Procedia Computer Science
Volume 51, 2015, Pages 670–679
ICCS 2015 International Conference On Computational Science

Optimization and Practical Use of Composition Based Approaches Towards Identification and Collection of Genomic Islands and Their Ontology in Prokaryotes

Rian Pierneef, Oliver Bezuidt and Oleg N. Reva
The University of Pretoria, Department Biochemistry, Bioinformatics and Computational Biology Unit, Pretoria 0002, South Africa. repierneef@live.com; bezuidt@gmail.com; oleg.reva@up.ac.za

Abstract

Motivation: Horizontally transferred genomic islands (islands, GIs) have been referred to as important factors which contribute towards bacterial evolution in general and particularly towards the emergences of pathogens and outbreak instances. The development of tools for identification of such elements and retracing their distribution will help to understand how such cases arise. Sequence composition has been used to identify GIs, infer their phylogeny; and determine their relative time of insertion. Collection of metadata on known GIs will enhance insight into horizontal gene transfer ontology and flow.
Results: This paper introduces the merger of SeqWord Genomic Islands Sniffer (SWGIS), which utilizes composition based approaches for identification of GIs in bacterial genomic sequences, and the Predicted Genomic Islands (Pre_GI) database, which houses 26,744 islands found in 2,407 bacterial plasmids and chromosomes. SWGIS is a standalone program that detects GIs using a set of optimized parametric measures with estimates of acceptable false positive and false negative rates. Pre_GI is a novel repository that includes island ontology and flux. This study furthermore illustrates the need for parametric optimization towards the prediction of GIs to minimize false negative and false positive predictions. In addition Pre_GI emphasizes the practicality of the compounded knowledge that the database affords in detection and visualization of ontological links between GIs.
Availability: SWGIS is freely available on the web at http://www.bi.up.ac.za/SeqWord/sniffer/index.html, and Pre_GI is freely accessible at http://pregi.bi.up.ac.za/index.php.

Keywords: genomic island, horizontal gene transfer, database, ontology

1 Introduction

Recurrent outbreaks of pathogens that possess new virulence factors and broad range antibiotic resistance gene cassettes reflect the importance of horizontal gene transfer (HGT) in the evolution of pathogenic bacteria (Smith *et al.*, 2000; Fernández-Gómez *et al.*, 2012). In many cases, the evolution of pathogens is mediated by mobile genetic elements, which can easily be interchanged between bacterial taxa inhabiting the same or different environments (Kelly *et al.*, 2009). Outbreaks of the suddenly emerged pathogens of unclear aetiology are characterized by an increased virulence and tolerance to many antibiotics. As a result of the latter, outbreaks of gastrointestinal and nosocomial infections take a heavy death toll (Potron *et al.*, 2011; Brzuszkiewicz *et al.*, 2011). The two major methods for genomic island (GI) identification use sequence similarity (mainly BLAST) and DNA/codon composition approaches. However, either approach has its own benefits and limitations. In this study we show that the composition based approaches may produce reliable predictions when optimal parameters are introduced. The aspect of base composition similarity among closely related species arises from their common origin (Sueoka, 1962), i. e. from the same lineage of plasmids or phages, or from the same former host organism. Similarity is also influenced by the species specific mutational pressure that acts upon the whole chromosome to maintain composition

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.183

stability. Comparative analysis between lineages have uncovered that genes acquired by HGT display atypical oligonucleotide usage (OU) biases, which are distinct from those of their host genomes (Hacker & Carniel, 2001; van Passel *et al.*, 2005, 2011). The principles mentioned above were brought into practical use by Karlin (1998). It was proved that frequencies of oligonucleotide as short as dinucleotides might possess a genomic signature. Thereafter, distribution of longer words was shown to be even better phylogenetic descriptors (Reva & Tümmler, 2004; Deschavanne *et al.*, 1999).

Current databases applicable in bacteriological research include IslandViewer (Langille & Brinkman, 2009), PAIDB (Yoon *et al.*, 2007) and ACLAME (Lima-Mendez *et al.*, 2010); all of which are constructed to facilitate specific and non-overlapping research. IslandViewer employs three methods of GI prediction to identify horizontally acquired genomic fragments of all types in sequenced bacterial genomes. PAIDB catalogues verified pathogenicity islands (PAIs) and ACLAME reconstructs reticulation events in bacterial genomes. Whilst all of the above mentioned resources have numerous applications, none of them allow identification of GI movement and ontology, including but not limited to PAI.

In this work we present the SeqWord Genomic Islands Sniffer (SWGIS) program developed to identify GIs in bacterial genome by the composition based approach and distinguishing them from other loci with alternative OU usage and the Predicted Genomic Islands (Pre_GI) database that house predicted GIs and their ontology. The idea to identify GIs by alterations in frequencies of oligonucleotide is not new. A number of computational tools based on this approaches have been proposed recently (Mrázek & Karlin, 1999; Pride & Blaser, 2002; Abe *et al.*, 2003; Dufraigne *et al.*, 2005; Chatterjee *et al.*, 2008; Ménigaud *et al.*, 2012). SWGIS uses a set of combinatorial parametric measures to improve sensitivity and specificity of the composition based methods and in this way it outperforms many other programs. Several sets of the SWGIS combinatorial parametric measures were revised to improve on rates of prediction of true GIs. SWGIS was then compared to IslandViewer tools to measure their predictive values on known and curated GI predictions.

Pre_GI serves as a reservoir for island ontology, similarity and flux to further island prediction and reason by affording the opportunity of compounded knowledge in a friendly and accessible format.

