Unsupervised detection and assessment of statistical significance of Genomic Islands

Raghunath Chatterjee[1], Keya Chaudhuri[1] and Probal Chaudhuri [2]*

[1]Molecular & Human Genetics Division, Indian Institute of Chemical Biology, Jadavpur, Kolkata -700 032, India.

[2]Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203, B.T. Road, Kolkata - 700 108, India.

* Corresponding author: Dr. Probal Chaudhuri
Theoretical Statistics and Mathematics Unit
Indian Statistical Institute
203, B.T. Road, Kolkata-700 108, India
Tel : (+91)-(33)-25753414 or 25753400
Fax: (+91)-(33)-25773071 or 25776680
Email: probal@isical.ac.in

ABSTRACT

Many of the available methods for detecting Genomic Islands (GIs) in prokaryotic genomes use markers such as transposons, proximal tRNAs, flanking repeats etc., or they use other supervised techniques requiring training datasets. Most of these methods are primarily based on the biases in GC content or codon and amino acid usage of the islands. However, either these methods do not use any formal statistical test of significance or use statistical tests for which the critical values and the P-values are not adequately justified. We propose a method, which is unsupervised in nature and uses Monte-Carlo statistical tests based on randomly selected segments of a chromosome. Such tests are supported by precise statistical distribution theory, and consequently, the resulting P-values are quite reliable for making the decision. Our algorithm runs in two phases. Some '*putative* GIs' are identified in the *first phase,* and those are refined into smaller segments containing horizontally acquired genes in the *refinement phase.* This method is applied to three prokaryotic genomes leading to the discovery of several new pathogenicity, antibiotic resistance and metabolic islands that were missed by earlier methods. Many of these islands contain mobile genetic elements like phage-mediated genes, transposons, integrase and IS elements confirming their horizontal acquirement.

Key words: Horizontal Gene Transfer, Putative GI, Integron Island, Monte-Carlo Statistical Test, Oligo-nucleotide Distributions, Pathogenicity Island, P-value.

INTRODUCTION

Horizontal gene transfer is an important mechanism for the evolution of microbial genomes. In 1990, it was first observed that large blocks of horizontally acquired foreign sequences occur in chromosomes of pathogenic bacteria, and those regions are highly correlated with pathogenicity (1-3). Some of these possess mobile elements consisting of a gene for specific recombinase and sequences having characteristics of integration sites. Some others, despite their apparently foreign nature, lack insertion sequences, recombinase genes and specific *att* sites, and they may contain only fragments of mobility genes. In the latter case, the mobility sequences were probably lost in course of evolution after their integration into the bacterial genome (1). The first known foreign DNA blocks that were proved to be associated with virulence genes of pathogenic bacteria were named as *pathogenicity islands* (4). Later on, genomes of non-pathogenic bacteria have been shown to contain foreign gene blocks, which are not associated with virulence. These gene blocks determine various accessory functions like secondary metabolic activities, antibiotic resistance, symbiosis and other special functions related to the survival in harsh environmental conditions (5). Subsequently, all foreign gene blocks are collectively named in the literature as genomic islands (GIs) (5,6). There is an extensive literature on the study of GIs in prokaryotic genomes (7,8). GIs in prokaryotic genomes often contain horizontally transferred genetic materials as evident from the presence of integrase, transposons, phage mediated genes, etc. in these islands. Consequently, they are critically important in the study of the evolution, the pathogenesis and other special features of prokaryotic genomes.

Several methods have been reported and discussed in the available literature for detecting GIs in prokaryotic genomes (9-13). Many of these methods use markers such as transposons, proximal tRNAs, flanking repeats etc. to identify GIs (9,11,14). Mantri and Williams (11) used tRNA and tmRNA as markers. They further searched for the phage integrase and passed through different filtration procedures for the identification of GIs. Ou *et al.* (9) also started with tRNA and tmRNA genes as primary markers, and after passing through different filtration techniques, the GIs were identified. In another previous study, the authors have identified the GIs after performing the cluster analysis of the chromosomal fragments, which are formed by fragmenting the chromosomes based on locations of transposons (14). Such methods, which are based on standard markers, are particularly useful for detecting GIs acquired by a genome from another compositionally close donor genome or those, which have become compositionally close to the host genome due to the amelioration process. In such cases, the islands may not bear any compositional signature that can be used to distinguish it from rest of the host genome.

Consequently, identification of such islands has to rely on possible presence of structural features, like tRNA, direct repeats (DR), integrase gene etc. However, there are limitations of such methods, which are based on standard markers. Firstly, the GIs, which are associated with standard markers, can only be identified by this method. Secondly, there may be intra-chromosomal rearrangements, and islands may no longer be in the proximity of those standard markers after such rearrangements. Consequently, many GIs may not be detected by marker-based methods (7).

In an earlier paper (15), the authors used discriminant analysis, a supervised statistical technique, based on a training data-set that was formed by the authors using the aggregation of known GIs from different organisms. However, unless there are several organisms with some statistical similarities in their genome sequences as well as in their known GIs, such an aggregation to form the training data-set may not be appropriate. Besides, the GIs available and known *a priori* for a single organism may be very few at the beginning of the investigation.

In this paper, we have developed a method that does not use any standard marker when islands are searched in the genome. Islands identified by this method may, however, be confirmed subsequently by supporting factors that include such markers as well as possible presence of known horizontally transferred genes (e.g., phage mediated genes). This will be clear in the section where we discuss the results. Further, the proposed method is unsupervised in nature, and it does not require any training data set for its implementation.

Our method searches for islands in a prokaryotic chromosome using a probing window that slides over the entire chromosome and also varies in its size. For a given size and a given position of that probing window, the segment of the chromosome captured by the window is compared with the rest of the chromosome by means of some statistical tests. The outcome of each such test is a statistical P-value that lies between zero and one. A low P-value, which indicates a significant difference between the segments captured by the probing window and the rest of the chromosome, bears evidence for the probing window having a substantial overlap with a GI. All these P-values obtained from statistical tests carried out at different locations and for different sizes of the probing window can be represented by a 3D plot, which enables visualization of locations and sizes of GIs in the chromosome. For the determination of GIs, window based methods have been used in some earlier studies. The GIs of *Pseudomonas putida* KT2440 were determined by analyzing the compositional bias of the mono-, di- and tetra-nucleotide contents in the segment of the genome under the probing window of 4000 bp that slides in steps of 1000 bp (16). These authors, however, have used windows with fixed lengths, and

there is no objective guideline for how to determine that length in practice. Zhang and Zhang (10) used a *windowless* method for displaying the distribution of genomic GC content, and the cumulative GC profile was used by them for the determination of GIs. *Abrupt jump* in cumulative GC profile, which is due to relatively different GC content of an island, enabled them to identify the GI. But this was done in a subjective manner and neither clear quantitative measure nor any formal statistical test for assessing the abrupt change in the cumulative GC profile was proposed by these authors.

Known methods for identifying GIs are primarily based on GC contents of the islands, their oligo-nucleotide usage patterns and the codon usage biases in the genes present in the island (10,12,13,16). When a fixed segment under the probing window is compared with the whole chromosome, which may contain several GIs (in some cases it might be as large as 20% of the whole chromosome (17)), such a comparison is likely to get influenced by those islands, and this reduces the resolution of the comparison. In our method, the comparison is based on randomly selected segments of the chromosome with the fixed segment under the probing window. Each randomly selected segment has the same size as the fixed segment under consideration, and there is no overlap between a random segment and the fixed segment under investigation. This will be discussed in detail in the section on methods.

Various procedures studied in the literature generally lack a formal and rigorous statistical treatment of the problem of comparing a segment of the chromosome with the rest of the chromosome in order to decide whether the segment is the part of a GI or not. Often no formal statistical test is carried out, and the decision to declare a segment as part of an island is done in a subjective way as mentioned earlier. In some other cases, statistical tests have been carried out in a way that is somewhat questionable in the sense that the determination of the critical values and the P-values is not adequately justified due to lack of a rigorous statistical distribution theory of the deviance measures used for such tests. Yoon *et al.* (18) used Mahalanobis distance to evaluate the deviation of the codon usage of a gene from the mean of that in the genome. They assumed normal distribution of codon frequencies without much justification for it, and converted the Mahalanobis distance into a P-value using the $\chi 2$ distribution function. They have considered a gene as extraneous in codon usage if its P-value was less than 0.05 (18). On the other hand, Zhang and Zhang (10) obtained their results based on codon usage and amino acid usage biases using different cut-offs for the P-values. In some earlier studies (19,20), authors used higher order motifs to capture the compositionally deviating regions from the genome. In another study by Vernikos et al. (21), authors used variable order motifs and relative entropy for the detection of

compositionally deviating regions. In our method, we have used a Monte-Carlo statistical test, which is partly motivated by the idea of the bootstrap method in statistics (22,23) for comparing the segment under the probing window with randomly selected segments from the rest of the chromosome. Such Monte-Carlo statistical tests based on randomly selected segments of the chromosome can be supported by simple and precise statistical distribution theory. Consequently, the P-values obtained in our method will be quite reliable for making the decision.

METHODS

Let us denote a whole chromosomal sequence of an organism by *S*, and *s* will denote a given segment of *S*. In order to assess whether *s* differs significantly from the rest of *S*, we need a measure of distance that can be used for quantitative comparison between the given segment *s* and any other segment *s´* of *S* not having any overlap with *s*. Such a distance measure, which we may denote as *d*(*s, s´*), can be based on GC contents of *s* and *s´* or their oligo-nucleotide distributions. For instance, one may use the absolute distance, the Euclidean distance or Kullback-Leibler divergence computed from oligo nucleotide frequencies. Alternatively, for annotated genomes, one may form the distance measure *d*(*s, s´*) by comparing the gene contents of *s* and *s´* and their codon and amino acid usage biases.

Merkl *et al.* (12) used codon usage analysis of two species assuming the similarity of codon usage in phylogenetically related species. Weinel *et al.* (16) analyzed the di-nucleotide usage and the tetra-nucleotide usage in sliding windows and compared them with the di-nucleotide usage of the whole genome and uniform tetra-nucleotide usage respectively. In the study by Zhang and Zhang (10), putative GIs detected by cumulative GC profile were further analyzed by codon usage and amino acid usage of those regions compared to the whole chromosome *S*. Comparison of the codon usage and oligo-nucleotide usage of the given segment *s* with those for the whole chromosome *S* has some drawbacks because *S* may contain several GIs. In some cases, the total size of the GIs in *S* would be much larger than the length of *s*, and it can be as large as 20% of the size of *S* (17). This may statistically contaminate values of various parameters related to GC content as well as oligo-nucleotide and codon usage biases when computed for the entire chromosomal sequence *S*. This is likely to reduce the resolution of the comparison.

In our method, the comparison between *s* and the rest of *S* is based on *N* randomly selected segments $s_{1,1}, s_{1,2}, s_{1,3,}.....s_{1,N}$ from the chromosome *S*, each of which has the same length as that of *s*, and none of them has any overlap with *s*. This substantially reduces the influence of various possible

islands present in *S* on any statistical comparison between *s* and the rest of *S*, and that in turn increases the resolution of the comparison. We also choose *N* random pairs of segments $\left(s_{2,1}, s_{3,1}\right), \left(s_{2,2}, s_{3,2}\right), \left(s_{2,3}, s_{3,3}\right), \ldots\ldots\ldots\ldots \left(s_{2,N}, s_{3,N}\right)$ from *S*, where for each $1 \le i \le N$, $s_{2,i}$ and $s_{3,i}$ are independently selected, and each of them has the same length as the given segment *s* and no overlap with *s*. Then, we can compute the distances $d_{1,i} = d\left(s, s_{1,i}\right)$ and $d_{2,i} = d\left(s_{2,i}, s_{3,i}\right)$ for $1 \le i \le N$ and form the following two sets of distance values:

$$D_1 = \left\{ d_{1,i} \,\middle|\, 1 \le i \le N \right\} \text{ and } D_2 = \left\{ d_{2,i} \,\middle|\, 1 \le i \le N \right\}.$$

If *s* happens to be a part of a GI with characteristics very different from the rest of *S*, the values in D_1 are expected to be larger than those in D_2. Otherwise, the values in the two sets are expected to be of the same order of magnitudes.

