An interactive web-based *Pseudomonas aeruginosa* genome database: discovery of new genes, pathways and structures

Larry Croft,¹,² Scott A. Beatson,¹,² Cynthia B. Whitchurch,¹ Bixing Huang,¹ Robert L. Blakeley² and John S. Mattick¹,²

Author for correspondence: John S. Mattick. Tel: +61 7 3365 4446. Fax: +61 7 3365 4388. e-mail: j.mattick@imb.uq.edu.au

Using the complete genome sequence of *Pseudomonas aeruginosa* PAO1, sequenced by the *Pseudomonas* Genome Project (ftp://ftp.pseudomonas.com/data/pacontigs.121599), a genome database (http://pseudomonas.bit.uq.edu.au/) has been developed containing information on more than 95% of all ORFs in *Pseudomonas aeruginosa*. The database is searchable by a variety of means, including gene name, position, keyword, sequence similarity and Pfam domain. Automated and manual annotation, nucleotide and peptide sequences, Pfam and SMART domains (where available), Medline and GenBank links and a scrollable, graphical representation of the surrounding genomic landscape are available for each ORF. Using the database has revealed, among other things, that *P. aeruginosa* contains four chemotaxis systems, two novel general secretion pathways, at least three loci encoding F17-like thin fimbriae, six novel filamentous haemagglutinin-like genes, a number of unusual composite genetic loci related to vgr/Rhs elements in *Escherichia coli*, a number of fix-like genes encoding a micro-oxic respiration system, novel biosynthetic pathways and 38 genes containing domains of unknown function (DUF1/DUF2). It is anticipated that this database will be a useful bioinformatic tool for the *Pseudomonas* community that will continue to evolve.

**Keywords:** proteome, gene families, regulators, adhesins, type II secretion

---

**INTRODUCTION**

The *Pseudomonas* Genome Project (incorporating the Cystic Fibrosis Foundation, the University of Washington Genome Center and the PathoGenesis Corporation) has collaborated to determine the complete genome sequence of *Pseudomonas aeruginosa* PAO1. During the sequencing process contig data were periodically released (as quarterly updates) to the public at http://www.pseudomonas.com/, culminating in the release of a complete sequence as a single contig on 15 December 1999. To provide a user-friendly web-based genome browser for the *Pseudomonas* research community, we undertook a preliminary annotation of the publicly available sequences. It was anticipated that this database would provide a useful platform for genomic analysis in the interim period between early release of the sequence data and final publication of the complete annotated genome, but it has now evolved into an interactive resource which will aid research into *P. aeruginosa* on a continuing basis. This database includes more sophisticated analyses, such as domain information, Medline and GenBank links, and is now searchable by a variety of criteria including BLAST and Pfam domains.

**METHODS**

The publicly available genome sequence was downloaded from the *Pseudomonas* Genome Project (ftp://ftp.pseudomonas.com/data/pacontigs.121599). ORFs were identified in the genome sequence using the gene-predicting software Glimmer 2.0 (http://www.tigr.org/softlab/glimmer/glimmer.html). All ORFs were aligned using BLAST (gapped BLAST 2.0; Altschul et al., 1997) against GenPept (release 115) and the incomplete *Pseudomonas putida* genome (for which preliminary sequence data were obtained from The Institute for
Genomic Research web site at http://www.tigr.orgnotations. The BLAST output was parsed using a nawk script and loaded into a Postgres SQL (http://www.postgresql.org/) database. A web-based genome viewer was written in PHP (http://www.php.net/), a server side language, to interrogate the SQL database and to generate HTML and graphics. PHP was mounted on an SGI Origin 200 web server. BLAST and HMMER were run on an SGI Origin 2000, an SGI Octane and an SGI Origin 2100.

A total of 6405 ORFs were surveyed manually to identify errors such as large overlaps, the majority of which were found to be opposite strand ORF shadows on the basis of weak BLAST matches to GenPept and/or the P. putida genome. Of these, 1162 apparently spurious ORFs were deleted. The remaining 5243 ORFs were loaded into a public web-browsable genome database. BLAST outputs were manually analysed and an annotation was added to the database for each ORF. The BLAST matches were also used to automatically name known P. aeruginosa genes and also the closest named protein from other organisms. Furthermore, additional data were parsed from the BLAST output using a perl script to produce a descriptive line, for example ‘n% identical to [GenBank accession no.] gene product [organism] in a 341 aa overlap’. Domain analysis of our ORF set was performed using SMART (http://smart.embl-heidelberg.de/) (Schultz et al., 2000) and the Pfam database (http://pfam.wustl.edu/) (Bateman et al., 2000) using HMMER (http://hmmer.wustl.edu/). Domain data were added to the SQL database and integrated into the genome browser in both tabular and graphical form.

RESULTS AND DISCUSSION

The database is searchable (see http://pseudomonas.bit.uq.edu.au/gene_browser.phtml) by the following criteria:

- **gene name**, restricted to known P. aeruginosa genes;
- **name of homologue**, found in ‘closest named protein match’ field (see below);
- **gene ID number**, a unique identifier exclusive to this database;
- **keyword**, found in manual annotation;
- **position in genome**, according to nucleotide number, minutes or a clickable genome map;
- **Pfam domains**, from a selection of 665 domains identified by HMMER analysis;
- **sequence similarity**, using BLASTX, BLASTN and BLASTP.