2 Methods

All programming for SWGIS was implemented in Python 2.5. Algorithms of OU pattern calculation and comparison were described in detail previously (Reva & Tümmler, 2004 and 2005; Ganesan *et al.*, 2008; Bezuidt *et al.*, 2011). Several sets of the SWGIS combinatorial parametric measures were revised to improve on rates of true GI predictions. SWGIS was then compared to IslandViewer (Langille & Brinkman, 2009) to measure their predictive values on known and curated GI predictions. SWGIS and LingvoCom utilities are available for download from www.bi.up.ac.za/SeqWord/sniffer/. Sequences of bacterial chromosomes and plasmids were obtained from GenBank FTP server. Optimization of program run parameters was performed by the factorial analysis technique (St-Pierre & Weiss, 2009).

SWGIS was used to identify GI housed in Pre_GI. GI compositional similarity was measured by OU pattern similarity (Reva & Tümmler, 2004 and 2005) and sequence similarity hits identified through BLAST. Clustering of GI were produced by the Markov Clustering Algorithm (MCL) (Enright *et al.*, 2002) with OU pattern hits serving as a measure of relational scores. Non-overlapping cluster representatives were identified as the nodes with the highest number of compositional similarity links. Flux determination was based on the assumption of amelioration changing in the GI nucleotide landscape from time of insertion to equate with that of the host in which it resides, yet for an extended period after insertion a GI may be traced back to its origin by preserving compositional homomorphism with the donor (Lawrence & Ochman, 1997). This approach was used in Pre_GI to predict donor-recipient relationships by comparing OU pattern similarity values calculated for homologous GIs hosted by different organisms. Significant OU pattern differences of homologous GIs to that of hosts would indicate possible donor-recipient relations. A high OU pattern similarity of both homologous GIs to one host with a lower OU pattern similarity to another one indicates likelihood that the latter host is the recipient of a given GI from the former one. Pre_GI was developed to ensure an interactive communication through the Web-based user interface and a regular updating.

3 SWGIS performance

The basic principle behind the SWGIS algorithm is to superimpose the values of several statistical parameters (Reva & Tümmler, 2004 and 2005) calculated for a sliding window that allows identification of loci with an alternative OU pattern and distinguishing between the different categories of these genomic fragments. Particularly, GIs were identified by an alternative oligonucleotide usage (increased D parameter) with lower internally normalized OU variance (RV) and an increase in globally normalized OU variance (GRV). The latter two parameters were combined into a parameter V = GRV/RV. The value of V stays closer to 1 in the core genome and significantly increases in loci covered by a GI. Pattern skew (PS) comparison is used to filter out *rrn* operons characterized with extreme values of PS. These parameters are calculated in SWGIS for genomic loci by the use of a sliding window approach, whereby values of genomic fragments of 8 kbp with a 2 kbp step are compared to the tetranucleotide usage pattern calculated for the whole genome. If the program recognizes a statistically reliable increase of the local distance D accompanied by a significant increase of V, the window shifts several steps back and repeats the analysis, this time with the steps of 0.2 kbp to identify exact borders of the foreign inserts. The thresholds of parameter deviations from the average values to be considered as significant pattern alterations may be specified by users. This paper is to instruct the users about setting the parameters in a way that will help them to achieve acceptable false negative and false positive ratios.

SWGIS was developed for identification of GIs in multiple genomes by a single run. It takes as input complete bacterial genomes in GenBank (preferable) or FASTA format. Several output files are created for each genome depending on the user's selected choice. One is a standard text file with extension OUT, which contains a list of identified GIs with their coordinates, OU statistical values calculated for each GI and annotation of genes within the borders of the GI, if it is available. The others are the FASTA file with DNA sequences of identified GI; GenBank files created for each GI to accommodate the annotation data; and lastly, users may instruct the program to create a graphical SVG file comprising a genomic atlas with indicated positions of predicted GIs. An HTML help file on how to use SWGIS is available at www.bi.up.ac.za/SeqWord/sniffer/. Also, from the same page the users may download the latest version of the program.

3.1 SWGIS parametric optimization

An empirical analysis was performed to generate optimal parametric threshold values to be used for D and V. Setting the D and V values below 1.5 resulted in an increased false positive rate, whereas setting these values above 2.0 overlooked many known GIs (preliminary data not shown). The next step was to use the factorial analysis to determine optimal combinations of threshold values for D and V to ensure minimal false positive and false negative rates.

3.1.1 False negative rate calculation. The parametric measures for SWGIS were optimized to attain better predictions through the re-identification of known GIs from PAIDB (Yoon *et al.*, 2007), which were used as training data. The SWGIS optimization and re-identification analysis was carried out on 51 pathogenicity islands (PAIs) possessed by 24 micro-organisms. The latter was conducted in comparison to the IslandViewer programs comprising IslandPick, SIGI-HMM, and IslandPath. From these comparisons the calculations for false negative rates (FNR) were determined. FNR in this instance is defined as the percentage of the known GIs that were overlooked by either of the programs used in the study. SWGIS was run for 4 times with different combinations of D and V: [D:1.5; V:1.5]; [D:2.0; V:2.0]; [D:1.5; V:2.0] and [D:2.0; V:1.5]. Results are shown in Fig. 1. From the comparison of the results attained from all the programs, SWGIS outperformed individual IslandViewer methods even when the most stringent threshold values [D:2.0; V:2.0] were set. Jointly the IslandViewer programs identified 69% of the 51 PAIs, while SWGIS identified 88% with [D:1.5; V:1.5], 78% with [D:2.0; V:1.5], 65% with [D:1.5; V:2.0] and 63% with [D:2.0; V:2.0]. All PAIs predicted by the IslandViewer programs except for 2, which were only predicted by IslandPath, were also predicted by SWGIS [D:1.5; V:1.5]. Four PAIs were not detected by any method.

