bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pangenome-wide and molecular evolution analyses of the *Pseudomonas aeruginosa* species

Jeanneth Mosquera-Rendón[1,2], Ana M. Rada-Bravo[3,4], Sonia Cárdenas-Brito[1], Mauricio Corredor[2], Eliana Restrepo-Pineda[3], Alfonso Benítez-Páez[1,5]

Affiliations:

1 Bioinformatics Analysis Group- GABi, Centro de Investigación y Desarrollo en Biotecnología – CIDBIO, 111221 Bogotá D.C., Colombia.

2 Grupo GEBIOMIC, FCEN, Universidad de Antioquia, Medellín, Colombia.

3 Grupo Bacterias y Cáncer, Universidad de Antioquia, Medellín, Colombia.

4. Grupo Biociencias, Institución Universitaria Colegio Mayor de Antioquia, Medellín, Colombia.

5 Corresponding author: Centro de Investigación y Desarrollo en Biotecnología, Calle 64A # 52-53 Int8 Of203, 111221 Bogotá D.C., Colombia. Tel +57 3105590633. E-mail abenitez@cidbio.org.

Additional author contact details:

Jeanneth Mosquera-Rendón: jmosquera@cidbio.org

Ana M. Rada-Bravo: ana.rada@colmayor.edu.co

Sonia Cárdenas-Brito: scardenas@cidbio.org

Mauricio Corredor-Rodriguez: mauricio.corredor@udea.edu.co

Elinana Restrepo-Pineda: elianarestrepo.pineda@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

25 **Running title:** Molecular evolution in the *P. aeruginosa* pangenome

Key words: Molecular evolution, *Pseudomonas aeruginosa*, pangenome, non-synonymous substitutions, synonymous substitutions, genetic variation, pathogenicity.

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Background. Drug treatments and vaccine designs against the opportunistic human pathogen *Pseudomonas aeruginosa* have multiple issues, all associated with the diverse genetic traits present in this pathogen, ranging from multi-drug resistant genes to the molecular machinery for the biosynthesis of biofilms. Several candidate vaccines against *P. aeruginosa* have been developed, which target the outer membrane proteins; however, major issues arise when attempting to establish complete protection against this pathogen due to its presumably genotypic variation at the strain level. To shed light on this concern, we proposed this study to assess the *P. aeruginosa* pangenome and its molecular evolution across multiple strains. *Results*. The *P. aeruginosa* pangenome was estimated to contain almost 17,000 non-redundant genes, and approximately 15% of these constituted the core genome. Functional analyses of the accessory genome indicated a wide presence of genetic elements directly associated with pathogenicity. An in-depth molecular evolution analysis revealed the full landscape of selection forces acting on the *P. aeruginosa* pangenome, in which purifying selection drives evolution in the genome of this human pathogen. We also detected distinctive positive selection in a wide variety of outer membrane proteins, with the data supporting the concept of substantial genetic variation in proteins probably recognized as antigens. Approaching the evolutionary information of genes under extremely positive selection, we designed a new Multi-Locus Sequencing Typing assay for an informative, rapid, and cost-effective genotyping of *P. aeruginosa* clinical isolates. *Conclusions*. We report the unprecedented pangenome characterization of *P. aeruginosa* on a large scale, which included almost 200 bacterial genomes from one single species and a molecular evolutionary analysis at the pangenome scale. Evolutionary information presented here provides a clear explanation of the issues associated with the use of protein

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

conjugates from pili, flagella, or secretion systems as antigens for vaccine design, which exhibit high genetic variation in terms of non-synonymous substitutions in *P. aeruginosa* strains.

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Background

Humans are frequently infected by opportunistic pathogens that take advantage of their compromised immunological status to cause persistent and chronic infections. The Gram-negative bacterium *Pseudomonas aeruginosa* is one of those recurrent human pathogens. *P. aeruginosa* remains one of the most important pathogens in nosocomial infections, and it is often associated with skin, urinary tract, and respiratory tract infections [1]. Respiratory tract infections are of major relevance in cystic fibrosis patients, given that *P. aeruginosa* deeply affects their pulmonary function, causing life-threatening infections [2]. One of the better-known adaptive resistance mechanisms of *P. aeruginosa* to evade either the host immune response and drug therapy is its ability to form biofilms. The *Pseudomonas aeruginosa* biofilm is an extremely stable capsule-like structure constituted primarily of polysaccharides, proteins, and DNA, in which PsI exopolysaccharide seems to be a key player for biofilm matrix stability [3]. Quorum sensing signals promote the formation of *P. aeruginosa* biofilms, which minimizes the entry of antimicrobial compounds inside bacterial cells and hinders the recognition of pathogen-associated molecular patterns (PAMPs) by the host immune system [4]. Consequently, current treatments against *P. aeruginosa* fail to resolve infections before tissue deterioration occurs. To address this concern, more efficient alternatives to abolish *P. aeruginosa* infection have produced promising but not definitive results. Accordingly, several candidate *P. aeruginosa* vaccines have been developed by targeting outer membrane proteins (Opr), lipopolysaccharides (LPS), polysaccharides (PS), PS-protein conjugates, flagella, pili, and single or multivalent live-attenuated cells [5-9]. However, major issues in the development of a successful *P. aeruginosa* vaccine arise from the probable genotypic variation at the strain level, making *P. aeruginosa* a

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

presumably antigenically variable organism. Results supporting this assumption have been reported, yielding genetic information from the *P. aeruginosa* genome. For example, genetic variability explored in multiple *P. aeruginosa* isolates from different regions of the world indicated that *pcrV*, a member of the type III secretion system, exhibits limited genetic variation in terms of non-synonymous substitutions [10]. Although this type of analysis is informative, it provides only a very limited view of the genetic and evolutionary processes occurring at the genome level in *P. aeruginosa* and does not completely explain the failure to design and develop a successful vaccine against this human pathogen. Although antigen selection to design a *P. aeruginosa* vaccine is not a reported problem [11], to date, no genomic studies have correlated antigen genetic structure and variation with the effectiveness of antibody immunotherapy or vaccines, the efficacy of which remains elusive [11]. Moreover, enormous variation in the response against *P. aeruginosa* immunogenic proteins in patients with *P. aeruginosa* infections [12] could indicate that genetic factors from the pathogen and/or host could be responsible for the incomplete efficacy of candidate vaccines tested. In this fashion, this study aimed to i) better understand the genome structure and genetic variation exhibited by *Pseudomonas aeruginosa*¸ ii) link the genome variation information with past and future *P. aeruginosa* vaccine designs, and iii) present and validate new molecular markers for Multi-Locus Sequence Typing (MLST) based on the study of genes exhibiting a higher ratio of non-synonymous over synonymous substitution rate. To achieve these aims, a combined pangenome-wide and molecular evolution analysis was performed using up-to-date and genome-scale genetic information publicly available in the Pathosystems Resource Integration Center (PATRIC) database [13].

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results and Discussion

Defining the Pseudomonas aeruginosa pangenome

A total of 181 genomes of *P. aeruginosa* strains were obtained through the public PATRIC database (see methods). The preliminary analysis of the *Pseudomonas aeruginosa* genome size variability is shown in Table 1. The *P. aeruginosa* chromosome contains 6,175 genes on average, with a distribution ranging from 5,382 to 7,170 genes per genome, indicating a variation of 13-16% in terms of gene content among all strains analysed. By using the genome-centred approximation to define the *P. aeruginosa* pangenome (see methods), a total of 16,820 non-redundant genes were retrieved from those 181 genomes analysed. Almost one-third of the full set of genes constituting the *P. aeruginosa* pangenome, 5,209 genes (31%), were found to be uniquely present, meaning that every strain approximately contributes 29 new genes to the *Pseudomonas aeruginosa* pangenome on average. Initially, these data fit well with a theoretical number of strain-specific new genes added to the pangenome when a new strain genome was sequenced, 33 for the *Streptococcus agalactiae* pangenome [14]. However, for a more precise calculation, our observed data were fitted to the decay function (Figure 1A), and we found a decay rate of 0.071, which indicated that every *P. aeruginosa* strain contributes approximately 14 new genes to the pangenome. We believe that not only the number of genomes analysed but also the genome complexity of the species analysed influenced this number. In our case, *P. aeruginosa* exceeded the average *S. agalactiae* genome size 3-fold, but every strain contributed fewer genes to its pangenome compared to *S. agalactiae*. To test and validate this reciprocal proportionality, further studies must be conducted in a wide variety of species with a high number of strain genomes available. Additionally, the decay function indicated that a plateau is reached with 160 strains/genomes, which means that

