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[1,5,6]

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. E-mail abenitez@cidbio.org.

6 Current address: Microbial Ecology, Nutrition & Health Research Unit. Agrochemistry and Food Technology Institute (IATA-CSIC). 46980 Paterna-Valencia, Spain. E-mail abenitez@iata.csic.es.

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

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

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

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 more than 16,000 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

55 use of protein conjugates from pili, flagella, or secretion systems as antigens for vaccine
56 design, which exhibit high genetic variation in terms of non-synonymous substitutions in
57 *P. aeruginosa* strains.
58
59
60

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* infections 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

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].

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 and Additional File 1). The preliminary analysis of the *P. 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 of genomic and functional features of the *P. aeruginosa* pangenome, we performed general methods described by Tettelin and co-workers to define bacterial pangenomes [15]. After an iterative and combinatorial process, our observed data was plotted as rarefaction curves following Heaps' law (Figure 1A). 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 *P. 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 graphical representation of the discovery rate for new genes at the core

133 genome across the iterative analysis of *P. aeruginosa* strains for pangenome reconstruction
134 is shown in Figure 1B. We analyzed such data with power law ($n = \kappa N^{-\alpha}$) finding and
135 averaged alpha parameter of 2.36 ± 0.49 (CI = 2.27 to 2.46) indicating the *P. aeruginosa*
136 pangenome is closed according to proposed postulates of Tettelin and co-workers [15].
137 Moreover, interpretation of these data with the exponential regression allowed to estimate
138 and horizontal asymptote (θ) 10 genes ± 2.99, indicating a small but finite number of new
139 genes expected to be discovered with the study of new *P. aeruginosa* studies. A
140 preliminary analysis regarding the distribution of some genes involved in lung infections
141 like the biofilm-associated (*mifS*, *mifR*, *bamI*, *bdlA*, *bfiS*, and *bfmR*) and antibiotic
142 resistance genes (*oprM, ampC, ampD*, and PIB-1), and functionally annotated in the
143 Pseudomonas Genome Database [16], has revealed that these genetic entities are present
144 between 95 to 100% of the strains studied here. This indicates that such functions, and
145 pathogenicity by extension, are encoded into the core genome of *P. aeruginosa*.
146
147 The set of genes, which were not included in the core genome or were unique (present
148 in 1 genome), were referred to as the accessory genome; it included the 54% of genes
149 found in the *P. aeruginosa* pangenome (Table 1). Interestingly, when we plotted the
150 frequency of all pangenome genes present in different strains/genomes analysed (Figure
151 1C), we found a similar distribution to that reported by Lapierre and Gogarten when they
152 estimated the pangenome for more than 500 different bacterial genomes [17]. This
153 distribution plot clearly demonstrated the characteristic distribution and frequency of
154 different groups of the above-stated genes. In general terms, the *P. aeruginosa* pangenome
155 exhibits a high level of genome variability, whereby only 40% (2,503/6,175) of its genome
156 is constant, on average. Thus, the remaining 60% of *P. aeruginosa* genome is presented as
157 a variable piece of DNA composed of a wide repertoire of genes and molecular functions.

158 A very recent study has partially characterized the *P. aeruginosa* pangenome using a total
159 of 20 different human- and environmental-derived strains. Their numbers in terms of
160 average genome size and ORFs per strains are very close to those we show in the present
161 study. However, they estimate the *P. aeruginosa* pangenome to have 13,527 with more
162 than 4,000 genes catalogued as the core genome [18]. The pangenome estimated in our
163 study exceeds by more than 3,000 genes to that reported by Hilker and co-workers as well
164 as to that reported by Valot and co-workers [19]. This is totally expected given that the
165 more genomes analyzed, the more probability to discover new genes, an assumption that is
166 clearly exemplified in the Figure 1A. Conversely, the core genome appear to be negatively
167 affected by addition of new strains because the probability of sharing genes among strains
168 decreases as new strains are incorporated to the study sample. This parameter intuitively is
169 directly dependant of the number of strains used to calculate the core genome and their
170 clonal relationship, which could strongly reduce gene diversity in the pangenome. Given
171 that the multi-strain, iterative and combinatorial process used here to estimate the *P.*
172 *aeruginosa* pangenome has produced a closed pangenome, we proposed that core genome
173 for *P. aeruginosa* is composed of approximately 2,500 genes. This number is notably
174 lower than those proposed in very recent studies aiming the characterization of the *P.*
175 *aeruginosa* pangenome as well [18-20]. However, none of those studies have produced a
176 proper metrics indicating that their proposed pangenomes are closed. Therefore, our data
177 represent the most accurate characterization of the *P. aeruginosa* pangenome supported in
178 the analysis of more than 180 different strains throughout iterative and combinatorial
179 approaches. Moreover, the metrics presented in here is very close to that early
180 characterized for *Escherichia coli*, for which a core genome was defined to account 2,200
181 genes [21].
182

Subsequently, we proceeded to perform a functional analysis with the full set of genes uniquely presented as well as other set of genes categorized by frequency in the *P. aeruginosa* pangenome. As a consequence, the nucleotide sequences of genes found to be present only in one *P. aeruginosa* strain were translated to amino acid sequences and then submitted to the Kyoto Encyclopedia of Genes and Genomes (KEGG) through the KEGG Automatic Annotation Server (KASS) for functional annotation at the protein level [22]. We retrieved only 14% (738 out of 5,209) of the functional annotation for this set of genes, of which more than 59% (3,075 out of 5,209) comprises ORFs, encoding putative peptides shorter less than 100 aa in length. We explored the predominance of functions present in the 738 ORFs annotated at the KEEG 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. Since such functional categories are found uniquely in different strains, this fact would support the idea that *P. aeruginosa* strains exhibit a wide variety of mechanisms to survive in several adverse environments being able to remain latently as reservoir of these genetic traits. Furthermore, this would have direct implication in emergence of multi-resistant and virulent strains since such genetic traits could all converge into single strains by horizontal transference mechanisms.

