Supporting Text

Genome Sequencing and Assembly. Initial shotgun libraries were generated and sequenced at the Broad by the Microbial Sequencing Center yielding 76,452 (PA2192) and 77,884 (C3719) sequences (paired-reads). The reads were assembled using ARACHNE (1, 2). After refinement, final assemblies contained 82 (PA2192) and 124 (C3719) contigs with a total sequence spanning single scaffolds of 6.83 Mb (PA2192) and 6.15 Mb (C3719). The average sequence coverage over all contigs was 7X for both assemblies. PA2192 contains 81 estimated gaps whereas C3719 contains 123 estimated gaps. Additional assembly statistics are shown in the table below.

Assembly analyses of the two genomes

	Pseudomonas aeruginosa PA2192	*Pseudomonas aeruginosa* C3719
Reads Assembled	76,452	77,884
Contigs	82	124
Contig N50	176,566	107,426
Largest Contig	398,738	242,903
Total Contig Length	6.83Mb	6.15Mb
Scaffolds	1	1
Scaffold N50	6.91Mb	6.22Mb
Largest Scaffold	6.91Mb	6.22Mb
Total Scaffold Length	6.91Mb	6.22Mb
Coverage	7x	7x
Gaps	81	123
%Q20	99.68%	99.50%
%Q40	98.52%	98.33%
Genome Status	Draft	Draft

Genome Assembly Annotation. The assembled draft genome sequences were processed through the Broad Institute annotation pipeline. Gene structures were assigned by the

automated gene calling algorithm that uses a combination of predictive (Glimmer (3, 4), GeneMark (5-7)), mapped PA01 ORFs from PseudoCAP ((8) and evidence-based features (BLAST (9)). We also used purely *ab initio* gene predictions with no BLAST evidence in cases where the predicted ORF size was at least 120 base pairs. No more than 200 base pair overlaps were allowed between adjacent genes. Minimal targeted manual editing was done to resolve major discrepancies between the automated gene calls and the BLAST evidence and to correct 3' and 5' ends. Further, we reviewed all intergenic regions longer than 1.0 kb in length containing any good BLAST evidence and manually created new ORFs if sufficient evidence was present. Problematic annotations containing recognizable sequence gaps, errors and frame shifts were flagged with appropriate curation flags. Locus IDs of the form PA2G_##### (PA2192) and PACG_##### (C3719) were assigned to provide unique identifiers for genes over different assemblies. Loci are simply identifiers and are not guaranteed to have any particular order or internal structure. Genomic sequences of the draft assembly and annotation used for *P. aeruginosa* strains PA2192 and C3719 in these comparative analyses are available at The Broad Institute website (10) and includes all data from the original NCBI Genbank submission (2006-02-04; PA2192, NZ_AAKW00000000; C3719, NZ_AAKV00000000). Sequences and annotations used for the other strains were downloaded from Genbank and imported into The Broad database for genomic comparisons (PACS2, NZ_AAQW01000001 (2006-08-10); PA01, NC_002516; and PA14, NC_008463(2006-10-20)).

Genes encoding tRNAs and rRNA operons were also predicted using tRNAscan-SE (11) and RFAM (12, 13) respectively and tracked separately from the final gene set.

Whole Genome Alignments and Ortholog Analysis. Pairwise syntenic blocks between strains were computed as follows. Local alignments were found between each pair using PatternHunter (14). Alignment blocks >10,000 bp on both sequences were merged to form collinear blocks. In cases where blocks overlapped to a high degree on either strain, the shorter block was discarded. Global alignments were computed over each block using ClustalW (15) for use in ortholog prediction.

Ortholog pairs were computed using two methods. In the first, the global alignments between syntenic blocks were used to map gene coordinates from each strain to every other strain. If the mapped gene coordinates overlapped a gene prediction on the target strain by at least 65% of the length of both genes, the two genes were identified as an ortholog pair. In the second method, the DNA sequence of each gene set was aligned to each other gene set using BLASTN. The alignments were filtered to require alignment length greater than 60% of both genes. Pairs of genes whose alignments met a reciprocal-best criterion were retained as predicted orthologs.

Pairwise ortholog predictions produced by both methods were used in computing ortholog clusters. For each gene in a given strain, we searched for a synteny-based ortholog on every other strain. If no synteny-based ortholog was found for a given target

strain, then we searched for a BLAST-based ortholog to that target strain. The resulting pairwise orthologs were clustered by single linkage.

