Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W97–W104***
doi:10.1093/nar/gkm380

MobilomeFINDER: web-based tools for *in silico* and experimental discovery of bacterial genomic islands

Hong-Yu Ou[1], Xinyi He[1], Ewan M. Harrison[2], Bridget R. Kulasekara[3], Ali Bin Thani[2], Aras Kadioglu[2], Stephen Lory[3], Jay C. D. Hinton[4], Michael R. Barer[2,5], Zixin Deng[1] and Kumar Rajakumar[2,5,*]

[1]Laboratory of Microbial Metabolism and School of Life Science & Biotechnology, Shanghai Jiaotong University, P. R. China, [2]Department of Infection, Immunity and Inflammation, Leicester Medical School, University of Leicester, Leicester LE1 9HN, UK, [3]Department of Microbiology and Molecular Genetics, Harvard Medical School, Boston, MA 02115, USA, [4]Molecular Microbiology Group, Institute of Food Research, Norwich Research Park, Norwich NR4 7UA and [5]Department of Clinical Microbiology, University Hospitals of Leicester NHS Trust, Leicester LE1 5WW, UK

Received January 30, 2007; Revised April 22, 2007; Accepted April 30, 2007

ABSTRACT

MobilomeFINDER (http://mml.sjtu.edu.cn/MobilomeFINDER) is an interactive online tool that facilitates bacterial genomic island or 'mobile genome' (mobilome) discovery; it integrates the ArrayOme and tRNAcc software packages. ArrayOme utilizes a microarray-derived comparative genomic hybridization input data set to generate 'inferred contigs' produced by merging adjacent genes classified as 'present'. Collectively these 'fragments' represent a hypothetical 'microarray-visualized genome (MVG)'. ArrayOme permits recognition of discordances between physical genome and MVG sizes, thereby enabling identification of strains rich in microarray-elusive novel genes. Individual tRNAcc tools facilitate automated identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites and other integration hotspots in closely related sequenced genomes. Accessory tools facilitate design of hotspot-flanking primers for *in silico* and/or wet-science-based interrogation of cognate loci in unsequenced strains and analysis of islands for features suggestive of foreign origins; island-specific and genome-contextual features are tabulated and represented in schematic and graphical forms. To date we have used MobilomeFINDER to analyse several *Enterobacteriaceae*, *Pseudomonas aeruginosa* and *Streptococcus suis* genomes. MobilomeFINDER enables high-throughput island identification and characterization through increased exploitation of emerging sequence data and PCR-based profiling of unsequenced test strains; subsequent targeted yeast recombination-based capture permits full-length sequencing and detailed functional studies of novel genomic islands.

INTRODUCTION

Comparative analyses of multiple bacterial genomes have revealed that some bacterial species possess an extremely plastic genome (1,2). Horizontal gene transfer events have led to the integration of foreign DNA segments into species-specific syntenic backbones, often within tRNA and tmRNA gene sites (3,4). This 'optional' genomic repertoire, termed 'mobilome' (mobile genome) (5,6), which can vary considerably between members of the same bacterial species, includes episomal plasmids, transposons, integrons, prophages and a growing list of pathogenicity islands (PAIs) or genomic islands (GIs) (2,7). Many *in silico* approaches for detecting mobile genetic elements in sequenced bacterial genomes have been developed recently. These include methods based on anomalous codon usage, G+C content, dinucleotide bias, and amino acid usage patterns (8–11), identification of archetypal GI-specific features (12) and comparative genomics (3,13); for excellent reviews see (1,5,14).

The main barrier to high-throughput prospecting of the mobilome has been a paucity of bacterial genome sequence information, and so it has become a major challenge to develop rapid and cost-effective approaches to discover strain-specific DNA that is dispersed amongst hundreds of members of bacterial species of principal interest to man (1). Recently we have developed a high-throughput strategy, dubbed MobilomeFINDER, for experimental

*To whom correspondence should be addressed. Tel: +44 116 2231498; Fax: +44 116 2525030; Email: kr46@le.ac.uk
Correspondence may also be addressed to Zixin Deng. Tel: +86 21 62933404; Fax: +86 21 62932418; Email: zxdeng@sjtu.edu.cn

and *in-silico* discovery of bacterial GIs (Figure 1). This approach combines the newly proposed 'MAmP' (6), 'tRNAcc' (3) and 'tRIP' (3) comparative genomics-based approaches with an experimental island capture step facilitated by island probing (15) and/or a yeast-based homologous recombination system (16). MAmP (Microarray-Assisted mobilome Prospecting) is underpinned by comparative genomic hybridization (CGH), ArrayOme (6) and pulsed-field gel electrophoresis (PFGE) genome sizing (Figure 1A) and is used to screen large numbers of isolates to identify strains that are particularly rich in mobilome DNA sequences to which the species meta-array would have been 'blind'. tRNAcc (tRNA gene contents and contexts analysis), complemented by an *in silico* PCR approach (Figure 1B), is used to identify putative GIs in closely related complete and near-complete genomes. Finally, the tRIP (tRNA site interrogation for pathogenicity islands, prophages and other GIs) (Figure 1C) strategy permits high-throughput experimental identification and characterization of new GIs through PCR-based profiling of MAmP-selected or otherwise chosen test strains, followed by large-scale targeted capture and full-length sequencing of GIs. We have now incorporated and improved the previously reported ArrayOme (6) and tRNAcc standalone tools (3) into a user-friendly MobilomeFINDER web-server as a public resource: http://mml.sjtu.edu.cn/MobilomeFINDER/.

