Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen

C. K. Stover*, X. Q. Pham†, A. L. Erwin*, S. D. Mizoguchi*, P. Warrener*, M. J. Hickey*, F.S. L. Brinkman‡, W. O. Hufnagle*, D. J. Kowalik*, M. Lagrou*, R. L. Garber*, L. Goltry*, E. Tolentino*, S. Westbrock-Wadman*, Y. Yuan*, L. L. Brody*, S. N. Coulter*, K. R. Folger*, A. Kas†, K. Larbig§, R. Lim†, K. Smith†, D. Spencer†, G. K.-S. Wong†, Z. Wu†, I. T. Paulsen‖¶, J. Reizer¶, M. H. Saier¶, R. E. W. Hancock‡, S. Lory# & M. V. Olson†

* *PathoGenesis Corporation, 201 Elliott Avenue West, Seattle, Washington 98119, USA*
† *Department of Medicine and Genetics, University of Washington Genome Center, Box 352145, University of Washington, Seattle, Washington 98195, USA*
‡ *Department of Microbiology and Immunology, University of British Columbia, 300-6174 University Blvd, Vancouver, British Columbia V6T 1Z3, Canada*
§ *Klinische Forschergruppe, Medizinische Hochschule Hannover, D-30623 Hannover, Germany*
‖ *The Institute for Genomic Research, 9712 Medical Center, Rockville, Maryland 20850, USA*
¶ *Department of Biology, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093-0116, USA*
# *Department of Microbiology, University of Washington School of Medicine, Seattle, Washington 98195, USA*

Pseudomonas aeruginosa **is a ubiquitous environmental bacterium that is one of the top three causes of opportunistic human infections. A major factor in its prominence as a pathogen is its intrinsic resistance to antibiotics and disinfectants. Here we report the complete sequence of *P. aeruginosa* strain PAO1. At 6.3 million base pairs, this is the largest bacterial genome sequenced, and the sequence provides insights into the basis of the versatility and intrinsic drug resistance of *P. aeruginosa*. Consistent with its larger genome size and environmental adaptability, *P. aeruginosa* contains the highest proportion of regulatory genes observed for a bacterial genome and a large number of genes involved in the catabolism, transport and efflux of organic compounds as well as four potential chemotaxis systems. We propose that the size and complexity of the *P. aeruginosa* genome reflect an evolutionary adaptation permitting it to thrive in diverse environments and resist the effects of a variety of antimicrobial substances.**

P. aeruginosa is a versatile Gram-negative bacterium that grows in soil, marshes and coastal marine habitats, as well as on plant and animal tissues[1]. It forms biofilms on wet surfaces such as those of rocks and soil[2,3]. The emergence of *P. aeruginosa* as a major opportunistic human pathogen during the past century may be a consequence of its resistance to the antibiotics and disinfectants that eliminate other environmental bacteria. *P. aeruginosa* is now a significant source of bacteraemia in burn victims, urinary-tract infections in catheterized patients, and hospital-acquired pneumonia in patients on respirators[4]. It is also the predominant cause of morbidity and mortality in cystic fibrosis patients, whose abnormal airway epithelia allow long-term colonization of the lungs by *P. aeruginosa*. These infections are impossible to eradicate, in part because of the natural resistance of the bacterium to antibiotics, and ultimately lead to pulmonary failure and death.

Here we report the sequencing of the genome of *P. aeruginosa*. The sequence is of interest because of the insights it provides into the role of this bacterium as a pathogen, and because it offers new information on the relationship between genome size, genetic complexity and ecological versatility in bacteria. At 6.3 million base pairs (Mbp), the *P. aeruginosa* genome is markedly larger than most of the 25 sequenced bacterial genomes. In fact, with 5,570 predicted open reading frames (ORFs), the genetic complexity of *P. aeruginosa* approaches that of the simple eukaryote *Saccharomyces cerevisiae*, whose genome encodes about 6,200 proteins[5]. In contrast, *P. aeruginosa* has only 30–40% of the number of predicted genes present in the simple metazoans *Caenorhabditis elegans* and *Drosophila melanogaster*[6].

Sequencing and assembly

Sequencing of the complete 6.3-Mbp genome of *P. aeruginosa* was accomplished by a straightforward implementation of whole-genome-shotgun sampling. The largest genome that has been completely sequenced by this approach is from *Deinococcus radiodurans* (2.6 Mbp)[7]. The other large bacterial genome sequences (*Bacillus subtilis*, 4.2 Mbp; *Synechocystis*, 3.6 Mbp; *Escherichia coli*, 4.6 Mbp; and *Mycobacterium tuberculosis*, 4.4 Mbp)[8–11] were all initially determined by sequencing overlapping sets of clones, polymerase chain reaction (PCR) products and gel-purified restriction fragments.

With one major exception, the assembled genome sequence is in excellent agreement with the physical map of the *P. aeruginosa* genome[12,13]. The exception is the inversion of more than one-quarter of the genome in the PAO1 isolate we sequenced, relative to DSM-1707, the PAO1-derived isolate previously mapped in the laboratory of B. Tümmler[12,13] (Fig. 1). As both of these isolates are clonally derived from the original PAO1 strain of *P. aeruginosa*, any differences in genome structure between PAO1 and DSM-1707 must have arisen during propagation. This inversion does not appear to be unique to the sequenced isolate as PAO1 stocks from other laboratories have the same inversion (H. Schweizer, personal communication). The inversion appears to have resulted from homologous recombination between the *rrnA* and *rrnB* loci, which are orientated in opposite directions and separated by 1.7 Mbp (Fig. 1). Comparative analysis of digests of PAO1 and DSM-1707 with *Sfo*I, *Swa*I and *Pac*I supported this possibility. Earlier observations of inversions of genomic segments between oppositely orientated ribosomal DNA loci in *E. coli* and *S. typhimurium* led to the proposal that these reversible genome rearrangements may have adaptive significance[14].