Statistical test for comparing *s* with the rest of *S*

In view of the way the distance values in D_1 and D_2 have been obtained by random sampling of segments of *S*, the values in each of these two sets can be viewed as independent and identically distributed random variables, and the values in D_1 will be completely independent from the values in D_2. The problem of comparing the values in the two sets D_1 and D_2 can be formulated as a statistical testing problem, where the null hypothesis can be taken as H_0: "the expected value of an element in D_1 is the same as that of an element in D_2", and the alternative hypothesis would be H_A: "the expected value of an element in D_1 is larger than that of an element in D_2". We set

$$m_1 = N^{-1} \sum_{i=1}^{N} d_{1,i}\,, \qquad s_1^2 = N^{-1} \sum_{i=1}^{N} \left(d_{1,i} - m_1 \right)^2,$$

$$m_2 = N^{-1} \sum_{i=1}^{N} d_{2,i}\,, \qquad s_2^2 = N^{-1} \sum_{i=1}^{N} \left(d_{2,i} - m_2 \right)^2.$$

Then, each of m_1 and m_2 is approximately normally distributed being an average of independent and identically distributed random variables by the well-known central limit theorem in probability theory if *N* is large. Further, m_1 and m_2 are independently distributed, and s_1^2 / N and s_2^2 / N will be the standard estimates for their variances respectively. Hence, the statistic

$$Z = \frac{m_1 - m_2}{\sqrt{\{(s_1^2/N) + (s_2^2/N)\}}}$$

will be approximately normally distributed for large N, and the mean of that normal distribution will be zero if H_0 is true, and it will be positive if H_A is true. The variance of that asymptotic normal distribution will be one under both hypotheses. Consequently, Z can be used as a test statistic for testing H_0 against H_A in a one-sided test. Here, the P-value can be computed using the observed value of Z for the given segment s under the probing window and the standard normal distribution. This way of assessing the statistical significance of the evidence for s being part of a GI in the chromosome S using a Monte-Carlo test based on random samples of segments from S is partly motivated by the idea of the bootstrap (22,23). In the present study, we have used $N = 200$. For larger values of N, the normal approximation will be more accurate for the distribution of the test statistic, but the corresponding computation time will also increase substantially. For some smaller chromosomes, we have tried values of N upto 500, but the results did not change significantly.

If for some reasons (e.g., computational constraints), one is forced to use smaller values of N, the normal approximation for the distribution of Z will not be valid. In that case, one may work with a different formulation of the statistical hypotheses as follows. The null hypothesis in that case can be formulated as H_0: "the statistical distribution of an element in D_1 is the same as that of an element in D_2", and the alternative hypothesis can be formulated as H_A: "the distribution of an element in D_1 is *stochastically larger* than that of an element in D_2". With these re-formulated hypotheses, one can carry out the test using two-sample Kolmogorov-Smirnov statistic (24) or Wilcoxon-Mann-Whitney statistic (24-26). These test statistics have been used by previous authors (14). However, the power of such non-parametric statistical tests for detecting GIs tends to be less than the preceding test based on normal distribution, which is applicable for relatively larger values of N.

Statistical analysis with segments having variable sizes and locations

In order to identify islands at different locations of the chromosome and to determine the stretches of those islands, it is necessary to carry out our statistical analysis using a probing window that slides across the chromosome and also varies in its size. The statistical test described above can be implemented for any location and size of the segment s under that probing window, and the P-value can be computed. It would be useful to plot these P-values so that one can visualize possible locations of the

islands in the chromosome as well as their stretches. Such a plot of P-values would also enable us to assess visually the statistical significance of the evidence for or against different segments of the chromosomes to be possible parts of GIs.

For visual presentation of the '*putative* GIs' identified by the analysis described above, a 3D plot for a chromosome can be generated. In this 3D plot, chromosomal locations of the probing window are plotted along the x-axis, corresponding probing window sizes are plotted along the y-axis, and the P-values in gray scale are plotted along the z-axis. Here, the P-value for a specific location and size of the window is plotted using a gray scale that changes gradually from black to white, where black corresponds to the extreme P-value = 0, and white corresponds to the other extreme P-value = 1. The white dots corresponding to higher P-values become almost invisible in the white background while dark dots corresponding to low P-values will be prominently visible marking the '*putative* GIs' in the chromosome.

For a specified value of P_0 ($0 < P_0 < 1$), one can determine all the segments of a chromosome that are associated with a P-value less than or equal to P_0. This will lead to the identification of some '*putative* GIs' having varying sizes and locations in the chromosome that are identifiable with P-values equal to P_0 or smaller. Ranges of the '*putative* GIs' in terms of their chromosomal locations can be determined using the cut-off value P_0 and considering a specified number r of overlapping windows of variable sizes having P-values smaller than or equal to P_0.

Further refinement of the '*putative* GIs' identified by the *first phase* of the algorithm.

'*Putative* GIs' obtained using our previous analysis, which may be considered to be the *first phase* of the algorithm, are always of larger size than what they are supposed to be because of the presence of many 'false positives' (i.e., segments of the genome that are statistically detected as GIs but are not biologically parts of any true island). To reduce the false positives and increase the specificity of our method, a *refinement phase* with a fixed overlapping probing window of size w over the regions detected as '*putative* GIs' by the *first phase* of the analysis has been performed. Random samples of genomic segments in the *refinement phase* were chosen from the genome excluding the regions detected as '*putative* GIs' in the *first phase*. The comparison between a probing window w and the rest of S is again based on N randomly selected segments $w_{1,1}, w_{1,2}, w_{1,3,}.....w_{1,N}$ from the chromosome S, each of which has the same length as that of w. The method is very similar to the method used in the *first phase*.

The P-values are generated using Monte-Carlo tests carried out at variable locations of the probing window with a fixed size.

In the *first phase* of our analysis, the presence of several GIs in the genome may statistically contaminate the randomly sampled segments by affecting their oligo-nucleotide distributions. Only after identifying '*putative* GIs' in the *first phase*, it would be possible to take random sample of segments from cleaner stretches of the genome that would exclude the '*putative* GIs' detected in the *first phase.*

A smaller probing window is recommended for the *refinement phase* as it will provide an accurate way of precisely detecting the GIs. Gene order conservation is rarely observed in distantly related species and several rearrangements and movement of genes occurs frequently. So, some genes, which are not horizontally acquired from other species, may be present within a '*putative* GI' identified in the *first phase,* and to some extent, this problem is taken care of by the use of a smaller probing window. However, the use of smaller probing window requires randomly sampled segments from non-contaminated stretches of the genome, and those stretches are available after running the *first phase*.

Smaller probing windows are not recommended in the *first phase* because it increases the computational cost during the *first phase*. The use of smaller probing windows that slide over the genome lead to a large number of statistical tests, and this may produce many false positive results. Further, there are high chances of substantial overlap of a randomly selected window in the *first phase* with an island in the genome containing horizontally acquired materials.

As in the *first phase* of the analysis, for a specified value of P_0 ($0 < P_0 < 1$), one can again determine all the segments of a '*putative* GI', which is identified in the *first phase*, that are associated with a P-value less than or equal to P_0.

Choice of different parameters associated with the algorithm

In the following section, we have presented results obtained using the absolute distance based on tetra-nucleotide frequencies. Those results are obtained using $P_0 = 0.05$ and $r = 5$ in *first phase* and $P_0 = 0.001$ in the *refinement phase*. The value of P_0 in the *first phase* was relaxed to 0.05, and it was chosen in such a way that most of the horizontally acquired stretches of the genome could be captured by the '*putative* GIs' detected in the *first phase*. After we obtain the '*putative* islands', we would be able to generate some statistically non-contaminated stretches of the genome (i.e., genomic regions excluding those putative islands). Those stretches can be used for random sampling of segments in the *refinement phase*. In order to determine the value of P_0 in the *refinement phase*, we have carried out a

performance assessment of our method for different values of P_0 based on a dataset related to *Salmonella typhi* CT18 generated by Vernikos et al. (21). Their method of constructing the dataset of putative horizontally transferred genes is discussed briefly in the section on results and discussion. We have calculated and plotted the sensitivity (SN), the specificity (SP) and the accuracy (AC) of our method for different values of P_0 ranging from P_0 = 0.05 to 0.00001 (Fig.1A). The slopes of the curve for SN, SP and AC were also plotted for different values of P_0 (Fig.1B). As this cut-off P-value increases, the specificity and the accuracy increase, but the sensitivity decreases. The specificity and the accuracy increase steadily up to P_0 = 0.001 (Fig.1A), and then the slope of each of the two curves decreases (Fig.1B). The sensitivity decreases with the increase in the cut-off P-value, but in the region from P_0 = 0.05 to P_0 = 0.001, the sensitivity decreases slowly, and then it decreases much more sharply. Alternatively, one can determine the value of P_0 using the ROC curve approach.

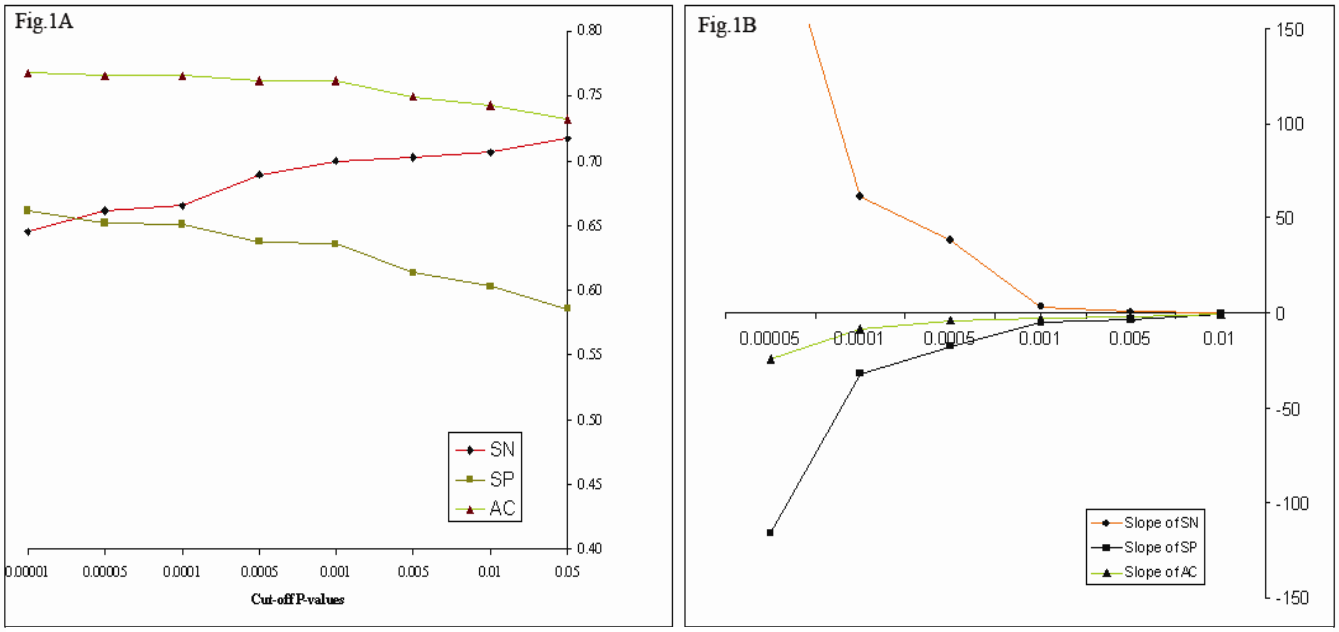


Fig.1: The influence of different choices of cut-off P-values (P_0) used in the *refinement phase* on the sensitivity (SN), the specificity (SP) and the accuracy (AC) of *Design-Island* applied to a manually curated data set of 1560 putative horizontally transferred genes of *Salmonella typhi* CT18 generated by Vernikos et al. (21) is shown in Fig.1A. Fig.1B shows corresponding variations of slopes of the curves for SN, SP and AC for different choices of cut-off P-values (P_0).

