Nucleic Acids Research, 2006, Vol. 34, No. 1 ***e3***
doi:10.1093/nar/gnj005

A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria

Hong-Yu Ou[1], Ling-Ling Chen[2], James Lonnen[1], Roy R. Chaudhuri[3], Ali Bin Thani[1], Rebecca Smith[1], Natalie J. Garton[1], Jay Hinton[4], Mark Pallen[3], Michael R. Barer[1,5] and Kumar Rajakumar[1,5,*]

[1]Department of Infection, Immunity and Inflammation, Leicester Medical School, University of Leicester, Leicester LE1 9HN, UK, [2]Laboratory for Computational Biology, Shandong Provincial Research Center for Bioinformatic Engineering and Technique, Shandong University of Technology, Zibo, 255049, China, [3]Bacterial Pathogenesis and Genomics Unit, Division of Immunity and Infection, Medical School, University of Birmingham, Birmingham B15 2TT, UK, [4]Molecular Microbiology Group, Institute of Food Research, Norwich Research Park, Norwich NR4 7UA, UK and [5]Department of Clinical Microbiology, University Hospitals of Leicester NHS Trust, Leicester LE1 5WW, UK

Received November 2, 2005; Revised and Accepted December 12, 2005

ABSTRACT

We devised software tools to systematically investigate the contents and contexts of bacterial tRNA and tmRNA genes, which are known insertion hotspots for genomic islands (GIs). The strategy, based on MAUVE-facilitated multigenome comparisons, was used to examine 87 *Escherichia coli* MG1655 tRNA and tmRNA genes and their orthologues in *E.coli* EDL933, *E.coli* CFT073 and *Shigella flexneri* Sf301. Our approach identified 49 GIs occupying ~1.7 Mb that mapped to 18 tRNA genes, missing 2 but identifying a further 30 GIs as compared with Islander [Y. Mantri and K. P. Williams (2004), *Nucleic Acids Res.*, 32, D55–D58]. All these GIs had many strain-specific CDS, anomalous GC contents and/or significant dinucleotide biases, consistent with foreign origins. Our analysis demonstrated marked conservation of sequences flanking both empty tRNA sites and tRNA-associated GIs across all four genomes. Remarkably, there were only 2 upstream and 5 downstream deletions adjacent to the 328 loci investigated. *In silico* PCR analysis based on conserved flanking regions was also used to interrogate hotspots in another eight completely or partially sequenced *E.coli* and *Shigella* genomes. The tools developed are ideal for the analysis of other bacterial species and will lead to *in silico* and experimental discovery of new genomic islands.

INTRODUCTION

The synteny or colinearity of bacterial chromosomal genes is generally well preserved between strains of the same species. Conserved regions along the chromosomes of individual strains are referred to as the genomic backbone (1). Horizontal gene transfer events have led to the integration of alien genomic islands (GIs) into these backbones (2). This additional complement of DNA, which can vary considerably between members of the same species, frequently lies within recognized insertion 'hotspots'; the commonest and most generic of these being tRNA and tmRNA sites (2), hereafter referred to collectively as 'tRNA' genes. Foreign DNA segments include chromosomally captured plasmids, bacteriophage genomes, archetypal genomic islands (GIs) and various mosaic and degenerate elements. Furthermore, several studies have confirmed that individual island-encoded integrases recognize specific short sequences that typically comprise the 3′ termini of a growing list of tRNA genes (3,4).

We hypothesized that acquired islands could be identified by locating pairs of conserved backbone regions flanking potential insertion hotspots. Simple pair-wise alignment of segments from two genomes were unlikely to detect accurate genomic island boundaries, as deletions involving core chromosomal genes in a single strain could mask island flanks (5).

*To whom correspondence should be addressed at Department of Infection, Immunity and Inflammation, Leicester Medical School, University of Leicester, Maurice Shock Building, University Road, PO Box 138, Leicester LE1 9HN, UK. Tel: +44 0 116 2231498; Fax: +44 0 116 2525030; Email: kr46@le.ac.uk

Ideally, flanking segments from three or more genomes representative of a species would need to be compared. However, the inspection of numerous hotspots across multiple genomes is very laborious.

We have devised an easy-to-use software package and used it to perform a high-throughput systematic interrogation of tRNA genes in four *Escherichia coli* and *Shigella* genomes. The method, termed tRNAcc for tRNA gene content and context analysis, was complemented by an *in silico* PCR approach that identified putative GIs in all the complete and near-complete *E.coli* and *Shigella* genomes. The utility of the proposed method for *in vitro* screening of test bacterial strains was also highlighted. Exponential growth in genome sequence data has resulted in major bottlenecks in the analysis process. We propose that tRNAcc will help address this challenge by facilitating rapid, high-throughput discovery of GIs, thereby focussing increased research effort on these important genomic entities.

MATERIALS AND METHODS

Databases

Four fully sequenced *E.coli* and *Shigella* genomes were employed for the primary tRNAcc analysis. Complete genome sequences and annotation information were downloaded from NCBI (ftp.ncbi.nih.gov/genomes/): *E.coli* K-12 MG1655 (NC_000913.2) (6), uropathogenic *E.coli* CFT073 (NC_004431) (7), enterohaemorrhagic *E.coli* O157:H7 EDL933 (NC_002655) (1) and *Shigella flexneri* 2a Sf301 (NC_004337) (8). Details of tRNA and tmRNA (*ssrA*) genes were obtained from NCBI annotations and the tmRNA website (9), respectively.

tRNA content and context (tRNAcc) analysis

All 86 tRNA genes and the 1 tmRNA gene (*ssrA*) in MG1655 and their corresponding orthologues in the other three genomes were investigated to verify whether their 3′ end regions were occupied by islands. The tRNAcc method is illustrated in Figure 1. For each locus a 4 kb upstream chromosomal block (UCB), the tRNA gene and a 250 kb downstream chromosomal block (DCB) were extracted. Virtually all identified GIs are smaller than 250 kb in size. Next, the UCB and DCB fragments from the four genomes were aligned separately using the multiple sequence aligner Mauve v1.2.2 (10) that calculated gapped alignment scores using MUSCLE 3.52 (11). The program was run using default parameters, except that the minimum backbone size was set to 500 bp. Conserved upstream and downstream flanking regions were then identified by parsing resulting backbone reports.

In this study, a GI was defined as the anomalous segment between the 3′ end of the tRNA gene and 5′ end of the corresponding conserved downstream flanking region. If the 3′ end of a tRNA site was adjacent to a putative island, the corresponding tRNA site was referred to as 'occupied' (Figure 2b); otherwise it was flagged as 'empty' (Figure 2a). IdentifyIsland (Table 1) combined the multiple sequence aligner Mauve 1.2.2 (10) with subsequent processing modules written using C++ to perform the above steps. A total of 328 tRNA sites were examined in turn.

Potential problems may arise during analysis of bacterial genomes that exhibit high levels of intra-species rearrangements. However, the vast majority of empty tRNA sites should be correctly recognized as re-arrangement events are very unlikely to have directly disrupted the short DNA segments of ~1–2 kb that contain the empty tRNA gene and its cognate conserved flanking sequences. This segment alone is sufficient to define the site as empty. Sites that are identified as occupied and any others that are flagged up as problematic by the algorithm, either because of a missing or inverted flank, should be re-examined manually to ensure confidence in the limits of the islands defined and detect any missed islands. In a few instances, it may be worth considering reordering orthologous regions of comparator genomes to match that of the reference genome, thus minimizing these problems.

