GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Prediction of genomic islands in seven human pathogens using the Z-Island method

W. Wei and F.-B. Guo

School of Life Science and Technology,
University of Electronic Science and Technology of China, Chengdu, China

Corresponding author: F.-B. Guo
E-mail: fbguo@uestc.edu.cn

Genet. Mol. Res. 10 (4): 2307-2315 (2011)
Received November 22, 2010
Accepted April 26, 2011
Published October 5, 2011
DOI http://dx.doi.org/10.4238/2011.October.5.1

ABSTRACT. We adopted the method of Zhang and Zhang (the Z-Island method) to identify genomic islands in seven human pathogens, analyzing their chromosomal DNA sequences. The Z-Island method is a theoretical method for predicting genomic islands in bacterial genomes; it consists of determination of the cumulative GC profile and computation of codon usage bias. Thirty-one genomic islands were found in seven pathogens using this method. Further analysis demonstrated that most have the known conserved features; this increases the probability that they are real genomic islands. Eleven genomic islands were found to code for products involved in causing disease (virulence factors) or in resistance to antibiotics (resistance factors). This finding could be useful for research on the pathogenicity of these bacteria and helpful in the treatment of the diseases that they cause. In a comparison of the distribution of mobility elements in genomic islands predicted by different methods, the Z-Island method gave lower false-positive rates. The Z-Island method was found to detect more known genomic islands than the two methods that we compared it with, SIGI-HMM and IslandPick. Furthermore, it maintained a better balance between specificity and sensitivity. The only inconvenience is that the steps for finding genomic islands by the Z-Island method are semi-automatic.

Key words: Genomic island; Z-Island method; Virulence factor

INTRODUCTION

The horizontal transfer of alien genes pervades the process of bacterial evolution even dating back to their origins (Ochman et al., 2000). The event of gene insertion may lead to the emergence of new features, such as pathogenicity, in recipient bacteria. Moreover, the event of horizontal gene transfer may contribute to the birth of new species (Gogarten et al., 2002; Monier et al., 2009). The genes integrated into the native genome stem from other biological entities such as phages. In particular, large exogenous gene fragments in bacterial genomes constitute genomic islands (GIs). The core genes in GI are required to originate by horizontal transfer. It is interesting that GIs often contain genes associated with the survival of the organism under adverse conditions. GIs can be classified as pathogenicity islands (PAIs), symbiosis islands, metabolic islands, secretion islands, and resistance islands (Hentschel and Hacker, 2001; Do and Miyano, 2008). Among them, PAIs contain clusters of genes encoding virulence factors (VFs) such as overt toxins, adherence factors, secretion proteins, and molecules required for entry into the host cell or for acquisition of limiting metabolites (Hentschel and Hacker, 2001). There is also another method for classifying GIs. By this method, there are tRNA, tmRNA and non-RNA integrated GIs. GIs originating from tRNA/tmRNA integration always lie near tRNA/tmRNA sites, which are known as “hotspots” for integration. However, there are no such sites around non-RNA integrated GIs (Ou et al., 2006). Recently, sRNA (non-coding small RNA)-integrated GIs have also been discovered (Sridhar and Rafi, 2007).

GIs possess the following set of highly conserved characters (Vernikos and Parkhill, 2008). i) Sequence composition is different from that of recipient DNA. ii) Transferred genes carried in the island are large (usually 10-200 kb). iii) The border of insertion is usually adjacent to tRNA/tmRNA site. iv) GIs often have limited phylogenetic distribution, i.e., they exist in some genomes but are lacking in other closely related ones. v) When recombining with the host gene sequence, there may be specific fragments such as repeat sequences flanking GIs. vi) GIs often have mobility genes (e.g., integrase and transposase). However, most GIs just show parts of the typical characteristics.

