INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms

Sakshi Shrivastava, Ch V Siva Kumar Reddy and Sharmila S Mande*

Bio-Sciences Division, Innovation Labs, Tata Consultancy Services, 1 Software Units Layout, Hyderabad 500 081, India

Corresponding author (Fax, +91-40-6667-2222; Email, sharmila.mande@tcs.com)

Genomic islands (GIs) are regions in the genome which are believed to have been acquired via horizontal gene transfer events and are thus likely to be compositionally distinct from the rest of the genome. Majority of the genes located in a GI encode a particular function. Depending on the genes they encode, GIs can be classified into various categories, such as 'metabolic islands', 'symbiotic islands', 'resistance islands', 'pathogenicity islands', etc. The computational process for GI detection is known and many algorithms for the same are available. We present a new method termed as Improved N-mer based Detection of Genomic Islands Using Sequence-clustering (INDeGenIUS) for the identification of GIs. This method was applied to 400 completely sequenced species belonging to proteobacteria. Based on the genes encoded in the identified GIs, the GIs were grouped into 6 categories: metabolic islands, symbiotic islands, resistance islands, secretion islands, pathogenicity islands and motility islands. Several new islands of interest which had previously been missed out by earlier algorithms were picked up as GIs by INDeGenIUS. The present algorithm has potential application in the identification of functionally relevant GIs in the large number of genomes that are being sequenced. Investigation of the predicted GIs in pathogens may lead to identification of potential drug/vaccine candidates.

[Shrivastava S, Reddy Ch V S K and Mande S S 2010 INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms; *J. Biosci.* **35** 351–364] **DOI** 10.1007/s12038-010-0040-4

1. Introduction

Bacterial genomes evolve through events such as mutations, rearrangements and horizontal gene transfer (HGT). Due to HGT, bacterial species acquire new traits from unrelated sources. There is substantial evidence to suggest that HGT events give rise to compositionally biased regions in the genomes, called genomic islands or GIs (Blum *et al.* 1994; Sullivan and Ronson 1998). It has been shown that the majority of genes located on the GI encode functions that are useful for survival of the bacteria within a population (Dobrindt *et al.* 2004). For instance, GI-containing genes encoding the iron-uptake system in *Yersinia pestis* enhances the capacity of this organism to grow and disseminate in the soil or in a host (Schubert *et al.* 2002).

Studies have indicated that some GIs are involved in metabolic functions. Such islands contain gene clusters belonging to various metabolic pathways, examples of which include the polyketide synthesis pathway in Streptomyces (Ginolhac *et al.* 2005), arginine biosynthesis and xanthan gum production pathways in Xanthomonas (Lima *et al.* 2008), etc. A GI in *Burkholderia pseudomallei* containing

Keywords. Genomic island; metabolic island; pathogenicity island; resistance island; secretion island; symbiotic island

The program is available on request from the authors.

Abbreviations used: FPI, Francisella pathogenicity island; GI, genomic island; HGT, horizontal gene transfer; HPI, high pathogenicity island; INDeGenIUS, Improved N-mer based Detection of Genomic Islands Using Sequence-clustering; LEE, locus of enterocyte effacement; LPS, lipopolysaccharide; MDR, multidrug resistant; MI, metabolic island; MoI, motility island; MSHA, mannose-sensitive haemagglutinin; PAI, pathogenicity island; RI, resistance island; SyI, symbiotic island; SI, secretion island; TCA, tricarboxylic acid

Supplementary tables and figures pertaining to this article are available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/sept2010/351-364-suppl.pdf*

possible sugar utilization and amino acid catabolism genes was termed as 'metabolic island' (MI) by Tumapa *et al.* (2008). Many metabolic pathways are indicated to have been acquired via HGT events. For example, the lipopolysaccharide biosynthesis pathway in Xanthomonas is known to have been acquired via HGT (Patil *et al.* 2004, 2007).

Some GIs have been observed to carry nitrogen fixation genes, and are widely known across various genera belonging to the rhizobial group of alpha-proteobacteria. These genera include Rhizobium, Bradyrhizobium, Sinorhizobium, Azorhizobium and Mesorhizobium (Gonzalez *et al.* 2003). It is known that these rhizobial lineages had diverged before the evolution of legumes and the symbiosis genes were then acquired by horizontal transfer (Sullivan *et al.* 2002; Finan 2002). Each of these organisms encodes genes which include either the *nif* or *nod* genes that function in the areas of nitrogen fixation and nodulation, respectively (Finan 2002; Gonzalez *et al.* 2003). Kaneko *et al*. (2000) identified a 611 kb DNA segment encoding 30 genes for nitrogen fixation and 24 genes for nodulation in *Mesorhizobium loti* and termed this region as a 'symbiotic island' (SyI). Burkholderia and Ralstonia, which belong to the beta-proteobacteria and are known to be nitrogen-fixing legume symbionts (Chen *et al.* 2003; Tumapa *et al.* 2008), also contain islands encoding the *nif* and *nod* genes. Thus, SyIs have been observed in both alpha- and beta-proteobacteria.

Certain GIs carry genes that encode factors which confer resistance to antimicrobial substances. Such islands have been found in a few organisms such as *Salmonella enterica, Shigella flexneri*, *Vibrio cholerae* and *Staphylococcus aureus* (Dobrindt *et al.* 2004). These include the well-characterized SXT island in *V. cholerae* and the multidrug-resistant (MDR) island SGI1 of *S. enterica* (Beaber *et al.* 2002; Doublet *et al.* 2002; Chiu *et al.* 2005). Most of the genes found in these islands correspond to specific antibiotic resistance, the examples being the *tetR* gene for tetracycline resistance in SGI-1 of *S. enterica* and SRL cluster *of S. flexneri 2a*, *floR* gene for chloramphenicol resistance in the SXT island of *V. cholerae,* etc. An MDR cluster of 45 genes has been discovered in *Acinetobacter baumannii* (Fournier *et al.* 2006). An 86 kb genomic region containing 45 resistance gene clusters in *A. baumannii AYE* is known to be the largest identified GI and was termed as 'resistance island' (RI) by Fournier *et al. (*2006).

A number of GIs, particularly in pathogenic bacteria, contain genes coding for toxins, virulence factors, adhesins, etc. These islands were the first group of GIs to be characterized and categorized as 'pathogenicity islands' (PAIs). Following the discovery of PAIs in uropathogenic *Escherichia coli* (Hacker *et al.* 1990), they have been identified and studied widely in other bacterial genomes such as the *Francisella* pathogenicity island (FPI) of *Francisella tularensis* (Nano *et al.* 2004) and high pathogenicity island (HPI) of *Yersinia spp.* (Carniel 1999).

Gene components of bacterial secretion systems have also been found to be located on some of the GIs. Six different types (types I to VI) of bacterial secretion systems are known to transport effector molecules outside the bacterial cell wall. Since the mode of functioning of these systems varies in different organisms, the genes that they encode also vary a great deal. The Type I secretion system is relatively simple and is composed of only four gene components. On the other hand, there exists a complex array of genes and components for all the other secretion systems. Types II, III, IV and VI secretion systems consist of 15–20 genes, each making up a complex secretion system (Kostakioti *et al.* 2005). The Type V secretion system is an autotransporter whose substrates can mediate their own transport across the outer lipid bilayer (Kostakioti *et al.* 2005). One of the examples of a GI containing a bacterial secretion system gene cluster is the CAG PAI of *Helicobacter pylori* which codes for the components of the type IV secretion system (Odenbreit *et al.* 2000).