To account for isolated strain-specific deletion events involving core chromosomal DNA immediately flanking tRNA sites (Figure 2), we used IdentifyIsland to individually analyse different subsets of the four genomes (Figure 1). MG1655 was used as a reference template in all cases. We then compared the sizes and boundaries of putative GIs identified using TabulateIsland and selected those corresponding to the set yielding the smallest GIs. These smaller entities did not include core chromosomal regions that had been deleted in one strain only. Clearly it would be impractical when analysing five or more genomes to analyse all subset permutations. However, as subset analysis is primarily performed to identify isolated instances of deleted upstream or downstream flanking sequences, it would be sufficient in the majority of instances to examine the limited number of subsets generated by omitting only one genome. Alternatively, subsets of a primary panel of four selected genomes could be analysed by tRNAcc, with cognate tRNA loci in additional available genomes interrogated using the *in silico* tRIP procedure described below. Following automated tRNAcc analysis potentially occupied sites were examined using the *coli*BASE online utility (http://colibase.bham.ac.uk/) (12) and/or the interactive Artemis Comparison Tool (ACT) sequence viewer (13). GIs <1 kb in size were excluded, while hemi-nested islands that had been multiply assigned to several adjacent, closely clustered tandem tRNA genes were re-assigned to a single tRNA locus.

In silico PCR-based interrogation of tRNA sites

Twenty tRNA genes that had been found to harbour an island in one of the four primary genomes were selected for *in silico* PCR-based interrogation. Based on the identified GI boundaries, 2 kb upstream (UF) and downstream (DF) flanking regions were extracted using ExtractFlank (Table 1). Primers specific for each flank were designed using Primaclade (14), with ClustalW-derived (15) multiple sequence alignments serving as inputs. Candidate primers were then screened by BLASTN (16) against the genomes under consideration to minimize the likelihood of non-specific amplification. Finally, selected primer pairs for each tRNA site were checked using a locally installed version of electronic PCR (e-PCR) (17). If the *in silico* tRIP amplicon that was obtained corresponded to the expected product, the primer pair was considered to be specific. Details of primers used in this study are listed in Supplementary Table S4.

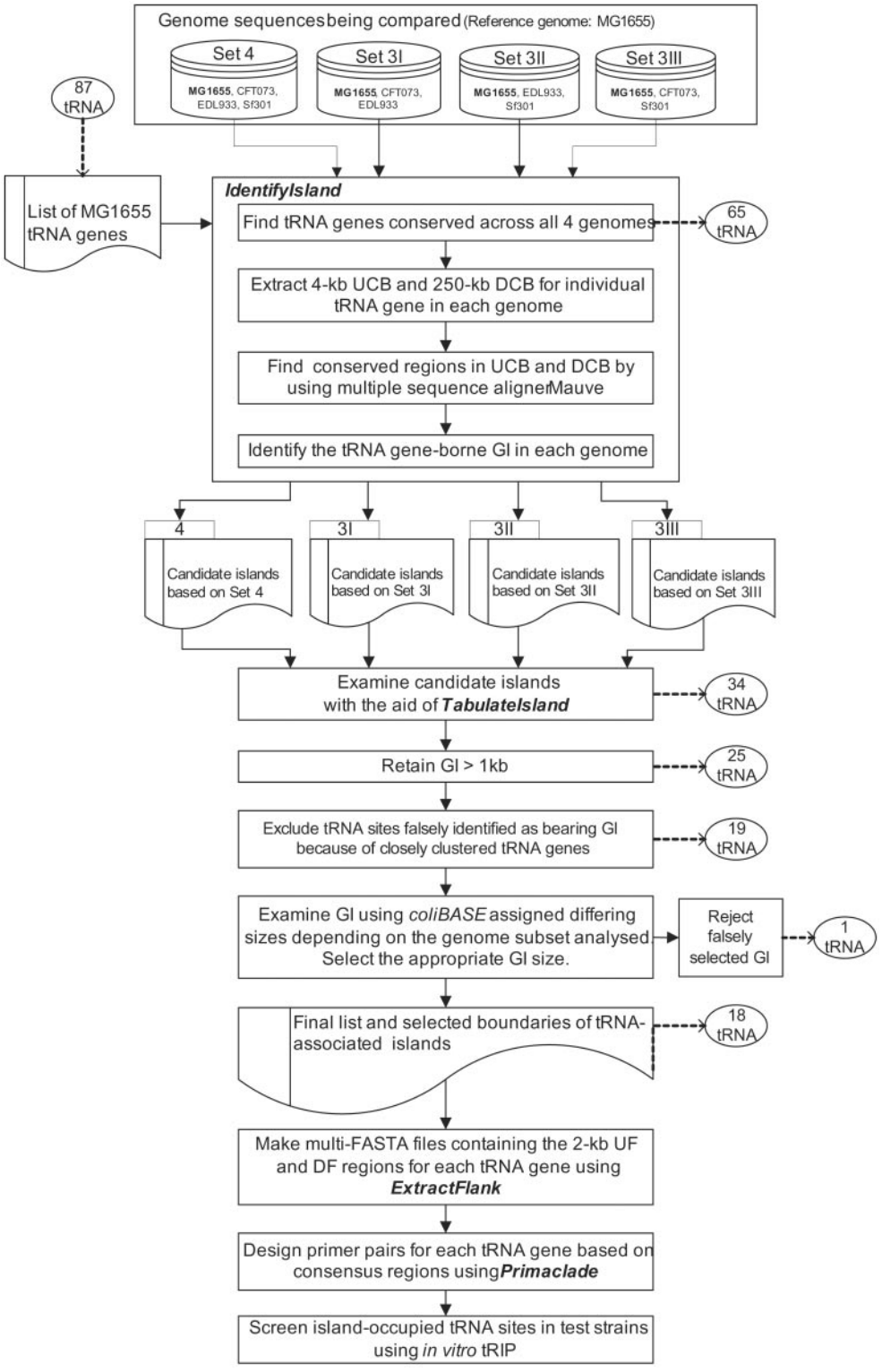

Figure 1. Flowchart depicting the tRNAcc high-throughput strategy developed and used to analyse the contents and contexts of tRNA genes in sequenced *E.coli* and *Shigella* genomes. Four stand-alone tools, indicated in bold italic font in the figure, were employed to identify islands (IdentifyIsland, TabulateIsland) and design primers (ExtractFlank, Primaclade) corresponding to the conserved upstream and downstream flanking regions of each tRNA site to be interrogated. See Table 1 for a summary of the programs features. In this study, four complete genomes were compared by the tRNAcc method: *E.coli* K-12 MG1655, *E.coli* UPEC CFT073, *E.coli* O157:H7 EDL933 and *S.flexneri* 2a Sf301. Four distinct genome subsets were analysed with the MG1655 genome being used as the reference template in each case. The numbers in the ovals above the word 'tRNA' indicate the number of tRNA genes still being considered at each stage in the analysis. The following abbreviations were used: UCB, upstream chromosomal block; DCB, downstream chromosomal block; GI, genomic island; UF, 2 kb upstream conserved flank; DF, 2 kb downstream conserved flank.

