The ISME Journal (2007) 1, 331–340

www.nature.com/ismej

ORIGINAL ARTICLE

Sequence-based analysis of pQBR103; a representative of a unique, transfer-proficient mega plasmid resident in the microbial community of sugar beet

Adrian Tett[1,2,10], Andrew J Spiers[1,3,10,11], Lisa C Crossman[4,10], Duane Ager[1], Lena Ciric[1], J Maxwell Dow[5], John C Fry[2], David Harris[4], Andrew Lilley[1], Anna Oliver[1], Julian Parkhill[4], Michael A Quail[4], Paul B Rainey[3,6,7], Nigel J Saunders[8], Kathy Seeger[4], Lori AS Snyder[8,12], Rob Squares[4], Christopher M Thomas[9], Sarah L Turner[1], Xue-Xian Zhang[6], Dawn Field[1] and Mark J Bailey[1]

[1]*Centre for Ecology and Hydrology-Oxford, Oxford, UK;* [2]*Cardiff School of Biosciences, Cardiff University, Park Place, Cardiff, UK;* [3]*Department of Plant Sciences, University of Oxford, Oxford, UK;* [4]*Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton Cambridge, UK;* [5]*Department of Microbiology, BIOMERIT Research Centre, BioSciences Institute, National University of Ireland, Cork, Ireland;* [6]*Institute of Molecular Bioscience, Massey University, Auckland, New Zealand;* [7]*Institute of Advanced Study, Massey University, Auckland, New Zealand;* [8]*Sir William Dunn School of Pathology, University of Oxford, Oxford, UK and* [9]*School of Biosciences, University of Birmingham, Birmingham, UK*

The plasmid pQBR103 was found within *Pseudomonas* populations colonizing the leaf and root surfaces of sugar beet plants growing at Wytham, Oxfordshire, UK. At 425 kb it is the largest self-transmissible plasmid yet sequenced from the phytosphere. It is known to enhance the competitive fitness of its host, and parts of the plasmid are known to be actively transcribed in the plant environment. Analysis of the complete sequence of this plasmid predicts a coding sequence (CDS)-rich genome containing 478 CDSs and an exceptional degree of genetic novelty; 80% of predicted coding sequences cannot be ascribed a function and 60% are orphans. Of those to which function could be assigned, 40% bore greatest similarity to sequences from *Pseudomonas* spp, and the majority of the remainder showed similarity to other γ-proteobacterial genera and plasmids. pQBR103 has identifiable regions presumed responsible for replication and partitioning, but despite being *tra*$^+$ lacks the full complement of any previously described conjugal transfer functions. The DNA sequence provided few insights into the functional significance of plant-induced transcriptional regions, but suggests that 14% of CDSs may be expressed (11 CDSs with functional annotation and 54 without), further highlighting the ecological importance of these novel CDSs. Comparative analysis indicates that pQBR103 shares significant regions of sequence with other plasmids isolated from sugar beet plants grown at the same geographic location. These plasmid sequences indicate there is more novelty in the mobile DNA pool accessible to phytosphere pseudomonas than is currently appreciated or understood.

The ISME Journal (2007) **1,** 331–340; doi:10.1038/ismej.2007.47; published online 5 July 2007
Subject Category: integrated genomics and post genomics approaches in microbial ecology
Keywords: *Pseudomonas*; phytosphere; environmental; plasmid; sequence

Introduction

The pace of genome sequencing continues to accelerate with over 300 bacterial genomes completed and more than 1000 expected within the next year (Overbeek *et al.*, 2005). Even with this vast complement of genomes, new genes are continually being revealed and the typical genome now has, on average, 20% orphans and up to 70% conserved but uncharacterized genes (Galperin and Koonin, 2004). The pace of discovery is set to accelerate rapidly

Correspondence: Professor MJ Bailey, Centre for Ecology and Hydrology-Oxford, Mansfield Road, Oxford OX1 3SR, UK.
E-mail: mbailey@ceh.ac.uk
[10]These authors contributed equally to this work.
[11]Current address: SIMBIOS Centre, University of Abertay Dundee, Bell Street, Dundee DD1 1HG, UK.
[12]Current address: Centre for Systems Biology, School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK.
Received 5 March 2007; revised 22 May 2007; accepted 22 May 2007; published online 5 July 2007

with the rate at which natural environments are being explored using genomic techniques (Field *et al.*, 2006).

Plasmids, ubiquitous within most bacterial communities, are known to carry a wide range of important traits such as antibiotic resistance, virulence factors and degradative pathways, and play an extensive role in bacterial evolution as agents of horizontal gene transfer (HGT) (Ochman *et al.*, 2000; Osborn and Boltner, 2002; van Elsas *et al.*, 2003; Frost *et al.*, 2005). Despite this, plasmids have received far less attention than their bacterial and archaeal hosts (Frost *et al.*, 2005). Over 900 plasmids have been sequenced, but most of these are relatively small (500 are under 10 kb) and many have been sequenced for their medical importance or coincidentally during bacterial and archaeal genome sequencing projects; of the first 353 bacterial genome projects, 117 have produced 231 of the 852 eubacterial plasmids available in the public domain (Molbak *et al.*, 2003; www.genomics@ceh.ac.uk/plasmiddb/). As we have yet to sample a significant proportion of natural bacterial diversity (Martiny and Field, 2005), we have barely glimpsed the likely diversity of plasmids, especially those that are of a more significant size (only 12% of sequenced plasmids are >100 kb) (Molbak *et al.*, 2003). The plasmid gene pool, along with other mobile genetic elements represents a vast component of DNA on earth that we know comparatively little about and that is rarely sampled (Frost *et al.*, 2005).

Conjugative plasmids are the most important agents of gene transfer in phytosphere bacteria (Miller and Levy, 1989; Lilley *et al.*, 1994; Bailey *et al.*, 1994, 1996) and here we characterize the genome of a plasmid obtained from one of the best-characterized long-term studies of plant-associated bacteria. In a 5-year study (Wytham Farm, Oxfordshire, UK) plasmids were isolated from bacteria isolated from the roots and leaves of crops (sugar beet, wheat and corn), grasses, nettles, thistles and daisies (Bailey *et al.*, 2001). Restriction fragment comparisons of these plasmids and repetitive (REP)-PCR comparisons of their bacterial hosts have identified five local plasmid (group I–V) types (Lilley *et al.*, 1996) among at least 15 pseudomonad types (Bailey *et al.*, 2001) colonizing the phytosphere (leaves and roots).