When a search is performed the results are listed on a separate page with links for each ORF to the gene viewer. The gene viewer (Fig. 1) contains the following information:

- **gene name**, automatically generated from BLASTP output;
- **gene ID number**;
- **gene description**, a manually annotated field;
- **gene position**, indicated by ORF start and ORF stop fields (note these do not indicate transcriptional orientation, merely the region of the genome spanned by the ORF);
- **peptide length**;
- **closest named protein match**, automatically generated from BLASTP output corresponding to the closest homologue in which the protein name has been defined in GenBank;
- **comment line**, automatically generated from BLASTP output corresponding to the closest homologue;
- **Medline links** to known P. aeruginosa genes and to the closest named protein match;

![Fig. 1. Pseudomonas Genome Database screen image showing the Genome Viewer page for the gene with ID number 4819. Features for each ORF shown on this page include the P. aeruginosa gene name (if known), Gene ID number, a manually annotated description, ORF start and stop (note these do not indicate transcriptional orientation, merely the region of the genome spanned by the ORF), predicted peptide length, comment line corresponding to the most similar protein obtained in a BLASTP search of GenBank, closest named protein match corresponding to the most similar named protein in the BLASTP search, a graphical view of surrounding ORFs in a 20 kb window, Medline links, SMART domains, Pfam domains, and nucleotide and peptide sequences of the ORF. The graphical viewer can be scrolled in both directions in increments of 20 kb (large-arrowed buttons) or 8 kb (small-arrowed buttons).](image-url)
smart domains and sequence features, such as coiled coils, signal peptides and transmembrane regions, in tabular form with links to additional information, alignments and structures;

- Pfam domains in graphical form and in tabular form with links to domain information and alignments;
- genomic context, a graphical view of surrounding ORFs in a 20 kb scrollable window including transcriptional orientation and reading frame;
- nucleotide and peptide sequence.

The genome database contains information for 5243 ORFs which we estimate represent more than 95% of the P. aeruginosa proteome. A total of 1291 ORFs had no or low similarity (defined here as expectation value E > 10^-10) to all other sequences in the current databases. While the formal annotation and publication of the P. aeruginosa genome is the prerogative of those who have done the sequencing and assembly, the genome database described herein provides a useful interactive tool for gene discovery and analysis of the publicly available genome sequence. Searching of the database enables the identification and retrieval of information on various categories of proteins, their genomic position and surrounding genomic landscape, with graphical outputs of domain homologies and links to other databases, including Medline.

This database can be used to analyse genes and gene clusters interactively and in more detail. We have identified a number of interesting loci which illustrate the usefulness of the database for probing P. aeruginosa molecular biology. Analysis of the database reveals that the P. aeruginosa genome contains, for example, 84 ORFs encoding proteins with a CheY-like domain (‘response regulator receiver domain’ in Pfam) and 55 ORFs encoding proteins with a histidine kinase domain (E < 10^-7) seventeen proteins were found to contain both domains, indicating that they could function as dual sensor-regulators, and two of these contained an additional C-terminal CheY-like domain. There are three orphan histidine kinases and 19 proteins containing receiver domains that are not associated with a cognate histidine kinase. Numerous other proteins with Pfam domains representing other well-characterized regulatory domains were identified, including 117 LysR-type and 58 AraC-type helix–turn–helix domains (E < 10^-7).

The genome database also reveals that P. aeruginosa contains 38 ORFs with DUF1 and/or DUF2 domains (SMART), which are domains of unknown function. DUF1 domain was first recognized in the Caulobacter crescentus signalling protein PleD which controls cell differentiation (Hecht & Newton, 1995; Aldridge & Jenal, 1999) and subsequently, along with DUF2, in enzymes which also contain putative oxygen-sensing domains and that are involved in cyclic di-GMP regulation of cellulose synthesis in Acetobacter xylinus (DUF1 and DUF2; Tal et al., 1998). These domains, originally referred to as GGDEF (DUF1) and EAL (DUF2), are found in a wide variety of bacterial proteins and occur in P. aeruginosa either singly (often with predicted transmembrane domains), together or in various combinations with other known or suspected signalling domains, including (i) CheY-like response regulator receiver domains; (ii) PAS/PAC domains (which are input-sensing domains found in many prokaryotic and eukaryotic regulatory proteins, including those regulating virulence, oxygen sensing, hydrogen sensing, sporulation, nitrogen metabolism, light reception and photopigments and circadian rhythms; see Ponting & Aravind, 1997); (iii) GAF domains (which are present in diverse phototransducing proteins and cGMP-specific phosphodiesterases; see Aravind & Ponting, 1997); (iv) HAMP domains (which are also known as ‘domain found in bacterial signalling proteins’ (Pfam) and which occur in numerous bacterial histidine kinases, adenyllyl cyclases, methyl-accepting proteins and other signalling proteins, possibly involved in sensing informational change; see Aravind & Ponting 1999); and (v) periplasmic substrate-binding domains, also known as ‘extracellular solute-binding proteins, family 3’ (Pfam). It is clear that DUF1- and DUF2-containing proteins are all members of a signal transduction or biochemical system whose precise function(s) and mode of action remains virtually unknown, but which indicates that there is still much to learn about signalling pathways and networks in P. aeruginosa.

Analysis of the P. aeruginosa genome with this database has also yielded other interesting observations, including the discovery of multiple chemotaxis-like systems, novel filamentous haemagglutinin (FHA)-like genes, novel type 1 fimbrial genes and composite genetic elements, among others, which were hitherto not known to exist in P. aeruginosa and which are presented below. The positions of loci are given as nucleotide numbers corresponding to the 15 December 1999 sequence release.

P. aeruginosa possesses multiple chemotaxis-like systems

The term chemotaxis is used to describe the ability of bacteria to adjust their swimming motility in response to environmental chemorepellents/attractants. The best characterized of the chemotaxis systems is that of Escherichia coli (Parkinson, 1993; Eisenbach, 1996). This system is composed of membrane-bound methyl-accepting chemotaxis proteins (MCPs) which sense specific environmental signals and are methylated or demethylated by the methyltransferase CheR or the methylesterase CheB, respectively, as part of a feedback circuit which adjusts the methylation state of the MCP allowing temporal sensing of chemical gradients. The MCPs are coupled via the adaptor protein CheW to the histidine kinase CheA which autophosphorylates in response to cues from the MCPs. CheA then phosphorylates the response regulator CheY which interacts directly with the flagellar motor switch to control the direction of flagellar rotation, conferring chemotaxis to the bacterium. CheA also competently phosphotransfers to the methylesterase CheR. This
system also contains another protein, CheZ, which serves to dephosphorylate the response regulator CheY. Related chemotaxis systems have been described in many other bacteria. These systems share the six core components of an MCP, and homologues of CheW, CheA, CheY, CheR and CheB, but may also include other proteins which are specific for each system.