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 unique genes and 49% in the accessory genome. Previous studies have shown similar results in terms of functional categories of core and accessory genomes partially defined for *P. aeruginosa*, where core genome is enriched in central metabolism functions and major cellular functions such as replication, transcription, and translation as well as other associated biosynthetic pathways [19, 20]. At the KEGG functional module level, we disclosed 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. Therefore, functional characterization of the *P. aeruginosa* core genome would indicate that the infectivity and resistance are features intrinsically

exhibited by any *P. aeruginosa* strain and that virulence and lethality would be confined to genetic traits encoded in the accessory genome.

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 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 [23]. Attempts to distinguish the specific role of these proteins through homologue searching in the Uniprot database have retrieved unclear results given that amino acid sequences of VirB2 and VirB9 from *P. aeruginosa* pangenome matched primarily with conjugal transfer proteins from *A. tumefaciens* (identity ~40% over 70% of the protein length), but also with toxin liberation protein F from *B. pertussis* (identity ~28% over 80% of the protein length). In any event, the VirB2 and VirB9 proteins must be exposed on the cell surface of pathogens making possible the *P. aeruginosa* be recognized by the host immune system and triggering 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. Notwithstanding, we cannot discard these high rates of non-synonymous substitutions appear as response of phage predation. In this last scenario, the information retrieved in the present study regarding the set of genes under strong positive selection can be also useful to design bacteriophage-based therapies which have already been tested in *P. aeruginosa* [24]. 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 [25, 26], 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 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 using the Simple Modular Architecture Research Tool (SMART) and the Protein Family database (Pfam) nomenclature systems. 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

306 them, we observed a high frequency of membrane-bound proteins acting as transporters or
307 receptors. Some of the functions over-represented in Table 3 agree with some stated from
308 previous analyses in which membrane proteins (transporters and/or receptors) as well as
309 the Sigma-54 factor seem to be under positive selection in *P. aeruginosa*. Interestingly, we
310 observed the presence of proteins related with either 16S RNA and ribosomal protein
311 methylation (Table 3). We detected such patterns of molecular evolution in this class of
312 proteins previously, but in different human pathogens [27]. Although we cannot shed light
313 on the meaning of this type of evolution in these proteins given their function, we
314 hypothesized that they might influence the translation process to modulate the expression
315 of a certain set of proteins directly or indirectly involved in pathogenesis. Recent studies
316 on rRNA methylation indicate that they play a meaningful role in decoding function [28-
317 30]. Indeed, some of them have been directly involved with virulence [31].
318
319 When we attempted a similar analysis in a counterpart set of proteins under purifying or
320 negative selection ($\omega < 1$), the biased distribution of omega values across the *P.*
321 *aeruginosa* pangenome (Figure 3A) made it difficult to set up a suitable threshold to
322 recover proteins under this type of selection. Therefore, we obtained Z-scores of both the
323 dN and dS rates (Figure 3A, light red histogram), thus reaching a normal distribution
324 around $\omega = 1$ (neutrality). Using this normalized distribution of ω values, we could
325 determine those genes with evolution significantly different ($p \leq 0.05$) from neutrality ($\omega =$
326 1) towards a strong negative selection (lowest ω values). As a result, we found a group of
327 268 proteins/genes under negative selection, the dN and dS rates of which are plotted in
328 Figure 3B (see the blue points distribution). The quantitative assessment to determine
329 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

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 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 ≥ 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

355 information with that completely available for 170 *P. aeruginosa* strains, thus building a
356 multiple sequence alignment almost 3,000 bp in length for a total of 176 different strains.
357 Using maximum likelihood approaches, we reconstructed the phylogenetic relationships
358 among all strains and retrieved the phylogenetic tree showed in Figure 4. Our six local
359 isolates were positioned in three different clades, where isolate 49 was closely related to
360 the highly virulent *P. aeruginosa* PA14 strain, representing the most common clonal group
361 worldwide [32]. By contrast, isolate 77 was related to several strains, including the multi-
362 drug-resistant *P. aeruginosa* NCGM2.S1 [33] and the cytotoxic corneal isolate *P.*
363 *aeruginosa* 6077 [34]. Finally, the 30-1, 42-1, 45, and 04 isolates presented a close
364 relationship and were related to the multi-drug resistant *P. aeruginosa* VRFPA02 isolate
365 from India [35].
366
367 Based on the best evolutionary model fitted to the nucleotide substitution pattern
368 observed for these markers (TrN+I+G), a proportion of invariable sites of 0.9080 was
369 obtained, thus indicating that more than 250 polymorphic sites are present in our MLST
370 approach. Moreover, gamma distribution parameters (0.5060) is indicative of few hot-spots
371 with high substitution rates [36]. In this fashion, we provided support to use the highly
372 variable genetic markers reported here for MLST to produce an initial, fast, and cost-
373 effective genotyping for *P. aeruginosa* strains of clinical interest. To compare if the
374 evolutionary history of *P. aeruginosa* strains is equally represented by of our proposed
375 MLST markers in comparison with that inferred by using common MLST markers [37,
376 38], we reconstructed a phylogeny using similar approaches and DNA sequences
377 corresponding to seven housekeeping genes: *acsA, aroE, guaA, mutL, nuoD, ppsA,* and
378 *trpE*. The resulting tree showed not deep topology differences when compared to that
379 created from our proposed MLST approach (data not shown). This indicate that the new

molecular markers proposed in this study for genotyping aims could be used to infer the evolutionary history of *P. aeruginosa* strains.