Properties of the genome and relationship to other bacteria

Basic features. These are summarized in Table 1 and Fig. 1. Most of the predicted ORFs have the high G+C content (66.6%)

characteristic of the genome as a whole and have codon usage similar to previously described *P. aeruginosa* genes. However, ten regions of 3.0 kilobases (kb) or greater exhibit significantly lower G+C content and unusual codon usage (Fig. 1), possibly indicative of recent horizontal transfer. In addition, there are two regions (PA616–PA648, PA715–PA728) containing probable bacteriophages.

Comparative analysis. To gain insight into the significance of the size of this genome, we concentrated on comparative analysis of the *P. aeruginosa* and *E. coli* genomes. Not only is *E. coli* the most intensively studied of all bacteria, but it is also the closest relative of *P. aeruginosa* among the bacteria with fully sequenced genomes: for example, when each of the 5,570 predicted ORFs for *P. aeruginosa* was compared with the pooled ORFs for 22 other bacterial genomes by the sequence-alignment algorithm BLASTP[44], nearly half of the best hits above a stringent comparison threshold (an expect value of 10^{-5}) were to *E. coli*, and no other organism accounts for even 10% (see Supplementary Information). The relative prominence of *E. coli* increases moderately at more stringent thresholds although a noisy pattern of weak 'best hits' to phylogenetically distant bacteria emerges at lower thresholds. Although this test confirms that *E. coli* is a sensible comparison partner for *P. aeruginosa*, the median amino-acid identity within the aligned region of the *P. aeruginosa*–*E. coli* orthologues is only 40%.

Comparison of the *P. aeruginosa* and *E. coli* genomes indicates that the large genome of *P. aeruginosa* is the result of greater genetic complexity rather than differences in genome organization. Distributions of ORF sizes and inter-ORF spacings are both nearly identical in the two genomes (see http://www.pseudomonas.com), and the extent of evolutionarily recent duplications appears comparable. The longest repeats in the *P. aeruginosa* genome are the four rDNA loci and one duplicated gene cluster that spans a few thousand base pairs (PA1899–PA1905; PA4210–PA4216). At the level of amino-acid sequence conservation, residual stretches of locally conserved gene order between *P. aeruginosa* and *E. coli* are far more evident than are internal duplications in either genome. At the same BLASTP threshold employed for the comparisons between the *P. aeruginosa* ORFs and those of other bacterial genomes, we searched for clusters of five or more ORFs that are conserved between *P. aeruginosa* and *E. coli*, allowing single-ORF insertions or deletions within the clusters. Thirty-three distinct clusters were identified, which included 256 ORFs; seven of these clusters involve ten or more ORFs. This analysis showed only a few gene clusters duplicated within either the *E. coli* or the *P. aeruginosa* genome. Hence, with respect to local gene order, evidence of the common ancestry of segments of the *P. aeruginosa* and *E. coli* genome is far more abundant than are the vestiges of more recent duplication events of comparable size within either genome.

Evidence for increased functional diversity. The apparent lack of recent gene duplication indicates that the size of the *P. aeruginosa* genome is due to greater gene and functional diversity. When we analysed the *P. aeruginosa*, *E. coli*, *B. subtilis* and *M. tuberculosis* genomes by BLASTP comparisons between all predicted ORFs within each organism, we found that the *P. aeruginosa* genome has significantly more distinct gene families (paralogous groups) than the other large bacterial genomes (see Supplementary Information). There are nearly 50% more paralogous groups in *P. aeruginosa* than predicted on the basis of a simple comparison of genome sizes.

If the larger genome of *P. aeruginosa* arose by recent gene duplication, we would have expected it to have a similar number of paralogous groups to the other large bacterial genomes, with a larger number of ORFs in each group. In fact, the number of ORFs in the paralogous groups in PAO1 is similar to the other genomes. These data indicate that selection for environmental versatility has favoured expansion of genetic capability through the development of numerous small paralogous gene families whose members encode distinct functions.

Annotation of *P. aeruginosa* genes

Prediction of gene function. The predicted ORFs were examined individually for (1) identity with known genes of *P. aeruginosa* with sequences deposited in GenBank, (2) similarity with well-characterized genes from other bacteria, or (3) presence of known functional motifs (Table 1; see http://www.pseudomonas. com for complete list). In each case the literature was searched to ensure that the proteins encoded by the homologous genes were functionally characterized to avoid the perpetuation of poorly supported functional assignments. In addition, 61 researchers who were members of the *P. aeruginosa* research community or had experience in particular aspects of bacterial physiology were enlisted for the *Pseudomonas* Community Annotation Project (PseudoCAP) to provide expert assistance and confirmatory information for the identification of ORFs and assigned functions.

We were able to assign a functional class to 54.2% of ORFs (Table 2). As in other bacterial genomes, a large proportion of the genome (45.8% of ORFs) consists of genes for which no function could be determined or proposed (confidence level 4; see Table 1). Of these, nearly a third (769 ORFs) possess homology to genes of unknown function predicted in other bacterial genomes, and the remainder (32% of ORFs) does not have strong homology with any reported sequence.