The chemotaxis system which controls swimming motility in P. aeruginosa has been described by Masduki et al. (1995) and Kato et al. (1999). This system is encoded by the genes cheY, cheZ, cheA, cheB and cheW, which are located in a cluster downstream of the flagellar structural genes, and cheR, which is reportedly situated at a distinct locus (> 1·8 Mb away; Kato et al., 1999). Three MCPs encoded by pctA, pctB and pctC have been identified which feed into this chemotaxis pathway and are also situated at a distinct locus (Kuroda et al., 1995; Taguchi et al., 1997). A second related system has been described in P. aeruginosa which controls type IV fimbrial-mediated twitching motility (Darzins, 1993, 1994, 1995; Alm & Mattick, 1997; C. B. Whitchurch and others, unpublished). This system is composed of a putative MCP (PilJ), a methyl transferase CheR-like protein (PilK), two CheW homologues (PilL and ChpC), a methyl esterase CheB homologue (ChpB) and a CheA-like histidine kinase ChpA. This system also has three CheY-like response regulator/receiver modules: PilG, PilH and a domain located at the C terminus of ChpA.

Using the search word ‘chemotaxis’ in the ‘search by word in description’ field on the ‘Search’ page of the database yields 52 ORFs with this word in the description line. Of these, 26 encode putative MCPs which are scattered throughout the genome. Only four of these genes (pctA, pctB, pctC and pilJ) have been characterized to date, which means much remains to be learned about chemo-sensing in P. aeruginosa and the specificities of each of these MCPs, as well as the particular pathways into which they connect.

The remaining chemotaxis ORFs are situated in five distinct clusters (see genome positions ~194014, 449639, 1585440, 3759948 and 4143740 and surrounds; Fig. 2). The locus at ~1585440 contains the swimming chemotaxis genes cheY, cheZ, cheA, cheB and cheW. Situated between cheB and cheW are four ORFs (1, 2, 3 and 4; Fig. 2). ORF1 and ORF2 have been reported to encode proteins which are homologous to the bacterial flagellar motor proteins MotA and MotB, respectively (Kato et al., 1999). The Pfam domain analysis for ORF3 (Gene ID 1909), which is shown on the graphical viewer page of our database, indicates that the predicted product of this gene contains a ParA family ATPase-like domain.

**Fig. 2.** Genetic organization of chemotaxis gene clusters. The 6·3 Mb P. aeruginosa genome is represented by the vertical line and genomic positions for each of the chemotaxis gene clusters are indicated. Sequence positions correspond to the 15 December 1999 sequence release of the P. aeruginosa PA01 genome. Individual ORFs are boxed and their orientation is denoted by arrows. The predicted class of protein encoded by each ORF is indicated within the boxes (Y, CheY; W, CheW). ORFs encoding proteins of a similar type are represented in the same colour. Gene ID numbers used in our database (http://pseudomonas.bit.uq.edu.au/) are indicated below each ORF box. Names of previously described P. aeruginosa genes are indicated below the relevant ORF.
This domain is found in a family of proteins (ParA, Soj) which are involved in bacterial chromosome and plasmid partitioning, and is also found in the septum-site-determining protein MinD (Mottaleb-Veshareh et al., 1990; Wheeler & Shapiro, 1997). BLAST searches of GenBank with the predicted ORF3 protein sequence confirm that this protein is homologous to proteins belonging to this family. The Pfam and SMART analyses of ORF4 (Gene ID 1911) indicate that the predicted protein contains a C-terminally located CheW domain. Interestingly, BLAST searches of GenBank revealed that *P. putida* and *Vibrio parahaemolyticus* also contain proteins which are closely related to the predicted ORF3 and ORF4 protein sequences. These homologues in *P. putida* and *V. parahaemolyticus* are situated within chemotaxis clusters with the same genetic arrangement as that of this swimming chemotaxis gene cluster of *P. aeruginosa*, except that in *V. parahaemolyticus* the *motA* and *motB* genes are situated elsewhere (see GenBank AF031898 and AF069392). The roles of the ORF3 and ORF4 proteins and their *P. putida* and *V. parahaemolyticus* homologues are unknown but are predicted to be chemotaxis-related.

The gene *cheR*, which is required for swimming chemotaxis in *P. aeruginosa*, is situated around 2 Mb away, as predicted by Kato et al. (1999), at ~3759948 (Fig. 2). Upstream of *cheR* is an ORF which is predicted to encode a protein homologous to CheV of *Campylobacter jejuni*, *Helicobacter pylori*, *Bacillus subtilis* and *V. parahaemolyticus* (Fig. 2). These proteins are composed of two domains, an N-terminal CheW domain and a C-terminal CheY domain. Interestingly, the *P. aeruginosa* CheV homologue is very closely related to that of *V. parahaemolyticus* (57% identity, 74% similarity) and, as is the case in *P. aeruginosa*, the cheV gene of *V. parahaemolyticus* is also located directly upstream of *cheR* at a site remote from that of the remaining chemotaxis genes (see GenBank U12817). The conserved genetic arrangement of the swimming chemotaxis genes of *P. aeruginosa* and *V. parahaemolyticus* suggests lateral gene transfer and/or a common ancestry, and it will be interesting to compare not just gene and predicted protein sequences, but also the conservation of gene order and genomic landscape between species as a means of trying to unravel their histories. The role of the *P. aeruginosa* CheV homologue is unknown but we predict that it may also play some role in the chemotactic control of swimming motility, or a related function. The genes encoding the MCPs PctA, PctB and PctC, which are known to control swimming motility in response to amino acids, are located at ~481252.

The cluster of twitching-motility-related chemotaxis-like genes, *pilG, pilH, pilI, pilJ, pilK, chpA, chpB* and *chpC*, is situated at ~449639 (Fig. 2). These genes are similar in arrangement and structure to the frz genes controlling gliding motility in *Myxococcus xanthus* (Ward & Zusman, 1999), except that they are more complex. In particular, the central *chpA* gene encodes a protein with six predicted Hpt (histidine phospho-transfer) domains, one of which has a serine in place of the conserved histidine, as well as a C-terminal CheY receiver domain, making this one of the most complex signal transduction proteins yet described in nature (C. B. Whitchurch and others, unpublished). It is also worth noting that the current release of the *P. aeruginosa* genome sequence appears to contain a sequence error at this point, or at least a difference from the pre-existing GenBank entry covering this region, which results in the loss of a stop codon and the apparent fusion of two adjacent genes (*pilL* and *chpA*; see Gene ID 551, cf. GenBank U79580). Given the magnitude of the sequence and the high G+C content of the genome, it is not necessarily surprising that this sequence release may have errors which will affect the accuracy of the annotation at various sites.