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

our data likely described the complete gene variability of the *Pseudomonas aeruginosa* species. Further information was extracted from the pangenome analysis regarding gene categorization. The core genome or extended core of genes was characterized as the set of genes present in all or almost all genomes analysed; in this manner, we established that the *Pseudomonas aeruginosa* core genome contains approximately 2,503 genes that are present in all 181 genomes studied, and they account for 15% of the pangenome. The remaining set of genes, which were not included in the core genome or were unique (present in 1 genome), were referred to as the accessory genome; it included the 54% of genes found in the *P. aeruginosa* pangenome (Table 1). Interestingly, when we plotted the frequency of all pangenome genes present in different strains/genomes analysed (Figure 1B), we found a similar distribution to that reported by Lapierre and Gogarten when they estimated the pangenome for more than 500 different bacterial genomes [15]. This distribution plot clearly demonstrated the characteristic distribution and frequency of different groups of the above-stated genes. In general terms, the *P. aeruginosa* pangenome exhibits a high level of genome variability, whereby only 40% (2,503/6,175) of its genome is constant, on average. Thus, the remaining 60% of *P. aeruginosa* genome is presented as a variable piece of DNA composed of a wide repertoire of genes and molecular functions that are tightly linked to different levels of pathogenesis, virulence, and resistance. To test this hypothesis, we proceeded to perform a functional analysis with the full set of genes uniquely presented. As a consequence, the nucleotide sequences of genes found to be present in only in one *P. aeruginosa* strain were translated to amino acid sequences and then submitted to the KASS server for functional annotation at the protein level [16]. We retrieved only 14% of the functional annotation for this set of genes, of which more than 59% comprise ORFs, potentially encoding peptides less than 100 aa in length. We explored the predominance of functions present in the 738 ORFs annotated at the KEEG

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Pathways level. Consequently, we found that in addition to proteins involved in more general functions, such as metabolic pathways (ko01100, 103 proteins) and the biosynthesis of secondary metabolites (ko01110, 30 proteins), proteins participating in more specific molecular tasks, such as the biosynthesis of antibiotics (ko01130, 22 proteins), the bacterial secretion system (ko03070, 20 proteins), ABC transporters (ko02010, 17 proteins), and two-component system (ko02020, 36 proteins), were frequently present as well. Among all of these proteins, we highlighted the presence of several members of the type II and IV secretion systems responsible for the secretion of bacterial toxins, proteins of the macrolide exporter system, and beta-lactamases and efflux pump proteins associated with beta-lactam resistance.

We further assessed the molecular functions of the portion of the *P. aeruginosa* accessory genome comprising genes between the 5th and 95th percentile of frequency (9 < accessory genome < 172) among all the genomes analysed. A total of 2,605 proteins were submitted again to the KASS server, retrieving functional annotation for 735 (28%) of them. We found a similar predominance of the above-stated pathways, but we expanded our analysis to include the biosynthesis of amino acids (ko01230, 37 proteins) and amino sugar and nucleotide sugar metabolism (ko00520, 13 proteins). Strikingly, we found additional proteins involved in vancomycin resistance as well as proteins of the type I and VI secretion systems associated with the export of toxins, proteases, lipases and other effector proteins. A general view of the molecular functions confined to different categories of the *P. aeruginosa* pangenome is shown in Figure 2. Comparison at the orthology level (Figure 2) indicated that a high level of functional specificity exists in all gene categories of the *P. aeruginosa* pangenome, whereby 79% of annotated genes in the core genome are not present in other categories. This percentage remains high at 47% in

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

unique genes and 49% in the accessory genome. At the KEGG pathway level, we found the opposite effect, but we found some molecular pathways to be distinctive for every gene category in the pangenome. Table 2 summarizes those molecular pathways in which the *P. aeruginosa* core genome was found to contain a wide range of genes involved in either antibiotic biosynthesis and resistance and genes involved in infectious and other human diseases.

Molecular evolution in the Pseudomonas aeruginosa pangenome

In addition to uncovering the genes and functions that confer distinctive features to *P. aeruginosa* strains, we explored the genetic variability in every gene family retrieved from its pangenome. This approach could provide evidence of how the *P. aeruginosa* genome evolves to evade the immune response as well as depict the level of variability thought to be the major cause of the lack of success in designing a effective vaccine. For more than 10,000 gene families containing at least 2 members, we calculated the synonymous (dS) and non-synonymous (dN) rates, parameters indicative of the selection pressure on coding genes. The global distribution of dS and dN rates expressed as the omega value (ω = dN/dS) across the *P. aeruginosa* pangenome is presented in Figure 3A. Although the distribution of ω values fits well into a unimodal distribution, globally, it shows a shift-to-left distribution towards values lower than 1 with ω median = 0.1. These data suggest that the *P. aeruginosa* coding genome is under purifying selection as a whole, in which synonymous substitutions are predominantly higher than non-synonymous substitutions. The coding genes considered under positive selection must present $\omega > 1$ (dN > dS); however, at the initial stage, we performed more restrictive filtering, thus considering those genes that exhibited at least a 2-fold greater non-synonymous substitution rate than the

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

synonymous substitutions ($\omega \geq 2$). As a result, we retrieved a total of 230 genes (1.4% of pangenome) for which 71 functional annotations (31%) were recovered from the KASS server. We found a wide variability in terms of the molecular pathways for the genes under positive selection. Notably, among all genes under positive selection, we detected that some of them coded for proteins with remarkable functions, such as VirB2 and VirB9 (K03197 and K03204, respectively). Both proteins are components of the type IV secretion system and are localized at the outer membrane. In the case of VirB2 proteins, the T-pilus protein controls attachment to different receptors on the host cell surface to deliver toxin effector molecules [17]. VirB2 and VirB9 proteins exposed on the cell surface of pathogens make *P. aeruginosa* a proper candidate for recognition by the host immune system to trigger a specific response against these potential antigens, thus promoting immune memory against this pathogen. The antigenicity of VirB2 and VirB9 proteins is further supported by their high rate of non-synonymous substitutions observed across different strains analysed, which would be result of the strong selection forces from the host immune system. Similarly, other outer membrane-bound proteins, such as the flippase MurJ (K03980) and the flagellin FlgF (K02391), which have been associated with virulence and pathogenicity [18, 19], exhibited a higher rate of non-synonymous substitutions than synonymous substitutions .

Strong selection forces from the immune response or environmental pressure were also detected in a set of *P. aeruginosa* genes tightly linked with virulence in other human pathogens. Therefore, we observed positive selection in the following genes: the PsrtC (K08303) homologue, a protease involved in mucus degradation during *H. pylori* infection (pathway ko05120); the MprF and ParR homologues (KO14205 and K18073, respectively), proteins involved in the cationic antimicrobial peptide (CAMP) resistance in

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gram-positive and Gram-negative bacteria (ko1503), respectively; the PstS homologue (K02040), an integral membrane ABC phosphate transporter that modulates the TLR4 response during *M. tuberculosis* infection (ko5152); the *T. brucei* ICP homologue (14475), a protein involved in immunosupression by modulating the degradation of IgGs (ko5143); and the RNA polymerase sigma-54 factor (K03092), which is associated with the *V. cholera* pathogenic cycle to control the expression of motor components of flagella (ko5111).

Given the low level of functional annotation for genes under positive selection, we performed an additional quantitative assessment to determine protein domain enrichment in the group of proteins under positive selection. Once the inventory of SMART and Pfam domains contained in the entire *P. aeruginosa* pangenome was assessed, we performed a Fisher's exact test for 2 x 2 contingency tables to verify the significant over-representation of Pfam/SMART domains in the proteins under positive selection with respect to the pangenome. We observed the presence and prevalence of 4,090 different protein domains from both the SMART and Pfam classification in the *P. aeruginosa* pangenome. Forty-four of these 4,090 domains were found to be over-represented in the proteins exhibiting positive selection (Table 3). Among them, we observed a high frequency of membrane-bound proteins acting as transporters or receptors. Some of the functions over-represented in Table 3 agree with some stated from previous analyses in which membrane proteins (transporters and/or receptors) as well as the Sigma-54 factor seem to be under positive selection in *P. aeruginosa*. Interestingly, we observed the presence of proteins related with either 16S RNA and ribosomal protein methylation (Table 3). We detected such patterns of molecular evolution in this class of proteins previously, but in different human pathogens [20]. Although we cannot shed light on the meaning of this type of evolution in these

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

proteins given their function, we hypothesized that they might influence the translation process to modulate the expression of a certain set of proteins directly or indirectly involved in pathogenesis. Recent studies on rRNA methylation indicate that they play a meaningful role in decoding function [21-23]. Indeed, some of them have been directly involved with virulence [24].

