OXFORD

Briefings in Bioinformatics, 18(3), 2017, 357–366

doi: 10.1093/bib/bbw019
Advance Access Publication Date: 18 March 2016
Paper

Zisland Explorer: detect genomic islands by combining homogeneity and heterogeneity properties

Wen Wei,* Feng Gao,* Meng-Ze Du, Hong-Li Hua, Ju Wang and Feng-Biao Guo

Corresponding author: Feng-Biao Guo, Key Laboratory for NeuroInformation of the Ministry of Education and Center for Information in BioMedicine, University of Electronic Science and Technology of China, Chengdu 610054, China. E-mail: fbguo@uestc.edu.cn
*These authors contributed equally to this work.

Abstract

Genomic islands are genomic fragments of alien origin in bacterial and archaeal genomes, usually involved in symbiosis or pathogenesis. In this work, we described Zisland Explorer, a novel tool to predict genomic islands based on the segmental cumulative GC profile. Zisland Explorer was designed with a novel strategy, as well as a combination of the homogeneity and heterogeneity of genomic sequences. While the sequence homogeneity reflects the composition consistence within each island, the heterogeneity measures the composition bias between an island and the core genome. The performance of Zisland Explorer was evaluated on the data sets of 11 different organisms. Our results suggested that the true-positive rate (TPR) of Zisland Explorer was at least 10.3% higher than that of four other widely used tools. On the other hand, the new tool did not lose overall accuracy with the improvement in the TPR and showed better equilibrium among various evaluation indexes. Also, Zisland Explorer showed better accuracy in the prediction of experimental island data. Overall, the tool provides an alternative solution over other tools, which expands the field of island prediction and offers a supplement to increase the performance of the distinct predicting strategy. We have provided a web service as well as a graphical user interface and open-source code across multiple platforms for Zisland Explorer, which is available at http://cefg.uestc.edu.cn/Zisland_Explorer/ or http://tubic.tju.edu.cn/Zisland_Explorer/.

Key words: genomic islands; prediction; homogeneity; cumulative GC profile

Introduction

Horizontal gene transfer among genomes undoubtedly plays an important role in expanding genetic material and driving incipient speciation [1, 2]. A cluster of these alien genes constitutes the genomic islands in bacterial or archaeal genomes, which can be discovered in a certain genome but be absent from closely related genomes. A genomic island can be involved in various functions, including those related to symbiosis or

Wen Wei is a lecturer at School of Life Sciences, Chongqing University. His research focuses on computational models for genetic and genomic field.
Feng Gao is a professor at Department of Physics, Tianjin University. His researches are performed within the fields of bioinformatics and microbial genomics.
Meng-Ze Du is a doctoral student at Bioinformatics Center, University of Electronic Science and Technology of China. Her research interest is computational genomics and proteomics.
Hong-Li Hua is a masters student at Bioinformatics Center, University of Electronic Science and Technology of China. Her current research interest focuses on computational genomics.
Ju Wang is a professor in School of Biomedical Engineering, Tianjin Medical University. His research interest is to develop theoretical methods and apply computational tools to address some biological or medical issues.
Feng-Biao Guo is a full professor and director of Bioinformatics Center, University of Electronic Science and Technology of China. His current research interests include developing computational models and tools for gene (or genomic region) prediction in bacterial genomes, investigating nucleotide composition/structural variation or similarity across bacterial genomes, revealing potential factors associated with the variation and spectrum of substitutional rates and spontaneous mutational rates in microbes, analyzing the basic gene components of prokaryotic genomes and applying it into the field of synthetic biology.
Submitted: 24 November 2015; **Received (in revised form):** 19 January 2016

Table 1. Comparison of core algorithm of each method

Core feature Method	Sequence heterogeneity	Structural heterogeneity	Homogeneity	Comparative genome
Zisland Explorer	Codon usage bias Amino acid bias		Similarity within a island	
Cumulative GC profile			Similarity within a island	
Islander		tRNA site Mobility gene		
IslandPath-DIMOB	Di-nucleotide bias	Mobility gene		
SIGI-HMM	Codon usage bias			
IslandPick				Comparison with closely related genomes

pathogenesis, and may help an organism's adaptation. For example, some genomic islands encode genes associated with the improvement of organism survival under adverse conditions that can cooperate to confer the organism with novel phenotypes such as the capacity to cause disease to infect the host cell [1, 3].

Since the discovery of the pathogenic islands coding for hemolysin and fimbrial determinants in uropathogenic *Escherichia coli* strains [4], intensive studies have been dedicated to identifying genomic islands with conserved features such as compositional bias, mobility elements and transfer RNA (tRNA) hotspots [5–9]. IslandPick was developed to identify unique island regions in a genome by comparing it against closely related genomes [10]. This method automatically selects the comparative genomes for each query genome according to the phylogenetic innate characters via an evolutionary distance function. IslandPick identifies regions unique to the query genome as genomic islands; thus, the outcome of the method inevitably depends on the selection of reference genomes.

Among the typical features of genomic islands, the compositional bias has been demonstrated to be the most important feature [7]. In fact, assessing the nucleotide composition difference (heterogeneity) between the foreign gene fragments and native genome is a more usual way of detecting island transfer events [6]. IslandPath-DIMOB [11] and SIGI-HMM [12] are representative tools for sequence composition approaches. The former identifies a fragment to be a genomic island if it contains eight or more consecutive open reading frames with dinucleotide bias and one or more mobile genes. The latter is based on the analysis of codon usage of each gene by comparing it against a carefully selected set of codon tables representing microbial donors or highly expressed genes. Using the Viterbi algorithm, SIGI-HMM classifies all genes into the most probable codon usage states, native or nonnative. Finally, genes with the states of nonnative are considered as island regions. Islander uses structural heterogeneity-searching strategy and splits DNA fragments by tRNA/transfer-messenger RNA signatures. Then, it restrictedly finds island candidates through several filters, including tests of an integrase gene, correct fragment/transfer DNA orientation and sequence length [13].

A windowless method for calculating the cumulative GC profile (*Z'*) has been proposed to describe the GC content variation in a genome [14, 15]. Intuitively, for a genome containing genomic islands, leaps should show up in the cumulative GC profile because the GC content is homogenous within the island [16]. Thus, the homogeneity of a DNA fragment can be quantified to describe the leap in cumulative GC profiles and used to predict genomic islands. So far, this approach has been successfully used to identify genomic islands in many organisms [17–22]. However, to identify genomic islands with the cumulative GC profile, the plot depicting the GC variation of the genome sequence has to be analyzed manually, which is not only fairly inaccurate in many cases, but also difficult to apply in a high-throughput way.