Murasaki, a fast tool for finding locally similar regions across the global scope of the genomes (16) was used to view the five genomes. The five genomes are represented as scaled thick black lines. The positions of the rRNA operons are shown with bright green arrows. Before providing as input to Murasaki, the genomes were modified as follows: PAC3719 was reverse complemented and then transposed at 1.94 Mb; PA2192 was transposed at 4.37 Mb. The leftmost end of all the genomes corresponds to the origins of replication (*ori*). The regions around ori are well conserved across all genomes. Three of the five genomes show a large inversion. As shown in the image, PA2192, PAO1, and PACS2 have a large inversion with respect to the other two genomes, PA14 and PAC3719. The left endpoints of all the inversions coincide with the same rRNA operon, while the right endpoints occur at three different rRNA operons. The regions in the reference genome (PA2192) are given a color gradient from red through blue in order to visualize where the corresponding segments in the other genomes are occurring. Murasaki was run using 24-bit hash keys and a randomly chosen mismatch pattern (111010111100111001110101011101110110111011101011101110).

CGView (17), a java-based image generation tools was used to generate the Pangenome image (Fig. 1) and circular maps (SI Fig. 4). Input was a XML-formatted annotation generated from a Perl script that takes GFF-formatted input. The scale (in kb) corresponds to the coordinates on the PA14 genome, adjusted to accommodate RGPs absent from PA14. Note that the inversions in PA2192, PAO1 and PACS2 were

straightened out for Figure 1. High resolution figure can also be found at BioRG website (18)

Phylogenetic Analysis. To assess the genotypic diversity among the six *P. aeruginosa* strains PAO1, PA14, C3719, PA2192, PACS2, and LES analysis was performed using the maximum parsimony method (19) and rooted using *P. fluorescens* as the outgroup. To obtain a more reliable phylogeny, a total of 1,836 ORFs with orthologs in all seven genomes were identified, the sequences were aligned, concatenated and analyzed to obtain the phylogenetic tree (total length of about 2.12 Mb; (20)). Individual sequences for the ORFs were also used to generate 1,836 individual *gene trees*. These gene trees were analyzed to investigate which ones agreed with the concatenated phylogenetic tree described above.

The analyzed sequence had 480,000 variable sites with an average of about 10 parsimony-informative sites (PIS) per ORF. About 55% of the ORFs had at most six informative sites, while 92% of the ORFs had at most 16 informative sites. A list of twenty ORFs (*PA0155, PA0690, PA0692, PA0752, PA2363, PA2386, PA2393, PA2403, PA2424, PA2886, PA3923, PA4373, PA4503, PA4514, PA4526, PA4542,* and *PA5040*) with more than 60 informative sites is referred to as the MaxPIS list. Of 1836 individual gene trees, 204 showed no discrepancy with the concatenated-consensus phylogenetic tree. A list of 144 pyhlogentically useful genes (PUG) with at least three PIS (SI Table 2) has been identified. Among the ORFs commonly used to infer phylogeny (*recA, gyrB, oriC, citS, nucP, xdhB, bdhA* and *mtlD*) and for multilocus sequence typing (*acsB, aroE,*

guaA, mutL, nuoD, ppsA and *trpE*; 21), only *recA* and *guaA* were on the PUG list, suggesting that the other genes may not be as phylogenetically useful as previously thought.

1. Batzoglou S, *et al.* (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Res* **12,** 177-189.

2. Jaffe DB, *et al.* (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 13:91-96.

3. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27:**4636-4641.

4. Salzberg SL, Delcher AL, Kasif S, White O (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **26:**544-548.

5. Borodovsky M, *et al.* (1995) Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic acids research* **23,** 3554-3562.

6. Borodovsky M, Peresetsky A (1994) Deriving non-homogeneous DNA Markov chain models by cluster analysis algorithm minimizing multiple alignment entropy. *Comput Chem* **18:**259-267.

7. Borodovsky M, Rudd KE, Koonin EV (1994) Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res* **22:**4756-4767.

8. Stover CK, *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406:**959-964.

9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215:**403-410.

10. http://www.broad.mit.edu/annotation/genome/pseudomonas_group/MultiHome.html

11. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25:**955-964.

12. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. *Nucleic Acids Res* **31:**439-441.

13. Griffiths-Jones S, *et al.* (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33:**D121-D124.

14. Ma B, Tromp J, Li M (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18:**440-445.

15. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:**4673-4680.

16. Popendorf K (2007) Murasaki: language-theory based homology detection tool across multiple large scale genomes [http://murasaki.dna.bio.keio.ac.jp/].

17. Stothard P, Wishart DS (2005) Circular genome visualization and exploration using CGView. *Bioinformatics* **21:**537-539.

18. http://biorg.cs.fiu.edu/genomics/PA/supplemental/

19. Swofford DL (1998) *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)* (Sinauer, Sunderland, MA).

20. Rokas A, Williams BL, King N, Carroll, SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425:**798-804.

21. Curran B, Jonas D, Grundmann H, Pitt T, Dowson CG (2004) Development of a multilocus sequence typing scheme for the opportunistic pathogen *Pseudomonas aeruginosa*. *J Clin Microbiol* **42:**5644-5649.