DOI: 10.1111/J.1364-3703.2004.00263.X

# Lineage-specific regions in *Pseudomonas syringae* pv tomato DC3000

VINITA JOARDAR[1], MAGDALEN LINDEBERG[2], DAVID J. SCHNEIDER[3], ALAN COLLMER[2] AND C. ROBIN BUELL[1]*

[1] *The Institute for Genomic Research, 9712 Medical Center Dr, Rockville, MD 20850, USA*
[2] *Department of Plant Pathology, Cornell University, Ithaca, NY 14853, USA*
[3] *USDA/ARS, Ithaca, NY 14853, USA*

## SUMMARY

Comparative analyses of the chromosome of *Pseudomonas syringae* pv *tomato* DC3000 with the finished, complete genomes of *Pseudomonas aeruginosa* PAO1, an animal pathogen, and the non-pathogenic soil inhabitant *Pseudomonas putida* KT2440 revealed a high degree of sequence conservation in genes involved in 'housekeeping functions'. However, divergence is present among these three fluorescent pseudomonads, yielding 'suites' of species-specific genes that may provide the genetic basis for adaptation to an ecological niche and lifestyle. For DC3000, 1053 genes located on the chromosome were specific to DC3000 and not present in PAO1 or KT2440. The majority of these DC3000-specific genes either lack a known function or are mobile genetic elements. However, these genes do share features among themselves such as association with regions of atypical trinucleotides, unusual G+C content and localization within large tracts of DC3000-specific sequence, suggestive of lateral gene transfer events. Indeed, a comparison of syntenic blocks among these three complete *Pseudomonas* genomes revealed that a substantial portion (533) of the DC3000-specific chromosomal genes (1053) were located in lineage-specific regions (defined as being larger than 2 kb and enriched in mobile genetic elements and/or genes specific to DC3000 in this three-way comparison). A large proportion of mobile genetic elements (199 of 318 genes; 63%), which are highly enriched in DC3000, were present within such regions. Similarly, most of the genes encoding type III secretion system virulence effectors were located in lineage-specific regions. Consistent with the plasticity of the DC3000 genome, a putative chromosomal inversion mediated by identical copies of ISPsy6 involving 2838 kb (44%) of the DC3000 genome was detected. These data suggest that a substantial portion of the differentiation of DC3000, a plant pathogen, from an animal pathogen and a soil inhabitant has involved transfer of a large number of novel genes coupled with amplification of mobile genetic elements.

## INTRODUCTION

Pathovars of *Pseudomonas syringae* represent an important group of plant pathogenic bacteria, encompassing a collection of isolates that cause disease on a range of taxonomically diverse flowering plants. Characterization of these isolates through plant host specificity assays has resulted in the assignment of isolates to more than 50 pathovars, which can be classified into nine genomospecies by DNA–DNA hybridization (Gardan *et al*., 1999). An important attribute of the *P. syringae* pathovar group is the presence of a type III secretion system (TTSS), which injects effector proteins into host cells, thereby conditioning the outcome of the host–pathogen interaction (for reviews see Chang *et al*., 2004; Collmer *et al*., 2002; Greenberg and Vinatzer, 2003). One isolate of *P. syringae* that has been extensively studied is *P. syringae* pv *tomato* DC3000 (DC3000), which causes bacterial speck of tomato and can infect the model plant *Arabidopsis thaliana* (Whalen *et al*., 1991).

Genes associated with pathogenesis and other niche-specific functions are often laterally acquired and clustered by function in genetic modules (Hacker and Kaper, 2000). The *hrp* (*h*ypersensitive *r*esponse and *p*athogenicity) and *hrc* (*hrp c*onserved) genes, which encode the TTSS, are examples of such genes. In DC3000, the *hrp*/*hrc* genes are clustered at the centre of a tripartite pathogenicity island and are flanked by an exchangeable effector locus and a conserved effector locus, which harbour a subset of the TTSS effector genes (Alfano *et al*., 2000). The exchangeable effector locus is enriched in mobile genetic elements (MGEs) and is linked to a tRNA-Leu locus. Importantly, operons with *hrp*/*hrc* genes and almost all of the known *avr* (*avi*rulence) and *hop* (*H*rp *o*uter *p*rotein) effector genes are preceded by Hrp box promoters that are activated by the HrpL alternative sigma factor encoded within the *hrp* gene cluster. The effector genes have been comprehensively identified in the DC3000 genome through multiple bioinformatic and experimental approaches. Reflecting its status as the premiere model for plant–pathogen interactions, over 40 confirmed effector proteins have been identified in the DC3000 genome, the highest number for any bacterial pathogen of either animals or plants.

The ability of comparative genomics to identify open reading frames (ORFs) that are present in one species or isolate but

**Correspondence*: Tel.: +1 301 795 7558; fax: +1 301 838 0208; e-mail: rbuell@tigr.org

lacking in other closely related bacteria provides an efficient approach to exploring the genetics of specialized bacterial functions such as pathogenesis. Lineage-specific regions (LSRs) can be identified by comparison of complete genomes of different species belonging to the same genus or even strains within the same species (da Silva *et al.*, 2002; Perna *et al.*, 2001). Not only can sets of genes that are unique be identified, but the location and distribution of these genes among bacteria can also shed light on the origins and functions of these specific genes. For example, genomic islands (GIs) are large LSRs (typically 10–200 kb) that are present in one species and absent in closely related species (Bhattacharyya *et al.*, 2002; Perna *et al.*, 2001; da Silva *et al.*, 2002). GIs are typically associated with MGEs, have unusual G+C content/atypical nucleotide composition and an increased frequency of insertion into tRNA genes, suggesting lateral gene transfer events (Hacker and Carniel, 2001). Perhaps the most interesting subset of LSRs and GIs are the pathogenicity islands (PAIs), which contain genes involved in host–pathogen interactions (Hacker and Kaper, 2000).

The complete genome sequence has been published for three *Pseudomonas* species, the opportunistic human pathogen *Pseudomonas aeruginosa* PAO1 (PAO1; 6.2 Mb, 5565 ORFs, Stover *et al.*, 2000), the plant pathogen *P. syringae* pv *tomato* DC3000; 6.5 Mb, 5763 ORFs, Buell *et al.*, 2003) and the soil saprophyte *Pseudomonas putida* KT2440 (KT2440; 6.2 Mb, 5420 ORFs, Nelson *et al.*, 2002). The KT2440 and DC3000 genomes were annotated at The Institute of Genomic Research via manual curation of all the ORFs with the exception of hypothetical proteins that constitute 11% of both genomes (597 in KT2440; 629 in DC3000). Information on the gene naming conventions is available on the Comprehensive Microbial Resource website (http://www.tigr.org/CMR2/db_assignmentextver2.shtml). The PAO1 genome has been annotated by the *P. aeruginosa* community and can be obtained from PseudoCAP (http://www.pseudomonas.com). Thus, these three pseudomonads each have a high quality of annotation.

The availability of these sequences enables us to compare and contrast the genetic complement of DC3000, a sophisticated plant pathogen, with that of members of the same genus with distinct lifestyles. Specifically, the PAO1 and KT2440 genomes can be used as a comparative filter to identify candidate genes potentially associated with the specialized DC3000 phytopathogenic lifestyle. We previously reported that BLASTP analysis of all three *Pseudomonas* species revealed 3797 DC3000 ORFs that were conserved among all three species (Buell *et al.*, 2003), with 1159 ORFs specific to DC3000. The conserved ORFs include orthologues (reciprocal best hits) and paralogues (ORFs belonging to the same gene family that have arisen by gene duplication). In a three-way BLASTP comparison between these pseudomonads it was observed that, based on a percentage of total ORFs DC3000 contains the smallest core genome (ORFs shared by all species) and the largest fraction of genes specific to a species (data not shown). These DC3000-specific genes primarily encode ORFs of unknown function (conserved domain protein, conserved hypothetical protein, hypothetical protein and unknown function), MGEs and virulence factors. Here, we performed a more detailed analysis of the species-specific ORFs within the DC3000 chromosome with respect to their association with genomic features that may represent lateral gene transfer events in this isolate, such as atypical nucleotides, unusual G+C content, and location within LSRs. As the MGEs and TTSS effector genes represent a significant portion of the DC3000 genome and especially the species-specific portion of the DC3000 genome, we have performed a more detailed analysis of DC3000 MGEs and TTSS effector genes and their distribution within the genome.

Draft/unfinished sequence is available for several other *Pseudomonas* species, and it is likely that homologues of some DC3000-specific genes identified in the context of this study may be present in other *Pseudomonas* species. However, these sequences have not been included in this study owing to the limitations imposed by the incomplete nature of the available sequence and restrictions on whole genome analyses placed by the generators of these sequences. Thus, we have chosen to restrict our analyses to the three finished, annotated and published *Pseudomonas* genomes: DC3000, PAO1 and KT2440. As additional complete genomes become available, the presence and distribution of DC3000-specific genes can be refined, leading to additional clues regarding the origin and evolution and authoritative identification of LSRs. However, the catalogue of LSRs we present here will provide a useful guide for researchers seeking those regions of the DC3000 genome that are most likely involved in the special lifestyle of a plant pathogen.