We have chosen to sequence the pQBR103 plasmid, because it is a typical member of one of the most common plasmid types found at Wytham, the genetically distinct group-I plasmids. Group-I plasmids confer mercuric resistance (Hg^R) and their transfer is known to be promoted seasonally in the phytosphere of crop plants and field weeds (Lilley *et al.*, 1994, 2003; Lilley and Bailey, 1997a, b). pQBR103 is a self-transferable (conjugative) plasmid isolated in a landmark experiment replicated over 2 years assessing plasmid transfer in the phyllosphere. In these experiments, the naturally indigenous strain *Pseudomonas fluorescens* SBW25 was marked and reintroduced to sugar beet crops as a seed dressing. These SBW25 populations colonized the leaves and roots of the crop and 2 months later, in high summer, acquired a variety of plasmids including pQBR103 by transfer from indigenous hosts (Lilley and Bailey, 1997b). Although the original host of pQBR103 was not recovered from the sugar beet site, the plasmid can transfer and be maintained in a wide range of phytosphere *Pseudomonas* spp, but appears not to transfer to Enterobacter, α- or β-proteobacteria recipients (MJ Bailey, personal communication). Pseudomonads carrying group-1 plasmids show substantial reductions in fitness up to the mid-season stage of plant development (sugar beet and chickweed). However, in green-house, field and mesocosm studies, as plants mature they recover the full fitness of the plasmid-free control (Lilley and Bailey, 1997b; Lilley *et al.*, 2003).

pQBR103 persists in SBW25 in the laboratory and in field studies, and can be experimentally manipulated. Using *In Vivo* Expression Technology (IVET) (Rainey, 1999; Zhang *et al.*, 2004a), 37 transcriptional fusions from pQBR103 were recovered that were induced in the plant environment. Interestingly, such fusions were recovered, relative to genome size, six times more frequently in the plasmid than similarly active fusions from the SBW25 chromosome (Zhang *et al.*, 2004a). Some of these regions have been investigated further, but as yet no individual gene has been shown to be essential for phytosphere survival or fitness (Zhang *et al.*, 2004a, b). To date, these enigmatic, large environmental plasmids have mostly defied attempts at phenotypic characterization in laboratory experiments and other than the evidence for replication, maintenance and transfer, the only characterized traits of these plasmids are Hg and UV resistance (Lilley *et al.*, 1996). Genome sequencing plus post-genomic techniques such as comparative genomic hybridization (CGH) and expression-based microarrays offer complementary methods for exploring the genetic potential of these plasmids. Here, we report the complete sequence of pQBR103 and its comparison to diverse Hg^R plasmids from the sugar beet phytosphere using a CGH microarray.

Materials and methods

Plasmids, culturing and isolation

The pQBR plasmids used in this work (pQBR4, pQBR29, pQBR41, pQBR42, pQBR44, pQBR47, pQBR55, pQBR57 and pQBR103) were maintained in *P. fluorescens* SBW25 or *P. putida* UWC1 (Lilley *et al.*, 1996; Lilley and Bailey, 1997a). Cultures were grown using *Pseudomonas*-selective agar (PSA, Oxoid, UK) at 28°C. When appropriate, plasmids were maintained with 27 μg/ml $HgCl_2$. Broad-range

Hg^R was determined using 0.2–10 μg/ml phenylmercuric acetate (PMA). Plasmid DNA was isolated using methods as described previously (Lilley *et al.*, 1994).

Sequencing strategy and annotation
pQBR103 DNA was obtained from SBW25 (Lilley *et al.*, 1994) and further purification was achieved by gel electrophoresis and recovery. The enrichment of pQBR103 DNA relative to SBW25 chromosomal DNA is described in the Supplementary Information. Plasmid DNA was sonicated by two 10 s bursts at 15% maximum power using a Model CL4 sonicator (Misonix Inc., Farmingdale, NY, USA) and selected size fractions used to construct libraries in pUC19 (New England Biolabs, UK). The finished assembly was based on 4508 paired end-reads from one pUC19 library with insert sizes of 2.0–4.0 kb, 357 paired end-reads from a second library with inserts of 1.4–2.0 kb, and completed by gap filling to give an 8.64-fold sequence coverage. Amersham Big-Dye terminator sequencing chemistry (Amersham, Little Chalfont, UK) was used on ABI3700 sequencing machines. Phrap (http://www.phrap.org/) and GAP4 (Bonfield *et al.*, 1995) were used for sequence assembly, and Artemis (Rutherford *et al.*, 2000) used to annotate the finished sequence. The pQBR103 predicted proteome was submitted to the GNARE system (Sulakhe *et al.*, 2005) to assign EC numbers and map potential Kegg pathways. EMBOSS (http://emboss.sourceforge.net/) and REPUTER (Kurtz *et al.*, 2001) were used to detect repeats. Homologous coding sequence (CDS) in the genome were clustered with the mcl algorithm (Enright *et al.*, 2002). tRNAscan-SE (Lowe and Eddy, 1997) was used to identify transfer RNA genes. Mapping and analysis of the IVET sequences (Zhang *et al.*, 2004a, b) are described in the Supplementary Information. The data have been submitted to the EMBL database under accession number AM235768. A genome report compliant with the 'Minimum Information about a Genome Sequence specification' (MIGS) (http://gensc.sf.net) has been submitted to the Genome Catalogue (GCat identifier: 000021_GCAT) (Field *et al.*, 2007).

Microarray construction, hybridization and analysis
The design of the CGH microarray and probe production are described in the Supplementary Information. Briefly, 122 PCR-amplified probes were spotted six times onto glass slides with negative controls provided by pUC19, and *Escherichia coli* DH10 (Gibco-BRL, UK) and UWC1 chromosomal DNA. Plasmid and chromosomal DNA was extracted (McAllister and Stephens, 1993), labelled with either Cy3 or Cy5-dCTP (Amersham Pharmacia Biotech, UK) and hybridized individually as described (Snyder *et al.*, 2005). Slides were scanned using a ScanArray Express HT microarray scanner (Perkin Elmer, UK). Fluorescence data were extracted form the slide images using GenePix Pro 6 software (Molecular Devices, UK) and the mean median value from six replicate spots was determined (hybridization signal) for each probe hybridized.