When we used that technique with a range of P_0 values from 0.05 to 0.00001, it again led to the same value of P_0 as the optimal, and we have chosen the cut-off P-value as P_0 = 0.001 for the *refinement phase.* This has been used in all our analysis of different bacterial genomes. It is possible that for some other bacterial genomes a different choice of P_0 would be optimal depending on the nucleotide

compositions of those genomes. However, the empirical results obtained for all the three different bacterial genomes using this choice of P_0, which is *S. typhi* CT18 specific, demonstrate good performance of our algorithm.

We have carried out our analysis with distance measures based on oligo-nucleotides of different orders (i.e., sizes). The islands detected by methods based on different orders of oligo-nucleotides did not differ considerably. Only in some cases either the boundaries of the segments of the '*putative* GIs' slightly differed or a single '*putative* GI' broke into two or more segments. In most of the organisms, the '*putative* GIs' detected using tetra-nucleotide analysis include those detected by other analysis based on other oligo-nucleotides, and the later analysis sometimes missed some of the important segments of the genomes containing known horizontally acquired materials. As we will see in the section containing a comparative study of different methods, our method outperformed the method W8 (20), which is a method based on octa-nucleotides, in many cases.

We have considered three types of distance measure computed using oligo-nucleotide frequencies. These are the absolute distance, the Euclidean distance and the Kullback-Leibler divergence. But all these distances lead to almost the same result. The '*putative* GIs' detected by methods based on different distances tend to differ in their boundaries to a small extent. We have finally decided to use the absolute distance, which is computationally the simplest among all the distances considered.

It will be appropriate to note here that our method as well as some of the earlier methods (e.g., IVOM, HGT-DB, IslandPath, W8 etc.) are designed to identify those GIs, which exhibit oligo-nucleotide compositions that deviate significantly from the rest of the genome. Consequently, GIs, which do not bear any compositional signature that can distinguish them from rest of the genome, may not be detected by such methods. As we have already mentioned in the introduction, for such GIs, which are compositionally similar to the rest of the genome, structural features such as tRNA, DR, integrase genes etc. may be used to identify them.

One may, in principle, use distances computed using codon usage or amino acid usage biases instead of oligo-nucleotide distributions. However, that will require the use of complete annotation of the entire chromosome and the gene content of each and every randomly selected segment for our Monte-Carlo test. This makes the implementation of the method computationally challenging, and we have not pursued that here.

The entire methodology is presented in the form of a flow chart in Fig.2A, 2B, and we have named our method as *Design-Island* (an acronym for *Detection of Statistically Significant Genomic Island*). The computer programs for *Design-Island* will be uploaded on the Internet in near future. It can also be obtained from the authors by sending a request by e-mail.

Fig.2A

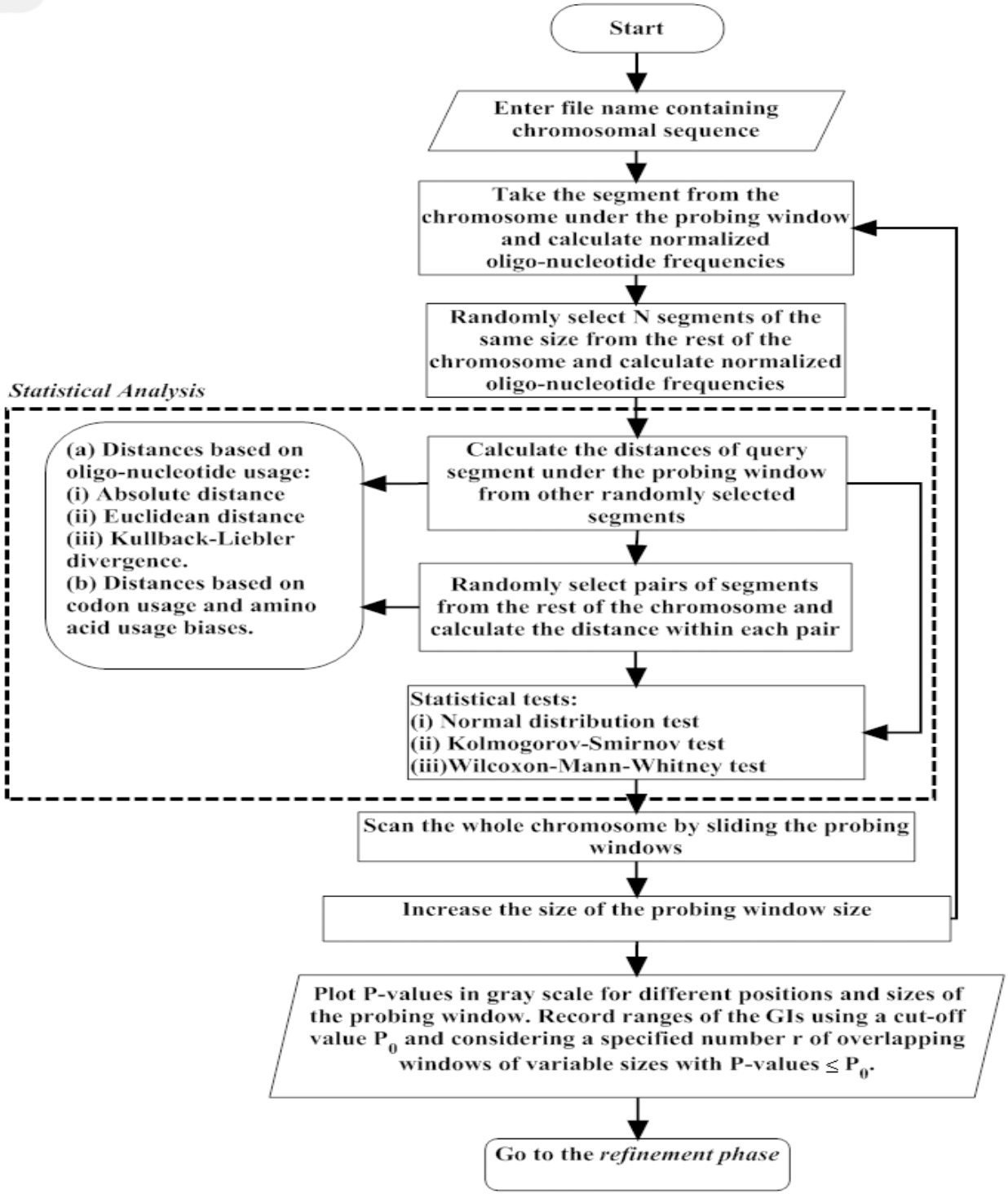


Fig.2B

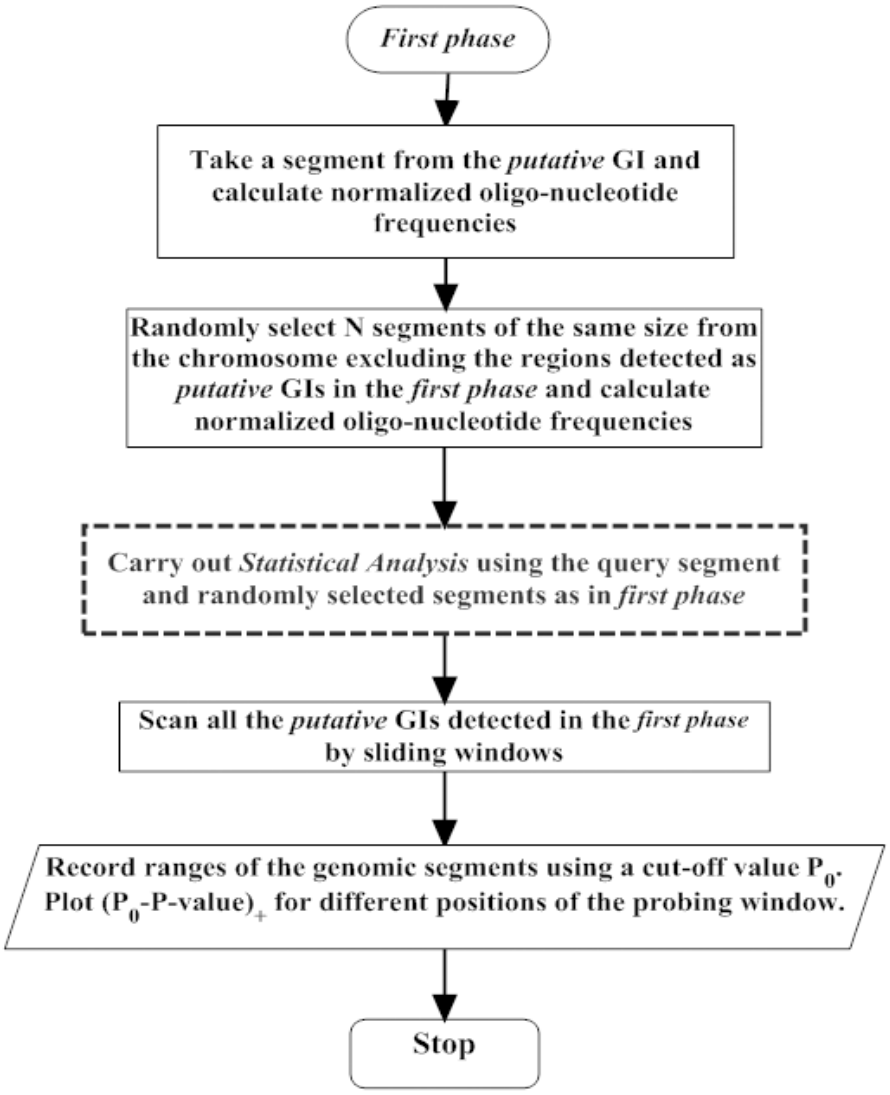


Fig.2: Algorithmic flow-charts of the *first phase* (Fig.2A) and the *refinement phase* (Fig.2B) of *Design-Island*.