By integrating different features (Table 1), the five methods all have high prediction precision. However, because genomic islands usually share only limited conserved features, these methods may fail to achieve ideal sensitivity in some cases. In this work, we design a *de novo* strategy to predict genomic islands according to the cumulative GC profile and the GC-Profile, a segmentation method proposed by us and collaborators. The tool, so-called Zisland Explorer, combines the homogeneity and the heterogeneity of a sequence for the first time. The new method can automatically split DNA fragments according to the sequence composition homogeneity and advance the predicting procedure for classifying genomic islands from these split fragments. Compared with the widely used tools, Zisland Explorer is able to detect genomic islands by relying on the genomic sequence only and without time-consuming homology analysis.

Methods

Zisland explorer algorithm

Zisland Explorer adopts a multistep strategy (Figure 1) and the details can be described as follows.

Step 1: Split genomic sequence into segments using GC-Profile

The original GC-Profile is a general tool for splitting DNA sequences according to the Jensen–Shannon divergence between the left and right subsequences in the occurrence frequencies of A, C, G and T [23, 24]. The order index can be changed into different forms according to the segmentation purposes. Here, we aim to segment sequences according to GC homogeneity shown in the cumulative GC profile and hence change the order index S as:

$$S(P) = (A + T)^2 + (G + C)^2. \quad (1)$$

Let $P_L = (A_L + T_L, C_L + G_L)$ and $P_R = (A_R + T_R, C_R + G_R)$ be the frequency vectors in the left and right subsequences, respectively. The divergence between two subsequences was defined as:

$$\Delta S(P_L, P_R) = \omega_1 S(P_L) + \omega_2 S(P_R) - S(\omega_1 P_L, \omega_2 P_R). \quad (2)$$

The coefficients ω_1 and ω_2 were the length weights of the two subsequences. We here use the halting parameter 50 and the minimum length 1000 bp as the divergence standard. When the count of the DNA fragments to be split is <15, the procedure of GC-Profile segmentation is repeated with a halting parameter of 25.

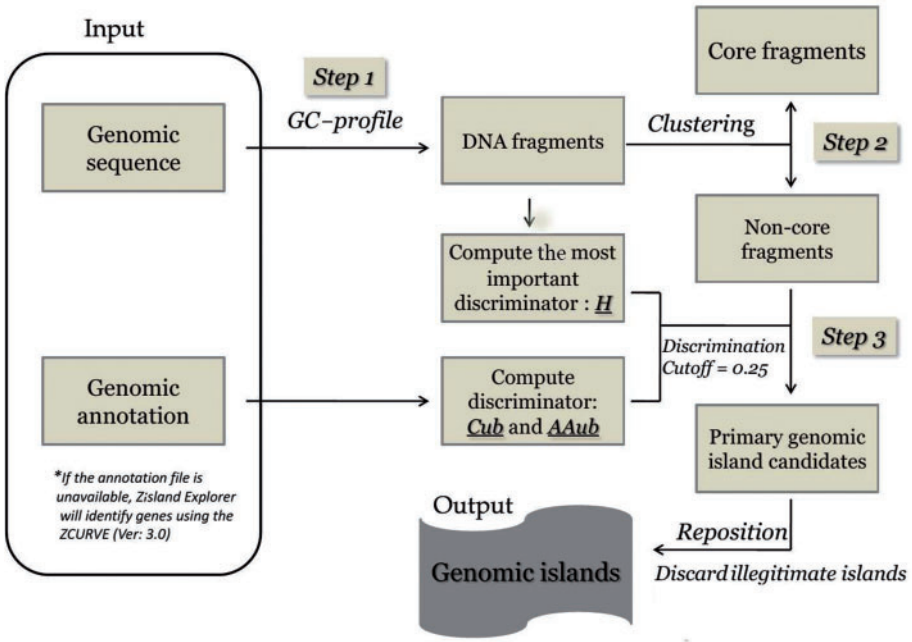

Figure 1. Zisland Explorer workflow.

Step 2: Exclude core segments

After the above step, DNA fragments are identified. Next, we need to exclude the core fragments before predicting genomic islands.

1. Cluster potential core set 1 using GC heterogeneity between them and the whole genome

Because of the GC divergence (heterogeneity) between the genomic islands and the core genome, occurrence frequency vector $[(G+C)^2, (A+T)^2]$ is used in the next clustering step. The initial core vector is set as the average vector of the whole genome, and then we find the most similar fragment using the Euclidean distance. We reset the core vector as the average of the initial core vector and the one most similar to it. The iteration stops when the length of the core set 1 is over a cutoff of the whole genome. According to the results of greedy searching in L-positive and L-negative data sets from published literature (see 'Algorithm Validation'), we identify this cutoff as 80%.

2. Detect potential core set 2 using GC homogeneity within each segment

Compared with core fragments, genomic islands usually have a more similar GC content along the sequence. We used an index *H* to describe the GC homogeneity (similarity) of a fragment.

Every DNA sequence can be represented uniquely in a three-dimensional space (*X*, *Y*, *Z*) using a *Z*-curve [25], with *X*, *Y* and *Z* measuring the cumulative distributions in purine/pyrimidine, amino/keto and weak/strong hydrogen bonds along the DNA sequence, respectively:

$$\begin{aligned} X_n &= (A_n + G_n) - (C_n + T_n) \\ Y_n &= (A_n + C_n) - (G_n + T_n) \\ Z_n &= (A_n + T_n) - (G_n + C_n), \\ n &= 1, 2, 3, \ldots N \end{aligned} \quad (3)$$

To amplify the deviations in the cumulative GC variations, a windowless technique called the cumulative GC profile (*Z'*) has been developed. The element *Z* in Equation (3) is fitted into a straight line with a slope *k* by using the least-squares method, and the cumulative GC profile of the sequence analyzed can be described as the difference between the real and fitted GC variations [Equation (4)]. Because the GC content of a certain genomic island is homogeneous, it can be represented as an approximate straight line in the cumulative GC profiles:

$$Z' = Z - kn. \quad (4)$$

GC homogeneity (*H*) is defined by Equation (5), in which *M* and *N* are the lengths of the fragment and the chromosome, respectively, and symbol *d* denotes the deviation of *Z'* from a constant for a whole genome or a fragment. A larger *H*-value indicates higher GC homogeneity:

$$H = 1 - \frac{d_f}{d_g} = 1 - \frac{\sqrt{\sum_{n=1}^{M} (Z')^2 / M}}{\sqrt{\sum_{n=1}^{N} (Z')^2 / N}}. \quad (5)$$

Note that, the here used *H* index is a little different with the original *h* index, in that their sum equals to 1. We calculate the GC homogeneity of DNA fragments and then sort them. A fragment with a minimum *H*-value (minimum GC homogeneity) is classified into core set 2, until the length of core set 2 is over 80% of the whole genome.