Results

General features of the pQBR103 sequence
pQBR103 is a circular plasmid of 425 094 bp, significantly larger than the original 330 kb estimate (Lilley and Bailey, 1997a), making it the largest self-transmissible plasmid to be sequenced to date (www.genomics.ceh.ac.uk/plasmiddb/). A circular plot of the general features of pQBR103 is presented in Figure 1. The plasmid has an average G + C content of 53.15%, which is lower than that seen for other *Pseudomonas* spp chromosomes and most large plasmids (Table 1). A total of 478 predicted coding sequences (CSDs) have been annotated with an average size of 246 amino acids accounting for 83.4% of the coding capacity of the plasmid (coding density is 1.124 CDSs per kb). The distribution of predicted CDSs was found to be heavily biased, with 357 (76%) on the forward strand compared with 121 (24%) on the reverse strand. Only 95 (20%) of the predicted proteins could be ascribed a putative function through homology with known proteins or functional domains in public databases (Supplementary Table 1). Enzyme commission (EC) numbers could be assigned to 12 CDSs, one of which belongs to a described Kegg pathway (CDS 104: ubiquinone biosynthesis; EC 2.1.1). According to Kegg pathway annotations in the GNARE system (Sulakhe *et al.*, 2005), *P. aeruginosa* PAO1 (5,566 CDSs), *P. fluorescens* Pf-5 (6,137 CDS), and *P. fluorescens* PfO-1 (5,736 CDSs) have representatives of 78, 95 and 61 Kegg pathways respectively (using a GNARE cutoff of 10). Compared to these, a proteome of the same size as pQBR103 would be expected to have 5–7.3 pathways. The phylogenetic distribution of these homologues indicate that many, but not all, of the best matches come from members of the genus *Pseudomonas* (Supplementary Figure 1). A further 100 (21%) predicted proteins are conserved hypotheticals with homology to uncharacterized proteins and 283 (59%) are orphans with no significant level of detectable homology to sequences in public databases. Only a single conserved hypothetical pseudogene was found. pQBR103 was not found to carry any tRNA genes, or REP-like repeat elements found in *Pseudomonas* spp (Tobes and Pareja, 2006).

The maintenance genes of pQBR103: replication, partitioning and transfer
pQBR103 contains the 300 bp minimal origin of replication (*ori*V) identified experimentally in another group-I plasmid pQBR11 (Viegas *et al.*, 1997) (located at 259 339–259 639 bp with a 1 bp insertion)

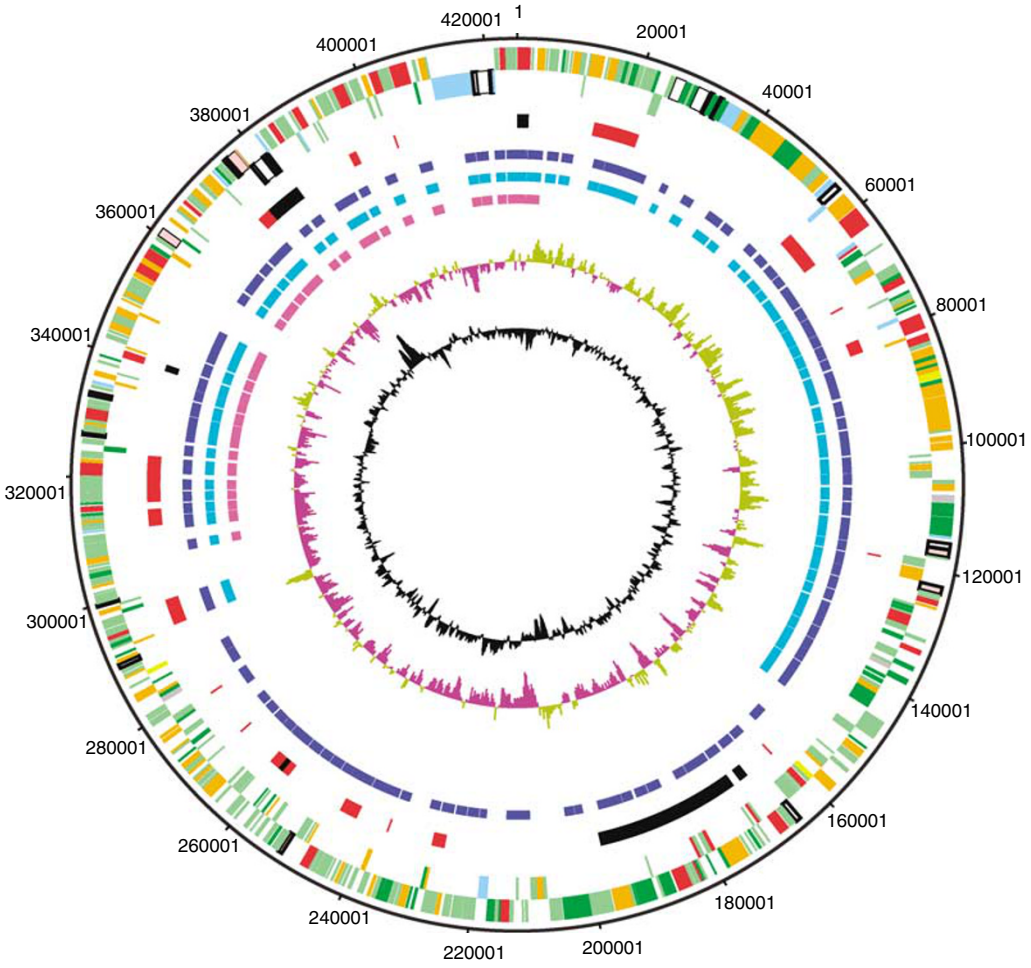

Figure 1 The 425 094 bp genome sequence of pQBR103 is presented in a circular plot, along with markers indicating regions of plant-specific transcription, and regions of conservation between pQBR103 and other group I plasmids, pQBR44 and pQBR47. Nine concentric circles are shown (from outer to innermost): 1, base pair coordinates. Base pair 1 of the genome has been arbitrarily defined as the replication mode of this plasmid is uncharacterized. 2–3, Annotated CDSs regions in the forward and reverse strands, respectively (functional CDSs, red: DNA associated, yellow: metabolism, pink: phage and transposon, white: environmental/survival and transmission, blue: regulators, dark green: transmembrane, grey: domain match only; pseudogene, brown: conserved hypothetical CDSs, orange: orphan CDS, light green). 4, Regions of interest (black, clockwise from 0 bp): ParAB, RulAB, potential Tra region, *ori*V, RepA replicon and Tn*5042*-like transposon; and (red) IVET regions of potential plant-induced transcriptional activity. 5–7: Microarray analyses of pQBR103 CDS distribution: probe regions in pQBR103 (blue), positive hybridization from pQBR44 (cyan) and pQBR47 (magneta). 8, GC skew. 9, GC deviation from the mean% G+C. CDS, coding sequence.