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 performed a reverse vaccinology approach using a large amount of genetic information available in the PATRIC database to determine the genetic elements of *Pseudomonas aeruginosa* to be probably targeted in future clinical studies aiming new vaccine designs. We have extensively described the *P. aeruginosa* pangenome in terms of the effective number of non-redundant genes present in this bacterial species by analysing more than 180 different strain genomes. We outlined the genomic variability of this human pathogen, demonstrating that approximately 60% of the *P. aeruginosa* genome is 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

404 from a wide spectrum of antibiotic resistances to a specialized secretion system delivering
405 toxins and effector proteins potentially harmful for host cells. However, pathogenicity
406 traits were also observed in the distinctive KEGG pathways revealed for the core genome.
407
408 Although this is not the first report to describe the pangenome for a single bacterial
409 species [14, 21, 39, 40], and other very recent studies have attempted to determine the *P.*
410 *aeruginosa* pangenome [18-20], this report is the first to describe a closed *P. aeruginosa*
411 pangenome at very large scale, including almost 200 bacterial genomes from this human
412 pathogen and performing a pangenome-scale molecular evolutionary analysis. Our study
413 fits well with previous and general genomic characterizations of this human pathogen [18-
414 20, 41], and it definitely expands our knowledge about the evolutionary mechanisms of *P.*
415 *aeruginosa* pathogenesis. This study aimed to reveal the evolutionary processes occurring
416 at the pangenome level in *P. aeruginosa* that could explain the failure to design and
417 develop of a successful vaccine against this human pathogen as well as provide an
418 understanding of the molecular mechanisms that drive the evasion of the host immune
419 system. We observed that the *P. aeruginosa* genome is globally under purifying selection,
420 given the distribution of omega values (ω = dN/dS, median ~0.1) discerned for every gene
421 family present in its pangenome. This result was further supported by the finding that there
422 are 10-fold more genes under strong purifying selection than strong positive selection
423 (significantly different to neutrality, $p \leq 0.05$). Although we found that the *P. aeruginosa*
424 pangenome evolves to purifying selection as a whole, we distinguished some genes and
425 functions predominantly present in the reduced set of genes under positive selection. As a
426 consequence, a considerable number of proteins located at the outer membrane, such as
427 those associated with receptor and transporter functions, were identified to have an
428 increased rate of non-synonymous substitutions. 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.

For the first time, pangenome-scale evolutionary information is presented to support the design of new *P. aeruginosa* vaccines. 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, a fact that has been described previously but at very lower scale [42, 43].

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. We demonstrated that these markers contain an adequate proportion of hotspots for variation, exhibiting high nucleotide substitution rates. Using these four loci, we discerned the genetic identity of 6 local isolates of *P. aeruginosa* and related them with the resistance and virulence traits carried in reference strains.

Methods

Pangenome-wide analysis

Genome information from *P. 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 using iterative and combinatorial approaches [14, 15, 17]. Briefly, a BLAST-based iterative method was used to extract the full set of non-redundant

453 genes representing the *P. aeruginosa* pangenome. A single iteration consisted in a random
454 selection of a strain as pangenome primer, then the remaining set of strain were randomly
455 incorporated to the pangenome. The above process was calculated over 200 iterations with
456 random permutation of the strain order in every iterative step. A rarefaction curve was
457 plotted with all data generated and consisted in 200 different measures throughout a
458 sequential addition of 181 different strains. Pangenome metrics was also obtained in
459 iterative manner with data fitting to the power law as previously stated [15]. Power
460 regression was calculated individually for each iteration of pangenome reconstruction
461 (n=200) and plotted in R v.3.1.2 with the *"igraph"* package (https://cran.r-project.org).
462 Alpha parameter from the $n = \kappa N^{-\alpha}$ power regression, indicating whether pangenome is
463 open or closed, was calculated individually with least squares fit of the power law to the
464 number of new genes discovered at core genome according to tettelin and coworkers [15].
465 Therefore, the global alpha value for the *P. aeruginosa* pangenome was determined as the
466 mean of all 200 different alpha values generated ± sd with the confidence interval at 0.95
467 level. Finally, the set of non-redundant genes obtained was used to explore their
468 occurrence pattern in the 181 *P. aeruginosa* genomes through BLASTN-based
469 comparisons [44, 45].
470
471 *Molecular evolution analysis*
472 The full set of ORFs constituting the *P. aeruginosa* pangenome was used to search
473 homologues in all genomes analysed, and multiple sequence alignments were built using
474 refined and iterative methods [46, 47]. The synonymous and non-synonymous substitution
475 rates were calculated in a pairwise manner using approximate methods [48] and by
476 correcting for multiple substitutions [49]. Omega values (ω) were computed as the
477 averaged ratio of dN/dS rates from multiple comparisons, and genes under strong positive

478 selection were selected when $\omega \geq 2$. The Z-score of ω values was computed to depict
479 functions of genes under strong purifying selection and potential MLST genetic markers
480 under strong positive selection ($p \leq 0.05$). Large-scale analyses of pairwise comparisons,
481 statistical analysis, and graphics were performed using R v3.1.2 (https://cran.r-project.org).
482
483 *Functional genomics analysis*
484 Functional annotation of genes was performed using the KEGG Automatic Annotation
485 Server for KEGG Orthology [22]. KEGG functional modules and ontologies were explored
486 in the KEGG BRITE database [50]. Functional domains present in genes of interest were
487 assigned using Perl scripting for batch annotation (http://smart.embl-
488 heidelberg.de/help/SMART_batch.pl) against the Simple Modular Architecture Research
489 Tool (SMART) together with Pfam classification [51, 52]. Fisher's exact test with a false
490 discovery rate (FDR) for 2 x 2 contingency tables to measure enrichment of Pfam/SMART
491 domains was performed using R v3.1.2 (https://cran.r-project.org). Venn diagrams were
492 drawn using the *jvenn* server [53].
493
494 *Multi-locus sequence typing*
495 The six *P. aeruginosa* strains (labelled as 04, 30-1, 42-1, 45, 49, and 77) were
496 isolated from patients who acquired healthcare-associated infections at a clinic located
497 outside the metropolitan area of Medellin, Antioquia, Colombia. This study was
498 approved by the ethics committee of the Fundación Clínica del Norte Hospital (Bello,
499 Antioquia, Colombia). The six isolates, previously characterized for multi-drug
500 resistance, were kindly donated to the scientist of the Bacteria & Cancer Researching
501 Group of the Faculty of Medicine, University of Antioquia, Colombia. The genomic
502 DNA from *P. aeruginosa* multi-drug resistant strains was extracted using a