The remaining two clusters of putative chemotaxis-like genes (clusters 4 and 5; Fig. 2) are entirely novel and have no known function. Cluster 4 is a complete set of chemotaxis-like genes encoding homologues of CheY, CheA, CheW, CheR and CheB, and has two associated MCPs. This cluster appears to be related to the chemotaxis genes of a number of bacteria including Caulobacter crescentus, Rhodobacter sphaeroides and Bacillus subtilis in that it also encodes a homologue of the chemotaxis protein CheD which facilitates the methylation of MCPs by CheR in these systems (Rosario & Ordal, 1996).

Cluster 5 encodes homologues of CheR, CheB, CheA and CheW as well as an MCP and a second protein containing a C-terminally located CheW domain. Unlike the other chemotaxis-like gene clusters of *P. aeruginosa*, cluster 5 does not encode an individual CheY protein. However, SMART and Pfam analyses show the CheA homologue of this cluster (Gene ID 4817) is actually a CheA/CheY hybrid protein, like ChpA, as it also possesses a response regulator receiver domain situated at its C terminus. SMART and Pfam analyses also show that a CheY-like response regulator receiver domain is found at the N terminus of the protein encoded by the ORF situated at the end of this cluster (Gene ID 4814). These analyses also show that this protein contains a DUF1 domain at its C terminus (see above). Other proteins with a similar domain structure include the C. crescentus protein PleD, which is part of a signal transduction pathway controlling cell differentiation (Hecht & Newton, 1995; Aldridge & Jenal, 1999). We predict that this protein in *P. aeruginosa* participates in a phosphotransfer-dependent signal transduction pathway involving the cluster 5 chemotaxis-like system. Interestingly, this cluster shows very close similarity and syntony to a locus of seven genes in *Pseudomonas fluorescens*, termed the *wsp* (‘wrinkly spreader’) operon (E. Bantanaki & P. B. Rainey, personal communication). Mutations in the *wsp* operon affect the production of cellulose (which is encoded by the *wss* operon) and abolish rapid surface spreading of colonies (A. J. Spiers & P. B. Rainey, unpublished). Intriguingly, however, PAO1 lacks a homologue of the *wss* operon.

It is also interesting that the twitching motility chemotaxis-like system, the swimming motility chemotaxis system, and the swimming chemotaxis system have no known homologues in *P. aeruginosa*. However, it is possible that the *P. aeruginosa* PAO1 lacks a homologue of the *wss* operon.
system and the _wsp_-like chemotaxis system contain more than one _CheW_ module. Considering the large number of ORFs encoding putative MCPs (26) scattered around the genome, it is tempting to speculate that the presence of multiple _CheW_ homologues in these systems may facilitate the interaction of a greater number of MCPs into the various signal transduction pathways, thereby diversifying the range of chemical signals to which these pathways can respond.

**Novel general secretion pathways**

*P. aeruginosa* contains at least 35 genes required for the biosynthesis and function of type IV fimbriae (Alm & Mattick, 1997; Semmler _et al_., 2000), which are polar filaments involved in epithelial attachment and twitching motility (Semmler _et al_., 1999). These filaments are composed of a structural subunit, PilA or pilin, which has a highly conserved and highly hydrophobic N-terminal domain that forms the core of a helical structure (Parge _et al_., 1995). Assembly of type IV fimbriae also requires a number of accessory proteins, including other pilin-like proteins, a specific prepilin peptidase, inner- and outer-membrane proteins and nucleotide-binding proteins. Homologues of these proteins are also involved in type II protein secretion in bacteria, suggesting that these are all subsets of a supersystem which shares a common evolutionary origin, architecture and functional basis (Hobbs & Mattick, 1993; Mattick & Alm, 1995). Given our laboratory’s long-standing interest in this system we used the genome database to see if any other type IV prepilin subunits of locus A are more difficult to categorize according to family class. Prepilin-like subunits from those encoding type IV pilin subunits.

**Overall, the prepilin-like subunits of locus B** showed somewhat less similarity to the respective _P. aeruginosa_ Xcp homologue (Tommassen _et al_., 1992; Bleves _et al_., 1998). In contrast, with the exception of prepilin-like subunits A1 and A5 (which share homology with the GspK and GspP subfamilies, respectively) the remaining prepilin-like subunits of locus A are more difficult to differentiate according to family class. Prepilin-like subunits A2 and A4 exhibit a variant PilD-dependent N-terminal motif (A−/F+1/E−6) instead of G−/F+1/E−6, although this variation has been shown to be partially conservative by site-directed mutagenesis of _P. aeruginosa_ PilA (Strom & Lory, 1991), and a leader peptide terminating in an alanine residue has been subsequently identified in _P. aeruginosa_ homologues FimT of _P. aeruginosa_ (Alm & Mattick, 1996) and CglC/CglD of _Streptococcus pneumoniae_ (Pestova & Morrison, 1998). The absence of +24 and +27 tyrosine residues (Fig. 3) and obvious C-terminal cysteine residues capable of forming disulfide bonds distinguished all 10 novel subunits from those encoding type IV pilin subunits. Notably, each locus contained at least one prepilin-like subunit with similarity to GspK-like protein subunits that lack the conserved E−6 residue characteristic of typical prepilin-like leader sequences (Alm _et al_., 1996).