In the past few years, intensive studies on the identification of GIs have been performed, based on characteristics of horizontally transferred genes. These distinct characteristics include GC content, codon usage, amino acid usage, dinucleotide usage, and tetranucleotide relative abundance values (Greub et al., 2004; Do and Miyano, 2008; Kanhere and Vingron, 2009). Assessing the change in GC content is an effective way to detect the horizontal gene transfer events. Traditionally, the distribution of GC content in genomes is calculated by a sliding-window method, which has been widely used. However, the proper window size is hard to adjust. Large window size brings about low resolution, whereas small window size leads to large statistical fluctuations. Recently, the cumulative GC profile, a windowless method for computing the GC content, has been proposed (Zhang and Zhang, 2004). Compared with the sliding-window method, higher resolution is obtained in detecting the change in genomic GC. The GIs in genomes of *Vibrio vulnificus*, *Corynebacterium glutamicum*, *C. efficiens*, and *Bacillus cereus* have been identified by using cumulative GC profile analysis (Charkowski, 2004; Zhang and Zhang, 2004, 2005). In this study, the cumulative GC profile method, combined with the computation of codon usage bias, was used to detect GIs in seven bacterial pathogens.

The GIs identified, particularly the PAIs, will be useful in research on the pathogenicity of these bacteria and helpful in the treatment of the diseases they cause. Here, seven important human pathogens were used to identify GIs, namely *Acinetobacter baumannii* AYE, *Brucella melitensis* 16M (genome containing two chromosomes), *Enterococcus faecalis* V583, *Helicobacter pylori* 26695, *Mycobacterium tuberculosis* CDC1551, *Neisseria meningitidis* Z2491, and *Streptococcus pneumoniae* TIGR4. *A. baumannii* invades the body by invasive devices and can cause many kinds of symptoms depending on which body site is infected, such as pneumonia (the lungs). *B. melitensis* causes brucellosis, a disease affecting humans, sheep and cattle. *E. faecalis* can cause a variety of nosocomial infections, of which urinary tract infections are the most common. *H. pylori* bacteria can cause digestive illnesses, including gastritis and peptic ulcer disease. *M. tuberculosis* is the causative agent in most cases of tuberculosis. *N. meningitidis* plays an important role in endemic bacterial meningitis. *S. pneumoniae* is the most common pathogen of pneumonia and meningitis.

MATERIAL AND METHODS

Databases

Full DNA sequences and related annotation information of genomes for seven human pathogens were downloaded from the NCBI ftp site (ftp.ncbi.nih.gov/genomes/). They are *A. baumannii* AYE (NCBI accession: NC_010410), *B. melitensis* 16M chromosome I and chromosome II (NC_003317, NC_003318), *E. faecalis* V583 (NC_004668), *H. pylori* 26695 (NC_000915), *M. tuberculosis* CDC1551 (NC_002755), *N. meningitidis* Z2491 (NC_003116), and *S. pneumoniae* TIGR4 (NC_003028). The above eight chromosomes were used to identify GIs. Futhermore, a dataset of virulence factors was obtained from the virulence factor database VFDB (Yang et al., 2008; http://www.mgc.ac.cn/VFs/). The results of three GI predictors, IslandPath-DIMOB, SIGI-HMM and IslandPick GI are available at the web site (http://www.pathogenomics.sfu.ca/islandviewer/query.php), which were used to compare with the Z-Island method.

A systematic method used to predict GIs

A systematic method, called Z-Island here, combining the cumulative GC profile and the computation of codon usage bias has been previously proposed by Zhang and Zhang (2004). In this study, this systematic method was used to predict GIs in the seven human pathogens. Here, we briefly summarize the method. Let,

$$Z_n = (A_n + T_n) - (C_n + G_n),\ n = 0,1,2,\ldots,N,\ Z_n \in [-N, N] \quad \text{(Equation 1)}$$

In the equation, A_n, C_n, G_n, and T_n are the cumulative numbers of the nucleotides A, C, G, and T, respectively, occurring in the subsequence from the first base to the n-th base in the DNA sequence under study. $Z_n \sim n$ could be plotted as a 2-D curve and we could use a straight line to fit it by using the least-squares approach. The slope of the so generated fitted line is denoted by *k*. We then get the Z'-curve, or cumulative GC profile, the coordinate of which is calculated as follows.

$$Z_n' = Z_n - kn \qquad \text{(Equation 2)}$$

The so-called cumulative GC profile could reflect the GC content variation along the DNA sequence. A spike in the cumulative GC profile indicates a decrease in GC content; otherwise, a drop in the curve indicates an increase in GC content.