## RESULTS AND DISCUSSION

The general features of the DC3000 genome have been reported previously (Buell *et al.*, 2003). The presence of two plasmids, pDC3000A and pDC3000B, in DC3000 further distinguishes the plant pathogen from PAO1 and KT2440, which do not possess any extrachromosomal elements. Although the plasmids do harbour virulence genes, they do not have a major role in the virulence of DC3000 (Buell *et al.*, 2003). Therefore, the DC3000 plasmids have not been included in this comparative study of the *Pseudomonas* genomes. Previous BLASTP analyses revealed 1159 ORFs in DC3000 not present in either PAO1 or KT2440 (Buell *et al.*, 2003). Of these, 1053 are present on the chromosome. The majority of the DC3000-specific chromosomal ORFs have no known function (770 ORFs, 73%; hypothetical protein, conserved hypothetical protein, conserved domain protein or unknown function). Hypothetical proteins are ORFs predicted by the gene finder GLIMMER 2.0 (Delcher *et al.*, 1999) that have no significant matches to proteins in the GenBank database or to domains in the Pfam or TIGRFAM databases (Bateman *et al.*,

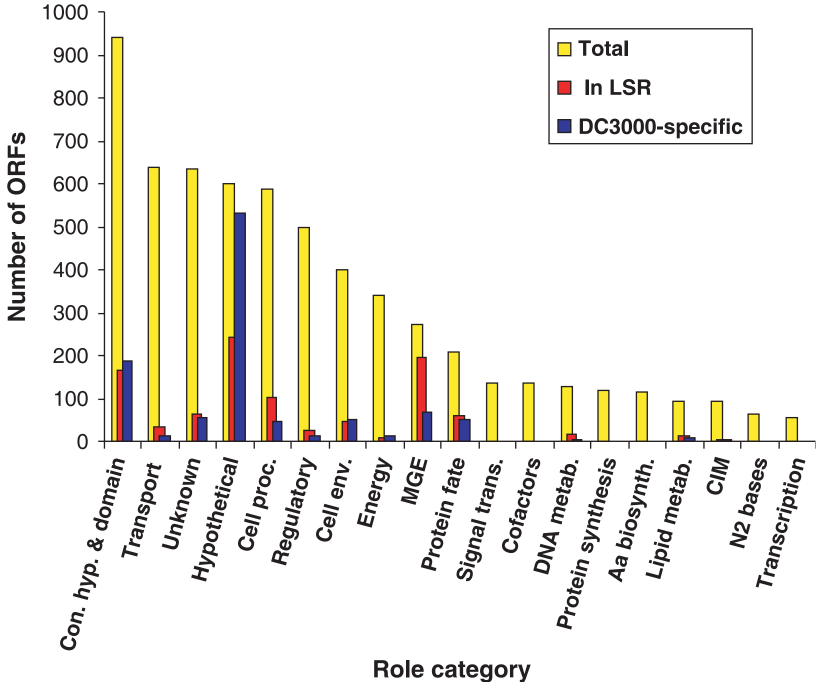

**Fig. 1** Functional classification of ORFs located in lineage-specific regions (LSRs) and ORFs specific to DC3000. The bars represent the total number of chromosomal DC3000 ORFs in the role category (yellow), the ORFs present in LSRs (red) and ORFs that are specific to DC3000 in comparison with *P. aeruginosa* PAO1 and *P. putida* KT2440 (blue). The ORFs were assigned to role categories adapted from Riley (1993) with only the broader categories represented here. Con hyp & domain: conserved hypothetical and conserved domain; Transport: transport and binding function; Unknown: unknown function; Hypothetical: hypothetical proteins; Cell proc.: cellular processes; Regulatory: regulatory functions; Cell env.: cellular envelope; Energy: energy metabolism; MGE: mobile and extrachromosomal element functions; Signal trans.: signal transduction; Cofactors: biosynthesis of cofactors. prosthetic groups and carriers; DNA metab.: DNA metabolism; Aa biosynth: amino acid biosynthesis; Lipid metab.: fatty acid and phospholipid metabolism; CIM: central intermediary metabolism; N2 bases: purines, pyrimidines, nucleosides and nucleotides.

2004; Haft *et al*., 2003). Conserved hypothetical genes have database matches to hypothetical proteins identified in other species whereas conserved domain proteins contain partial (domain) matches to database proteins in which the function of the domain is unknown. The genes that fall under the category of unknown function include enzymes of unknown specificity and other protein families with undefined function. The other main groups of DC3000-specific genes are virulence factors (88 ORFs), MGEs (69 ORFs) and cell envelope proteins (44 ORFs). A more detailed breakdown of the DC3000-specific ORFs and their functional role categories is given in Fig. 1.

## Features of lateral gene transfer and lineage-specific regions

Features such as unusual G+C content, atypical nucleotide composition, the presence of MGEs, and the localization of DC3000-specific genes in clusters and/or integration into tRNA genes are suggestive of lateral gene transfer events and were investigated further. The distribution of these chromosomal features and their associations are summarized in Table 1.

The mean G+C content for the ORFs in the DC3000 chromosome is 58.6% with a standard deviation (SD) of 4.2%. A total of 269 ORFs with a G+C composition below 50.2% (mean – 2SD; 260 ORFS) or above 67.0% (mean + 2SD; nine ORFs) were identified as having unusual G+C content (Table 1). The transposase, ISPsy5, Orf1 is the predominant MGE with low G+C content, accounting for 40 ORFs out of a total of 52 MGEs in this category. In addition, the majority of the virulence factors exhibiting atypical G+C content (15 of 20) are related to the TTSS.

Trinucleotide skew has been used to analyse regions of bacterial genomes that may have been acquired by horizontal transfer (Paulsen *et al*., 2003; Seshadri *et al*., 2004). Probability values for $\chi^2$ analysis to determine the atypical trinucleotide composition within a genome are based on the assumptions that the DNA composition is relatively uniform throughout the genome and that the trinucleotide composition is independent. High $\chi^2$ values indicate regions of the genome that appear unusual and require further investigation. The $\chi^2$ values for the DC3000 chromosome ranged from 23 to 3500 with a mean value of 174 (supplementary Table S1). Regions encoding rRNA tend to exhibit atypical trinucleotide content and the $\chi^2$ values observed for the rRNA-encoding regions on the DC3000 chromosome were between 163 and 1033 (mean $\chi^2$ value of 659 ± 257). All $\chi^2$ values above 1173 (mean + 2SD; 43 windows of 2000 bp) were identified as regions of atypical nucleotide composition based on the trinucleotide skew (supplementary Fig. S1). A total of 56 ORFs were either fully or partially contained within these regions (Table 1).

Random distribution of DC3000-specific ORFs would result in one ORF per every five ORFs or one DC3000-specific ORF every 6 kb. However, preliminary results clearly indicated a non-random distribution of DC3000-specific ORFs (Buell *et al*., 2003). Acquisition of new ORFs in a genome can occur through uptake of single genes or blocks of genes. Collinearity in the three *Pseudomonas* chromosomes was examined using the Position Effect algorithm (Carlton *et al*., 2002) and LSRs were identified in the

**Table 1** *P. syringae* pv. *tomato* DC3000 chromosomal ORFs and their association with features of the DC3000 chromosome that may represent lateral gene transfer.

| Features | No. of ORFs (% of total)[a] | No. of ORFs in LSRs (% of ORFs in LSRs)[b] | ORFs of unknown function [in LSRs][a] | ORFs specific to DC3000 [in LSRs][b] | Virulence-related ORFs [in LSRs][b] | ORFs with unusual G+C% [in LSRs][b] | Mobile genetic elements [in LSRs][b] | Atypical nucleotide composition[c] [in LSRs][b] |
|---|---|---|---|---|---|---|---|---|
| No. of ORFs | 5615<br>(100%) | 983<br>(100%) | 2188<br>— | 1053<br>— | 286<br>— | 269<br>— | 318<br>— | 56<br>— |
| No. of ORFs in LSRs | 983<br>(17.5%) | | 473<br>— | 533<br>— | 112<br>— | 135<br>— | 199<br>— | 37<br>— |
| ORFs of unknown function | 2188<br>(39.0%) | 473<br>(48.1%) | | 770<br>[357] | 229[d]<br>[96] | 150<br>[72] | 593[e]<br>[325] | 33<br>[20] |
| ORFs specific to DC3000 | 1053<br>(18.8%) | 533<br>(54.2%) | 770<br>[357] | | 88<br>[73] | 146<br>[81] | 69<br>[48] | 25<br>[18] |
| Virulence-related ORFs | 286<br>(5.1%) | 112<br>(11.4%) | 229[d]<br>[96] | 88<br>[73] | | 20<br>[17] | 115[e]<br>[89] | 1<br>[1] |
| ORFs with unusual G+C% | 269<br>(4.8%) | 135<br>(13.7%) | 150<br>[72] | 146<br>[81] | 20<br>[17] | | 52<br>[30] | 37<br>[25] |
| Mobile genetic elements | 318<br>(5.7%) | 199<br>(20.2%) | 593[e]<br>[325] | 69<br>[48] | 115[e]<br>[89] | 52<br>[30] | | 12<br>[9] |
| Atypical nucleotide composition[c] | 56<br>(1.0%) | 37<br>(3.8%) | 33<br>[20] | 25<br>[18] | 1<br>[1] | 37<br>[25] | 12<br>[9] | |

[a] Percentages are not expected to add up to 100% due to overlaps between the classes of ORFs and exclusion of certain classes of ORFs in the table such as ORFs conserved among the species that do not fall into any of the classes shown.