3. Identify all core fragments

These core fragments have both less GC heterogeneity and less GC homogeneity. Thus, to decrease the false-positive rate of core segments, we define core fragments as a set which covers both potential core sets 1 and 2.

Step 3: Identify genomic island candidates

1. Codon and amino-acid heterogeneity

The codon/amino-acid usage bias (*Cub*/*AAub*) of each gene is estimated by a variant cosine value between the average and individual usage vector [Equation (6)]. The occurrence frequencies of all sense codons (amino acids) in a gene could be deemed as

a usage vector. We denote the codon (amino acid) usage vector for the ith gene in the investigated genome as $\mathbf{C}_i$ ($\mathbf{AA}_i$). The average codon (amino acid) usage vector determined for all genes is denoted by $\mathbf{C}$ ($\mathbf{AA}$):

$$Xub = 1 - \frac{\mathbf{X_i} \bullet \mathbf{X}}{|\mathbf{X_i}| \times |\mathbf{X}|}. \quad X = C, AA \tag{6}$$

The *Cub*/*AAub* of each fragment is estimated by the average *Cub*/*AAub* of the embedded genes.

2. Scale discriminators by minimum vector

Before determining primary candidates, we need to scale three discriminators (*H*, *Cub*, *AAub*) for each fragment:

$$\mathrm{Scale}(X) = \frac{X - \mathrm{Min}(X)}{X + \mathrm{Min}(X)}. \quad X = H, Cub, AAub \tag{7}$$

The Min(*X*) is defined by the average of *m* minimum *X*-values. We assign the number *m* to 2 when a 10th of the fragment number does not exceed 2 and to 4 otherwise.

3. Genomic island prediction

First, the primary genomic island candidate is defined at a cutoff by the following equation. Here, ω is a weight index of three factors, and we define the homogeneity as the most important factor in the discrimination ($\omega_1 = 1$, $\omega_2 < 1$ and $\omega_3 < 1$):

$$\mathrm{Score} = \sqrt{\omega_1 \mathrm{Scale}(H)^2 + \omega_2 \mathrm{Scale}(Cub)^2 + \omega_3 \mathrm{Scale}(AAub)^2}. \tag{8}$$

Next, neighboring primary candidates join into genomic islands. The borders of the genomic islands are relocated at the nearest genes. Ribosomal protein clusters are known to be highly optimized, usually with a sequence composition different from the rest of the genome. To avoid these genomic fragments might be as false-positive predictions, the candidates encoding five ribosomal proteins in succession are abandoned according to the annotation file. Finally, we only keep the islands with a length of between 2 and 400 kb. Using the L-positive and L-negative data sets (see 'Algorithm Validation') from published literature, we confirm the best performance in sensitivity and specificity when assigning ω_2 as 0.6 and ω_3 as 0.5, which defines a discriminant score cutoff at 0.25.

Compared with the cumulative GC profile and the GC-Profile, this work is novel in sense that it proposed a systematic strategy to identify the genomic islands integrating these two methods and appending the step of filtering core regions. Furthermore, it is the first report to consider both sequence homogeneity and heterogeneity in the issue of island identification. Because of the above points, Zisland Explorer could automatically identify genomic islands from genomic sequence with high accuracy.

Algorithm validation

Data set from the literature (L-data set)

To assess the performance of Zisland Explorer, the genomic islands in 11 genomes identified using a genome-wide comparative approach were collected from published literature (Supplementary Table S1). These distinct data are from seven orders, namely *Burkholderiales* (*Burkholderia cenocepacia* J2315: NC_011000, NC_011001, NC_011002; *Bordetella petrii* DSM 12804: NC_010170), *Corynebacteriales* (*Corynebacterium diphtheriae* NCTC 13129: NC_002935), *Enterobacteriales* (*Cronobacter sakazakii* ATCC BAA-894: NC_009778; *Escherichia coli* CFT073: NC_004431; *Proteus mirabilis* HI4320: NC_010554; *Salmonella typhi* CT18: NC_003198), *Micrococcales* (*Clavibacter michiganensis* NCPPB 382: NC_009480), *Rhizobiales* (*Bartonella tribocorum* CIP 105476: NC_010161), *Lactobacillales* (*Streptococcus equi* 4047: NC_012471) and *Vibrionales* (*Vibrio cholerae* N16961: NC_002505, NC_002506). If genes were in a genomic island, they were labeled as the L-positive data set, otherwise they were labeled as the L-negative data set.

Data set from a comparative analysis (C-data set)

There is no 'gold standard' when choosing referential genomes, so genomic islands from different studies could have differential standards within evolutionary time. We thus reidentified the positive and negative data sets of those 11 genomes ourselves within a similar evolutionary scale using a comparative analysis [7, 10]. The Composition Vector (CV) method is used to calculate evolutionary relatedness for building referential genomes [26]. We remove closely related genomes (CV distance < 0.15) and distant queries (CV distance > 0.45), and then randomly choose at least two and at most five referential genomes, ensuring that references are not closely related to each other (CV distance > 0.1). The referential genomes and CV distances for the 11 genomes are listed in Supplementary Table S2.

We use all-against-all BLASTp to search homologs (E-value $< 1e^{-30}$, coverage > 0.8 and identity > 0.3). Three proteins in succession that are conservative in all referential genomes are considered to be a C-negative data set, whereas three successive proteins absent in all referential genomes are considered to be a C-positive data set (Supplementary Table S3; http://cefg.uestc.edu.cn/Zisland_Explorer/download.html).