The occurrence frequencies of 61 sense codons in a gene could be deemed as a 61-D vector. The average codon usage vector determined for all genes in a genome is denoted by *c*. Suppose the codon usage vector for the i-th gene in the investigated genome is denoted by c_i. The codon usage bias of this gene relative to the average vector could then be calculated by using the index of codon usage bias, cub_i.

$$cub_i = 1 - c_i \bullet c / (|c_i| \times |c|), cub_i \geq 0 \qquad \text{(Equation 3)}$$

In Equation 3, $| c_i |$ and $| c |$ are the modules of vectors c_i and *c*, respectively.

A nearly straight linear region in the Z'-curve denotes the deviation of the GC composition from a constant for a whole genome and it could be a candidate GI. The so-defined cub measures the relative codon usage variations in a potential GI compared with that of the whole genome. In this study, $P < 0.01$ is taken as the criterion for *t*-testing.

RESULTS AND DISCUSSION

GIs in seven human pathogens

With diversity and variety of microbial genomes sequenced, abundant GIs of probable horizontal origins have been discovered. In particular, it is worth noting that many GIs are associated with VFs. The establishment of infection is mediated by VFs, which are bacterial products that contribute to the ability of pathogens to cause disease. Therefore, we investigated GIs and their composition information in seven important human pathogens (Ho et al., 2009). As a consequence of varied sequence composition of different bacterial lineages, GIs usually have remarkably distinct sequence composition from a new host genome (Langille et al., 2010). Depending on this fact, the Z-Island method had been used to successfully identify eight GIs on three chromosomes since its first proposal by Zhang and Zhang (2004). As shown by the Z'-curve of *A. baumannii* AYE (Figure 1), most regions have large fluctuation due to their inhomogeneous GC content, but some regions show nearly straight lines with abrupt slopes, which could represent potential GIs. Further stastistical analysis showed that cub values of two candidates were significantly ($P < 0.01$) different from that of the rest of the chromosome. That is, these two fragments were predicted to be GIs by the Z-Island method. The Z'-curves for the other seven chromosomes showing the same situation are avaliable at http://cobi.uestc.edu.cn/resource/GIs/. After analyzing genomic sequences with the Z-Island method, 31 GIs were detected in seven microbes based on the fact that their GC content and codon usage tended to be different compared to the native backbone. Detail information of these GIs is listed in Table 1.

