GI-Cluster: detecting genomic islands in newly sequenced microbial genomes via consensus clustering on multiple features

Bingxin Lu and Hon Wai Leong
{bingxin and leonghw}@comp.nus.edu.sg
Department of Computer Science, National University of Singapore

Introduction

Lateral gene transfer (LGT), the transfer of genetic materials between two reproductively isolated organisms, is an important process in evolution. A large continuous genomic region acquired by LGT is often called a **genomic island (GI)**.

GIs can promote microbial genome evolution and adaptation of microbes to environments. They may also contain genes involved in pathogenesis and antibiotic resistance. Thus, the accurate inference of GIs is important for both evolutionary study and medical research.

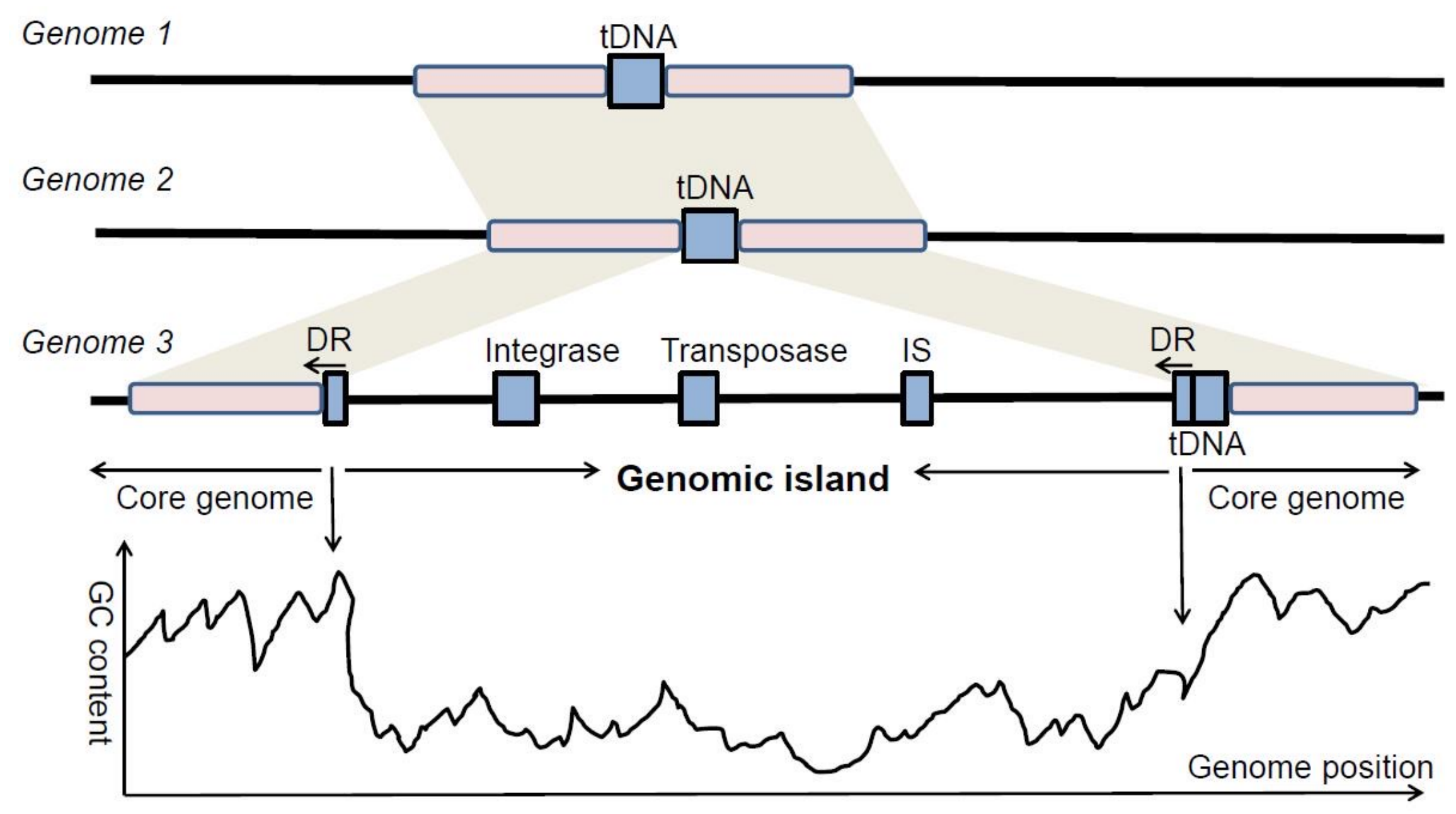


Figure 1. The schematic representation of features associated with a GI [1].