Some GIs have been found to code for genes that are responsible for bacterial motility. These genes include flagellary, pili and fimbrial genes. Such islands are often found to be acquired via HGT. For instance, it was hypothesized that one of the two flagellary systems in *Rhodobacter sphaeroides* has been acquired horizontally from alpha-proteobacteria (Poggio *et al.* 2007). Similarly, a flagellum-coding gene cluster has been predicted to be transferred horizontally between Euryarchaeota and Crenarchaeotais (Desmond *et al.* 2007).

A number of methods have been developed for the identification of GIs. Each of these methods utilizes different features of the islands. For instance, the information that most GIs are flanked by direct/inverted repeats and have tRNA or tmRNA in their proximity was utilized by Mantri and Williams (2004) to identify GIs. Similar feature-based methods include those by Ou *et al*. (2006) and Nag *et al.* (2006). One of the limitations of these annotation-based approaches is that these methods require fully annotated genomes. They also fail to identify GIs that are devoid of the flanking features such as direct repeats and tRNA/tmRNA. In order to overcome these limitations, methods based on genomic composition such as %G+C, codon usage, amino acid usage and oligonucleotide frequencies have been used for the identification of GIs (Karlin 2001; Rajan *et al.* 2007). These methods first obtain a signature of the genome which can be considered as 'native' to the genome. Then, after partitioning the genome into regions, the difference between the signature of each region and the signature of the 'native' genome is computed. The regions whose signatures are most distant from the signature of the 'native' genome indicate compositionally distinct regions and can be referred to as 'non-native' or GIs.

Based on the presence of functional genes in GIs, function-specific terms such as 'pathogenicity islands' (PAIs), 'antibiotic resistance islands' (RIs), 'metabolic islands' (MIs) and 'symbiotic islands' (SyIs) have been reported (Dobrindt *et al.* 2004). The present paper describes a method called INDeGenIUS for GI detection and reports classification of the predicted GIs into not only the above types of islands, but also into 'motility islands' (MoIs) and 'secretion islands' (SIs). SIs were further classified as Types I–VI SIs, depending on the type of secretion system genes contained in them. Identification of GIs may help in understanding bacterial evolution. In addition, investigation of GIs in pathogens may lead to identification of potential drug/vaccine candidates.

2. Methods

2.1 *INDeGenIUS: algorithm to identify genomic islands*

The Improved N-mer based Detection of Genomic Islands Using Sequence-clustering (INDeGenIUS) method has been developed using the principles of hierarchical clustering.

The steps in the INDeGenIUS method are schematically illustrated in figure 1 and described below. Details of steps 1–8 and 9–13 are illustrated in figure 1.

1 Divide the genome into {n} overlapping bins of equal sizes.
2 For a given oligonucleotide of length k, i.e. word length {k}, for each bin, compute a vector with frequency of all possible 4^k words as components.
3 Create {n} clusters with each bin assigned to one cluster. The frequency vector of the bin is the representative vector of the corresponding cluster.
4 Compute the distance between all possible pairs of clusters.
5 Choose the pair with the least distance and merge them into a single cluster. Each component in the representative vector for the merged cluster is obtained by averaging the corresponding frequency vector components of the bins that are merged.
6 Compute the percentage of the genome covered by each cluster.
7 Repeat from step 5 till the percentage of the genome covered by any one cluster exceeds a given threshold. This cluster is called a 'major cluster' and the remaining clusters are called 'minor clusters'.
8 Compute the centroid of the 'major cluster'.
9 Divide the genome into {m} non-overlapping bins of equal sizes.
10 For each bin, compute a vector with frequency of all possible 4^k words as components.
11 Compute the distance between the centroid of the 'major cluster' and vector corresponding to each of the m bins.
12 Plot the computed distances of all the bins.
13 The prominent peaks constitute the 'non-native' regions of the genome.

The distance between the clusters is calculated using the Manhattan distance metric (L1 norm) of the corresponding representative vectors. As the clustering proceeds, bins which are 'native' to the genome are grouped under 'major cluster'. Thus, the representative vector of the major cluster can be considered as the true signature of the genome.

Use of higher k-values will result in increased computational time as well as inaccurate frequencies of one or more k-mers. On the other hand, use of smaller k-values will lead to reduced chances of finding similar fragments. Since high accuracy has been reported earlier (Rajan *et al.* 2007) using a bin size of 10 kb and a k-value of 5, the same parameters were considered in the present study. Since the percentage of 'non-native' regions in a given genome may vary between 1% and 30%, we consider a cluster to be a 'major cluster' when it covers 70% of the genome.

2.2 *Validation of the INDeGenIUS method*

Completely sequenced genomes of 50 diversely distributed bacterial organisms were downloaded from the NCBI GenBank (*http://www.ncbi.nlm.nih.gov*). A fragment of length 20 kb was extracted from the genome of *Brucella melitensis 16M* and randomly inserted into each of the 50 genomes under study. The sensitivity of the method was calculated as the percentage of the inserted region that is predicted as 'non-native'. The effect of length of the inserted sequence on the sensitivity of the method was checked by inserting fragments of varying lengths (40 kb, 60 kb, 80 kb, 120 kb) from the *B. melitensis 16M* genome into each of the genomes under study. Also, in order to assess the sensitivity of the method in identifying fragments inserted from an organism belonging to the same family or genera, fragments of various sizes (40 kb, 60 kb, 80 kb, 120 kb) from *B. melitensis 16M* were inserted into the genomes of *O. anthropi* and *B. suis*.

2.3 *Identification and classification of genomic islands in proteobacteria*

Completely sequenced genomes of 400 proteobacterial organisms were downloaded from the NCBI (*http://www.ncbi.nlm.nih.gov*) and the INDeGenIUS method was used to identify the GIs in each of these organisms.

The distance of each bin from the major cluster was computed and plotted for each organism. The peaks