[b] Subset of ORFs present in LSRs.

[c] ORFs that are wholly contained within, or partially overlap regions of atypical nucleotide composition.

[d] ORFs of unknown function were considered to be associated with virulence factors if the ORFs/cluster of ORFs was immediately adjacent to a virulence factor.

[e] ORFs of unknown function and virulence factors were considered to be associated with MGEs if there were not more than two ORFs between the MGE and ORFs/cluster of ORFS in these categories.

DC3000 chromosome. A total of 44 LSRs representing ~1.1 Mb (17.5%) were documented in the DC3000 chromosome (Table 2; supplementary Tables S2 and S3). These include 34 LSRs that are larger than 10 kb. The LSRs have been classified based on their putative function. The ten LSRs that are smaller than 10 kb are associated with virulence or fitness. LSRs may have either been acquired by DC3000 after speciation or been deleted from the reference genomes, PAO1 and KT2440. Chromosomal rearrangements can preclude the recognition of the sites of insertion/deletion in the reference/DC3000 genome. However, in 28 of the 44 LSRs the putative sites of insertion/deletion have been identified (supplementary Table S3) by examination of gene synteny in regions flanking the LSR. The putative insertion/deletion site of virulence-related LSR6 is illustrated in Fig. 2. The ORFs in KT2440 and PAO1 that border putative insertion/deletion sites in the reference genomes are highlighted in yellow in supplementary Table S3.

A total of 983 ORFs are present in the LSRs, of which 533 (54%) are specific to DC3000 (Table 1, supplementary Tables S2 and S3). A substantial percentage of ORFs with atypical nucleotides (66%, 37 of 56 ORFs) and/or with an unusual G+C content (50%, 135 of 269 ORFs) are located within LSRs. The ORFs located within LSRs were not evenly distributed throughout the functional role categories (Fig. 1). In particular, a total of 473 ORFs within LSRs have no known function (244 hypothetical proteins, 166 conserved hypothetical proteins and 63 proteins of unknown function). Other large classes of genes within the LSRs are MGEs (199 total, 48 DC3000-specific) and virulence-related ORFs (112 total, 73 specific). Apart from the TTSS-related LSRs, there are LSRs that are phage-related, involved in toxin synthesis, and implicated in adaptation to the environment; these are discussed in detail below. Four LSRs identified in DC3000 have also been listed as islands integrated at tRNA sites in the islander database (Mantri and Williams, 2004; http://129.79.232.60/cgi-bin/islander/islander.cgi; integrative islands Psy8F, Psy10R, Psy11I and Psy14S). Two of these are virulence-type LSRs (LSR7 and LSR41, corresponding to Psy11I and Psy8F, respectively) and one is a phage LSR (LSR26, corresponding to Psy14S). However, the fourth integrative island (LSR3, corresponding to Psy10R) is of unknown function. In addition, seven LSRs, all of which are larger than 10 kb, comprise genes of unknown function.

Some regions of the DC3000 genome are enriched in LSRs and may be the preferred sites for laterally transferred genetic material. LSR5–LSR9 are present in a section of the DC3000 genome that shows substantial collinearity with the KT2440 genome (PP0353–PP0597; coordinates 429 668–697 071). Similarly, LSR31–LSR33 are clustered in a region of the DC3000 genome

**Table 2** Lineage-specific regions in the DC3000 chromosome.

| LSR no. | Type of LSR[a] | Length (bp) | Start locus PSPTO# | End locus PSPTO# | Total ORFs | DC3000-specific ORFs | Un-known function | Tn[b]-related | Other MGE | Viru-lence | Atyp.nucl.[c] | Unusual G+C% | Disrupted reading frame | tRNA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | virulence | 68 852 | 0009 | 0061 | 53 | 36 | 33 | 14 | 0 | 2 | 0 | 8 | 3 | |
| 2 | fitness | 21 632 | 0189 | 0208 | 20 | 7 | 8 | 8 | 0 | 0 | 6 | 9 | 7 | |
| 3 | unknown | 22 819 | 0274 | 0298 | 25 | 17 | 17 | 2 | 2 | 0 | 0 | 0 | 1 | Arg-1 |
| 4 | virulence | 9 625 | 0368 | 0374 | 7 | 3 | 3 | 1 | 0 | 1 | 0 | 0 | 0 | |
| 5 | virulence | 7 244 | 0473 | 0479 | 7 | 6 | 6 | 0 | 0 | 1 | 0 | 1 | 0 | |
| 6 | virulence | 3 304 | 0501 | 0504 | 4 | 3 | 0 | 1 | 0 | 3 | 0 | 3 | 1 | |
| 7 | virulence/secretion | 17 786 | 0520 | 0535 | 16 | 11 | 8 | 1 | 1 | 2 | 0 | 0 | 1 | Met-1 |
| 8 | phage/virulence | 23 610 | 0570 | 0591 | 22 | 10 | 9 | 0 | 7 | 2 | 0 | 4 | 0 | |
| 9 | virulence | 15 055 | 0702 | 0719 | 18 | 12 | 13 | 0 | 0 | 1 | 0 | 2 | 0 | |
| 10 | virulence/fitness | 98 941 | 0830 | 0916 | 87 | 48 | 35 | 12 | 1 | 9 | 0 | 12 | 7 | Pro-2 |
| 11 | fitness/resistance | 14 877 | 0930 | 0946 | 17 | 13 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 12 | fitness | 63 223 | 1055 | 1100 | 46 | 22 | 14 | 5 | 1 | 0 | 5 | 5 | 2 | Met-2 |
| 13 | virulence | 47 839 | 1367 | 1411 | 45 | 28 | 4 | 1 | 0 | 37 | 0 | 2 | 1 | Leu-1 |
| 14 | virulence | 8 111 | 1451 | 1455 | 5 | 3 | 4 | 0 | 0 | 1 | 0 | 0 | 0 | |
| 15 | virulence | 10 063 | 1566 | 1577 | 12 | 9 | 9 | 1 | 0 | 1 | 0 | 3 | 0 | |
| 16 | unknown | 15 973 | 1651 | 1662 | 12 | 9 | 9 | 2 | 0 | 0 | 0 | 0 | 0 | |
| 17 | fitness | 2 797 | 1929 | 1932 | 4 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | |
| 18 | unknown | 11 117 | 2006 | 2013 | 8 | 3 | 5 | 2 | 0 | 0 | 0 | 0 | 0 | |
| 19 | phage | 19 701 | 2067 | 2096 | 30 | 24 | 25 | 0 | 3 | 0 | 0 | 3 | 0 | Ser-2 |
| 20 | regulation | 17 613 | 2318 | 2330 | 13 | 5 | 4 | 4 | 1 | 0 | 2 | 4 | 2 | |
| 21 | virulence/resistance | 60 542 | 2809 | 2842 | 34 | 15 | 14 | 5 | 0 | 3 | 0 | 2 | 2 | |
| 22 | virulence | 38 730 | 2866 | 2899 | 34 | 18 | 22 | 3 | 0 | 1 | 0 | 5 | 3 | |
| 23 | virulence | 48 457 | 3208 | 3230 | 23 | 8 | 8 | 11 | 0 | 4 | 1 | 5 | 2 | |
| 24 | phage | 44 681 | 3384 | 3434 | 51 | 15 | 26 | 4 | 16 | 0 | 0 | 2 | 2 | |
| 25 | fitness | 26 559 | 3478 | 3499 | 22 | 11 | 8 | 0 | 0 | 0 | 0 | 2 | 0 | |
| 26 | phage/fitness | 18 410 | 3925 | 3949 | 25 | 11 | 15 | 1 | 5 | 0 | 4 | 5 | 5 | Ser-4 |
| 27 | virulence | 7 682 | 3960 | 3967 | 8 | 4 | 3 | 4 | 0 | 1 | 0 | 2 | 2 | Lys-2 |
| 28 | phage/virulence | 21 213 | 3995 | 4017 | 23 | 15 | 11 | 7 | 3 | 1 | 1 | 4 | 2 | |
| 29 | phage | 20 264 | 4035 | 4057 | 23 | 16 | 19 | 0 | 4 | 0 | 0 | 1 | 0 | |
| 30 | virulence | 5 848 | 4184 | 4190 | 7 | 7 | 5 | 0 | 0 | 1 | 0 | 5 | 0 | |
| 31 | virulence | 10 586 | 4321 | 4332 | 12 | 10 | 9 | 2 | 0 | 1 | 0 | 0 | 0 | |
| 32 | toxin | 21 113 | 4340 | 4348 | 9 | 4 | 3 | 1 | 0 | 0 | 0 | 2 | 1 | |
| 33 | unknown | 10 940 | 4384 | 4391 | 8 | 4 | 5 | 3 | 0 | 0 | 0 | 1 | 1 | |
| 34 | virulence | 7 511 | 4588 | 4599 | 12 | 11 | 3 | 1 | 0 | 8 | 0 | 3 | 0 | |
| 35 | unknown | 33 168 | 4603 | 4630 | 28 | 16 | 14 | 2 | 3 | 0 | 0 | 4 | 0 | |
| 36 | virulence | 139 720 | 4670 | 4772 | 103 | 49 | 36 | 25 | 4 | 29 | 4 | 20 | 16 | |
| 37 | virulence | 8 268 | 4993 | 4996 | 4 | 1 | 0 | 2 | 0 | 1 | 0 | 1 | 0 | |
| 38 | fitness | 17 842 | 5088 | 5108 | 21 | 10 | 11 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 39 | unknown | 20 234 | 5199 | 5215 | 17 | 7 | 12 | 4 | 0 | 0 | 7 | 8 | 0 | |
| 40 | fitness | 4 187 | 5232 | 5237 | 6 | 4 | 3 | 0 | 0 | 0 | 0 | 1 | 0 | |
| 41 | virulence | 13 168 | 5344 | 5355 | 12 | 9 | 7 | 0 | 1 | 2 | 0 | 0 | 1 | Phe-1 |
| 42 | regulation | 16 075 | 5362 | 5377 | 16 | 6 | 6 | 7 | 0 | 0 | 7 | 4 | 3 | Thr-2 |
| 43 | unknown | 18 447 | 5437 | 5450 | 14 | 6 | 10 | 4 | 0 | 0 | 0 | 1 | 2 | |
| 44 | fitness | 22 209 | 5564 | 5583 | 20 | 6 | 5 | 6 | 0 | 0 | 0 | 1 | 4 | |
| Total | | 1118 250 | | | 983 | 533 | 473 | 147 | 52 | 112 | 37 | 135 | 71 | |