ANALYSIS TOOLS

The MobilomeFINDER web-server comprises a flexible suite of individual interactive tools that greatly facilitate high-throughput experimental and *in-silico* discovery of bacterial GIs in closely related bacteria (Table 1; Figure 1).

ArrayOme web-interface: a tool for identification of strains rich in novel mobilome DNA

Microarray-derived CGH data identifies genes common to both the microarray used and the genome of the strain

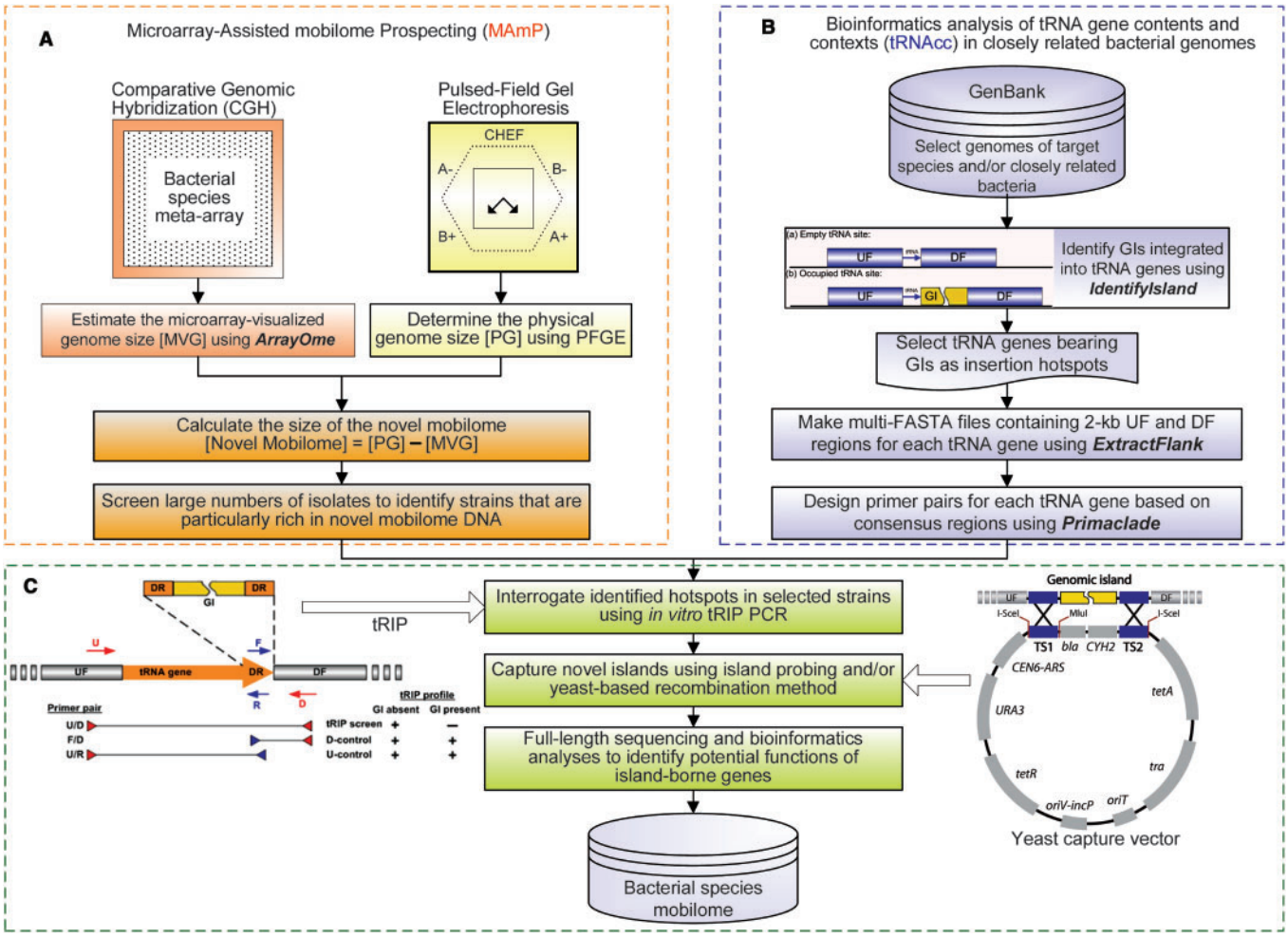

Figure 1. MobilomeFINDER island discovery strategy: (**A**) MAmP facilitates identification of strains rich in microarray-elusive novel DNA (6), (**B**) tRNAcc identifies putative GIs in complete and near-complete closely related genomes (3), (**C**) tRIP PCR permits high-throughput identification and characterization of new GIs in test strains which can then be systematically targeted for capture (15,16) and full-length sequencing. 'Island probing' refers to a technique of tagging GIs with a dual positive–negative selectable cassette to allow for subsequent marker rescue and/or GI deletion experiments to size and further characterize islands (15). The yeast recombinational method is based on an *E. coli*–yeast shuttle vector that is constructed to carry targeting sequence homologues to conserved sequences flanking the putative island (16, 28). See text for further details.

under investigation. This complement of genes represents an entity we previously defined as the microarray-visualized genome (MVG) (6). Briefly, ArrayOme produces an accurate estimate of MVG sizes based on microarray-CGH data alone (Figure 1A; Table 1). The sizes of novel mobilomes borne by individual strains can be estimated by comparing MVG sizes with PFGE-measured chromosome sizes to identify isolates rich in novel or highly divergent genetic material that are likely to carry unique GIs or prophages (Figure 1A).

The ArrayOme web-interface that we have now developed provides a universally accessible biologist-friendly tool that is complemented by a broad repertoire of DNA-prospecting accessories within the wider MobilomeFINDER web-based resource. To highlight the ease of use and benefits of this new tool we present the example of MVG size determination for *Helicobacter pylori* strain 87A300 using a previously published two-strain array (26695 and J99) CGH data set (17). The ArrayOme web-interface (Figure 2E) utilizes input data comprising: (i) a microarray-derived CGH dataset (Figure 2A), (ii) a microarray-specific index file in which array probes are mapped to CDS on either the major or minor reference templates (Figure 2B), (iii) major DNA template (*H. pylori* 26695 genome) files that provide location and size information for all CDS detectable with the particular microarray used (Figure 2C), and (iv) minor template (*H. pylori* J99 genome) files (Figure 2D). Guidance on creation and/or sourcing of files is readily accessible via 'Format' and 'Example' links. Similarly, the selection of options to maximize the utility and adaptability of ArrayOme for individual user applications is facilitated through push buttons, file-browsers and tick-boxes and explanatory notes. A major enhancement is the hyperlink-embedded graphical virtual genome map constructed by stringing together inferred contigs (ICs) that have been produced by merging adjacent genes classified as 'present'. The ICs are ordered based on major template coordinates followed by a tail made up of ICs derived from each of the minor templates in turn. The hypothetical MVG size and the contribution of each template to the virtual genome of a test strain is displayed (Figure 2F). The hyperlinks allow visualization of individual ICs using the NCBI Sequence Viewer v2.0 utility; subsequent selection of the NCBI 'Protein coding genes' link opens up the wider knowledge repository specific to genes within the selected IC. ArrayOme web-server also generates, a second graphical output comprising a circular map of the major template only that highlights the locations and sizes of identified ICs and inferred gaps (IGs) within this template (Figure 2G); IGs refer to contiguous regions defined as 'absent' following ArrayOme analysis. Hyperlinks within this output also allow access to individual NCBI Entrez Gene entries. High-quality image files in a PNG format, text files listing coordinates and CDS contents of ICs/IGs and a detailed ArrayOme result file are also generated.

Table 1. Interactive tools available at the MobilomeFINDER web-server that facilitate the process of bacterial genomic island or mobile genome (mobilome) discovery

Tool	Description
Microarray-assisted mobilome prospecting	Identify strains that are particularly rich in novel mobilome DNA; underpinned by CGH, ArrayOme and PFGE (Figure 1A)
ArrayOme	Estimate the cumulative size of parts of the genome that harbour genes identified as 'present' by microarray-based CGH; the size of the novel mobilome is thus predicted by comparing the microarray-visualized genome size with the PFGE-measured chromosome size
Island identification[a]	Identify GIs by tRNAcc analysis of closely related bacteria (Figure 1B)
IdentifyIsland	Identify putative islands based on conserved flanking blocks recognized by the multiple aligner Mauve 1.2.2 (18)
TabulateIsland	Tabulate islands identified by Identify-Island following analysis of different subsets of genomes
LocateHotspots	Locate proposed hotspots in non-annotated chromosomal sequences using BLASTN-based searches
IslandScreen	Identify putative islands by single-step crude analysis with combination of IdentifyIsland, LocateHotspots and DNAnalyser.
Primer design[a]	
ExtractFlank	Generate ClustalW-derived MSA files that contain upstream or downstream flanking regions of identified islands to serve as inputs for Primaclade-facilitated (21) design of conserved PCR primers
*insilico*tRIP	Interrogate identified hotspots for the presence or absence of an integrated element using a locally installed version of Electronic PCR (22); generates 0–500 kb virtual fragments
Island analysis[a]	
DNAnalyser	Calculate the GC content and dinucleotide bias of identified islands, and plot the negative cumulative GC profile of genomes
GenomeSubtrator	High throughput BLASTN-based comparison of CDS sequences against test genomes to identify strain-specific CDS based on the level of nucleotide similarity

[a]These tools 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.
GI, genomic island; CGH, comparative genomic hybridization; PFGE, pulsed-field gel electrophoresis; MSA, multiple sequence alignment; CDS, protein coding sequence.

