MTGIpick

MTGIpick allows the robust identification of genomic islands from a single genome

---

Qi Dai, Chaohui Bao, Yabing Hai, Sheng Ma, Tao Zhou, Cong Wang, Yunfei Wang, Xiaoqing Liu, Yuhua Yao, Wenwen Huo, Zhenyu Xuan, Min Chen and Michael Q Zhang. MTGIpick allows the robust identification of genomic islands from a single genome.
To be submitted to Briefings in Bioinformatics.

Qi Dai
College of Life Sciences,
Zhejiang Sci-Tech University, Hangzhou 310018,
China

Contents

1. Overview

MTGIpick is a software that implements multiscale statistical algorithm to predict genomic islands (GIs) from a single genome. It uses small-scale test with large-scale features to score small region deviating from the host and large-scale statistical test with small-scale features to identify multi-window segments for identification of GIs. MTGIpick can identify GIs from a single genome, without annotated information of genomes or prior knowledge from other datasets. In simulations with alien fragments from artificial and real genomes, MTGIpick reported robust results across different experiments. From real biological data, MTGIpick demonstrated better performance compared with existing methods, and identified GIs with more accurate size.

MTGIpick was written in Matlab and Java, compiled in Windows and Linux, and run on those platforms. We have supplied a version of MTGIpick. The output of MTGIpick consists of genomic signatures, conserved score of each predicted GIs and the total scale predict GIs. In addition, MTGIpick provides a new interactive genome visualization tool, which uses zoomable sunburst and sunburst partition to represent predicted GIs with conserved scores along the whole genome.

MTGIpick version 1.2.0 has been released with some new features:

- The input format follows the standard Fasta format. Multiple DNA sequences are supported. The file suffix is recommended to be .fasta or .fa.

- The uploading progress is monitored using a progress bar.

- New options to process the input file is added.

- The software efficiency has been improved.

- Reduce parameter selection on prediction.

- Warning messages are added to notify the users when anything wrong occurs.

- Output file is more readable and more details.

- An interactive visualization tool is added for predicted genomic islands along the whole genome.

Contact: daiailiu04@yahoo.com

2. Obtain and install

2.1. Requirements

MTGIpick has been compiled and tested under Sun Java interpreter and Matlab. MTGIpick can be used in Windows- and Linux- platforms. Java Virtual Machine and MATLAB Compiler Runtime (MCR) are required for MTGIpick setup on your platform. However, we strongly advise the use of openjdk (JDK) instead of the Oracle version of java virtual machine when working in linux-based machines as the Oracle version may result in some exceptions during the analyses.

Software	On window(x64)	On Linux(x86_64)
Java Virtual Machine	JDK 1.8	JDK 1.8
MATLAB Compiler Runtime	MCR 8.4	MCR 8.1

2.2. Download

There are two ways to download the MTGIpick:

1) You can download the MTGIpick package with JDK and MCR from our web (http://bioinfo.zstu.edu.cn/MTGI) .

A) Windows (MTGIpick.zip)

- MCRInstaller.exe # MCR 8.4 for Windows
- MTGI_setup.exe # Main program
- Example.fasta # Two sequences in FASTA format
- README.txt # Documentation
- help.htm (help.files) # Introduction

B) Linux (MTGIpick.zip)

- jdk-8u102-linux-x64.tar.gz # JDK 1.8 for Linux
- MCRInstaller.zip # MCR 8.1 for Linux
- MTGI_linux.jar # Main program

- install_jdk-mcr_linux.sh # Install JDK 1.8 and MCR 8.1

- run_MTGI_linux.sh # Run MTGIpick software

- Example.fasta # Two sequences in FASTA format

- README.txt # Documentation

- help.htm (help.files) # Introduction

- jsonFile (d3) # Visualization files

2) If you download the MTGIpick package without JDK 1.8 or MCR 8.4 (8.1) from our webs (http://bioinfo.zstu.edu.cn/MTGI , https://github.com/bioinfo0706/MTGIpick) , download the JDK and MCR for your platform from the following Web:

JDK 1.8: http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html#jdk-8u102-oth-JPR

MCR: http://www.mathworks.com/products/compiler/mcr/?refresh=true

Make sure that the JDK and MCR are saved into the folder of the MTGIpick, and please rename the MCR as MCRInstaller.