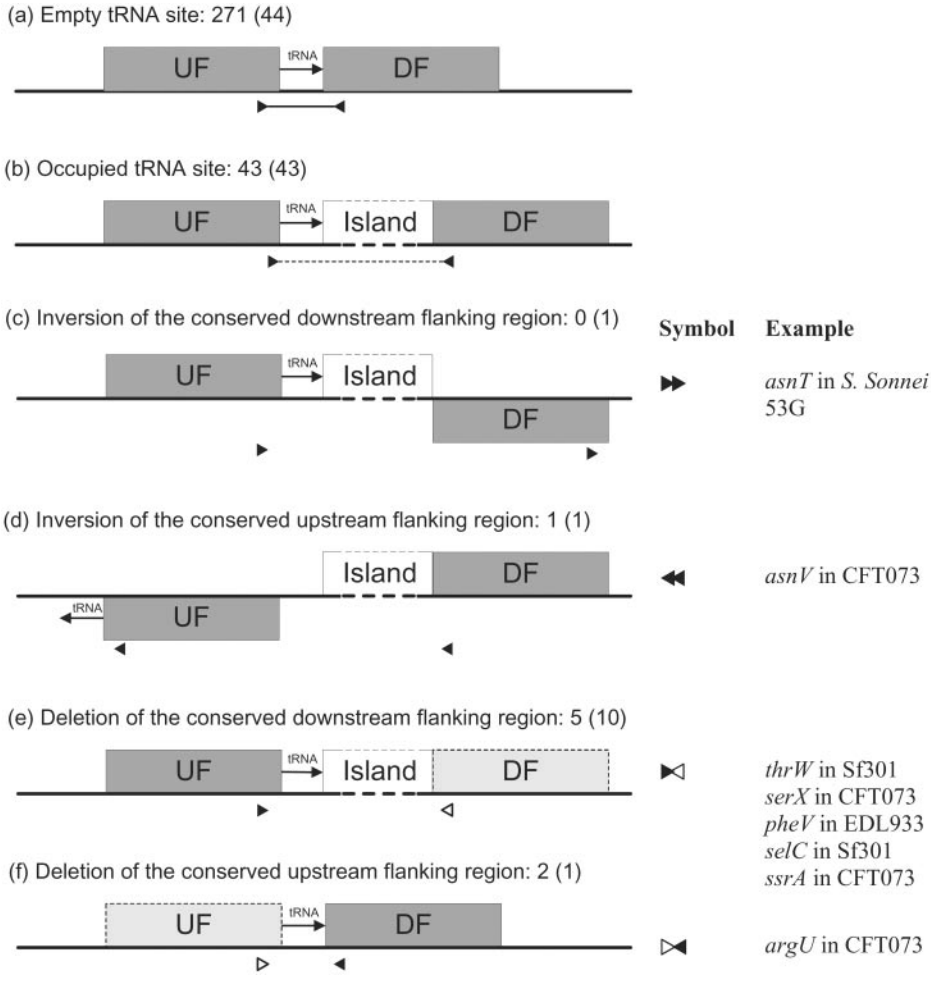

Figure 2. Schematic representation of a range of hypothetical tRNA site configurations present in the four complete genomes (MG1655, CFT073, EDL933 and Sf301) (**a**–**f**). The conserved UF and DF regions flanking tRNA genes are shown as dark grey filled boxes. UF and DF boxes drawn below the line indicate inversions with respect to the reference template MG1655 (**c** and **d**). The UF and DF boxes shown in pale grey with a broken outline represent deletions with respect to MG1655 (**e** and **f**). Genomic islands, where present, are indicated as broken boxes to emphasize the relatively large size of these regions. Arrowheads shown below each sub-figure indicate the location and orientation of primers specific to the UF and DF regions. Hollow arrowheads indicate the absence of matching complementary sequence. The solid line between the arrowheads shown in (**a**) indicates a likely successful *in vitro* PCR amplification; while the dotted line in (**b**) indicates a successful e-PCR-based 'amplification' that would typically yield a product of size far in excess of that that could be generated through standard *in vitro* PCR. The numbers shown above each configuration after the colon symbol represent the number of examples observed in the four genomes tested based on the 87 MG1655 tRNA genes and the total complement of orthologues present in the other three genomes (Table 2). The numbers of examples observed in the five unpublished genomes (*E.coli* EAEC O42, EPEC E2348/69, ETEC E24377A, *E.coli* HS and *S.sonnei* 53G), with respect to the subset of 20 tRNA genes only (Supplementary Table S3), are shown in parentheses. The symbols shown alongside the drawings are used in Table 2 and Supplementary Table S4 to highlight tRNA loci affected by inversions and/or deletions. Examples of the various atypical configurations observed in the four genomes are shown to the right. The figure is not drawn to scale.

Table 1. Stand-alone tools developed and used for high throughput analyses of the contents and contexts of tRNA genes in bacterial genomes

Software tool[a]	Description	Reference
Island identification		
IdentifyIsland	Identify putative islands based on conserved flanking blocks recognized by the multiple aligner Mauve 1.2.2 (10)	This work
TabulateIsland	Tabulate the identified islands following analysis of different subsets of genomes	This work
LocateHotspots	Locate proposed hotspots in non-annotated chromosomal sequences using BLASTN-based searches	This work
Primer design		
ExtractFlank	Generate multi-FASTA files containing the upstream or downstream flanking regions for the identified islands	This work
Primaclade	Design conserved PCR primers for the upstream or downstream flanking regions found in multiple bacterial genomes being compared. This program is available at http://www.umsl.edu/services/kellogg/primaclade.html	(14)
Island analysis		
DNAnalyser	Calculate the GC content and dinucleotide bias of identified islands, and the negative cumulative GC profile of genomes	This work
GenomeSubtractor	High-throughput BLASTN-based comparison of CDS sequences against test genomes to identify strain-specific CDS based on the level of nucleotide similarity	This work

[a]These programs can also be used for the generic identification and preliminary characterization of putative genomic islands located at other user-specified hotspots and for the analysis of cognate flanking sequences.

Determination of signatures of foreign acquisition

Strain-specific CDS were identified using a BLASTN-based procedure and a homology score described by Fukiya *et al*. (18). This *H*-value (18) was based on the length of match and degree of identity. For each query, the *H*-value was calculated as follows: $H = i \times l_m / l_q$, where i was the level of identity of the region with the highest Bit score expressed as a frequency of between 0 and 1, l_m the length of the highest scoring matching sequence (including gaps), and l_q the query length. If there were no matching sequences with a BLASTN *E*-value < 0.01, the *H*-value assigned to that query sequence was defined as zero (18). Therefore, *H* belonged to the set, $H \in [0,1]$. Details of the procedure used are included in the Supplementary Data.

DNAnalyser (Table 1) was written to facilitate computation of GC content and dinucleotide bias values for the identified tRNA-associated GIs. Dinucleotide bias analysis was performed using the method of Karlin (19). The genome-averaged dinucleotide relative abundance difference (δ^*) value (19) was obtained by using a 20 kb non-overlapping, sliding window along the entire genome sequence.

tRNAcc software package

We developed the tRNAcc 1.0 software package to facilitate the process of analysing the tRNA gene contents and contexts across multiple closely related bacterial genomes. tRNAcc comprises a suite of individual tools listed in Table 1. The software is divided into three sections by function: (i) identification of tRNA-associated GIs and their boundaries, (ii) design of primers specific to conserved UF and DF regions and (iii) analysis of putative islands for evidence of foreign origin. The codes written in C++, Perl or Bioperl (20) modules were tested under MS Windows 2000. tRNAcc is available as open-source software from the following URL: http://www.le.ac.uk/iii/staff/kr46/tRNAcc/index.htm.