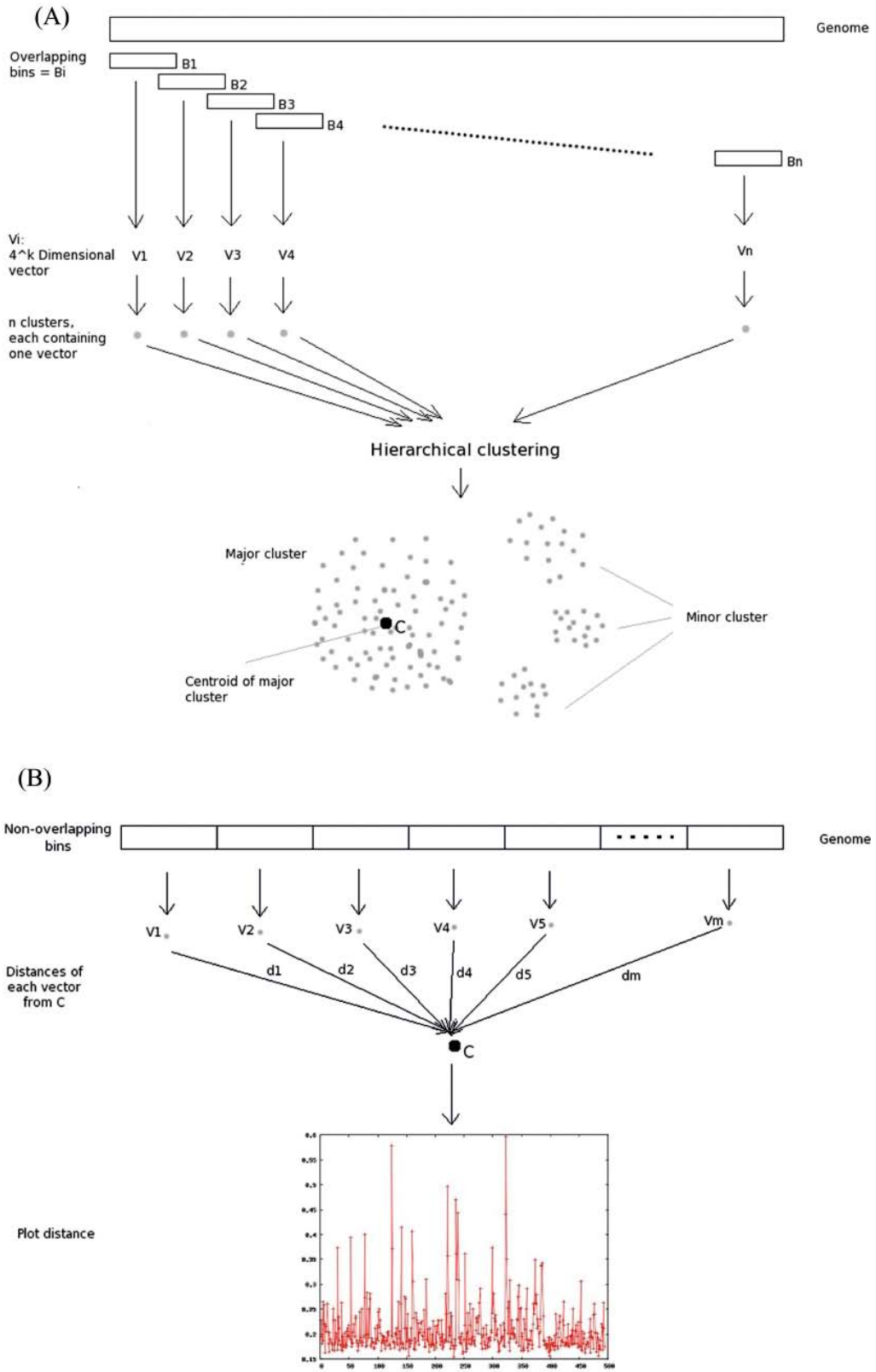

Figure 1. Schematic diagrams illustrating the various steps of the INDeGenIUS method. **(A)** Creation of 'major' and 'minor' clusters (Steps 1–8). **(B)** Prediction of the 'non-native' regions of the genome (Steps 9–13).

in the plot indicate the corresponding regions that are compositionally distinct as compared to the native genome. The prominent peaks correspond to probable GIs. These prominent peaks are manually studied and annotated. Since the percentage of 'non-native' regions in a given genome may vary between 1% and 30%, the number of prominent peaks will also vary. Genes present in these GIs were then analysed as follows.

The genes in each of the predicted GIs were analysed in order to classify the GIs into six categories: probable MI, probable SyI, probable RI, probable SI, probable PAI and probable MoI. In order to classify the GIs into these categories, a list of genes that characterize a particular island was first created by extracting information on them from the literature. Each GI was characterized into a specific class based on the type and number of genes present in the list. Thus, if the major part of a GI region consisted of genes encoding a particular metabolic property, it was classified as a probable MI. These properties included a wide range of pathways, examples of which include metabolic synthesis, metabolic degradation and metabolic utilization pathways. Likewise, if a GI encoded mostly nitrogen fixing and nodulating genes or resistant genes or genes for motility/ transport, it was classified as a probable SyI, RI or MoI, respectively. Similarly, if a GI encoded virulence-specific genes such as toxins, adhesins, modulins, invasins, phage-related genes, secretion system components, etc. then it was classified as a probable PAI. The information on known PAIs was taken from the Pathogenicity Island Database (PAI-DB: *http://www.gem.re.kr/paidb/about_paidb.php?m=h*) and Virulent Factor Database (VfDB: *http://www.mgc.ac.cn/VFs/main.htm*). If the predicted PAI region contained genes specific to any of the known secretion systems, it was subclassified as an SI. Thus, in this study, PAIs were considered as islands without SIs. An in-depth analysis for the predicted SIs was carried out in order to classify the SIs under six heads – Type I SI, Type II SI, Type III SI, Type IV SI, Type V SI and Type VI SI, based on the type of the secretion system (Types I–VI) encoded in them. For the recently discovered Type VI secretion system, predicted gene components (Shrivastava and Mande 2008), in addition to known ones, were looked for in the predicted SI regions. The predicted islands which coded mostly hypothetical or putative genes were classified as 'unknown islands'. All the classified GIs were further grouped under two heads – those that are known from the literature and others that are not.

3. Results

The known computational methods for GI detection are based on the identification of genomic regions (which can be thought of as 'non-native') deviating from the 'genomic signature' that characterizes a given genome. Since the percentage of the genome occupied by the 'non-native' regions influences the calculation of the genomic signature, incorrect signatures could result in wrong prediction of the 'non-native' regions. The present method constructs the likely signature of the genome by applying clustering techniques, thereby increasing the prediction accuracy of the 'non-native' regions significantly.

3.1 *Validation of the algorithm*

The sensitivity of prediction of the inserted fragment of length 20 kb from *Brucella melitensis 16M* in 50 genomes by INDeGenIUS was found to be 100%. The inserted fragments from *B. melitensis 16M* was chosen since this organism is evolutionarily distant from the 50 organisms under study. In other words, the present method correctly identified the 'non-native' region of length 20 kb in all the 50 genomes. This is in contrast to the results obtained by other methods such as Centroid, %G+C and Genomic-signature, which showed average sensitivity values of 89.8%, 57.7% and 78.4%, respectively (Rajan *et al.* 2007). The method also predicts inserted fragments of various lengths (40 kb, 60 kb, 80 kb and 120 kb) with 100% sensitivity. In addition, INDeGenIUS was able to predict most of the inserted fragments of various lengths (20 kb, 40 kb, 60 kb, 80 kb and 120 kb) from *B. melitensis 16M* into *Ochrabacterium anthropi* as against those into *Brucella suis* (supplementary table 1), indicating the ability of the method to predict foreign fragments originating from an organism belonging to the same family.

3.2 *Identification and classification of genomic islands*

Out of the total number of 2905 GIs identified by INDeGenIUS in 400 completely sequenced proteobacterial genomes, a total of 1308 GIs could be classified into the 6 categories (supplementary tables 2 and 3). Of these, 416 GIs could be classified as MIs, 27 as SyIs, 22 as RIs, 373 as SIs, 222 as PAIs and 249 as MoIs. Among these, while 29 known GIs across 19 organisms were picked up as probable GIs, 4 known GIs were not detected by INDeGenIUS (table 1). The known GIs which were not detected by INDeGenIUS include the vrl locus of *Dichelobacter nodosus*, LIPI2 locus of *Listeria ivanovii,* SPI6 of *Salmonella enterica* and VSPI and II of *Vibrio cholerae.*

3.3 *Analysis of genomic islands in proteobacteria*

Supplementary table 2 gives the descriptions of the functional islands obtained in various organisms. Information about the predicted island type in each of the 400 organisms along with information on their species and strains is listed in

Figure 2. Distances of all possible bins (genomic regions) from the 'native genome' (major cluster) plotted against the complete genome in 6 representative organisms. **(A)** *Yersinia pestis*, **(B)** *Escherichia coli*, **(C)** *Acinetobacter baumannii*, **(D)** *Mesorhizobium loti*, **(E)** *Xanthomonas campestris* and **(F)** *Burkholderia pseudomallei*. The predicted GIs encoding various properties correspond to the peaks in the plots. Some of the properties encoded by the predicted islands are marked with the following abbreviations: T2SI, Type II SI; T3SI, Type III SI; T4SI, Type IV SI; T6SI, Type VI SI; LPS, lipopolysaccharide; FIM, fimbrial proteins; PAI, pathogenicity island; RI, resistance island; MI, metabolic island; MDR, multidrug-resistant genes UI, unknown island encoding hypothetical/unknown genes.

supplementary table 3. Plots indicating genomic regions that deviate from the signature of the native genome in 6 representative organisms are shown in figure 2 and the identified 'non-native' regions in each are described below.