When we attempted a similar analysis in a counterpart set of proteins under purifying or negative selection ($\omega < 1$), the biased distribution of omega values across the *P. aeruginosa* pangenome (Figure 3A) made it difficult to set up a suitablethreshold to recover proteins under this type of selection. Therefore, we obtained Z-scores of both the dN and dS rates (Figure 3A, light red histogram), thus reaching a normal distribution around $\omega = 1$ (neutrality). Using this normalized distribution of ω values, we could determine those genes with evolution significantly different ($p \leq 0.05$) from neutrality ($\omega = 1$) towards a strong negative selection (lowest ω values). As a result, we found a group of 268 proteins/genes under negative selection, the dN and dS rates of which are plotted in Figure 3B (see the blue points distribution). The quantitative assessment to determine protein domain enrichment indicated that more than 130 SMART and/or Pfam domains were over-represented in this set of proteins, and as expected, most of them were related to the central functions of cell maintenance, such as translation (ribosome proteins, tRNA biogenesis, amino acid starvation response), carbohydrate metabolism, amino acid biosynthesis and transport, and respiration.

New high variability markers for multi-locus sequence typing of P. aeruginosa strains

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Characterization of the *P. aeruginosa* pangenome offers not only critical information about the molecular functions and prevalence of certain genes across multiple strains analysed but also information about the level of genetic variability at the strain level. A molecular evolution approach retrieved a large set of genes/proteins under positive selection in *P. aeruginosa*. At the same time, such genes/proteins could be used for genotyping aims to associate certain genetic variants with pathogenicity and virulence traits. As a consequence, we selected and tested some *P. aeruginosa* genes in a MLST strategy to discern phylogenetic relationships among a large number of PATRIC reference strains analysed and six *P. aeruginosa* aminoglycoside and carbapenem-resistant strains isolated from patients who acquired healthcare-associated infections in a clinic located outside the metropolitan area of Medellin, Antioquia, Colombia.

We narrowed down the list of MLST candidates by selecting the genes that had the following characteristics: i) present in at least 95% of the strains explored at the sequence level (frequency $\geq$ 172); ii) exhibiting omega values significantly higher than 1 (Figure 3B, $p \leq 0.05$, $\omega > 15$); and iii) short enough to facilitate Sanger sequencing in a few reactions. Of the 27 genes/proteins showing significant positive selection, we finally selected four genes, the features of which are depicted in Table 4. After amplification and Sanger sequencing of selected genes in our six *P. aeruginosa* isolates, we combined that genetic information with that completely available for 170 *P. aeruginosa* strains, thus building a multiple sequence alignment almost 3,000 bp in length for a total of 176 different strains. Using maximum likelihood approaches, we reconstructed the phylogenetic relationships among all strains and retrieved the phylogenetic tree showed in Figure 4. Our six local isolates were positioned in three different clades, where isolate 49 was closely related to the highly virulent *Pseudomonas aeruginosa* PA14 strain, representing the most common

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

clonal group worldwide [25]. By contrast, isolate 77 was related to several strains, including the multi-drug-resistant *Pseudomonas aeruginosa* NCGM2.S1 [26] and the cytotoxic corneal isolate *Pseudomonas aeruginosa* 6077 [27]. Finally, the 30-1, 42-1, 45, and 04 isolates presented a close relationship and were related to the multi-drug resistant *Pseudomonas aeruginosa* VRFPA02 isolate from India [28].

Based on the best evolutionary model fitted to the nucleotide substitution pattern observed for these markers (TrN+I+G), a proportion of invariable sites of 0.9080 was obtained, thus indicating that more than 250 polymorphic sites are present in our MLST approach. Moreover, gamma distribution parameters (0.5060) is indicative of few hot-spots with high substitution rates [29]. In this fashion, we provided support to use the highly variable genetic markers reported here for MLST to produce an initial, fast, and cost-effective genotyping for *P. aeruginosa* strains of clinical interest.

Conclusions

High-throughput sequencing technology has permitted the analysis of the genetic identity of a vast number of microorganisms, an applied science especially relevant to studying human pathogens and their virulence and pathogenicity traits in depth. Here, we have utilized the large amount of genetic information available in the PATRIC database to determine the genetic elements directly associated with pathogenesis of *Pseudomonas aeruginosa*. We have extensively described the *P. aeruginosa* pangenome in terms of the effective number of non-redundant genes presentin this bacterial species by analysing more than 180 different strain genomes (Table 1). We outlined the genomic variability of this human pathogen, demonstrating that approximately 60% of the *P. aeruginosa* genome is

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variable across strains, with the remaining genome encoding genes that are involved in central functions, such as virulence, resistance, toxicity and pathogenicity.

We have identified major genetic pieces of the core and accessory genome in *P. aeruginosa*. Approximately 15% (2,503/16,820 genes) of the pangenome was found to constitute the core genome and was present in 100% of the strains studied, accomplishing general molecular functions for cell maintenance such as replication, translation, transcription, central metabolism, electron transport chain, amino acid biosynthesis and transport, nucleotide biosynthesis, cell wall synthesis and maintenance, and cell division. Conversely, the accessory genome exhibited a comprehensive variety of functions, ranging from a wide spectrum of antibiotic resistances to a specialized secretion system delivering toxins and effector proteins potentially harmful for host cells. However, pathogenicity traits were also observed in the distinctive KEGG pathways revealed for the core genome (Table 2, Figure 2).

Although this is not the first report to describe the pangenome for a single bacterial species [14, 30-32], this report is the first to describe it on such a large scale, including almost 200 bacterial genomes from one single species and performing a pangenome-scale molecular evolutionary analysis. Our study fits well with previous and general genomic characterizations of this human pathogen [33], and it definitively expands our knowledge about the evolutionary mechanisms of *P. aeruginosa* pathogenesis. This study aimed to reveal the evolutionary processes occurring at the pangenome level in *P. aeruginosa* that could explain the failure to design and develop of a successful vaccine against this human pathogen as well as provide an understanding of the molecular mechanisms that drive the evasion of the host immune system. We observed that the *P. aeruginosa* genome is

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

globally under purifying selection, given the distribution of omega values (ω = dN/dS, median ~0.1) discerned for every gene family present in its pangenome (Figure 3A). This result was further supported by the finding that there are 10-fold more genes (268 vs. 27, Figure 3B) under strong purifying selection than strong positive selection (significantly different to neutrality, $p \leq 0.05$). Although we found that the *P. aeruginosa* pangenome evolves to purifying selection as a whole, we distinguished some genes and functions predominantly present in the reduced set of genes under positive selection. As a consequence, a considerable number of proteins located at the outer membrane, such as those associated with receptor and transporter functions, were identified to have an increased rate of non-synonymous substitutions (Table 3). These data corroborated our results based on KEGG functional analysis, which described an ample group of surface-exposed proteins under strong selection forces from the immune response or environmental pressure. Therefore, they could be the major antigens recognized by the human immune system.

For the first time, pangenome-scale evolutionary information is presented to support the design of a *P. aeruginosa* vaccine. In this fashion, failures when using protein conjugates from pili, flagella, or secretions systems [5, 7, 9, 11] are partially explained by the data presented here, which indicates the presence of a high genetic variation in this class of proteins in terms of non-synonymous substitutions.

Finally, we further explored the genetic information derived from our molecular evolution analyses and proposed a set of four new polymorphic genetic markers for MLST (Table 4). We demonstrated that these markers contain an adequate proportion of hotspots for variation, exhibiting high nucleotide substitution rates. Using these four loci, we

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

discerned the genetic identity of 6 local isolates of *P. aeruginosa* and related them with the resistance and virulence traits carried in reference strains (Figure 4).

Methods

Pangenome-wide analysis

Genome information from *Pseudomonas aeruginosa* strains was downloaded via the ftp server from the PATRIC database [13]. A set of 181 available genomes (*ffn* files) was retrieved from the PATRIC database, April 2014 release. Estimation of the *Pseudomonas aeruginosa* pangenome size was assessed in a similar manner to that previously reported as genome- and gene-oriented methods [14, 15]. Briefly, a BLAST-based iterative method was used to extract the full set of non-redundant genes representing the *P. aeruginosa* pangenome. Then, the set of non-redundant genes obtained was used to explore their occurrence pattern in the 181 *Pseudomonas aeruginosa* genomes through BLASTN-based comparisons [34, 35].