Table 1 Features of *Pseudomonas* plasmid and chromosomal genomes

Genome	*Size (kb)*	*% G+C*	*% Coding*[a]	*% Unknown*[b]	*IS/Tn*[c]
Pseudomonas chromosomes	5889–7075	58–66 (62.2±3.1)	84–90 (86.9±1.9)	26–45 (36±6.7)	2–170 (0.009±0.011)
Largest plasmids	251–2095	45–66 (54.6±8.3)	72–89 (82.7±5.5)	25–50 (37±9.0)	0–55 (0.054±0.056)
Pseudomonas plasmids	40–199	54–62 (56.0±2.4)	62–87 (75.9±7.8)	27–48 (39±7.6)	0–26 (0.070±0.049)
pQBR103	425	53	83	80	3 (0.007)

Genome characteristics compiled were known from 13 *Pseudomonas* chromosomes, 15 of the largest (circular) non-*Pseudomonas* plasmids and the 16 largest *Pseudomonas* plasmids sequenced in Molbak *et al.* (2003). Names and accession numbers for all genomes are supplied in the Supporting text. Means±s.d. are provided in parentheses, and as per kb.
[a]% Genome as CDSs.
[b]% Unknown CDSs annotated as conserved hypothetical (CH) (having homology to proteins of unknown function) and orphan (O) (with no homologues or function).
[c]Measure of the number of Insertion sequences or transposons (complete/partial) measured as the number of CDSs annotated as transposases.

but not the minimal replicon of the group-III plasmid pQBR55 (Turner *et al.*, 2002), which is known to share an overlapping host range *in situ*. pQBR103 also contains two plasmid replication initiator genes, both sharing significant homology with RepA from the IncA/C–IncP3 RA1 plasmid of *E. coli/Pseudomonas* spp. (Llanes *et al.*, 1994). However, only one, CDS 383, was found to have associated 32 copies of a 22 bp repeat (5′-GTTGTAGGTTTG(A/G)TG(G/C)GCCCTA-3′) and two DnaA boxes, and was therefore likely to represent a functional minimal replicon similar to that found in RA1 (Llanes *et al.*, 1994) (Figure 2).

CDS 001 (parA) and 002 (parB) are parAB active-partitioning system homologues and predicted to fulfil this role in pQBR103 (Hayes and Barilla, 2006), since there is a degenerate 48 bp inverted repeat upstream of *parA* potentially involved in autoregulation of *parAB* expression, and a 500 bp AT-rich region containing ~30 iterations of a degenerate 6 bp repeat downstream of *parB* that may represent a *parS*-like site (Figure 2). There are a further 3 parB homologues, none of which has proximal ParA homologues or repeats (Supplementary Table 2) and may or may not be involved in partitioning; at least one member of the ParB family is known simply as a regulatory protein (McKenna *et al.*, 2003). No recognizable coupled-cell death anti-segregation or site-specific multimer resolution systems were identified, although two Int-type recombinases without adjacent resolution sites (inverted repeats) are present.

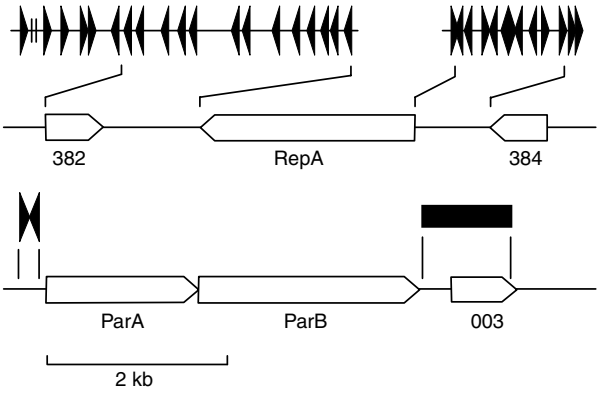

Figure 2 pQBR103 contains a putative RepA minimal replicon with tandem repeat elements separate from the putative ParAB partitioning cassette. The putative minimal replicon contains CDS383, a homologue of the plasmid replication initiator gene *repA* from the IncA/C-IncP3 plasmid RA1, 32 copies of a 22 bp repeat element (triangles) and two DnaA boxes (vertical lines) (338 642–341 577 bp) (top). The origin of replication would be expected to be located in the region defined by the DnaA boxes. The putative partitioning cassette contains CDS001 and 002, *parA* and *parB* homologues, respectively. Located upstream of *parA* is a degenerate 48 bp inverted repeat (424 970–425 069 bp) (triangles), and downstream of *parB* is a relatively AT-rich region (2100–2600 bp) (box) containing up to 30 copies of a 6 bp repeat (5′-TGC TTT-3′) element.

pQBR103 is known to be self-transmissible; however, the transfer functions are not readily identifiable through strong similarity to classic conjugal transfer systems (Peabody *et al.*, 2003; Frost *et al.*, 2005; Schroder and Lanka, 2005). A number of potential transfer-associated CDSs are identifiable within a 27 kb portion of pQBR103 containing CDSs 160–191 (170 292–197 117 bp). These share limited similarity to 6 of the 11 VirB/D4 proteins required for the archetypal type IV secretion system (T4SS) of *Agrobacterium tumefaciens* pTiC58 (Figure 3 and Supplementary Table 3). A putative DNA primase (CDS 209) and transfer inhibition protein (CDS 289) were also identified, either of which might be expected to be close to the origin of transfer. The limited extent of these homologies and the dispersed nature of the CDS could suggest that pQBR103 conjugation is mediated by a highly divergent T4SS-like transfer mechanism (but see below for comparison to other group-I plasmids).

Accessory traits

pQBR103 carries a near-perfect copy of the Tn*5042* Hg^R type II transposon (375 222–382 212 bp), which has been found to be highly conserved across contemporary plasmids and those isolated from permafrost grounds from the Upper Pleistocene (Mindlin *et al.*, 2005). Beyond the presence of the Hg^R operon (CDSs 430–435), a common characteristic of environmental plasmids (Barkay *et al.*, 2003), pQBR103 contains conspicuously few identifiable accessory genes of known function. RulAB homologues can be identified, which explains the enhanced UV-resistance the plasmid confers upon *P. fluorescens* SBW25 (Zhang *et al.*, 2004a), a factor known to impact the survival of *P. syringae* in the

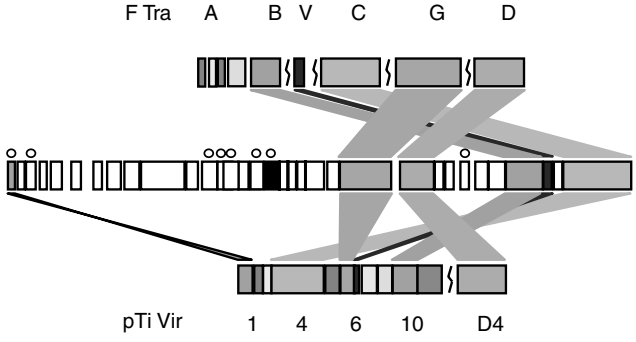

Figure 3 Components of a conjugal transfer apparatus sharing homology with classical Type IV secretion systems (T4SS) are located in the 170 292–197 117 bp region of pQBR103. Shown are the F plasmid Tra (top) (TraA, L, E, K, B, V, C, G and D), putative pQBR103 Tra region (middle) (CDSs 160–191, ~27 kb) and pTi (pTiC58) VirB/D4 (bottom) (VirB1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and VirD4) regions to scale. pQBR103 CDS in white have no recognizable role in transfer. Transfer components sharing the same functional role and sequence homology are indicated by colors. Transcription is from left to right, except for those CDSs in pQBR103 marked by a circle. Adapted from Schroder and Lanka (2005). CDS, coding sequence.