The known HPI-containing genes for biosynthesis, transport and regulation of the siderophore yersiniabactin were obtained as the most deviant peak in *Y. pestis kim* (figure 2A). Among the other peaks, a probable MI coding for lipopolysaccharide (LPS) biosynthesis genes, 4 predicted SIs (corresponding to two Type VI SI, one Type II SI and one Type III SI) and one probable MoI coding for flagellary system genes were predicted by INDeGenIUS.

Similarly, the known locus of enterocyte effacement (LEE) PAI of *E. coli O157* coding for a Type III secretion system was picked up as one of the peaks by INDeGenIUS (figure 2B). The tallest peak was also predicted as a Type III SI, different from LEE. Apart from these, many phage genes

Table 1 List of known genomic islands and their predictions by INDeGenIUS.

Organism name	Property encoded	Known Genomic Islands	Predicted by INDeGenIUS
Acinetobacter	Multi drug resistance	Resistance island	√
Dichelobacter nodosus	VAP locus	Pathogenicity island	√
	Vrl locus	Pathogenicity island	
Erwinia	Hrp PAI	Pathogenicity island	√
Escherichia	LEE (T3SS)	Pathogenicity island	√
Francisella	FPI	Pathogenicity island	√
*Neisseria**	GGI	Pathogenicity island	√
Photorhabdus	Mcf PAI	Pathogenicity island	
	Tcd island	Pathogenicity island	√
Pseudomonas	Hrp PAI	Pathogenicity island	√
	HSI	Secretion island	√
Salmonella	SPI-I	Pathogenicity island	√
	SPI-II	Pathogenicity island	√
	SPI-III	Pathogenicity island	√
	SPI-IV	Pathogenicity island	√
	SPI-V	Pathogenicity island	√
Shigella	SHI-I	Pathogenicity island	√
	SHI-II	Pathogenicity island	√
	SHI-III	Pathogenicity island	√
	SRL	Resistance island	
	SHI-O	Metabolic island	√
Vibrio	VPI-I	Pathogenicity island	√
	VPI-II	Pathogenicity island	√
	VSP I & II	Genomic islands	
Xanthomonas	Type III secretion system	Pathogenicity island	√
	Xanthum gum producing proteins	Metabolic island	√
Yersinia	HPI	Pathogenicity island	√
Helicobacter	CAG	Pathogenicity island	√
Azorhizobium	Symbiotic genes	Symbiotic island	√
Bradyrhizobium	Symbiotic genes	Symbiotic island	√
Mesorhizobium	Symbiotic genes	Symbiotic island	√
Rhizobium	Symbiotic genes	Symbiotic island	√
Sinorhizobium	Symbiotic genes	Symbiotic island	√

*Only in *N. gonorrhoeae.*

encoding PAIs, two MIs containing genes coding for LPS biosynthesis genes and urease synthesis genes were also predicted.

In addition, the well-known MDR island coding for several antibiotic resistance genes in *A. baumannii* was obtained as the second-highest peak (figure 2C), reflecting its compositionally biased nature. The other predicted islands included a Type VI SI and two MIs coding for the phenyl acetic acid degradation and urease synthesis pathways.

The known SyI in the symbiotic organism *Mesorhizobium loti* was predicted to be the top peak (figure 2D). The other predicted islands in this symbiotic organism included a Type III SI, Type VI SI, a flagellary system coding an MoI and two MIs coding for genes belonging to the thiamine biosynthesis and peptidoglycan synthesis pathways.

The known MI coding for xanthan gum production was one among the two predicted MIs in *X. campestris* (figure 2E), the other being an LPS biosynthesis-specific MI. Apart from MIs, a Type III SI, Type VI SI and many phage-related PAIs were also predicted.

The identified GI corresponding to the topmost peak in *B. pseudomallei* (figure 2F) contained gene components of the Type VI secretion system and thus was predicted to be a Type VI SI. Predicted MIs included four pathways, namely, the peptidoglycan synthesis, capsular polysaccharide synthesis, LPS biosynthesis and phenylacetic acid degradation pathways. Several PAIs with phage-related genes were also predicted.

The present method was compared with three genomic composition-based methods, namely, %G+C, genomic signature and centroid (Karlin 2001; Rajan *et al.* 2007). Comparison of plots obtained for 2 representative organisms using all the 4 methods is given in supplementary figure 1. As can be seen from supplementary figure 1A, plots obtained for *Xanthomonas oryzae* using the present method clearly indicated well-defined peaks as compared to those obtained with the other methods. Similarly, for *Yersinia pestis,* three prominent peaks corresponding to T2SI, T6SI and HPI were obtained by the present method (supplementary figure 1B). Some of these peaks were not picked up by the other methods. The results obtained by the present method for *E. coli* were also comparable to those reported by Lawrence and Ochman (1998).

3.3.1 *Metabolic islands:* The predicted MIs had gene clusters mainly from 10 metabolic pathways. These included pathways for LPS biosynthesis, polysaccharide synthesis, urease synthesis, cobalamine synthesis, peptidoglycan synthesis, thiamine biosynthesis, flagellary biosynthesis, tricarboxylic acid (TCA) cycle, propanediol utilization and phenylacetic acid degradation (supplementary table 4). Likely MIs encoding these pathways were predicted across 59, 37, 16, 46, 102, 19, 53, 15, 14 and 22 organisms, respectively (supplementary table 4). As can be observed, peptidoglycan synthesis was seen to be the most widely predicted pathway among the proposed MIs. Pathways for propanediol utilization, TCA cycle and urease synthesis were predicted across very few organisms.

Three assumptions were made to validate the above predictions. First, since it is widely known that GIs are compositionally biased regions and are mostly acquired through HGT (Mario *et al.* 2009), an evidence of a horizontal transfer event in each of these pathways was looked for. Second, it is believed that HGT is a common phenomenon in genes which are organized in clusters, as transfer and acquisition of entire gene clusters is easier than that of individual genes (Lawrence and Roth 1996). Thus, if all or the majority of genes belonging to a pathway were identified within a predicted GI, then the predicted GI was considered as a likely MI. Figure 3 highlights the genes that occur within the predicted MIs in each of these pathways. Third, since it is known that horizontally transferred genes are usually present in only a few lineages (Garcia-Vallvé *et al.* 2000), the interstrain and interspecies differences occurring in these pathways were evaluated. The supporting evidence for the predicted MIs is discussed below.