[a] LSRs were classified based on the functional categories of the ORFs present in the LSR. Virulence-type LSRs contained at least one ORF involved in the virulence of DC3000. Phage-type LSRs contained regions of the genomes that have been annotated as prophages or bacteriocins (Buell *et al.*, 2003). Fitness-type LSRs contain ORFs that allow DC3000 to adapt to its ecological niche. The regulation-type LSRs contain at least one ORF involved in regulation. The toxin-type LSR contains ORFs that are involved in synthesis of an insecticidal toxin. LSRs can be classified under more than one functional category. The remainder of the LSRs are classified as unknown.

[b] Transposon.

[c] Atypical trinucleotide composition.

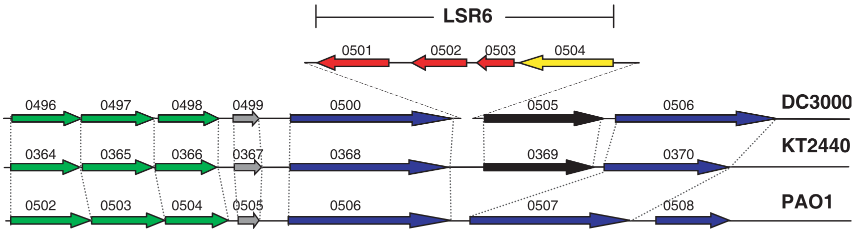

**Fig. 2** Insertion/deletion site of LSR6 in DC3000. The region of the DC3000 genome flanking LSR6 is aligned with the corresponding regions in the reference genomes, KT2440 and PAO1. The solid black lines represent the genomes and filled arrows represent ORFs with the arrowheads indicating direction of transcription. The locus numbers are shown above the ORFs, which are colour coded by role category (biosynthesis of cofactors, prosthetic groups and carriers: green; conserved hypothetical protein: grey; fatty acid and phospholipid metabolism: blue; unknown function: black; cellular processes: red; mobile genetic elements: yellow). The dashed black lines in the DC3000 genome represent the site of insertion/deletion of LSR6. Reciprocal best hits are indicated by dotted grey lines.

(PSPTO4314–PSPTO4478; coordinates 4864 702–5042 784) syntenic with the PAO1 genome (PA4314–PA4486; coordinates 4842 700–5017 834).

## Lineage-specific regions: virulence

PAIs are LSRs that are usually associated with MGEs and contain multiple virulence-related ORFs that are not present in non-pathogenic species (Groisman and Ochman, 1996). In addition, PAIs may carry complete genetic modules for a virulence-related function. For example, the TTSS PAIs of *P. syringae* and enteropathogenic *E. coli* can confer to non-pathogens the ability to inject effector proteins into host cells (Huang *et al*., 1988; McDaniel and Kaper, 1997). Two large PAIs, the Hrp PAI (LSR13) described above, and the bipartite cluster encoding genes for the synthesis of the phytotoxin coronatine (LSR36), have been extensively described (Alfano *et al*., 2000; Buell *et al*., 2003), and they will not be discussed further here. There are 20 other LSRs that presently can be assigned to the virulence category (Table 2), but it is important to note that this number may increase as new bioassays improve our ability to detect subtle virulence phenotypes. The largest of these known virulence-related LSRs is LSR10, which is ~99 kb in length. Virulence determinants present in LSR10 include five confirmed TTSS effectors. LSR10 is bounded by tRNA-Pro-2 with a truncated site-specific recombinase (PSPTO0830) at one end and a cluster of chemotaxis-related genes at the other end. LSR10 has a high density of MGEs, which is reflected in seven ORFs with disrupted reading frames.

LSR10 is located between ClpB (PSPTO0829) and NADH dehydrogenase (PSPTO0917) genes. Homologues of these two genes are contiguous in *P. aeruginosa* PAO1; however, an analogous PAI, PAPI-1, is present in *P. aeruginosa* strain UCBPP-PA14 (accession number AY273869). Although similarly integrated at a tRNA site (tRNA-Lys) downstream of *clpB*, PAPI-1 (111 332 bp) is larger than LSR10 (98 940 bp), and of the 25 ORFs with putative homologues in LSR10, none is a known virulence factor in DC3000. However, mutations in three of the PAPI-1 genes (RL003, RL016 and RL022), which have closest homologues in DC3000 (PSPTO0831, 0848 and 0859, respectively) resulted in reduced virulence of PA14 on animal and plant hosts (He *et al*., 2004). It is interesting that PAPI-1 does not contain the DC3000 fitness-related genes present in LSR10 (discussed below), consistent with the mosaic nature of the PA14 pathogenicity island. Although we have limited our systematic analysis to a comparison of strains DC3000, PAO1 and K2440, whose complete genomes have been published, this example involving strain UCBPP-PA14 further illustrates the power of comparative genomics to reveal potential hot spots for acquisition of virulence-related genes in pseudomonad genomes.

Four other virulence LSRs are associated with tRNA genes. These integrative LSRs are LSR7 (tRNA-Met1), LSR41 (tRNA-Phe-1), LSR13 (tRNA-Leu1, the Hrp PAI) and LSR27 (tRNA-Lys2). Although DC3000 does not appear to have a complete, functional type IV secretion system (Buell *et al*., 2003), two putative type IV secretion genes are present in LSR7 with one being truncated. A homologue of PSPTO0524 (peptidase, M20/M25/M40 family) in LSR7 is present on the plasmid pPSR1 of *P. syringae* pv. *syringae* A2 (Sundin *et al*., 2004). The TTSS effector HopPsyA$_{Pto}$ and its chaperone, ShcA$_{Pto}$, are located in LSR41. Three TTSS genes (PSPTO0501–0503), two confirmed effector genes (*hopPtoS2* and *hopPtoF*) and a TTSS chaperone gene (*shcF*) are present only in DC3000 and are located adjacent to a degenerate copy of the transposase, ISPsy5 (LSR6, Fig. 2).