2.3. Install and Run

1) Windows (Tested on win7 x64)

Before installing MTGIpick, make sure that the MCR 8.4 for windows is saved into the same folder of MTGIpick software. Install MCR 8.4 first, and run MTGIpick setup directly.

2) Linux (Tested on CentOS7)

Before installing MTGIpick, make sure that the JDK 1.8 (jdk-8u102-linux-x64.tar.gz) and MCR 8.1 for Linux are saved into the same folder of MTGIpick software. Please follow the following steps for installing and running MTGIpick:

Step 1

To install JDK 1.8 (jdk-8u102-linux-x64.tar.gz) and MCR 8.1, it requires a simple command line as follow:

> bash install_jdk-mcr_linux.sh

Step 2

To run the MTGIpick software, just type a simple command line as follow (Once the first step has run, execute the second step to run MTGIpick):

>bash run_MTGI_linux.sh

3. Tutorial

3.1. Functions and features

1. IST-LFS: an iteration of small-scale t-test with large-scale feature selection to quantifying the compositional difference between a region and the 'native' genome regions.

2. MSA: A multiscale segmentation algorithm to investigate the variability of genomic signals and to identify large, multi-window segments.

3. CG-MJSD: A boundary detection method based on CG-based segmentation and Markovian Jensen-Shannon divergence.

4. The methods and parameters could be chosen according to your purpose.

5. Using Zoomable Sunburst and Sunburst Partition to represent the predicted GIs with conserved scores along the whole genome.

3.2. Input

3.2.1 Input file format

MTGIpick accepts DNA sequences, and the input file has to be in fasta, fa, or fna format. For example, the file name is example.fasta or example.fa, its content looks like this:

>cya_GI1_93960_99000_GI2_208020_223020_GI3_408480_410520

CCCCATTCCCCCCATTCCCTCCTTTTCCACCATACCCTCTTTTCCCCTCGTTGCCCCCAA

ATTTTTACGCATTTCCCCATTAATGCGATGATCCCAGCGCGAAAGCATCTGTGATTAAGA

CGTCTATCAATTTATACTCGTTAGGGTTTTTTCTTCGGTGGTACCATCTGGGCGCCTACG

>……

Example DNA input (.fna format), its content looks like this:

>cya_GI1_93960_99000_GI2_208020_223020_GI3_408480_410520

CCCCATTCCCCCCATTCCCTCCTTTTCCACCATACCCTCTTTTCCCCTCGTTGCCCCCAA

ATTTTTACGCATTTCCCCATTAATGCGATGATCCCAGCGCGAAAGCATCTGTGATTAAGA

>ctt_ GI2_208020_223020_GI1_93960_99000_GI3_408480_410520

CCCCATTCCCCCCATTCCCTCCTTTTCCACCATACCCTCTTTTCCCCTCGTTGCCCCCAA

ATTTTTACGCATTTCCCCATTAATGCGATGATCCCAGCGCGAAAGCATCTGTGATTAAGA

Please note GenBank (from NCBI) will not work with MTGIpick. If using sequences from NCBI be sure to save them as FASTA format first. Sequence format conversion tools are available at http://www.ebi.ac.uk/Tools/sfc/.

The sequences can be uploaded to MTGIpick in a file. It is very important that each of the sequences has a unique name. If they do not, the software will fail. There must be no empty lines, white spaces or control characters between sequences or at the top of the file. This will also cause the software to fail.

3.2.2 Process Input file

There will be an upload progress bar to monitor upload progress when clicking the button to upload a file. If the input file you uploaded contains at least two sequences, a dialog box appears to tell you to select a way to process the input file.

Predict each sequence separately. For each DNA sequence of the input file, MTGIpick will process it separately and predict its GIs.

Assemble and predict. The sequences from the input file are assembled into a sequence according to the order of these sequences in the input file, and MTGIpick will process it as a sequence and predict its GIs.