RESULTS AND DISCUSSION

Identification of genomic islands in published *E.coli* and *Shigella* genomes with tRNAcc

The tRNAcc method was used for high-throughput interrogation of tRNA sites in four published *E.coli* and *Shigella* genomes: *E.coli* K-12 MG1655, *E.coli* UPEC CFT073, *E.coli* O157:H7 EDL933 and *S.flexneri* 2a Sf301. All 86 tRNA genes and the sole tmRNA gene in MG1655 were screened along with 241 orthologues in the remaining three genomes. A total of 49 GIs mapping to 18 tRNA genes and spanning a total length of 1666 kb were detected; 10–15 sites were found to be occupied per genome. Two GIs, missed by tRNAcc but detected by Islander (21), are also included in our analysis for completeness (Tables 2 and 3).

Key features of the 51 tRNA-associated GIs strongly supported a foreign origin. These included the possession of large numbers of strain-specific CDS, anomalous GC contents and significant dinucleotide biases (Supplementary Table S1). Using a conservative *H*-value cut-off of <0.42 (Supplementary Figure S3), where a *H*-value = 0 indicates no significant match and a *H*-value = 1 reflects 100% DNA sequence identity over the full length of the CDS, 988 CDS present within these 51 GIs are unique to a single strain, while a further 437 CDS are only common to two strains. Less than 8% (147/1783) of the encoded CDS possess homologues in all four genomes. The *H*-value distribution plot shown in Figure 3 highlights the restricted strain distribution of the majority of CDS borne on the identified GIs as 3824 of the 5349 calculated *H*-values ranged between 0 and 0.2, clearly demonstrating that most CDS exhibited little or no similarity to CDS in the other three strains (see Supplementary Figure S1 for additional *H*-value plots). Furthermore, a significant proportion of the total numbers of CDS unique to each strain were shown to lie within these 51 GIs. However, there was considerable variation between strains with between 26% of Sf301 strain-specific CDS and 52% of CFT073 unique CDS lying within 10 and 15 tRNA-associated islands, respectively. Clearly, these data emphasize that many strain-specific CDS lie within islands mapping to non-tRNA loci. This is entirely consistent with the recent findings of Chiapello *et al*. (22) who identified in excess of 800 chromosomal strain-specific 'loops' or GIs in each of four *E.coli* strains investigated. These loops spanned a very wide range of sizes and mapped to many distinct loci.

As horizontally acquired DNA often has a distinct CG base composition, the negative form of the cumulative GC profile of Zhang and colleagues (23) was used to visualize the locations of identified islands within the contexts of complete genomes. With these plots GC-rich regions appear as sharp upward swings while AT-abundant segments show up as abrupt downward deflections. As an example the plot for EDL933 is shown in Figure 4. Identified GIs are shown in green with cognate tRNA sites indicated as blue spots. Consistent with their anomalous GC contents, tRNA-associated GIs principally mapped to regions of sharp transitions in this plot (see Supplementary Figure S4 for additional negative cumulative GC plots).

The 2 kb upstream (UF) and downstream (DF) flanking regions corresponding to the 20 tRNA sites shown to be occupied in one or more of the 4 strains are highly conserved (Table 2); only 5 DF and 2 UF segments are missing out of a potential total of 160 (80 UF and 80 DF) segments. These chromosomal backbone segments were identified following MAUVE-based multi-genome alignments to detect conserved collinear genomic blocks of >0.5 kb. Furthermore, with the exception of the seven DF and UF segments mentioned above, a minimum of 0.3 kb of the 2 kb segments were common to all four strains based on a >90% ClustalW-calculated identity cut-off. All instances of deletions and/or inversions of flanking segments are indicated in Table 2. Conservation of the immediate contexts of tRNA genes also extended to sites that were shown to be unoccupied in all four strains. Forty-five of the sixty-seven remaining MG1655 tRNA genes possessed counterparts in the other three genomes. All 90 flanking regions associated with these sites were conserved across the four genomes. Even the remaining 22 tRNA genes that lacked an orthologue in one or more genomes possessed conserved flanks with there being no instances of UF or DF deletions among the corresponding 62 sites in the four genomes.

Comparison of tRNAcc and Islander for the detection of tRNA-borne islands

Mantri and Williams (21) have recently used a novel algorithm, Islander, to scan bacterial genomes for tRNA-associated

Table 2. Sizes of genomic islands identified by the tRNAcc method that map to tRNA sites in four sequenced *E.coli* and *Shigella* genomes[a]

No.	tRNA gene	*E.coli* K-12 MG1655	*E.coli* UPEC CFT073	*E.coli* O157:H7 EDL933	*S.flexneri* 2a Sf301	Identity of 2.0 kb UF[b]	Identity of 2.0 kb DF[b]
1	*aspV*	**2.4**	**100**	**36.9**	**57.7**	96%	96%
2	*thrW*	**39.9** [34.3]	**7.7**	**35.2** [10.6]	**21.6** ▶◁	96%	93%
3	*serW*	0.3	0.3	**87.8** [87.6]	0.3	98%	97%
4	*serT*	0	0	**45.2** [45.2]	0	97%	97%
5	*serX*	0.5	**113.8** [113.5] ▶◁	**87.5** [87.6]	0	97%	81% (1.5 kb, 97%)
6	*tyrT*	1.0	**7.3**	0.8	0.4	98%	96%
7	*leuZ*	0	0	**21.1** [21.1]	0	98%	94%
8	*serU*	**1.4**	**23.2** [22.5]	**46.6** [45.2]	**22.3** [21.6]	71% (1.1 kb, 98%)	95%
9	*asnT*	**10.1**	**37.2**	**11.0**	**4.5**	62% (1.1 kb, 98%)	96%
10	*argW*	**12.6** [10.2]	**14.6**	**14.1** [8.6]	**5.3**	96%	94%
11	*lysV*	**4.6**	**4.4**	0.5	0.6	92%	97%
12	*metV*	0	**32.7**	0	0[c]	83% (1.5 kb, 96%)	93%
13	*glyU*	**11.7**	0.1	**27.7**	**10.0** [7.6]	96%	96%
14	*pheV*	**9.1**	**127.9** [104.6]	**23.5** ▶◁	**55.1** [46.7]	96%	92%
15	*selC*	**1.9**	**68.6**	**43.7** [5.1]	**29.9** ▶◁[f]	97%	95%
16	*pheU*	0	**52.1** [52.1]	0	0	97%	51% (0.3 kb, 94%)
17	*leuX*	**40.1**	**15.9**	**44.4** [10.2]	**7.5**	72% (1.2 kb, 93%)	97%
18	*ssrA*	**29.6**	**48.4** [48.4] ▶◁	**29.2**	**3.7**	94%	95%
19	*asnV*[d]	0	**54.4** [54.4] ◀◀	0	0	96%	95%
20	*argU*[e]	**21.3** [21.3]	0 ▷◀	0	0 ▷◀	55% (0.4 kb, 96%)	98%