The similarity of type IV fimbriae and type II protein secretion apparatus and DNA uptake extends beyond the common use of prepilin-like subunits to include the accessory proteins: GspE-like nucleotide-binding proteins PilB/XcpR/ComG-1; GspF-like inner-membrane proteins PilC/XcpS/ComG-2; GspO-like prepilin peptidase PilD (shared by fimbrial and type II secretion prepilin subunits)/ComC; and GspD-like outer-membrane proteins PilQ/XcpQ (Alm & Mattick, 1997; Tommassen _et al_., 1992; Albano _et al_., 1989). Using the graphical viewer of our database it was clear that locus A contained ORFs encoding a nucleotide-binding protein and inner-membrane protein, whilst

---

**Fig. 3.** Multiple alignment of prepilin-like subunits. Sequences were aligned with CLUSTAL W using the MacVector application. Identical residues are shaded black and conserved residues are boxed. The conserved G−/F+1/E−6 motif is denoted with symbols (**). The Y−24Y−27 motif for type IV prepilin subunits is presented in lower case letters for the _P. aeruginosa_ PilA sequence. Gene ID numbers used in our database are 3478, 3477, 3476, 3474 and 3473, and 889, 887, 8002, 885 and 884 for prepilin-like subunits A1–A5 and B1–B5, respectively.
locus B contained three ORFs encoding a nucleotide-binding protein and inner- and outer-membrane proteins (Fig. 4). Furthermore, a closer examination of locus B revealed the existence of GspC, GspL and GspM homologues (Fig. 4). The presence of an apparently complete set of type II secretion genes paralogous to the xcp cluster suggests a possible role for locus B in protein secretion, independent of the Xcp gene products.

The extent to which other components of type IV fimbriae and type II secretion systems interact with the products of the novel loci remains to be seen, but it is reasonable to assume that, in the absence of an ORF for another prepilin peptidase, PilD is required to process the novel prepilin-like subunits. Interestingly, there is an additional homologue of an XcpR (nucleotide-binding protein) at an unrelated position in the genome (~5.865 Mb) (Fig. 4). Although the overall identity with XcpR was only 32%, the predicted protein contained an intact Walker Box (GxxGxGKT; Turner et al., 1993), indicating that it was capable of binding ATP. This protein may intersect with one or other of the homologous systems in P. aeruginosa or alternatively may even contribute to type IV fimbrial function (like PilB, PilT and PilU). Three additional ORFs encoding proteins with homology to GspD were identified in the genome: two with low homology to GspD at positions 1498583 and 4828262 (Gene ID 1830 and 5605, respectively), and the type III secretion protein PscC (position 1858372, Gene ID 2236) contains a C-terminal domain with homology to GspD (Yahr et al., 1996). Given that XqhA has been shown to be responsible for residual protein export in P. aeruginosa ΔxcpQ (Martinez et al., 1998), it is conceivable that either of the two orphan GspD homologues (or indeed XqhA) interacts with the components of one or more of these putative type II secretion systems. Mutational studies will be required to show if P. aeruginosa is capable of producing more than two structures related to the general secretory pathway (Pugsley, 1993) and if so, the nature of their cargo and how their respective components interact (if at all) with the established systems.

**Novel F17-like fimbriae loci**

The P. aeruginosa genome contains at least three loci encoding type 1-like fimbriae, which are quite distinct from type IV fimbriae. These fimbriae are commonly found in enteric pathogens and include classical type 1 fimbriae, P pili and F17 fimbriae, all of which belong to the group of fimbrial structures assembled by the chaperone/usher pathway and which appear to function strictly as adhesion devices (Hung & Hultgren, 1998; Soto & Hultgren, 1999). This type of fimbria has not previously been identified in P. aeruginosa. The chaperone/usher pathway requires the interaction of a periplasmic immunoglobulin-like chaperone protein with a fimbrial subunit, targeting of the chaperone--
subunit complex to an outer-membrane usher protein, disassociation of the complex and assembly into pili across the outer membrane (Hung & Hultgren, 1998). The simplest fimbriae assembled by this pathway are the ‘thin, flexible pili’ (including F17) which generally require only a chaperone, usher, subunit and an adhesin (assembled at the distal end of the structure) (Soto & Hultgren, 1999).

Initial BLAST searches of our database identified three genes which showed similarity to the major subunit of E. coli F17-A fimbriae (see positions ∼1073235, 4569005 and 2342419 for loci A, B and C, respectively; Fig. 5). Each fimbrial subunit contained a signal sequence required for export from the cytoplasm (a feature of known fimbrial subunits) and a typical β-zipper motif (a penultimate tyrosine, alternating hydrophobic residues at positions 4, 6 and 8, and a glycine at position 14 from the C terminus) known to be involved directly in the interaction with a chaperone (Holmgren et al., 1992). By using the Pfam search tool to identify ORFs in our database containing the ‘Gram-negative pili assembly chaperone’ Pfam domain (Bateman et al., 2000), we found that all three fimbrial subunit genes were associated with one (or two) gene(s) encoding a FGS-type fimbrial chaperone protein (Fig. 5). A multiple alignment of putative P. aeruginosa chaperone peptide sequences with PapD (the prototype FGS-type chaperone; Lindberg et al., 1989) revealed that the P. aeruginosa homologues shared the majority of residues found to be conserved in other members of the chaperone family (Holmgren et al., 1992) (data not shown). Similarly, a Pfam search using the ‘fimbrial usher protein’ Pfam domain (Bateman et al., 2000) revealed that all three loci were accompanied by genes encoding typical fimbrial usher proteins (Hung & Hultgren, 1998; Fig. 5). These searches also revealed other, apparently orphan proteins with similarity to fimbrial chaperones and ushers (at position 558380 and 5217449; Gene ID 665 and 6046, respectively) which are adjacent to ORFs encoding proteins with low similarity to other fimbrial proteins and that may represent other as yet unrecognized fimbrial loci or surface structures.

