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

Statistical analyses of conserved features of genomic islands in bacteria

F.-B. Guo, Z.-K. Xia, W. Wei and H.-L. Zhao

Center of Bioinformatics and Key Laboratory for NeuroInformation of the Ministry of Education, 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. 13 (1): 1782-1793 (2014)
Received December 7, 2012
Accepted August 15, 2013
Published March 17, 2014
DOI http://dx.doi.org/10.4238/2014.March.17.6

ABSTRACT. We performed statistical analyses of five conserved features of genomic islands of bacteria. Analyses were made based on 104 known genomic islands, which were identified by comparative methods. Four of these features include sequence size, abnormal G+C content, flanking tRNA gene, and embedded mobility gene, which are frequently investigated. One relatively new feature, G+C homogeneity, was also investigated. Among the 104 known genomic islands, 88.5% were found to fall in the typical length of 10-200 kb and 80.8% had G+C deviations with absolute values larger than 2%. For the 88 genomic islands whose hosts have been sequenced and annotated, 52.3% of them were found to have flanking tRNA genes and 64.7% had embedded mobility genes. For the homogeneity feature, 85% had an *h* homogeneity index less than 0.1, indicating that their G+C content is relatively uniform. Taking all the five features into account, 87.5% of 88 genomic islands had three of them. Only one genomic island had only one conserved feature and none of the genomic islands had zero features. These statistical results should help to understand the general structure of known genomic islands. We found that larger genomic islands tend to have relatively small G+C deviations relative to absolute values. For example, the absolute G+C deviations of 9 genomic islands longer than 100,000 bp were all less than 5%. This is a

novel but reasonable result given that larger genomic islands should have greater restrictions in their G+C contents, in order to maintain the stable G+C content of the recipient genome.

Key words: Genomic islands (GIs); Conserved feature; G+C deviation; Homogeneity index

INTRODUCTION

Genomic island (GI), as a general term, refers to any cluster of genes that has been acquired by horizontal transfer, and they generally are 10-200 kb in length (Dobrindt et al., 2004). Due to the alien origin, GIs appear in only a few isolates or strains of one specific bacterial species. However, there may still be some GIs that exist in almost all sequenced strains of one species because they have been integrated into the ancestral host before its split or have been integrated into different hosts one after another during the post-split course (Guo et al., 2012). According to their functions, GIs could be divided into pathogenicity islands, secretion islands, antimicrobial resistance islands, metabolic islands and symbiotic islands (Hentschel et al., 2000). Among them, pathogenicity islands obviously attract the most attention because of their importance to human pathogens. GIs usually encode dispensable and accessory functions. They are often associated with microbial adaptations when infecting hosts or living in specific niches, and they have had a substantial impact on bacterial evolution (Hacker and Carniel, 2001). Therefore, there is a growing interest to efficiently identify GIs in newly sequenced bacterial genomes. To gain insight into differences between closely related bacterial species or strains, the identification of GIs in newly sequenced genomes is becoming a common first step (Langille et al., 2010).

Generally, there are two categories of methods to identify genomics islands in bacterial hosts: those that are based on sequence composition bias and those that use comparative genomics (Langille et al., 2010). Both types have their pros and cons. For example, the prominent fault of the composition based method is that it usually generates more false-positive predictions. As for the latter, it will be unfeasible if there are not too many genomes of closely related strains and suitable outgroup references. In fact, in most studies where the detection of GIs is the main topic, both composition biases and the presence of fundamental features are investigated to confirm the presence of a GI. During the actual annotation of GIs, the presence of one or several fundamental features could provide additional proof for a newly predicted GI. Those features include the typical size of 10-200 kb, the appearance of mobility genes (e.g., integrases and transposases), proximal transfer RNAs, and abnormal G+C content (Vernikos and Parkhill, 2008; Soares et al., 2012). In addition, a homogeneity feature has been noticed by Zhang and colleagues (Chen, 2006; Zhang and Zhang, 2004a,b, 2005, 2008; Wei and Guo, 2011; Guo and Wei, 2012). In this study, we performed a statistical analysis of these features based on 104 known GIs that were identified by comparative genomics and put emphasis of G+C deviation and the homogeneous index.

MATERIAL AND METHODS

Data source

All the GIs involved in this work are known GIs that have been well studied. These

GIs are frequently used in related works and they could constitute fairly reliable datasets. From the database PAIDB (http://www.gem.re.kr/paidb/) (Yoon et al., 2007), we obtained 63 such GIs. In addition, 41 known GIs were obtained by automatic PubMed search and subsequent manual check. In total, 104 well-documented GIs were prepared to perform through analysis in this work. Details of all these GIs are listed in Table 1. DNA sequences and annotation information for bacterial hosts of these GIs were downloaded from NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/).