Performance

The following quotas are measured to assess the performance of the Zisland Explorer tool:

$$\begin{aligned} TPR &= \frac{TP}{TP+FN} \\ TNR &= \frac{TN}{TN+FP} \\ OACC &= \frac{TP+TN}{TP+TN+FN+TP} \\ ACC &= \frac{TPR+TNR}{2} \\ F1 &= \frac{2TP}{2TP+FP+FN} \\ MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{aligned} \tag{9}$$

TP, FN, FP and TN denote the true positives, false negatives, false positives and true negatives, respectively. Sensitivity (or true-positive rate, TPR) measures the correctly identified proportion in positives, while specificity (or true-negative rate, TNR) represents this proportion in negatives. Overall accuracy (OACC) is defined as the percentage of samples correctly found. Accuracy (ACC) is adopted to represent the arithmetic balance between TPR and TNR, whereas the F1 score measures the harmonic balance between TPR and precision. In addition, the Matthews correlation coefficient (MCC) is used to describe a correlation coefficient between the truth and predicted genomic islands.

Results

Validating performance

To evaluate the performance of Zisland Explorer in the prediction of genomic islands, we collected genomic islands from

Table 2. Performance of Zisland Explorer

	L-data set			C-data set		
	TPR	TNR	OACC	TPR	TNR	OACC
B. tribocorum	0.338	0.988	0.754	0.454	0.996	0.818
B. petrii	0.291	0.980	0.841	0.313	0.995	0.799
B. cenocepacia Chr. 1	0.628	0.982	0.948	0.553	0.990	0.940
B. cenocepacia Chr. 2	0.930	0.962	0.960	0.720	0.987	0.934
B. cenocepacia Chr. 3	1.000	0.917	0.923	0.713	1.000	0.730
C. michiganensis	0.786	0.962	0.956	0.297	0.961	0.931
C. diphtheriae	0.305	0.983	0.913	0.214	1.000	0.759
C. sakazakii	0.523	0.969	0.930	0.379	0.995	0.827
E. coli	0.459	0.965	0.880	0.353	0.989	0.778
P. mirabilis	0.521	0.955	0.905	0.307	0.995	0.788
S. typhi	0.614	0.954	0.926	0.454	0.987	0.894
S. equi	0.374	0.984	0.897	0.356	1.000	0.750
V. cholerae Chr. 1	0.472	0.998	0.972	0.226	0.997	0.933
V. cholerae Chr. 2	1.000	0.994	0.995	0.831	1.000	0.911
Average	0.589	0.971	0.914	0.441	0.992	0.842

published literature (L-data set) and also identified putative genomic islands by comparative analysis (C-data set) in 11 bacterial genomes. These genomes were phylogenetically varied enough to assess the power of the prediction fairly. We downloaded files of genomic sequences (*.fna) and annotations (*.ptt) from the RefSeq (ftp://ftp.ncbi.nlm.nih.gov/genomes/).

Self-validation using the L-data set

In total, 97 genomic islands in these genomes were predicted by Zisland Explorer (Supplementary Table S4). We found that most of the putative islands (64.9%) covered the literature-based data set. The performances are listed in Table 2. As mentioned above (see 'Algorithm Validation'), we identified the weight indexes and discriminant score cutoffs using the L-data set. An average OACC of 91.4% was obtained and 8 of 14 chromosomes had higher accuracy than average. The averages of ACC, F1 score and MCC were 0.780, 0.597 and 0.582, respectively.

We also compared four widely used tools, namely IslandPick, Islander, SIGI-HMM and IslandPath-DIMOB, with Zisland Explorer (Supplementary Table S5, Figure 2A). The predictions of these tools were obtained from the references of IslandViewer 3 [27] and Islander [13]. Generally, the prediction outcomes of methods based on the empirical features show highly similar distribution in GC bias ($|GC_{island}\text{-}GC_{host}|$), codon usage bias and amino-acid bias (Supplementary Table S6), and also are better than that of the comparative genomics-based method, IslandPick. Furthermore, the average island sizes of the four composition-based methods have similar distribution, whereas the IslandPick and SIGI-HMM find islands with much smaller sizes. In addition, we took one of the commonly identified islands by the five methods as an example to illustrate its typical features of mobile elements (Supplementary Figure S1; http://cefg.uestc.edu.cn/Zisland_Explorer/download.html). For the five methods compared using the L-data set, the averages of the TPR were 9.6%, 23.7%, 39.0%, 39.9% and 58.9% for IslandPick, Islander, SIGI-HMM, IslandPath-DIMOB and Zisland Explorer, respectively; the values of OACC were 87.9%, 89.7%, 89.6%, 89.7% and 91.4% for each of the five methods, respectively. Thus, Zisland Explorer was able to find more true island genes with a little higher OACC. Specifically, Zisland Explorer achieved a TPR of 100% and an OACC of 92.3% when used on *B. cenocepacia* Chr. 3. For comparison, the TPRs were only 0.0%, 0.0%, 25.9% and 38.9% for the same chromosome when IslandPick, Islander, SIGI-HMM and IslandPath-DIMOB, respectively, were used. Occasionally, the other four tools failed to identify any island in places where Zisland Explorer did, showing that Zisland Explorer performed better in analysis of these 11 genomes. In the case where TP = 0 and FP = 0, we defined MCC as 0 because the correlation tended to be random. Our test suggested that Zisland Explorer had the best TPR/TNR balance, TPR/precision

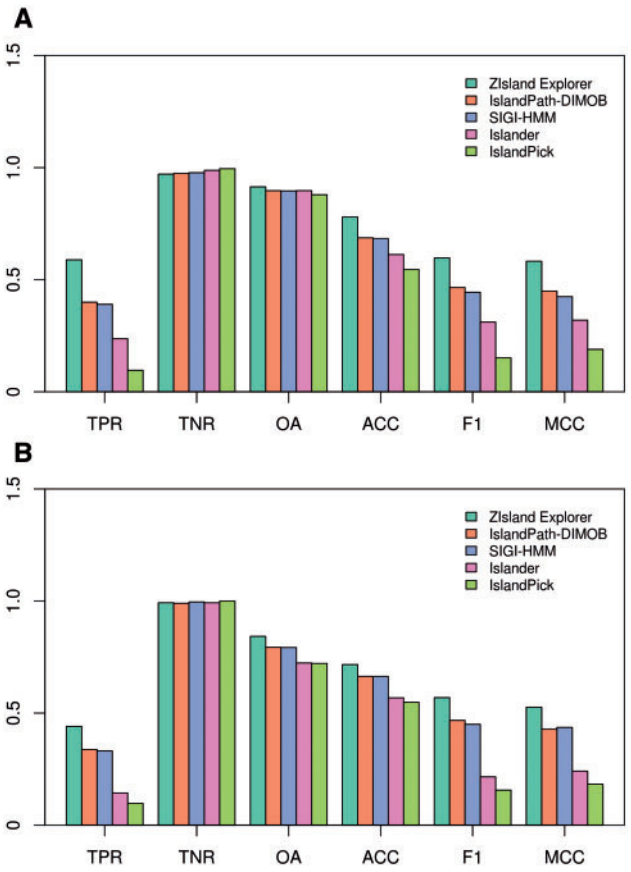

Figure 2. Performance comparison of Zisland Explorer with other tools. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