All three novel fimbrial loci have operon-like arrangements with a conserved gene order, encoding fimbrial subunit, chaperone and outer-membrane usher protein, respectively (Fig. 5). This pattern is suggestive of locus duplication, as reported for the fimbriae of Haemophilus influenzae biogroup aegyptius (Read et al., 1996), although the observation that there is a similar degree of divergence between orthologues and paralogues in P. aeruginosa indicates that duplication may have been very ancient and horizontal transfer cannot be discounted. Interestingly, immediately 3’ of locus C is an ORF encoding a protein of unknown function and an additional chaperone gene. The closest homologue of the unknown protein is a Yersinia pestis hypothetical protein (yp39) which is also found in a cluster of putative fimbrial genes with an identical arrangement to locus C (Buchrieser et al., 1998, 1999). In P. aeruginosa, fimbrial locus B also contains an ORF (that may be co-transcribed) encoding an FHA-like protein (see following section). Although the genes for related fimbriae are associated with the genes for FHA expression in Bordetella pertussis, no functional relationship has been established. However, their close linkage suggests a common regulatory pathway (Willems et al., 1994).
Loci B and C and the homologous locus in \textit{Y. pestis} each have an additional chaperone, a feature not previously described in other organisms. Fimbrial chaperones and outer-membrane ushers are known to operate in parental pairs, which are able to assemble foreign fimbrial subunits (albeit less efficiently than their natural partners) provided they possess the necessary motifs (i.e. a \(\beta\)-zipper) (Jones et al., 1993). However, the partners within chaperone/usher pairs are not thought to be as readily interchangeable (Klemm et al., 1995; Jones et al., 1993). Given that the order of subunit assembly is proposed to be determined by kinetic partitioning (Hung & Hultgren, 1998), it is possible that additional chaperones may allow more efficient incorporation of fimbrial subunits from heterologous loci via heterologous membrane usher proteins. Elucidating the role (if any) that an additional chaperone plays in fimbrial biogenesis awaits mutational studies.

The adhesive specificity of type 1 and related fimbriae is usually determined by the structure of the tip protein(s) known as fimbrial adhesins (Hultgren et al., 1993). Fimbrial adhesins can be thought of as having a C-terminal region with the conserved features of the major and minor fimbrial subunits that allow interaction with a chaperone (i.e. \(\beta\)-zipper), and a variant N-terminal receptor-binding domain (Hung & Hultgren, 1998). Generally, fimbrial adhesins are encoded in the same cluster as the other structural components of the fimbriae (Hung & Hultgren, 1998). Gene 2776 of locus C (corresponding to \textit{y39} of \textit{Y. pestis}) has a signal sequence and similarity to fimbrial subunits in the C-terminal domain (including a \(\beta\)-zipper), indicating that it probably encodes a fimbrial adhesin. Similarly, gene 5316 of locus B also appears to be a fimbrial adhesin. Although there was no apparent adhesin gene associated with locus A, a gene with low similarity to type 1 fimbrial subunits was identified elsewhere in the genome (position 5948579; Gene ID 6870). Closer analysis revealed a signal sequence, \(\beta\)-zipper and an N-terminal domain with no similarity to other fimbrial subunits, suggesting that this ORF is likely to encode a fimbrial adhesin. It should also be noted that genes 663 and 664, associated with the aforementioned orphan chaperone, could also encode fimbrial adhesins. Furthermore, the existence of other, as yet unidentified adhesin ORFs in the genome cannot be ruled out. A single fimbrial adhesin, \textit{FimD}, is known to be shared by two fimbrial types in \textit{Bordetella pertussis} to facilitate antigenic variation (Willems et al., 1992) and thus it is possible that this type of sharing also occurs in \textit{P. aeruginosa}.

Finally, we undertook a preliminary search for physical evidence of thin fimbrial structures by electron microscopy. Fig. 6 shows the presence of at least one type, and possibly two types, of peritrichous thin fimbriae in \textit{P. aeruginosa}. This is the first time, to our knowledge, that this type of fimbriae has been reported in \textit{P. aeruginosa}. Which loci encode which particular fimbriae and whether these fimbriae are involved in \textit{P. aeruginosa} virulence, and under what circumstances, will require gene knockout and \textit{in vivo} and \textit{in vitro} analyses, but the presence of such fimbriae suggests that a new aspect of the biology of \textit{P. aeruginosa} remains to be explored.

\textbf{FHA-like genes}

During annotation of the genome we also identified two very large ORFs, encoding proteins with similarity to FHA (Stibitz et al., 1988; Locht et al., 1993), which we have termed FHA-like A (FHA-A) and FHA-like B (FHA-B) at positions 2761873 and 42914, respectively (Fig. 7). FHA is a major attachment factor produced by virulent \textit{Bordetella} spp. (Stibitz et al., 1988; Dominghini et al., 1990). The 370 kDa FHA precursor is exported and activated by a specific accessory protein (FhaC) in a manner analogous to that observed for the export and activation of the haemolysins, ShlA and HpmA, of \textit{Serratia marcescens} and \textit{Proteus mirabilis}, respectively (Poole et al., 1988; Uphoff & Welch, 1990). FHA, ShlA and HpmA and some other outer-surface proteins [e.g. HMW1, HMW2 and HxuA of \textit{H. influenzae} (Barenkamp & Leininger, 1992; Cope et al., 1994)] share a conserved N-terminal domain important for export and activation. Using the N-terminal region of FHA in a secondary \textsc{blast} search against our database we identified a further four ORFs encoding FHA-like proteins (C, D, E and F at positions 749886, 5186294, 5082323 and 4561491 respectively; Fig. 7). Although all six \textit{P. aeruginosa} FHA-like proteins share low similarity with FHA across their entire sequence, they generally share a number of conserved features of FHA, including (i) large size (from 1000 to 5600 aa); (ii) repeat regions (from 400 to 3000 aa); (iii) homologous N-terminal \(\sim 250\) aa sequences, including a conserved activation/secretion motif, NPNGI, at \(\sim 180\) aa (Locht et al., 1993); and (iv) the presence of at least one CB3 integrin

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Fig6.png}
\caption{Electron micrograph of \textit{P. aeruginosa} showing peritrichously located thin fimbriae (indicated by arrows). Cells were negatively stained using uranyl acetate (1% uranyl acetate, 0.4% sucrose) and examined using a JEOL 1010 transmission electron microscope. Bar, 0.1 \(\mu\)m.}
\end{figure}
adhesion motif, RGD, indicating the potential for each novel FHA-like protein to bind eukaryotic cells (Relman et al., 1990; Fig. 7).