There are essentially two main theoretical approaches for identifying GIs (Langille et al., 2010). One is based on sequence composition and the other on comparative genomics. Compared with the former, the latter relies on the very definition of HGT (rather than on its outcome) and hence has lower rates of false-positive predictions. To achieve reliable analyses and results, all GIs involved in this work are those identified by the comparative genomic method.

Conserved features

Four types of widely accepted conserved features are analyzed and compared. First, GIs have the typical length of 10-200 kb (Vernikos and Parkhill, 2008; Soares et al., 2012). We wanted to know what proportion of them would fall out this range. Second, GIs, particularly those recently inserted, tend to have a distinct composition with the host genome and this feature is generally measured by G+C deviation (Khrustalev and Barkovsky, 2011; Soares et al., 2012; Vernikos and Parkhill, 2008), which is defined as

$$G{+}C\ deviation = G{+}C_{host} - G{+}C_{gi} \qquad \text{(Equation 1)}$$

However, there are still some GIs that have similar G+C content with the host and we wanted to determine the specific figure. Third, transposases and integrases, as mobility genes, could aid the integration of GIs into the hosts (Vernikos and Parkhill, 2008; Soares et al., 2012). Hence, most GIs contain mobility genes and the particular proportion was investigated in this study. Finally, tRNA genes, as another type of marker gene, often flank GI regions (Vernikos and Parkhill, 2008; Soares et al., 2012).

Cumulative GC profile and h index

Besides the above frequently-mentioned features of GIs, we also wanted to investigate one relatively new feature, which has been used as one of the standards for predicting novel GIs by Zhang and colleagues (Chen, 2006; Zhang and Zhang, 2004a,b, 2005, 2008; Wei and Guo, 2011; Guo and Wei, 2012). The homogeneity feature was obtained based on the cumulative GC profile (Zhang et al., 2001). The method of the cumulative GC profile proposed by Zhang and colleagues has been used to identify GIs in dozens of prokaryotic genomes (Charkowski, 2004; Greub et al., 2004; Chen, 2006; Do and Miyano, 2008; Zhang and Zhang, 2004a,b, 2005, 2008; Guo and Wei, 2012; Wei and Guo, 2011). The method is described briefly as follows.

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

Table 1. Details of 104 known genomic islands (GIs): host name, GI name, GI length, GC deviation, h index, mobility gene percentage, tRNA gene presence, the number of meeting features.