3.1.2 False positive rate calculation. Diverse native loci of bacterial genomes including *rrn* gene clusters; operons of ribosomal proteins; giant genes; and local tandem repeats are also characterized by alternative OU patterns and resemble horizontally acquired genes (Reva & Tümmler, 2005 and 2008). SWGIS uses superimposition of different OU statistical parameters to distinguish between different types of atypical genomic loci. The comparisons of GIs predicted by SWGIS and the IslandViewer tools were carried out in order to determine the rates of false positives. The estimation of the false positive rate (FPR) of predictions of GIs is problematic as there is no any formal way to prove that a given genomic fragment has not been acquired horizontally. As FPR cannot be estimated straight away, we first calculated the statistics of unconfirmed predictions, i. e. the frequencies of GIs, which were predicted only by one program and not the others. Sets of pre-calculated GIs predicted in 164 bacterial chromosomes were downloaded from the IslandViewer web resource (www.pathogenomics.sfu.ca/islandviewer/download.php) and included in the analysis. SWGIS searched for GIs in the same chromosomes with the run parameters set for [D:1.5; V:1.5]; [D:2.0; V:2.0]; [D:1.5; V:2.0] and [D:2.0; V:1.5]. It was stipulated that a GI is confirmed, if the genomic loci selected by different programs at least partly overlapped. Numbers of predicted GIs and frequencies of unconfirmed GIs for each program are summarized in Fig. 2.

A great deal of unconfirmed GIs predicted by different methods was observed. Many false positives might be expected among these unconfirmed GIs. SWGIS identified more GIs than the other methods with the less stringent parameter [D:1.5; V:1.5], and also resulted in the highest rate of unconfirmed predictions.

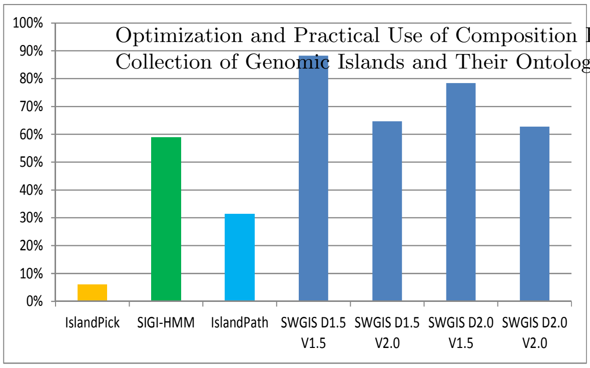


Fig. 1. Re-identification of known PAIs by IslandViewer tools and SWGIS with different threshold parameters.

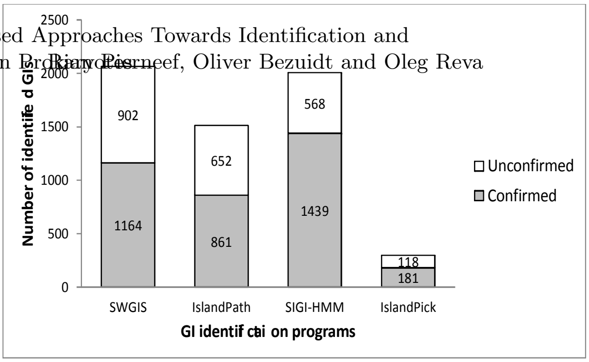


Fig. 2. Frequencies of GIs predicted only by one of the four programs (unconfirmed) and confirmed by the others. SWGIS was used with the most relaxed parameters [D:1.5; V:1.5].

For the assessment of SWGIS's performance and selecting the optimal parametric criterion, we performed an estimate for FPR based on the rate of unconfirmed predictions. First, the ratio of unconfirmed GIs comprising mobile elements associated genes was performed by a key word search through gene annotation. Predicted loci comprising at least one of mobile element associated genes ("integrase", "transposase", "phage" and "IS-element") were termed 'unconfirmed key positives'. Search for the same key words in gene annotations of 1,252 previously identified true positive GIs (Bezuidt *et al.*, 2011) showed that only 56% of them possessed genes associated with mobile elements. From this observation the amount of true positives was roughly estimated as 'Number of unconfirmed key positive GIs'×100/56. Estimated FNR, reduced FNR and FPR calculated for the training set are given in Table 1.

Table 1. Prediction of GIs by SWGIS with different program run parameters and estimated FPR and FNR.

SWGIS	[D:1.5; V:1.5]	[D:1.5; V:2.0]	[D:2.0; V:1.5]	[D:2.0; V:2.0]
Total GIs	2066	928	1571	809
Unconfirmed	902	280	545	188
Key positive	137	44	92	28
Estimated FPR*	657	201	381	138
Reduced FPR†	0.318	0.217	0.243	0.171
FNR	0.118	0.353	0.216	0.373

*Number of false positives was calculated as: "Unconfirmed GIs" – "Unconfirmed key positive GIs"×100/56;
†FPR was calculated as "False positive estimation"/ "Total GIs predicted".

3.2 Optimization of parametric values by factorial experiment

Factorial experiment design was applied to fit a model of two regression equations 1 and 2 to estimate FNR and FPR for given D and V thresholds. Sensitivity and specificity parameters were calculated by equations 3 and 4.

$$FNR = -0.628 + 0.118D + 0.392V \quad (1)$$

$$FPR = 0.752 - 0.121D - 0.173V \quad (2)$$

$$Sensitivity = \frac{1 - FNP - FPR}{1 - FPR} \quad (3)$$

$$Specificity = \frac{1}{1 + FPR} \quad (4)$$

Fig. 3A-C show expected FNR, FPR and FNR+FPR values that are likely to occur when different parametric combinations are in use. Although [D:1.5; V:1.5] resulted in smaller FNR and the highest sensitivity, it however generated an increased FPR and low specificity. And contrary, the setting [D:2.0; V:2.0] confered the highest specificity but decreased sensitivity. Changes in the cumulative FNR+FPR, which depend on D and V, are shown in Fig. 3C. It was observed that an increase in V gradually increased FNR+FPR, while a change in D had no effect as the increase in FNR was compensated by a similar decrease in FPR. Thus, optimization of specificity and sensitivity of GI identification by this approach may be achieved by an adjustment of D and keeping V threshold constant and minimal. It was calculated that the optimal specificity and sensitivity combination is achieved when the parameters are set for [D:1.7; V:1.5]. This setting serves as a default parameter for SWGIS.