The 372 ORFs that are known *P. aeruginosa* genes with demonstrated functions (confidence level 1) are primarily genes encoding lipopolysaccharide biosynthetic enzymes, virulence factors, such as exoenzymes and the systems that secrete them,

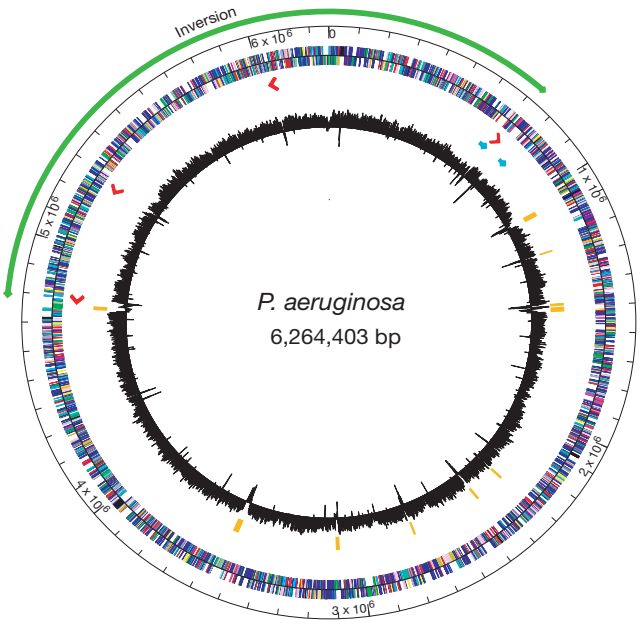

Figure 1 Circular representation of the *P. aeruginosa* genome. The outermost circle indicates the chromosomal location in base pairs (each tick is 100 kb). The distribution of genes is depicted by coloured boxes according to functional category and direction of transcription (outer band is the plus strand; inner band is the minus strand). Red arrows, the locations and direction of transcription of ribosomal RNA genes; green arrow, the inverted region that resulted from a homologous recombination event between *rrnA* and *rrnB*; blue arrows, location of two regions containing probable bacteriophages. The black plot in the centre is percentage G+C content plotted as the average for non-overlapping 1-kb windows spanning one strand for the entire *P. aeruginosa* genome. Yellow bars, regions of ≥ 3.0 kb with G+C content of two standard deviations ($< 58.8\%$) below the mean (66.6%) (see Supplementary Information). A linear map of the genes, with the colour code for functional categories, is available at http://www.pseudomonas.com.

and proteins involved in motility and adhesion. ORFs with strong homology to genes in other organisms with demonstrated functions (confidence level 2; 1,059 ORFs) include those required for DNA replication, protein synthesis, cell-wall biosynthesis and intermediary metabolism. *P. aeruginosa* is able to grow on minimal medium, and as we expected, we identified most of the genes required for biosynthesis of amino acids, nucleic acids and cofactors.

The ORFs that provided the most new information about *P. aeruginosa* biology are those that could be assigned a probable function on the basis of similarity to established sequence motifs, but could not be assigned a definite name (confidence level 3; 1,590 ORFs). Most of these genes encode products that are in one of three functional classes: putative enzymes (405 genes), transcriptional regulators (341 genes) or transporters of small molecules (408 genes). In some cases genomic context provided additional information, allowing us to identify loci that appear to encode systems such as metabolic pathways and secretion systems, although the substrates for such systems could not be identified. These and other features of the *P. aeruginosa* genome that may shed light on its biology are discussed below. Additional details are available in the Supplementary Information and at http://www.pseudomonas.com.

Regulation. *P. aeruginosa* has the highest proportion of predicted regulatory genes observed in the sequenced bacterial genomes. Analysis using relevant Pfam 5.2 family models[15] and HMMER 2.1.1 (http://hmmer.wustl.edu/) shows 468 genes containing motifs characteristic of transcriptional regulators or environmental sensors (see Supplementary Information). This analysis predicts that 8.4% of *P. aeruginosa* genes are involved in regulation, a far higher proportion than is found in other sequenced genomes. (Manual annotation of the genome identified 521 genes (9.4%) as encoding either transcriptional regulators or two-component regulatory system proteins (Table 2). Thus the parameters we employed gave somewhat conservative predictions.)

Similar computational analysis of regulatory motifs in 22 genomes indicates that as bacterial genome size increases, the proportion of the genome devoted to regulatory proteins increases as well (Fig. 2). This trend appears most prominent in prototrophic bacteria that can survive in diverse environments. For example, motifs characteristic of regulatory proteins are found in 5.8% of *E. coli* genes and 5.3% of *B. subtilis* genes, but only in 3.0% of genes in *M. tuberculosis*, a highly specialized pathogen with a comparable genome size. *Helicobacter pylori*, another highly specialized bacterial pathogen with a much smaller genome, possesses even less regulatory potential (1.1% of genes). When we compared *P. aeruginosa* transcriptional regulators with other bacterial systems, the most striking over-representations in were in the LysR, AraC, ECF-σ and two-component regulator families. There is an extraordinary number of putative two-component regulatory system proteins, with 55 sensors, 89 response regulators and 14 sensor–response regulator hybrids, far more than found in the other genomes analysed. Such systems permit organisms to respond to changes in their environment, and are often associated with global regulatory systems as well as with regulation of virulence.

Outer membrane proteins. Outer membrane proteins (OMPs) are of particular interest in *P. aeruginosa* due to their cell-surface exposure and their involvement in transport of antibiotics, in export of extracellular virulence factors, and in anchoring the structures that mediate adhesion and motility. About 150 genes are predicted to encode OMPs, a disproportionately large number compared with other genomes. Three large paralogous families were identified: the OprD family of specific porins (19 genes), the TonB-family of gated porins, which includes proteins involved in iron-siderophore uptake (34 genes), and the OprM family of outer membrane proteins involved in efflux or secretion (18 genes). These large families of proteins were unexpected, as single members of these families (for example, OprD) had been well studied with no appreciation that these proteins were members of a large paralogous group. To date, the only other genome that is known to contain a large paralogous family of OMPs is *H. pylori*[16]. The identification of these families could have a significant impact on the focus of antimicrobial and vaccine research.