The genes belonging to the predicted MIs encoding cobalamine synthesis in *Salmonella typhimurium* and *E. coli* were earlier suggested to have been acquired via HGT (Roth *et al.* 1996; Lawrence and Roth 1995). It was also suggested that the cobalamine operon and the propanediol operon (also predicted by INDeGenIUS), which are adjacent to each other, were acquired in a single horizontal transfer event in these two organisms (Lawrence and Roth 1996; Roth *et al.* 1996). In addition, genes of both the cobalamine synthesis and propanediol utilization pathways have been listed in the database of HGTs (HGT-DB: Garcia-Vallvé *et al.* 2000, 2003), which further supports our prediction. Further analysis indicated that all the genes involved in these two pathways were obtained within the predicted MI regions (figure 3), suggesting that the entire gene cluster might have been acquired via HGT. Thus, the predicted GIs containing genes that belong to these two pathways were considered as probable MIs.

All the genes in the predicted MI involved in the lipopolysaccharide biosynthesis pathway (figure 3) in *Xanthomonas oryzae, V. cholerae* and *S. enterica* were reported to have been acquired by HGT events (Mooi and Bik 1997; Patil *et al.* 2004, 2007; Vernikos *et al.* 2007). Examination of the different organisms in which the LPS biosynthesis gene cluster was predicted as a GI indicated a clear interspecies and interstrain differences, which hinted at the occurrence of an HGT event. For example, in the case of Vibrio, the LPS biosynthesis gene cluster was predicted in three strains of *V. cholerae,* namely, *O395, O1* and *M66*, but was absent from the strain *V. cholerae MJ* and species

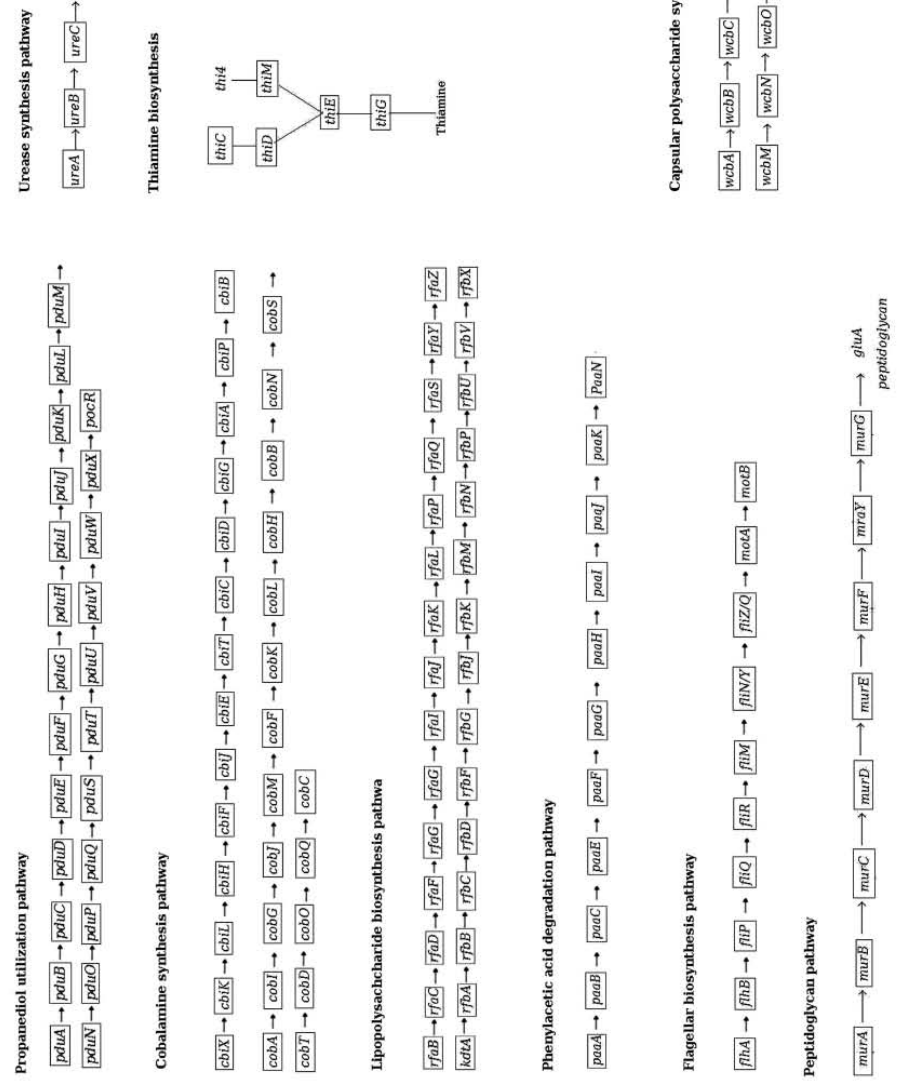

Figure 3. Metabolic pathways that have been identified as metabolic islands (MIs) in certain bacteria. Genes in each of the pathways which have been identified in the predicted MIs are highlighted inside the boxes.

V. vulnificus, V. fischeri, V. harvey and *V. parahemolyticus*. Similarly, in the case of Xanthomonas, the LPS biosynthesis gene cluster was obtained in *X. campestris* but not in *X. citri* and *X. oryzae*. A similar trend was observed in most of the other organisms (supplementary table 4), clearly indicating that the genes for this predicted MI were acquired horizontally. We observed that, in some cases, some or all of the genes present in the MI in certain bacteria were missing in its other strains and, as a result, our method did not pick up any GIs in these strains. For example, a few genes in the LPS gene cluster of *X. campestris* were observed to be missing in *X. citri.*

All 12 genes involved in the phenylacetic acid degradation pathway (figure 3) were identified as 'non-native' and thus classified as probable MIs by INDeGenIUS in 22 organisms. This pathway was earlier reported to have been acquired by HGT events in a few organisms which include *A. baumannii, Bordetella parapertussis, B. pseudomallei,* etc. (Garcia-Vallvé *et al.* 2000, 2003). Abe-yoshizumi *et al.* (2004) also suggested that all the genes belonging to this pathway are subjected to a frequent HGT within and across bacterial phyla. Furthermore, an interstrain difference in this pathway was also observed in *A. baumannii.* In this organism, this pathway was predicted as an MI in *A. baumannii AB0057, AB307* and *AYE,* but not in the other strains (supplementary table 4). Similarly, in the case of Bordetella, the pathway was predicted as an MI in *B. parapertussis* but not in the other strains.

All the genes involved in the biosynthesis of flagella (figure 3) in 53 organisms were predicted as MIs by INDeGenIUS. Flagellary systems are known to be acquired by HGT in a few of these organisms. For example, phylogenetic analysis suggested that one of the two flagellar systems in *R. sphaeroides* was acquired by horizontal transfer from a gamma-proteobacterium (Poggio *et al.* 2007). Other alpha-proteobacteria closely related to *R. sphaeroides* were also seen to have acquired a similar set of flagellar genes, suggesting that a common ancestor received this gene cluster (Poggio *et al.* 2007). According to Liu (2009), the bacterial flagella had arisen from a simple secretion system to become a well-honed structure, and various evolutionary forces such as HGT, gene loss and gene duplication events had contributed to this step-wise formation.

All genes in the urease pathway (figure 3) were identified within the predicted GIs by INDeGenIUS, indicating that the urease pathway was a probable MI. The gene cluster in this pathway in *E. coli O157*:*H7* was earlier proposed to have been acquired by HGT (Friedrich *et al.* 2005). The results of the INDeGenIUS method also predicted MIs in only two strains of *E. coli*, *E .coli O157* :*H7* and *E. coli O157: EDL933,* indicating an interstrain difference in the presence of an MI in *E. coli* (supplementary table 4).