Host	Genomic island	Length	G+C deviation	h	Mobility percentage (%)	tRNA	Combining presence number
Acinetobacter baumannii AYE	AbaR1	86190	-0.142	0.063	16.7		4
Bacillus cereus ATCC 10987	BCEGI-1	38294	0.045	0.009	31.3		4
	BCEGI-2	28732	0.040	0.013	0		3
Bacillus cereus ATCC 14579	BCGI-1	15929	0.050	0.004	7.1	+	5
	BCGI-2	62220	-0.024	0.009	0		3
	BCGI-3	48461	0.051	0.016	11.1		4
Bordetella petrii DSM 12804	BPGI-1	255480	0.038	0.189	11.9	+	3
	BPGI-2	143396	0.049	0.086	11.3		4
	BPGI-3	102094	0.025	0.034	16.2	+	5
	BPGI-4	47050	0.017	0.047	0		2
	BPGI-5	67704	0.055	0.057	15.3		4
	BPGI-6	159080	0.043	0.088	7.0	+	5
	BPGI-7	88795	0.080	0.093	5.7	+	5
Enterococcus faecalis V583	EFGI	137505	0.047	0.222	3.3	+	4
Escherichia coli 536	PAI I	76906	0.045	0.141	19.4	+	4
	PAI II	101767	0.035	0.296	11.3	+	4
	PAI III	75148	0.034	0.093	23.0	+	5
	PAI V	106245	0.028	0.214	5.6	+	4
Escherichia coli CFT073	PAI I	44642	0.039	0.064	9.1	+	5
	PAI II	58294	0.032	0.058	10.4	+	5
Escherichia coli O157:H7 EDL933	LEE	42794	0.094	0.176	3.8	+	4
Francisella tularensis SCHU S4	FPI	33579	0.012	0.199	5.6		2
Helicobacter hepaticus ATCC 51449	HHGI1	71027	0.028	0.075	1.4		4
Helicobacter pylori 26695	cag	38023	0.030	0.088	0		3
Neisseria meningitidis MC58	IHT-A	34084	0.054	0.362	0		2
	IHT-C	32553	0.086	0.054	3.7		4
Pseudomonas aeruginosa UCBPP-PA14	PAPI-1	115486	0.062	0.042	3.4	+	5
	PAPI-2	14670	0.085	0.021	16.3	+	5
Pseudomonas syringae tomato DC3000	Hrp	49467	0.009	0.035	2.0	+	4
Salmonella typhi CT18	SPI-1	41851	0.064	0.035	0		3
	SPI-2	41605	0.047	0.105	0	+	3
	SPI-3	16941	0.047	0.034	0	+	5
	SPI-4	23391	0.072	0.125	0		2
	SPI-5	7496	0.084	0.009	12.5	+	4
	SPI-6	58666	0.006	0.162	0	+	2
	SPI-7	133638	0.024	0.128	4.2	+	4
	SPI-8	6885	0.138	0.031	0	+	3
	SPI-9	15696	-0.056	0.011	0		3
	SPI-10	32934	0.055	0.058	4.3	+	5
	SPI-15	6364	0.030	0.004	20.0	+	4
	SPI-16	4478	0.100	0.013	0	+	3
	SPI-17	5122	0.136	0.021	0	+	3
Shigella flexneri 2a 2457T	SHI-1	45308	0.016	0.070	9.7	+	4
	SHI-2	28795	0.025	0.026	41.7	+	5
Shigella sonnei 53G	SSGI-1	39431	-0.008	0.079	9.1	+	4
	SSGI-2	81961	0.005	0.148	18.8	+	3
	SSGI-3	52985	-0.003	0.074	24.6	+	4
	SSGI-4	10962	0.104	0.025	35.7	+	5
	SSGI-5	26466	0.030	0.018	44.4	+	5
	SSGI-6	10172	0.112	0.046	0		3
	SSGI-7	16296	-0.015	0.035	55.0	+	4
	SSGI-8	13471	-0.029	0.020	14.3		4
Staphylococcus aureus COL	vSa1	15316	0.015	0.013	3.7		3
Staphylococcus aureus MRSA252	SaPI4	14394	0.016	0.018	4.8		3
Staphylococcus aureus Mu50	vSa3	14652	0.013	0.016	4.5		3
	vSa4	15111	0.023	0.025	5.0		4
	vSaα	26626	0.035	0.050	7.7	+	5
	vSaß	21541	0.031	0.030	4.4	+	5
	vSaγ	29415	0.027	0.010	3.2		4

Continued on next page

Table 1. Continued.

Host	Genomic island	Length	G+C deviation	h	Mobility percentage (%)	tRNA	Combining presence number
Staphylococcus aureus RF122	SaPIbov	15964	0.014	0.008	4.3		3
	SaPIbov2	17934	0.027	0.026	3.8		4
Staphylococcus epidermidis ATCC 12228	vSe1	4250	0.044	0.011	20.0		3
	vSe2	39376	0.060	0.059	2.7	+	5
	vSeγ	2665	0.014	0.003	0		1
Staphylococcus haemolyticus JCSC1435	vSh1	10870	0.027	0.020	14.3		4
	vSh2	16326	0.039	0.034	3.6		4
	vSh3	14707	0.015	0.019	20.0		3
Staphylococcus pneumoniae TIGR4	PPI-1	28349	0.066	0.076	11.4		4
Streptococcus agalactiae NEM316	SAPAI-1	18462	0.017	0.027	4.2	+	4
	SAPAI-2	14571	0.045	0.016	0	+	4
	SAPAI-3	47068	-0.020	0.042	0	+	4
	SAPAI-4	18779	0.027	0.008	0	+	4
	SAPAI-5	9439	0.063	0.022	0	+	3
	SAPAI-6	57739	0.009	0.016	0		2
	SAPAI-7	47068	-0.020	0.042	0		3
	SAPAI-8	47070	-0.020	0.018	0		3
	SAPAI-9	25536	0.014	0.015	0		2
	SAPAI-10	33445	-0.025	0.011	0		3
	SAPAI-11	7497	0.050	0.008	0	+	3
	SAPAI-12	81525	-0.016	0.069	0		2
	SAPAI-13	44563	0.042	0.06	0	+	4
	SAPAI-14	22494	0.030	0.022	0		3
Vibrio cholerae O1 biovar El Tor N16961	VPI	40883	0.121	0.033	6.9		4
	VPI-2	57170	0.054	0.035	6.0	+	5
	VSP-I	14038	0.083	0.029	9.1		4
	VSP-II	7447	0.084	0.014	0		2
Xanthomonas campestris vesicatoria 85-10	Hrp	23095	0.006	0.048	0		2
Yersinia pestis KIM	HPI	35904	-0.089	0.032	15.8	+	5
Bacteroides fragilis 86-5443-2-2	BfPAI	8592	0.033				
Clostridium difficile VPI10463	PaLoc	26039	0.024				
Dichelobacter nodosus A198	vap locus	12828	0.034				
	vrl locus	28106	-0.136				
Neisseria gonorrhoeae MS11	GGI	57358	0.082				
Photorhabdus luminescens W14	mcf	35876	-0.068				
	PAI I	24401	-0.134				
	PAI II	35280	0.033				
	PAI III	47740	0.011				
	tcd	127816	0.010				
Pseudomonas aeruginosa C	PAGI-2(C)	158230	0.007				
Pseudomonas aeruginosa SG17M	PAGI-3(SG)	128136	0.055				
Pseudomonas syringae phaseolicola 1302A	PPHGI-1	113527	0.049				
Salmonella typhimurium DT104	SGI1	47723	0.037				
Yersinia enterocolitica W22703	tc-PAIYe	20403	0.015				
Yersinia pseudotuberculosis 32777	YAPI	102752	0.003				