[a]Island sizes are shown to the nearest 0.1 kb. Predicted insertions at these loci of >1 kb in size are highlighted in bold type to indicate putative genomic islands. The sizes of the 21 tRNA-borne islands identified by Islander (21) are shown in square brackets. Details of the arrowhead symbols used are explained in Figure 2.
[b]The identities of the 2 kb upstream flanking regions (UF) and the 2 kb downstream flanking regions (DF) across all the four genomes are calculated by the multiple alignment program ClustalW 1.82 (15). Note that genomes exhibiting deletions of particular flanking regions were excluded from the corresponding multiple sequence alignments. If the identity of the complete 2 kb flanking sequences was <90%, a highly conserved region within the UF or DF region was further investigated. The sizes and identities of these shorter highly conserved regions present within the 2 kb segments themselves are shown in parentheses.
[c]As the *metV* gene was not annotated within the Sf301 genome, the Sf301 data shown relate to the sequence-identical *metW* gene that is immediately adjacent.
[d]A 4.3 kb DNA fragment with termini corresponding to the 3′ ends of the inversely orientated, sequence identical *asnW* and *asnV* genes was inverted in CFT073, with respect to the other three genomes. Consequently, the 54.4 kb island identified by Islander as being integrated into the gene annotated as *asnW* in CFT073 was missed by tRNAcc analysis. In this study, we have re-labelled this latter CFT073 gene as '*asnV*' to maintain the synteny of this tRNA gene and its cognate DF sequence in all four genomes. Hence, the CFT073 NCBI annotated *asnV* gene was now known as '*asnW*'. The upstream identities for this locus were calculated using the 2 kb upstream flanking regions of *asnV* in MG1655, EDL933 and Sf301 and the 2 kb UF fragment corresponding to the newly termed *asnW* gene in CFT073. The downstream identities were computed based on the downstream flanking regions of *asnV* in MG1655, EDL933 and Sf301 and the conserved downstream flanking region lying distal to the 54.4 kb island in CFT073.
[e]Owing to a 6.9 kb deletion of the upstream flanking region in CFT073 and a 4.1 kb deletion of the upstream region in Sf301, with respect to MG1655 and EDL933, no island was identified in the *argU* sites of the four strains using the tRNAcc method. As Islander had identified a 21.3 kb *argU*-borne island in MG1655, tRNAcc was re-run using only the two genomes (MG1655 and EDL933), resulting in successful identification of the MG1655 island.
[f]The secondary conserved downstream flanking region was inverted with respect to MG1655.

islands. As of July 5, 2005, the Islander database (www.indiana.edu/~islander) lists 141 GIs dispersed among 106 completely sequenced bacterial genomes (21). For a given bacterial genome, the Islander algorithm detects candidate islands adjacent to tRNA sites that are bounded by direct repeats and contain an integrase gene homologue. These candidate GIs are then further scrutinized through a series of filters to select the final set of archetypal integrative islands. The Islander database listed 21 tRNA-borne islands in MG1655, CFT073, EDL933 and Sf301. In contrast, the tRNAcc-facilitated multi-genome comparative approach identified more than twice as many tRNA-associated GIs in these same four genomes. However, it should be borne in mind that Islander was designed primarily to identify additional instances of tRNA-associated integrase site-specificity and consequently only identified GIs bearing integrase gene homologues. Nineteen islands, with equivalent tRNA-proximal boundaries, were identified by both methods. However, only 10 of these GIs had common distal termini that mapped to within a 1 kb window on the cognate genome by both methods. The remaining nine GIs were found to possess distal extensions ranging from 2.4 to 38.6 kb by the tRNAcc method (Supplementary Table S2). Most significantly, Islander missed 30 tRNAcc-identified entities. Given their novel CDS content and associated signatures of foreign origin (Supplementary Table S1), these elements are likely to be true genomic islands that have arisen following DNA acquisition events.

The *asnW*-borne GI in CFT073 and the *argU*-associated prophage in MG1655 were not identified by tRNAcc but detected by Islander. Subsequent *coli*BASE-facilitated examination of these tRNA sites in all four genomes revealed the basis of the tRNAcc false-negative results. The 54.4 kb *asnW*-associated island was missed due to an inversion of a 4.3 kb DNA fragment between the sequence-identical, indirectly orientated *asnW* and *asnV* genes in CFT073. This led to an inversion and translocation of the upstream flanks associated with the CFT073 *asnV* and *asnW* genes and prevented tRNAcc-mediated recognition of the corresponding conserved flanks. In order to maintain synteny and because of the arbitrary nature of the original designation, we have re-labelled the CFT073 *asnW* gene as '*asnV*' and the *asnV* as '*asnW*' for our analysis (Table 2). The *argU* island, on the other hand, was undetected because of deletions of the conserved upstream flank in CFT073 (6.9 kb deletion) and Sf301 (4.1 kb deletion). We had attempted to take account of deletions in isolated strains by analysing subsets of three genomes, but had not examined individual pairs of genomes. As expected, when only the two genomes that harboured conserved *argU* flanks

Table 3. Summary of tRNA-borne genomic islands in four published *E.coli* and *Shigella* genomes identified with tRNAcc

Strain	Annotated genome Chromosome size (kb)	No. of annotated CDS	No. of strain-specific CDS[a]	No. of horizontally transferred CDS[b]	Islands identified No. of islands	Total size of islands (kb)	No. annotated island-borne CDS	No. of island-borne strain-specific CDS[c]	No. of island-borne horizontally transferred CDS[d]
E.coli K-12 MG1655	4640	4242	234	N/A	12	184.9	188	89 (47.3%)	N/A
E.coli UPEC CFT073	5231	5379	884	N/A	15	708.8	742	458 (61.7%)	N/A
E.coli O157:H7 EDL933	5528	5324	916	593	14	554.1	621	380 (61.2%)	245 (39.5%)
S.flexneri 2a Sf301	4607	4180	236	N/A	10	217.7	232	61 (26.3%)	N/A

[a]Identification of CDS as strain-specific among the genomes compared is based on the level of nucleotide similarity. See Supplementary Data for details.
[b]The horizontally transferred CDS in EDL933 were taken from Horizontal Gene Transfer Database (HGT-DB) (http://www.fut.es/~debb/HGT/) (28). The horizontally transferred CDS for CFT073, Sf301 and the updated version of the MG1655 genome are currently not available (N/A) in the HGT-DB.
[c]The percentages of island-borne CDS that are defined as strain-specific in this study are shown in parentheses.
[d]The percentages of island-borne CDS that are listed in the HGT-DB are shown in parentheses.

(MG1655 and EDL933) were analysed by tRNAcc, the 21.3 kb MG1655 island that had been identified by Islander was found. At the same time, the EDL933 *argU* site was shown to be empty. Manual inspection of the *asnV* and *argU* sites in the four genomes and pairwise analysis of the remaining two-genome subsets yielded no other GIs.

Archetypal genomic islands are thought to have arisen following site-specific integration of precursor elements into tRNA or tmRNA genes. Resulting islands are flanked by short direct repeats (DR) that match the 3′ ends of target genes [see refs (2) and (24) for excellent reviews]. In the Islander database, island endpoints are defined as corresponding to the extremities of likely DR sequences that are identified using BLASTN and a series of logic filters (21). In this study, GIs were defined as variable DNA segments across the genomes being compared that lay between the 3′ termini of tRNA genes and the proximal ends of cognate conserved downstream flanking regions. The *pheV*-borne islands identified by tRNAcc and Islander highlight this point (Figure 5). The 104.6 and 46.7 kb DR-flanked islands in CFT073 and Sf301, respectively, were defined by Islander. In contrast, the tRNAcc-defined entity at the *pheV* locus in Sf301 possessed an 8.4 kb tRNA-distal extension, resulting in an entirely novel and distinct boundary. The GC content of this 8.4 kb segment was 44.2%, markedly different from that of the 46.7 kb Islander-defined entity (49.1%) or the Sf301 genome itself (50.9%). In addition, this segment exhibited no significant similarity with the genomes of MG1655, CFT073 and EDL933 at a DNA level, and contained 11 annotated CDS that were predicted to code for transposases, fimbrial proteins and several hypothetical proteins. Similarly, tRNAcc identified a larger *pheV*-borne island in CFT073 than did Islander. The additional segment, measuring 23.3 kb, once again mapped to the tRNA-distal end of the identified structure. Its DNA sequence showed no significant similarity with the other three genomes. The 18 CFT073-specific CDS contained within this region are predicted to code for diverse proteins including enzymes, a transposase, at least two transport proteins and several hypothetical proteins. These data are strong evidence that the 8.4 kb (Sf301) and 23.3 kb (CFT073) extensions are of foreign origins and support the notion that these regions should be regarded as part of the *pheV*-borne islands present in these strains. We hypothesize that these GIs represent composite elements that have arisen following sequential horizontal DNA acquisition events followed by reorganization and rationalization of tandem entities.