All the genes belonging to the capsular polysaccharide synthesis pathway (figure 3) were identified by INDeGenIUS as 'non-native' and thus classified as probable MIs in 37 organisms. Cieslewicz *et al*. (2005) have suggested an HGT event as a mechanism for capsule variation in Group B Streptococcus. One of the genes in this pathway, *kfiD,* was suggested earlier to have been introduced into *E. coli* from an exogenous streptococcal species (Munoz *et al.* 1998).

All genes involved in the peptidoglycan synthesis pathway and all but one gene of the thiamine biosynthesis pathway were predicted in the GIs (figure 3) by INDeGenIUS in 102 organisms, suggesting that these two pathways were probably MIs in these bacteria. These two pathways were also obtained as seed sets in the studies by Borenstein *et al.* (2008), who had predicted a set of genes called metabolic network's 'seed set', which are exogenously acquired by the bacteria. Thus, the prediction of the above-mentioned pathways as MIs by INDeGenIUS is in line with the work of Borenstein *et al*. (2008).

Interestingly, 9 genes in the TCA cycle pathway (figure 3) were predicted as compositionally distinct in 15 organisms by INDeGenIUS. These genes, which form part of this predicted MI, correspond to *gltA, sdhCDAB* and *sucABCD.* In addition, interspecies and interstrain differences were observed in Pseudomonas, where this MI has been predicted in *Pseudomonas aeruginosa LESB58* and *UCBPP*, but not in the other two strains of *P. aeruginosa*. Similarly, this gene cluster is predicted as an MI in *Pseudomonas putida GB_1* and *W619*, and not in the *FI* strain. Thus, although a case of known HGT was not observed for this gene cluster, the region can be classified as a likely MI.

Acquisition of MIs plays a crucial role in the survival of bacteria since acquisition of additional metabolic traits increases their adaptability and competitiveness, such as colonization of a new niche or growth/survival under rapidly changing growth conditions. Certain metabolic pathways such as the LPS and capsule production pathways provide virulence characteristics to the bacteria and hence play a crucial role in the pathogenicity of these organisms. Genes in such pathways could act as drug targets for infections against these bacteria.

3.3.2 *Symbiotic islands:* INDeGenIUS predicted SyIs in 27 organisms distributed across different groups of proteobacteria. The organisms included *Azoarcus sp*. from beta-proteobacteria, *Acidithiobacillus ferroxidans, Azotobacter vinilandii* and *Klebsiella pnemoniae* from gamma-proteobacteria and *Rhizobium japonicum* from alpha-proteobacteria (supplementary table 3). The SyIs in these organisms encode either the nitrogen fixation genes (*nif* and *fix* genes) or the nodulation genes (*nod* genes). Among these organisms, SyIs are known to occur in members of the Rhizobial group which include Rhizobium, Bradyrhizobium, Azorhizobium, Mesorhizobium and

Sinorhizobium (Gonzalez *et al.* 2003). Other nitrogen-fixing organisms such as *K. pneumoniae* and *A. vinilandii,* predicted to have SyIs by INDeGenIUS, are known to be capable of association with the root system of graminaceous plants (Yan *et al.* 2008). In addition, Yan *et al.* (2008) have shown that Azoarcus and Gluconacetobacter are also capable of developing endophytic associations and survive within plant tissues without causing disease symptoms and carry out nitrogen fixation. It was proposed that *nif* genes were lost in most lineages of bacteria and archaea, and that HGT played a role in the acquisition of *nif* genes in some of these lineages (Yan *et al*. 2008).

Genes within the SyI are known to enhance bacterial fitness. Also, structures of the SyIs emphasize the remarkable similarities in the evolutionary strategies adopted by symbionts and pathogens in their quest for interaction with eukaryotic hosts. The results of the present study could help one in understanding the ecotypic polymorphism (stable coexistence of symbiotic and asymbiotic genotypes) adapted by some organisms such as Klebsiella.

3.3.3 *Resistance islands:* RIs were predicted in 22 organisms by INDeGenIUS. These organisms included various species and strains of Acinetobacter, Rickettsia, Acidovorax, Burkholderia, Coxiella, Erwinia, Escherichia, Gluconacetobacter, Klebsiella, Salmonella, Shigella, Vibrio and Yersinia (supplementary tables 2 and 3). These probable RIs encoded either specific antibiotic resistance or MDR genes. Most of the predicted RIs were seen in pathogenic organisms with the exception of the non-pathogenic Gluconoacetobacter which contains genes for MDR in its GI. All the above-mentioned organisms have been listed in the Antibiotic Resistance Genes Database (*http://ardb.cbcb.umd.edu*; Liu and Pop 2009). Some of these resistance gene clusters have already been reported in the literature as GIs. These include the SXT gene cluster of Vibrio, SGI1 cluster of Salmonella and SRL cluster of Shigella (Turner *et al.* 2001; Beaber *et al.* 2002, 2004; Doublet *et al.* 2002). An MDR cluster of 45 genes was picked up as one of the prominent peaks in *Acinetobacter baumannii AYE,* a resistant strain of Acinetobacter. However, no prominent peaks were obtained in the non-resistant strains (SDF or ATCC) of this bacterium.

With the emergence of increasingly resistant bacterial strains, management of several bacterial infections has become a major concern for the public health. Investigation of the predicted RIs in such bacterial strains would help in the identification and elimination of regions that have the potential to confer a drug-resistant phenotype to them.

3.3.4 *Secretion system islands:* Out of the SIs predicted by INDeGenIUS, 21, 111, 89, 80, 5 and 64 could be classified under Type I, Type II, Type III, Type IV, Type V and Type VI SIs, respectively (supplementary table 4). Analysis of the overall distribution of various predicted SIs indicated that while Types I and II SIs are equally present across all groups of proteobacteria, Types III and IV SIs are dominant in gamma- and alpha-proteobacteria, respectively (supplementary table 5). Although Type VI SI is predominant in gamma-proteobacteria, it is also observed in a large number in beta-proteobacteria. Overall, Type II SI is the most widely present SI in proteobacteria followed by Types III, IV, VI, I and V SIs. When prevalence was compared, Types I and II SIs were seen to be equally dominant in pathogenic as well as non-pathogenic organisms. On the other hand, proposed SIs encoding genes belonging to Types III, IV and VI secretion systems were seen to be predominant in pathogenic organisms, suggesting the importance of secretion systems in imparting pathogenicity to these organisms (supplementary table 5).

The Type I SI predicted in *S. enterica* by INDeGenIUS contained genes corresponding to the Salmonella pathogenicity island SPI-4 (Gerlach *et al.* 2007). Similarly, among the predicted Type II SIs, a gene *pilD* of this system is known to be acquired via HGT in *V. cholerae* (Sandkvist 2001). Likewise, the predicted Type III SI in many organisms is known to be a part of the PAI in organisms such as *E. coli*, *S. typhimurium* and *Yersenia spp.* (Mecsas and Strauss 1996). The predicted Type IV SI in *H. pylori* contained genes encoding the known *CAG* PAI (Odenbreit *et al.* 2000). Finally, the proposed Type VI SI in *V. cholerae* is known to be a part of the PAI (Pukatzki *et al.* 2006).

Since pathogenic bacteria use their specialized secretion system machineries to export virulent factors outside their cell walls, identification of the gene components of these systems is crucial for the development of new vaccines and treatment for a number of human pathogens. Thus, the proposed SIs in several human pathogens may have wide therapeutic application for infectious diseases.