3.3 Case study of SWGIS failures and problem resolving strategies

The performance of SWGIS may be improved by further analyzing the patterns of genomes in which it performed poorly. Genomes in Fig. 4 are those in which SWGIS identified too many or too little GIs as compared to the IslandViewer tools.

A

FNR = -0.628 + 0.118D + 0.392V

D \ V	1.5	1.55	1.6	1.65	1.7	1.75	1.8	1.85	1.9	2
1.5	0.14	0.16	0.18	0.2	0.22	0.24	0.25	0.27	0.29	0.33
1.55	0.14	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.34
1.6	0.15	0.17	0.19	0.21	0.23	0.25	0.27	0.29	0.31	0.35
1.65	0.15	0.17	0.19	0.21	0.23	0.25	0.27	0.29	0.31	0.35
1.7	0.16	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.32	0.36
1.75	0.17	0.19	0.21	0.23	0.25	0.26	0.28	0.3	0.32	0.36
1.8	0.17	0.19	0.21	0.23	0.25	0.27	0.29	0.31	0.33	0.37
1.85	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.32	0.34	0.37
1.9	0.18	0.2	0.22	0.24	0.26	0.28	0.3	0.32	0.34	0.38
2	0.2	0.22	0.24	0.25	0.27	0.29	0.31	0.33	0.35	0.39

B

FPR = 0.752 - 0.121D - 0.173V

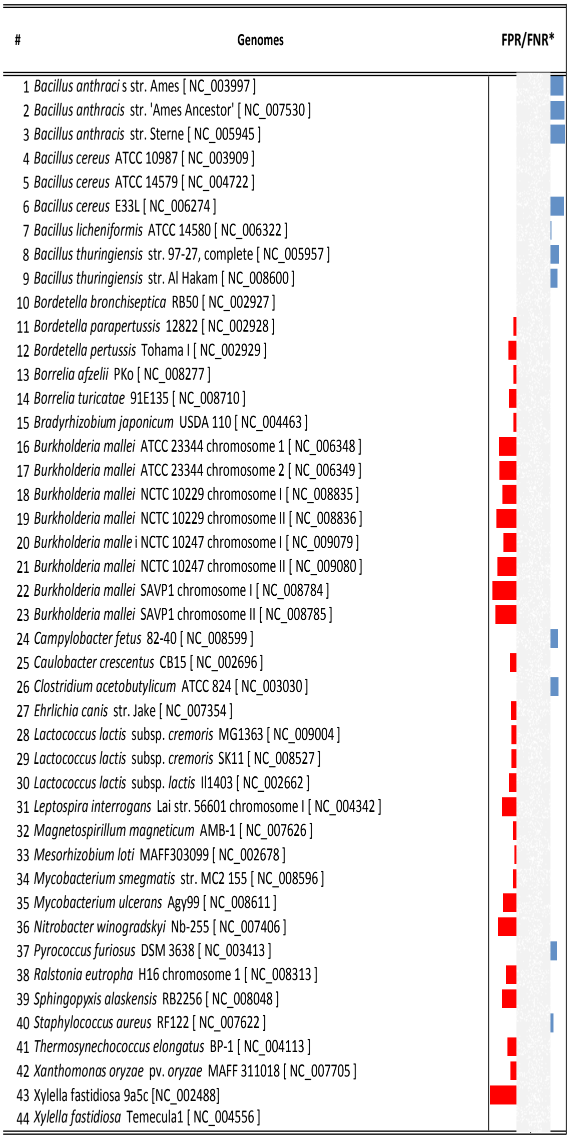
D \ V	1.5	1.55	1.6	1.65	1.7	1.75	1.8	1.85	1.9	2
1.5	0.31	0.3	0.29	0.28	0.28	0.27	0.26	0.25	0.24	0.22
1.55	0.3	0.3	0.29	0.28	0.27	0.26	0.25	0.24	0.24	0.22
1.6	0.3	0.29	0.28	0.27	0.26	0.26	0.25	0.24	0.23	0.21
1.65	0.29	0.28	0.28	0.27	0.26	0.25	0.24	0.23	0.22	0.21
1.7	0.29	0.28	0.27	0.26	0.25	0.24	0.23	0.23	0.22	0.2
1.75	0.28	0.27	0.26	0.25	0.25	0.24	0.23	0.22	0.21	0.19
1.8	0.27	0.27	0.26	0.25	0.24	0.23	0.22	0.21	0.21	0.19
1.85	0.27	0.26	0.25	0.24	0.23	0.23	0.22	0.21	0.2	0.18
1.9	0.26	0.25	0.25	0.24	0.23	0.22	0.21	0.2	0.19	0.18
2	0.25	0.24	0.23	0.22	0.22	0.21	0.2	0.19	0.18	0.16

C

FNR+FPR = 0.124 - 0.003D + 0.219V

D \ V	1.5	1.55	1.6	1.65	1.7	1.75	1.8	1.85	1.9	2
1.5	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.54	0.56
1.55	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.54	0.56
1.6	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.54	0.56
1.65	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.54	0.56
1.7	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56
1.75	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56
1.8	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56
1.85	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56
1.9	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56
2	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.56

Fig. 3. Parts A and B show FNR and FPR calculated for different combinations of D and V, respectively; and their sums are in the part C.