phyllosphere (Sundin and Murillo, 1999). Similar homologues enhance the fitness of *P. putida* under conditions of environmental stress (Tark *et al.*, 2005). Fifteen CDSs showed homology to proteins involved in the regulation of gene expression including a catabolite regulatory protein (Vfr/Crp), a RsmA/CrsA carbon storage regulator, a cold-shock DNA-binding domain protein, two RNA polymerase sigma factors, a cyclic diguanylate (c-di-GMP)-associated nucleotide cyclase and phosphodiesterase domain-containing response regulator-type proteins, and a number of other response regulator-type receiver domains, of which one (CDS 475) is part of a putative chemosensory/chemotaxis cluster (CDSs 440–475). Although *Pseudomonas* spp appear not to have *E. coli* H-NS-like homologues (Tendeng *et al.*, 2003), the presence of the putative DNA-binding NdpA and Hu homologues (CDSs 151 and 178) raises the possibility that pQBR103 may be able to influence DNA packaging and gene expression in a similar manner to that seen with H-NS (Dorman, 2007). Three CDSs (CDS 037, 038 and 051) also have weak homology to the AslB/AtsB arylsulfatase post-translational, though pQBR103 carries no cognate arylsulfatases.

Plant-inducible genes

The availability of the complete sequence enabled the mapping of IVET sequences reporting plant-induced transcriptional activity (Zhang *et al.*, 2004a) to 17 regions of the plasmid (Supplementary Table 4). Analysis of these suggest that 65 (14%) CDSs may be expressed in the sugar beet phytosphere from 11 regions. Included are the functional CDSs helA, helB, helC (putative DNA helicases) and Orn (oligoribonuclease) previously reported (Zhang *et al.*, 2004a, b), plus a ribonuclease, an AlgZ-like transcriptional regulator, a response regulator receiver domain protein, a restriction enzyme-related protein and the three Tn*5042* transposase subunits. It is notable that a further 54 CDSs without functional annotation are potentially transcribed as part of the same set of transcriptionally active regions of the plasmid genome. It is unlikely that any CDSs are transcribed in the remaining six regions as the orientation of IVET-reported transcriptional activity is the opposite to that suggested by the annotation of CDSs. Finally, the IVET insertions appear to indicate areas of complicated convergent/divergent/overlapping transcription in 4 of the 17 regions.

The evolution of large genome size

The large size of pQBR103 might have arisen through the accumulation of phage, insertion sequences and non-coding DNA, extensive internal duplications or the capture of novel sequences by a smaller ancestral plasmid. Compared to many other *Pseudomonas* plasmids pQBR103 contains evidence of only a single transposon and little evidence of bacteriophage remnants (Table 1). There are no extensive regions of apparent non-coding DNA in the annotation; the largest gap between adjacent CDSs on the same strand was 754 bp, and the largest gap between adjacent CDSs on opposing strands was 986 bp. Nor does pQBR103 contain extensive DNA repeat regions or unusual numbers of duplicated CDSs (Supplementary Table 2). Clustering of the predicted plasmid proteome revealed 20 protein families, the majority of which had homology to proteins outside pQBR103. The locations of these homologues suggest they are a mix of paralogues (that is, duplications within pQBR103 or an ancestral donor genome) and xenologues (acquisitions from different genomes).

Significant regions of pQBR103 are conserved in other group-I plasmids

To understand the relationship between the group-I plasmid pQBR103 and other plasmids that were also isolated from fluorescent pseudomonad phytosphere communities colonizing plants grown at the same geographic site, a comparative analysis of representatives of the three most common groups (I, III and IV) was performed. A PCR survey developed before the completion of the genome showed that three of the selected plasmids similar in size to pQBR103 shared all probes examined (Supplementary Table 5). In contrast, two much smaller and divergent group-I plasmids (pQBR44 and pQBR47) did not react with several of the probes. CGH microarray analysis shows that these plasmids are, overlapping subsets of contiguous regions of pQBR103 (Figures 1 and 4). Although this method can only provide information on the distribution of sequences present in pQBR103, previous estimates of size and similarities in REN profiles of pQBR44 and pQBR47 suggest they share their entire genetic content with pQBR103 (Lilley and Bailey, 1997a, b). Such analysis may define the conserved core region of group-I plasmids, although the putative origin, the potential transfer region and UV resistance determinant in pQBR103 appear to be absent from the smaller plasmids. The unshared region of pQBR103 contains a high proportion of orphan genes (70%) when compared to 50% for the shared region and 60% for the plasmid as a whole. This unshared region in pQBR103 also contains a smaller than expected number of functional genes and fewer with best matches to sequences of *Pseudomonas* spp origin.

Furthermore, the array study indicated that the group-I plasmids are completely distinct from group-III and group-IV plasmids. pQBR55 and pQBR57 showed no hybridization to the pQBR103 probes, with the exception of the Hg^R operon. This contains an organomercuric lyase, which is characteristic of broad-range inorganic–organic Hg^R (Barkay *et al.*, 2003). We empirically confirmed that all of the plasmids in this study confer up to 10 μg/

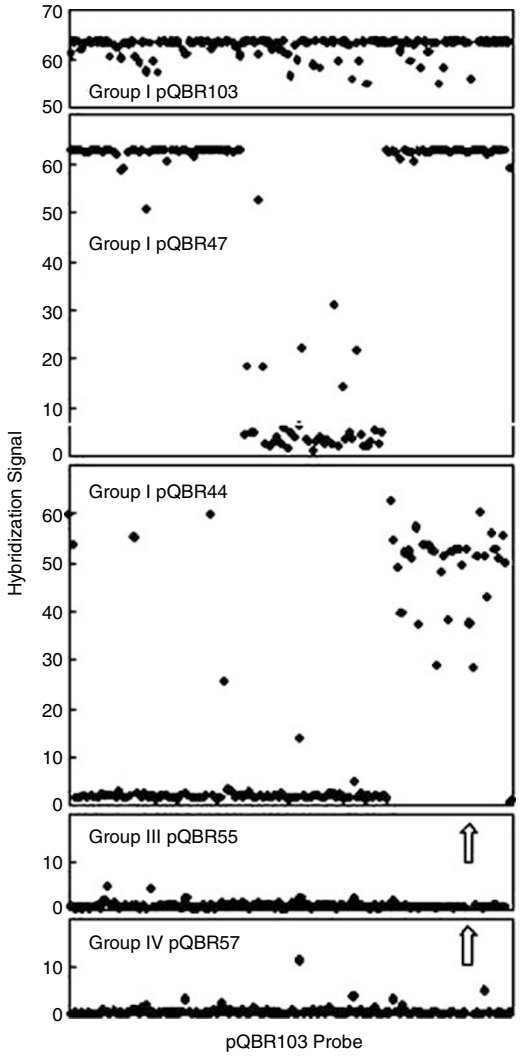

Figure 4 CGH microarray results show large regions of conservation and apparent deletions between group I plasmids. The microarray was used to test the group-I plasmids pQBR44 and pQBR47, the group-III plasmid pQBR55 and the group-IV plasmid pQBR57. The microarray used 122 pQBR103 probes, which are arranged in order along the *x* axis. Plasmid DNA used for hybridization was labelled with either Cy3 or Cy5-dCTP and the hybridization signal reported is the mean median fluorescence value from six replicate spots for each probe. The arrows indicate the position of strong Hg^R-probe signals for pQBR57 and pQBR55. No signals were obtained using labelled *P. putida* UWC1 chromosomal DNA, and the negative control probes did not hybridized to any of the labelled plasmid preparations (data not shown). CGH, comparative genomic hybridization.