Table 3. The proportion of ICEs predicted by five tools

ICE	Zisland Explorer	IslandPath-DIMOB	SIGI-HMM	IslandPick	Islander
BPGI2	21.9%	42.3%	48.0%	0.0%	0.0%
BPGI3	0.0%	56.2%	59.7%	0.0%	0.0%
BPGI4	0.0%	56.5%	26.7%	0.0%	0.0%
BPGI7	30.8%	48.2%	42.0%	0.0%	0.0%
PMGI6	100.0%	0.0%	49.3%	59.9%	0.0%
PMGI7	54.2%	0.0%	39.8%	95.6%	100.0%
STGI12 (SPI-7)	90.2%	74.0%	18.8%	22.8%	99.9%
SEGI4	100.0%	93.9%	0.0%	19.7%	0.0%
Average	49.6%	46.4%	35.5%	24.8%	25.0%

balance and MCC. The ACC, F1 score and MCC of Zisland Explorer were 9.3%, 13.2% and 13.3% higher, respectively, than those of the second-best tool, IslandPath-DIMOB.

Validation using the C-data set

The L-data sets collected from different studies may have differential scales within evolutionary time. By following the weight indexes and cutoffs identified above, we re-estimated performance by using the other data set (C-data set). Compared with the L-data set, the negatives of the C-data set were more conservative in species divergence, which showed a higher TNR (99.2%) when using the C-data set. We observed that 77 (79.4%) predicted genomic islands were covered by positive bases from comparative analysis-based data sets (Supplementary Table S4). Compared with other tools, Zisland Explorer showed improvements of at least 10.3% in TPR and 4.9% in OACC (Figure 2B). Our Zisland Explorer data showed similar performance in detecting genomic islands when compared with adapted L- and C- data sets, with Zisland Explorer giving better results than the other tools tested.

Checking accuracy of experimental data

Integrative and conjugative elements (ICEs) are kinds of self-transmissible islands. We confirmed eight experimental ICE data from the ICEberg database [28] among the 11 genomes and calculated the percentage of bases that were exactly predicted by the five tools (Table 3). Using the weight indexes and cutoff (0.25) from the L-data set, although Zisland Explorer failed to maintain good accuracy for every ICE, the overall performance (49.6%) was significantly better than SIGI-HMM, IslandPick and Islander, and comparable with IslandPath-DIMOB. Two Zisland Explorer islands completely and exactly overlapped the ICEs, PMGI6 and SEGI4, in individual bases, respectively. The 133 kb tRNA-PheU-associated island, SPI-7, had been covered by Zisland Explorer islands on 90.2% of the bases, which was lower than Islander (99.9%), but significantly higher than the others. The ICEs, BPGI3 and BPGI4, were missed by Zisland Explorer. However, they could be picked up if we slightly adjusted the cutoff from 0.25 to 0.245. BPGI3 showed deficiency in both homogeneity and amino-acid bias, and BPGI4 displayed a weaker amino-acid bias.

Testing additional islands predicted by Zisland Explorer

In total, 82 Zisland Explorer predictions can be covered by either C-positives or L-positives, and 34 of them completely overlapped to positives. To test additional positive rates, we focused on 15 predictions without positives covered, which could be prime candidates for false predictions. Other common positive features (tRNA, integrase and phage) of islands were investigated according to GenBank annotation (Supplementary Table S4). The function of 'hypothetical protein' had been confirmed by BLASTp against the nonredundant database.

A machine-learning approach revealed that 'phage' and 'integrase' were two important structural features for classifying genomic islands and negatives, except for compositional deviations [7]. We also found that 9 of 15 additional islands encode 'phage' or 'integrase'-associated proteins, and display other empirical features. On the other hand, the six fragments that missed both 'phage' and 'integrase' features were also absent from the tRNA hotspot site and lost by at least three other island predictors, which could be considered as false predictions. We also checked a common non-island feature, ribosomal RNA (rRNA) operons. However, the false predictions could not be attributed to rRNA operons. Thus, we can safely conclude that most of the false positives are novel islands based on the test of prime candidates for false predictions.

Improving prediction by combining Zisland Explorer

Every predictor is designed to predict islands by relying on several conserved features, including tRNA integrations, compositional bias and mobility elements. However, the conserved features vary among genomic islands; for example, some islands do not have a clear codon bias, detectable mobility elements or tRNA sites. Thus, we may miss some real islands by only using a single method. In fact, Zisland Explorer is designed based on principles different from IslandPick, IslandPath-DIMOB, SIGI-HMM and Islander. Therefore, it could serve as a complementary method of these tools, and their prediction outputs are significantly different. Here, we conflate the predicted islands of Zisland Explorer and each tool in the 11 genomes, respectively, (Table 4). With such joint strategy, either the L-data set or the C-data set could sharply increase the TPR, leading to improvements in OACC. When combining Islander with Zisland Explorer, we improve the TPR by 44.9% compared with Islander alone and 9.7% with Zisland Explorer alone for the L-data set. Simultaneously, this combined prediction increases the OACC by at least 1.6% for each single tool. Thus, the joint application of Zisland Explorer and each distinct tool could find more true island genes than any single method, with similar OACC, in the 11 genomes. On the basis of the above analysis, the TPR will be sharply improved by joint application. Therefore, we strongly suggest that researchers should use multiple genomic island finders for obtaining more genuine island genes.