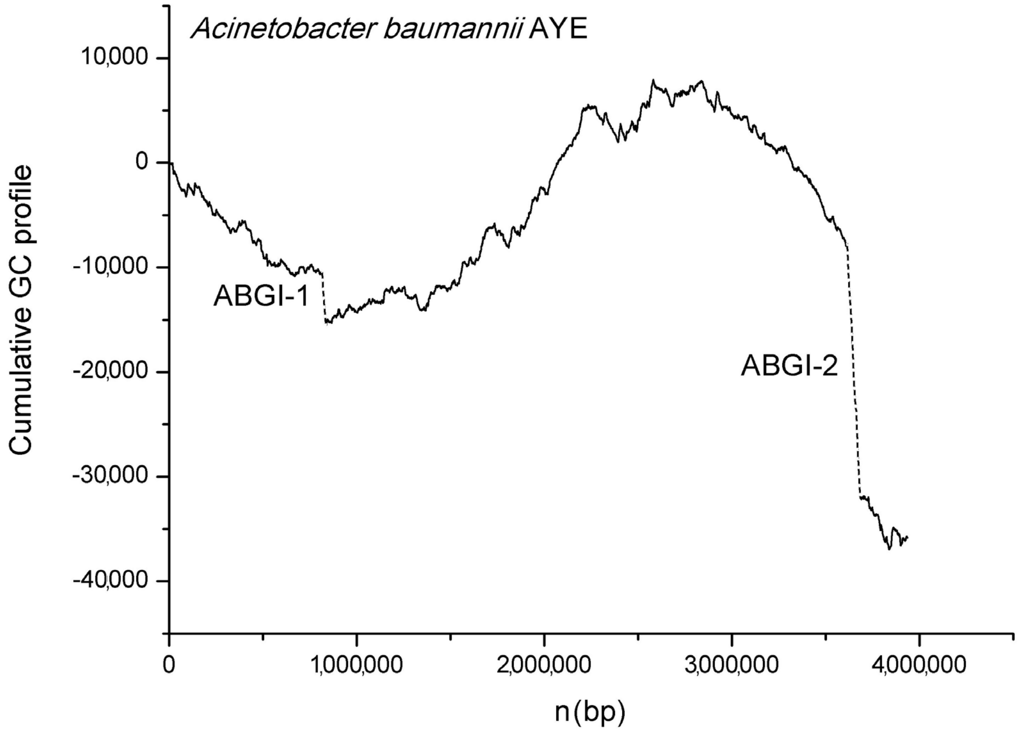

Figure 1. The Z'-curve (or cumulative GC profile) for genome of *Acinetobacter baumannii* AYE. Two identified genomic islands are denoted by dotted lines.

True GIs usually share some other apparently highly conserved features besides large size and composition bias. Here, we investigated two of them as follows. Mobility element is frequently involved in the mobilization of genomic DNA, and tRNA loci may act as targets for the integration of foreign DNA. Vernikos and Parkhill (2008) found that a mobility element is very important for a GI structural model and that tRNA-integrated loci are also important. The GIs predicted by the Z-Island method were found to fit this fact: 20 mobility element-embedded GIs were found, and nine GIs were associated with tRNA. Only eight GIs lacked these two conserved features. To conclude, all Z-Island GIs had at least the two features of large size (exceeding 8 kb) and abnormal composition, and most (74%) of them involved embedded mobility elements and/or an integrating tRNA locus.

The Z-Island method predicts not only novel GIs but also the already characterized ones. According to the pathogenicity island database PAIDB (Yoon et al., 2007; http://www.gem.re.kr/paidb/), there are five well-documented GIs, namely AbaR1, EF PAI, cag PAI, PPI-1, and IHT-A, in these seven pathogens. A comparision of predicted GIs and well-documented ones indicates that ABGI-2, EFGI-2, SPGI-3, and NMGI-1 overlap with four known GIs, AbaR1, EF PAI, PPI-1, and IHT-A, respectively. Only the cag PAI was missed by the Z-Island method.

Table 1. Genomic islands identified in seven human pathogens and their composition information.