ml PMA-resistance to *P. putida* UWC1 (which is resistant to 0.2 μg/ml PMA). Analysis of the Hg^R region has shown that it is highly conserved (Turner *et al.*, 2002; Mindlin *et al.*, 2005) and the partial sequencing of *merA* and *merR* revealed complete sequence conservation between the each of the different group- I, III and IV plasmids in this study (data not shown). On the basis of hybridization to microarray probes adjacent to the Tn, we infer that there are at least two insertion sites, one shared by the group-I plasmids and at least one in the group-III and group-IV plasmids.

Discussion

pQBR103 homology with known sequences is low

Other than megaplasmids sequenced as part of bacterial genome sequencing projects, pQBR103 is the largest plasmid sequenced to date for which independent transfer, replication and maintenance in different hosts has been demonstrated. Compared to other published plasmid sequence information which originate from the phytosphere environment pQBR103 contains the largest proportion of novel CDSs, indicating the potential untapped pool of genetic diversity within the horizontal gene pool for this specialized habitat. Eighty percent of the CDSs in pQBR103 cannot be ascribed putative function through homology and 60% share no significant similarity to known sequences. The lack of public database sequence information from large plasmids isolated from the phytosphere may, in part, explain the large number of orphan genes found in pQBR103. There are over 900 completely sequenced Eubacterial plasmids available in the public databases (Molbak *et al.*, 2003). However, only 2.6% (22) of these are above 100 kb and isolated from the phytosphere. About half of these large phytosphere plasmids are isolated from plant pathogens, and the other half are from rhizobia (from only five unique hosts). pQBR103 is the first example of a sequenced plasmid larger than 100 kb to be isolated from a phytosphere *Pseudomonas* species.

In addition to Hg^R, this CDS-rich genome has an identifiable putative RepA minimal replicon, origin of replication (*oriV*), and partition system (*par*AB), but apparently lacks the expected full complement of Tra functions and a significant set of accessory genes with obvious ecological value. In addition to Hg^R, the genome reveals one other obvious candidate fitness determinant, *rul*AB, which confers UV resistance that is of known value to bacteria found on plant leaves (Sundin and Murillo, 1999) and enhanced fitness under conditions of environmental stress (Tark *et al.*, 2005). Strikingly, pQBR103 lacks genes of the types commonly found in other large environmental plasmids (Galibert *et al.*, 2001; Gonzalez *et al.*, 2006) such as nutrient uptake and utilization genes associated with *Pseudomonas* phytosphere fitness (Gal *et al.*, 2003; Silby and Levy, 2004) or the complex organic compound metapathways carried by the smaller *Pseudomonas* plasmids (Greated *et al.*, 2002; Maeda *et al.*, 2003).

pQBR103 may provide fitness advantage by adapting the host to the chemically-complex phytosphere environment, perhaps through additional response and regulatory capacity. Preliminary proteomic expression studies (2D gel analysis) suggest that up to 48 *P. fluorescens* SBW25 polypeptides are upregulated or downregulated by pQBR103 in low

nutrient and pea exudates media (unpublished data, work in progress), in comparison, the expression of up to 9% of the *P. aeruginosa* PA01 transcriptome is affected by sugar beet exudates (Mark *et al.*, 2005). Only 10 possible pQBR103 polypeptides have been detected using 2D gel analysis, suggesting that the plasmid proteome is tightly regulated and probably highly sensitive to host and environmental conditions. Ultimately, site-specific mutagenesis, phenotypic, sugar beet and fitness assays will be required to determine the function and relative fitness value of each of the CDSs carried by pQBR103. The complete annotated sequence will greatly simplify the design and construction of such experiments.

The significance of intra-group mixing and inter-group compartmentalization of genetic material

Comparison to other characterized plasmids isolated from sugar beet grown at Wytham farm, UK, using PCR surveys and a CGH microarray, confirms that some pQBR103 sequences are shared with other group-I plasmids from this environment (Figure 4).

Whereas the putative pQBR103 sequences responsible for replication and partitioning are shared with the two smaller group-I plasmids pQBR44 and pQBR47, the candidate transfer region is not. As all three group-I plasmids are self-transmissible, this observation indicates that there may be more than one highly divergent and/or non-classical transfer mechanism. An alternative option is that there is a single novel conjugative transfer system within the conserved region and that the T4SS homologues might be involved in plant surface interactions; both theories need further investigation.

The relationship between the three group-I plasmids examined suggests a recent, common heritage where only a single recombination (deletion) event was required to explain the structure of each genome. Within our collection pQBR103 is not uniquely large, as previous fragment length polymorphism (FRLP) studies (Lilley *et al.*, 1996; Lilley and Bailey, 1997a) and the PCR survey in this study suggest high similarity between this plasmid and pQBR4, pQBR41 and pQBR42. Still, the distribution of CDSs with core plasmid functions throughout pQBR103 and the intermixing of these CDSs with those of presumed phytosphere function suggest that none are exclusively linked, and that the plasmid genome may be relatively free to recombine with other plasmids (and genomes) to generate derivatives in which different combinations of genes are produced. An implication of this is that pQBR103 and related plasmids may not be readily described in terms of an essential, minimal replicative backbone (in which replicative, maintenance and transfer functions are conserved, but may be linked or dispersed) and accessory genes distinct from those normally encoded by bacterial chromosomes (Frost *et al.*, 2005), not the least because so many of the pQBR103 CDSs remain to be functionally identified.