In the equation above, A_n, C_n, G_n and T_n, are the cumulative numbers of the bases *A*, *C*, *G*, and *T*, respectively, occurring in the subsequence from the first base to the n-th base in the inspected DNA sequence with length *N*. Z_n is one of the components of the *Z* curve (Guo et al., 2003). To amplify the deviations of Z_n, the curve of Z_n~n is fitted by a straight line using the least-squares approach.

$$Z = k \times n \qquad \text{(Equation 3)}$$

In equation (Equation 3), (Z, n) are the coordinates of a point on the straight line fitted and k is its slope. Instead of using the curve of $Z_n \sim n$, we will use the Z'_n curve, or cumulative GC profile, hereafter, where

$$Z'_n = Z_n - Z = Z_n - k \times n \quad \text{(Equation 4)}$$

Two basic characteristics of the Z' curve are described as follows. (i) If a region in the Z' curve looks like a straight line, the GC content would stay nearly constant within this region. (ii) An up jump (a drop) in the curve means a decrease (or increase) in GC content (Zhang and Zhang, 2004b).

GIs usually have a fairly homogeneous GC content, and this fact could be reflected by the corresponding Z' curve being a nearly straight line (Guo et al., 2003; Charkowski, 2004; Greub et al., 2004; Chen, 2006; Do and Miyano, 2008; Zhang and Zhang, 2004a,b, 2005, 2008; Guo and Wei, 2012; Wei and Guo, 2011). An index called *h* (Zhang and Zhang, 2004b), which quantitatively describes the homogeneity of the GC content of a GI, is defined by the following equation,

$$h = \frac{d_{gi}}{d_c} = \frac{\sqrt{\sum_{n=1}^{M}(z'_n)^2 / M}}{\sqrt{\sum_{n=1}^{N}(z'_n)^2 / N}} \quad \text{(Equation 5)}$$

In this equation, M and N are the lengths of the GI and chromosome, respectively. Symbol d denotes the deviation of the GC content from a constant for a whole genome or a GI. The so-defined h index measures the relative magnitude of the GC content variations in a GI compared with that of the whole genome. If *h* is much less than 1, the variations in GC content of GIs may be considered small.

RESULTS

G+C deviations of the 104 known GIs

As is well accepted, the G+C content of GIs, particularly those inserted recently, is different from that of their host genome (Hacker and Kaper, 2000). It would be important to investigate how far the G+C content of the 104 known GIs deviates from their hosts. Accordingly, the G+C deviations between the 104 GIs and their recipient genomes were calculated and shown as the histogram in Figure 1. As can be seen, only 20 GIs have an absolute G+C deviation less than 2%, and this means that 80.8% of GIs have relatively larger G+C deviations (>2%). The minimum value of absolute G+C deviations among the 104 GIs was only 0.3%. That is to say, that particular GI (SSGI-3 in the host *Shigella sonnei* 53G) and its host genome have almost the same G+C content. In contrast, the largest absolute value of G+C deviation was 14.2%. Next, the 104 GIs were divided into two classes, one was A+T-richer and the other was G+C-richer. The mean absolute value of G+C deviations for the 16 G+C-richer islands was 5.0% , whereas the mean absolute value was 4.0% for the 88 A+T-richer islands. Therefore, on average, the absolute value of G+C deviations of A+T-richer islands is quite similar to that of G+C-richer islands.