#	Genomes	FPR/FNR*
1	*Bacillus anthraci* s str. Ames [NC_003997]	🟦
2	*Bacillus anthracis* str. 'Ames Ancestor' [NC_007530]	🟦
3	*Bacillus anthracis* str. Sterne [NC_005945]	🟦
4	*Bacillus cereus* ATCC 10987 [NC_003909]	
5	*Bacillus cereus* ATCC 14579 [NC_004722]	
6	*Bacillus cereus* E33L [NC_006274]	🟦
7	*Bacillus licheniformis* ATCC 14580 [NC_006322]	🟦
8	*Bacillus thuringiensis* str. 97-27, complete [NC_005957]	🟦
9	*Bacillus thuringiensis* str. Al Hakam [NC_008600]	🟦
10	*Bordetella bronchiseptica* RB50 [NC_002927]	
11	*Bordetella parapertussis* 12822 [NC_002928]	🟥
12	*Bordetella pertussis* Tohama I [NC_002929]	🟥
13	*Borrelia afzelii* PKo [NC_008277]	🟥
14	*Borrelia turicatae* 91E135 [NC_008710]	🟥
15	*Bradyrhizobium japonicum* USDA 110 [NC_004463]	🟥
16	*Burkholderia mallei* ATCC 23344 chromosome 1 [NC_006348]	🟥
17	*Burkholderia mallei* ATCC 23344 chromosome 2 [NC_006349]	🟥
18	*Burkholderia mallei* NCTC 10229 chromosome I [NC_008835]	🟥
19	*Burkholderia mallei* NCTC 10229 chromosome II [NC_008836]	🟥
20	*Burkholderia malle* i NCTC 10247 chromosome I [NC_009079]	🟥
21	*Burkholderia mallei* NCTC 10247 chromosome II [NC_009080]	🟥
22	*Burkholderia mallei* SAVP1 chromosome I [NC_008784]	🟥
23	*Burkholderia mallei* SAVP1 chromosome II [NC_008785]	🟥
24	*Campylobacter fetus* 82-40 [NC_008599]	🟦
25	*Caulobacter crescentus* CB15 [NC_002696]	🟥
26	*Clostridium acetobutylicum* ATCC 824 [NC_003030]	🟦
27	*Ehrlichia canis* str. Jake [NC_007354]	🟥
28	*Lactococcus lactis* subsp. *cremoris* MG1363 [NC_009004]	🟥
29	*Lactococcus lactis* subsp. *cremoris* SK11 [NC_008527]	🟥
30	*Lactococcus lactis* subsp. *lactis* Il1403 [NC_002662]	🟥
31	*Leptospira interrogans* Lai str. 56601 chromosome I [NC_004342]	🟥
32	*Magnetospirillum magneticum* AMB-1 [NC_007626]	🟥
33	*Mesorhizobium loti* MAFF303099 [NC_002678]	🟥
34	*Mycobacterium smegmatis* str. MC2 155 [NC_008596]	🟥
35	*Mycobacterium ulcerans* Agy99 [NC_008611]	🟥
36	*Nitrobacter winogradskyi* Nb-255 [NC_007406]	🟥
37	*Pyrococcus furiosus* DSM 3638 [NC_003413]	🟦
38	*Ralstonia eutropha* H16 chromosome 1 [NC_008313]	🟥
39	*Sphingopyxis alaskensis* RB2256 [NC_008048]	🟥
40	*Staphylococcus aureus* RF122 [NC_007622]	🟦
41	*Thermosynechococcus elongatus* BP-1 [NC_004113]	🟥
42	*Xanthomonas oryzae* pv. *oryzae* MAFF 311018 [NC_007705]	🟥
43	Xylella fastidiosa 9a5c [NC_002488]	🟥
44	*Xylella fastidiosa* Temecula1 [NC_004556]	

Fig. 4. Genomes in which numbers of GIs predicted by SWGIS were significantly over-ranged regarding to the predictions by other programs that may indicate large FNR (red leftward bars) or large FPR (blue rightward bars) in the column FPR/FNR.

These genomes are graphically marked in the column FPR/FNR by red leftward and blue rightward bars depicting FNR and FPR over-ranges, respectively. FPR/FNR was calculated by the equation 5:

$$FPR / FNR = (N_{SWGIS} - N_{IV}) / N_{av} \quad (5)$$

where N_{SWGIS} is the number of GIs predicted by SWGIS with the parameters [D:1.5; V:1.5]; N_{IV} is the maximum number of GIs predicted by one of the IslandViewer programs and N_{av} is the average number of predicted GIs by the all programs.

3.3.1 False positives. To investigate possible causes of failures, predictions of GIs in several genomes were investigated. These genomes were searched for commonalities, which may explain the excessive number of GIs identified in them. These genomes showed to exhibit a common compositional polymorphism. Large parts of their chromosomes were characterized by alternative OU-bias. Compositional polymorphism of genomes of *Bacillus cereus* and related organisms has been previously reported by Bohlin *et al.* (2012). To avoid this increase in FPR, more stringent parameter settings should be set, preferably by an increase in the D threshold (see Fig. 3C).

3.3.2 False negatives. Composition based methods are customized to identify GIs as regions with atypical OU patterns in a given genome. This approach overlooks GIs, which share OU similarity with host organisms, or ancient acquisitions, which have already been affected by amelioration of their DNA. SWGIS also suffers from such a drawback. SWGIS was able to detect only a few GIs in genomes of *Bordetella*, *Borrelia*, *Burkholderia mallei, Lactococcus* and several others. Predictions in these were found to be inconsistent with those of IslandViewer.

These organisms did not resemble any taxonomic links between themselves. Even in two different strains of the same species prediction of GIs may suffer in one strain but be normal in another. For example, the reason for GI prediction failure in *X. fastidiosa* 9a5c is that this organism has developed a mutator phenotype that eroded its chromosomal OU pattern specificity (Reva & Tümmler, 2004). It was thus impossible for SWGIS to make predictions. In the contrary, there were no problems with GI identification in *X. fastidiosa* Temecula1, which shows a stable chromosomal OU pattern (Fig. 4).

Another example of an overlooked GI is in *Thioalkalimicrobium cyclicum* ALM1, as shown in Fig. 5. There is a large 87,608 bp long viral filamentous hemagglutinin gene with multiple constituent repeats, which can clearly be seen on the genomic atlas (Fig. 5). The reason for discarding this region was that SWGIS considers giant genes with multiple repeats as a separate category of genomic elements with alternative OU patterns (Reva & Tümmler, 2008). This special case of a false negative prediction may be resolved by a visual inspection of the genome maps provided by SWGIS and SeqWord Genome Browser (Ganesan *et al*., 2008; and visit www.bi.up.ac.za/SeqWord/mhhapplet.php). Including these giant genes by default to the SWGIS prediction output would result in too many false positives as these genes are usually resistant to HGT.