The distribution of experimentally confirmed TTSS effector genes and *hrp* box regulatory elements relative to the locations of the LSRs is shown in Fig. 3. Disproportionate numbers of effectors and *hrp* boxes result from the fact that not all candidate *hrp* boxes have been confirmed, and that many of those confirmed regulate expression of structural elements of the TTSS secretion apparatus. In DC3000, 83% of effectors and 85% of *hrp* promoters were found in association with LSRs, supporting the potential usefulness of LSRs for identification of additional virulence factors. Genes encoding confirmed effectors that were not assigned to LSRs were not clustered with other DC3000-specific genes and

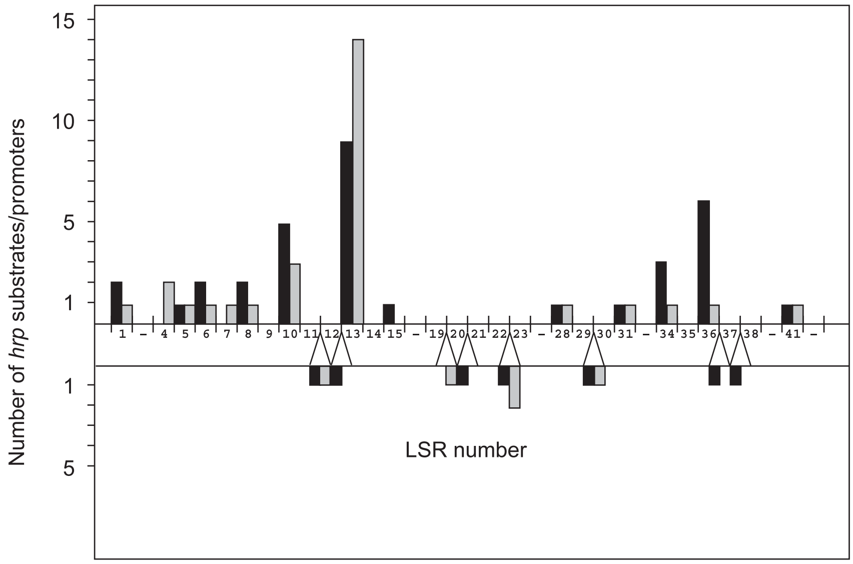

**Fig. 3** Distribution of experimentally confirmed Hrp pathway substrates and *hrp* promoters within and between LSRs. The numbers of Hrp pathway substrates (dark) and *hrp* promoters (light) found within LSRs are indicated in the top part of the figure, and those falling between the LSRs are shown below. LSRs are indicated by number and are shown in the order they appear in the genome. Hyphens indicate regions where two or more LSR numbers have been omitted because they did not contain substrates or *hrp* promoters.

therefore failed to meet the criteria for inclusion. Figure 3 reveals that the majority of TTSS effector genes were found in association with LSRs, and moreover, the majority of these were linked with other effector genes. As more complete *P. syringae* genomes become available, it will be interesting to determine if certain combinations of effector genes in LSRs are conserved, which might suggest a co-ordinated function for these effectors.

Variable segments have been identified by Wolfgang *et al.* (2003) in the *P. aeruginosa* PAO1 genome by comparison with other *P. aeruginosa* strains. It is interesting to note that some ORFs that flank the variable segments in PAO1 are also found in similar positions in the DC3000 genome, suggesting that certain regions of the core *Pseudomonas* genome are preferred sites for horizontally acquired genes. For example, PSPTO3229 and PSPTO3230 (annotated as a putative filamentous haemagglutinin, intein-containing and haemolysin activator protein; HlyB family protein, respectively) are at one end of LSR23. The orthologues of these two DC3000 virulence genes in *P. aeruginosa* PAO1, PA2462 and PA2463 are associated with variable segment 15 in PAO1.

## Lineage-specific regions: phage

The six phage-related LSRs in DC3000 (Table 2) include a pyocin LSR (LSR8), which is sandwiched between *trpE* (PSPTO0568) and *trpG* (PSPTO0592), anthranilate synthase components I and II, respectively. Apart from the bacteriocin (PSPTO0572), LSR8 also encodes two TTSS effectors, located just outside the pyocin, that are specific to DC3000. However, most of the ORFs in this phage LSR are not DC3000-specific, and the majority of these have putative homologues in KT2440. Although a phage is also present in PAO1 between *trpE* and *trpG* (PA0609 and PA0649), many of the ORFs in this PAO1 phage have putative homologues in another phage-type LSR (LSR24). Phage LSR28 harbours the TTSS effector AvrPto$_{DC3000}$, which is clustered with DC3000-specific genes and IS elements. Apart from virulence, phage LSRs also contain ORFs that could contribute to the fitness of DC3000. Two proteins, a cold shock domain family protein and a putative gas vesicle protein that could function in adaptation to atypical conditions, are found in the phage LSR26. Similar to the example in the section above, PSPTO3994 is adjacent to the phage LSR28 and the orthologue in PAO1, PA1940, lies just outside variable segment 10 (Wolfgang *et al.*, 2003).

## Lineage-specific regions: fitness

LSRs that contribute to increased bacterial fitness are referred to as 'fitness islands' (Hacker and Carniel, 2001). In this analysis, the putative fitness LSRs encompass a subset of LSRs that are likely to facilitate survival of DC3000 and/or adaptation to its ecological niches but are not directly responsible for pathogenicity. Two DC3000-specific genes in LSR10 are predicted to be involved in the uptake and utilization of sucrose, a carbon source found abundantly in the plant apoplast (Gottwald *et al.*, 2000), which is the site of pathogen growth and multiplication. PSPTO0890 is the sucrose porin precursor, ScrY, which transports sucrose into the cell. Sucrose, which is phosphorylated via the phopshoenolpyruvate-mediated phosphotransferase system, can then be metabolized by sucrose-6-phosphate hydrolase (ScrB; PSPTO0885) to yield fructose and glucose. A transcriptional regulator of the *Lac*I family (PSPTO0884) is adjacent to ScrB and is similar to the sucrose operon repressor of *Yersinia pestis* CO92 (YPO1642; accession number CAC90464). PAO1 and KT2440 do not have the ability to transport or to metabolize sucrose. Although PAPI-1 in *P. aeruginosa* UCBPP-PA14 is a PAI similar to LSR10, it does not have homologues of the sucrose utilization genes.

LSR17 is inserted into a flagella/chemotaxis gene cluster in DC3000 and contains an IS element (ISPsy6) and genes encoding cyanate lyase and two proteins of unknown function. Some bacteria are capable of overcoming the toxicity of cyanate in the environment via hydrolysis. Sodium cyanate is used as a herbicide and is thought to exert its toxicity by causing oxidative stress. The enzyme cyanate lyase (EC 4.2.1.104; cyanase, CynS), which is also found in other bacteria and plants, catalyses the reaction of cyanate with bicarbonate to produce ammonia and carbon dioxide. In *E. coli*, cyanate lyase functions in detoxification and in the utilization of cyanate as the sole source of nitrogen (Sung *et al.*, 1987), with a requirement for bicarbonate as a substrate (Johnson and Anderson, 1987). As DC3000 contains a cyanate MFS transporter (PSPTO1260) and the orthologue of the carbonic anhydrase, CynT (PSPTO5255), which regenerates the bicarbonate substrate required by CynS, cyanate lyase may function similarly in DC3000. The ability to exploit exogenous cyanate as a nitrogen source and/or to resist cyanate toxicity could contribute significantly to the survival of DC3000 in the soil and in cyanogenic plants.

Resistance to antimicrobial agents confers an ecological advantage by enhancing survival of pathogens in hostile environments. LSR11 and LSR21 are involved in such resistance. Genes directing resistance to the toxic metal anion tellurite (PSPTO0940-0946) are present in LSR11 along with genes of unknown function, and are similar to the *ter*ZABCDEF cluster found in two genomic islands in *E. coli* O157:H7 strains (Taylor *et al.*, 2002) and enterobacterial IncHI2 and IncHII plasmids (Taylor, 1999). Although it is known that tellurite is reduced to metallic tellurium, the functions of the *ter* genes, and the mechanism by which they mediate tellurite resistance, remain unknown. It has been proposed that the *ter* genes function primarily in survival of the bacterium and that detoxification of tellurite is a secondary function (Taylor, 1999). The macrolide efflux proteins, MacA and MacB (PSPTO2831 and 2832), and β-lactamase (PSPTO2834) in LSR21 have putative homologues but do not have orthologues in PAO1 and KT2440. These proteins, which are involved in resistance to antibiotics, are clustered with IS elements, two non-ribosomal peptide synthases and other DC3000-specific ORFs in the resistance LSR. In addition, a homologue of an auxin-responsive GH3-related protein is also present in LSR21.

LSR2 contains a putative nitrilase (PSPTO0189) and a truncated version of phenylacetaldoxime dehydratase (PSPTO0190), which metabolizes phenylacetaldoxime to phenylacetonitrile. Phenylacetaldoxime is an intermediate in the biosynthesis of benzylglucosinolate, which belongs to a class of compounds known as glucosinolates—natural plant products that function in defence against plant pests, and are of interest as cancer-preventing agents (Wittstock and Halkier, 2000). A functional version of phenylacetaldoxime dehydratase could potentially decrease the synthesis of glucosinolates by diverting intermediates. Nitrilases can use phenylacetonitrile and other nitriles (organic cyanides; RCN) as a source of carbon and nitrogen by converting the nitriles to the corresponding carboxylic acids. Also present in LSR2 is the cysteine synthase encoded by PSPTO0203, which catalyses the synthesis of cysteine from serine. There are two other cysteine synthases in the DC3000 genome, *cysK* and *cysM* (PSPTO3902 and PSPTO1692), which have orthologues in PAO1 and KT2440. However, PSPTO0203 is most similar to a cysteine synthase present on the pCC7120beta plasmid of *Nostoc* sp. 7120 (gene alr7586; accession number BAB77229).