Typical length of GIs

For the 104 known GIs, a histogram of the length distribution is shown in Figure 2.

As can be seen, 33 GIs were longer than 20 kb. Only 11 GIs were shorter than 10 kb. Comparatively, only one GI had a length greater than 200 kb. In summary, 12 GIs fell outside the range of typical length of 10 to 200 kb. If the range between 0 and 200 kb was divided into 10 parts, the number of GIs falling in each part decreased gradually, which can be seen in Figure 2. Therefore, there were more short GIs than long GIs based on the characteristics of the 104 GIs. In summary, the mean length of GIs was 46.3 kb (± 42.6 SD). The largest GI had a size of 255 kb and the shortest GI was only 2.7 kb.

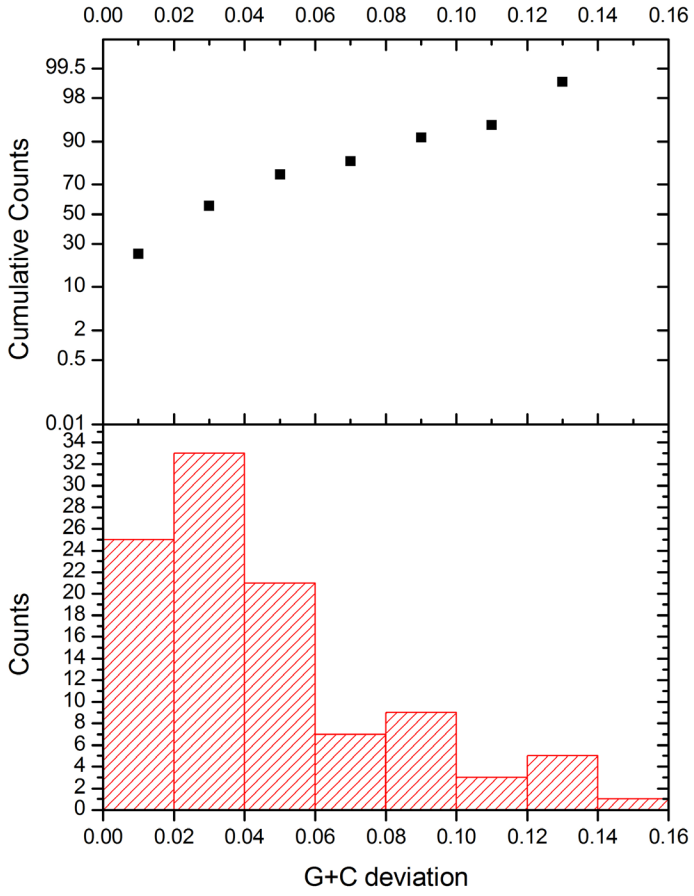

Figure 1. Histogram and cumulative probability of G+C deviation distribution among 104 known GIs.

Mobility genes and tRNA genes

For the 88 GIs whose hosts had been sequenced and annotated, we investigated the presence of mobility genes and tRNA genes. Details of this analysis are listed in Table 1. In this table, if one GI had flanking tRNA genes, it was marked with "+" and otherwise not marked. However, we marked detailed numbers for mobility genes because one GI contained many mobility genes. Two types of mobility genes are considered and they are integrases and

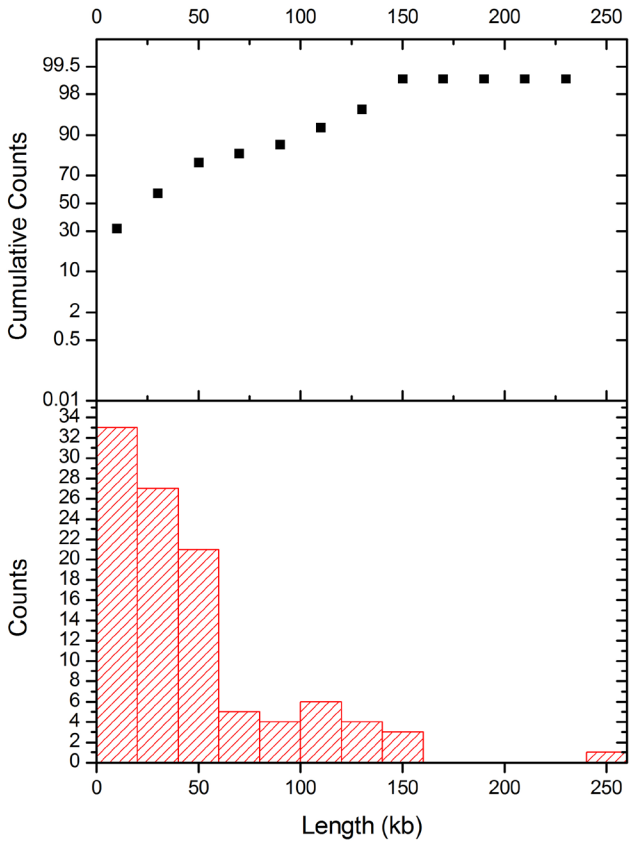

Figure 2. Histogram and cumulative probability of length distribution among 104 known GIs.