3.3. Methods

3.3.1 IST-LFS method

1) Framework

IST-LFS is a proposed small-scale t-test with large-scale feature selection that was used to quantify the compositional differences of a region from the host in the MTGIpick. It is efficient at detecting horizontal gene transfers or GIs with small sizes. The steps are described below:

a) Split a genome into n non-overlapping windows of size 1kb.

b) Calculate the frequencies of the tetranucleotides in each window as genomic signatures.

c) Extract the signatures of the host with the help of the confidence intervals on the windows' variances.

d) Calculate the kurtosis of each tetranucleotide across n windows and select the windows with a larger kurtosis as informative signatures.

e) Measure the divergence of the ith window from the host using the two-sample t-test.

f) Select windows whose scores are large enough to be considered to be statistically significant.

g) Delete the selected windows and update all of windows of the genome; then, repeat steps d-f until there is no window to be found.

h) Refine the boundaries of the predicted GIs using the CG-MJSD method.

2) Parameters of IST-LFS Software

Word size: the length of k-mer, and default is 4.

Windowed transform: the total number of the windows used in genomic transformation, and default is 1.

Iteration time: the periods of time that are repeated to select windows whose scores are large enough to be considered statistically significant, and default is 5.

Core feature size: the size of selected features by the proposed kurtosis, and default is 256.

Eye window size: the size of eye windows used in the proposed divergence measure based on two-sample t-test, and default is 2.

Time standard error: the standard deviation of the mean of the window scores to select windows associated with putative GIs, and default is 0.05.

Upstream/downstream of 'raw' GIs: the length of sequences around 'raw' GIs to refine the boundaries of predicted GIs, and default is 5kb.

3.3.2 MTGIpick method

1) Framework

MTGIpick is a novel method for the robust identification of GIs using the multiscale statistical testing. The steps are described below:

a) Split a genome into n non-overlapping windows of size 1kb.

b) Calculate the frequencies of the tetranucleotides in each window as genomic signatures.

c) At a smaller scale, we propose an iteration of small-scale t-tests with large-scale feature selection (IST-LFS) to quantify the compositional differences of a region from the host.

d) At a large scale, we investigated the variability of genomic signals and used multiscale segmentation algorithm (MSA) to identify large, multi-window segments.

e) Calculate conserved score of each nucleotide according to the total number of appearances in the selected segments, from which GIs are detected with respect to their conserved scores.

f) Refine the boundaries of the predicted genomic islands using the CG-MJSD method.

2) Parameters of MTGIpick Software

Word size: the length of k-mer, and default is 4.

Windowed transform: the total number of the windows used in genomic transformation, and default is 4.

Iteration time: the periods of time that are repeated to select windows whose scores are large enough to be considered statistically significant, and default is 5.

Core feature size: the size of selected features by the proposed kurtosis, and default is 256.

Eye window size: the size of eye windows used in the proposed divergence measure based on two-sample t-test, and default is 5.

Time standard error: the standard deviation of the mean of the window scores to select windows associated with putative GIs, and default is 0.05.

Total scale: the total number of the scales in the multiscale segmentation algorithm, and default is 45.

Minimum GI size: the minimum size of predicted genomic islands, and default is 10kb. A smaller value is recommended if you want to predict GIs with small size.

Time standard error: the standard deviation of the mean of enrichment scores to select segments associated with putative GIs, and default is 0.3.

Upstream/downstream of 'raw' GIs: the length of sequences around 'raw' GIs to refine the boundaries of predicted GIs, and default is 5kb.

3.4. Results

3.4.1 Outputs

A dialog box appears to tell you to select a way to download the results once your project is complete. There are two ways to download the results:

1). Download the results by clicking save button.

2). Find the results in the same directory where the input file is stored.

The outputs of the MTGIpick consist of genomic signatures, conserved scores of predicted GIs and predicted GIs. They are stored in the same directory where the input file is stored. Output of the genomic signatures is a Zip file whose name is created by the input file name and signature. If the input file contains at least two sequences, the Zip file contains more signature files for all the sequences.