The potential of DC3000 to move towards or away from a remarkable array of chemical attractants and repellants is supported by the presence of ~60 genes involved in chemotaxis. Methyl-accepting chemotaxis proteins (MCPs; chemoreceptors that respond to stimuli in the environment) and other components of the chemotactic machinery are present in seven genomic LSRs. A putative operon of chemotaxis genes in LSR10 (PSPTO0908–0916) includes *cheB* (methylesterase), *cheD* (stimulates methylation of MCP), *cheR* (methyl transferase), *cheW* (receptor coupling protein), *cheA* (sensor histidine kinase), *cheY* (signal transduction protein), two MCP genes and an STAS domain protein gene of unknown function. All the genes in this cluster have paralogues in DC3000 except for *cheD*. LSR12 is the only other LSR that contains multiple *mcp* genes (three) whereas five other LSRs each have a single MCP component.

## Lineage-specific regions: extrachromosomal and mobile genetic elements

Genetic material acquired via lateral gene transfer not only paves the way for rapid adaptation of the organism to its environment but can also serve as a mechanism for genome expansion. The DC3000 genome contains a large complement of proteins involved in transposition and proteins that are phage- or plasmid-related (382 ORFs, 6.6%; Buell *et al.*, 2003), of which 318 ORFs are on the chromosome. Within the mobile and extrachromosomal element functional category, IS elements account for 210 of the total number of ORFs on the chromosome (Buell *et al.*, 2003), with 55 specific to DC3000. The most abundant DC3000-specific IS element is the ISPsy4 transposase and its transposition helper protein (43 ORFs; 21 complete copies and one incomplete copy). ISPsy4, an IS21 family element, is most closely related to transposases found in *Y. pestis* strains CO92 (transposase for insertion sequence IS100, *ypmt1*/*y1093*) and KIM. In DC3000, most of the ISPsy4 elements are located adjacent to genes specific to the pathogen and/or in clusters of IS elements, with 17 copies present in 14 LSRs. ISPs1 is also found only in DC3000, with three incomplete copies and one full-length copy that is clustered with the TTSS effector gene *hopPtoE* and species-specific genes of unknown function in LSR31. Although the OrfA components of ISPsy8 and ISPsy9 are present only in DC3000, the OrfB components have

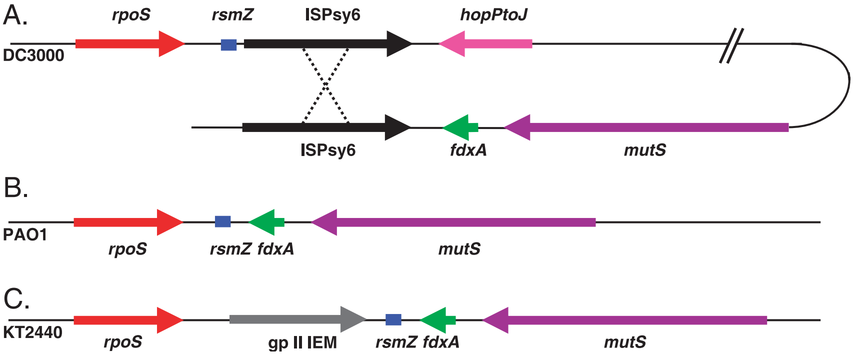

**Fig. 4** Proposed chromosomal inversion in DC3000. The *rpoS-mutS* regions of the three *Pseudomonas* chromosomes, A. DC3000, B. PAO1 and C. KT2440. Homologous recombination between two identical copies of the mobile genetic element ISPsy6 may have resulted in an inversion in the DC3000 genome. The thin black lines represent the genomes and filled arrows represent genes, with the arrowheads indicating direction of transcription [*rpoS*, red; *fdxA*, green; *hopPtoJ*, pink; *mutS*, purple; ISPsy6, black; gp II IEM (group II intron encoding maturase), grey]. The regulatory RNA *PsrsmZ* is shown as a filled blue rectangle. The dashed lines in A show the potential site of homologous recombination in DC3000 and the parallel lines (bold) represent the block of the DC3000 genome in between the two copies of ISPsy6.

putative homologues in PAO1 and KT2440. Neither of these IS elements is associated with LSRs. The ISPsy14 transposase and its transposition helper protein are not present in PAO1 or KT2440 but have been reported in other pseudomonads (Yeo *et al*., 1998). In DC3000, the single copy of ISPsy14 is found in LSR21.

### Chromosomal inversion in DC3000

One impact of the high frequency of IS elements within the DC3000 genome may be a putative chromosomal inversion that apparently arose by homologous recombination between identical copies of ISPsy6 (PSPTO1567 and PSPTO4060) at widely separated insertion sites. This potential DNA rearrangement is revealed by inspection of the duplicated target site sequences (direct repeats) that flank some of the IS elements in DC3000 and is supported by an examination of gene synteny with other pseudomonads in the *surE-mutS* region of the genomes (Fig. 4). In PAO1, the six genes in the syntenic region (PA3625–3620) are *surE-pcm-nlpD-rpoS-fdxA-mutS*. In DC3000, a hypothetical protein and the insertion sequence ISPsy6 (PSPTO1566, 1567) are located downstream of *surE-rpoS* (PSPTO1562–1565), followed by the TTSS effector *hopPtoJ* (PSPTO1568) with these three ORFs being constituents of LSR15. Approximately 2838 kb (44%) of the DC3000 chromosome appears to be inverted between the two copies of ISPsy6 with *fdxA* (PSPTO4059) and *mutS* (PSPTO4058) located adjacent to the second copy ISPsy6 (PSPTO4060). Because an inversion about the origin does not invert the direction of transcription relative to replication of genes on the segment, such an inversion is a commonly observed feature in bacterial genomes (Eisen *et al*., 2000; Suyama and Bork, 2001). Interestingly, the *surE-mutS* region in *P. putida* KT2440 (PP1620–1629) contains a group II intron-encoding maturase (PP1624), suggesting that this region of the chromosome in pseudomonads may be a hotspot for insertion of MGEs.

The proposed chromosomal inversion appears to be a relatively recent event as it is not observed in other *Pseudomonas* species sequenced to date. The same gene order as in PAO1 is seen in *P. s.* pv. *syringae* B728a (Psyr2162–2167; accession number NZ_AABP00000000), *P. s.* pv. *phaseolicola* 1448A (V. Joardar, unpublished data), *P. aeruginosa* UCBPP-PA14 (*mutS-fdxA-rpoS-pcm-surE*; accession number NZ_AABQ07000001), *P. fluorescens* PfO-1 (*mutS-fdxA-rpoS-pcm-surE*; accession number NZ_AAAT02000037) and *P. fluorescens* CHA0 (*fdxA-rpoS*; Heeb *et al*., 2002). A putative homologue of the TTSS effector gene *hopPtoJ* is present in *P. s.* pv. *syringae* B728a (accession number NZ_AABP02000008; TBlastN, *E*-value = 1.0E-135) and it appears to be present in a region of the chromosome not linked to the *surE-mutS* cluster. Another feature of interest in the *surE-mutS* region of the DC3000 chromosome is *rsmZ*$_{Ps}$, the homologue of a small regulatory RNA, *rsmZ* described in *P. fluorescens* CHA0 (accession number AF245440; Heeb *et al*., 2002). In the CHA0 strain, *rsmZ* is located between *rpoS* and *fdxA*, whereas in DC3000, *rsmZ*$_{Ps}$ is present immediately upstream of the proposed inversion point ISPsy6 (PSPTO1567).

## CONCLUSIONS

Comparative genomic approaches revealed that the genetic complement specific to the plant pathogen *P. syringae* pv *tomato* DC3000, in comparison with the opportunistic animal pathogen *P. aeruginosa* and the saprophyte *P. putida*, was highly enriched in genes with unknown function and MGEs. A further investigation of the DC3000-specific ORFs suggests a majority of them share features with laterally transferred genes and that these are likely to contribute to various aspects of the fitness and pathogenicity of DC3000. For example, more than 80% of the known virulence genes that are specific to DC3000 (73 of 88) are localized in LSRs in the chromosome and a majority of the TTSS effector genes are contained in LSRs that have clusters of such genes. With respect to potential fitness and adaptation to the environment, DC3000-specific ORFs that are present within LSRs appear to aid in utilization of plant-derived energy sources such as sucrose and cyanate. In addition, an array of MCPs, including ten in LSRs, are likely to enhance the ability of the bacterium to respond to chemical stimuli in the environment. However, approximately half the ORFs present in LSRs (473 of 983, 48%) have yet to be assigned a function. One prominent feature of the DC3000 genome is the expansion of MGEs in comparison with the genomes of *P. aeruginosa* and *P. putida*. In addition to impacting single ORFs within the genome, MGEs also have the potential to mediate larger rearrangements such as the putative inversion of 2.8 Mb within the DC3000 genome. These data suggest that although the DC3000 genome has adapted to its environment and lifestyle as a pathogen primarily through attainment of new ORFs, large-scale rearrangements may permit another level of genome plasticity.

