Table 2
The summary of selected programs for predicting genomic islands.

Program	Form	Availability
Methods based on gene composition of one genome		
PAI-IDA [30]	Command line	Upon request
SIGI-HMM [31]	Graphical interface	https://www.uni-goettingen.de/en/research/185810.html
Methods based on DNA composition of one genome		
Window-based methods		
AlienHunter [32]	Command line	http://www.sanger.ac.uk/resources/software/alien_hunter
Centroid [33]	Command line	Upon request
Design-Island [34]	Command line	http://www.isical.ac.in/~rchatterjee/Design-Island.html
INDeGenIUS [35]	Command line	Upon request
GI-SVM [36]	Command line	https://github.com/icelu/GI_Prediction
Windowless methods		
GC Profile [37,38,60]	Web-based	http://tubic.tju.edu.cn/GC-Profile
MJSD [39]	Command line	http://cbio.mskcc.org/~aarvey/mjsd/
Methods based on GI structure of one genome		
Direct integration methods		
IslandPath [40]	Web-based	http://www.pathogenomics.sfu.ca/islandpath/
Machine learning methods		
GIDetector [41]	Command line	http://www5.esu.edu/cpsc/bioinfo/software/GIDetector
GIHunter [42]	Command line	http://www5.esu.edu/cpsc/bioinfo/software/GIHunter
Methods base on multiple genomes		
tRNAcc [43]	Web-based	http://db-mml.sjtu.edu.cn/MobilomeFINDER/
IslandPick [27]	Command line	http://www.pathogenomics.sfu.ca/islandviewer/download/
Ensemble methods		
IslandViewer [44–46]	Web-based	http://www.pathogenomics.sfu.ca/islandviewer
EGID [47]	Command line	http://www5.esu.edu/cpsc/bioinfo/software/EGID
GIST [48]	Graphical interface	http://www5.esu.edu/cpsc/bioinfo/software/GIST
PredictBias [49]	Web-based	http://www.bioinformatics.org/sachbinfo/predictbias.html
PIPS [50]	Command line	http://www.genoma.ufpa.br/lgcm/pips
Methods for incomplete genome		
GI-POP [51]	Web-based	http://gipop.life.nthu.edu.tw

Both AlienHunter and GI-SVM use a fixed-size overlapping window of fixed step size. AlienHunter is the first program for GI detection on raw genomic sequences. It measures segment atypicality via relative entropy based on interpolated variable order motifs (IVOM). The threshold can be obtained by either k-means clustering or standard deviation (when there are fewer samples). GI-SVM is a recent method using either fixed or variable order k-mer frequencies. It detects atypical windows via one-class SVM with spectrum kernel. An automatic threshold can be obtained from one dimensional k-means clustering.

Centroid partitions the genome by a non-overlapping window of fixed size. The average of k-mer frequency vectors for all the windows is seen as the centroid. Based on the Manhattan distances from each frequency vector to the centroid, outlier windows are selected by a threshold derived from standard deviation. INDeGenIUS is a method similar to Centroid. But it uses overlapping windows of fixed size and computes the centroid via hierarchical clustering.

Design-Island is a two-phase method utilizing k-mer frequencies. It incorporates statistical tests based on different distance measures to determine the atypicality of a segment via pre-specified thresholds. In the first phase a variable-size window is used to obtain initial GIs, whereas in the refinement phase a smaller window of fixed size is used to scan over these putative GIs for getting final GI predictions.

Some of these methods are designed to alleviate the problem of genome contamination. Design-Island excludes the initially obtained putative GIs when computing parameters for the entire genome in the second phase. GI-SVM measures the atypicality of all the windows simultaneously via one-class SVM, and only some windows contribute to the genomic signature.

To deal with the imprecise GI boundaries that result from a large step size, AlienHunter uses HMM to further localize the boundaries between predicted GIs and non-GIs. But most other programs do not consider this issue.

The few windowless methods mainly include GC Profile [37,60] and MJSD [39].

GC Profile is an intuitive method to calculate global GC content distribution of a genome with high resolution. The abrupt drop in the profile indicates the sharp decrease of GC content and thus the potential presence of a GI. This method was later developed into a web-based tool which is used for analyzing GC content in genome sequences [38]. However, other features have to be used together with GC Profile for GI prediction due to the poor discrimination power of GC content.

MJSD is a recursive segmentation method based on Markov Jensen-Shannon divergence (MJSD) measure. The genome is recursively cut into two segments by finding a position where the sequences to its left and to its right have statistically significant compositional differences. Subsequently, each segment is compared against the whole genome to check its atypicality via a predefined threshold.

Methods based on DNA sequence composition have the similar advantages and disadvantages as methods based on gene sequence composition.

Specifically, window-based methods can be highly sensitive with appropriate implementations. For example, AlienHunter was reported to have the highest recall in previous evaluation [27], and GI-SVM was recently shown to have even higher sensitivity than AlienHunter [36]. But their precisions are quite low due to the limited input information. They are also inherently incapable of identifying the precise boundaries between regions with compositional differences [39].

In contrast, windowless methods can delineate the boundaries between GIs and non-GIs more accurately [39]. GC Profile has successfully discovered a few reliable GIs in several genomes [60]. But it seems subjective to access the abruptness of jump in the GC profile, and only GIs with low GC content can be detected. MJSD is better at predicting GIs of size larger than 10 kb [39], but the procedure to determine segment atypicality still suffers from the contamination of the whole genome.