garbe J, Schobert M, Häussler S, et al. *Pseudomonas aeruginosa* population structure revisited under environmental focus: impact of water quality and phage pressure. Environ Microbiol. 2012;14:1952–67.
20. Donlan RM, Costerton JW. Biofilms: survival mechanisms of clinically relevant microorganisms. Clin Microbiol Rev. 2002;15:167–93.
21. Martin N, Singh J, Aballay A. Natural genetic variation in the *Caenorhabditis elegans* response to *Pseudomonas aeruginosa*. G3 Bethesda Md. 2017;7: 1137–47.
22. Reddy KC, Andersen EC, Kruglyak L, Kim DH. A polymorphism in *npr-1* is a behavioral determinant of pathogen susceptibility in *C. elegans*. Science. 2009;323:382–4.
23. Stroustrup N, Ulmschneider BE, Nash ZM, López-Moyado IF, Apfeld J, Fontana W. The *Caenorhabditis elegans* lifespan machine. Nat Methods. 2013;10:665–70.
24. Feinbaum RL, Urbach JM, Liberati NT, Djonovic S, Adonizio A, Carvunis A-R, et al. Genome-wide identification of *Pseudomonas aeruginosa* virulence-related genes using a *Caenorhabditis elegans* infection model. PLoS Pathog. 2012;8:e1002813.
25. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput Biol. 2018;14:e1005958.
26. Kay E, Humair B, Dénervaud V, Riedel K, Spahr S, Eberl L, et al. Two GacA-dependent small RNAs modulate the quorum-sensing response in *Pseudomonas aeruginosa*. J Bacteriol. 2006;188:6026–33.
27. Zhang Y-F, Han K, Chandler CE, Tjaden B, Ernst RK, Lory S. Probing the sRNA regulatory landscape of *P. aeruginosa*: post-transcriptional control of determinants of pathogenicity and antibiotic susceptibility. Mol Microbiol. 2017;106(6):919–37
28. Franklin MJ, Nivens DE, Weadge JT, Howell PL. Biosynthesis of the *Pseudomonas aeruginosa* extracellular polysaccharides, alginate, Pel, and Psl. Front Microbiol. 2011;2:167.
29. Rocchetta HL, Burrows LL, Lam JS. Genetics of O-antigen biosynthesis in *Pseudomonas aeruginosa*. Microbiol Mol Biol Rev MMBR. 1999;63:523–53.
30. van Tilburg BE, Charron-Mazenod L, Reading DJ, Reckseidler-Zenteno SL, Lewenza S. Exopolysaccharide-repressing small molecules with antibiofilm and antivirulence activity against *Pseudomonas aeruginosa*. Antimicrob Agents Chemother. 2017;61(5):e01997–16.
31. Aires JR, Köhler T, Nikaido H, Plésiat P. Involvement of an active efflux system in the natural resistance of *Pseudomonas aeruginosa* to aminoglycosides. Antimicrob Agents Chemother. 1999;43:2624–8.
32. Westbrock-Wadman S, Sherman DR, Hickey MJ, Coulter SN, Zhu YQ, Warrener P, et al. Characterization of a *Pseudomonas aeruginosa* efflux pump contributing to aminoglycoside impermeability. Antimicrob Agents Chemother. 1999;43:2975–83.
33. Köhler T, Ouertatani-Sakouhi H, Cosson P, van Delden C. QsrO a novel regulator of quorum-sensing and virulence in *Pseudomonas aeruginosa*. PLoS One. 2014;9:e87814.
34. Kim DH, Feinbaum R, Alloing G, Emerson FE, Garsin DA, Inoue H, et al. A conserved p38 MAP kinase pathway in Caenorhabditis elegans innate immunity. Science. 2002;297:623–6.
35. Loh B, Haase M, Mueller L, Kuhn A, Leptihn S. The transmembrane morphogenesis protein gp1 of filamentous phages contains walker A and walker B motifs essential for phage assembly. Viruses. 2017;9.