## EXPERIMENTAL PROCEDURES

### Sequence sources

The complete genome sequence and annotation of *P. aeruginosa* PAO1 (accession number AE004091; Stover *et al.*, 2000) and *P. putida* KT2440 (accession number AE015451; Nelson *et al.*, 2002) were obtained from GenBank. *P. syringae* pv *tomato* DC3000 sequence and annotation were obtained from databases within The Institute for Genomic Research and can be obtained through GenBank (accession number AE016853; Buell *et al.*, 2003). The PAO1 data used represent the 8 October 2003 version of the continually updated, reviewed, *P. aeruginosa* PAO1 genome annotation from PseudoCAP (http://www.pseudomonas.com).

### Bioinformatic analyses

Functional role categories were defined using a modified system by Riley (1993). Regions of atypical trinucleotide composition were identified by $\chi^2$ analysis (Heidelberg *et al.*, 2000). The distribution of all 64 trinucleotides (3-mers) was computed for the complete genome in all six reading frames, followed by a 3-mer distribution in 2000-bp windows. The sliding windows overlap by 1000 bp. For each window, the $\chi^2$ statistic on its 3-mer content as well as of the whole genome was computed, with a large $\chi^2$ value indicating that the 3-mer composition of that window is different from the rest of the chromosome. G+C values were calculated for each ORF and the mean G+C and standard deviation were calculated for all the ORFs in the chromosome. ORFs with unusual G+C content were identified as a variation from the mean G+C by two standard deviations (mean ± 2SD).

Conservation of gene order in the chromosomes of the three pseudomonads was established using the Position Effect algorithm (Carlton *et al.*, 2002), which utilized similarity at the protein level coupled with location within the genome. In brief, putative homologues were identified in the three *Pseudomonas* predicted proteomes using BlastP ($E < 10^{-15}$). The homologous proteins of each match pair were defined as anchor points in a sliding window of ten adjacent proteins on each side of the anchor point. The sliding windows were compared and scored with a scoring function that sums the per cent similarity (0–100) of the BlastP matches for matching genes and penalizes for gaps (e.g. non-matching genes within the window). Conserved gene order was identified by the presence of a minimum of four matching proteins within windows that scored above zero. Blocks of collinearity were identified in the reference chromosomes and only regions of the DC3000 chromosome not present in syntenic blocks were investigated further. LSRs were defined as regions of the DC3000 genome larger than 2 kb that are enriched in MGEs and/or ORFs specific to DC3000 in comparison with the two completely sequenced *Pseudomonas* species, PAO1 and KT2440. Specifically, at least 50% of ORFs within a LSR must be either MGEs or DC3000-specific genes, with no more than a stretch of four shared genes in between them.

## SUPPLEMENTARY MATERIAL

The following are available as supplementary material at http://www.blackwellpublishing.com/products/journals/suppmat/MPP/MPP263/MPP263sm.htm: **Fig. S1** $\chi^2$ values for the trinucleotide composition within the DC3000 genome. The $\chi^2$ values that are above the cutoff of 1173, representing regions of the genome with atypical trinucleotide composition, are shown in red (present within an LSR) and blue (not present within an LSR). The $\chi^2$ values below the cutoff of 1173 are shown in grey. **Table S1.** Trinucleotide skew and G+C content of the *P. syringae* pv *tomato* DC3000 chromosome. **Table S2.** *P. syringae* pv *tomato* DC3000 chromosomal ORFs present in Lineage-specific regions (LSRs). **Table S3.** *P. syringae* pv *tomato* DC3000 chromosome gene list and BlastP analyses with *P. aeruginosa* PA01 and *P. putida* KT2440.

## ACKNOWLEDGEMENTS

We thank Robert T. Deboy (Microbial Genomics) and Samuel V. Angiuoli (Bioinformatics) at The Institute for Genomic Research for their contribution to the work. Funding for the work described in this study was provided by National Science Foundation Grant DBI-0077622.

## REFERENCES

**Alfano, J.R., Charkowski, A.O., Deng, W.L., Badel, J.L., Petnicki-Ocwieja. T., van Dijk, K. and Collmer. A.** (2000) The *Pseudomonas syringae* Hrp pathogenicity island has a tripartite mosaic structure composed of a cluster of type III secretion genes bounded by exchangeable effector and conserved effector loci that contribute to parasitic fitness and pathogenicity in plants. *Proc. Natl Acad. Sci. USA*, **97**, 4856–4861.

**Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R.** (2004) The Pfam protein families database. *Nucleic Acids Res.* **32** (Database issue), D138–141.

**Bhattacharyya, A., Stilwagen, S., Ivanova, N., D'Souza, M., Bernal, A., Lykidis, A., Kapatral, V., Anderson, I., Larsen, N., Los, T., Reznik, G., Selkov, E. Jr, Walunas, T.L., Feil, H., Feil, W.S., Purcell, A., Lassez, J.L., Hawkins, T.L., Haselkorn, R., Overbeek, R., Predki, P.F. and Kyrpides, N.C.** (2002) Whole-genome comparative analysis of three phytopathogenic *Xylella fastidiosa* strains. *Proc. Natl Acad. Sci. USA*, **99**, 12403–12408.

**Buell, C.R., Joardar, V., Lindeberg, M., Selengut, J., Paulsen, I.T., Gwinn, M.L., Dodson, R.J., Deboy, R.T., Durkin, A.S., Kolonay, J.F., Madupu, R., Daugherty, S., Brinkac, L., Beanan, M.J., Haft, D.H., Nelson, W.C., Davidsen, T., Zafar, N., Zhou, L., Liu, J., Yuan, Q., Khouri, H., Fedorova, N., Tran, B., Russell, D., Berry, K., Utterback, T., Van Aken, S.E., Feldblyum, T.V., D'Ascenzo, M., Deng, W.L., Ramos, A.R., Alfano, J.R., Cartinhour, S., Chatterjee, A.K., Delaney, T.P., Lazarowitz, S.G., Martin, G.B., Schneider, D.J., Tang, X., Bender, C.L., White, O., Fraser, C.M. and Collmer, A.** (2003) The complete genome sequence of the *Arabidopsis* and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000. *Proc. Natl Acad. Sci. USA*, **100**, 10181–10186.

**Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., Peterson, J.D., Pop, M., Kosack, D.S., Shumway, M.F., Bidwell, S.L., Shallom, S.J., Van Aken, S.E., Riedmuller, S.B., Feldblyum, T.V., Cho, J.K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L.M., Florens, L., Yates, J.R., Raine, J.D., Sinden, R.E., Harris, M.A., Cunning-ham, D.A., Preiser, P.R., Bergman, L.W., Vaidya, A.B., van Lin, L.H., Janse, C.J., Waters, A.P., Smith, H.O., White, O.R., Salzberg, S.L., Venter, J.C., Fraser, C.M., Hoffman, S.L., Gardner, M.J. and Carucci, D.J.** (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*, **419**, 512–519.

**Chang, J.H., Goel, A.K., Grant, S.R. and Dangl, J.L.** (2004) Wake of the flood: ascribing functions to the wave of type III effector proteins of phytopathogenic bacteria. *Curr. Opin. Microbiol.* **7**, 11–18.

**Collmer, A., Lindeberg, M., Petnicki-Ocwieja, T., Schneider, D.J. and Alfano, J.R.** (2002) Genomic mining type III secretion system effectors in *Pseudomonas syringae* yields new picks for all TTSS prospectors. *Trends Microbiol.* **10**, 462–469.

**Delcher, A.L., Harmon, D., Kasif, S., White, O. and Salzberg, S.L.** (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641.

**Eisen, J.A., Heidelberg, J.F., White, O. and Salzberg, S.L.** (2000) Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* **1**, RESEARCH0011.1–11.9.

**Gardan, L., Shafik, H., Belouin, S., Broch, R., Grimont, F. and Grimont, P.A.** (1999) DNA relatedness among the pathovars of *Pseudomonas syringae* and description of *Pseudomonas tremae* sp. *nov.* and *Pseudomonas cannabina* sp. *nov.* (ex Sutic and Dowson 1959). *Int. J. Syst. Bacteriol.* **49**, 469–478.

**Gottwald, J.R., Krysan, P.J., Young, J.C., Evert, R.F. and Sussman, M.R.** (2000) Genetic evidence for the in planta role of phloem-specific plasma membrane sucrose transporters. *Proc. Natl Acad. Sci. USA*, **97**, 13979–13984.

**Greenberg, J.T. and Vinatzer, B.A.** (2003) Identifying type III effectors of plant pathogens and analyzing their interaction with plant cells. *Curr. Opin. Microbiol.* **6**, 20–28.