transposases. In each GI, the total number of the two kinds of genes was determined and it was divided by the number of all genes in that GI. The ratio obtained was used to denote the probability of mobility genes existing in one specific GI. In total, 46 (52.2%) GIs had a flanking tRNA as their integration sites. Comparatively, 57 GIs (64.7%) contained mobility genes and the mean ratio of the mobility gene number to the total gene number for them was 12%. As an extreme case, even 55% of genes in the GI SSGI-7 had the mobility function.

h index of GI homogeneity

The homogeneity of G+C content of a GI is one feature only mentioned and employed by Zhang and colleagues (Chen, 2006; Zhang and Zhang, 2004a,b, 2005, 2008; Wei and Guo, 2011; Guo and Wei, 2012). According to them, GIs usually have fairly constant G+C content, and hence, they show almost straight lines in the cumulative GC profile. Zhang and Zhang (2004b) proposed one index to measure the homogeneity as in Equation (5). Here, we calculated the *h* index for 88 GIs with annotation information, and the distribution histogram

is shown in Figure 3. As can be seen, 56 GIs had an *h* index less than 0.05, whereas 75 GIs showed values less than 0.1. Zhang and Zhang (2004b) suggested 0.05 as the threshold for being a genuine GI. Chen (2006) changed this threshold to 0.1, and Guo and Wei (2012) then followed her revision. Based on the present analysis, the value of 0.1 may be one more reliable choice because only 63.6% of known GIs had an *h* index less than 0.05. However, if a lower rate of false positives is desired and if completeness of the prediction is not critical, the threshold of 0.05 may be preferred.

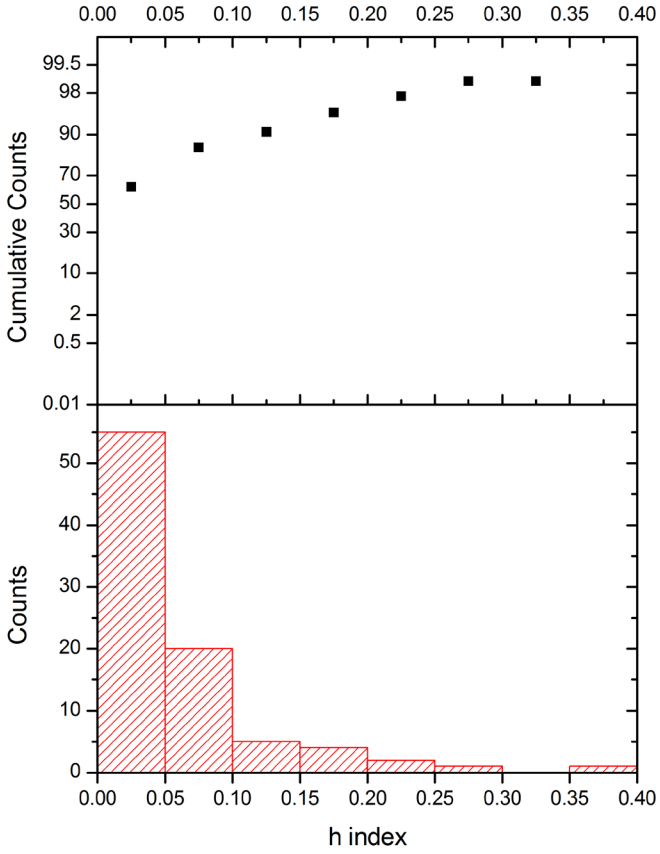

Figure 3. Histogram and cumulative probability of *h* index distribution among 88 known GIs with annotation information.

G+C deviations of GIs and their sizes

To our knowledge, the relationship between the G+C deviation of GIs and their size has not been investigated in any previous study. We performed the following analysis to look at this problem. First, G+C deviation and size of the 104 known GIs were calculated. A scatter plot was then drawn, as shown in Figure 4. In this figure, values of the horizontal axis denote the sizes of GIs, whereas the vertical axis is marked by the absolute values of G+C deviations.