503 GeneJER^{TM}GenomicDNA Purification Kit (Thermo Scientific, Waltham, MA, USA).
504 The reference sequences of *P. aeruginosa* PA01 for the four markers selected to
505 perform MLST were downloaded from a public database [GenBank: AE004091.2:
506 region 930623 to 931822 (family 3333), region 1167488 to 1168237 (family 3675),
507 region 2230183 to 2229425 (familyc 4766), region 2935851 to 2936423 (family 5348)].
508 Primers were designed to amplify the complete sequence of each gene, and the
509 Polymerase Chain Reaction (PCR) proceeded with 28 cycles of amplification using
510 Phusion® High-Fidelity DNA Polymerase (Thermo Scientific, Waltham, MA, USA)
511 and 50 ng of genomic DNA. PCR products were isolated using a GeneJet PCR
512 Purification Kit (Life technologies, Carlsbad, CA, USA), and both strands were
513 sequenced by the Sanger automatic method in an ABI 3730 xl instrument (Stab Vida
514 Inc., Caparica, Portugal). Base calling and genetic variants were manually explored
515 using the delivered *ab1* files and FinchTV viewer (Geospiza Inc. Perkin Elmer,
516 Waltham, MA, USA). Assembled sequences from both strands were obtained and
517 concatenated to respective reference sequences obtained from the PATRIC genomes
518 analysed. Sequences belonging to the respective gene family were aligned using
519 iterative methods [46, 47], and alignments were concatenated to perform phylogenetic
520 analysis. The sequential likelihood ratio test was carried out to detect the evolutionary
521 model that better explained genetic variation in all genes integrated in the MLST
522 approach. For that reason, we used the jModelTest tool [54], and model selection was
523 completed by calculating the corrected Akaike Information Criterion (cAIC). The
524 MLST tree was constructed using the Interactive Tree Of Life (iTOL) tool [51, 55] and
525 the phylogeny obtained using the TrN+I+G model. For comparisons aims, we compiled
526 genetic information from seven MLST markers commonly used in *P. aeruginosa*
527 genotyping being the housekeeping genes: *acsA, aroE, guaA, mutL, nuoD, ppsA,* and

trpE [37, 38]. Aligned sequences were concatenated and phylogenetically analyzed with the jModelTest tool as well. Tree topology generated from this conventional MLST markers was compared with that obtained using the new MLST markers proposed in this study.

Availability of supporting data

The features of the Pseudomonas aeruginosa strains used in this study are included in the Additional File 1. All the DNA sequences derived from PCR amplification and Sanger sequencing of the four MLST studied here for the *P. aeruginosa* clinical isolates were submitted to the GenBank through BankIt server [GenBank: KU214214 to KU214237].

Competing interests

The authors declare that they have no competing interests.

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 performed PCR techniques. JMR and SCB curated the sequences from Sanger automatic sequencing. JMR, SCB, and ABP prepared the manuscript. 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 laboratory staff, who donated the clinical isolates for the MLST studies. The JMR M.Sc. fellowship was supported by the Colombian Agency for Science, Technology, and Innovation (Colciencias) with funds of the 5817-5693-4856 grant.

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.
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. Tettelin H, Riley D, Cattuto C, Medini D: Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 2008, 11(5):472-477.
16. Winsor GL, Lam DK, Fleming L, Lo R, Whiteside MD, Yu NY, Hancock RE, Brinkman FS: Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. *Nucleic Acids Res* 2011, 39(Database issue):D596-600.
17. Lapierre P, Gogarten JP: Estimating the size of the bacterial pan-genome. *Trends Genet* 2009, 25(3):107-110.
18. Hilker R, Munder A, Klockgether J, Losada PM, Chouvarine P, Cramer N, Davenport CF, Dethlefsen S, Fischer S, Peng H *et al*: Interclonal gradient of virulence in the Pseudomonas aeruginosa pangenome from disease and environment. *Environ Microbiol* 2015, 17(1):29-46.
19. Valot B, Guyeux C, Rolland JY, Mazouzi K, Bertrand X, Hocquet D: What It Takes to Be a Pseudomonas aeruginosa? The Core Genome of the Opportunistic Pathogen Updated. *PLoS One* 2015, 10(5):e0126468.
20. Ozer EA, Allen JP, Hauser AR: Characterization of the core and accessory genomes of Pseudomonas aeruginosa using bioinformatic tools Spine and AGEnt. *BMC Genomics* 2014, 15:737.
21. 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.
22. 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.
23. 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.
24. Hraiech S, Bregeon F, Rolain JM: Bacteriophage-based therapy in cystic fibrosis-associated Pseudomonas aeruginosa infections: rationale and current status. *Drug Des Devel Ther* 2015, 9:3653-3663.
25. 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.
26. 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.
27. 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.
28. 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.
29. Benitez-Paez A, Villarroya M, Armengod ME: Regulation of expression and catalytic activity of Escherichia coli RsmG methyltransferase. *Rna* 2012, 18(4):795-806.