Import of nutrients. Consistent with its environmental versatility, *P. aeruginosa* has nearly 300 cytoplasmic membrane transport systems, about two-thirds of which appear to be involved in the import of nutrients and other molecules (http://www-biology.ucsd.edu/~ipaulsen/transport). The overall substrate specificities of the *P. aeruginosa* transporters are similar to those of *E. coli* and *B. subtilis* with certain significant exceptions (see Supplementary Information). *P. aeruginosa* has a large variety of transporters for mono-, di-, and tri-carboxylates, but it appears to be conspicuously deficient in sugar transporters. For example, it possesses four

Table 1 Genome features

General features			
Genome size (bp)	6,264,403		
G+C content	66.6%		
Coding regions	89.4%		
Stable RNA	0.4%		
RNA			
rRNA	16S	23S	5S
rrnA	722,096–723,631	724,103–726,993	727,136–727,255
rrnB	4,793,731–4,792,196	4,791,724–4,788,836	4,788,693–4,788,574
rrnC	5,269,259–5,267,724	5,267,252–5,264,362	5,264,219–5,264,100
rrnD	6,044,743–6,043,208	6,042,736–6,039,846	6,039,703–6,039,584
Transfer RNA*	63 species		
Non-classical RNA	4 species		
Coding sequences			
Confidence level†	ORFs (%)	Definition	
1	372 (6.7)	*P. aeruginosa* genes with demonstrated function	
2	1059 (19.0)	Strong homologues of genes with demonstrated function from other organisms	
3	1590 (28.5)	Genes with proposed function based on motif searches or limited homology	
4	769 (13.8)	Homologues of reported genes of unknown function	
4	1780 (32.0)	No homology to any reported sequences	
Total	5570 (100)		

*Transfer RNAs were identified with tRNAscan-SE[40].

†Each annotation includes a numerical confidence level indicating the basis of determination of the protein name. A complete list of predicted genes and their annotations is available at http://www.pseudomonas.com.

dicarboxylate permeases of the TRAP-T type (*E. coli* has only one), and has only two phosphotransferase system (PTS) sugar transporters—for fructose and *N*-acetylglucosamine (*E. coli* has more than twenty)[17]. Also, *P. aeruginosa* has no predicted sugar transporters of the major facilitator superfamily (MFS), although *E. coli* has more than twenty. The apparent lack of sugar transporters in *P. aeruginosa* correlates with the absence of an intact glycolytic pathway and with its aerobic, oxidative metabolism[18].

β-Oxidative metabolism. In contrast to its limited ability to grow on sugars, *P. aeruginosa* can use a wide variety of other carbon compounds, and its genome provides insight into the molecular basis of this metabolic versatility. In addition to known oxidative enzymes and pathways, we found a substantial number of other genes encoding putative enzymes characteristic of β-oxidation, such as acyl-CoA dehydrogenase (25 genes) and enoyl-CoA hydratase/isomerase (16 genes). In contrast, *E. coli* contains four genes for acyl-CoA dehydrogenase and seven for enoyl-CoA hydratase/isomerase. With the exception of *M. tuberculosis*, no other sequenced genome contains such large numbers of these enzymes. The β-oxidative genes are often clustered with other genes encoding proteins that may have related functions, such as probable acyl-CoA thiolases, short-chain dehydrogenases, flavin-containing monooxygenases, or other oxidoreductases. In several cases, these gene clusters also contained genes for MFS transport proteins and outer membrane porins of the OprD family (see Supplementary Information).

Intrinsic drug resistance and efflux systems. *P. aeruginosa* is noted for its intrinsic resistance to many front-line antibiotics, due mainly to its low outer membrane permeability and to active efflux of antibiotics[19]. Four *P. aeruginosa* multidrug efflux systems have been reported, all of which are members of the resistance-nodulation-cell division (RND) family[20,21]. We used BLASTP analysis to identify potential export systems in the PAO1 genome, and probable multidrug efflux systems were identified by a phylogenetic analysis of each family[17]. The *P. aeruginosa* genome appears to contain a large number of undescribed drug efflux systems, predominantly of the RND and MFS families (Fig. 3). The number of predicted drug efflux systems from the MFS, small multi-drug resistance (SMR), ATP-binding cassette (ABC) and multidrug and toxic compound extrusion (MATE) families is similar to other organisms such as *E. coli*, *B. subtilis* and *M. tuberculosis*. However, *P. aeruginosa* contains many more predicted AcrB/Mex-type RND multidrug efflux systems (10 genes) than *E. coli* (4), *B. subtilis* (1) and *M. tuberculosis* (0). Each of the *P. aeruginosa* genes encoding a putative RND transport protein is adjacent to a gene for a probable membrane fusion protein; most RND loci also contain genes for outer membrane proteins of the OprM family (see Supplementary Information).

Protein secretion. *P. aeruginosa* secretes several virulence factors, including toxins, lipases and proteases. Four pathways of protein secretion have been described for Gram-negative bacteria[22], and three of these were evident in *P. aeruginosa*. The prototypic type I system in *P. aeruginosa*, which directs secretion of alkaline protease (encoded by *aprA*), consists of the ABC transport protein AprD, the membrane fusion protein AprE and the OprM-family outer membrane protein AprF. The PAO1 genome appears to contain four additional type I systems. One of these clusters (PA3404–PA3408) is homologous to the haem acquisition system (Has) of *Serratia marcescens*. A sixth *aprF* homologue (PA4974) was not clustered with genes for other putative transport proteins. This gene was most similar in sequence to *tolC*, which encodes an *E. coli* outer membrane protein involved in haemolysin secretion.

A type II secretion system (general secretion pathway) is encoded by the *xcp* gene cluster (PA3095–PA3105) and the unlinked *pilD*/*xcpA* gene (PA4528)[23]. An unexpected finding was that many of the *xcp* genes have homologues elsewhere on the chromosome, including four additional homologues of *xcpQ* and *xcpR*. Type III secretion systems are found in many plant and human pathogens and are responsible for contact-dependent delivery of proteins into the cytoplasm of host cells. *P. aeruginosa* contains a single type III secretion system (PA1690–PA1725) which secretes several proteins including exoenzymes S, T and Y[24].