As tRNAcc and Islander are based on different principles, we suggest that combined application of these complementary strategies will permit the ready detection of the vast majority of tRNA-borne islands. Other generic GI discovery algorithms that are not confined to interrogating tRNA loci are also of major value as these approaches identify many other non-tRNA-associated elements and may provide evidence of alternative favoured integration sites. The IslandPath strategy described by Hsiao *et al*. (25) identifies putative GIs by profiling GC contents and dinucleotide composition of individual CDS and/or clusters of CDS and flagging putative islands based on user-defined deviations from genome-wide mean values. This approach may even detect horizontally acquired islands that are shared by the full set of genomes being compared, provided these entities originated from organisms with

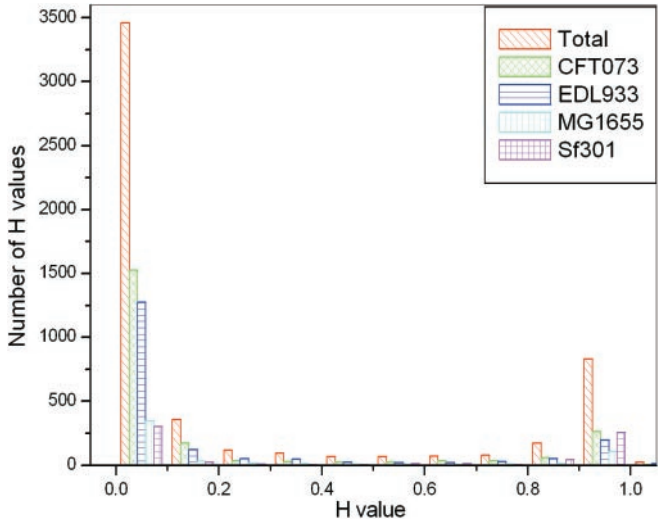

Figure 3. The distribution of *H*-values corresponding to the island-borne CDS in *E.coli* K-12 MG1655, *E.coli* UPEC CFT073, *E.coli* O157:H7 EDL933 and *S.flexneri* 2a Sf301 identified by tRNAcc and/or Islander methods. This homology score had been proposed by Fukiya *et al*. (18) and reflected the degree of similarity between the matching reference genome sequence and the CDS itself in terms of the length of match and the degree of identity at a DNA level. See Supplementary Data for details. Red, green, blue, cyan and magenta bars represent total CDS, CFT073 CDS, EDL933 CDS, MG1655 CDS and Sf301 CDS, respectively. Note that each CDS in a given genome has three *H*-values that were obtained by BLASTN searches against the other three genomes in turn.

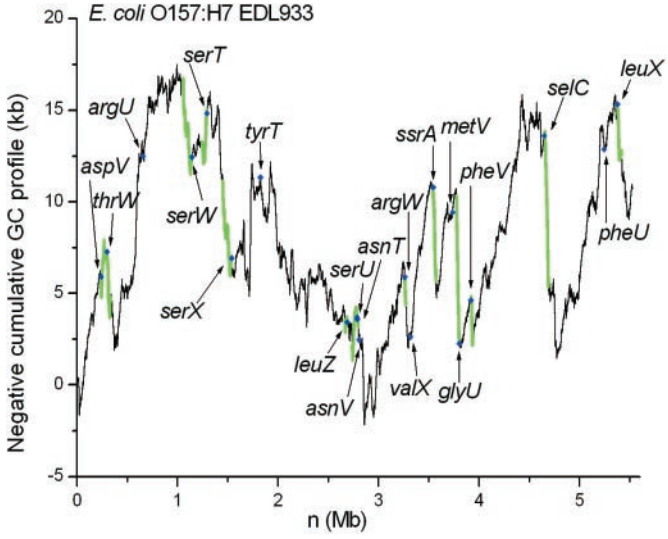

Figure 4. Negative cumulative GC profile (23) highlighting the genomic context of islands identified by tRNAcc in *E.coli* O157:H7 EDL933. A sharp upward spike in the negative cumulative GC profile indicates a relatively sharp increase in GC content, whereas an abrupt fall indicates a relatively sharp decrease in GC content. The locations of tRNA-associated genomic islands are shown in green and the tRNA and tmRNA genes are represented as blue diamonds. Details of this plot are specified in the supplementary material.

distinct DNA signatures. Similar to tRNAcc, the very recently described MOSAIC algorithm also utilizes a multi-genome comparative approach (22). However, unlike our approach it scans entire genomes for exact matching sequences and proceeds to segment chromosomes into core backbone sequences and strain-specific loops. It is a very powerful tool that performs an *in silico* version of DNA heteroduplex analysis, recognizing strain-specific sequences as 'non-hybridizing' loops. The MOSAIC strategy has just been applied to analyse genomes of 13 bacterial species, yielding an abundance of data (22). Clearly, the combined application of multiple strategies could be of value in many instances to maximize identification of GIs and ensure accurate delineation of their boundaries.

In silico PCR screening of unpublished *E.coli* and *Shigella* genomes

Given the high level of conservation of UF and DF segments, we used a simple *in silico* PCR approach to interrogate the 20 identified tRNA hotspots for the presence or absence of an integrated element. We named the strategy 'tRIP' for tRNA site interrogation for pathogenicity islands, prophages and other GIs. Specific primer pairs were designed based on conserved UF and DF segments. Additionally, to account for instances of deletions involving particular DF segments, alternate primers were designed to correspond to secondary downstream flanking regions (DF′) for the *thrW*, *pheV* and *ssrA* loci. *In silico* tRIP facilitated by the e-PCR tool of Schuler (17) was then employed to re-examine the four genomes that had previously been analysed by tRNAcc and three other published *E.coli* and *Shigella* genomes (*E.coli* K-12 W3110, *E.coli* O157:H7 Sakai and *S.flexneri* 2a 2457T) that had not been used to train the tRNAcc algorithm. Complete or near complete genomes of enteroaggregative *E.coli* (EAEC) O42, enteropathogenic *E.coli* (EPEC) E2348/69, enterotoxigenic *E.coli* (ETEC) E24377A, *E.coli* O9 HS and *Shigella sonnei* 53G were also analysed by *in silico* tRIP to fully exploit this island-finding strategy. The sizes of the resulting virtual amplicons are shown in Supplementary Table S3. The occupancy of tRNA sites in test strains was then inferred by comparison of amplicon sizes obtained with those corresponding to equivalent empty sites. If the predicted sizes differed by <1 kb the tRNA site in the test strain was classified as empty; otherwise the site was categorized as occupied. The size of the putative associated genomic island was estimated based on this discrepancy.

When the O42 genome was subjected to *in silico* tRIP to investigate the 20 tRNA sites, 18 virtual PCR products were obtained. The only negative e-PCR results were for *serX* and *argW*, sites that were subsequently found to be associated with DF deletions in O42 (Supplementary Table S3). Ten loci were predicted to harbour large GIs. These islands, spanning a total of about 440 kb, carried 485 putative CDS, 205 of which were EAEC O42-specific with respect to MG1655, CFT073, EDL933 and Sf301 (Table 4). Furthermore, 42% of the O42-specific CDS were located within these 10 chromosomal segments alone. Details of the identified putative GIs are shown in Supplementary Table S5 and their locations within the genomic context indicated in Supplementary Figure S4.