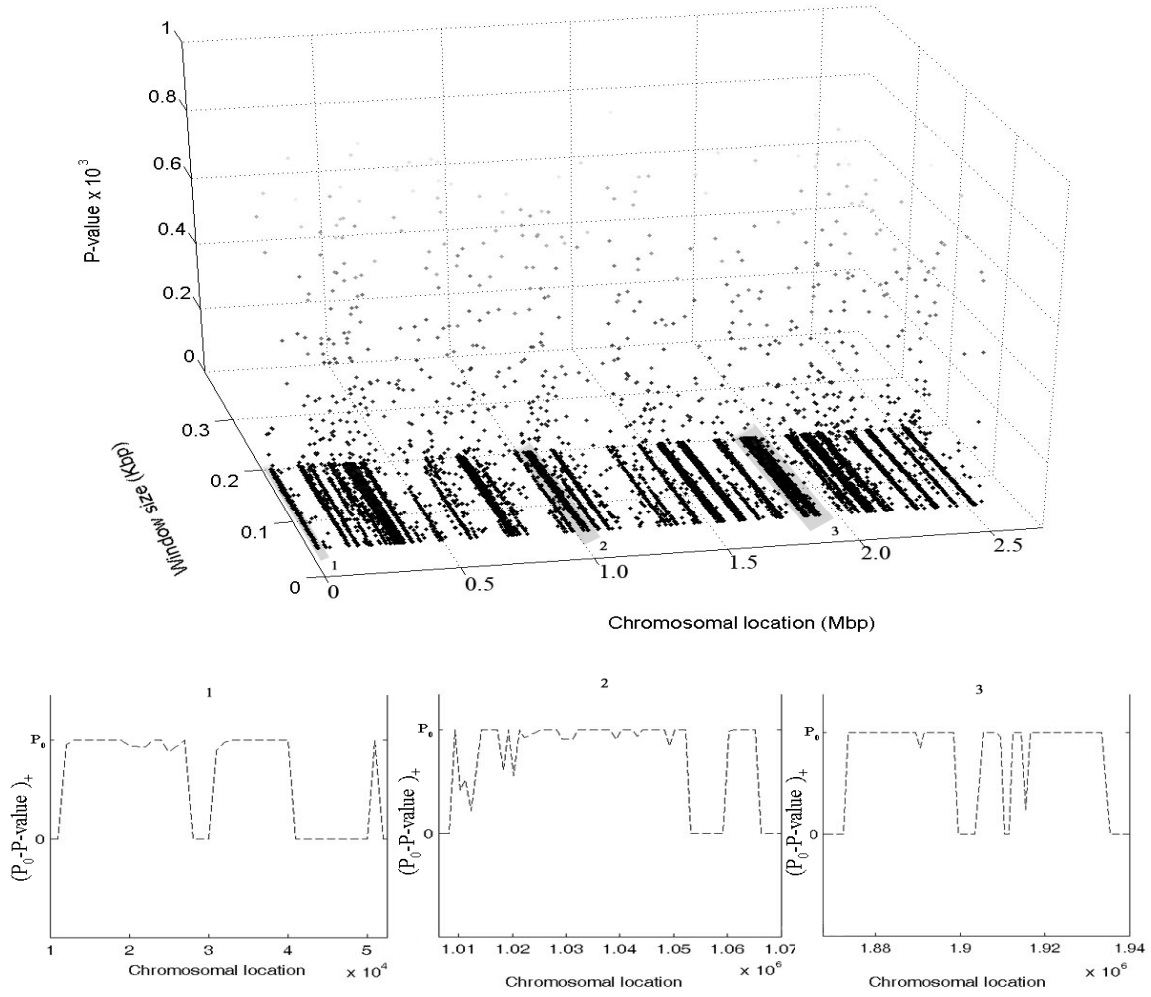
RESULTS AND DISCUSSION

We have implemented *Design-Island* on chromosomal sequences of three prokaryotic genomes, obtained from NCBI database (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). The co-ordinates of statistically significant genomic segments detected by *Design-Island* and their gene contents in five chromosomes of these three prokaryotes are presented in Supplementary Table-1, and these segments are discussed below.

***Salmonella typhi* CT18**

Salmonella enterica serovar Typhi (*S. typhi*), an aetiological agent of typhoid fever, is a serious invasive bacterial disease of human. Many *S. enterica serovars* actively invade the mucosal surface of the intestine but are normally contained in healthy individuals by the local immune defense mechanism. However, *S. typhi* has evolved the ability to spread to the deeper tissues of human including liver, spleen and bone marrow (27). In *S. typhi*, thirteen pathogenicity islands (popularly known as SPIs – *Salmonella* Pathogenicity Islands) and five islands containing bacteriophages related genes have been reported (21,27).

Fig.3A(i)

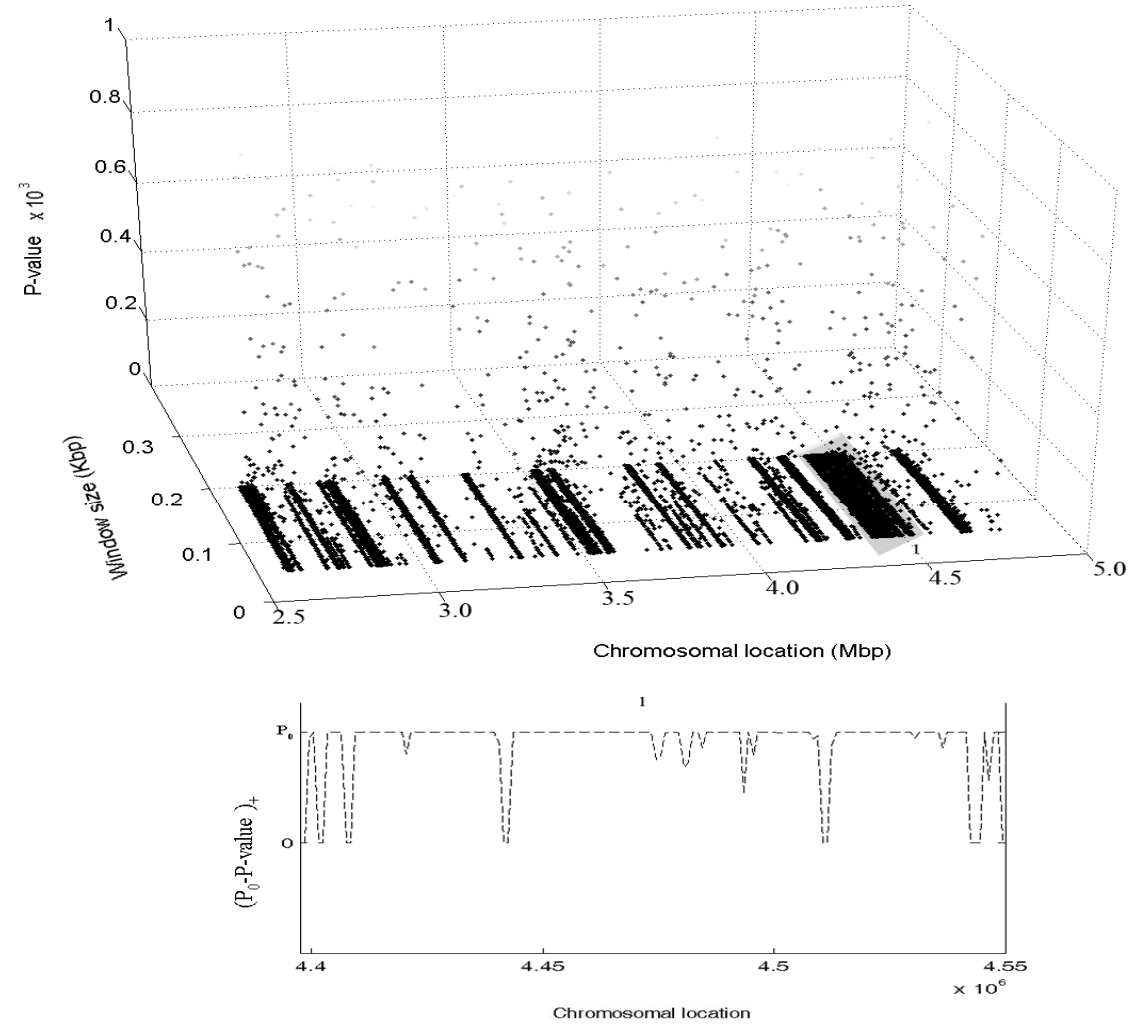


In *S. typhi* CT18 *Design-Island* detected ninety seven '*putative* GIs' in the *first phase,* and after refinement, these islands are broken into two hundreds and twenty-one statistically significant genomic segments that include all of the GIs detected in the previous studies. Major gene content of these segments code

for phage proteins, putative pathogenicity island proteins, virulence associated secretory protein, Vi polysaccharide proteins, integrase, phage integrase, putative bacteriophage proteins, IS element transposases, flagellar proteins, UV protection protein, type III secretion system, type III restriction-modification system, killing factor KicA and B, different chains of NADH dehydrogenase and heat shock proteins. Among the newly detected genomic segments, the major genes present are those, which code for putative toxin like proteins, putative virulence proteins, putative phage proteins, integrase, type III restriction modification system, some pseudo genes, some transporters, flagelins, different chains of NADH dehydrogenase and ATP synthase, peniciline binding protein, fimbrial subunits, lipopolysaccharide core biosynthesis protein, heat shock and cold shock proteins.

Two 3D plots generated from the *first phase* of our algorithm and some representative 1D plots generated from the *refinement phase* of the algorithm applied to the chromosome of *S. typhi* CT18 are shown here in Fig.3A(i), 3A(ii). The first plot corresponds to the stretch of the chromosome from the start of the chromosome up to 2.5 Mbp position (Fig.3A(i)), and the other plot corresponds to the stretch of the chromosome from 2.5 Mbp position up to the end (Fig.3A(ii)). Representative 1D plots for four of the '*putative* GIs' detected in the *first phase* and enclosed in gray blocks are shown in the lower panel of the figures. The '*putative* GI' that stretches from 10000 to 52500 is fragmented into three segments, namely 11000-28000 bp, 30000-41000 bp and 50000-52000 bp. The '*putative* GI' that stretches from 1006250 to 1070000 bp is fragmented into two segments, namely 1008250-1053250 bp and 1059250-1066250 bp. The '*putative* GI' that stretches from 1867500 to 1940000 bp is fragmented into three segments, namely 1872500-1899500 bp, 1903500-1910500 bp and 1911500-1934500 bp (Fig.3A(i)). In Fig.3A(ii), the 1D plot for the '*putative* GI' that stretches from 4397500 to 4550000 bp is shown in the lower panel. This '*putative* GI' is fragmented into six segments, namely 4398500-4401500 bp, 4402500-4407500 bp, 4408500-4441500 bp, 4442500- 4510500 bp, 4511500-4542500 bp and 4544500-4549500 bp. The third, the fourth and the fifth segments of the above mentioned '*putative* GI' contain mainly phage genes, some pseudo genes and the Vi polysaccharide, which is the major virulence determinant in *S. typhi*. After running the *refinement phase,* the genes excluded from the above mentioned '*putative* GIs' are mainly DNA polymerase III, theta subunit, transcriptional activator protein, putative transcriptional regulator, exodeoxyribonuclease X, ribosome modulation factor (protein E), possible sulfatase regulatory protein, serine/threonine protein phosphatase 1, putative ion and/or amino acid symporter, aminopeptidase N and some hypothetical and conserved hypothetical proteins.

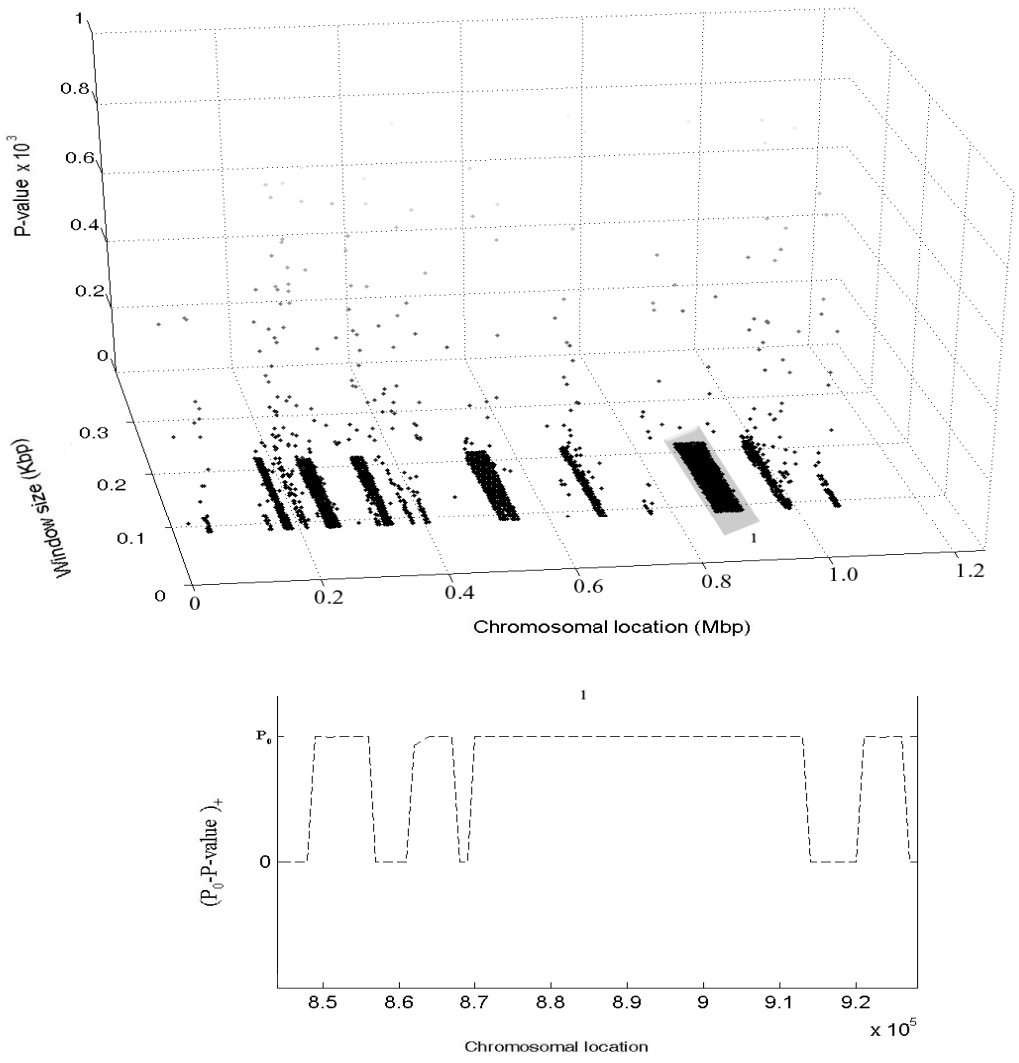
Fig.3A(ii)



***Vibrio cholerae* (Chromosome-I)**

Vibrio cholerae is the etiological agent of cholera, which is a well-known diarrheal disease that often occurs in the form of an epidemic. There is a wide variety of strains and biotypes of *V. cholerae*. They receive and transfer genes for toxins, colonization factors, antibiotic resistance, capsular polysaccharides and new surface antigens (28). The lateral or horizontal transfer of some of the phage mediated virulence genes, pathogenicity islands and other accessory genetic elements have already been described in the literature (29,30). In chromosome-I of *V. cholerae* N16961, Hsiao et al. (31) used two different methods (DINUC and DIMOB), and for each method, they have used two different datasets. In one, they considered all ORFs, and in the other, they considered only the ORFs having length more than 300 bp. Their four methods detected twenty-one, eighteen, five and six GIs respectively.