MobilomeFINDER web-tools for genomic island identification by comparative analysis

The MobilomeFINDER web-server streamlines identification of GIs by comparative analysis of tRNAcc and other integration hotspots (Figure 1B). Briefly, IdentifyIsland (Table 1) exploits the multiple sequence aligner Mauve 1.2.2 (18) to investigate whether tRNA sites across multiple genomes are occupied by anomalous

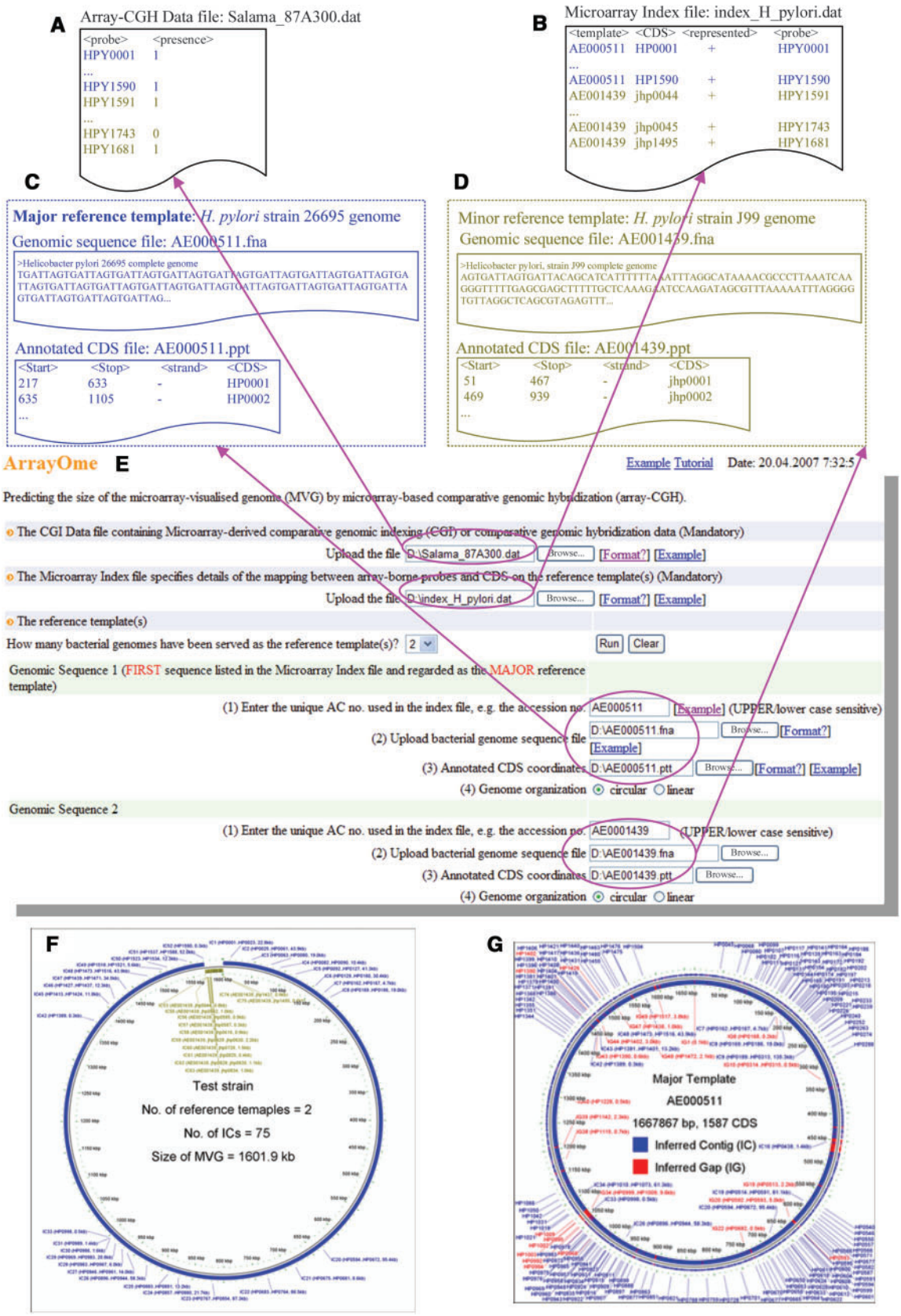

Figure 2. The ArrayOme web-tool. (**A**) Test strain CGH data file (17); (**B**) microarray index file that maps array probes to template CDS; (**C–D**) genome sequence and the annotated CDS files for major and minor reference templates, respectively; (**E**) input screenshot; (**F**) virtual MVG map of test strain; (**G**) major reference template map showing inferred contigs and gaps. See text for details.