Molecular evolution analysis

The full set of ORFs constituting the *Pseudomonas aeruginosa* pangenome was used to search homologues in all genomes analysed, and multiple sequence alignments were built using refined and iterative methods [36, 37]. The synonymous and non-synonymous substitution rates were calculated in a pairwise manner using approximate methods [38] and by correcting for multiple substitutions [39]. Omega values (ω) were computed as the averaged ratio of dN/dS rates from multiple comparisons, and genes under strong positive selection were selected when $\omega \geq 2$. The Z-score of ω values was computed to depict functions of genes under strong purifying selection and potential MLST genetic markers

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

under strong positive selection ($p \leq 0.05$). Large-scale analyses of pairwise comparisons, statistical analysis, and graphics were performed using R v3.1.2 (http://www.R-project.org).

Functional genomics analysis

Functional annotation of genes was performed using the KEGG Automatic Annotation Server for KEGG Orthology [16]. KEGG pathways and ontologies were explored in the KEGG BRITE database [40]. Functional domains present in genes of interest were assigned using Perl scripting for batch annotation (http://smart.embl-heidelberg.de/help/SMART_batch.pl) against the Simple Modular Architecture Research Tool (SMART) together with Pfam classification [41, 42]. Fisher's exact test with a false discovery rate (FDR) for 2 x 2 contingency tables to measure enrichment of Pfam/SMART domains was performed using R v3.1.2 (http://www.R-project.org). Venn diagrams were drawn using the *jvenn* server [43].

Multi-locus sequence typing

The six *Pseudomonas aeruginosa* strains were isolated from patients who acquired healthcare-associated infections at a clinic located outside the metropolitan area of Medellin, Antioquia, Colombia. Such strains, clinically characterized for multi-drug resistance, were kindly donated by the scientific staff of the Bacteria & Cancer Researching Group of the Faculty of Medicine, University of Antioquia, Colombia. The genomic DNA from *P. aeruginosa* multi-drug resistant strains was extracted using a GeneJER^{TM}GenomicDNA Purification Kit (Thermo Scientific, Waltham, MA, USA). The reference sequences of *Pseudomonas aeruginosa* PA01 for the four markers selected to perform MLST were accessed via GenBank using the following accession number

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AE004091.2: for gene family 3333, region 930623 to 931822; for gene family 3675, region 1167488 to 1168237; for gen family 4766, region 2230183 to 2229425; and for gene family 5348, region 2935851 to 2936423. Primers were designed to amplify the complete sequence of each gene, and PCR proceeded with 28 cycles of amplification using Phusion® High-Fidelity DNA Polymerase (Thermo Scientific, Waltham, MA, USA) and 50 ng of genomic DNA. PCR products were isolated using a GeneJet PCR Purification Kit (Life technologies, Carlsbad, CA, USA), and both strands were sequenced by the Sanger automatic method in an ABI 3730 xl instrument (Stab Vida Inc., Caparica, Portugal). Base calling and genetic variants were manually explored using the delivered *ab1* files and FinchTV viewer (Geospiza Inc. Perkin Elmer, Waltham, MA, USA). Assembled sequences from both strands were obtained and concatenated to respective reference sequences obtained from the PATRIC genomes analysed. Sequences belonging to the respective gene family were aligned using iterative methods [36, 37], and alignments were concatenated to perform phylogenetic analysis. The sequential likelihood ratio test was carried out to detect the evolutionary model that better explained genetic variation in all genes integrated in the MLST approach. For that reason, we used the jModelTest tool [44], and model selection was completed by calculating the corrected Akaike Information Criterion (cAIC). The MLST tree was constructed using iTOL [41, 45] and the phylogeny obtained using the TrN+I+G model.

Competing Interests

The authors declare that they have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Authors' contributions

ABP designed and directed this study. JMR and ABP performed the pangenome, molecular evolution, and phylogenetic analyses. ERP and AMR obtained the *P. aeruginosa* clinical isolates. JMR, AMR, and MC provided PCR techniques. JMR and SCB curated the sequences from Sanger automatic sequencing. JMR, SCB, and ABP prepared the manuscript preparation. All authors read and approved the final version of this manuscript.

Acknowledgements

The authors give thanks to the Colombian Agency for Science, Technology, and Innovation (Colciencias) and the National Fund for Science, Technology, and Innovation "Francisco José de Caldas" for grant 5817-5693-4856 to ABP and grant 1115-5693-3375 to ERP. The authors also thank the "Clinica Antioquia" microbiology staff, Medellín, Colombia, who donated the clinical isolates for the MLST studies.

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Lavoie EG, Wangdi T, Kazmierczak BI: **Innate immune responses to Pseudomonas aeruginosa infection**. *Microbes Infect* 2011, **13**(14-15):1133-1145.
2. Hauser AR, Jain M, Bar-Meir M, McColley SA: **Clinical significance of microbial infection and adaptation in cystic fibrosis**. *Clin Microbiol Rev* 2011, **24**(1):29-70.
3. Ma L, Conover M, Lu H, Parsek MR, Bayles K, Wozniak DJ: **Assembly and development of the Pseudomonas aeruginosa biofilm matrix**. *PLoS Pathog* 2009, **5**(3):e1000354.
4. Alhede M, Bjarnsholt T, Givskov M, Alhede M: **Pseudomonas aeruginosa biofilms: mechanisms of immune evasion**. *Adv Appl Microbiol* 2014, **86**:1-40.
5. Doring G, Meisner C, Stern M: **A double-blind randomized placebo-controlled phase III study of a Pseudomonas aeruginosa flagella vaccine in cystic fibrosis patients**. *Proc Natl Acad Sci U S A* 2007, **104**(26):11020-11025.
6. Lang AB, Rudeberg A, Schoni MH, Que JU, Furer E, Schaad UB: **Vaccination of cystic fibrosis patients against Pseudomonas aeruginosa reduces the proportion of patients infected and delays time to infection**. *Pediatr Infect Dis J* 2004, **23**(6):504-510.
7. Horn MP, Zuercher AW, Imboden MA, Rudolf MP, Lazar H, Wu H, Hoiby N, Fas SC, Lang AB: **Preclinical in vitro and in vivo characterization of the fully human monoclonal IgM antibody KBPA101 specific for Pseudomonas aeruginosa serotype IATS-O11**. *Antimicrob Agents Chemother* 2010, **54**(6):2338-2344.
8. Kamei A, Coutinho-Sledge YS, Goldberg JB, Priebe GP, Pier GB: **Mucosal vaccination with a multivalent, live-attenuated vaccine induces multifactorial immunity against Pseudomonas aeruginosa acute lung infection**. *Infect Immun* 2011, **79**(3):1289-1299.
9. Campodonico VL, Llosa NJ, Bentancor LV, Maira-Litran T, Pier GB: **Efficacy of a conjugate vaccine containing polymannuronic acid and flagellin against experimental Pseudomonas aeruginosa lung infection in mice**. *Infect Immun* 2011, **79**(8):3455-3464.
10. Lynch SV, Flanagan JL, Sawa T, Fang A, Baek MS, Rubio-Mills A, Ajayi T, Yanagihara K, Hirakata Y, Kohno S *et al*: **Polymorphisms in the Pseudomonas aeruginosa type III secretion protein, PcrV - implications for anti-PcrV immunotherapy**. *Microb Pathog* 2010, **48**(6):197-204.
11. Doring G, Pier GB: **Vaccines and immunotherapy against Pseudomonas aeruginosa**. *Vaccine* 2008, **26**(8):1011-1024.
12. Montor WR, Huang J, Hu Y, Hainsworth E, Lynch S, Kronish JW, Ordonez CL, Logvinenko T, Lory S, LaBaer J: **Genome-wide study of Pseudomonas aeruginosa outer membrane protein immunogenicity using self-assembling protein microarrays**. *Infect Immun* 2009, **77**(11):4877-4886.
13. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R *et al*: **PATRIC, the bacterial bioinformatics database and analysis resource**. *Nucleic Acids Res* 2014, **42**(Database issue):D581-591.