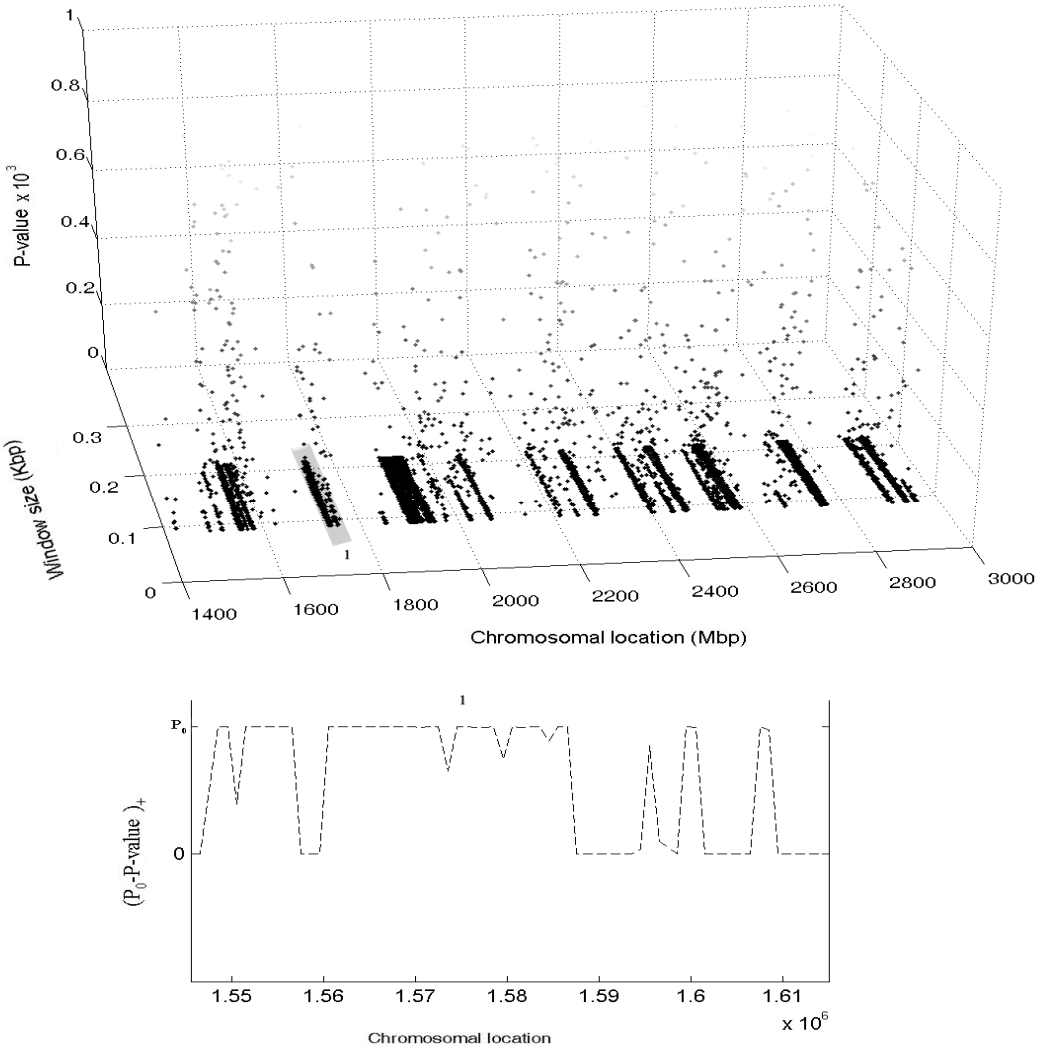
Fig.3B(i)



In chromosome-I of *V. cholerae,* our algorithm *Design-Island* has detected thirty seven '*putative* GIs' in the *first phase.* After the *refinement phase,* these islands are fragmented into one hundred and sixty-two statistically significant genomic segments, and each of the GIs detected by earlier authors has substantial overlaps with one or several of these segments. Many of these segments were missed by earlier authors though they clearly contain horizontally acquired genetic materials. The major gene contents of some of these genomic segments are pathogenic and pathogenicity related genes like cholera enterotoxin, zonula occludens toxin, TCP pilus virulence regulatory protein, type IV pilin, MSHA biogenesis proteins, MSHA pilin proteins, flagella and flagella associated genes, RTX toxin transporter, RTX toxin activating protein, RTX toxin RtxA, agglutination protein, surface antigen and iron acquisition system genes. Besides, there are antibiotic resistance islands containing multi-drug resistance genes and mobile genetic elements like phage genes and phage related genes, integrase or recombinase and transposase. In

some of these segments, the major gene contents are thiamin biosynthesis proteins, sulfite reductase (NADPH) flavoproteins, methyltransferase, ABC transporters, 3-isopropylmalate dehydratases and acetyltransferases, which have been predicted to be parts of the GIs in *V. cholerae* and other micro-organisms (31). The GIs containing virulent genes like RTX toxin transporter, RTX toxin activating protein, RTX toxin RtxA genes, MSHA biogenesis proteins and MSHA pilin proteins have been missed in most of the previous studies.

Fig.3B(ii)



Two 3D plots generated from the *first phase* of our algorithm and some representative 1D plots generated from the *refinement phase* of the algorithm applied to chromosome-I of *Vibrio cholerae* are shown here in Fig.3B(i), 3B(ii). We have presented the '*putative* GIs' detected in the *first phase* of the algorithm using two plots for convenient visualization. The first plot corresponds to the stretch of the chromosome from the start of the chromosome up to 1.4 Mbp position (Fig.3B(i)), and the other plot corresponds to

the stretch of the chromosome from 1.4 Mbp position up to the end (Fig.3B(ii)). The 1D plot for the '*putative* GI' enclosed in the gray block that stretches from 844000 to 928000 bp and detected as *Pathogenicity Island* in earlier studies in the 3D plot is shown in the lower panel. It is fragmented into four genomic segments, namely 848000-857000 bp, 861000-868000 bp, 869000-914000 bp and 920000-927000 bp (Fig.3B(i)). The genes excluded from the '*putative* GI' after the *refinement phase* are mainly some transcriptional regulators and hypothetical genes. The 1D plot for the '*putative* GI' that stretches from 1545500 to 1615000 bp is also fragmented in five segments, namely 1547500-1557500 bp, 1559500-1587500 bp, 1594500-1596500 bp, 1598500-1601500 bp and 1606500-1609500 bp (Fig.3B(ii)). In the 1D plot, the first two segments contain Rtx genes and CtxA and B, Zot, phage replication protein and some transposase. The remaining three segments contain hypothetical genes, tail-specific protease and putative solute/DNA competence effector. After the *refinement phase*, the genes, which are excluded from the '*putative* GI', are ribosome modulation factor, aminopeptidase N and some hypothetical proteins.

***Vibrio cholerae* (Chromosome-II)**

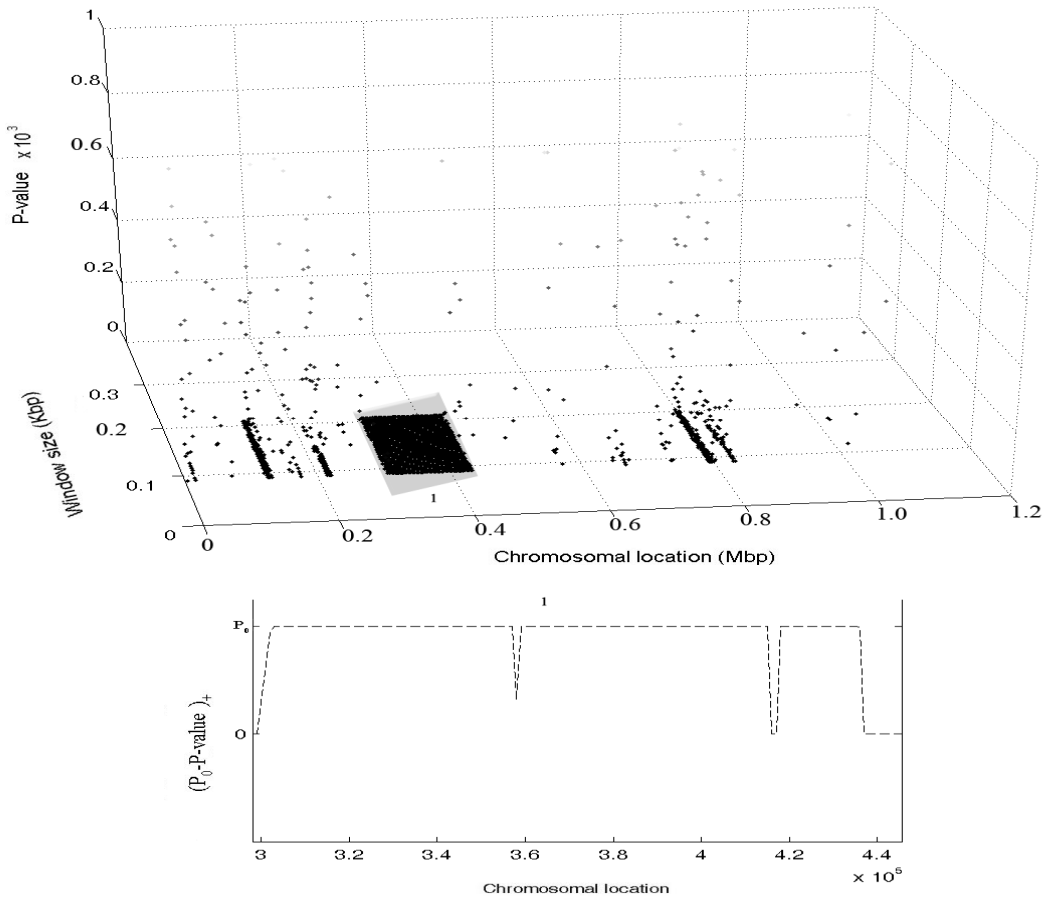
In chromosome-II of *V. cholerae* N16961, Hsiao et al. (31) detected ten, eight, five and five GIs by their four methods discussed earlier. Eight of the ten GIs detected by their first analysis are positioned either as adjacent segments or nearly adjacent segments in the chromosome-II of *V. cholerae* genome. The region containing these eight GIs were determined and named as an *integron island* by other authors (28,32,33).