strain-specific DNA segments lying between the 3'-ends of tRNA genes and downstream conserved flanks. In addition to the original multi-step, manually curated procedure (Figure 3), MobilomeFINDER supports single-step, entirely automated crude analysis using IslandScreen, a newly available tool that has been developed by integrating LocateHotspots, IdentifyIsland and DNAnalyser. We have described the tRNAcc strategy and individual tools in detail previously (3). Table 1 outlines key features of the tools. The web-interface varies with each tool but typically consists of an input and run parameter page and status and results pages. For example, with IdentifyIsland the input page permits uploading of the genome sequence data and details of tRNA sites to be interrogated (Figure 3A), whilst with DNAnalyser results pages display feature tables and graphical representations of islands (Figure 3G–H). In the example shown, the complete and near-complete genome sequences of *Klebsiella pneumoniae* MGH 78578 and Kp342, respectively (Figure 3A–C), were submitted. In addition, the 87 tRNA/tmRNA genes to be interrogated and their locations in the two genomes were also entered into the interface. After automated IdentifyIsland analysis (Figure 3D), potentially occupied sites were manually examined to exclude misassignments [see Ou *et al.* (3) for details].

MobilomeFINDER web-tools for detection of foreign DNA signatures

The manually edited IdentifyIsland output file is then fed into DNAnalyser (Figure 3E), which generates a tabulated output of key island features including location, size, GC content and dinucleotide frequency distribution (10) (Figure 3F). Furthermore, the web-based DNAnalyser is an intuitive tool that: (i) draws a circular genome map with the locations of identified GIs marked (Figure 3G) and (ii) plots the negative cumulative GC profile, in which a sharp upward spike indicates a relatively sharp increase in GC content whereas an abrupt fall indicates a relatively sharp decrease in GC content (11) (Figure 3H), thus highlighting the wider genomic context of islands. Hyperlinks to the NCBI database in output tables and schematics permit visualization of islands and access to annotation data. In the example presented, amongst other entities we identified a 33 kb prophage-like mobile element inserted into a *met* tRNA gene site (56_Met) in the *K. pneumoniae* MGH 78578 chromosome that possessed an integrase gene, recognizable flanking direct repeats and exhibited distinct GC content and dinucleotide usage; all features typical of archetypal integrative elements. The web-based GenomeSubtractor performs *in silico* 'subtractive hybridization' and outputs data in a tabulated form showing the locations of single and clustered strain-specific CDS based on the length of match and degree of identity (3,19).

Web-tools for high-throughput design and validation of tRIP primers

The newly developed ExtractFlank (Table 1) web-interface facilitates one-step extraction and ClustalW (20) mediated alignment of upstream (UF) and downstream (DF) hotspot flanking regions for the set of genomes being investigated. The multiple sequence alignment files are then fed into Primaclade (http://www.umsl.edu/services/kellogg/primaclade.html) (21) to facilitate the design of tRIP PCR primers specific for each flank. The newly developed *insilico*tRIP tool, supported by Electronic PCR (22), allows multiple primer pairs to be checked simultaneously for specificity versus numerous hotspots in multiple genomes (Table 1). With ready browser access to primer set and genome template files, a 500 kb cap on virtual amplicon size and an output comprising amplicon sizes and FASTA-formatted sequences, the *insilico*tRIP tool serves as a generic multiplex-PCR tool that is ideal for large-scale primer validation and/or *in silico* PCR-based genomic exploration.

IMPLEMENTATION

MobilomeFINDER runs on a Linux platform and has integrated and improved the ArrayOme (6) and tRNAcc standalone packages (3) that we reported previously. Specific enhancements include: (i) a newly developed *insilico*tRIP tool, (ii) an intuitive DNAnalyser tool that generates additional outputs comprising a circular genome map with the locations/sizes of GIs marked and a plot of the negative cumulative GC profile of the genome, (iii) a schematic ArrayOme output, (iv) hyperlinks to visualize DNA fragment details using NCBI Sequence Viewer, (v) an ExtractFlank tool that automatically generates ClustalW multiple sequence alignment files and (vi) 'example' and 'format' prompts, ready access to tutorial files and enhanced warnings re file formatting errors aid file construction and entry. In addition, the following freely available components were employed: Mauve 1.2.2 (18); NCBI Blast 2.2.9 (23); ClustalW (20); Primaclade (21); Electronic PCR (22); CGview (24); gnuplot (http://www.gnuplot.info) and Bioperl (25). Each run is assigned a job-id and the output files are kept on the server for 7 days allowing the user to inspect the results at any given time. The server web site includes a step-by-step tutorial for general users as well as detailed technical documentation and the open source codes of tRNAcc and ArrayOme for software developers. In addition, users can download the standalone versions of tRNAcc and ArrayOme to run locally.