In view of their relatively high similarity with the N-terminal domain of FHA and related proteins we examined whether the *P. aeruginosa* FHA-like ORFs were also associated with sequences encoding accessory proteins. All *P. aeruginosa* FHA-like ORFs, except FHA-F, were associated with ORFs encoding FhaC homologues (the *Bordetella pertussis* accessory protein for FHA export and activation; see Gene ID 3200, 51, 903, 6012 and 5904 for FHA accessory proteins A–E, respectively). A multiple alignment of these potential accessory proteins (data not shown) indicated that, like FhaC, all contained repeating amphiphilic β-strands of approximately 10 aa, including a conserved phenylalanine residue at the C terminus (Willems et al., 1994). FHA-F is found within novel fimbrial locus B (see previous section).

Despite the similarity of *P. aeruginosa* FHA-like proteins to *Bordetella pertussis* FHA, there are significant differences, particularly in the nature of their repeat regions. *Bordetella pertussis* FHA contains a set of short repeats in the N-terminal domain prior to the RGD box (Locht et al., 1993; Fig. 7). *P. aeruginosa* FHA-like A, B, D, E and F also contain repeats which vary in length and position across the family. FHA-A and FHA-B are nearly 100% identical across much of their nucleotide sequence and the additional nucleotides of sequence present in FHA-A are largely due to a string of additional repeats near the N-terminal domain. FHA-E and FHA-F share ~40% amino acid identity overall but FHA-E has a unique N-terminal region of ten 18-aa repeats as well as patches of additional sequence in the C-terminal repeat region. The close similarity seen between FHA-A and FHA-B, and FHA-E and FHA-F, respectively, suggests that that each pair arose from duplication events (more recent in the case of FHA-A and FHA-B). The remaining two FHA-like proteins are remarkably different. FHA-C contains no obvious repeats whereas FHA-D contains a unique set of 13 near-identical 81-aa repeats and one partial repeat in its C-terminal domain. We consider it likely that the FHA-like ORFs in *P. aeruginosa* encode a group of large surface-expressed proteins that are exported in a manner similar to *Bordetella pertussis* FHA, are probably antigenic and may have a role in adhesion to eukaryotic cells. These are all experimentally testable hypotheses.

### Composite genetic elements

*E. coli* contains a class of complex genetic composites known as Rhs elements whose evolution has been studied extensively (Wang et al., 1998). In *E. coli* the Rhs element generally consists of (i) a 3.7 kb G + C-rich core encoding a large protein (composed of a core element and variable 100–200 aa C-terminal core extension) thought to associate with the cell surface, (ii) a small downstream ORF, (iii) a downstream insertion sequence and sometimes (iv) one or more ORFs at a 5' position (such as vgr). Vgr is a large protein of unknown function distinguished by 18–19 repetitions of a Val-Gly dipeptide occurring with an 8-residue periodicity (Wang et al.,...
1998). It has been speculated that, as both Vgr and the Rhs core ORF are hydrophilic and are characterized by a regularly repeated peptide motif, both ORFs may encode ligand-binding proteins found either on the bacterial cell wall or secreted into the medium (Hill et al., 1995). Although a precise role for these elements has yet to be determined, the degree of sequence conservation and their relative abundance (0.8% of E. coli K-12 genome) suggest that maintenance of Rhs elements provides some advantage to the host cell (Wang et al., 1998).

As far as we are aware, there has been no report of these elements in bacterial species other than E. coli, although their unusually high G+C content suggests that they may have originated elsewhere and been transferred into the E. coli genome (Wang et al., 1998). However, they do appear to be present in other unfinished bacterial genomes (see below).

During annotation of the P. aeruginosa genome we identified several ORFs encoding proteins with significant similarity to Vgr. Using BLAST we were able to identify 10 complete homologues of Vgr in the P. aeruginosa genome (Table 1). Using the graphical viewer of our genome database we were also able to identify adjacent ORFs that, together with these Vgr-like ORFs, might be considered composite genetic elements (Table 1). Interestingly, only one Vgr-like ORF was found to be associated with a complete Rhs element (which shares 32% identity to the E. coli Rhs core element and features a unique 149-aa core extension). The Rhs core ORF at Vgr locus I contains more than 20 repetitions of the motif YDxxGR, concentrated (and most periodic) between residues 400 and 800, reflecting the pattern observed in the E. coli core ORF (Wang et al., 1998). No additional Rhs core elements were identified in the genome, although the remnants of such a sequence exist at a position not linked with a Vgr ORF (~2758000). All but one of the remaining Vgr-like sequences were found to be upstream of ORFs encoding (generally large) proteins that, despite sharing no similarity, are nearly all hydrophilic proteins of unknown function and characterized by repeating peptide motifs (i.e. of a similar nature to the Vgr and the Rhs core ORF) (Table 1). Although no other downstream ORFs exhibit the periodicity of the Rhs core ORF at Vgr-I, we consider it likely that (at least) some of these ORFs may form a composite genetic element with their associated Vgr-like ORF. Interestingly, Vgr-A, Vgr-C and Vgr-D are each associated with downstream ORFs that are present in other unfinished genome sequences, notably those genomes that also encode homologues of Rhs core ORFs and Vgr (data not shown). Some composite genetic elements in E. coli contain a small ORF, encoding a product with high similarity to Vibrio cholerae haemolysin co-regulated protein of unknown function (hcp), upstream of vgr (Williams et al., 1996; Wang et al., 1998). There are three hcp homologues (58% identity with both E. coli and V. cholerae Hcp) in the P. aeruginosa genome which share almost identical nucleotide sequence and are found upstream of Vgr homologues (Table 1). Therefore, in P. aeruginosa, the Vgr-like ORFs appear to exist as part of larger composite elements, similar to the arrangement in E. coli.