Webserver and application

We also provide an online service for Zisland Explorer (http://cefg.uestc.edu.cn/Zisland_Explorer/). When performing island prediction, the users only need to submit a standard FASTA sequence file of the genome (Figure 3I) and optionally upload an annotation file in GenBank ptt style (Figure 3II). If the annotation file is unavailable, Zisland Explorer will identify genes using ZCURVE (Ver: 3.0) [29], a gene annotation tool based on Z-curve theory and having been validated on hundreds of genomes. The users will obtain a file of predicted islands (Figure 3III) and an image of the cumulative GC profile displaying the predicted islands (Figure 3IV).

Furthermore, we also provide a version of the tool with graphical user interface (GUI) to run locally, which can work on Windows, Linux or Mac OS X systems. The use of the GUI version is same as for the web service (Figure 3). Zisland Explorer is an open-source software, and its source code is available at http://cefg.uestc.edu.cn/Zisland_Explorer/ or http://tubic.tju.edu.cn/Zisland_Explorer/ for free. The code runs dependent on Python, Biopython and Matplotlib.

Table 4. Improvement of each tool by combing with Zisland Explorer

	Increase of each index					
	TPR	TNR	OACC	ACC	F1	MCC
L-data set						
IslandPick	51.59%	–2.60%	3.63%	24.50%	45.86%	40.58%
IslandPath-DIMOB	29.30%	–1.97%	1.61%	13.66%	16.51%	15.63%
SIGI-HMM	28.97%	–1.94%	1.66%	13.52%	18.65%	18.07%
Islander	44.85%	–2.75%	2.67%	21.05%	33.98%	30.86%
C-data set						
IslandPick	36.73%	–0.73%	12.69%	18.00%	43.14%	36.04%
IslandPath-DIMOB	19.71%	–0.62%	6.44%	9.54%	17.12%	14.29%
SIGI-HMM	19.96%	–0.72%	6.69%	9.62%	20.09%	15.43%
Islander	34.84%	–0.75%	12.82%	17.04%	39.07%	30.59%

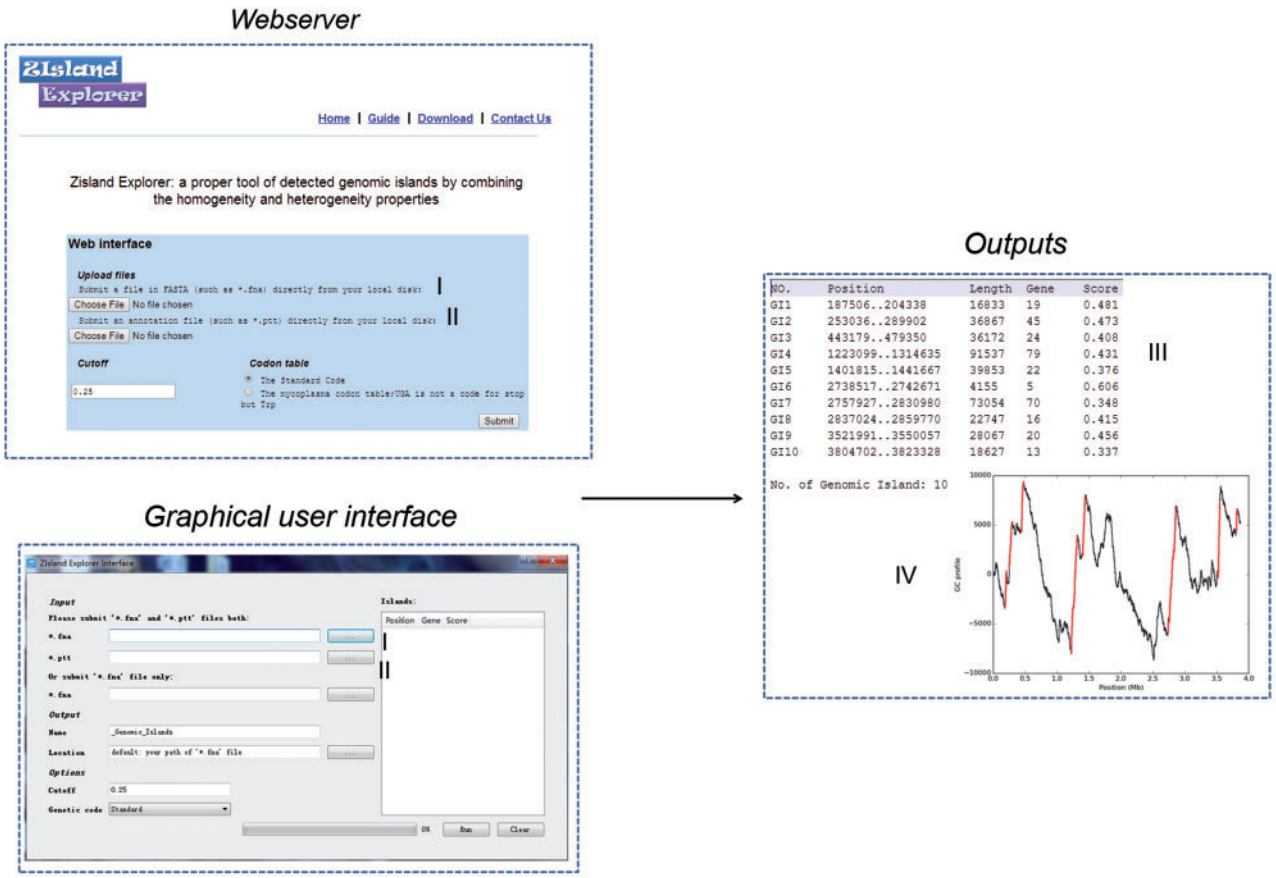

Figure 3. Webserver and GUI.

Zisland Explorer was used to predict all sequenced bacteria and archaea in RefSeq (2763 genomes until May 2015). The predictions are available at http://cefg.uestc.edu.cn/Zisland_Explorer/list.html. We found that over 68% of these genomes have embedded genomic islands. There is no difference between bacteria and archaea in the percentage of embedded islands ($P > 0.5$). Most genomes hold ~0–5% genomic islands (Figure 4).

Discussion

About a decade ago, Zhang *et al.* proposed a systematic method to identify genomic islands [14], which was based on the idea that the GC content of a certain genomic island was homogeneous within the island itself but was different from that of the core genome. Thus, a genomic island is shown as a straight line abruptly 'hopping' up (or down) within an otherwise zig-zag cumulative GC profile. An index of h has been proposed to quantify this homogeneity feature. In our previous work, we confirmed that genuine genomic islands have smaller h values [30]. This method has been applied in a few island-finding studies [19–21].