4 Pre_gi database

Pre_GI is an interactive database freely accessible at http://pregi.bi.up.ac.za. The database allows users to browse current GIs and/or compare newly predicted GIs against the entries in Pre_GI. An analytical resource for GI ontology and the deconstruction of MGE fluxes was the driving force behind the development of Pre_GI. The availability of all sequence and compositional comparison results allows users the opportunity to inspect ontological links between GIs and the donor-recipient relations to identify fluxes of GIs. The inclusion of host lineages and other metadata, including but not limited to habitat and isolation, aims at highlighting the biological reasoning and logic behind GI presence in the current genome and its movement through bacterial species.

4.1 Pre_GI content and GI browsing

SWGIS was used for a semi-automated search of GIs in multiple GenBank files of bacterial chromosomes and plasmids obtained from the NCBI to populate Pre_GI. SWGIS parameters were set at D = 1.7 and V = 1.5 to ensure an optimal sensitivity/specificity ratio. Currently Pre_GI contains 26,744 GIs identified in 2,407 bacterial chromosomes and plasmids. GIs are accessible by means of various browse functions, i. e. host accession, host strain description, host taxonomy and host information. GIs may furthermore be located by means of gene content and physical location on the host genome.

Each GI is individually represented by means of location on the host genome and all information relating to said GI is clearly displayed or easily accessible by means of hyperlinks. GI metadata includes positional attributes, SWGIS parameter statistics, compositional and sequence similarities. Gene content confirmation of HGT events by keyword search and prediction of the GIs by other methods, i. e. IslandViewer and PAIDB, validate true positive prediction of GIs.

All-against-all composition and similarity comparisons of GIs resulted in 69,176,627 significant OU pattern similarity links (above 75%, see Bezuidt *et al.*, 2011) and 3,692,401 BLASTN hits with an e-value threshold of 1 × 10^{-5}. All-against-all gene similarity search for genes contained in GIs was detected by BLASTP with an e-value cut-off 1 × 10^{-5}. This resulted in 138,590,509 hits stored in the database.

Gene annotations are searchable to identify GIs containing genes with similar annotation. Annotations are linked to the QuickGO browser from EMBL-EBI (http://www.ebi.ac.uk/QuickGO/).

The ability to predict donor-recipient flux of GIs with added information on bacterial host lineage and other host related metadata allows for the logical explanation of evolutionary impact of HGT on strain and species levels to fit to ever changing environment. Detected fluxes are displayed in the corresponding tables in the Pre_GI interface by means of colored arrows in the direction of movement between pairs of organisms sharing homologous GIs.

4.2 Novel island comparison to Pre_GI entries

Current sequencing technologies and the ever increasing speed and affordability of these technologies require a dynamic database to allow for sequence and composition comparison of novel islands identified in newly sequenced bacterial genomes to known GIs. Sequence similarity search for a novel GI stored in FASTA format may be performed by using BLASTN page. High scoring hits are hyperlinked to the subject Pre_GI entries and may be graphically visualized. Compositional similarity may be obtained by calculating OU pattern similarity. A novel GI sequence in FASTA format is first compared to the 420 MCA cluster representatives to identify corresponding clusters and then the sequence is compared to all GIs housed in the hit clusters.

More detailed ontology search to compare multiple novel GIs is possible if the sequences of predicted GIs are stored in GenBank files (a default SWGIS output option). Both sequence and compositional similarity comparisons are performed on all loaded GIs against the database to determine novel island ontology and origin. These applications enable quick and efficient investigation of novel GIs against the wealth of biological data contained in Pre_GI to determine their position in the general network of the horizontal gene exchange between bacteria.

4.3 The amalgamation of SWGIS and PRE_GI

Let's consider a case study to demonstrate the interplay between SWGIS and Pre_GI. A confirmed outbreak of canine brucellosis in Sweden during August of 2013 was caused by *Brucella canis* strain SVA13 (Kaden *et al.*, 2014). Sweden is officially free of brucellosis with the outbreaks acquired abroad. The outbreak in 2013 was caused by a male canine imported from Spain for breeding. The whole genome of the causative agent was sequenced, assembled and analyzed. Whole sequences were assembled with SeqMan 8.0.2 and aligned against the reference sequence *B. canis* ATCC 23365, accession CP007629 for the chromosome 1 and CP007630 for the chromosome 2. SWGIS was used to identify possible GIs in both chromosomes. The parameters by default were chosen and resulted in 6 GI predicted in CP007629, displayed in Fig. 6, and 1 GI was found in CP007630. All 7 SWGIS predicted GIs in composed GenBank file format were uploaded and compared to all entries in Pre_GI. An automated search for sequence similarity between the 7 uploaded GIs against the Pre_GI entries was performed in a batch by BLASN with an e-value cut-off of 1 × 10^{-5}. Plurality of GIs found in CP007629 and CP007630 showed a high sequence similarity to GIs hosted by *B. canis* ATCC 23365 (NC_010103), while the fifth GI on CP007629 indicated in Fig. 7 by an arrow showed the best hit to the GI predicted in *Bartonella grahamii* as4aup (NC_012846). It may be learnt from the host related data stored in Pre_GI that *B. grahamii* as4aup was isolated from a wood mouse (*Apodemus sylvaticus*) in central Sweden and that *Bartonella* comprises human and animal pathogens spread by the bite of a blood-sucking arthropod. It may be possible that the outbreak strain of *B. canis* resulted from an acquisition of virulence factors from the zoonotic bacterium *B. grahamii.* Compositional comparison of other GIs found in *B. canis* strain SVA13 suggested possible acquisition of GI 6 from *Desulfovibrio aespoeensis* through *B. ovis*.