3.3.5 *Pathogenicity islands:* The majority of genes within the PAIs predicted across 222 organisms corresponded to phage- and prophage-related genes (supplementary table 2). Such genes included the phage major/minor tail proteins, prophage/phage maintenance proteins and phage capsid proteins. This observation is supported by the fact that prophage and phage genes occurring within bacteria are acquired via HGT and are associated with virulence (Garcia-Vallvé *et al.* 2000). Bacteriophages are known to be responsible for promoting HGT between bacteria (Garcia-Vallvé *et al.* 2000). A number of bacterial genomes, including *E. coli O157, Bordetella pertussis* and *Burkholderia mallei,* contain multiple prophage-related genes (Asadulghani *et al.* 2009). The predicted PAIs (supplementary table 3) belonged to mostly pathogenic organisms. Examples of non-pathogenic proteobacteria identified to possess probable PAIs include *Gluconobacter oxydans, Magnetospirillum*

magneticum, Nitrobacter hamburgensis, etc. PAIs were also detected in most of the pathogenic strains of a number of bacteria. For example, prominent peaks corresponding to the VPI-I and VPI-II PAIs in *V. cholerae* were absent in the non-pathogenic strains *V. fischeri* and *V. harvey*.

The prophage-related genes in the proposed PAIs in several bacteria might help in understanding their role in these bacteria and might throw some light on the interstrain variability due to the integration of viral phages in these bacteria.

3.3.6 *Motility island:* Apart from the above island types, MoIs containing genes for flagellary, fimbriae and pili machinery were predicted across 249 organisms (supplementary table 2). The predicted MoIs in *Agrobacterium tumafaciens* and *Vibrio vulnificus* contained genes coding for type IV pili and mannose-sensitive haemagglutinin (MSHA) pilin proteins. The MoIs predicted in other organisms had a gene cluster of the flagellary system (supplementary tables 2 and 3). Although these were predicted widely across all proteo-bacteria, gamma-proteobacteria contained the maximum numbers.

Apart from providing motility to the bacteria, since flagella can act as an adhesive organelle, they can also contribute to the virulence of organisms. It is also seen that motility is essential in some phases of the life cycle of certain pathogens, and that virulence and motility are often intimately linked. This could help in exploiting the possibility of bacterial motility as a specific therapeutic antibacterial target to cure or prevent diseases.

3.3.7 *Unknown islands:* Apart from the above-mentioned categories of GIs, the 'unknown islands' category mostly contained hypothetical genes. Detailed analysis of such islands may throw some light on their probable biological role.

4. Discussion

In this paper, we describe a method called INDeGenIUS for the identification of GIs in a given genome sequence. This method outperforms the other known methods in detecting 'non-native' regions in any genome. The predicted GIs in 400 completely sequenced proteobacterial genomes contained functionally important regions. Based on these encoded functions, the GIs could be classified into 6 functional categories.

With an increasing number of completely sequenced bacterial genomes, the present algorithm will be useful in identifying functionally relevant GIs in these genomes. In addition, knowledge of GIs would allow one to better manipulate the useful functions contained in GIs for beneficial bacterial genetic engineering purposes and identification of candidate drug targets.

Acknowledgements

We would like to thank Hannah Rajasingh, Mohammed Monzoorul Haque and Tarini Shankar Ghosh for providing valuable suggestions.

References

Abe-yoshizumi R, Kamei U, Yamada A, Kimura M and Ichihara S 2004 The evolution of the phenylacetic acid degradation pathway in bacteria; *Biosci. Biotechnol. Biochem*. **68** 746–748

Asadulghani M, Ogura Y, Ooka T, Itoh T, Sawaguchi A, Iguchi A, Nakayama K and Hayashi T 2009 The defective prophage pool of *Escherichia coli O157*: prophage–prophage interactions potentiate horizontal transfer of virulence determinants; *PLoS Pathog*. **5** e1000408

Beaber J W, Hochhut B and Waldor M K 2002 Genomic and functional analysis of SXT, an integrating antibiotic resistance gene transfer element derived from *Vibrio cholerae*; *J. Bacteriol.* **184** 4259–4269

Beaber J W, Hochhut B and Waldor M K 2004 SOS response promotes horizontal dissemination of antibiotic resistance genes; *Nature (London)* **427** 72–74

Blum G, Ott M, Lischewski A, Ritter A, Imrich H, Tschäpe H and Hacker J 1994 Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen; *Infect. Immun.* **62** 606–614

Borenstein E, Kupiec M, Feldman M W and Ruppin E 2008 Large-scale reconstruction and phylogenetic analysis of metabolic environments; *Proc. Natl. Acad. Sci. USA* **105** 14482–14487

Carniel E 1999 The Yersinia high-pathogenicity island; *Int. Microbiol.* **2** 161–167

Chen W M, Moulin L, Bontemps C, Vandamme P, Béna G and Boivin-Masson C 2003 Legume symbiotic nitrogen fixation by β-proteobacteria is widespread in nature; *J. Bacteriol*. **185** 7266–7272

Chiu C H, Tang P, Chu C, Hu S, Bao Q, Yu J, Chou Y Y, Wangand H S *et al*. 2005 The genome sequence of *Salmonella enterica* serovar *Choleraesuis*, a highly invasive and resistant zoonotic pathogen; *Nucleic Acids Res.* **33** 1690–1698

Cieslewicz M J, Chaffin D, Glusman G, Kasper D, Madan A, Rodrigues S, Fahey J, Wessels M R *et al.* 2005 Structural and genetic diversity of Group B Streptococcus capsular polysaccharides; *Infect. Immun*. **73** 3096–3103

Desmond E, Brochier-Armanet C and Gribaldo S 2007 Phylogenomics of the archaeal flagellum: rare horizontal gene transfer in a unique motility structure; *BMC Evol. Biol*. **7** 106

Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev Microbiol*. **2** 414–424

Doublet B, Butaye P, Imberechts H, Boyd D, Mulvey M R, Chaslus-Dancla E and Cloeckaert A 2002 Salmonella genomic island 1 multi drug resistance gene clusters in *Salmonella enterica* serovar *Agona* isolated in Belgium in 1992 to 2002; *Antimicrob. Agents Chemother.* **48** 2510–2517

Finan T M 2002 Evolving insights: symbiosis islands and horizontal gene transfer; *J. Bacteriol*. **184** 2855–2856

Fournier P E, Vallenet D, Barbe V, Audic S, Ogata H, Poirel L, Richet H, Robert C *et al*. 2006 Comparative genomics of multi drug resistance in *Acinetobacter baumannii*; *PLoS Genet.* **2** e7

Friedrich A W, Köck R, Bielaszewska M, Zhang W, Karch H and Mathys W 2005 Distribution of the urease gene cluster among and urease activities of enterohemorrhagic *Escherichia coli O157* isolates from humans; *J. Clin. Microbiol.* **43** 546–550

Garcia-Vallvé S, Romeu A and Palau J 2000 Horizontal gene transfer in bacterial and archaeal complete genomes; *Genome Res*. **10** 1719–1725

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

Gerlach R G, Jackel D, Stecher B, Wagner C, Lupas L, Hardt W D and Hensel M 2007 Salmonella pathogenicity island 4 encodes a giant non-fimbrial adhesin and the cognate type I secretion system; *Cell Microbiol*. **9** 1834–1850

Ginolhac A, Jarrin C, Robe P, Perrière G, Vogel T M, Simonet P and Nalin R 2005 Type I polyketide synthases may have evolved through horizontal gene transfer; *J. Mol. Evol*. **60** 716–725