However, the original method needs to pick 'hops' in cumulative GC profiles with manual intervention. In fact, it is hard to determine the straight 'hops' precisely, especially for genomes

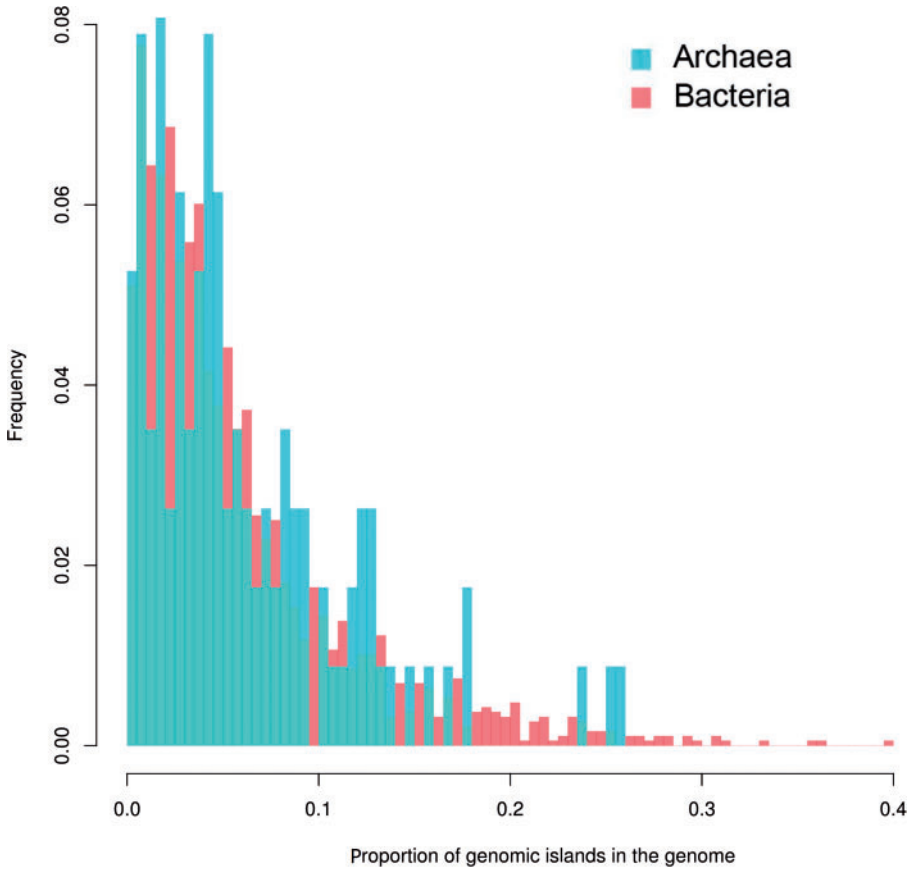

Figure 4. Proportion of islands in each genome. Light red is for bacteria and light blue for archaea. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

with high GC variation. On the other hand, another method, GC-Profile, was developed to split genome according to composition similarity of the four bases. Because this method was not designed specifically for genomic island identification, its performance on this issue had not been evaluated.

In this study, we proposed a systematic strategy to identify the genomic islands by combining these two methods. First, we split the input genome into segments using GC-Profile. Next, we picked out the core genome with the clustering program and the *H* index from the cumulative GC profile. Then the genome fragments were filtered to obtain the island candidates. Finally, we identified the genomic islands based on the combination of amino-acid usage bias, codon usage bias and *H* value. The GC-Profile method in the first step splits a genome based on the GC homogeneity in the same region and the GC divergence between two adjacent regions. The clustering program in the second step uses the composition similarity among the core genome; and the *H* index is calculated based on the observation that genomic islands are more GC homogeneous than the nearby native sequences in the genome. In the last step, both the *Cub* and *AAub* reflect the composition heterogeneity between island region and the core genome. As two kinds of island identifiers, the cumulative GC profile uses the GC homogeneity of islands, and all the other four tools are mainly based on the composition or phylogenetic bias between islands and the host. Here, the Zisland Explorer for the first time combines the two features using a systematic strategy.

In the original cumulative GC profile method, Zhang *et al.* used index *h* to identify genomic islands with a cutoff. When evaluating on the L-data set, we found it works well in some genomes such as *V. cholerae* Chr. 2 ($h = 0.050$ and TPR = 1.000 with the best OACC = 0.973). However, predictions are unsatisfactory for genomes with high GC variation such as *S. typhi* ($h = 0.009$ and TPR = 0.110 with the best OACC = 0.923). The Zisland Explorer does not depend on the GC variation of the investigated genome. Its TPR rises to 0.614 and OACC to 0.926 for *S. typhi*, and has improved OACC for *V. cholerae* (Table 2) than the original cumulative GC profile.

To find less additional positives, we discard core segments from the input genome with two complementary ways. Clustering segments based on the Euclidean distance is one way and *H* index evaluation is the second way. The former uses the property that islands are composition heterogenic with the core genome, whereas the latter supposes that each island itself is much composition homogeneous. To show they reflect different contents of sequence composition, we calculated the Euclidean distance from each segment to the whole genome and the *H* index, respectively, for each segment. Consequently, average correlation between the two measures in 11 genomes is only 0.466 (R), and hence they are not well correlated. Furthermore, we evaluate their different effect on the actual results in the genome *S. typhi*. The first method identifies 18 core segments, whereas the second method finds five ones, and the intersection between them has only three ones. Based on the above two analyses, both composition heterogeneity and homogeneity could play roles in filtering core segments. Compared with direct identification of islands, the procedure of filtering core segments and then identifying islands could improve 2% TNR in the 11 genomes.

In the performance test using 11 genomes, Zisland Explorer finds more genuine islands than the widely used tools (IslandPick, IslandPath-DIMOB, SIGI-HMM and Islander), with improvement in OACC. Moreover, Zisland Explorer provides a better TPR/TNR balance, TPR/precision balance and MCC. Compared with other tools, the unique GC homogeneity is introduced into Zisland Explorer to promote performance. However, Zisland Explorer and other tools are not good in TPR tests because of the relatively varied features among real islands. Joint application among different methods is a better way to improve prediction power. As successful pioneer work, Brinkman and Langille's group compiled a web service, IslandViewer, to combine the outputs of multiple genomic island predictors [27, 31, 32]. We find that the performance of these widely used tools is indeed further improved by integrating them with Zisland Explorer. To optimize the user's experience, we additionally provide a web service, GUI and open-source code across multiple platforms for Zisland Explorer.