The P. aeruginosa Vgr paralogues can be divided into three subgroups that each share amino acid identity of ~50% within subgroups and 20–35% between subgroups (Table 1). At least two pairs of Vgr paralogues have arisen from recent duplications judging from their high nucleotide similarity (A and B, and G and H). A comparison of the P. aeruginosa Vgr paralogues indicates that they contain a conserved core and a C-terminal extension of greater variability, highly reminiscent of the E. coli Rhs ORF structure (see above). In contrast, both copies of E. coli vgr are highly similar (Wang et al., 1998). The greater number of Vgr paralogues and their greater divergence relative to the E. coli Vgr paralogues suggests that Vgr is likely to have

### Table 1. Summary of Vgr loci in P. aeruginosa PAO1

<table>
<thead>
<tr>
<th>Vgr locus</th>
<th>Type*</th>
<th>Position (Mb)†</th>
<th>Gene ID</th>
<th>Associated features</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>I</td>
<td>1.63</td>
<td>1973</td>
<td>3’ ORF, 5’ hcp</td>
</tr>
<tr>
<td>B</td>
<td>I</td>
<td>2.84</td>
<td>343</td>
<td>3’ ORF, 5’ hcp</td>
</tr>
<tr>
<td>C</td>
<td>III</td>
<td>0.11</td>
<td>124</td>
<td>3’ ORF, locus D</td>
</tr>
<tr>
<td>D</td>
<td>III</td>
<td>0.11</td>
<td>118</td>
<td>3’ ORF, locus C</td>
</tr>
<tr>
<td>E</td>
<td>II</td>
<td>5.92</td>
<td>6849</td>
<td>3’ ORF, 5’ hcp</td>
</tr>
<tr>
<td>F</td>
<td>II</td>
<td>5.72</td>
<td>6623</td>
<td>3’ ORF</td>
</tr>
<tr>
<td>G</td>
<td>II</td>
<td>3.68</td>
<td>4296</td>
<td>3’ ORF</td>
</tr>
<tr>
<td>H</td>
<td>II</td>
<td>3.89</td>
<td>4537</td>
<td>3’ ORF</td>
</tr>
<tr>
<td>I</td>
<td>III</td>
<td>3.03</td>
<td>3494</td>
<td>3’ Rhs core with unique extension</td>
</tr>
<tr>
<td>J</td>
<td>(III)</td>
<td>2.62</td>
<td>3081</td>
<td></td>
</tr>
</tbody>
</table>

* Type is based on amino acid similarity. Only Vgr-J was not able to be classified easily; however, it is most similar to paralogues of type III.

† Sequence positions correspond to the 15 December 1999 sequence release of the P. aeruginosa PAO1 genome.

‡ ORF is in the opposite orientation to vgr.
existed in the *P. aeruginosa* genome earlier than *E. coli*. Evolutionary studies have established that each of the three *E. coli* *Rhs* element subfamilies moved independently from a G+C-rich background into an A+T-rich background (Wang *et al.*, 1998) foreshadowing the possibility that an *Rhs* element may have been originally transferred into an ancient *E. coli* genome from *P. aeruginosa* or similar species. However, homologues of Vgr and the *Rhs* core element can be found in several of the unfinished microbial genomes of GenBank (data not shown) and the evolutionary history and function of these composite genetic elements should become significantly clearer as more bacterial genomes are completed and comparative analyses are undertaken.

**Other features of the genome**

The previous sections outline a few of the findings which have arisen from our analysis of the *P. aeruginosa* genome using the interactive database. All of these reveal hitherto unsuspected aspects of *P. aeruginosa* biology and open up new avenues of experimental investigation. There are many more, for which space does not permit a full discussion, but which include the following:

- A 15 kb cluster of *fix*-like genes encoding a micro-oxic respiration system, similar to that used by nitrogen-fixing bacteria (Fischer, 1994), including two adjacent homologous operons encoding membrane-bound *cbb* type cytochrome oxidase complexes (*fixNOQP*), an adjacent operon encoding a membrane-bound cation pump complex (*fixGHIS*) and genes encoding homologues of nitrogen fixation/micro-oxic regulatory cascade proteins (*fixL, fixJ* and *fixK/anr*) (see ~1678000).
- An almost complete copy of bacteriophage *pf1* (see ~784500).
- A cluster of three genes encoding a homogentisate catabolism pathway that demonstrates remarkable (~50%) identity to homologues in both humans and fungi (Fernandez-Canon & Penalva, 1995) (see ~2186000).
- A number of novel colicins, bacteriocins and accessory proteins (see ~1066500, 3490500, 4634000 and 4245000).
- Six almost identical copies of a ~1300 nt insertion element encoding a 38 kDa transposon and conserved flanking regions (see ~500500, 2557000, 3044000, 3843000, 4473500 and 5382000).

It is already clear that the study of any bacterium is greatly assisted by the availability of the genome sequence, which *ipsa facto* reveals the full repertoire of its genes and its proteome, and which can vastly accelerate mutational and molecular genetic analyses. Conversely, it is also becoming apparent that the lack of a genome sequence is a great impediment to the experimental investigation of those bacteria for which the sequence is not available and renders such studies, by comparison, inefficient and incomplete, to say the least. This is amply illustrated by the case studies given in this paper. We hope that this database, integrating a variety of bioinformatic analyses and links, will be a useful platform for the ongoing genomic and molecular biological analysis of *P. aeruginosa*. It is anticipated that future versions of the database will be based upon the complete annotated *P. aeruginosa* genome when this becomes publicly available and will incorporate experimental information and linked clone resources, such as a minimal cosmid tiling path covering the genome (B. Huang and others, unpublished). The platform on which this database has been developed also makes it amenable to the generation of similar genome databases using other complete or near complete microbial genomes.

**NOTE ADDED IN PROOF**

The full *P. aeruginosa* PAO1 genome sequence has now been formally published (Stover *et al.*, 2000). Recently another two MCPs encoded by *ctpH* and *ctpL* have been identified which confer swimming chemotaxis to inorganic phosphate (Wu *et al.*, 2000).

**ACKNOWLEDGEMENTS**

L. Croft, S. A. Beatson and C. B. Whitchurch contributed equally to this work. We thank Francis Clark for recompilation of the BLAST source code to generate links in BLAST output. We thank James Lever for continued system support. We also thank Chris Ponting for helpful comments and Andrew Leech for electron microscopy. We thank the *Pseudomonas* Genome Project for making the *P. aeruginosa* PAO1 genome sequence publicly available. Sequencing of *P. putida* was accomplished with support from DOE and BMBF. This work was supported by a grant to J.S.M. by the Australian Research Council. The Centre for Functional and Applied Genomics is a Special Research Centre of Australian Research Council.

**REFERENCES**


Regulate methylation of taxis proteins.


Received 15 May 2000; revised 10 July 2000; accepted 13 July 2000.