Using *Design-Island*, we have detected seventeen '*putative* GIs' in the *first phase.* After refinement, these islands are broken into forty-six statistically significant genomic segments some of which have substantial overlap with the GIs detected by earlier authors. These segments include all of the above mentioned eight GIs detected by the first method of Hsiao et al. (31) that are parts of the *integron island.* They also include a putative integrase as well as some transposase missed by their analysis. The remaining two disjoint GIs, detected by their analysis, are also included in the segments detected in our analysis. In the newly detected genomic segments, the major gene contents are different pathogenic or pathogenicity related genes like hemagglutinin or protease, IcmF-related protein, putative hemolysin, methyl accepting chemotaxis proteins and integrase.

One 3D plot generated from the *first phase* of our algorithm and one representative 1D plot generated from the *refinement phase* of the algorithm applied to chromosome-II of *Vibrio cholerae* are

shown here in Fig.3C. The 3D and the representative 1D plots corresponding to the '*putative* GI' that stretches from 298000 to 445500 bp are presented in the upper and the lower panels respectively in Fig.3C. The above '*putative* GI', detected as *Integron Island* in previous studies (14,28,32-35), is fragmented into two genomic segments after we ran the *refinement phase* of our algorithm. These are located at 301000-416000 bp and 417000-437000 bp. The first segment contains the major genes of *Integron Island,* and the second one contains mainly some transposase and putative acetyltransferases. The genes excluded from the above '*putative* GI' after the *refinement phase* are mainly transcriptional regulator, anaerobic ribonucleoside triphosphate reductase, aspartate aminotransferase and some hypothetical genes.

Fig.3C



***Vibrio vulnificus* CMCP6 (Chromosome-I)**

Vibrio vulnificus, an opportunistic pathogen, experiences a dramatic environmental change during its infection process. *V. vulnificus* is an estuarine bacterium that preferentially affects individuals, who are heavy drinkers of alcohol, who are patients with underlying hepatic disease, and who have

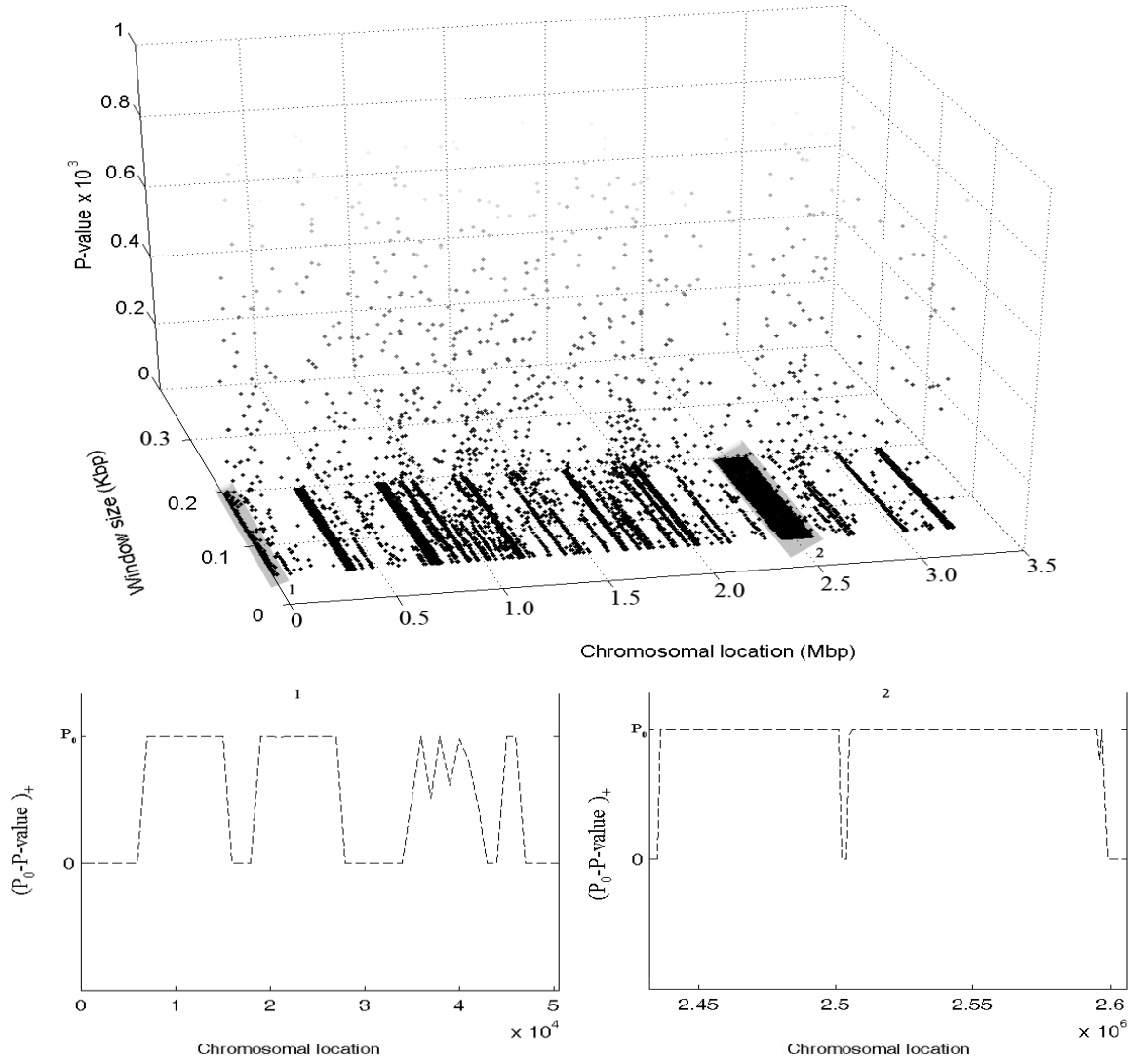
immuno-compromised conditions. The putative virulence factors of *V. vulnificus* reported so far include a hemolysin, a protease, phospholipase A2, siderophores, capsular polysaccharides and a transmembrane signal-transducing transcription activator ToxRS system. *V. vulnificus*, while infecting the susceptible host, passes through gastric acidity, experiences an abrupt pH increase in the duodenum, receives bile secretion, invades into intestinal mucosa, and eventually enters the bloodstream, where the pathogen multiplies. During this complicated infection process, *V. vulnificus* should be able to sense changes in the environmental parameters in the host milieu (36). Earlier studies by Mantri and Williams (11) reported two GIs and Zhang and Zhang (10) reported three GIs in chromosome-I of *V. vulificus* CMCP6.

In chromosome-I of *V. vulnificus* CMCP6, *Design-Island* detected fifty-three '*putative* islands' in the *first phase,* and after refinement, these islands were fragmented into one hundred and fifty-two statistically significant genomic segments. These segments include all the GIs detected in earlier studies. The major gene contents of the segments are phage proteins, phage integrase, hemolysin, antitoxin of toxin-antitoxin stability systems, MSHA biogenesis proteins, MSHA pilin proteins, plasmid stabilization systems, components of type II secretory pathway, flagellin and some transporters. In the newly detected segments, the major genes present are MSHA biogenesis protein, MSHA pilin protein, polysaccharide export-related protein, Flp pilus assembly protein, flagellin, flagellar biosynthesis protein, chemotaxis protein, different components of Type II secretory pathway, different subunits of ATP synthase, multiple antibiotic transporters and some other ABC transporters. All these were missed by earlier authors.

One 3D plot generated from the *first phase* of our algorithm and two representative 1D plots generated from the *refinement phase* of the algorithm applied to chromosome-I of *Vibrio vulnificus* are shown here in Fig.3D. The 3D and the two 1D plots corresponding to the '*putative* GIs' that stretch from 1 to 50500 bp and from 2432000 to 2606000 bp respectively are shown in Fig.3D. The first '*putative* GI' enclosed in a gray block is fragmented into four segments, namely 6001-16001 bp, 18001-28001 bp, 36001-43001 bp and 44001-47001 bp. The main gene contents of these segments are RTX toxin, DNA uptake protein, ABC-type transport systems and methyl-accepting chemotaxis protein. After the *refinement phase,* the genes excluded from the above '*putative* GI' are mainly transglutaminase-like enzyme, putative beta-lactamase class A, putative GTPase and TRAP-type C4-dicarboxylate transport systems. The second '*putative* GI', also detected in earlier studies (10,14), is fragmented into two segments namely 2437000-2502000 bp and 2504000-2599000 bp. These fragments contain *Super-integron* integrase, antitoxin of toxin-antitoxin stability system, plasmid stabilization system protein, TPR repeat containing protein, transposase and inactivated derivatives, multidrug efflux protein, cold

shock domain family protein and some acetyltransferases. The genes excluded from the above '*putative* GI' after passing through the *refinement phase* are histidine utilization repressors and some hypothetical genes.

Fig.3D



***Vibrio vulnificus* CMCP6 (Chromosome-II)**

In chromosome-II of *V. vulnificus* CMCP6, *Design-Island* detected thirty-three '*putative* islands' in the *first phase,* and after refinement, these islands are fragmented into one hundred and nine statistically significant genomic segments. These segments mainly include transposase, integrase, toxin secretion ATP binding protein, type II secretory proteins, agglutination protein, protein of avirulence locus, multidrug resistance proteins, flagellin, methyl-accepting chemotaxis protein, bacterial surface protein containing Ig-like domain, chaperonin GroEL, Flp pilus assembly protein, biopolymer transport

protein, ABC-type Fe3+ transport system, ABC-type phosphate transport system and some other ABC transporters.