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

14. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS *et al*: **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"**. *Proc Natl Acad Sci U S A* 2005, **102**(39):13950-13955.
15. Lapierre P, Gogarten JP: **Estimating the size of the bacterial pan-genome**. *Trends Genet* 2009, **25**(3):107-110.
16. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W182-185.
17. Backert S, Fronzes R, Waksman G: **VirB2 and VirB5 proteins: specialized adhesins in bacterial type-IV secretion systems?** *Trends Microbiol* 2008, **16**(9):409-413.
18. Ulland TK, Buchan BW, Ketterer MR, Fernandes-Alnemri T, Meyerholz DK, Apicella MA, Alnemri ES, Jones BD, Nauseef WM, Sutterwala FS: **Cutting edge: mutation of Francisella tularensis mviN leads to increased macrophage absent in melanoma 2 inflammasome activation and a loss of virulence**. *J Immunol* 2010, **185**(5):2670-2674.
19. Wong HC, Liu SH, Chen MY: **Virulence and stress susceptibility of clinical and environmental strains of Vibrio vulnificus isolated from samples from Taiwan and the United States**. *J Food Prot* 2005, **68**(12):2533-2540.
20. Mosquera-Rendon J, Cardenas-Brito S, Pineda JD, Corredor M, Benitez-Paez A: **Evolutionary and sequence-based relationships in bacterial AdoMet-dependent non-coding RNA methyltransferases**. *BMC Res Notes* 2014, **7**:440.
21. Benitez-Paez A, Villarroya M, Armengod ME: **The Escherichia coli RlmN methyltransferase is a dual-specificity enzyme that modifies both rRNA and tRNA and controls translational accuracy**. *Rna* 2012, **18**(10):1783-1795.
22. Benitez-Paez A, Villarroya M, Armengod ME: **Regulation of expression and catalytic activity of Escherichia coli RsmG methyltransferase**. *Rna* 2012, **18**(4):795-806.
23. Kimura S, Suzuki T: **Fine-tuning of the ribosomal decoding center by conserved methyl-modifications in the Escherichia coli 16S rRNA**. *Nucleic Acids Res* 2010, **38**(4):1341-1352.
24. Kyuma T, Kimura S, Hanada Y, Suzuki T, Sekimizu K, Kaito C: **Ribosomal RNA methyltransferases contribute to Staphylococcus aureus virulence**. *Febs J* 2015.
25. Wiehlmann L, Wagner G, Cramer N, Siebert B, Gudowius P, Morales G, Kohler T, van Delden C, Weinel C, Slickers P *et al*: **Population structure of Pseudomonas aeruginosa**. *Proc Natl Acad Sci U S A* 2007, **104**(19):8101-8106.
26. Miyoshi-Akiyama T, Kuwahara T, Tada T, Kitao T, Kirikae T: **Complete genome sequence of highly multidrug-resistant Pseudomonas aeruginosa NCGM2.S1, a representative strain of a cluster endemic to Japan**. *J Bacteriol* 2011, **193**(24):7010.
27. Allewelt M, Coleman FT, Grout M, Priebe GP, Pier GB: **Acquisition of expression of the Pseudomonas aeruginosa ExoU cytotoxin leads to increased bacterial virulence in a murine model of acute pneumonia and systemic spread**. *Infect Immun* 2000, **68**(7):3998-4004.
28. Malathi J, Murugan N, Umashankar V, Bagyalakshmi R, Madhavan HN: **Draft Genome Sequence of Multidrug-Resistant Pseudomonas aeruginosa Strain**

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

VRFPA02, Isolated from a Septicemic Patient in India. *Genome Announc* 2013, **1**(4).

29. Yang Z: **Among-site rate variation and its impact on phylogenetic analyses**. *Trends Ecol Evol* 1996, **11**(9):367-372.
30. D'Auria G, Jimenez-Hernandez N, Peris-Bondia F, Moya A, Latorre A: **Legionella pneumophila pangenome reveals strain-specific virulence factors**. *BMC Genomics* 2010, **11**:181.
31. Kittichotirat W, Bumgarner RE, Asikainen S, Chen C: **Identification of the pangenome and its components in 14 distinct Aggregatibacter actinomycetemcomitans strains by comparative genomic analysis**. *PLoS One* 2011, **6**(7):e22420.
32. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R *et al*: **The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates**. *J Bacteriol* 2008, **190**(20):6881-6893.
33. Klockgether J, Cramer N, Wiehlmann L, Davenport CF, Tummler B: **Pseudomonas aeruginosa Genomic Structure and Diversity**. *Front Microbiol* 2011, **2**:150.
34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
35. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
36. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.
37. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.
38. Korber B: **HIV signature and sequence variation analysis**. In: *Computational analysis of HIV molecular sequences.* Edited by Rodrigo A, Learn G. Dordrecht, Netherlands: Kluwer Academic Publishers; 2000: 55-72.
39. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions**. *Mol Biol Evol* 1986, **3**(5):418-426.
40. Aoki-Kinoshita KF, Kanehisa M: **Gene annotation and pathway mapping in KEGG**. *Methods Mol Biol* 2007, **396**:71-91.
41. Letunic I, Doerks T, Bork P: **SMART 7: recent updates to the protein domain annotation resource**. *Nucleic Acids Res* 2012, **40**(Database issue):D302-305.
42. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J *et al*: **Pfam: the protein families database**. *Nucleic Acids Res* 2014, **42**(Database issue):D222-230.
43. Bardou P, Mariette J, Escudie F, Djemiel C, Klopp C: **jvenn: an interactive Venn diagram viewer**. *BMC Bioinformatics* 2014, **15**:293.
44. Darriba D, Taboada GL, Doallo R, Posada D: **jModelTest 2: more models, new heuristics and parallel computing**. *Nat Methods* 2012, **9**(8):772.
45. Letunic I, Bork P: **Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W475-478.

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

605 Table 1. Main features of the *Pseudomonas aeruginosa* pangenome

Features analysed	***P. aeruginosa* pangenome**
Genomes	181
Total genes	1,117,803
Average genome size	6,175
Pangenome size (non-redundant genes)	16,820
Core genome	2,503
Accessory genome	9,108
Unique genes	5,209
Average unique genes/strain	16
Gene families under positive selection	233

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

610 Table 2. KEGG pathways distinctive for the *P. aeruginosa* pangenome gene category

Pathway ko number	Description	Genes
Core genome		
00073	Cutin, suberine and wax biosynthesis	1
00590	Arachidonic acid metabolism	1
00592	alpha-Linoleic acid metabolism	1
00471	D-Glutamine and D-glutamate metabolism	3
00740	Riboflavin metabolism	3
00785	Lipoic acid metabolism	2
00908	Zeatin biosynthesis	1
01051	Biosynthesis of ansamycins	1
01055	Biosynthesis of vancomycin group antibiotics	1
00332	Carbapenem biosynthesis	2
00401	Novobiocin biosynthesis	2
00642	Ethylbenzene degradation	1
00791	Atrazine degradation	1
00983	Drug metabolism	4
03008	Ribosome biogenesis in eukaryotes	1
04011	MAPK signalling pathway - yeast	1
04070	Phosphatidylinositol signalling system	3
04152	AMPK signalling pathway	1
04210	Apoptosis	1
04910	Insulin signalling pathway	1
04920	Adipocytokine signalling pathway	1
04918	Thyroid hormone signalling pathway	1
04964	Proximal tubule bicarbonate reclamation	1
05230	Central carbon metabolism in cancer	1
05231	Choline metabolism in cancer	1
05205	Proteoglycans in cancer	1
05204	Chemical carcinogenesis	2
05203	Viral carcinogenesis	1
05010	Alzheimer's disease	3
05012	Parkinson's disease	2
05014	Amyotrophic lateral sclerosis	1
04930	Type II diabetes mellitus	1
04932	Non-alcoholic fatty liver disease	2
05111	Vibrio cholerae pathogenic cycle	3
05132	Salmonella infection	2
05142	Chagas disease	1
05143	African trypanosomiasis	1
01502	beta-lactam resistance	13
Unique genes		
00120	Primary bile acid biosynthesis	2
00534	Glycosaminoglycan biosynthesis - heparan sulfate	1
00901	Indole alkaloid biosynthesis	1
00621	Dioxin degradation	5
00984	Steroid degradation	2
03450	Non-homologous end-joining	1
04010	MAPK signalling pathway	1
04150	mTOR signalling pathway	1

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

04114	Oocyte meiosis	1
04728	Dopaminergic synapse	1
04726	Serotonergic synapse	1
04720	Long-term potentiation	1
04722	Neurotrophin signalling pathway	1
05030	Cocaine addiction	1
05031	Amphetamine addiction	1
05034	Alcoholism	1
Accessory genome		
00906	Carotenoid biosynthesis	1

Catalogue of the KEGG pathways (ko) distinctively found in three gene categories of the *P. aeruginosa* pangenome: core, accessory, and unique genes. The number of pathways correlated with those numbers presented in Figure 2 (Venn diagram on the right).