Example the genomic signature file, its content looks like this:

>gi|16758993|ref|NC_003198.1| Salmonella enterica subsp. enterica serovar Typhi str. CT18

NO	kmer	kurtosis score
1	TCCC	1.509693e+01

2 AGTT 1.461676e+01
3 CTAG 1.180616e+01
4 CCCC 1.132821e+01
5 GGGG 1.041196e+01
6 TATA 9.547606e+00
7 GTTC 8.937067e+00
8 AAGT 8.544586e+00
9 CAGT 8.471593e+00
10 CCCA 7.966514e+00
.....

The first line is sequence name, and all the genomic signatures are sorted according to their kurtosis scores.

Output of the predicted GIs of total scales is a Zip file whose name is created by the input file name and predict GIs. If input file contains at least two sequences, the Zip file contains more predict GIs files for all the sequences.

Here is a brief description of the predicted GIs of the total scales:

1. seqid - The name of the sequence.

2. source - The program MTGIpick

3. type- The name of this type of feature, genomic island.

4. start - The starting position of the predicted genomic island in the sequence.

5. end - The ending position of the predicted genomic island in the sequence.

6. score - The conserved scores of each nucleotide according to the total number of appearances in the predicted genomic island along total scales in MTGIpick.

7. strand - The strand of the feature, + for positive strand.

8. phase - There's no annotation information and be replaced with "." charcater.

9. attributes - The numbers of the genomic islands.

Example the total scale predict GIs looks like this:

##gff-version 3
##sequence-region Example 1 757540

Example	MTGIpick	genomic_island	125001	137000	1	+	.	ID=GI1
Example	MTGIpick	genomic_island	266001	269000	6	+	.	ID=GI2
Example	MTGIpick	genomic_island	269001	272000	7	+	.	ID=GI3
Example	MTGIpick	genomic_island	272001	287000	30	+	.	ID=GI4
Example	MTGIpick	genomic_island	287001	288000	9	+	.	ID=GI5
Example	MTGIpick	genomic_island	288001	305000	30	+	.	ID=GI6
Example	MTGIpick	genomic_island	305001	306000	28	+	.	ID=GI7
Example	MTGIpick	genomic_island	328001	329000	2	+	.	ID=GI8
Example	MTGIpick	genomic_island	329001	340000	3	+	.	ID=GI9
Example	MTGIpick	genomic_island	340001	342000	1	+	.	ID=GI10
Example	MTGIpick	genomic_island	430001	431000	1	+	.	ID=GI11
Example	MTGIpick	genomic_island	431001	447000	2	+	.	ID=GI12
Example	MTGIpick	genomic_island	546001	558000	1	+	.	ID=GI13

.....

Output of each scale predict GIs is a Zip file whose name is created by the input file name and predict. If the input file contains at least two sequences, the Zip file contains more predicted GI files for all the sequences.

Here is a brief description of the predicted GIs of each scale:

1. seqname - The name of the sequence.

2. source - The program MTGIpick

3. type - The name of this type of feature, genomic island.

4. start - The starting position of the predicted genomic island in the sequence.

5. end - The ending position of the predicted genomic island in the sequence.

6. score - The conserved scores of each nucleotide according to the total number of appearances in the predicted genomic island along total scales in MTGIpick.

7. strand - The strand of the feature, + for positive strand.

8. phase - There's no annotation information and be replaced with "." charcater.

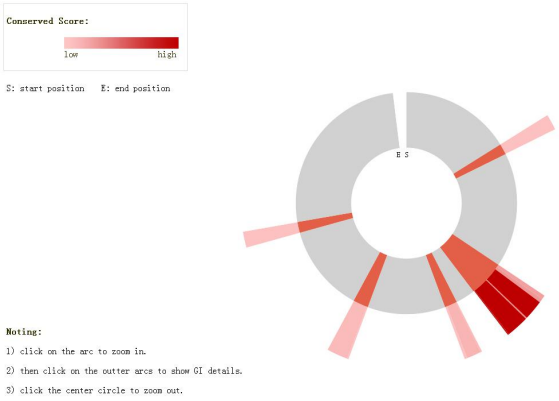
9. attributes - the numbers of the genomic islands.