Fig.3E

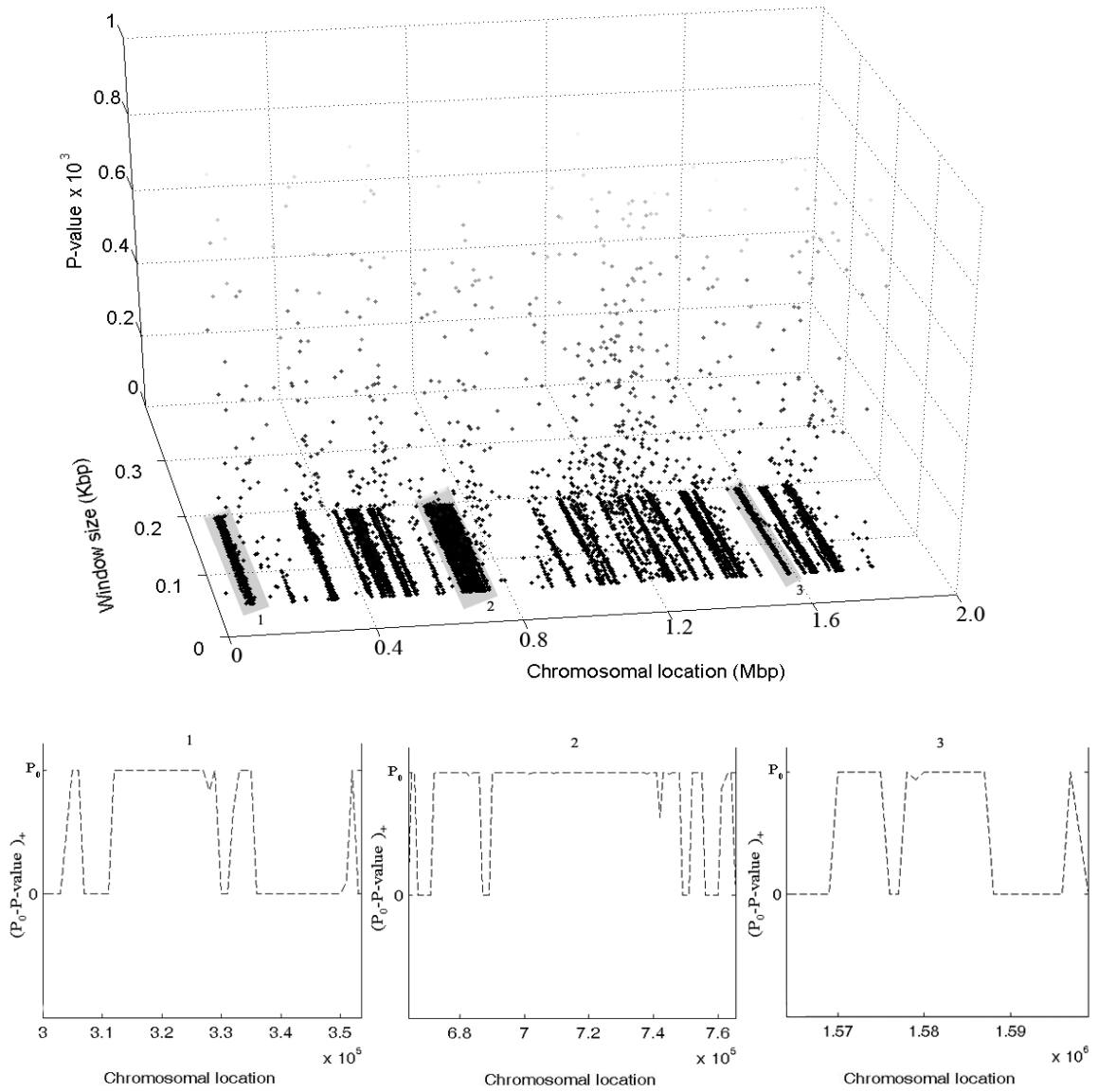


Fig.3: In the upper panel, 3D plots of the P-values for a window with variable size that slides across (i) the chromosome of *Salmonella typhi* CT18 from 1 bp, i.e., the start to 2.5 Mbp (Fig.3A(i)), (ii) the chromosome of *Salmonella typhi* CT18 from 2.5Mbp to 4.8 Mbp, i.e., end (Fig.3A(ii)). (iii) the chromosome-I of *Vibrio cholerae* from 1 bp, i.e., the start, to 1.4 Mbp (Fig.3B(i)), (iv) the chromosome-I of *Vibrio cholerae* from 1.4 to 3.1 Mbp, i.e., the end (Fig.3B(ii)), (v) the chromosome-II of *Vibrio cholerae* (Fig.3C), (vi) the chromosome-I of *Vibrio vulnificus* (Fig.3D) and (vii) the chromosome-II of *Vibrio vulnificus* (Fig.3E). The P-value at a specific location and for a specific size of the window is plotted using a gray scale that changes gradually from black to white with black corresponding to the extreme P-value = 0 and white corresponding to the other extreme P-value = 1. The white dots corresponding to higher P-values are almost invisible in the white background while dark dots corresponding to low P-values are prominently visible marking the GIs in the chromosome. Lower panel in each figure gives some representative 1D plots generated from the *refinement phase* for some of the '*putative* GIs' (enclosed in gray blocks and labeled as 1,2,… in the 3D plots) detected in the *first phase* of *Design-Island*. The quantity $(P_0\text{-P-value})_+$ for the region of a GI detected in the *first phase* is plotted. Here, for P-value > P_0, $(P_0\text{-P-value})_+ = 0$, and for P-value < P_0, $(P_0\text{-P-value})_+ = (P_0\text{-P-value})$.

One 3D plot generated from the *first phase* of our algorithm and three representative 1D plots generated from the *refinement phase* of the algorithm applied to chromosome-II of *Vibrio vulnificus* are shown here in Fig.3E. The 3D plot and three 1D plots are shown in Fig.3E in the upper and the lower panels respectively. 1D plots corresponding to the three '*putative* GIs' marked in the gray blocks stretching from 300000 to 353500 bp, from 664000 to 765500 bp and 1564000 to 1599000 bp are shown in the lower panel. Among the above mentioned '*putative* GIs', the first one is fragmented into four segments (303000-307000 bp, 311000-330000 bp, 331000-336000 bp and 351000-353000 bp), and the genes present in these segments are mainly outer membrane receptor protein, methyl-accepting chemotaxis protein, chaperonin and alpha and beta subunits of NAD/NADP transhydrogenase. The main genes excluded from the '*putative* GI' after running the *refinement phase* are allophanate hydrolase subunit 1, transcriptional regulators, signal transduction histidine kinase, response regulator and cobyric acid synthase. The second '*putative* GI' mentioned above is fragmented into five segments (664000-667000 bp, 672000-687000 bp, 689000-749000 bp, 751000-756000 bp and 760000-765000 bp), and the main gene contents of these segments are integrases, transposase, methyl-accepting chemotaxis proteins and DNA repair proteins. After the *refinement phase,* the genes excluded from this '*putative* GI' are mainly chromosome segregation ATPase, nucleoside diphosphate kinase regulator and some hypothetical genes. The third '*putative* GI' mentioned above is fragmented into three segments (1574000-1576000 bp, 1577000-1588000 bp and 1596000-1598000 bp). The genes present in these segments are mainly membrane-associated phospholipid phosphatase, cell wall-associated hydrolase and some hypothetical genes. The genes excluded from this '*putative* GI' after the *refinement phase* are mainly transcriptional regulators, phosphomethylpyrimidine kinase, Valyl-tRNA synthetase and some hypothetical genes.

Co-ordinates of the detected segments and the percentages of the genome covered by (i) the '*putative* islands' identified in the *first phase* of the algorithm, (ii) genomic segments detected after the *refinement phase* are given in Supplementary Table-1. Further, in the last column of Supplementary Table-1, the genes included in our identified segments along with the percentage of those genes in the entire collection of genes present in the annotated chromosome are presented. The percentages of HTGs identified by different methods are reported in Supplementary Table-2.

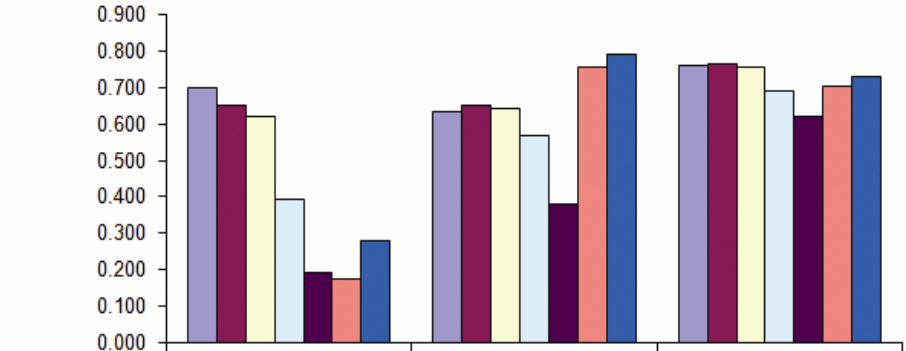
Ribosomal proteins, a group of highly expressed genes, deviate compositionally from the genomic background. However, those genes are widely believed to have limited mobility, and they do not tend to transfer across species (37). For this reason, ribosomal proteins and the stretches with heavy

loads of ribosomal proteins are excluded from the segments obtained in the *refinement phase* of the algorithm following a similar approach taken by some earlier authors (20, 38). In the following section, where compare our method with the other methods available in the literature, we have excluded ribosomal proteins from the prediction made by a method in order to maintain proper comparability of different methods.

Performance comparison with other methods

For performance assessment of *Design-Island*, a dataset of 1560 manually curated putative horizontally transferred genes in *S. typhi* CT18, generated by Vernikos et al. (21) were used. *S. typhi* CT18 is a well-studied prokaryote in terms of its HGT events. Vernikos et al. (21) selected *S. typhimurium* LT2 as a sister lineage to *S. typhi,* and the genome of *E. coli* K12 was chosen as an outgroup of *S. typhi* and *S. typhimurium*. Their main idea was that the genes present in all the three genomes form a set of core genes, while the rest of the genes represent either species or strain specific genes, and thus they may be considered as putative HTGs. The sensitivity (SN), the specificity (SP) and the accuracy (AC) of *Design-Island* have been compared with those of six other methods available in the literature, namely W8 (20), IslandPath-GC (based on GC composition), IslandPath-DB (based on di-nucleotide bias) (31), Islander (11), HGT-DB (38) and IVOM (21). The results are summarized in Fig.4 accompanied with its data table. The sensitivity of *Design-Island* is the highest (70%) among the methods considered for comparison, the second in the list being IVOM (64.9%). Regarding the accuracy, *Design-Island* is in the second position with an accuracy = 76.2%, and it is slightly below IVOM that has accuracy = 76.4%. The third method in the list is W8 with accuracy = 75.4%. The specificity of *Design-Island* (63.5%) is comparable with that of IVOM (65.3%) and W8 (64.3%).

Fig.4



	SN	SP	AC
Design-Island	0.700	0.635	0.762
IVOM	0.649	0.653	0.764
W8	0.621	0.643	0.754
IslandPath-GC	0.392	0.569	0.692
IslandPath-DB	0.193	0.381	0.619
Islander	0.176	0.755	0.701
HGT-DB	0.279	0.791	0.730

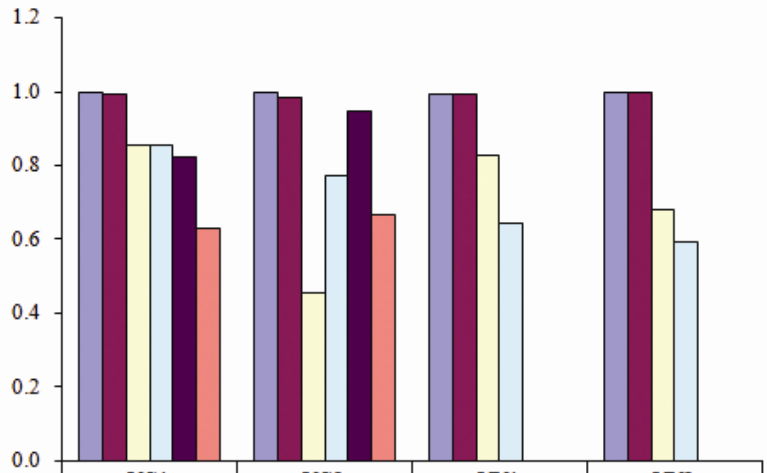
Fig.4: The bar diagram and the corresponding data table for the sensitivity (SN), the specificity (SP) and the accuracy (AC) of *Design-Island* along with the other methods using a manually curated data set of 1560 putative horizontally transferred genes of *Salmonella typhi* CT18 generated by Vernikos et al. (21).

However, the specificity of *Design-Island* is low when compared with that of HGT-DB (78.9%) and Islander (75.5%). Note that design-Island, IVOM and W8 predicted a much larger number of putative horizontally transferred genes compared to the number of such horizontally transferred genes predicted by HGT-DB and Islander, and this largely explains the behavior of different methods in terms of their accuracies as pointed out by earlier authors (21).