Summarily, we just aim at developing a systematic strategy or protocol to integrate the GC-Profile segmentation method and the cumulative GC profile (island identifying) method to automatically identify genomic islands. These two methods are two tools with different purposes designed by us and/or our collaborators. In this work, they are integrated into the novel tool, Zisland Explorer by appending the step of filtering core regions. Compared with the two original methods, the new tool is improved mainly in its functions. First, it replaced the manually intervened mode into a completely automatic way, and currently it is easily used by experimental researchers. Second, more stable prediction results could be obtained by combining with features of composition bias.

Key Points

- We describe a tool for detecting genomic islands by combining homogeneity and heterogeneity properties.
- Zisland Explorer can obtain more genuine islands with improvement in accuracy.
- Zisland Explorer offers a supplement to increase the performance of other predicting strategy.
- Zisland Explorer is open source and easily to use.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

This work was supported by the National Natural Science Foundation of China (grant numbers 31470068, 31501063, 31171238 and 31571358); Sichuan Youth Science and Technology Foundation of China (2014JQ0051); China National 863 High-Tech Program (2015AA020101); Postdoctoral Science Foundation of China (2015M580211); and Fundamental Research Funds for the Central Universities of China (ZYGX2013J101, ZYGX2015Z006 and ZYGX2015J144). Funding for the open access charge was provided by the National Natural Science Foundation of China (31470068 and 31501063).

Acknowledgments

We are greatly indebted to Prof. Chun-Ting Zhang for inspiring discussion and invaluable assistance.

References

1. Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* 2005;**3**(9):679–87.
2. Dobrindt U, Hochhut B, Hentschel U, *et al*. Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* 2004;**2**(5):414–24.
3. Hentschel U, Hacker J. Pathogenicity islands: the tip of the iceberg. *Microbes Infect* 2001;**3**(7):545–8.
4. Hacker J, Bender L, Ott M, *et al*. Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various extraintestinal *Escherichia coli* isolates. *Microb Pathog* 1990;**8**(3):213–25.
5. Ou HY, Chen LL, Lonnen J, *et al*. 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* 2006;**34**(1):e3
6. Langille MG, Hsiao WW, Brinkman FS. Detecting genomic islands using bioinformatics approaches. *Nat Rev Microbiol* 2010;**8**(5):373–82.
7. Vernikos GS, Parkhill J. Resolving the structural features of genomic islands: a machine learning approach. *Genome Res* 2008;**18**(2):331–42.
8. Ou HY, He X, Harrison EM, *et al*. MobilomeFINDER: web-based tools for *in silico* and experimental discovery of bacterial genomic islands. *Nucleic Acids Res* 2007;**35**:W97–104.
9. Bi D, Jiang X, Sheng ZK, *et al*. Mapping the resistance-associated mobilome of a carbapenem-resistant *Klebsiella pneumoniae* strain reveals insights into factors shaping these regions and facilitates generation of a 'resistance-disarmed' model organism. *J Antimicrob Chemother* 2015;**70**(10):2770–4.
10. Langille MG, Hsiao WW, Brinkman FS. Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* 2008;**9**:329
11. Hsiao W, Wan I, Jones SJ, *et al*. IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* 2003;**19**(3):418–20.
12. Waack S, Keller O, Asper R, *et al*. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* 2006;**7**:142
13. Hudson CM, Lau BY, Williams KP. Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes. *Nucleic Acids Res* 2015;**43**:D48–53.
14. Zhang R, Zhang CT. A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. *Bioinformatics* 2004;**20**(5):612–22.
15. Zhang CT, Zhang R. Genomic islands in *Rhodopseudomonas palustris*. *Nat Biotechnol* 2004;**22**(9):1078–9.
16. Charkowski AO. 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* 2004;**16**(2):180–1.
17. Zhang R, Zhang CT. 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* 2008;451930
18. Zhang R, Ou HY, Gao F, *et al*. Identification of Horizontally-transferred Genomic Islands and Genome Segmentation Points by Using the GC Profile Method. *Curr Genomics* 2014;**15**(2):113–21.
19. Guo FB, Wei W. Prediction of genomic islands in three bacterial pathogens of *pneumonia*. *Int J Mol Sci* 2012;**13**(3):3134–44.

20. Wei W, Guo FB. Prediction of genomic islands in seven human pathogens using the Z-Island method. *Genet Mol Res* 2011;**10**(4):2307–15.
21. Chen LL. Identification of genomic islands in six plant pathogens. *Gene* 2006;**374**:134–41.
22. Nakagawa S, Takaki Y, Shimamura S, *et al*. Deep-sea vent epsilon-proteobacterial genomes provide insights into emergence of pathogens. *Proc Natl Acad Sci USA* 2007;**104**(29):12146–50.
23. Gao F, Zhang CT. GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucleic Acids Res* 2006;**34**:W686–91.
24. Zhang CT, Gao F, Zhang R. Segmentation algorithm for DNA sequences. *Phys Rev E Stat Nonlin Soft Matter Phys* 2005;**72(4 Pt 1)**:041917
25. Zhang CT, Zhang R. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res* 1991;**19**(22):6313–17.
26. Xu Z, Hao B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Res* 2009;**37**:W174–8.
27. Dhillon BK, Laird MR, Shay JA, *et al*. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. *Nucleic Acids Res* 2015;**43**(W1):W104–8.
28. Bi D, Xu Z, Harrison EM, *et al*. ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Res* 2012;**40**:D621–626.
29. Hua ZG, Lin Y, Yuan YZ, *et al*. ZCURVE 3.0: identify prokaryotic genes with higher accuracy as well as automatically and accurately select essential genes. *Nucleic Acids Res* 2015;**43**(W1):W85–90.
30. Guo FB, Xia ZK, Wei W, *et al*. Statistical analyses of conserved features of genomic islands in bacteria. *Genet Mol Res* 2014;**13**(1):1782–93.
31. Dhillon BK, Chiu TA, Laird MR, *et al*. IslandViewer update: Improved genomic island discovery and visualization. *Nucleic Acids Res* 2013;**41**:W129–132.
32. Langille MG, Brinkman FS. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 2009;**25**(5):664–5.