Other potential surface molecules. The genome of *P. aeruginosa* PAO1 contains two extremely long open reading frames, PA2462 (5,628 amino acids) and PA41 (3,536 amino acids). Each of these ORFs has sequence similarity to proteins that are adhesins in other bacterial pathogens, the filamentous haemagglutinin (FhaB) of *Bordetella pertussis*, and the HMW1A / HMW2A adhesins

Table 2 Functional classes of predicted genes

Function class	ORFS	% of ORFS
Adaptation, protection (for example cold shock proteins)	60	1.1
Amino acid biosynthesis and metabolism	150	2.7
Antibiotic resistance and susceptibility	19	0.3
Biosynthesis of cofactors, prosthetic groups and carriers	119	2.1
Carbon compound catabolism	130	2.3
Cell division	26	0.5
Cell wall, LPS	83	1.5
Central intermediary metabolism	64	1.1
Chaperones & heat shock proteins	52	0.9
Chemotaxis	43	0.8
DNA replication, recombination, modification and repair	81	1.5
Energy metabolism	166	3.0
Fatty acid and phospholipid metabolism	56	1.0
Membrane proteins	7	0.1
Motility & attachment	65	1.2
Nucleotide biosynthesis and metabolism	60	1.1
Protein secretion/export apparatus	83	1.5
Putative enzymes	409	7.3
Related to phage, transposon or plasmid	38	0.7
Secreted factors (toxins, enzymes, alginate)	58	1.0
Transcription, RNA processing and degradation	45	0.8
Transcriptional regulators	403	7.2
Translation, post-translational modification, degradation	149	2.7
Transport of small molecules	555	10.0
Two-component regulatory systems	118	2.1
Hypothetical	1,774	31.8
Unknown (conserved hypothetical)	757	13.6
Total	5,570	100.0

On the basis of known function or homology to genes of known function, PAO1 ORFs were assigned to one of 25 functional categories derived from those used for functional classification of *E. coli* ORFs[11]. Where no functional inferences were possible, genes with homology to genes of unknown function predicted in other genomes were designated 'unknown', and genes without strong homology to any previously reported sequence were designated 'hypothetical'.

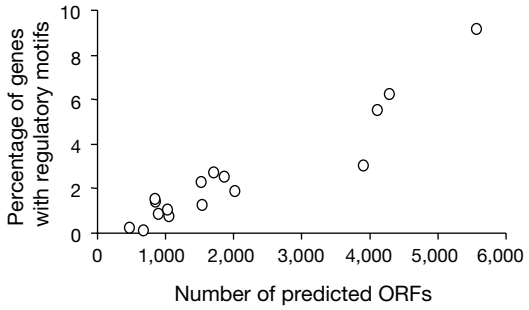

Figure 2 The percentage of genes with regulatory motifs increases with the size of the genome. Each sensor or regulatory family model was extracted from the Pfam 5.2 database[15] and analysed against a database containing the combined predicted ORFs of each of the 22 genome sequences listed below. For each genome, the total number of ORFs identified with a probability of less than 10^{-4} as containing any of the regulatory motifs was divided by the number of predicted ORFs in that genome to calculate the percentage regulatory genes. The genomes analysed were *M. genitalium* (480 predicted ORFs), *M. pneumoniae* (677), *R. prowazekii* (834), *B. burgdorferi* (850), *C. trachomatis* (894), *T. pallidum* (1031), *C. pneumoniae* (1052), *A. aeolicus* (1522), *H. pylori* 26695 (1553), *H. influenzae* (1709), *M. jannaschii* (1715), *P. abyssi* (1765), *T. maritima* (1846), *M. thermoautotrophicum* (1869), *N. meningitidis* MC58 (2025), *P. horikoshii* (2064), *A. fulgidus* (2407), *Synechocystis* PCC6803 (3169), *M. tuberculosis* (3918), *B. subtilis* (4100), *E. coli* (4289), *P. aeruginosa* (5570).

of *Haemophilus influenzae*[25,26]. PA2462 is adjacent to an ORF (PA2461) with strikingly abnormal codon usage and a G+C content of only 38.5%. Further, in addition to the genes known to be involved in synthesis of lipopolysaccharides, we noted three other genetic loci that may be involved in the synthesis of extracellular polysaccharides (see Supplementary Information). For example, a seven-gene cluster (PA1385–PA1391) adjacent to the *galE* gene includes genes for four probable glycosyl transferases and an ABC transport protein similar to putative carbohydrate-exporting proteins encoded in O-antigen biosynthetic gene clusters in other organisms. This seven-gene cluster is in a region with lower G+C content than the surrounding genes.

Chemosensing and chemotaxis. *P. aeruginosa* appears to have the most complex chemosensory systems of all the complete bacterial genomes, with four loci that encode probable chemotaxis signal-transduction pathways (see Supplementary Information). Of these, one (PA1456–PA1464) is similar in gene organization to the *Salmonella typhimurium* locus required for flagella-mediated swimming toward chemoattractants[27]. A second (PA173–PA180) more closely resembles the gene organization seen in *Rhodobacter sphaeroides*[28]. Each of the two remaining clusters has homology to the *che* genes from *E. coli* and to the *frz* genes from the non-flagellated gliding bacterium, *Myxococcus xanthus*[29]. PA408–PA417 is required for twitching motility[30] and PA3702–PA3708 is as yet uncharacterized. *P. aeruginosa* undergoes chemotaxis toward a variety of sugars, amino acids, and inorganic phosphate, and away from thiocyanic and isothiocyanic esters[31–34]. We identified 26 ORFs encoding probable chemotaxis sensory transducer proteins to mediate these responses.