5 Conclusion

Compositional comparison of bacterial genomes known also as genome linguistics is a prospective approach to cope with large scale genome comparison projects. Many computational tools based on composition similarity analysis have been proposed over the past decade and proved to be useful (Abe *et al.*, 2003; Dufraigne *et al.*, 2005; Chatterjee *et al.*, 2008; Ganesan *et al.*, 2008; Hasan *et al.*, 2012). These showed to be reliable in detection of GIs in complete genome sequences. SWGIS employs the revised OU statistics, which was introduced in our earlier papers (Reva & Tümmler, 2004 and 2005; Ganesan *et al.*, 2008). It is comparable to the other composition based methods for GI identification; particularly SIGI-HMM, which employs Hidden Markov Models (Langille & Brinkman, 2009), and GOHTAM, which uses both the chaos game model and codon bias statistics (Ménigaud *et al*., 2012). SWGIS has been scaled to analyze multiple genomes in a single run and was customized to meet the user's needs. A total of 2,407 GenBank files of bacterial chromosomes and plasmids downloaded from NCBI database were used as an input to search for GIs. A total of 26,744 GIs were identified. All these GIs and relevant supplementary information were stored in Pre_GI database (http://pregi.bi.up.ac.za/index.php).

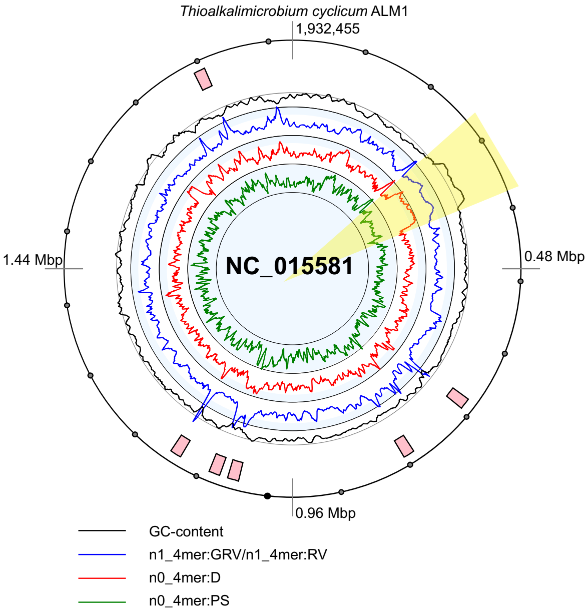


Fig. 5. An insertion of a giant viral gene into chromosome of *Thioalkalimicrobium cyclicum* ALM1 that was overlooked by SWGIS is highlighted on the atlas.

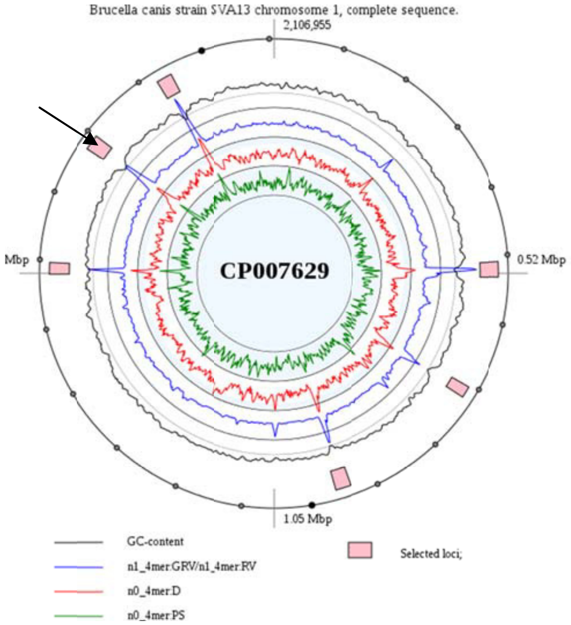


Fig. 6. Graphical representation of 6 islands found in *Brucella canis* strain SVA13 on the chromosome 1.

On average it took approximately 5-10 min for SWGIS to predict GIs in one bacterial chromosome. FNR/FPR statistics were also implemented to aid with the selection of optimal parameters for GI identification. A case study was performed to investigate the failures of the composition based GI detection and to consider possible ways to overcome these failures. The comparison of different approaches of GI identification is rather problematic. The predictions retrieved by different programs overlap only partly (see Fig. 1 and 2). This discrepancy results from the extreme versatility of HGT, which occurs through three different mechanisms: conjugation, transduction and transformation. Having been inserted, the integrated elements fall under the pressure of fragmentation and amelioration. Efficiency of different methods strongly depends on the lengths of islands, their genetic content and the time passed after inserting. The best result may be achieved when the outputs of several programs are combined, as it was implemented in the IslandViewer web-portal (Langille & Brinkman, 2009) and later in GIST (Hasan *et al*., 2012). In this work it was shown that SWGIS may significantly contribute towards the identification of GIs, which in most cases remained undetected by the IslandViewer programs.

An important issue of identification of GIs is the ability to distinguish and filter out false predictions. It has been reported that not all genomic loci showing alternative DNA compositions were horizontally transferred (Koski *et al.*, 2001; Reva & Tümmler, 2004). Nevertheless, no attempts have been made until now to estimate the rates of false negative and false positive predictions attributed to different methods. There is no consensus at the moment, which predicted GIs should be designated as false positives. Prophinder (Lima-Mendez *et al.*, 2008), Islander, SIGI-HMM and IslandPath/DIMOB (Mantri & Williams, 2004; Langille *et al.*, 2008) search for genes associated with horizontally transferred islands (transposases, integrases, viral capsid proteins, etc) to confirm the lateral origin of corresponding genomic fragments. Whereas, GOHTAM (Ménigaud *et al.*, 2012) simply returns a whole list of atypical regions found in the given genome together with their annotation data to allow users to decide themselves which of them were horizontally transferred. SWGIS employs the superposition of OU statistical parameters to distinguish between GIs and other categories of atypical genomic loci (Reva & Tümmler, 2005; Bezuidt *et al.*, 2011). It additionally performs BLASTN similarity search of the predicted DNA fragments against an incorporated database of 16S rRNA sequences to discard false selected *rrn* operons. A drawback of all these discriminating approaches is that they unavoidably increase the percentage of overlooked GIs. The factorial analysis of the proposed GI identification algorithm was performed in this work to allow users to make an informative choice in selecting of customizable parameters to ensure acceptable FNR and FPR.