As with previous tRNAcc-identified GIs, these *in silico* tRIP identified entities also exhibited signatures consistent with horizontal acquisition. Space does not permit a full enumeration of the potential functions of the genes and islands discovered in this analysis. Links from the URL http://colibase.bham.ac.uk/cgi-bin/tRNAcc.cgi to each gene cluster in the online genomics resource *coli*BASE (12) provide the reader

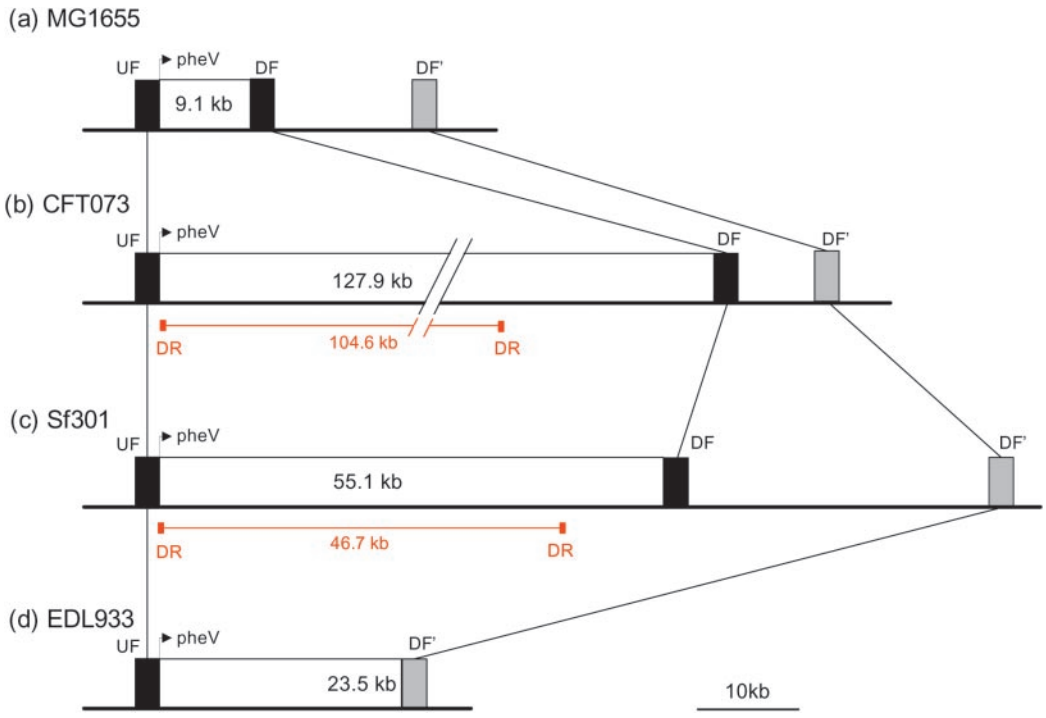

Figure 5. The four *pheV*-borne islands in MG1655 (**a**), CFT073 (**b**), Sf301 (**c**) and EDL933 (**d**) genomes identified by the tRNAcc method. The 9.1 kb island in MG1655, 127.9 kb island in CFT073 and 55.1 kb island in Sf301 are flanked by conserved upstream (UF) and downstream (DF) backbone segments. However, the DF region, common to the other three genomes, is absent in EDL933. Instead, the first instance of a conserved chromosomal block common to the other genomes occurs 23.5 kb downstream of the EDL933 *pheV* gene. This secondary conserved block has been designated as DF′. Matching 2 kb flanking regions are represented as connected blocks. In this study, genomic island-like regions were defined as anomalous segments between the 3′ end of tRNA genes and the 5′ end of the conserved downstream flank. Consequently, the tRNAcc-identified GIs in MG1655, CFT073 and Sf301 lay between the *pheV* and DF loci, while that in EDL933 was defined as the segment between the *pheV* gene and the proximal boundary of the DF′ conserved segment. The Islander-defined 104.6 and 46.7 kb islands at the *pheV* locus in CFT073 and Sf301, respectively, are shown as red lines flanked by DR sequences (red rectangles).

Table 4. Summary of putative islands in the five unpublished *E.coli* and *Shigella* genomes identified by *in silico* tRIP[a]

Strain	Unpublished genome[a] Chromosome size (kb)	No. of CDS predicted	No. of strain- specific CDS[c]	Islands identified[b] No. of islands	Total size of islands (kb)	No. of islands-borne predicted CDS	No. of island-borne strain-specific CDS[d]
E.coli EAEC O42	5242	4899	490	10	440.0	485	205 (42.3%)
E.coli EPEC E2348/69	5075	5313	606	10	208.0	241	115 (41.7%)
E.coli ETEC E24377A	4980	4254	261	11	411.1	283	120 (42.2%)
E.coli O9 HS	4644	3989	146	9	140.6	79	29 (36.7%)
S.sonnei 53G	4989	5118	344	6	228.1	225	16 (7.1%)

[a]The two unpublished genomes *E.coli* ETEC E24377A and *E.coli* O9 HS sequenced by TIGR were downloaded from NCBI. The other three unpublished chromosomal sequences of *E.coli* EAEC O42, *E.coli* EPEC E2348/69 and *S.sonnei* 53G were downloaded from the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/pathogens/Escherichia_Shigella/). The putative CDS of *E.coli* EPEC E2348/69 and *S.sonnei* 53G were identified using GLIMMER 2.13 (29).
[b]Putative islands were identified when *in silico* tRIP PCR amplicons were at least 1 kb greater in size than those corresponding to empty matching sites. These islands are highlighted in bold type in Table S3 in the supplementary materials.
[c]Determination of strain-specific CDS among the genomes compared was based on the level of nucleotide similarity with respect to MG1655, CFT073, EDL933 and Sf301. See the supplementary materials for details.
[d]The percentages of island-borne CDS that were defined as strain-specific in this study are shown in parentheses.

with a starting point for fuller analyses. However, even a cursory glance at the results of homology searches on islands from the unfinished genomes reveals interesting biological vignettes. For example, in EAEC *E.coli* O42, the *aspV* and *pheU* GIs are related to a large gene cluster conserved in many different pathogens/symbionts (26), which may encode a novel secretion system. The *serU* islands in EAEC O42 and in ETEC E24377A appear to be lambdoid prophages, while the *serU* island in the commensal strain HS represents a mu-like phage. The *glyU* island from O42 represents ETT2, a recently described type III secretion gene cluster (27), while components of type V or type II secretion systems occur on several other islands. As a striking example the 34 kb *ssrA* island in EAEC O42 carried at least 30 O42-specific genes among the 34 putative CDS encoded. This island possessed an integrase gene homologue but lacked recognizable flanking DR sequences. The results of a PSI-BLAST similarity search for the 34 predicted proteins are shown in Supplementary

Table S6. Remarkably, 14 CDS only yielded matches with other hypothetical proteins and a further 12 CDS had no significant homology at amino acid level with sequences in the databases. The *ssrA* GI from O42 appears to be derived from an integrated plasmid. O42 is an important agent of acute and persistent diarrhoea in children in developing countries; given parallels with other diarrhoeagenic *E.coli*, it will be important to investigate possible roles in pathogenesis for the *ssrA*-associated and/or other identified GIs.