Discussion

We propose that the large genome size and genetic complexity of *P. aeruginosa* reflect evolutionary adaptations permitting it to thrive in diverse ecological niches. Analysis of the complete genome sequence of *P. aeruginosa* reveals many clues regarding the basis of this versatility. *P. aeruginosa* has broad capabilities to transport, metabolize and grow on organic substances, numerous iron-siderophore uptake systems, and the enhanced ability to export compounds (for example, enzymes and antibiotics) by a large number of protein secretion and RND efflux systems. *P. aeruginosa* potentially possesses four chemotaxis systems, at least one of which contributes to its ability to form biofilms[35]. Thus this organism can readily move to more favourable conditions or consolidate and 'dig in' for persistent colonization of a particular microenvironment. Consistent with its increased genetic complexity, *P. aeruginosa* has the greatest percentage of genes devoted to command and control systems (for example, environmental sensors and transcriptional regulators) observed in a bacterial genome. These regulatory genes presumably modulate the diverse genetic and biochemical capabilities of this bacterium in changing environmental conditions.

P. aeruginosa infections are particularly difficult to treat because of intrinsic resistance to antibiotics. It would appear that, in the course of evolving the functional diversity required to compete with other microorganisms in a variety of environments, it developed mechanisms for resisting naturally occurring antimicrobial compounds. The efflux systems we identified could contribute to this intrinsic resistance. The effects of antimicrobials could also be mitigated by modulating expression of drug targets, enzymatic modifiers, transport systems and compensatory pathways. Indeed, the unusually large regulatory capability in *P. aeruginosa* may provide greater latitude for adaptive drug resistance through gene regulation than exists in other bacteria with smaller genomes. Furthermore, given its capacity to metabolize a wide variety of organic substrates, it is also possible that *P. aeruginosa* possesses greater potential for enzymatic modification and degradative drug resistance mechanisms than was thought. Therefore, the metabolic diversity, transport capabilities and regulatory adaptability that enable *P. aeruginosa* to thrive and compete with other microorganisms probably all contribute to its high intrinsic resistance to antibiotics. Knowledge of the complete genome sequence and encoded processes provides a wealth of information for the discovery and exploitation of new antibiotic targets, and hope for the development of more effective strategies to treat the life-threatening opportunistic infections caused by *P. aeruginosa* in humans. □

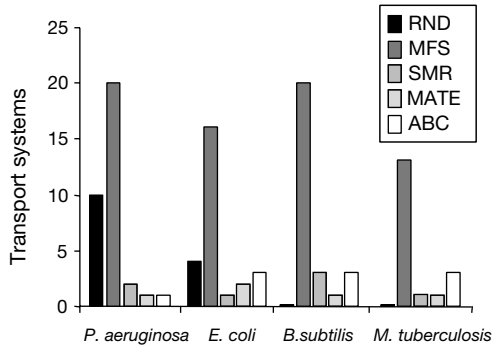

Figure 3 Comparison of the number of predicted drug efflux systems in *P.aeruginosa*, *E. coli*, *B. subtilis* and *M. tuberculosis*. For the last three organisms, these numbers are based on predictions taken from http://www-biology.ucsd.edu/~ipaulsen/transport/. Five types of multidrug efflux systems are analysed: resistance/nodulation/cell division family (RND; for example, *E. coli* AcrB), major facilitator superfamily (MFS; for example, *B. subtilis* Bmr), small multidrug resistance family (SMR; for example, *E. coli* EmrE), multidrug and toxic compound extrusion family (MATE; for example, *Vibrio parahaemolyticus* NorM)[41] and ATP-binding cassette family (ABC; for example, *Lactococcus lactis* LmrA)[17]. Only family members that clearly clustered with known multidrug efflux systems were counted as probable multidrug efflux systems. For example, the number of RND multidrug efflux systems does not include members of this family that belong to the SecD/SecF protein excretion, the Czc metal efflux or the *M. tuberculosis* MmpL glycolipid efflux[42] protein clusters, but only includes proteins belonging to the AcrB/Mex multidrug efflux protein cluster[43].

Methods

Sequencing and assembly

Strain PAO1, a wound isolate[36], was chosen as a strain prototype for sequencing because it is the most widely used *P. aeruginosa* laboratory strain and because physical and genetic maps were available[12,13]. Strain PAO1 was obtained from B. Holloway's collection maintained in the laboratory of P. Phibbs, Univ. of Georgia. Our approach to the *P. aeruginosa* sequencing relied on standard data-collection and sequence-assembly methods. Most of the data comprised individual sequencing traces acquired from randomly picked M13 templates. The data were processed with the phred/phrap/consed package of base-calling, sequence-assembly, and finishing/editing software (http://bozeman.mbt.washington.edu)[37,38]. At frequent intervals throughout the project, *de novo* phrap assemblies were carried out using all available data. The main requirement for achieving practical computation times was adequate real memory: on the full data set, the assembly required 4 h on a workstation with 4 gigabytes of memory. We employed dye-primer and dye-terminator chemistry in a 51:49 ratio to acquire 94,847 usable shotgun-sequencing traces. The shotgun traces provided 6.9-fold coverage of the genome in 'high-quality' base calls (that is, those with phred scores > 20, which corresponds to an error rate $< 1\%$). The final raw-data set included an additional 1,604 'finishing' traces, most of which were obtained by priming with custom primers on M13 templates selected from the random-template collection. The purpose of most 'finishing' reads was to improve data quality in regions where the phred/phrap/consed software recognized that the consensus sequence had inadequate support. We also acquired 1,672 cosmid-end sequences from a collection of 836 cosmids that contained 40-kb inserts. The inserts in these cosmids covered 97% of the genome; hence, their end sequences provided a strong check on the validity of the final assembly.