Gonzalez V, Bustos P, Ramirez-Romero M A, Medrano-Soto A, Salgado H, Hernandez-Gonzalez I, Hernandez-celis J C, Quintero V *et al*. 2003 The mosaic structure of the symbiotic plasmid of *Rhizobium etli CFN42* and its relation to other symbiotic genome compartment; *Genome Biol*. **4** R36

Hacker J, Bender L, Ott M, Wingender J, Lund B, Marre R and Goebel W 1990 Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal *Escherichia coli* isolates; *Microb. Pathog.* **8** 213–225

Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, Sasamoto S, Watanabe A, Idesawa K *et al*. 2000 Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*; *DNA Res.* **7** 331–338

Karlin S 2001 Detecting anamolous gene clusters and pathogenicity islands in diverse bacterial genomes; *Trends Microbiol*. **9** 335–343

Kostakioti M, Newman C L, Thanassi D G and Stathopoulos C 2005 Mechanisms of protein export across the bacterial outer membrane; *J. Bacteriol*. **187** 4306–4314

Lawrence J G and Ochman H 1998 Molecular archaeology of the *Escherichia coli* genome; *Proc. Natl. Acad. Sci. USA* **95** 9413–9417

Lawrence J G and Roth J R 1995 The cobalamin (coenzyme B12) biosynthetic genes of *Escherichia coli*; *J. Bacteriol.* **177** 6371–6380

Lawrence J G and Roth J R 1996 Selfish operons: horizontal transfer may drive the evolution of gene clusters; *Genetics* **143** 1843–1860

Lawrence J G and Roth J R 1996 Evolution of coenzyme B(12) synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex; *Genetics* **142** 11–24

Lima W C, Paquola A C M, Varani A M, Sluys M A V and Menck C F M 2008 Laterally transferred genomic islands in Xanthomonadales related to pathogenicity and primary metabolism; *FEMS Microbiol Lett*. **281** 87–97

Liu B and Pop M 2009 ARDB – Antibiotic Resistance Genes Database; *Nucleic Acids Res*. **37** D443– D447

Liu R 2009 *Origin and evolution of the bacterial flagellar system, pili and flagella: current research and future trends* (ed.) K Jarrell (Ontario, Canada: Caister Academic Press)

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

Mario J, Meer V D, Roelof J, Murriel G, Rosalind M H, Derek W H and Derrick W C 2009 Genomic islands: tools of bacterial horizontal gene transfer and evolution; *FEMS Microbiol Rev.* **33** 376–393

Mecsas J and Strauss E J 1996 Molecular mechanisms of bacterial virulence: Type III secretion and pathogenicity islands; *Emerg. Infect. Dis.* **2** 270–288

Mooi F R and Bik E M 1997 The evolution of epidemic *Vibrio cholerae* strains; *Trends Microbiol.* **5** 161–165

Muñoz R, García E and López R 1998 Evidence for horizontal transfer from *Streptococcus* to *Escherichia coli* of the *kfiD* gene encoding the K5-specific UDP-glucose dehydrogenase; *J. Mol. Evol*. **46** 432–436

Nag S, Chatterjee R, Chaudhuri K and Chaudhuri P 2006 Unsupervised statistical identification of genomic islands using oligonucleotide distributions with application to *Vibrio* genomes; *Sadhana* **31** 105–115

Nano F E, Zhang N, Cowley S C, Klose K E, Cheung K K M, Roberts M J, Ludu J S, Letendre G W *et al*. 2004 A *Francisella tularensis* pathogenicity island required for intramacrophage growth; *J. Bacteriol.* **186** 6430–6436

Odenbreit S, Püls J, Sedlmaier B, Gerland E, Fischer W and Haas R 2000 Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion; *Science* **287** 1497–1500

Ou H Y, Chen L L, Lonnen J, Chaudhuri R R, Thani A B, Smith R, Garton N J, Hinton J *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

Patil P B and Sonti R V 2004 Variation suggestive of horizontal gene transfer at a lipopolysaccharide (lps) biosynthetic locus in *Xanthomonas oryzae pv. oryzae,* the bacterial leaf blight pathogen of rice; *BMC Microbiol.* **4** 40

Patil P B, Bogdanove A J and Sonti R V 2007 The role of horizontal transfer in the evolution of a highly variable lipopolysaccharide biosynthesis locus in xanthomonads that infect rice, citrus and crucifers; *BMC Evol. Biol*. **7** 243

Poggio S, Abreu-Goodger C, Fabela S, Osorio A, Dreyfus G, Vinuesa P and Camarena L 2007 A complete set of flagellar genes acquired by horizontal transfer coexists with the endogenous flagellar system in *Rhodobacter sphaeroides*; *J. Bacteriol*. **189** 3208–3216

Pukatzki S, Amy T Ma, Derek S, Bryan K, David S, Nelson W C, Heidelberg JF and Mekalanos J J 2006 Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the Dictyostelium host model system; *Proc. Natl. Acad. Sci. USA* **103** 1528–1533

Rajan I, Aravamuthan S and Mande S S 2007 Identification of compositionally distinct regions in genomes using the centroid method; *Bioinformatics* **23** 2672–2677

Roth J R, Lawrence J G and Bobik T A 1996 COBALAMIN (COENZYME B12): synthesis and biological significance; *Annu. Rev. Microbiol*. **50** 137–181

Sandkvist M 2001 Type II secretion and pathogenesis; *Infect. Immun.* **69** 3523–3535

Schubert S, Picard B, Gouriou S, Heesemann J and Denamur E 2002 Yersinia high-pathogenicity island contributes to virulence in *Escherichia coli* causing extraintestinal infections; *Infect. Immun*. **70** 5335–5337

Shrivastava S and Mande S S 2008 Identification and functional characterization of gene components of type VI secretion system in bacterial genomes; *PLoS ONE* **3** e2955

Sullivan J T and Ronson C W 1998 Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates in to a phe-tRNA gene; *Proc. Natl. Acad. Sci. USA* **95** 5145–5149

Sullivan J T, Trzebiatowski J R, Cruickshank R W, Gouzy J, Brown S D, Elliot R M, Fleetwood D J, McCallum N G *et al*. 2002. Comparative sequence analysis of the symbiosis island of *Mesorhizobium loti* strain R7A; *J. Bacteriol.* **184** 3086–3095

Tumapa S, Holden M T, Vesaratchavest M, Wuthiekanun V, Limmathurotsakul D, Cheirakul W, Feil E J, Currie B J *et al*. 2008 *Burkholderia pseudomallei* genome plasticity associated with genomic island variation; *BMC Genomics* **9** 190

Turner S A, Luck S N, Sakellaris H, Rajakumar S and Adler B 2001 Nested deletions of the SRL pathogenicity island of *Shigella flexneri* 2a; *J. Bacteriol*. **183** 5535–5543

Vernikos G S, Thomson N R and Parkhill J 2007 Genetic flux over time in the *Salmonella* lineage; *Genome Biol.* **8** R100

Yan Y, Yang J, Dou Y, Chen M, Ping S, Peng J, Lu W, Zhang W *et al*. 2008 Nitrogen fixation island and rhizosphere competence traits in the genome of root-associated *Pseudomonas stutzeri A1501*; *Proc. Natl. Acad. Sci. USA* **105** 7564–7569

MS received 11 May 2010; accepted 16 July 2010

*e*Publication: 10 August 2010

Corresponding editor: Reiner A Veitia