A total of 46 GIs, mapping to 15 tRNA genes and spanning a total length of 1489 kb, were identified by *in silico* tRIP in the five unpublished *E.coli* and *Shigella* genomes analysed (Table 4). Initial analyses of these regions support their assignment as true GIs (Supplementary Table S5). Thus, it is clear that tRNAcc has captured some of the most dynamic components of the *E.coli* gene pool.

CONCLUSION

In this study, we have undertaken a systematic examination of tmRNA and tRNA sites in *E.coli* and *Shigella* genomes. Our results confirm earlier reports that these genes serve as integration hotspots for a diverse repertoire of foreign DNA. Furthermore, core chromosomal sequences immediately flanking tRNA genes in all four genomes investigated were shown to be highly conserved in the vast majority of instances. This permitted use of a simple *in silico* PCR approach to identify putative GIs in both complete genomes and incomplete genomes that had yet to even come off the 'sequencing pipeline'. More importantly, our approach will identify suitable specific primers that correspond to the conserved UF and DF segments and that can readily be used to interrogate tRNA loci in test strains for the presence or absence of GIs. We are currently undertaking an experimental tRIP study with a collection of *E.coli* and *Shigella* strains and have already identified numerous tRNA-associated GIs using this method (K. Rajakumar, J. Lonnen, A.B. Thani and H.-Y. Ou, unpublished data). We have also utilized the tRNAcc and tRIP algorithms to investigate selected non-tRNA loci that had been identified as putative integration hotspots following analysis of microarray-defined highly variable genomic regions in a collection of *E.coli* strains (Supplementary Table S7; see the brief description in the Supplementary Data). The tRNAcc and tRIP methodologies are also applicable to other bacterial species provided at least two distinct strains have been sequenced. To emphasize this point, we have performed analyses using four complete genomes of *Salmonella enterica* and the complete and near-complete genome sequences of *Pseudomonas aeruginosa* strains PAO1 and PA14, respectively. The putative GIs identified by the tRNAcc method are listed in Supplementary Tables S8, S9 and S10. However, it should be emphasized that the accuracy of the predictions made and the power of the tRNAcc method to identify primer sequences for wet-science based tRIP exploration increases significantly if subsets comprising permutations of three or more available genomes are analysed. The bioinformatics tools developed in this study will facilitate the early and high-throughput discovery of GIs through increased exploitation of emerging sequence data and PCR-based profiling of large collections of bacterial strains.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Prof. Chun-Ting Zhang at Tianjin University for helpful discussions and critical reading of the manuscript. The authors are also grateful to Sarah Fandrich for support. The authors thank the Sanger Institute for their policy of making preliminary sequence data publicly available and acknowledge the use in this study of unpublished genome data corresponding to *E.coli* EAEC O42, *E.coli* EPEC E2348/69 and *S.sonnei* 53G. The authors also thank the Institute for Genomic Research for their policy of making preliminary sequence data publicly available and acknowledge the use in this study of unpublished genome data corresponding to *E.coli* ETEC E24377A and *E.coli* O9 HS. Permission to use unpublished *P.aeruginosa* PA14 sequence available via the MGH ParaBioSys resource is gratefully acknowledged. M.P. would like to thank the BBSRC for funding R.R.C. and the *coli*BASE project as part of University of Birmingham Exploiting Genomics project on *E.coli* (grant number EGA16107). This study was supported by a *medi*Search grant from The Leicestershire Medical Research Foundation, United Kingdom to K.R. and M.R.B. and by the BBSRC Core Strategic Grant to J.H. J.L. was supported by a BBSRC PhD studentship and A.B.T. by the University of Bahrain. Funding to pay the Open Access publication charges for this article was provided by *medi*Search.

Conflict of interest statement. None declared.

REFERENCES

1. Perna,N.T., Plunkett,G.,III, Burland,V., Mau,B., Glasner,J.D., Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.
2. Hacker,J. and Kaper,J.B. (2000) Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.*, **54**, 641–679.
3. Hou,Y.M. (1999) Transfer RNAs and pathogenicity islands. *Trends Biochem. Sci.*, **24**, 295–298.
4. Williams,K.P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.*, **30**, 866–875.
5. Simillion,C., Vandepoele,K. and Van de Peer,Y. (2004) Recent developments in computational approaches for uncovering genomic homology. *Bioessays*, **26**, 1225–1235.
6. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
7. Welch,R.A., Burland,V., Plunkett,G.,III, Redford,P., Roesch,P., Rasko,D., Buckles,E.L., Liou,S.R., Boutin,A., Hackett,J. *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 17020–17024.
8. Jin,Q., Yuan,Z., Xu,J., Wang,Y., Shen,Y., Lu,W., Wang,J., Liu,H., Yang,J., Yang,F. *et al.* (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res.*, **30**, 4432–4441.
9. Gueneau de Novoa,P. and Williams,K.P. (2004) The tmRNA website: reductive evolution of tmRNA in plastids and other endosymbionts. *Nucleic Acids Res.*, **32**, D104–D108.

10. Darling,A.C.E., Mau,B., Blattner,F.R. and Perna,N.T. (2004) Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Res*., **14**, 1394–1403.
11. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.
12. Chaudhuri,R.R., Khan,A.M. and Pallen,M.J. (2004) coliBASE: an online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics. *Nucleic Acids Res*., **32**, D296–D299.
13. Carver,T.J., Rutherford,K.M., Berriman,M., Rajandream,M.-A., Barrell,B.G. and Parkhill,J. (2005) ACT: the Artemis comparison tool. *Bioinformatics*, **21**, 3422–3423.
14. Gadberry,M.D., Malcomber,S.T., Doust,A.N. and Kellogg,E.A. (2005) Primaclade—a flexible tool to find conserved PCR primers across multiple species. *Bioinformatics*, **21**, 1263–1264.
15. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*., **31**, 3497–3500.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
17. Schuler,G.D. (1997) Sequence Mapping by Electronic PCR. *Genome Res*., **7**, 541–550.
18. Fukiya,S., Mizoguchi,H., Tobe,T. and Mori,H. (2004) Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* Strains revealed by comparative genomic Hybridization microarray. *J. Bacteriol*., **186**, 3911–3921.
19. Karlin,S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol*., **9**, 335–343.
20. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*., **12**, 1611–1618.
21. 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**, D55–D58.
22. Chiapello,H., Bourgait,I., Sourivong,F., Heuclin,G., Gendrault-Jacquemard,A., Petit,M.-A. and El Karoui,M. (2005) Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. *BMC Bioinformatics*, **6**, 171.
23. Zhang,R. and Zhang,C.T. (2004) A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. *Bioinformatics*, **20**, 612–622.
24. Dobrindt,U., Hochhut,B., Hentschel,U. and Hacker,J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nature Rev. Microbiol*., **2**, 414–424.
25. Hsiao,W., Wan,I., Jones,S.J. and Brinkman,F.S.L. (2003) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, **19**, 418–420.
26. Das,S. and Chaudhuri,K. (2003) Identification of a unique IAHP (IcmF associated homologous proteins) cluster in *Vibrio cholerae* and other proteobacteria through *in silico* analysis. *In Silico Biol*., **3**, 287–300.
27. Ren,C.P., Chaudhuri,R.R., Fivian,A., Bailey,C.M., Antonio,M., Barnes,W.M. and Pallen,M.J. (2004) The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli*, is present in the majority of strains but has undergone widespread mutational attrition. *J. Bacteriol*., **186**, 3547–3560.
28. Garcia-Vallve,S., Guzman,E., Montero,M.A. and Romeu,A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res*., **31**, 187–189.
29. Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res*., **26**, 544–548.