The enigma of pQBR103

Knowledge of the abundance, distributions and diversity of plasmids in bacterial communities from nonclinical environments is limited (Smalla *et al.*, 2000). pQBR103 is only one of a collection of hundreds of plasmids isolated from this environment on the basis of inorganic mercury resistance (Hg^R) and *ex situ* and *in situ* exogenous capture using a *Pseudomonas* spp (Lilley *et al.*, 1996; Lilley and Bailey, 1997a). The sequence of pQBR103 confirms that 40% of functionally characterized CDSs have greatest similarity to the chromosomal sequences of *Pseudomonas* spp, and a further 26% to closely related γ-proteobacteria many of which also have homology to *Pseudomonas* spp. This finding, plus the high proportion of orphan CDSs, suggest two models for the origin and current host-range of pQBR103 within the phytosphere community. In the first, pQBR103 is largely confined to the *Pseudomonas* but may have resided within other genera, and in the second, pQBR103 originally had a wide host range but has recently begun to specialize as a *Pseudomonas* plasmid. Although speculative, the relatively low G+C content of pQBR103 might also reflect a past history of residence in lower G+C bacteria such as *Erwinia* and *Klebsiella* spp, which are important members of this sugar beet phytosphere community known to support large numbers of plasmids (Powell *et al.*, 1993; Kobayashi and Bailey, 1994). The vast number of orphan genes suggests an unidentified genetic reservoir in the community, which could be shared with other plasmids or other bacteria which have yet to be characterized at the genomic level. Determining how large this genetic component of the pan-genome is and where else it resides is essential to understand the ecology of the microbial phytosphere community.

The sequence information provided by the analysis of this large, environmental plasmid indicates that the extent of novel genetic diversity in the phytosphere is extensive. While the relevance of HGT (Frost *et al.*, 2005) and the importance of the pan-genome (Rodriguez-Valera, 2002; Medini *et al.*, 2005) have been recognized, a broader genomic view, recognizing the interconnection of taxonomically more diverse bacterial genomes is only just emerging. We propose that 'pan-community' genomics should include the specific study of HGT networks that link the genomes of bacterial 'species'. The pQBR and related plasmids provide a suitable model system to study this phenomenon. Sequencing of pQBR103 has provided the first insight into the molecular landscape of a potentially new class of large, environmental, non-pathogenic, low copy number plasmids that appear to lack hallmark accessory genes. The sequencing of additional pQBR

plasmids will help to elucidate the role these plasmids play in the ecology of phytosphere microbial communities.

Acknowledgements

We thank G Wilson and C Ekeke for their help in the analysis of the pQBR103 sequence and Bela Tiwari for consultation on the microarray analysis. AT was funded by a UK NERC studentship. The sequencing of pQBR103 was funded via a UK BBSRC grant (P16729) to PR and CT. The unpublished *P. fluorescens* Pf0-1 and SBW25 genomes were accessed from the USA DOE Joint Genome Institute and the UK Wellcome Trust Sanger Institute, respectively.

References

Bailey MJ, Kobayashi N, Lilley AK, Powell BJ, Thompson IP. (1994). Potential for gene transfer in the phytosphere: isolation and characterisation of naturally occurring plasmids. In: Bazin MJ, Lynch JM (eds). *Environmental Gene Release*. Chapman & Hall: London, UK, pp 77–98.

Bailey MJ, Lilley AK, Diaper JD. (1996). Gene transfer in the phyllosphere. In: Morris CE, Nicot P, Nguyen-the C (eds). *Microbiology of Aerial Plant Surfaces*. Plenum Publishing: New York, USA, pp 103–123.

Bailey MJ, Rainey PB, Zhang X-X, Lilley AK. (2001). Population dynamics, gene transfer and gene expression in plasmids, the role of the horizontal gene pool in local adaptation at the plant surface. In: Lindow SE, Hecht-Poinar I, Elliott VJ (eds). *Microbiology of Aerial Plant Surfaces*. American Phytopath. Soc. Press: St Paul, USA, pp 171–189.

Barkay T, Miller SM, Summers AO. (2003). Bacterial mercury resistance from atoms to ecosystems. *FEMS Microbiol Rev* **27**: 355–384.

Bonfield JK, Smith K, Staden R. (1995). A new DNA sequence assembly program. *Nucl Acid Res* **23**: 4992–4999.

Dorman CJ. (2007). H-NS, the genome sentinel. *Nat Rev Microbiol* **5**: 157–161.

Enright AJ, Van Dongen S, Ouzounis CA. (2002). An efficient algorithm for large-scale detection of protein families. *Nucl Acid Res* **30**: 1575–1584.

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P *et al.* (2007). Towards a richer description of our complete collection of genomes and metagenomes: the 'Minimum Information about a Genome Sequence' (MIGS) specification. *Nat Biotechnol* (in press).

Field D, Wilson G, van der Gast C. (2006). How do we compare hundreds of bacterial genomes? *Curr Opin Microbiol* **9**: 499–504.

Frost LS, Leplae R, Summers AO, Toussaint A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* **3**: 722–732.

Gal M, Preston GM, Massey RC, Spiers AJ, Rainey PB. (2003). Genes encoding a cellulosic polymer contribute toward the ecological success of *Pseudomonas fluorescens* SBW25 on plant surfaces. *Mol Ecol* **12**: 3109–3121.

Galibert F, Finan TM, Long SR, Puhler A, Abola P, Ampe F *et al.* (2001). The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293**: 668–672.

Galperin MY, Koonin EV. (2004). 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucl Acid Res* **32**: 5452–5463.

Gonzalez V, Santamaria RI, Bustos P, Hernandez-Gonzalez I, Medrano-Soto A, Moreno-Hagelsieb G *et al.* (2006). The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. *Proc Natl Acad Sci USA* **103**: 3834–3839.

Greated A, Lambertsen L, Williams PA, Thomas CM. (2002). Complete sequence of the IncP-9 TOL plasmid pWW0 from *Pseudomonas putida*. *Environ Microbiol* **4**: 856–871.

Hayes F, Barilla D. (2006). The bacterial segrosome: a dynamic nucleoprotein machine for DNA trafficking and segregation. *Nat Rev Microbiol* **4**: 133–143.

Kobayashi N, Bailey MJ. (1994). Plasmids isolated from the sugar beet phyllosphere show little or no homology to molecular probes currently available for plasmid typing. *Microbiol* **140**: 289–296.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucl Acid Res* **29**: 4633–4642.

Lilley AK, Bailey MJ. (1997a). The acquisition of indigenous plasmids by a genetically marked pseudomonad population colonizing the sugar beet phytosphere is related to local environmental conditions. *Appl Environ Microbiol* **63**: 1577–1583.

Lilley AK, Bailey MJ. (1997b). Impact of plasmid pQBR103 acquisition and carriage on the phytosphere fitness of *Pseudomonas fluorescens* SBW25: burden and benefit. *Appl Environ Microbiol* **63**: 1584–1587.

Lilley AK, Bailey MJ, Barr M, Kilshaw K, Timms-Wilson TM, Day MJ *et al.* (2003). Population dynamics and gene transfer in genetically modified bacteria in a model microcosm. *Mol Ecol* **12**: 3097–3107.

Lilley AK, Bailey MJ, Day MJ, Fry J. (1996). Diversity of mercury resistance plasmids obtained by exogenous isolation from the bacteria of sugar beet in three successive seasons. *FEMS Microbiol Ecol* **20**: 211–228.