The collection of identified GIs in appropriate and accessible format aids research on GI ontology, origin and biological logic of their existence. The collaboration of SWGIS and Pre_GI in GI research offers a valuable addition to other available GI detection tools with numerous advantages.

Acknowledgements

Funding: This work was funded by the grant for bioinformatics and genomics research #86941 provided by the National Research Foundation (NRF) of South Africa.

References

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T (2003) Informatics for unveiling hidden genome signatures. *Genome Res*., **13***,* 693-702.

Bezuidt O, Pierneef R, Mncube K, Lima-Mendez G, Reva ON (2011) Mainstreams of horizontal gene exchange in enterobacteria*,* consideration of the outbreak of enterohemorrhagic *E. coli* O104*:*H4 in Germany in 2011. *PLoS One*., **6***,* e25702.

Bohlin J, van Passel MW, Snipen L, Kristoffersen AB, Ussery D, Hardy SP (2012) Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands. *BMC Genomics*., **13***,* 66.

Brzuszkiewicz E, Thürmer A, Schuldes J, Leimbach A, Liesegang H, *et al.* (2011) Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). *Arch Microbiol*., **193***,* 883-891.

Chatterjee R, Chaudhuri K, Chaudhuri P (2008) On detection and assessment of statistical significance of genomic islands. *BMC Genomics*., **9***,*150.

Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol*., **16***,* 1391-1399.

Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res*., **33***,* e6.

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large- scale detection of protein families. *Nucleic Acids Res*., **30***,* 1575-1584.

Fernández-Gómez B, Fernàndez-Guerra A, Casamayor EO, González JM, Pedrós-Alió C, Acinas SG (2012) Patterns and architecture of genomic islands in marine bacteria. *BMC Genomics*., **13***,* 347.

Ganesan H, Rakitianskaia AS, Davenport CF, Tümmler B, Reva ON (2008) The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics*., **9***,* 333.

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

Hasan MS, Liu Q, Wang H, Fazekas J, Chen B, Che D (2012) GIST: genomic island suite of tools for predicting genomic islands in genomic sequences. Bio*information*., **8***,* 203-205.

Kaden R, Agren J, Ferrari S, Lindberg M, Backman S, Wahab T (2014) Whole-Genome Sequence of *Brucella canis* Strain SVA13, Isolated from an Infected Dog. *Genome Announc*.. **2**(4), e00700-00714.

Karlin S (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol*., **1***,* 598-610.

Kelly BG, Vespermann A, Bolton DJ (2009) The role of horizontal gene transfer in the evolution of selected foodborne bacterial pathogens. *Food Chem Toxicol*., **47***,* 951-968.

Koski LB, Morton RA, Golding GB (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol*., **18***,* 404-412.

Langille MG, Brinkman FS (2009) IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics*., **25***,* 664-665.

Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.,* **44***,* 383-397.

Lima-Mendez G, Van Helden J, Toussaint A, Leplae R (2008) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*., **24***,* 863-865.

Lima-Mendez G, Toussaint A, Leplae R (2010) ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res.,* **35***,* D395-D400.

Mantri Y, Williams KP (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. *Nucleic Acids Res*., **32***,* D55-D58.

Ménigaud S, Mallet L, Picord G, Churlaud C, Borrel A, Deschavanne P (2012) GOHTAM: a website for 'Genomic Origin of Horizontal Transfers, Alignment and Metagenomics'. *Bioinformatics*, **1***,* 1270-1271.

Mrázek J, Karlin S (1999) Detecting alien genes in bacterial genomes. *Ann NY Acad Sci*., **870***,* 314-329.

Potron A, Kalpoe J, Poirel L, Nordmann P (2011) European dissemination of a single OXA-48-producing *Klebsiella pneumoniae* clone. *Clin Microbiol Infect*., **17***,* E24-E26.

Pride DT, Blaser MJ (2002) Identification of horizontally acquired elements in *Helicobacter pylori* and other prokaryotes using oligonucleotide difference analysis. *Genome Let*., **1***,* 2-15.

Reva ON, Tümmler B (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics*., **5***,* 90.

Reva ON, Tümmler B (2005) Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics*., **6***,* 251.

Reva ON, Tümmler B (2008) Think big – giant genes in bacteria. *Environ Microbiol*., **10***,* 768-777.

Smith JM, Feil EJ, Smith NH (2000) Population structure and evolutionary dynamics of pathogenic bacteria. *Bioessays*., **22***,* 1115-1122.

St-Pierre NR, Weiss WP (2009) Technical note: designing and analyzing quantitative factorial experiments. *J Dairy Sci*., **92***,* 4581-4588.

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci U.S.A*., **48***,* 582-592.

van Passel MW, Bart A, Thygesen HH, Luyf AC, van Kampen AH, van der Ende A (2005) An acquisition account of genomic islands based on genome signature comparisons. *BMC Genomics*., **6***,* 163.

van Passel MW (2011) Tracing common origins of genomic islands in prokaryotes based on genome signature analyses. *Mob Genet Elements*., **1***,* 247-249.

Yoon SH, Park Y-K, Lee S, Choi D, Oh TK, *et al.* (2007) Towards pathogenomics: a web-based resource for pathogenicity islands. *Nucleic Acids Res*., **35***,* D395-400.