Predicted genomic island	Start	End	Length	No. of genes	Codon usage bias ± SD	P	RNA	Mobility element	Type of VF/RF	Number of VF and RF
Acinetobacter baumannii AYE										
ABGI-1	813035	837637	24603	1	0.333 ± 0	P < 0.001				
ABGI-2	3612223	3689884	77662	80	0.331 ± 0.143	P < 0.001		○	Antibiotic resistance (D)	8
Brucella melitensis 16M chromosome I										
BMIGI-1	46200	74693	28494	26	0.182 ± 0.113	P < 0.01	Ala tRNA			
BMIGI-2	1033326	1075186	41861	46	0.174 ± 0.089	P < 0.001	Val tRNA	○	Intracellular survival (NS)	2
BMIGI-3	1263121	1281312	18192	18	0.183 ± 0.067	P < 0.001	Pro tRNA	○		
BMIGI-4	1448164	1469946	21783	32	0.205 ± 0.063	P < 0.001	Gln tRNA	○	Intracellular survival (NS)	11
BMIGI-5	1708913	1745717	36805	51	0.214 ± 0.096	P < 0.001		○		
Brucella melitensis 16M chromosome II										
BMIIGI-1	170442	198894	28453	30	0.161 ± 0.081	P < 0.01		○		
BMIIGI-2	743174	759720	16547	21	0.170 ± 0.082	P < 0.01		○		
Enterococcus faecalis V583										
EFGI-1	125857	164286	38430	45	0.146 ± 0.061	P < 0.01		○	Adherence (O)	1
EFGI-2	434221	600450	166230	150	0.143 ± 0.071	P < 0.001		○	Adherence and toxin (O)	5
EFGI-3	2198027	2261565	63539	60	0.297 ± 0.112	P < 0.001		○	Antibiotic resistance (D)	2
Helicobacter pylori 26695										
HPGI-1	1040628	1072332	31705	31	0.142 ± 0.068	P < 0.01	Pro tRNA	○		
Mycobacterium tuberculosis CDC1551										
MTGI-1	332822	342155	9334	10	0.312 ± 0.176	P < 0.01				
MTGI-2	831137	852854	21718	27	0.188 ± 0.100	P < 0.001	Thr tRNA	○		
MTGI-3	1431151	1488246	57096	51	0.186 ± 0.141	P < 0.001		○	Nutrient acquisition (NS)	1
MTGI-4	2535626	2639909	104284	117	0.155 ± 0.100	P < 0.001		○	Toxin (O)	3
Neisseria meningitidis Z2491										
NMGI-1	178197	186639	8443	8	0.270 ± 0.093	P < 0.01			Capsule (D)	8
NMGI-2	298110	311587	13478	21	0.304 ± 0.128	P < 0.001				
NMGI-3	392992	401893	8902	6	0.231 ± 0.049	P < 0.01	Ile tRNA			
NMGI-4	664694	690186	25493	19	0.276 ± 0.139	P < 0.001				
NMGI-5	757457	776712	19256	20	0.244 ± 0.097	P < 0.001		○		
NMGI-6	1022454	1035910	13457	21	0.281 ± 0.129	P < 0.001	Thr tRNA			
NMGI-7	1521454	1532377	10924	9	0.268 ± 0.114	P < 0.01		○	Toxin (O)	1
NMGI-8	1732700	1745056	12357	11	0.237 ± 0.082	P < 0.01		○		
NMGI-9	2066691	2075326	8636	12	0.266 ± 0.094	P < 0.001				
Streptococcus pneumoniae TIGR4										
SPGI-1	504613	518366	13754	22	0.244 ± 0.112	P < 0.001		○		
SPGI-2	661717	673075	11359	16	0.210 ± 0.074	P < 0.01				
SPGI-3	972795	1002507	29713	36	0.187 ± 0.080	P < 0.01		○	Iron uptake (NS)	1
SPGI-4	1227766	1274595	46830	55	0.184 ± 0.084	P < 0.01	Arg tRNA	○		
SPGI-5	1678227	1692557	14331	1	0.680 ± 0	P < 0.001				

t-test. VF = virulence factors; RF = resistance factors; "○" denotes the presence of mobility element. VFs are categorized, according to the VFDB database as O = offensive; D = defensive; NS = nonspecific.

Virulence factors in GIs predicted by the Z-Island method

Due to the different functions of proteins that are encoded by GIs, they may be classified as PAIs, symbiosis islands, metabolic islands, and resistance islands. Here, PAIs and resistance islands in these seven pathogens were investigated, and this could be helpful in the study of the pathogenesis of the bacteria or of the treatment of related diseases. PAIs are regarded as typical GIs, which encode clusters of genes that contribute to virulence. For increasing chances of survival from the effects of antibiotics, a resistance island encodes one or more proteins with antibiotic resistance function. VFs are generally categorized as "offensive", "defensive", "nonspecific", and "regulation". After retrieval with the VFDB database, 33 VFs are detected in 9 PAIs found by the Z-Island method. Among them, PAIs EFGI-1 and EFGI-2 contain five VFs that are associated with adherence. PAIs EFGI-2, MTGI-4 and NMGI-7 contain eight toxin proteins. The two types of VFs carry out an "offensive" function. Comparatively, the expressions of eight "defensive" capsule proteins in NMGI-1 have the functions to mediate the resistance to phagocytosis and block the Opa- or Opc-mediated invasion into host cells. "Nonspecific" comprises another class of VFs. LPS containing PAIs BMIGI-2 and BMIGI-4, lysA gene encoding GI MTGI-3 and iron uptake function encoding GI SPGI-3 correspond to this class. Ho et al. (2009) found that most VFs present in GIs have "offensive" functions. However, among the 33 VFs occurring in nine Z-Island PAIs, only 10 encode "offensive" functions. The consistency may be due to the few GI samples and the absence of the most commom VFs, type III/IV secretion factors, in this study. Furthermore, another search for antibiotic resistance genes shows that two GIs (ABGI-2 and EFGI-3) predicted by the Z-Island method possess resistance factor genes, which encode "defensive" products for surviving in the presence of antibiotics.