36. Ledsgaard L, Kilstrup M, Karatt-Vellatt A, McCafferty J, Laustsen AH. Basics of antibody phage display technology. Toxins. 2018;10.
37. Stern A, Keren L, Wurtzel O, Amitai G, Sorek R. Self-targeting by CRISPR: gene regulation or autoimmunity? Trends Genet TIG. 2010;26:335–40.
38. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science. 2018;359.
39. Pawluk A, Davidson AR, Maxwell KL. Anti-CRISPR: discovery, mechanism and function. Nat Rev Microbiol. 2017.
40. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 2015;43:D298–9.
41. He J, Baldini RL, Déziel E, Saucier M, Zhang Q, Liberati NT, et al. The broad host range pathogen *Pseudomonas aeruginosa* strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes. Proc Natl Acad Sci U S A. 2004;101:2530–5.
42. Liu M, Li X, Xie Y, Bi D, Sun J, Li J, et al. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. Nucleic Acids Res. 2019;47: D660–5.
43. Sánchez-Diener I, Zamorano L, López-Causapé C, Cabot G, Mulet X, Peña C, et al. Interplay among resistance profiles, high-risk clones, and virulence in the Caenorhabditis elegans *Pseudomonas aeruginosa* infection model. Antimicrob Agents Chemother. 2017;61.
44. Hoffman LR, D'Argenio DA, MacCoss MJ, Zhang Z, Jones RA, Miller SI. Aminoglycoside antibiotics induce bacterial biofilm formation. Nature. 2005; 436:1171–5.
45. Brenner S. The genetics of *Caenorhabditis elegans*. Genetics. 1974;77:71–94.
46. McEwan DL, Feinbaum RL, Stroustrup N, Haas W, Conery AL, Anselmo A, et al. Tribbles ortholog NIPI-3 and bZIP transcription factor CEBP-1 regulate a *Caenorhabditis elegans* intestinal immune surveillance pathway. BMC Biol. 2016;14:105.
47. Djonović S, Urbach JM, Drenkard E, Bush J, Feinbaum R, Ausubel JL, et al. Trehalose biosynthesis promotes *Pseudomonas aeruginosa* pathogenicity in plants. PLoS Pathog. 2013;9:e1003217.
48. Prentki P, Krisch HM. In vitro insertional mutagenesis with a selectable DNA fragment. Gene. 1984;29:303–13.
49. West SE, Schweizer HP, Dall C, Sample AK, Runyen-Janecky LJ. Construction of improved *Escherichia-Pseudomonas* shuttle vectors derived from pUC18/ 19 and sequence of the region required for their replication in *Pseudomonas aeruginosa*. Gene. 1994;148:81–6.
50. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016; 44:6614–24.
51. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30:772–80.
52. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.
53. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490.
54. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015;11:e1004041.
55. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242–5.
56. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015;43: D130–7.
57. Bartell JA, Blazier AS, Yen P, Thøgersen JC, Jelsbak L, Goldberg JB, et al. Reconstruction of the metabolic network of *Pseudomonas aeruginosa* to interrogate virulence factor synthesis. Nat Commun. 2017;8:14631.
58. Marino ND, Zhang JY, Borges AL, Sousa AA, Leon LM, Rauch BJ, et al. Discovery of widespread type I and type V CRISPR-Cas inhibitors. Science. 2018;362:240–2.
59. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
60. Bikard D, Hatoum-Aslan A, Mucida D, Marraffini LA. CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. Cell Host Microbe. 2012;12:177–86.
61. Vercoe RB, Chang JT, Dy RL, Taylor C, Gristwood T, Clulow JS, et al. Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. PLoS Genet. 2013;9:e1003454.
62. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8:209.
63. Vasquez-Rifo A, Veksler-Lublinsky I, Cheng Z, Ausubel FM, Ambros V. Github. 2019. https://github.com/isana18/PaeGWAS
64. Vasquez-Rifo A, Veksler-Lublinsky I, Cheng Z, Ausubel FM, Ambros V. Zenodo. 2019. https://doi.org/10.5281/zenodo.3534092.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.