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

615 Table 3. Domain enrichment in proteins under positive selection

SMART/Pfam Domain	Description	Fisher's test
Chromate_transp	Probably act as chromate transporters in bacteria	0.0000
Sulfatase	Present in esterases hydrolysing steroids, carbohydrates and proteins	0.0020
PepSY_TM	Conserved transmembrane helix found in bacterial protein families	0.0041
PrmA	Present in the Ribosomal protein L11 methyltransferase	0.0123
Cons_hypoth95	Present in 16S RNA methyltransferase D	0.0166
MTS	Present in the 16S RNA methyltransferase C	0.0182
DUF1329	Putative outer membrane lipoprotein	0.0215
DUF4102	Putative phage integrase	0.0235
CHASE	Extracellular domain of bacterial transmembrane receptors	0.0284
G3P_acyltransf	Enzymes converting glycerol-3-phosphate into lysophosphatidic acid	0.0284
AceK	Bacterial isocitrate dehydrogenase kinase/phosphatase protein	0.0284
Choline_sulf_C	C-terminus of enzyme producing choline from choline-O-sulfate	0.0284
DUF2165	Unknown function	0.0284
DUF2909	Unknown function	0.0284
DUF3079	Unknown function	0.0284
DUF444	Unknown function	0.0284
DUF533	Unknown function; integral membrane protein	0.0284
DUF791	Unknown function	0.0284
DUF972	Unknown function	0.0284
Glu_cys_ligase	Enzyme carrying out the first step of glutathione biosynthesis	0.0284
Herpes_UL6	Present in proteins similar to herpes simplex UL6 virion protein	0.0284
His_kinase	Membrane sensor, a two-component regulatory system	0.0284
Inhibitor_I42	Protease inhibitor	0.0284
PPDK_N	Present in enzymes catalysing the conversion of pyrophosphate to PEP	0.0284
Sigma54_AID	Activating interacting domain of the Sigma-54 factor	0.0284
Sigma54_CBD	Core binding domains of the Sigma-54 factor	0.0284
Sigma54_DBD	DNA binding domain of the Sigma-54 factor	0.0284
PAS, PAS 4/9	Present in signalling proteins working as signal sensors	0.0330
MFS	Major Facilitator Superfamily of small molecule transporters	0.0359
Autoind_synth	Autoinducer synthase involved in quorum-sensing response	0.0423
AzIC	Putative protein involved in branched-chain amino acid transport	0.0423
Chitin_bind	Present in carbohydrate-active enzymes (glycoside hydrolases)	0.0423
DUF3299	Unknown function	0.0423
PTS_EIIC / IIB	Phosphoenolpyruvate-dependent phosphotransferase system	0.0423
TctC	Member of the tripartite tricarboxylate receptors	0.0423
UPF0004	Domain found in tRNA methythiotransferases	0.0423

The SMART and Pfam domains are presented in a non-redundant manner. Function description was recovered from annotations in SMART or Pfam databases. Fisher's test values correspond to p-values ($p \leq 0.05$), supporting the over-representation of the corresponding domain in the set of proteins under positive selection.

620

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 4. Potential genetic markers for MLST in *P. aeruginosa* strains

Gene Family[a]	Function[b]	Omega (ω)	Length (bp)	Strain Frequency[c]
3333	Chitin binding protein	108	1,170	98.9% (179)
3675	Flagellar basal-body rod protein FlgF	5,884	750	99.5% (180)
4766	Predicted branched-chain amino acid permease AzIC	86	763	96.7% (175)
5348	Unknown function	32	573	99.5% (180)

a Nomenclature according to pangenome gene inventory.

b Function inferred from KEGG, SMART, and/or BLAST-based search.

c Number of strains carrying respective genes are denoted in parenthesis.

625

bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure legends

Figure 1. Genes in the *P. aeruginosa* pangenome. A - Plot for the number of new genes contributed to the pangenome as new strains are analysed. Data perfectly fit the decay function (R^2 = 0.9519), indicating a constant rate of genes introduced to the pangenome (0.071) and reaching a functional plateau at ~160. The dashed line indicates the function's non-linear regression curve. B - Histogram for the prevalence of different gene families of the pangenome. The 16,820 non-redundant gene families determined to be present in the *P. aeruginosa* pangenome were distributed according to their frequency across all strains analysed. Three gene categories are clearly distinguished, highlighting the core genome (gene families present in all strains analysed), the unique genes (genes present in only one strain), and the accessory genome (gene families exhibiting a variable frequency).

Figure 2. Functional annotation of the pangenome according to gene family categorization. Two Venn diagrams are presented, indicating the functional annotation at the orthology level (left diagram) and molecular pathway level (right diagram) for the three different categories established in concordance with gene frequency across strains. The redundancy of functions was predominantly at the pathway level and permitted to discern distinctive elements for each gene category. Those distinctive pathways are listed in Table 2.

Figure 3. Molecular evolution of the *P. aeruginosa* pangenome. A - Histogram showing the distribution of omega (ω) values across the *P. aeruginosa* pangenome. The light blue histogram shows the original distribution with the tendency towards values indicating purifying selection (shift to left from neutrality). The superposed light red histogram

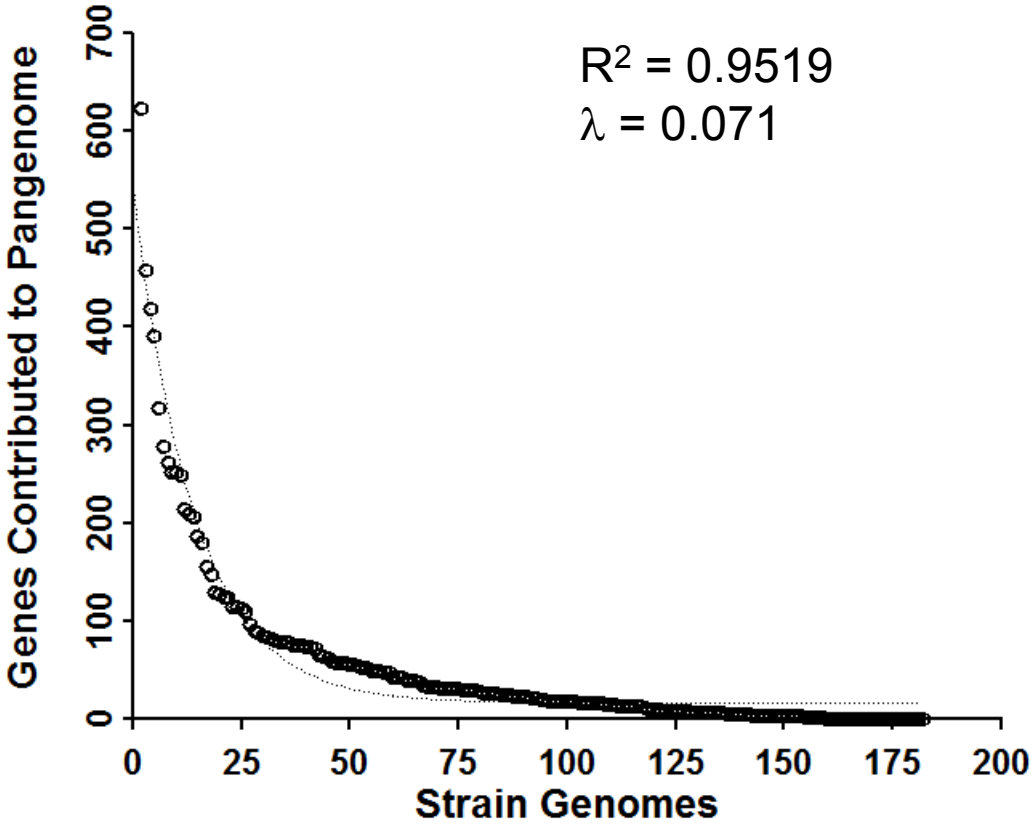
bioRxiv preprint doi: https://doi.org/10.1101/020305; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

650 indicates the Z-scores for the selection of genes with ω significantly different than 1. Those with significant ω < 1 were considered to be under strong purifying selection for functional analysis, and those with significant ω > 1 were selected to be under strong positive selection for the MLST approach. B - Scatter plot to represent the distribution of normalized dN and dS rates for all gene families detected in the *P. aeruginosa* pangenome.
655 Gene families under strong purifying selection are highlighted in blue, whereas gene families under positive selection (ω > 2) are highlighted in red. The set of gene family candidates for MLST under strong positive selection are highlighted in green. The diagonal dashed line indicates the boundary for neutrality.