Comparison of the Z-Island method with other existing methods for predicting GIs

Just like the Z-Island method, SIGI-HMM (Waack et al., 2006) and IslandPath-DIMOB (Hsiao et al., 2005) are two of the composition-based methods for predicting GIs. In contrast, IslandPick (Langille et al., 2008) is a comparative genomic approach. After thorough comparison, three composition-based methods obtained quite consistent results. Specifically, the Z-Island method detected 61% of SIGI-HMM GIs and 39% of Z-Island GIs was found by SIGI-HMM. Similarly, 38% of IslandPath-DIMOB GIs coincided with 52% of Z-Island GIs. However, the prediction of the Z-Island method overlaps very little with that of the comparative genomic method IslandPick. Exactly, just one Z-Island GI was found to overlap partly with two IslandPick GIs.

The distribution of mobility elements in GIs was examined to measure the false-positive rate of the different methods. In this section, only SIGI-HMM, IslandPick and the Z-Island methods are compared because the results of IslandPath-DIMOB exclude all the candidate GIs that do not have any mobility elements at the first step (Hsiao et al., 2005). Therefore, all GIs predicted by IslandPath-DIMOB harbor at least one mobility element. Comparison results show that a higher proportion of mobility elements were discovered in the Z-Island GIs. Exactly 65% of Z-Island GIs have mobility elements, whereas the percentages for SIGI-HMM and IslandPick are 39 and 38%, respectively. As widely accepted, the mobility element is one of the important and highly conserved features of GIs. Therefore, the above analysis suggests that the Z-Island method may have a low false-positive rate than the other two.

As mentioned above, there are five well-documented GIs in the seven pathogenes.

These known GIs could be used as test sets for eveluating the accuracy of the prediction methods. As shown in Table 2, the Z-Island method correctly found four of the five known GIs. Comparatively, IslandPath-DIMOB found two of the five, and only one known GI was found by SIGI-HMM and IslandPick. Furthermore, cag PAI as the only GI missed by the Z-Island method, was also not found by the other three methods. Therefore, the Z-Island method has the highest accuracy among the four methods, based on the prediction results on the five well-documented GIs. Obviously, we regard five well-documented GIs as a positive sample of the test set. To evaluate the specificity of the four methods, we will choose the n/2 bases just upstream of it and the same number of bases just downstream of it as negative samples, when one well-documented GI has the length of *n* bases. Similarly, we will generate negative samples for each correctly found GI. Consequently, true samples and negative samples in the test sets contain the same number of bases. Based on the test set, the specificities of the four methods were also calculated and shown in Figure 2. As can be seen from this figure, the Z-Island method gave the best balance between the sensitivity and specificity indices. In conclusion, the Z-Island was shown to be an excellent method for predicting GIs in bacterial genomes. The only fly in the ointment is that the steps of picking up GIs are semi-automatic.

Table 2. Finding or not of five well-decumented islands by the predicting method.

Well-decumented island	Z-Island	SIGI-HMM	IslandPick	IslandPath-DIMOB
AbaR1	○		○	○
EF PAI	○	○		○
PPI-1	○			
IHT-A	○			
cag PAI				

"○" denotes the correct prediction of the well-documented genomic islands by the corresponding method.