As can be seen, a good right triangle exists in the plot. That is to say, the mean size of GI increases while the range or variance of G+C deviation decreases with length. Indeed, the absolute G+C deviations of 9 GIs longer than 100,000 bp are all less than 5%. Therefore, larger GIs tend to have a smaller G+C deviation according to the absolute values. This is a novel but reasonable result given that larger GIs should have greater restrictions on their G+C content, to maintain the stable G+C content of the recipient genome.

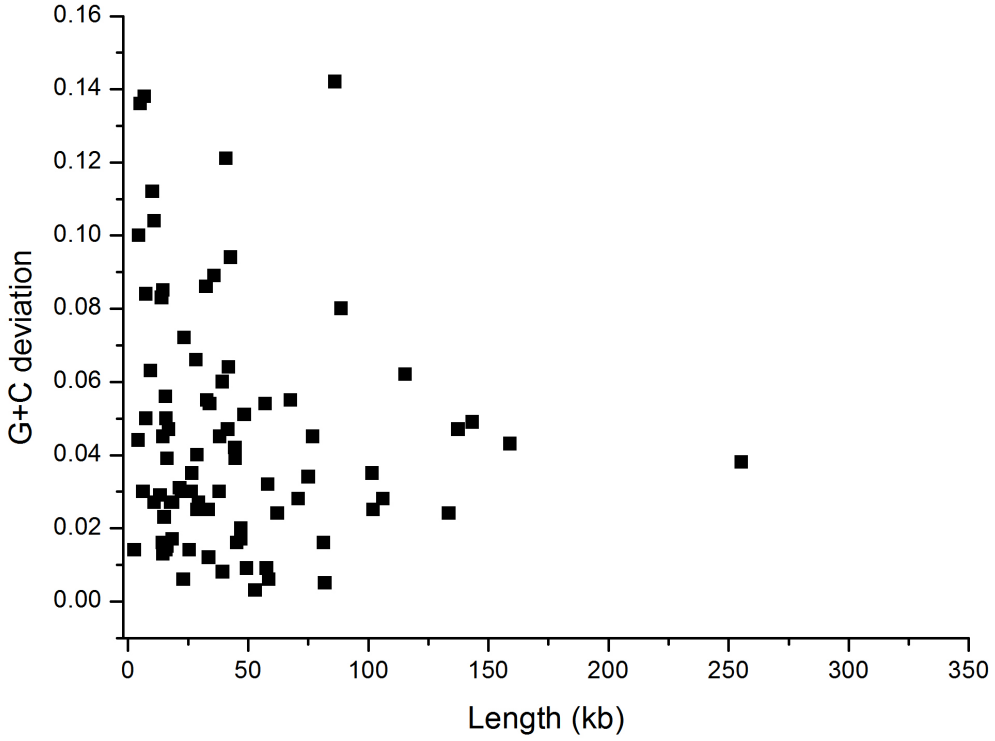

Figure 4. Plot of G+C deviation versus length among 104 known GIs.

DISCUSSION

GIs tend to be A+T richer

Horizontally transferred genes have been observed to be A+T richer in pathogens with medium and low G+C content (Jain et al., 2003). Daubin and Perriere (2003) suggested that either the donor species is always A+T richer than the recipient species or that there is a bias toward the internalization of A+T-rich alien DNA in the genome. As for GIs, there are few reports on the G+C deviation of GIs with respect to the host genome except that performed by Vernikos and Parkhill (2008). However, the authors paid attention to only the variation of G+C deviation values and not the direction of G+C deviation. Here, we addressed this problem based on the analysis on the 104 known GIs. As shown in Figure 4, 88 GIs were A+T richer

than their host genomes and only 16 were G+C richer. Furthermore, the hosts of the 104 GIs could be divided into three classes, high G+C (> 60%), medium G+C (40-60%) and low G+C (<40%), respectively. Consequently, 33 of the 40 GIs in the low G+C hosts were A+T richer. Among the 52 GIs detected in the hosts with medium G+C content, 43 were A+T richer. For the high G+C hosts, all of the 12 GIs were A+T richer.

In conclusion, most (88/104 = 84.6%) of the known GIs analyzed in this work were A+T richer than their host's genome. The AT-richness of GIs did not seem to depend on the G+C content of their recipients. This suggests that horizontally transferred single genes and GIs have similar evolutionary mechanisms.

Why do larger GIs tend to have a smaller G+C deviation?