660 **Figure 4.** Circular phylogenetic tree showing the genetic relationships among 170 reference PATRIC strains and our six *P. aeruginosa* isolates. The phylogenetic tree was built from the best evolutionary model explaining evolution at the concatenated gene families 3333, 3675, 4766, and 5348 after a sequential likelihood ratio test [44]. A total of 176 *P. aeruginosa* strains are located in the tree, and the localization of our clinical isolate
665 is indicated. A close view of this tree permitted us to infer relationships among our clinical isolates with virulent and multi-drug resistant strains.

Figure 1

A


R^2 = 0.9519
λ = 0.071
Genes Contributed to Pangenome
0
100
200
300
400
500
600
700
0
25
50
75
100
125
150
175
200
Strain Genomes

B

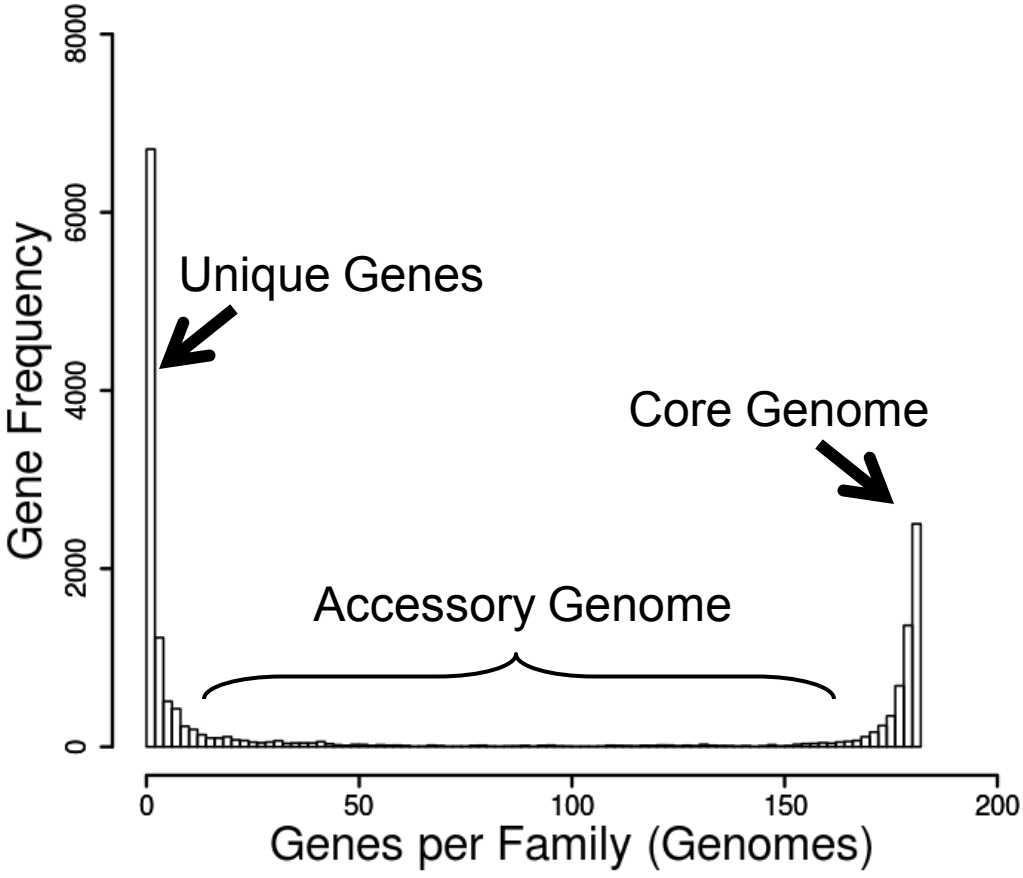

Gene Frequency
0
2000
4000
6000
8000
Unique Genes
Core Genome
Accessory Genome
0
50
100
150
200
Genes per Family (Genomes)

Figure 2

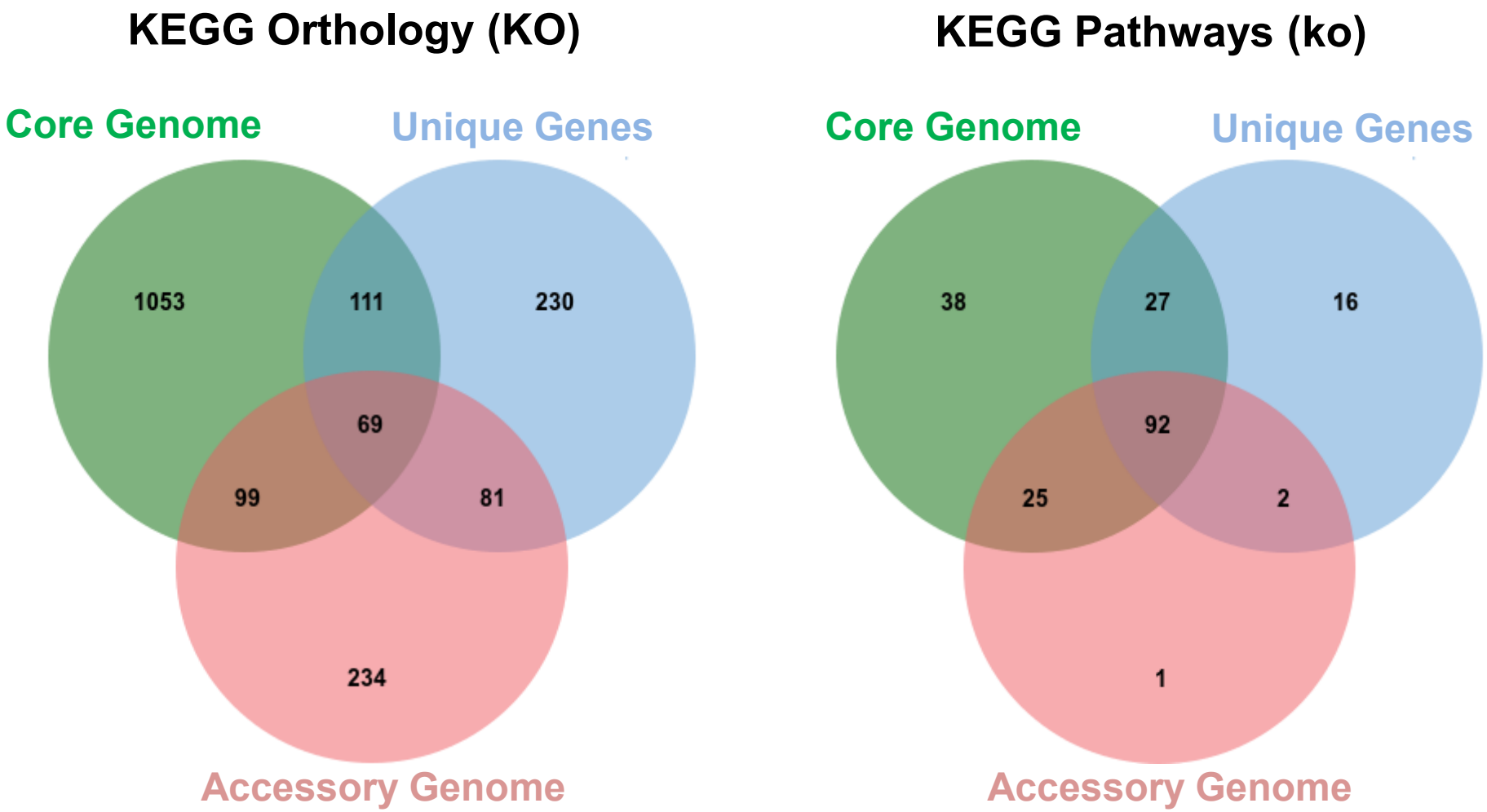
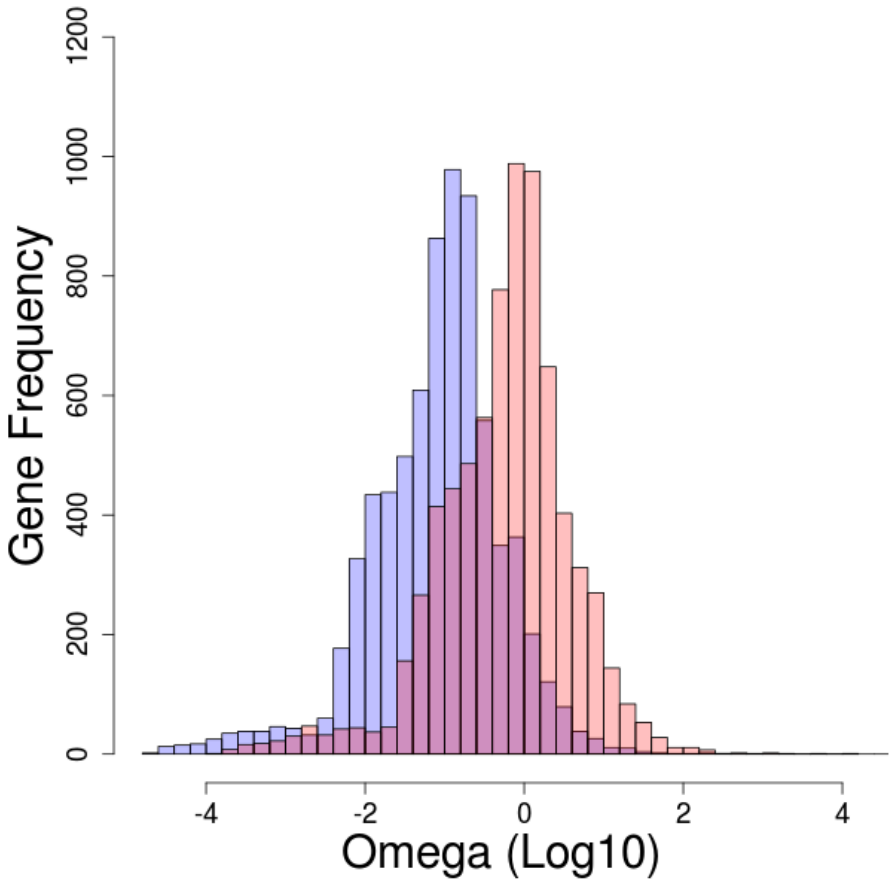