For a second round of comparative analysis and performance assessment of different methods, datasets of compositionally atypical genes, which are predicted as part of the GIs by different procedures, were generated for each of the two chromosomes of *V. cholerae* and *V. vulnificus* by the following method of '*majority votes'*. The genes, which were predicted as part of the GIs by more than half of the methods (i.e., some or all of IVOM, W8, HGT-DB, IslandPtah-DINUC, IslandPtah-DIMOB and *Design-Island*) applied to any of these genomes, were chosen as a dataset of '*TYPE-I genes*'. The genes, which were predicted as part of the detected GIs by exactly half of these methods applied to a genome, were excluded from the analysis. Finally, the genes, which were predicted as part of the detected GIs by less than half of the methods, were chosen as '*TYPE-II genes*'. Based on these datasets,

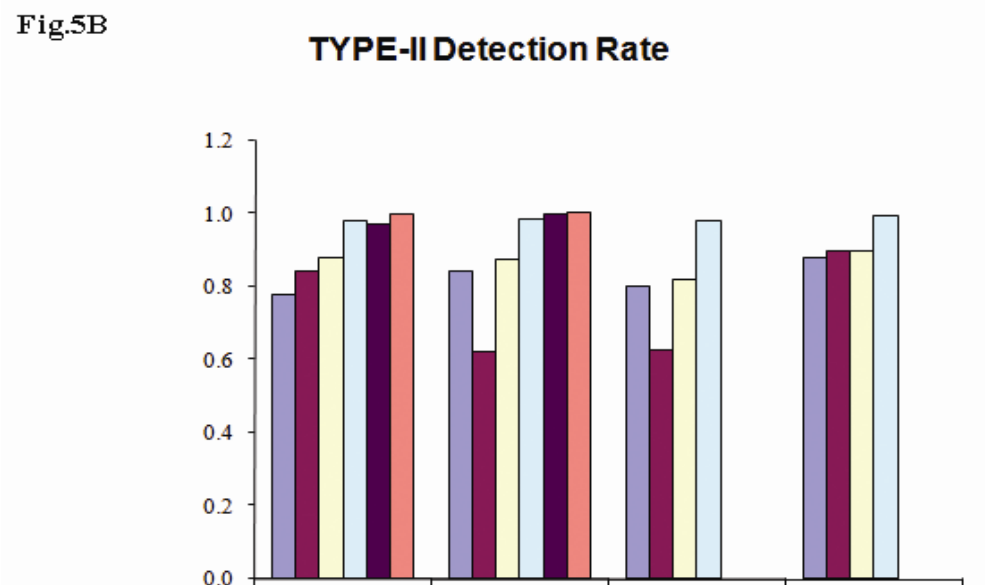
we calculated the '*TYPE-I detection rate'*, and the *'TYPE-II detection rate'* for different methods including ours for a comparative analysis. The results are summarized in Fig.5A and 5B along with corresponding data tables. As it is evident, *Design-Island* has the highest TYPE-I detection rate in each case. In the case of *V. vulnificus* chromosome-II *Design-Island* and IVOM have identical TYPE-I detection rate. In terms of the TYPE-II dtection rate, *Design-Island* performs slightly better than IVOM in the case of *V. cholerae* chromosome-II and *V. vulnificus* chromosome-I, and it performs slightly worse in the case of *V. cholerae* chromosome-I. In the case of *V. vulnificus* chromosome-II, TYPE-II detection rate of *Design-Island* is slightly lower than that of W8 and IVOM. As we have already noted, HGT-DB, Islander and IslandPath have predicted much smaller number of horizontally transferred genes compared to the number of such genes predicted by *Design-Island*, IVOM and W8. As a consequence, HGT-DB and IslandPath tend to outperform *Design-Island*, IVOM and W8 in terms of their TYPE-II detection rates.

Fig.5A



	VC1	VC2	VV1	VV2
Design-Island	1.000	1.000	0.995	1.000
IVOM	0.994	0.984	0.993	1.000
W8	0.855	0.453	0.826	0.682
HGT DB	0.855	0.771	0.642	0.594
Islandpath-DINUC	0.824	0.948		
Islandpath-DIMOB	0.630	0.667		

Fig.5B



	VC1	VC2	VV1	VV2
Design-Island	0.776	0.840	0.801	0.878
IVOM	0.843	0.621	0.626	0.898
W8	0.876	0.874	0.816	0.895
HGT DB	0.978	0.983	0.980	0.991
Islandpath-DINUC	0.969	0.999		
Islandpath-DIMOB	0.999	1.000		

Fig.5: Bar diagrams and corresponding data tables for the sensitivity (SN) (Fig.5A), the specificity (SP) (Fig.5B) and the accuracy (AC) (Fig.5C) of *Design-Island* along with the other methods using a data set of '*putative* HTGs' and '*putative* non-HTGs' of chromosomes-I and -II of *Vibrio cholerae* (VC1 and VC2) and *Vibrio vulnificus* CMCP6 (VV1 and VV2) generated by "*majority votes*". The cells in data tables are left blank in the cases where the data are not available.

CONCLUDING REMARKS

The method proposed and discussed in this paper is an unsupervised method in the sense that it does not require any training dataset to begin with. The method uses Monte-Carlo statistical tests that are implemented using randomly sampled segments, and normal critical values are used for the test statistic. In many of the earlier methods, no statistical test has been performed, and in some cases, where statistical tests were carried out, the determination of the critical values and the P-values were not adequately justified due to lack of rigorous statistical distribution theory. In *Design-Island,* such difficulties are effectively overcome by using Monte-Carlo statistical tests based on randomly selected segments from a chromosome.

We have carried out an elaborate comparative analysis involving different bacterial genomes, and it demonstrates that the performance of *Design-Island* is often comparable to many other well

known methods in terms of their sensitivity, specificity and accuracy. Further, in some cases, *Design-Island* outperforms many of those competing methods.

Our method *Design-Island* has detected several new segments of bacterial genomes as parts of some GIs that were missed by earlier methods. The analysis of the gene contents of these detected segments confirmed their horizontal acquirement. For example, in the case of *V. cholerae* chromosome-I, most of the previous methods were not able to detect the GI containing Cholerae toxin (Ctx), whereas it is experimentally verified that this region is CtxΦ phage mediated GI. *Design-Island* has detected that GI with high significance level. In chromosome-II of *V. cholerae, Design-Island* has detected some pathogenicity related genes and phage genes in some segments of the genome that were missed by earlier methods. Also in the case of *S. typhi* CT18, *Design-Island* has predicted some pathogenic or pathogenicity related genes like putative virulence proteins, putative phage proteins, integrase as horizontally acquired materials that were not detected by earlier methods.

ACKNOWLEDGEMENT

The research of Raghunath Chatterjee was supported by a fellowship from the Council of Scientific and Industrial Research, Government of India. The research of Probal Chaudhuri was supported by grants from the Council of Scientific and Industrial Research, Government of India and the Department of Biotechnology, Government of India. The authors are thankful to anonymous reviewers for their careful reading the earlier version of the paper and several useful comments and suggestions. The authors are also grateful to George Vernikos of Sanger Institute for several helpful discussions over email and providing their data on *Salmonella typhi* CT18 and computer programs related to their IVOM algorithm.

REFERENCES

1. Haker, J. and Kaper, J.B. (1999) In Kaper, J. B. and Haker, J. (eds.), . Am. Soc. Microbiol., Washington, DC, pp. 1-11.
2. Hacker, J., Blum-Oehler, G., Muhldorfer, I. and Tschape, H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol*, **23**, 1089-1097.
3. Groisman, E.A. and Ochman, H. (1996) Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*, **87**, 791-794.
4. Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R. and Goebel, W. (1990) Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates. *Microb Pathog*, **8**, 213-225.

5. Weinstock, G.M. (2000) Genomics and bacterial pathogenesis. *Emerg Infect Dis*, **6**, 496-504.
6. Hacker, J. and Kaper, J.B. (2000) Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol*, **54**, 641-679.
7. Il'ina, T.S. and Romanova Iu, M. (2002) Bacterial genomic islands: organization, function, and role in evolution. *Mol Biol (Mosk).* **36**, 228-239.
8. Dobrindt, U., Hochhut, B., Hentschel, U. and Hacker, J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol*, **2**, 414-424.
9. Ou, H.Y., Chen, L.L., Lonnen, J., Chaudhuri, R.R., Thani, A.B., Smith, R., Garton, N.J., Hinton, J., Pallen, M., Barer, M.R. *et al.* (2006) A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. *Nucleic Acids Res.*, **34**, e3.
10. Zhang, R. and Zhang, C.T. (2004) A systematic method to identify genomic islands and its applications in analyzing the genomes of Corynebacterium glutamicum and Vibrio vulnificus CMCP6 chromosome I. *Bioinformatics.*, **20**, 612-622.
11. Mantri, Y. and Williams, K.P. (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res.*, **32**, D55-58.
12. Merkl, R. (2004) SIGI: score-based identification of genomic islands. *BMC Bioinformatics.*, **5**, 22.
13. Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W.F., Surovcik, K., Meinicke, P. and Merkl, R. (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics.*, **7**, 142.
14. Nag, S., Chatterjee, R., Chaudhuri, K. and Chaudhuri, P. (2006) Unsupervised statistical identification of genomic islands using oligonucleotide distributions with application to Vibrio genomes. *Sadhana.*, **31**, 105-115.
15. Tu, Q. and Ding, D. (2003) Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. *FEMS Microbiol Lett.*, **221**, 269-275.
16. Weinel, C., Nelson, K.E. and Tummler, B. (2002) Global features of the Pseudomonas putida KT2440 genome sequence. *Environ Microbiol.*, **4**, 809-818.
17. Waterhouse, J.C., Swan, D.C. and Russell, R.R. (2007) Comparative genome hybridization of Streptococcus mutans strains. *Oral Microbiol Immunol.*, **22**, 103-110.
18. Yoon, S.H., Hur, C.G., Kang, H.Y., Kim, Y.H., Oh, T.K. and Kim, J.F. (2005) A computational approach for identifying pathogenicity islands in prokaryotic genomes. *BMC Bioinformatics.*, **6**, 184.
19. Tsirigos, A. and Rigoutsos, I. (2005) A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res.*, **33**, 922-933.
20. Tsirigos, A. and Rigoutsos, I. (2005) A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. *Nucleic Acids Res.*, **33**, 3699-3707.
21. Vernikos, G.S. and Parkhill, J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. *Bioinformatics.*, **22**, 2196-2203.
22. Efron, B. (1979) *Bootstrap methods: another look at the jackknife.*
23. Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, London.

24. Randles, R.H., and Wolfe, D. A. (1979) *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York.
25. Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bulletin.*, **1**, 80-83.
26. Mann, H.B. and Whitney, D.R. (1947) On a test of whether one of 2 random variables is stochastically larger than the other. *Annals of Mathematical Statistics.*, **18**, 50-60.
27. Parkhill, J., Dougan, G., James, K.D., Thomson, N.R., Pickard, D., Wain, J., Churcher, C., Mungall, K.L., Bentley, S.D., Holden, M.T. *et al.* (2001) Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. *Nature.*, **413**, 848-852.
28. Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Umayam, L. *et al.* (2000) DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. *Nature.*, **406**, 477-483.
29. Kovach, M.E., Shaffer, M.D. and Peterson, K.M. (1996) A putative integrase gene defines the distal end of a large cluster of ToxR-regulated colonization genes in Vibrio cholerae. *Microbiology.*, **142 (Pt 8)**, 2165-2174.
30. Karaolis, D.K., Johnson, J.A., Bailey, C.C., Boedeker, E.C., Kaper, J.B. and Reeves, P.R. (1998) A Vibrio cholerae pathogenicity island associated with epidemic and pandemic strains. *Proc Natl Acad Sci U S A.*, **95**, 3134-3139.
31. Hsiao, W.W., Ung, K., Aeschliman, D., Bryan, J., Finlay, B.B. and Brinkman, F.S. (2005) Evidence of a Large Novel Gene Pool Associated with Prokaryotic Genomic Islands. *PLoS Genet.*, **1**, e62.
32. Rowe-Magnus, D.A., Guerout, A.M. and Mazel, D. (1999) Super-integrons. *Res Microbiol.*, **150**, 641-651.
33. Hall, R.M., Brookes, D.E. and Stokes, H.W. (1991) Site-specific insertion of genes into integrons: role of the 59-base element and determination of the recombination cross-over point. *Mol Microbiol.*, **5**, 1941-1959.
34. Mazel, D., Dychinco, B., Webb, V.A. and Davies, J. (1998) A distinctive class of integron in the Vibrio cholerae genome. *Science.*, **280**, 605-608.
35. DiRita, V.J. (2000) Genomics happens. *Science.*, **289**, 1488-1489.
36. Kim, Y.R., Lee, S.E., Kim, C.M., Kim, S.Y., Shin, E.K., Shin, D.H., Chung, S.S., Choy, H.E., Progulske-Fox, A., Hillman, J.D. *et al.* (2003) Characterization and pathogenic significance of Vibrio vulnificus antigens preferentially expressed in septicemic patients. *Infect Immun.*, **71**, 5461-5471.
37. Jain, R., Rivera, M.C. and Lake, J.A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A.*, **96**, 3801-3806.
38. Garcia-Vallve, S., Guzman, E., Montero, M.A. and Romeu, A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.*, **31**, 187-189.