Lilley AK, Fry JC, Day MJ, Bailey MJ. (1994). *In situ* transfer of an exogenously isolated plasmid between indigenous donor and recipient *Pseudomonad* spp in sugar beet rhizosphere. *Microbiol* **140**: 27–33.

Llanes C, Gabant P, Couturier M, Michel-Briand Y. (1994). Cloning and characterization of the Inc A/C plasmid RA1 replicon. *J Bacteriol* **176**: 3403–3407.

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

Maeda K, Nojiri H, Shintani M, Yoshida T, Habe H, Omori T. (2003). Complete nucleotide sequence of carbazole/dioxin-degrading plasmid pCAR1 in *Pseudomonas resinovorans* strain CA10 indicates its mosaicity and the presence of large catabolic transposon Tn*4676*. *J Mol Biol* **326**: 21–33.

Mark GL, Dow JM, Kiely PD, Higgins H, Haynes J, Baysse C *et al.* (2005). Transcriptome profiling of bacterial responses to root exudates identifies genes involved in microbe-plant interactions. *Proc Natl Acad Sci USA* **102**: 17454–17459.

Martiny JBH, Field D. (2005). Ecological perspectives on the sequenced genome collection. *Ecol Lett* **8**: 1334–1345.

McAllister CF, Stephens DS. (1993). Analysis in *Neisseria meningitidis* and other *Neisseria* species of genes

homologous to the FKBP immunophilin family. *Mol Microbiol* **10**: 13–23.
McKenna S, Beloin C, Dorman CJ. (2003). In vitro DNA-binding properties of VirB, the *Shigella flexneri* virulence regulatory protein. *FEBS Lett* **545**: 183–187.
Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. (2005). The microbial pan-genome. *Curr Opin Genet Dev* **15**: 589–594.
Miller RV, Levy SB. (1989). Horizontal gene transfer in relation to environmental release of genetically engineered microorganisms. In: Levy SB, Miller RV (eds). *Gene Transfer in the Environment*. McGraw-Hill Publishing Company: New York, USA, pp 405–420.
Mindlin S, Minakhin L, Petrova M, Kholodii G, Minakhina S, Gorlenko Z *et al.* (2005). Present-day mercury resistance transposons are common in bacteria preserved in permafrost grounds since the upper pleistocene. *Res Microbiol* **156**: 994–1004.
Molbak L, Tett A, Ussery DW, Wall K, Turner S, Bailey M *et al.* (2003). The plasmid genome database. *Microbiology* **149**: 3043–3045.
Ochman H, Lawrence JG, Groisman EA. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299–304.
Osborn AM, Boltner D. (2002). When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. *Plasmid* **48**: 202–212.
Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M *et al.* (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucl Acid Res* **33**: 5691–5702.
Peabody CR, Chung YJ, Yen MR, Vidal-Ingigliardi D, Pugsley AP, Saier MH. (2003). Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. *Microbiology* **149**: 3051–3072.
Powell BJ, Purdy KJ, Thompson IP, Bailey MJ. (1993). Demonstration of tra$^+$ plasmid activity in bacteria indigenous to the phyllosphere of sugar beet; gene transfer to a genetically modified pseudomonad. *FEMS Microbiol Ecol* **12**: 195–206.
Rainey PB. (1999). Adaptation of *Pseudomonas fluorescens* to the plant rhizosphere. *Environ Microbiol* **1**: 243–257.
Rodriguez-Valera F. (2002). Approaches to prokaryotic biodiversity: a population genetics perspective. *Environ Microbiol* **4**: 628–633.
Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA *et al.* (2000). Artemis: sequence visualization and annotation. *Bioinformatics* **16**: 944–945.
Schroder G, Lanka E. (2005). The mating pair formation system of conjugative plasmids – a versatile secretion machinery for transfer of proteins and DNA. *Plasmid* **54**: 1–25.
Silby MW, Levy SB. (2004). Use of *in vivo* expression technology to identify genes important in growth and survival of *Pseudomonas fluorescens* Pf0-1 in soil: discovery of expressed sequences with novel genetic organization. *J Bacteriol* **186**: 7411–7419.
Smalla K, Krogerrecklenfort E, Heuer H, Dejonghe W, Top E, Osborn M *et al.* (2000). PCR-based detection of mobile genetic elements in total community DNA. *Microbiology* **146**: 1256–1257.
Snyder LA, Jarvis SA, Saunders NJ. (2005). Complete and variant forms of the 'gonococcal genetic island' in *Neisseria meningitidis*. *Microbiology* **151**: 4005–4013.
Sulakhe D, Rodriguez A, D'Souza M, Wilde M, Nefedova V, Foster I. (2005). GNARE: automated system for high-throughput genome analysis with grid computational backend. *J Clin Monit Comput* **19**: 361–369.
Sundin GW, Murillo J. (1999). Functional analysis of the *Pseudomonas syringae rulAB* determinant in tolerance to ultraviolet B (290–320 nm) radiation and distribution of *rulAB* among *P. syringae* pathovars. *Environ Microbiol* **1**: 75–87.
Tark M, Tover A, Tarassova K, Tegova R, Kivi G, Horak R *et al.* (2005). A DNA polymerase V homologue encoded by TOL plasmid pWW0 confers evolutionary fitness on *Pseudomonas putida* under conditions of environmental stress. *J Bacteriol* **187**: 5203–5213.
Tendeng C, Soutourina OA, Danchin A, Bertin PN. (2003). MvaT proteins in *Pseudomonas* spp.: a novel class of H-NS-like proteins. *Microbiology* **149**: 3047–3049.
Tobes R, Pareja E. (2006). Bacterial repetitive extragenic palindromic sequences are DNA targets for insertion sequence elements. *BMC Genomics* **7**: 62.
Turner SL, Lilley AK, Bailey MJ. (2002). Ecological and molecular maintenance strategies of mobile genetic elements. *FEMS Microbiol Ecol* **42**: 209–215.
Van Elsas J, Turner S, Bailey MJ. (2003). Horizontal gene transfer in the phytosphere. *New Phytologist* **157**: 525–537.
Viegas CA, Lilley AK, Bruce K, Bailey MJ. (1997). Description of a novel plasmid replicative origin from a genetically distinct family of conjugative plasmids associated with phytosphere microflora. *FEMS Microbiol Lett* **149**: 121–127.
Zhang XX, Lilley AK, Bailey MJ, Rainey PB. (2004a). The indigenous *Pseudomonas* plasmid pQBR103 encodes plant-inducible genes, including three putative helicases. *FEMS Microbiol Ecol* **51**: 9–17.
Zhang XX, Lilley AK, Bailey MJ, Rainey PB. (2004b). Functional and phylogenetic analysis of a plant-inducible oligoribonuclease (*orn*) gene from an indigenous *Pseudomonas* plasmid. *Microbiology* **150**: 2889–2898.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)