30. 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.
31. Kyuma T, Kimura S, Hanada Y, Suzuki T, Sekimizu K, Kaito C: Ribosomal RNA methyltransferases contribute to Staphylococcus aureus virulence. *Febs J* 2015.
32. 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.
33. 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.
34. 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.
35. Malathi J, Murugan N, Umashankar V, Bagyalakshmi R, Madhavan HN: Draft Genome Sequence of Multidrug-Resistant Pseudomonas aeruginosa Strain VRFPA02, Isolated from a Septicemic Patient in India. *Genome Announc* 2013, 1(4).
36. Yang Z: Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* 1996, 11(9):367-372.
37. Maatallah M, Bakhrouf A, Habeeb MA, Turlej-Rogacka A, Iversen A, Pourcel C, Sioud O, Giske CG: Four genotyping schemes for phylogenetic analysis of Pseudomonas aeruginosa: comparison of their congruence with multi-locus sequence typing. *PLoS One* 2013, 8(12):e82069.
38. Syrmis MW, Kidd TJ, Moser RJ, Ramsay KA, Gibson KM, Anuj S, Bell SC, Wainwright CE, Grimwood K, Nissen M *et al*: A comparison of two informative SNP-based strategies for typing Pseudomonas aeruginosa isolates from patients with cystic fibrosis. *BMC Infect Dis* 2014, 14:307.
39. 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.
40. 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.
41. Klockgether J, Cramer N, Wiehlmann L, Davenport CF, Tummler B: Pseudomonas aeruginosa Genomic Structure and Diversity. *Front Microbiol* 2011, 2:150.
42. Spangenberg C, Heuer T, Burger C, Tummler B: Genetic diversity of flagellins of Pseudomonas aeruginosa. *FEBS Lett* 1996, 396(2-3):213-217.
43. Winstanley C, Coulson MA, Wepner B, Morgan JA, Hart CA: Flagellin gene and protein variation amongst clinical isolates of Pseudomonas aeruginosa. *Microbiology* 1996, 142 (Pt 8):2145-2151.
44. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215(3):403-410.
45. 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.

46. Edgar RC: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004, 5:113.
47. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004, 32(5):1792-1797.
48. 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.
49. Nei M, Gojobori T: Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 1986, 3(5):418-426.
50. Aoki-Kinoshita KF, Kanehisa M: Gene annotation and pathway mapping in KEGG. *Methods Mol Biol* 2007, 396:71-91.
51. 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.
52. 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.
53. Bardou P, Mariette J, Escudie F, Djemiel C, Klopp C: jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* 2014, 15:293.
54. Darriba D, Taboada GL, Doallo R, Posada D: jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 2012, 9(8):772.
55. 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.

722 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 genes
Pangenome size (non-redundant genes)	16,820
Core genome	2,503 genes
Accessory genome	9,108 genes
Unique genes	5,209
Average unique genes/strain	16
Gene families under positive selection	233

723

724

725

726

727 Table 2. KEGG functional modules distinctive for the *P. aeruginosa* pangenome gene

728 category

KEGG Module Number	Description	Genes
Core Genome		
M00064	ADP-L-glycero-D-manno-heptose biosynthesis	2
M00493	AlgZ-AlgR (alginate production) two-component regulatory system	1
M00235	Arginine/ornithine transport system	3
M00531	Assimilatory nitrate reduction	1
M00475	BarA-UvrY (central carbon metabolism) two-component regulatory system	2
M00086	Beta-Oxidation	1
M00123, M00573, M00577	Biotin biosynthesis	3
M00364, M00366	C10-C20 isoprenoid biosynthesis	2
M00170, M00171	C4-dicarboxylic acid cycle	2
M00168	CAM (Crassulacean acid metabolism)	1
M00722, M00727, M00728	Cationic antimicrobial peptide (CAMP) resistance	3
M00256	Cell division transport system	1
M00010	Citrate cycle	4
M00120	Coenzyme A biosynthesis	3
M00338	Cysteine biosynthesis	1
M00154, M00155	Cytochrome c oxidase	3
M00417	Cytochrome o ubiquinol oxidase	3
M00552	D-galactonate degradation, De Ley-Doudoroff pathway	2
M00596	Dissimilatory sulfate reduction	1
M00542	EHEC/EPEC pathogenicity signature	2
M00008	Entner-Doudoroff pathway	1
M00445	EnvZ-OmpR (osmotic stress response) two-component regulatory system	3
M00515	FlrB-FlrC (polar flagellar synthesis) two-component regulatory system	1
M00729	Fluoroquinolone resistance	1
M00344, M00345	Formaldehyde assimilation	3
M00497	GlnL-GlnG (nitrogen regulation) two-component regulatory system	1
M00605	Glucose/mannose transport system	2
M00012	Glyoxylate cycle	1
M00050	Guanine ribonucleotide biosynthesis, IMP	3
M00259	Heme transport system	1
M00045	Histidine degradation	4
M00226	Histidine transport system	1
M00620	Incomplete reductive citrate cycle	3
M00131	Inositol phosphate metabolism	1
M00190	Iron(III) transport system	4
M00535	Isoleucine biosynthesis	3
M00113	Jasmonic acid biosynthesis	1
M00505	KinB-AlgB (alginate production) two-component regulatory system	2
M00080	Lipopolysaccharide biosynthesis	1
M00320	Lipopolysaccharide export system	2
M00255	Lipoprotein-releasing system	1
M00116	Menaquinone biosynthesis	1
M00740	Methylaspartate cycle	2
M00189	Molybdate transport system	3