Only 13 sites in the genome required specialized finishing procedures: four of these sites are the rDNA loci, which contain nearly exact copies of a 5.9-kb repeat, six are nearly identical copies of a 1.4-kb insertion sequence, and the other three also involve repeated sequences. The sequences at all 13 of these sites were obtained by sequencing PCR products that spanned the individual repeats. In addition to validating the assembly with cosmid-end sequences, we also monitored the base-pair accuracy of the final sequence in a variety of ways. One test involved full sequencing, by conventional methods, of two cosmids that contained widely spaced segments of the *P. aeruginosa* genome; these cosmids, which were selected at the beginning of the project, were brought to the highest

achievable quality standard by expert 'finishers'. In the final whole-genome assembly, which was entirely independent of the cosmid sequencing, we found no discrepancies with the 81,843 base pairs present in the two cosmids.

Gene predictions

Open reading frames were initially predicted by GeneMark.HMM (http://genemark.biology.gatech.edu/GeneMark/whmm.cgi)[39]. Additional ORFs with homology to known genes were identified by BLASTX analysis. Predicted ORFs were reviewed individually by gene annotators for start-codon assignment based on additional contextual information such as the proximity of ribosomal binding sequence motifs and predicted signal peptides. □

Received 16 May; accepted 10 July 2000.

1. Hardalo, C. & Edberg, S. C. *Pseudomonas aeruginosa*: assessment of risk from drinking water. *Crit. Rev. Microbiol.* **23,** 47–75 (1997).
2. Costerton, J. W., Stewart, P. S. & Greenberg, E. P. Bacterial biofilms: a common cause of persistent infections. *Science* **284,** 1318–1322 (1999).
3. Ahearn, D. G., Borazjani, R. N., Simmons, R. B. & Gabriel, M. M. Primary adhesion of *Pseudomonas aeruginosa* to inanimate surfaces including biomaterials. *Methods Enzymol.* **310,** 551–557 (1999).
4. Bodey, G. P., Bolivar, R., Fainstein, V. & Jadeja, L. Infections caused by *Pseudomonas aeruginosa*. *Rev. Infect. Dis.* **5,** 279–313 (1983).
5. Ball, C. A. *et al.* Integrating functional genomic information into the *Saccharomyces* genome database. *Nucleic Acids Res.* **28,** 77–80 (2000).
6. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25,** 232–234 (2000).
7. White, O. *et al.* Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286,** 1571–1577 (1999).
8. Kunst, F. *et al.* The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390,** 249–256 (1997).
9. Kaneko, T. *et al.* Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3,** 109–136 (1996).
10. Cole, S. T. *et al.* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome. *Nature* **393,** 537–544 (1998).
11. Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1474 (1997).
12. Schmidt, K. D., Schmidt-Rose, T., Römling, U. & Tümmler, B. Differential genome analysis of bacteria by genomic subtractive hybridization and pulsed field gel electrophoresis. *Electrophoresis* **19,** 509–514 (1998).
13. Schmidt, K. D., Tümmler, B. & Römling, U. Comparative genome mapping of *Pseudomonas aeruginosa* PAO with *P. aeruginosa* C, which belongs to a major clone in cystic fibrosis patients and aquatic habitats. *J. Bacteriol.* **178,** 85–93 (1996).
14. Mahan, M. J. & Roth, J. R. Ability of a bacterial chromosome segment to invert is dictated by included material rather than flanking sequence. *Genetics* **129,** 1021–1032 (1991).
15. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **28,** 262–266 (2000).
16. Hancock, R. E., Alm, R., Bina, J. & Trust, T. *Helicobacter pylori*: a surprisingly conserved bacterium. *Nature Biotechnol.* **16,** 216–217 (1998).
17. Paulsen, I. T., Sliwinski, M. K. & Saier, M. H. Jr Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol.* **277,** 573–592 (1998).
18. Temple, L. M., Sage, A. E., Schweizer, H. P., & Phibbs, P. V. Jr in *Pseudomonas* (ed. Montie, T. C.) 35–72 (Plenum, New York, 1998).
19. Hancock, R. E. Resistance mechanisms in *Pseudomonas aeruginosa* and other nonfermentative gram-negative bacteria. *Clin. Infect. Dis.* **27,** S93–S99 (1998).
20. Westbrock-Wadman, S. *et al.* Characterization of a *Pseudomonas aeruginosa* efflux pump contributing to aminoglycoside impermeability. *Antimicrob. Agents Chemother.* **43,** 2975–2983 (1999).
21. Nikaido, H. Antibiotic resistance caused by gram-negative multidrug efflux pumps. *Clin. Infect. Dis.* **27,** S32–S41 (1998).
22. China, B. & Goffaux, F. Secretion of virulence factors by *Escherichia coli*. *Vet. Res.* **30,** 181–202 (1999).
23. Bleves, S., Gerard-Vincent, M., Lazdunski, A. & Filloux, A. Structure-function analysis of XcpP, a component involved in general secretory pathway-dependent protein secretion in *Pseudomonas aeruginosa*. *J. Bacteriol.* **181,** 4012–4019 (1999).
24. Yahr, T. L., Vallis, A. J., Hancock, M. K., Barbieri, J. T. & Frank, D. W. ExoY, an adenylate cyclase secreted by the *Pseudomonas aeruginosa* type III system. *Proc. Natl Acad. Sci. USA* **95,** 13899–13904 (1998).
25. Domenighini, M. *et al.* Genetic characterization of *Bordetella pertussis* filamentous haemagglutinin: a protein processed from an unusually large precursor. *Mol. Microbiol.* **4,** 787–800 (1990).
26. Noel, G. J., Barenkamp, S. J., St. Geme, J. W. III, Haining, W. N. & Mosser, D. M. High-molecular-weight surface-exposed proteins of *Haemophilus influenzae* mediate binding to macrophages. *J. Infect. Dis.* **169,** 425–429 (1994).
27. Stock, J. B. & Surette, M. G. Chemotaxis. in *Escherichia coli and Salmonella: Cellular and Molecular Biology.* (ed. Neidhardt, F. C.) 1103–1129 (ASM, Washington, DC, 1996).
28. Armitage, J. P. & Schmitt, R. Bacterial chemotaxis: *Rhodobacter sphaeroides* and *Sinorhizobium meliloti*—variations on a theme? *Microbiology* **143,** 3671–82 (1997).
29. McBride, M. J., Weinberg, R. A. & Zusman, D. R. "Frizzy" aggregation genes of the gliding bacterium *Myxococcus xanthus* show sequence similarities to the chemotaxis genes of enteric bacteria. *Proc. Natl Acad. Sci. USA* **86,** 424–428 (1989).
30. Darzins, A. Characterization of a *Pseudomonas aeruginosa* gene cluster involved in pilus biosynthesis and twitching motility: sequence similarity to the chemotaxis proteins of enterics and the gliding bacterium *Myxococcus xanthus*. *Mol. Microbiol.* **11,** 137–153 (1994).
31. Ohga, T., Masduki, A., Kato, J. & Ohtake, H. Chemotaxis away from thiocyanic and isothiocyanic esters in *Pseudomonas aeruginosa*. *FEMS Microbiol. Lett.* **113,** 63–66 (1993).
32. Nelson, J. W. *et al.* Mucinophilic and chemotactic properties of *Pseudomonas aeruginosa* in relation to pulmonary colonization in cystic fibrosis. *Infect. Immun.* **58,** 1489–1495 (1990).
33. Kato, J., Sakai, Y., Nikata, T. & Ohtake, H. Cloning and characterization of a *Pseudomonas aeruginosa* gene involved in the negative regulation of phosphate taxis. *J. Bacteriol.* **176,** 5874–5877 (1994).
34. Moulton, R. C. & Montie, R. C. Chemotaxis by *Pseudomonas aeruginosa*. *J. Bacteriol.* **137,** 274–280 (1979).
35. O'Toole, G. A. & Kolter, R. Flagellar and twitching motility are necessary for *Pseudomonas aeruginosa* biofilm development. *Mol. Microbiol.* **30,** 295–304 (1998).
36. Holloway, B. W. Genetic recombination in *Pseudomonas aeruginosa*. *J. Gen. Microbiol.* **13,** 572–581 (1955).
37. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8,** 195–202 (1998).
38. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185 (1998).
39. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115 (1998).
40. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
41. Brown, M. H., Paulsen, I. T. & Skurray, R. A. The multidrug efflux protein NorM is a prototype of a new family of transporters. *Mol. Microbiol.* **31,** 394–395 (1999).
42. Cox, J. S., Chen, B., McNeil, M. & Jacobs, W. R. Jr. Complex lipid determines tissue-specific replication of *Mycobacterium tuberculosis* in mice. *Nature* **402,** 79–83 (1999).
43. Tseng, T.-T. *et al.* The RND permease superfamily: an ancient, ubiquitous and diverse family that includes human disease and development proteins. *J. Molec. Microbiol. Biotech.* **1,** 107–126 (1999).
44. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