Because sequence alignment algorithms, such as the multigenome comparison tool Mauve (18) and the pairwise alignment tool BLAST (23), are computationally intensive, it may not be possible to return results to users immediately when the input is large. With the current hardware configuration using two Dual-Core Intel Xeon 2.8GHz processors and 8GB RAM, the MAUVE-facilitated tool IdentifyIsland takes about 1 h to discover islands by comparative analysis of the contents and contexts of ~80 tRNA sites across three closely related bacterial genomes. Three tools, LocateHotspots and GenomeSubtractor that use BLASTN and IdentifyIsland that uses MAUVE, display a URL for subsequent retrieval of results if the job cannot be

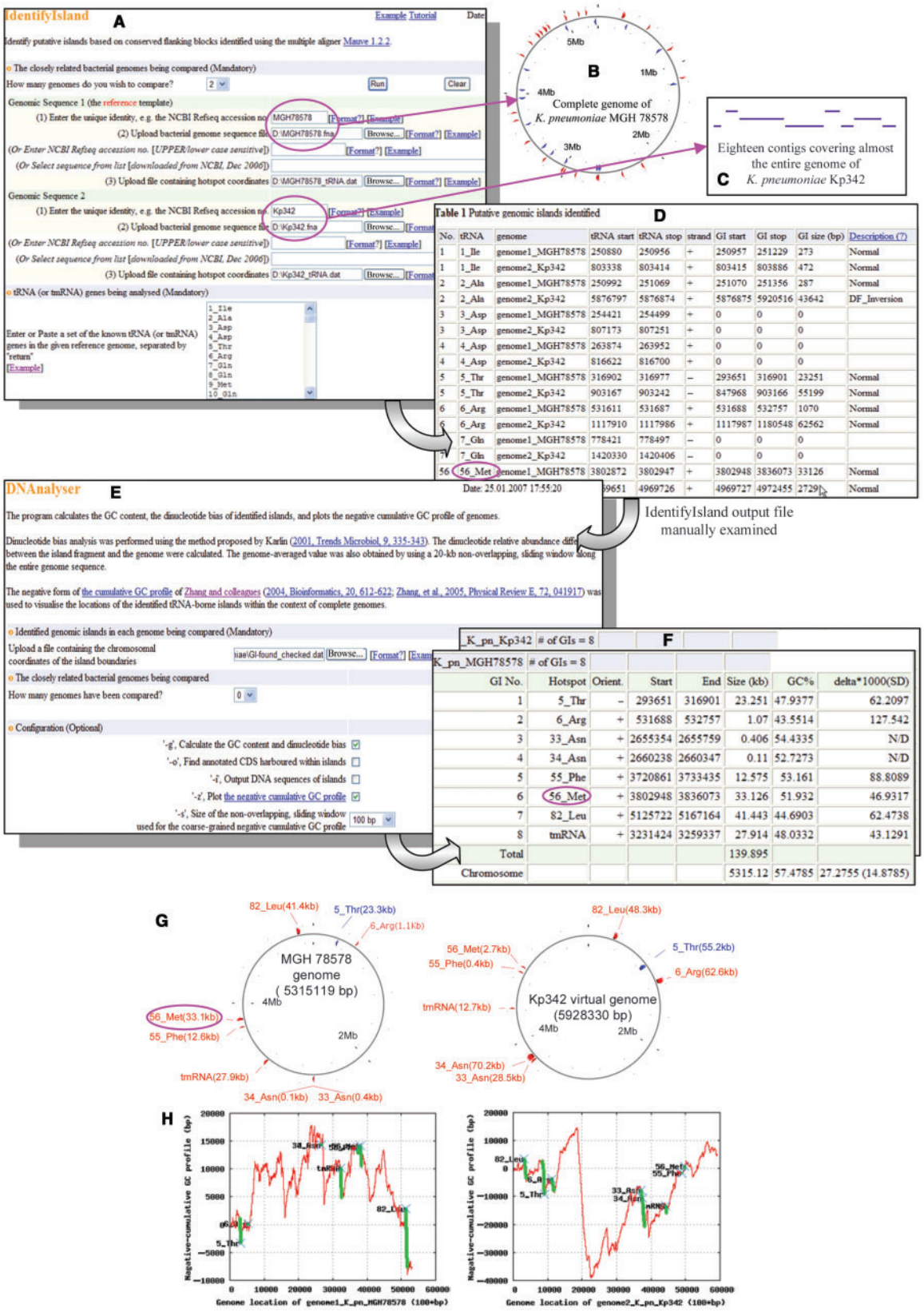

Figure 3. Web-based tRNAcc analysis of two *Klebsiella pneumoniae* genomes. (**A**) IdentifyIsland inputs specifying the *K. pneumoniae* genome sequences and the coordinates of tRNA genes. (**B**) MGH 78578 chromosomal map with tRNA genes marked. (**C**) Kp342 contigs were compared with the MGH 78578 genome using PipMaker (31) and ordered by coordinates of best alignment to produce a 'virtual' Kp342 genome. (**D**) IdentifyIsland output listing boundary coordinates of putative islands. (E) DNAnalyser input interface. (**F–H**) DNAnalyser output data for the two *K. pneumoniae* genomes analysed. See text for details.

completed promptly. Alternatively, if users supply their e-mail address, results will be emailed automatically upon completion of the job.