KEGG Orthology (KO)
Core Genome
Unique Genes
1053
111
230
69
99
81
234
Accessory Genome
KEGG Pathways (ko)
Core Genome
Unique Genes
38
27
16
92
25
2
1
Accessory Genome

Figure 3

A


Gene Frequency
0
200
400
600
800
1000
1200
-4
-2
0
2
4
Omega (Log10)

B

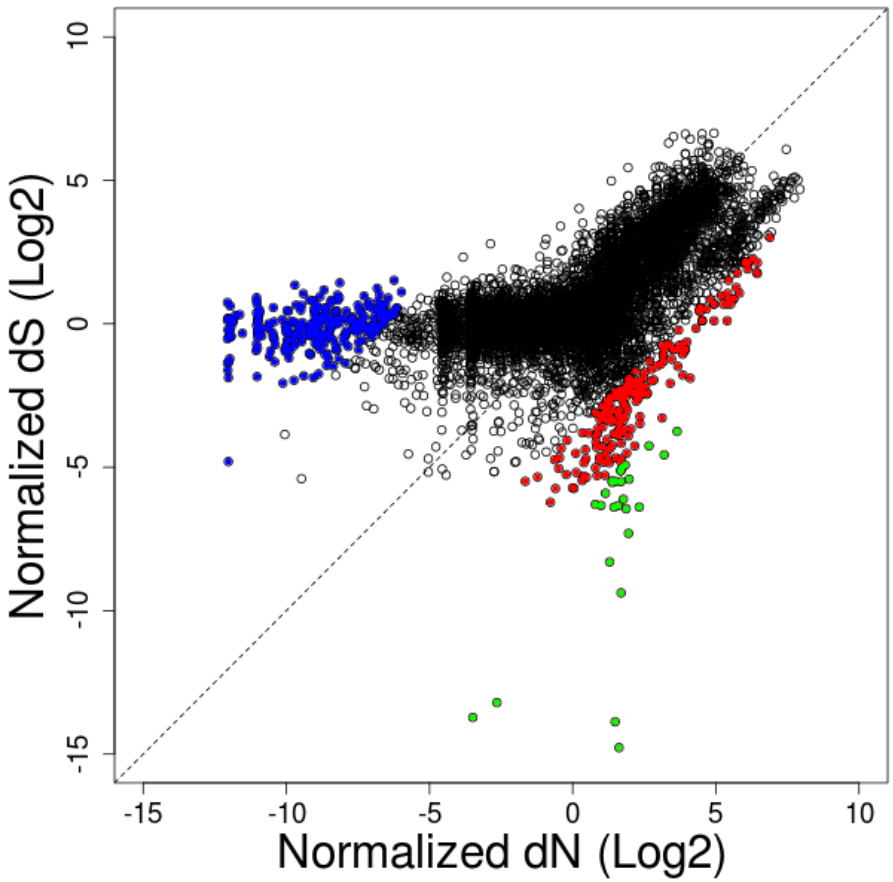
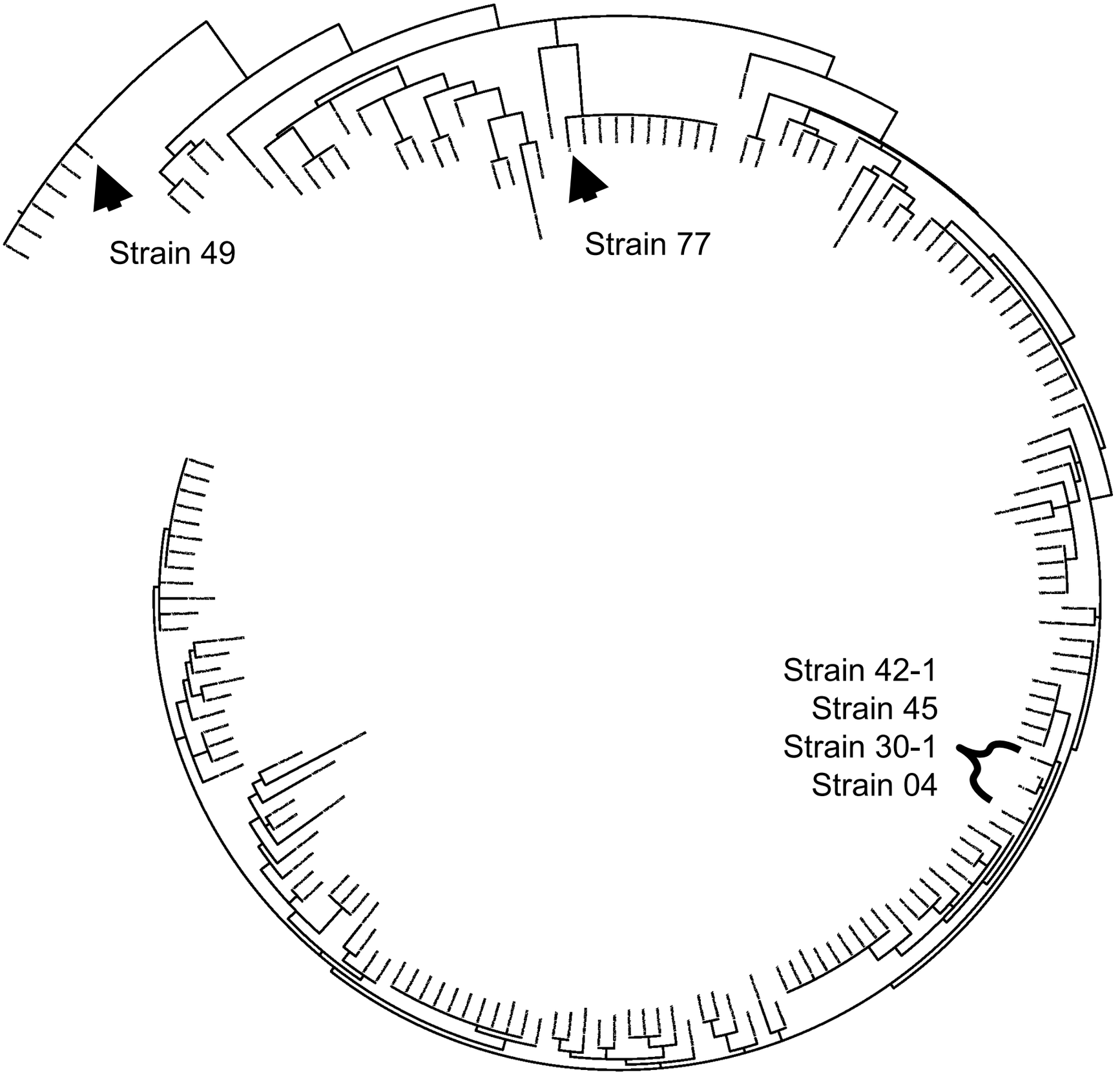

Normalized dS (Log2)
10
5
0
-5
-10
-15
-15
-10
-5
0
5
10
Normalized dN (Log2)

Figure 4


Strain 49
Strain 77
Strain 42-1
Strain 45
Strain 30-1
Strain 04