M00711	Multidrug resistance, efflux pump MdtIJ	1
M00115	NAD biosynthesis	3
M00144	NADH:quinone oxidoreductase	13
M00471	NarX-NarL (nitrate respiration) two-component regulatory system	1
M00622	Nicotinate degradation	1
M00615	Nitrate assimilation	2
M00438	Nitrate/nitrite transport system	1
M00439	Oligopeptide transport system	1
M00209	Osmoprotectant transport system	2
M00004, M00007	Pentose phosphate pathway	6
M00024	Phenylalanine biosynthesis	3
M00434	PhoR-PhoB (phosphate starvation response)	1
M00222	Phosphate transport system	4
M00501	PilS-PilR (type 4 fimbriae synthesis) two-component regulatory system	1
M00133	Polyamine biosynthesis	3
M00015	Proline biosynthesis	3
M00247, M00258	Putative ABC transport system	3
M00193	Putative spermidine/putrescine transport system	7
M00046	Pyrimidine degradation	1
M00053	Pyrimidine deoxyribonuleotide biosynthesis, CDP/CTP	5
M00052	Pyrimidine ribonucleotide biosynthesis, UMP	3
M00377	Reductive acetyl-CoA pathway (Wood-Ljungdahl pathway)	1
M00167	Reductive pentose phosphate cycle	5
M00523	RegB-RegA (redox response) two-component regulatory system	2
M00125	Riboflavin biosynthesis, GTP	2
M00394	RNA degradosome	1
M00308	Semi-phosphorylative Entner-Doudoroff pathway	2
M00185	Sulfate transport system	2
M00436	Sulfonate transport system	3
M00435	Taurine transport system	1
M00089	Triacylglycerol biosynthesis	2
M00332	Type III secretion system	2
M00025, M00040	Tyrosine biosynthesis	4
M00117, M00128	Ubiquinone biosynthesis	7
M00029	Urea cycle	3
M00651	Vancomycin resistance	2
M00241	Vitamin B12 transport system	1
M00660	Xanthomonas spp. pathogenicity signature	2
M00242	Zinc transport system	2
Accessory Genome		
M00502	GlrK-GlrR (amino sugar metabolism) two-component regulatory system	1
M00533	Homoprotocatechuate degradation	2
M00240	Iron complex transport system	3
M00005	PRPP biosynthesis	1
M00473	UhpB-UhpA (hexose phosphates uptake) two-component regulatory system	1
M00644	Vanadium resistance	1
Unique Genes		
M00653	AauS-AauR (acidic amino acids utilization) two-component regulatory system	1
M00500	AtoS-AtoC (cPHB biosynthesis) two-component regulatory system	1
M00450	BaeS-BaeR (envelope stress response) two-component regulatory system	1
M00104	Bile acid biosynthesis	1
M00581	Biotin transport system	1
M00569	Catechol meta-cleavage	4

M00582	Energy-coupling factor transport system	1
M00760	Erythromycin resistance	1
M00524	FixL-FixJ (nitrogen fixation) two-component regulatory system	1
M00713	Fluoroquinolone resistance	1
M00059	Glycosaminoglycan biosynthesis	1
M00499	HydH-HydG (metal tolerance) two-component regulatory system	1
M00714, M00645	Multidrug resistance	2
M00664	Nodulation	1
M00549	Nucleotide sugar biosynthesis	1
M00267	PTS system, N-acetylglucosamine-specific II component	1

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

733

734 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

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

739

740 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)

741 a Nomenclature according to pangenome gene inventory.

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

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

744

Figure legends

Figure 1. The *P. aeruginosa* pangenome. A - Rarefaction curve of the 200 different pangenomes calculated from random combinations of strains. Iterations and combinations are shown as the dots cloud indicating the total number of non-redundant genes included in the pangenome as genomes are included in the analysis. Red filled circles indicate the median of each iteration. B - Decay function for new genes discovered during pangenome reconstruction. Iterations and combinations are shown as dots cloud indicating the number of new genes incorporated to core genome. Red filled circles indicate the median of each iteration. The power law alpha parameter shown inside the plot is the average of such values retrieved individually in each iteration after fitting ± sd. The theta (θ) value was calculated from the horizontal asymptote where the exponential regression converges. C - 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

768 distinctive elements for each gene category. Those distinctive pathways are listed in Table
769 2.
770
771 **Figure 3.** Molecular evolution of the *P. aeruginosa* pangenome. A - Histogram showing
772 the distribution of omega (ω) values across the *P. aeruginosa* pangenome. The light blue
773 histogram shows the original distribution with the tendency towards values indicating
774 purifying selection (shift to left from neutrality). The superposed light red histogram
775 indicates the Z-scores for the selection of genes with ω significantly different than 1. Those
776 with significant $\omega < 1$ were considered to be under strong purifying selection for functional
777 analysis, and those with significant $\omega > 1$ were selected to be under strong positive
778 selection for the MLST approach. B - Scatter plot to represent the distribution of
779 normalized dN and dS rates for all gene families detected in the *P. aeruginosa* pangenome.
780 Gene families under strong purifying selection are highlighted in blue, whereas gene
781 families under positive selection ($\omega > 2$) are highlighted in red. The set of gene family
782 candidates for MLST under strong positive selection are highlighted in green. The diagonal
783 dashed line indicates the boundary for neutrality.
784
785 **Figure 4.** Circular phylogenetic tree showing the genetic relationships among 170
786 reference PATRIC strains and our six *P. aeruginosa* isolates. The phylogenetic tree was
787 built from the best evolutionary model explaining evolution at the concatenated gene
788 families 3333, 3675, 4766, and 5348 after a sequential likelihood ratio test [54]. A total of
789 176 *P. aeruginosa* strains are located in the tree, and the localization of our clinical isolate
790 is indicated. A close view of this tree permitted us to infer relationships among our clinical
791 isolates with virulent and multi-drug resistant strains.