APPLICATIONS

MobilomeFINDER and the related experimental methodologies are applicable to a wide range of bacterial species. To date the web-server has been used to perform comparative bacterial genomic analyses for several species including nine *Escherichia coli* genomes, four *Salmonella enterica* genomes, two *K. pneumoniae* genomes (Figure 3), two *Pseudomonas aeruginosa* genomes and two *Streptococcus suis* genomes; the resulting data are shown at http://mml.sjtu.edu.cn/MobilomeFINDER/database.htm.

We have used the MobilomeFINDER web-server to characterize the GI contents of blood culture-derived *E. coli* isolates obtained from patients with no laboratory evidence of concurrent urinary tract infections. CGH analyses using the *ShE.coli* metagenome microarray (http://www.ifr.ac.uk/safety/microarrays/) (6), together with PFGE-based genome sizing has been used to identify mobilome-rich strains by MAmP (Figure 1A). In addition, PCR-based tRNA site interrogation (tRIP) (Figure 1C) coupled with chromosome walking and sequencing has been used to investigate sixteen tRNA loci in ten selected *E. coli* isolates. Approximately half of the 85 GIs identified were related to UPEC strain CFT073 islands, with an equal number resembling elements in *Shigella* and EAEC, EHEC, EPEC pathotypes of *E. coli*. Based on a limited preview data, at least seven GIs contained sequences novel to *E. coli*, with six possessing stretches of sequence without any counterparts in the entire DNA database (K. Rajakumar, unpublished data). In addition to the 95 *E. coli* GIs we identified within sequenced genomes in our recent study (3), we have also discovered by *insilico*tRIP analysis a large *leuX* tRNA gene-associated GI that contains a likely DNA modifying *dnd* gene cluster (26,27) within the unfinished genome of enterotoxigenic (ETEC) *E. coli* B7A (RefSeq accession no. NZ_AAJT00000000).

YEAST RECOMBINATIONAL SYSTEM-BASED CAPTURE OF GENOMIC ISLANDS

The yeast recombinational capture system, originally described by Raymond *et al.* (28) and subsequently modified by Wolfgang *et al.* (16), is a laboratory complement of the MobilomeFINDER tool. A single capture vector is constructed carrying targeting sequence homologues to conserved sequences flanking the putative island insertion site (Figure 1C). It can subsequently be used to capture and characterize in detail chromosomal intervals from any number of strains (16). A set of plasmids has been engineered and these are available on request from Stephen Lory (Email: stephen_lory@hms.harvard.edu). Additional information about the system is provided online at http://mml.sjtu.edu.cn/MobilomeFINDER/ycv.htm.

CONCLUSION

The MobilomeFINDER web-server has been developed to facilitate high-throughput experimental and *in-silico* discovery of bacterial GIs by combining MAmP, tRNAcc, tRIP and other related approaches. We present it as a comprehensive, comparative-genomics-based mobilome discovery platform dedicated to biologists. It is clear that even with current high-throughput genomic sequencing facilities (29) it will not be feasible to sequence hundreds of isolates to identify and decode the novel gene pool accessible to each and every bacterial species. Furthermore, microarray CGH data alone is limited to genes represented on the array and provides no insight as to the extent of novel DNA in a test strain. We propose that a strategy such as MobilomeFINDER will help address this challenge by facilitating the identification of key mobilome-rich strains and the rapid, high-throughput discovery and characterization of GIs, thereby focussing increased research effort on understanding the role of the bacterial pan-genome. In combination with shotgun sample pyrosequencing (29,30), the MobilomeFINDER strategy could even aid prioritization of strains for full-length genome characterization, thus maximizing the 'genetic return' per genome sequenced.