The predict GIs of each scale looks like this:

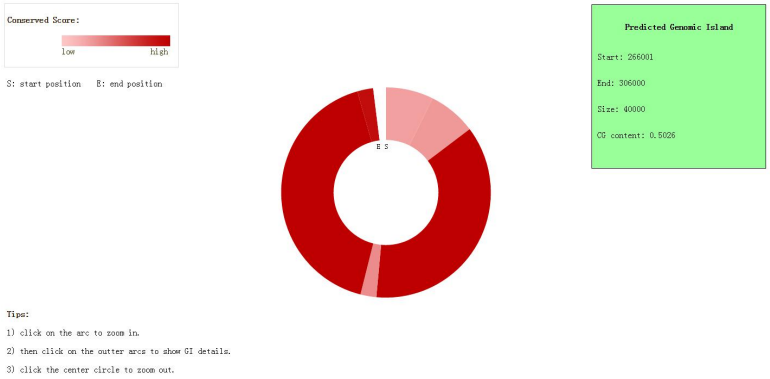
```
Scale 1
##gff-version 3
##sequence-region Example 1 757540
Example   MTGIpick  genomic_island      272001    287000    1         +         .         ID=GI1
Example   MTGIpick  genomic_island      288001    306000    1         +         .         ID=GI2

Scale 2
##gff-version 3
##sequence-region Example 1 757540
Example   MTGIpick  genomic_island      272001    287000    1         +         .         ID=GI1
Example   MTGIpick  genomic_island      288001    306000    1         +         .         ID=GI2

Scale 3
##gff-version 3
##sequence-region Example 1 757540
Example   MTGIpick  genomic_island      272001    287000    1         +         .         ID=GI1
Example   MTGIpick  genomic_island      288001    306000    1         +         .         ID=GI2
.....
```

3.4.2 Interactive visualization tool

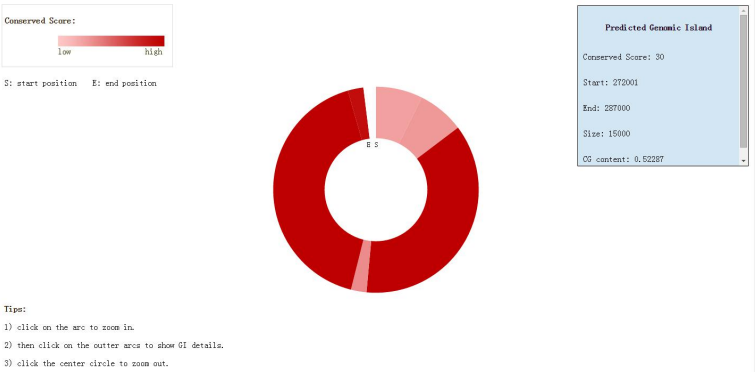
MTGIpick provides a new interactive genome visualization tool, which uses zoomable sunburst and sunburst partition to represent predicted GIs with conserved score along the whole genome. MTGIpick has generated a number of html files in the same directory where the input file is stored, you can open them directly and view the predicted GIs with conserved scores.



Orange regions in the first circle represent GIs predicted by MTGIpick method at all the scales. Click on any orange region in the first circle to zoom in, conserved scores of this predicted GI will be represented, where darker red colors are used to represent higher conserved score; light-to-dark for low-to-high.



Click on any region of the above predicted GI, more information on the predicted GI will be represented at the top right hand corner of the page, such as start position, end position, size, conserved score and GC content.



Click on the center circle to zoom out.

4. References

1. Dhillon BK, Chiu TA, Laird MR, et al. IslandViewer update: Improved genomic island discovery and visualization. Nucleic Acids Res 2013;41:W129-132.
2. Aaron JA, Rajeev K, Azad AR, et al. Detection of genomic islands via segmental genome heterogeneity. Nucleic Acids Res 2009;37:5255-5266.
3. Jaron KS, Moravec JC, Martinkova. SigHunt: horizontal gene transfer finder optimized for eukaryotic genomes. Bioinformatics 2014;30:1081-1086.
4. Langille MG, Hsiao WW, Brinkman. Detecting genomic islands using bioinformatics approaches. Nature Rev Microbiol 2010;8:373-382.

If you have any problems, please contact:
daiailiu04@yahoo.com