Figure 1

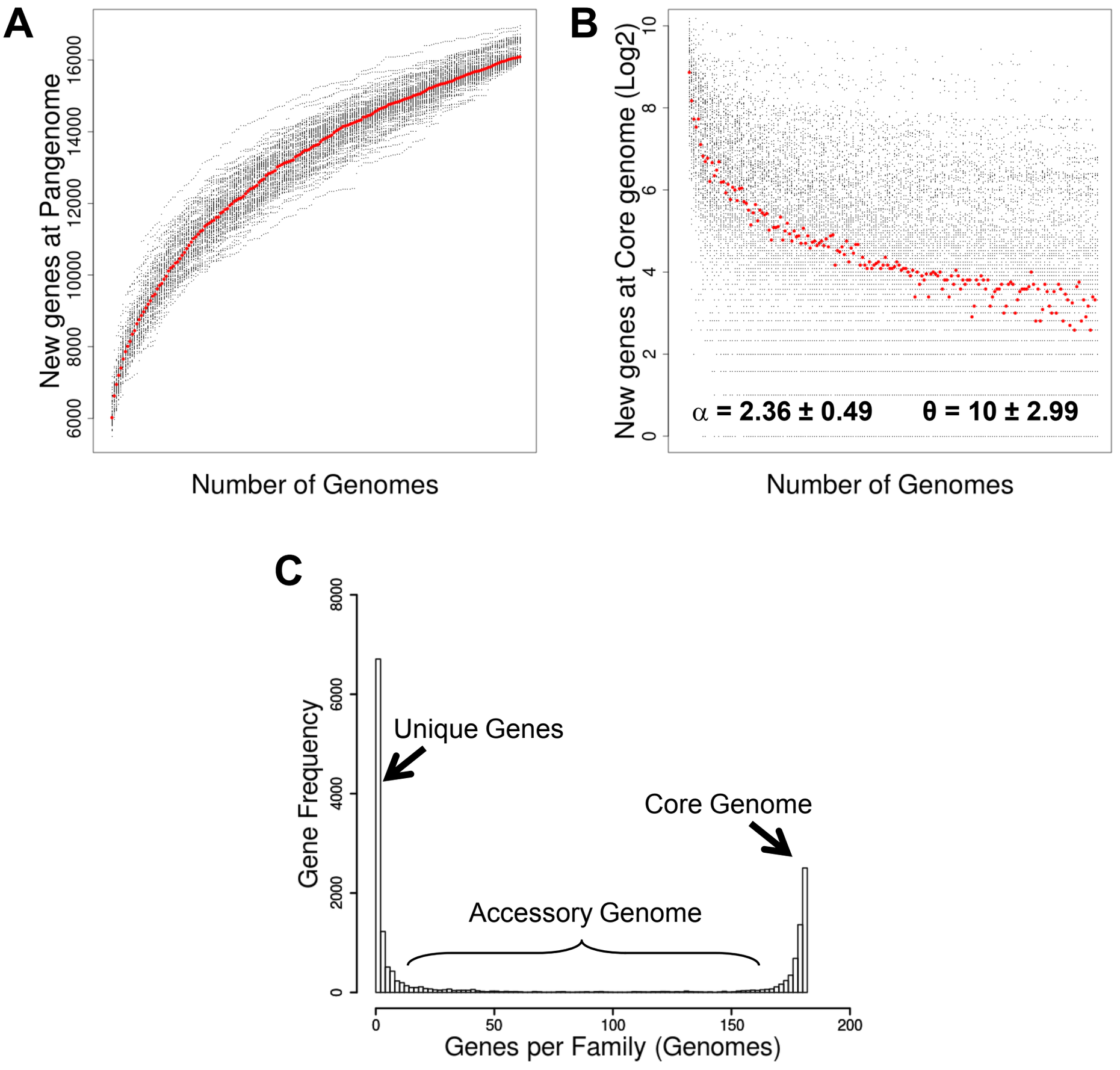

Figure 2

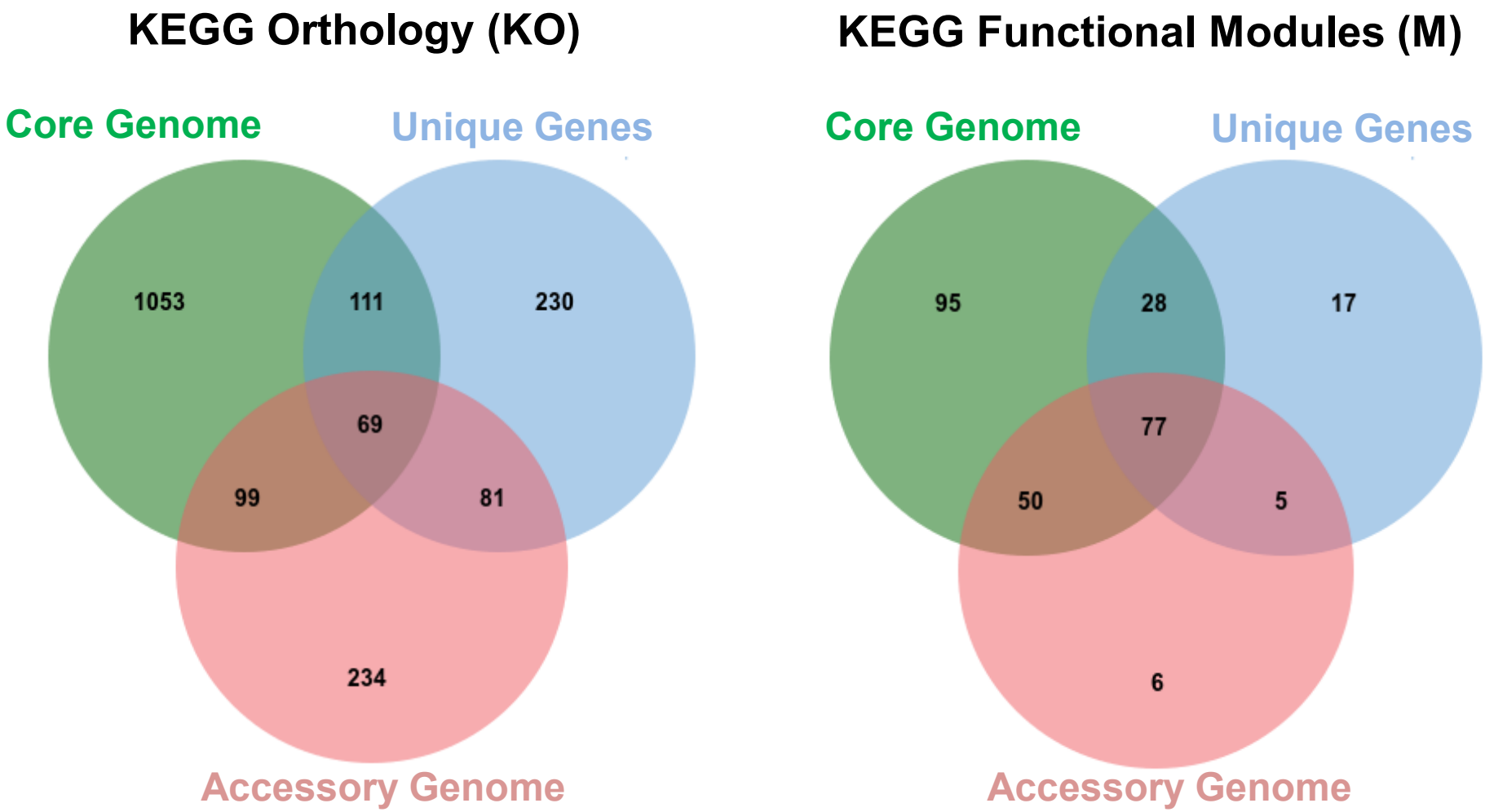
KEGG Orthology (KO)
Core Genome
Unique Genes
1053
111
230
69
99
81
234
Accessory Genome
KEGG Functional Modules (M)
Core Genome
Unique Genes
95
28
17
77
50
5
6
Accessory Genome