ACKNOWLEDGEMENTS

We are grateful to James Lonnen, Mansi Mukesh Patel and Jon van Aartsen for support in developing this resource, and to many other colleagues for testing and suggesting enhancements to MobilomeFINDER. We are also grateful to Dr Ling-Ling Chen at Shandong University of Technology, China for critical reading of the manuscript and many valuable comments. We thank the Institute for Genomic Research (TIGR) and the Genome Sequencing Centre at Washington University in St Louis for their policy of making preliminary sequence data publicly available and acknowledge the use in this study of unpublished genome data corresponding to *K. pneumoniae* strain Kp342 and MGH 78578, respectively. This study was supported by grants from the 863 program, Ministry of Science and Technology of China (Grant No. 2006AA02Z328) to H.Y.O.; National Science Foundation of China (30500285/c0110) to X.H.; and a mediSearch grant from The Leicestershire Medical Research Foundation to K.R. and M.R.B. Work in the Hinton lab was supported by a Core Strategic Grant from the BBSRC. E.H. was supported by a MRC/University of Leicester PhD studentship and A.B.T. by the University of Bahrain. Work in S.L.'s laboratory was supported by the NIH grant (GM068516). Funding to pay the Open Access publication charges for this article was provided by the 863 program from the Ministry of Science and Technology, China (2006AA02Z328).

Conflict of interest statement. None declared.

REFERENCES

1. Binnewies,T.T., Motro,Y., Hallin,P.F., Lund,O., Dunn,D., La,T., Hampson,D.J., Bellgard,M., Wassenaar,T.M. *et al.* (2006) Ten years

of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct. Integr. Genomics*, **6**, 165–185.
2. Dobrindt,U., Hochhut,B., Hentschel,U. and Hacker,J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.*, **2**, 414–424.
3. Ou,H.Y., Chen,L.L., Lonnen,J., Chaudhuri,R.R., Thani,A.B., Smith,R., Garton,N.J., Hinton,J., Pallen,M., Barer,M.R. *et al.* (2006) A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. *Nucleic Acids Res.*, **34**, e3.
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. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.*, **3**, 722–732.
6. Ou,H.Y., Smith,R., Lucchini,S., Hinton,J., Chaudhuri,R.R., Pallen,M., Barer,M.R. and Rajakumar,K. (2005) ArrayOme: a program for estimating the sizes of microarray-visualized bacterial genomes. *Nucleic Acids Res.*, **33**, e3.
7. Gal-Mor,O. and Finlay,B.B. (2006) Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol.*, **8**, 1707–1719.
8. 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.
9. Hsiao,W., Wan,I., Jones,S.J. and Brinkman,F.S. (2003) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics*, **19**, 418–420.
10. Karlin,S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol.*, **9**, 335–343.
11. 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.
12. 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.
13. 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.
14. Koonin,E.V., Makarova,K.S. and Aravind,L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.*, **55**, 709–742.
15. Rajakumar,K., Sasakawa,C. and Adler,B. (1997) Use of a novel approach, termed island probing, identifies the *Shigella flexneri she* pathogenicity island which encodes a homolog of the immunoglobulin A protease-like family of proteins. *Infect. Immun.*, **65**, 4606–4614.
16. Wolfgang,M.C., Kulasekara,B.R., Liang,X., Boyd,D., Wu,K., Yang,Q., Miyada,C.G. and Lory,S. (2003) Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA*, **100**, 8484–8489.
17. Salama,N., Guillemin,K., McDaniel,T.K., Sherlock,G., Tompkins,L. and Falkow,S. (2000) A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl Acad. Sci. USA*, **97**, 14668–14673.
18. 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.
19. 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.
20. 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.
21. 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.
22. Schuler,G.D. (1997) Sequence mapping by electronic PCR. *Genome Res.*, **7**, 541–550.
23. 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.
24. Stothard,P. and Wishart,D.S. (2005) Circular genome visualization and exploration using CGView. *Bioinformatics*, **21**, 537–539.
25. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
26. Zhou,X., He,X., Liang,J., Li,A., Xu,T., Kieser,T., Helmann,J.D. and Deng,Z. (2005) A novel DNA modification by sulphur. *Mol. Microbiol.*, **57**, 1428–1438.
27. Zhou,X., He,X., Li,A., Lei,F., Kieser,T. and Deng,Z. (2004) Streptomyces coelicolor A3(2) lacks a genomic island present in the chromosome of Streptomyces lividans 66. *Appl. Environ. Microbiol.*, **70**, 7110–7118.
28. Raymond,C.K., Sims,E.H., Kas,A., Spencer,D.H., Kutyavin,T.V., Ivey,R.G., Zhou,Y., Kaul,R., Clendenning,J.B. *et al.* (2002) Genetic variation at the O-antigen biosynthetic locus in Pseudomonas aeruginosa. *J. Bacteriol.*, **184**, 3614–3622.
29. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
30. Spencer,D.H., Kas,A., Smith,E.E., Raymond,C.K., Sims,E.H., Hastings,M., Burns,J.L., Kaul,R. and Olson,M.V. (2003) Whole-genome sequence variation among multiple isolates of *Pseudomonas aeruginosa*. *J. Bacteriol.*, **185**, 1316–1325.
31. Schwartz,S., Zhang,Z., Frazer,K.A., Smit,A., Riemer,C., Bouck,J., Gibbs,R., Hardison,R. and Miller,W. (2000) PipMaker–a web server for aligning two genomic DNA sequences. *Genome Res.*, **10**, 577–586.