X. Q. Pham managed the later stages of the shotgun, closure and finishing phases of the genome sequencing; A. Erwin did essential work in the annotation of the sequence and preparation of this manuscript; S. Mizoguchi managed the genome informatics databases, web site and analysis tools development; Kim Smith managed much of the primary-shotgun data collection; and F. Brinkman managed the *P. aeruginosa* collaborative annotation project. A total of 1741 genes were annotated by 61 expert volunteers from the *Pseudomonas* research community (L. Adewoye, A. M. Antonio, S. K. Arora, M. Bailey, S. Beatson, W. Bitter, R. Blakeley, C. O. Carranza, P. Cooke, P. Cornelis, L. Croft, T. de Kievit, R. De Mot, B. Erni, M. Eschbach, G. Fichant, C. J. Fields, A. Filloux, J. Campos-García, P. Hager, T. Hayashi, R. Herzog, B. Huang, V. Huang, B. H. Iglewski, Y. Itoh, K.-E. Jaeger, D. Jahn, Z. Jalil, J. Kato, M. Kertesz, A. M. Kropinski, J. Lam, I. Lamont, A. Lazdunski, P. C. Lau, R. Levesque, K. Mathee, J. Mattick, J. Nezezon, M. Ochs, G. O'Donovan, L. Passador, P. Phibbs, Y. Quentin, A. Rodrigue, P. H. Roy, M. Saier, F. Sanschagrin, M. J. Schurr, P. C. Seed, T. H. M. Smits, G. Soberón-Chávez, L. G. Treviño, B. Tümmler, J. van Beilen, M. Vasil, C. Whitchurch, K. Williams, K. Wong and B. Worobec). We thank P. Green and B. Ewing for assistance with the whole-genome phrap assemblies. Major funding for this project was provided by the Cystic Fibrosis Foundation and PathoGenesis Corporation. Funding for the organization of the PseudoCAP project was from the Canadian, French and German Cystic Fibrosis Foundations. R.E.W.H. was recipient of the Medical Research Council of Canada (MRC) Distinguished Scientist Award and received funding from the MRC.

Correspondence and requests for materials should be addressed to M.V.O. e-mail: mvo@u.washington.edu). The complete nucleotide sequence and sequences of predicted ORFs have been deposited with GenBank (accession no. AE004091). Sequence data, annotations and details of many of the analyses described here are also available on the *P. aeruginosa* genome web site, http://www.pseudomonas.com.