Figure 3

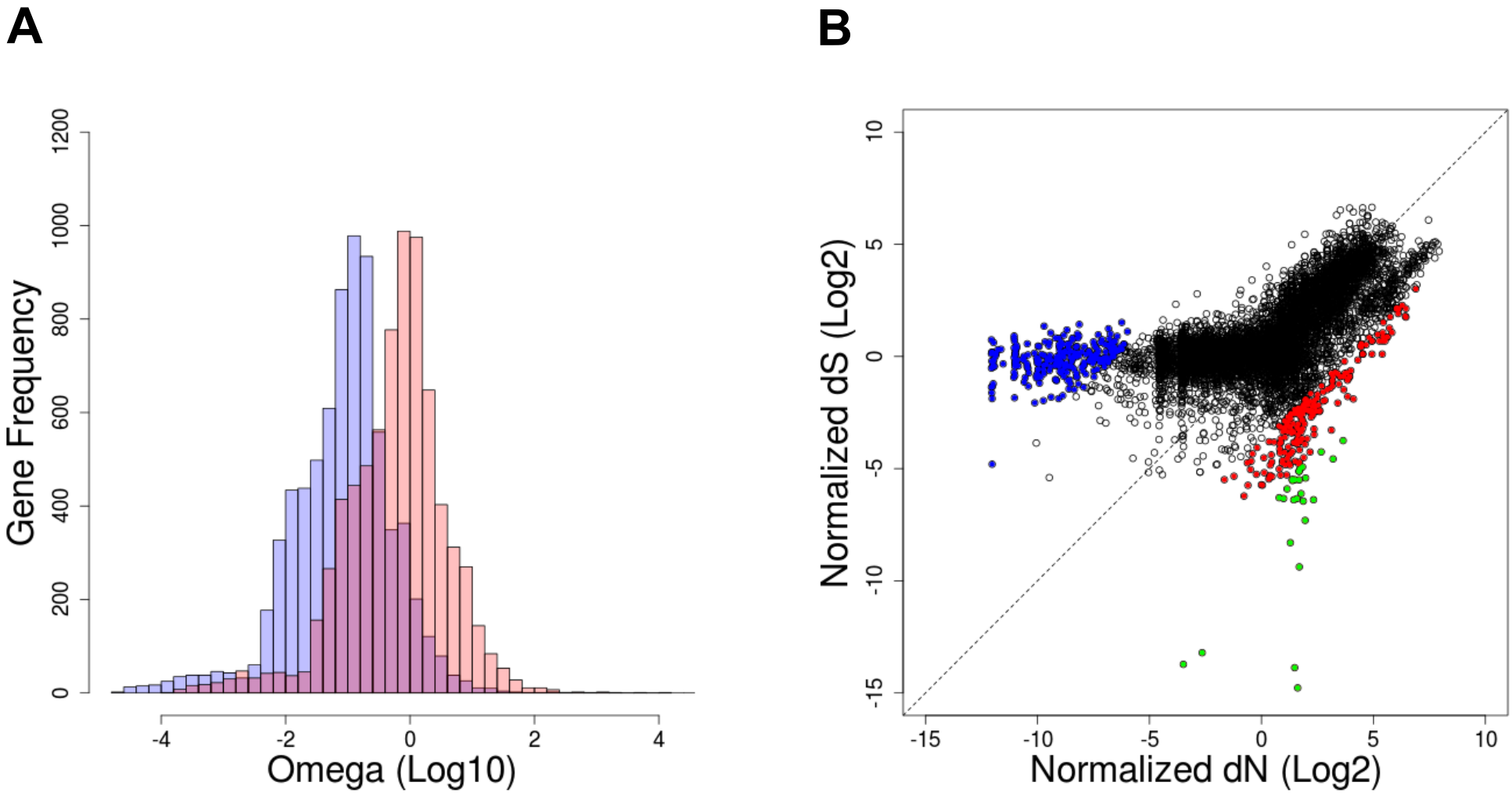
A
Gene Frequency
0
200
400
600
800
1000
1200
-4
-2
0
2
4
Omega (Log10)
B
Normalized dS (Log2)
10
5
0
-5
-10
-15
-15
-10
-5
0
5
10
Normalized dN (Log2)

Figure 4