One of the interesting results of this work is that larger GIs tend to have a relatively smaller G+C deviation. This is due to two possible alternative reasons. Either the donor genome of larger GIs is subject to more severe limitation, or the larger GIs evolve faster than shorter GIs during the post-insertion period. That is to say, only the donor having similar G+C as the recipient is able to donate larger GIs. Alternatively, during the post-insertion period, larger GIs exert stronger pressure, so they evolve faster to have similar G+C as the recipient. No matter how, larger GIs should be subject to more restrictions on their G+C content to maintain the unchangeable G+C content of the recipient genome. If the latter explanation is right, it means some larger GIs could have a larger G+C deviation at the time of introgression. After integration, the G+C content of larger GIs should evolve faster towards the status of the hosts.

Simultaneous presence of multiple conserved features

We analyzed the typical value or presence/absence of each single feature, and their simultaneous presence is also interesting. We take the range between 10 to 200 kb as the typical length of GI. If a GI has its length in this range, it will be regarded as meeting this feature. For the feature of G+C deviation, those having the absolute value of more than 2% are thought to meet it. The value of 0.1 is taken as the threshold of the *h* homogeneity index. Having an *h* index less than 0.1 is taken as the standard for a GI meeting this feature. Taking flanking tRNA gene and embedded mobility gene into account, there are a total of five features. Among the 88 GIs with annotation information, 19 (19/88, 21.6%) GIs exhibited all five features, 34 (38.5%) GIs showed four features, 24 (27.2%) GIs showed three features and 10 (11.4%) GIs displayed two features. Finally, only one GI exhibited only one feature, and the outlier was GI vSeγ in the genome of *Staphylococcus epidermidis* ATCC 12228. The only effective feature for vSeγ was the *h* index, which was only 0.003, the smallest value among all the 88 GIs. There were no GIs that did not exhibit any feature. On the basis of the analysis, we can conclude that all of the known GIs show the conserved features well. Therefore, the five conserved features could be taken as reliable evidence for predicting candidate GIs. Typical values obtained for them in this work would help to optimize parameters in devising or improving GI predicting methods.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (#

#31071109), the Program for New Century Excellent Talents in University (N#CET-11-0059) and the Special Fund of China Postdoctoral Science Foundation (#201104687 and #2013M540705). We are grateful to the reviewers for their valuable comments, which led to the improvement of this paper.

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.

Chen LL (2006). Identification of genomic islands in six plant pathogens. *Gene* 374: 134-141.

Daubin V and Perriere G (2003). G+C3 structuring along the genome: a common feature in prokaryotes. *Mol. Biol. Evol.* 20: 471-483.

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.

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

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.

Guo FB and Wei W (2012). Prediction of genomic islands in three bacterial pathogens of pneumonia. *Int. J. Mol. Sci.* 13: 3134-3144.

Guo FB, Ou HY and Zhang CT (2003). ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.* 31: 1780-1789.

Guo FB, Wei W, Wang XL, Lin H, et al. (2012). Co-evolution of genomic islands and their bacterial hosts revealed through phylogenetic analyses of 17 groups of homologous genomic islands. *Genet. Mol. Res.* 11: 3735-3743.

Hacker J and Kaper JB (2000). Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* 54: 641-679.

Hacker J and Carniel E (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.* 2: 376-381.

Hentschel U, Steinert M and Hacker J (2000). Common molecular mechanisms of symbiosis and pathogenesis. *Trends Microbiol.* 8: 226-231.

Jain R, Rivera MC, Moore JE and Lake JA (2003). Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.* 20: 1598-1602.

Khrustalev VV and Barkovsky EV (2011). "Protoisochores" in certain archaeal species are formed by replication-associated mutational pressure. *Biochimie* 93: 160-167.

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

Soares SC, Abreu VA, Ramos RT, Cerdeira L, et al. (2012). PIPS: pathogenicity island prediction software. *PLoS One* 7: e30848.

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

Wei W and Guo FB (2011). Prediction of genomic islands in seven human pathogens using the Z-Island method. *Genet. Mol. Res.* 10: 2307-2315.

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 CT and Zhang R (2004a). Genomic islands in *Rhodopseudomonas palustris*. *Nat. Biotechnol.* 22: 1078-1079.

Zhang R and Zhang CT (2004b). 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.

Zhang R and Zhang CT (2008). Accurate localization of the integration sites of two genomic islands at single-nucleotide resolution in the genome of *Bacillus cereus* ATCC 10987. *Comp Funct. Genomics* 451930.

Zhang CT, Wang J and Zhang R (2001). A novel method to calculate the G+C content of genomic DNA sequences. *J. Biomol. Struct. Dyn.* 19: 333-341.