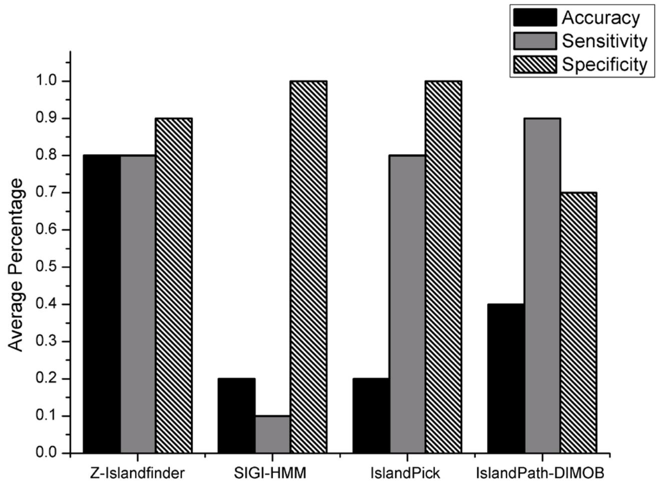

Figure 2. Result comparing five well-document genomic islands (GIs) with GIs predicted by four GI predictors. Black, gray and hatched bars represent average accuracy, sensitivity and specificity, respectively. All bases in well-documented GIs are donoted as positive group and bases in non-GI regions are regarded as negative group. Accuracy measures the percentage of the number of GIs correctly found by the GI predictor. If one well-documented GI partly overlaps with the result of the method, then it will be taken as correctly found. Specificity and sensitivity are caculated based on only those GIs correctly found by one method. Specificity is measured using the formula: true positives / (true positives + false positives). Sensitivity is measured using the formula: true positives / (true positives + false negatives).

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (grants #60801058 and #31071109), Committee of Science and Technology of Sichuan Province (grant #2008JY0053), and the Fundamental Research Funds for the Central Universities of China (grant #ZYGX2009J082).

REFERENCES

Charkowski AO (2004). Making sense of an alphabet soup: the use of a new bioinformatics tool for identification of novel gene islands. Focus on "identification of genomic islands in the genome of *Bacillus cereus* by comparative analysis with *Bacillus anthracis*". *Physiol. Genomics* 16: 180-181.

Do JH and Miyano S (2008). The GC and window-averaged DNA curvature profile of secondary metabolite gene cluster in *Aspergillus fumigatus* genome. *Appl. Microbiol. Biotechnol.* 80: 841-847.

Gogarten JP, Doolittle WF and Lawrence JG (2002). Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* 19: 2226-2238.

Greub G, Collyn F, Guy L and Roten CA (2004). A genomic island present along the bacterial chromosome of the *Parachlamydiaceae* UWE25, an obligate amoebal endosymbiont, encodes a potentially functional F-like conjugative DNA transfer system. *BMC Microbiol.* 4: 48.

Hentschel U and Hacker J (2001). Pathogenicity islands: the tip of the iceberg. *Microb. Infect.* 3: 545-548.

Ho Sui SJ, Fedynak A, Hsiao WW, Langille MG, et al. (2009). The association of virulence factors with genomic islands. *PLoS One* 4: e8094.

Hsiao WW, Ung K, Aeschliman D, Bryan J, et al. (2005). Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet.* 1: e62.

Kanhere A and Vingron M (2009): Horizontal gene transfers in prokaryotes show differential preferences for metabolic and translational genes. *BMC Evol. Biol*. 9: 9.

Langille MG, Hsiao WW and Brinkman FS (2008). Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* 9: 329.

Langille MG, Hsiao WW and Brinkman FS (2010). Detecting genomic islands using bioinformatics approaches. *Nat. Rev. Microbiol.* 8: 373-382.

Monier A, Pagarete A, de Vargas C, Allen MJ, et al. (2009). Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Res.* 19: 1441-1449.

Ochman H, Lawrence JG and Groisman EA (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405: 299-304.

Ou HY, Chen LL, Lonnen J, Chaudhuri RR, 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.

Sridhar J and Rafi ZA (2007). Identification of novel genomic islands associated with small RNAs. *In Silico Biol.* 7: 601-611.

Vernikos GS and Parkhill J (2008). Resolving the structural features of genomic islands: a machine learning approach. *Genome Res.* 18: 331-342.

Waack S, Keller O, Asper R, Brodag T, et al. (2006). Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* 7: 142.

Yang J, Chen LH, Sun LL, Yu J, et al. (2008). VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res.* 36: D539-D542.

Yoon SH, Park YK, Lee S, Choi D, et al. (2007). Towards pathogenomics: a web-based resource for pathogenicity islands. *Nucleic Acids Res.* 35: D395-D400.

Zhang R and Zhang CT (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.

Zhang R and Zhang CT (2005). Genomic islands in the *Corynebacterium efficiens* genome. *Appl. Environ. Microbiol.* 71: 3126-3130.