**Groisman, E.A. and Ochman, H.** (1996) Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*, **87**, 791–794.

**Hacker, J. and Carniel, E.** (2001) Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Report*, **2**, 376–381.

**Hacker, J. and Kaper, J.B.** (2000) Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* **54**, 641–679.

**Haft, D.H., Selengut, J.D. and White, O.** (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373.

**He, J., Baldini, R.L., Deziel, E., Saucier, M., Zhang, Q., Liberati, N.T., Lee, D., Urbach, J., Goodman, H.M. and Rahme, L.G.** (2004) The broad host range pathogen *Pseudomonas aeruginosa* strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes. *Proc. Natl Acad. Sci. USA*, **101**, 2530–2535.

**Heeb, S., Blumer, C. and Haas, D.** (2002) Regulatory RNA as mediator in GacA/RsmA-dependent global control of exoproduct formation in *Pseudomonas fluorescens* CHA0. *J. Bacteriol.* **184**, 1046–1056.

**Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Umayam, L., Gill, S.R., Nelson, K.E., Read, T.D., Tettelin, H., Richardson, D., Ermolaeva, M.D., Vamathevan, J., Bass, S., Qin, H., Dragoi, I., Sellers, P., McDonald, L., Utterback, T., Fleishmann, R.D., Nierman, W.C. and White, O.** (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*, **406**, 477–483.

**Huang, H.C., Schuurink, R., Denny, T.P., Atkinson, M.M., Baker, C.J., Yucel, I., Hutcheson, S.W. and Collmer, A.** (1988) Molecular cloning of a *Pseudomonas syringae* pv. *syringae* gene cluster that enables *Pseudomonas fluorescens* to elicit the hypersensitive response in tobacco plants. *J. Bacteriol.* **170**, 4748–4756.

**Johnson, W.V. and Anderson, P.M.** (1987) Bicarbonate is a recycling substrate for cyanase. *J. Biol. Chem.* **262**, 9021–9025.

**Mantri, Y. and Williams, K.P.** (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res.* **32** (Database issue), D55–58.

**McDaniel, T.K. and Kaper, J.B.** (1997) A cloned pathogenicity island from enteropathogenic *Escherichia coli* confers the attaching and effacing phenotype on *E. coli* K-12. *Mol. Microbiol.* **23**, 399–407.

**Nelson, K.E., Weinel, C., Paulsen, I.T., Dodson, R.J. and Hilbert, H., Martins dos Santos, V.A., Fouts, D.E., Gill, S.R., Pop, M., Holmes, M., Brinkac, L., Beanan, M., DeBoy, R.T., Daugherty, S., Kolonay, J., Madupu, R., Nelson, W., White, O., Peterson, J., Khouri, H., Hance, I., Chris Lee, P., Holtzapple, E., Scanlan, D., Tran, K., Moazzez, A., Utterback, T., Rizzo, M., Lee, K., Kosack, D., Moestl, D., Wedler, H., Lauber, J., Stjepandic, D., Hoheisel, J., Straetz, M., Heim, S., Kiewitz, C., Eisen, J.A., Timmis, K.N., Dusterhoft, A., Tummler, B. and Fraser, C.M.** (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ. Microbiol.* **4**, 799–808.

**Paulsen, I.T., Banerjei, L., Myers, G.S., Nelson, K.E., Seshadri, R., Read, T.D., Fouts, D.E., Eisen, J.A., Gill, S.R., Heidelberg, J.F.,**

**Tettelin, H., Dodson, R.J., Umayam, L., Brinkac, L., Beanan, M., Daugherty, S., DeBoy, R.T., Durkin, S., Kolonay, J., Madupu, R., Nelson, W., Vamathevan, J., Tran, B., Upton, J., Hansen, T., Shetty, J., Khouri, H., Utterback, T., Radune, D., Ketchum, K.A., Dougherty, B.A. and Fraser, C.M.** (2003) Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science*, **299**, 2071–2074.

**Perna, N.T., Plunkett, G. 3rd, Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A. and Blattner, F.R.** (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.

**Riley, M.** (1993) Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* **57**, 862–952.

**Seshadri, R., Myers, G.S., Tettelin, H., Eisen, J.A., Heidelberg, J.F., Dodson, R.J., Davidsen, T.M., DeBoy, R.T., Fouts, D.E., Haft, D.H., Selengut, J., Ren, Q., Brinkac, L.M., Madupu, R., Kolonay, J., Durkin, S.A., Daugherty, S.C., Shetty, J., Shvartsbeyn, A., Gebregeorgis, E., Geer, K., Tsegaye, G., Malek, J., Ayodeji, B., Shatsman, S., McLeod, M.P., Smajs, D., Howell, J.K., Pal, S., Amin, A., Vashisth, P., McNeill, T.Z., Xiang, Q., Sodergren, E., Baca, E., Weinstock, G.M., Norris, S.J., Fraser, C.M. and Paulsen, I.T.** (2004) Comparison of the genome of the oral pathogen *Treponema denticola* with other spirochete genomes. *Proc. Natl Acad. Sci. USA*, **101**, 5646–5651.

**da Silva, A.C., Ferro, J.A., Reinach, F.C., Farah, C.S., Furlan, L.R., Quaggio, R.B., Monteiro-Vitorello, C.B., Van Sluys, M.A., Almeida, N.F., Alves, L.M., do Amaral, A.M., Bertolini, M.C., Camargo, L.E., Camarotte, G., Cannavan, F., Cardozo, J., Chambergo, F., Ciapina, L.P., Cicarelli, R.M., Coutinho, L.L., Cursino-Santos, J.R., El-Dorry, H., Faria, J.B., Ferreira, A.J., Ferreira, R.C., Ferro, M.I., Formighieri, E.F., Franco, M.C., Greggio, C.C., Gruber, A., Katsuyama, A.M., Kishi, L.T., Leite, R.P., Lemos, E.G., Lemos, M.V., Locali, E.C., Machado, M.A., Madeira, A.M., Martinez-Rossi, N.M., Martins, E.C., Meidanis, J., Menck, C.F., Miyaki, C.Y., Moon, D.H., Moreira, L.M., Novo, M.T., Okura, V.K., Oliveira, M.C., Oliveira, V.R., Pereira, H.A., Rossi, A., Sena, J.A., Silva, C., de Souza, R.F., Spinola, L.A., Takita, M.A., Tamura, R.E., Teixeira, E.C., Tezza, R.I., Trindade dos Santos, M., Truffi, D., Tsai, S.M., White, F.F., Setubal, J.C. and Kitajima, J.P.** (2002) Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature*, **417**, 459–463.

**Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K., Wu, Z., Paulsen, I.T., Reizer, J., Saier, M.H., Hancock, R.E., Lory, S. and Olson, M.V.** (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature*, **406**, 959–964.

**Sundin, G.W., Mayfield, C.T., Zhao, Y., Gunasekera, T.S., Foster, G.L. and Ullrich, M.S.** (2004) Complete nucleotide sequence and analysis of pPSR1 (72,601 bp), a pPT23A-family plasmid from *Pseudomonas syringae* pv. *syringae* A2. *Mol. Genet. Genomics*, **270**, 462–476.

**Sung, Y.C., Parsell, D., Anderson, P.M. and Fuchs, J.A.** (1987) Identification, mapping, and cloning of the gene encoding cyanase in *Escherichia coli* K-12. *J. Bacteriol.* **169**, 2639–2642.

**Suyama, M. and Bork, P.** (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet.* **17**, 10–13.

**Taylor, D.E.** (1999) Bacterial tellurite resistance. *Trends Microbiol.* **7**, 111–1115.

**Taylor, D.E., Rooker, M., Keelan, M., Ng, L.K., Martin, I., Perna, N.T., Burland, N.T. and Blattner, F.R.** (2002) Genomic variability of O islands encoding tellurite resistance in enterohemorrhagic *Escherichia coli* O157:H7 isolates. *J. Bacteriol.* **184**, 4690–4698.

**Whalen, M.C., Innes, R.W., Bent, A.F. and Staskawicz, B.J.** (1991) Identification of *Pseudomonas syringae* pathogens of *Arabidopsis* and a bacterial locus determining avirulence on both *Arabidopsis* and soybean. *Plant Cell*, **3**, 49–59.

**Wittstock, U. and Halkier, B.A.** (2000) Cytochrome P450 CYP79A2 from *Arabidopsis thaliana* L. catalyzes the conversion of 1-phenylalanine to phenylacetaldoxime in the biosynthesis of benzylglucosinolate. *J. Biol. Chem.* **275**, 14659–14666.

**Wolfgang, M.C., Kulasekara, B.R., Liang, X., Boyd, D., Wu, K., Yang, Q., Miyada, C.G. and Lory, S.** (2003) Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA*, **100**, 8484–8489.

**Yeo, C.C., Wong, D.T. and Poh, C.L.** (1998) IS1491 from *Pseudomonas alcaligenes* NCIB 9867: characterization and distribution among *Pseudomonas* species. *Plasmid*, **39**, 187–195.