The rapid increase of newly sequenced microbial genomes calls for better GI detection methods. To get better predictions, it seems necessary to utilize multiple GI-related evidence. But it is hard to get a systematic integration of different features, due to the extreme variety of GIs.

Method

GI-Cluster utilizes **consensus clustering** [2] to detect GIs from a newly sequenced genome by combining separate clusterings of genomic segments obtained on multiple GI-related features.

The assumption that consensus clustering works for GI prediction: given a set of genomic segments in a genome, different features delineating a segment may lead to different partitions of these segments. By identifying segments in the same group in most partitions, we can get more stable pairwise relationship and hence more robust clustering of these segments.

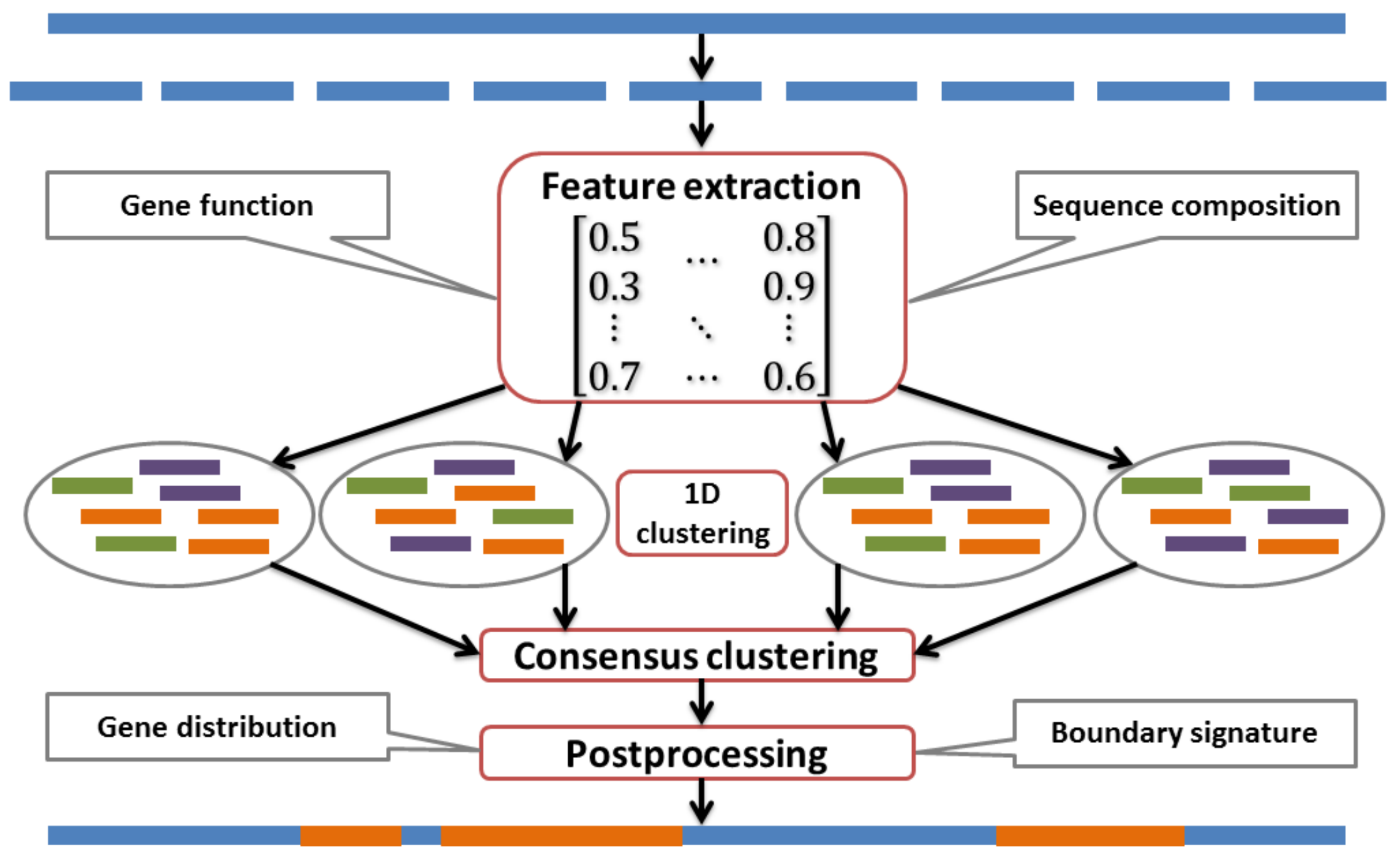
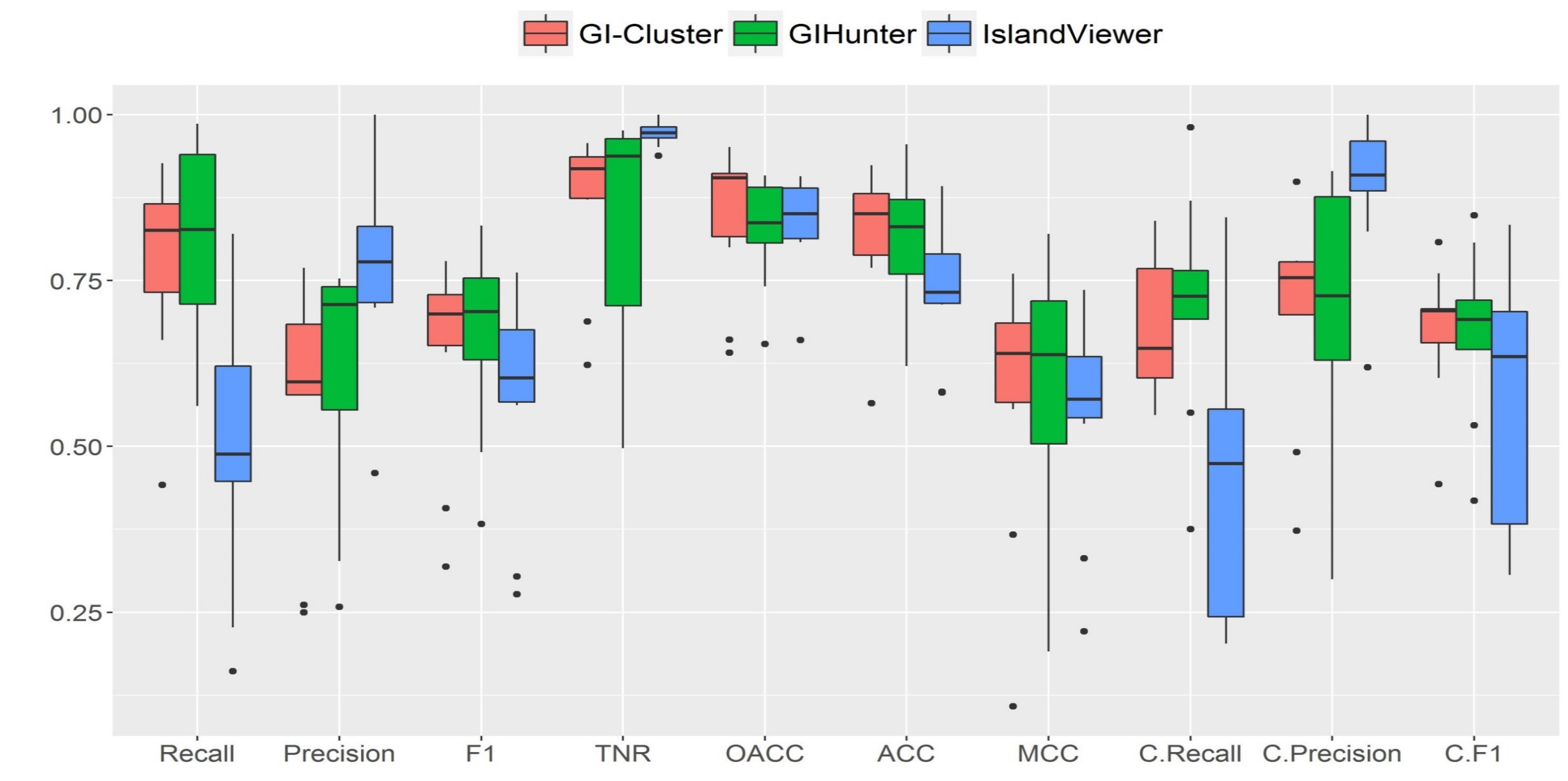


Figure 2. The major steps in GI-Cluster for identifying GIs in a genome.

Table 1. The methods used for extracting GI-related features in a genome.

Category	Feature	Measure	Computing method
Sequence composition	GC content	χ2	Python scripts
	Codon usage	χ2,Cub, AAub, CAI	Python scripts, codonW
	k-mer frequency	Covariance	Python scripts
Gene function	Mobility-related gene	Percentage	HMMer against Pfam
	Phage-related gene	Percentage	Blast against PHAST
	Virulence factor	Percentage	Blast against VFDB
	Antibiotics resistance gene	Percentage	Blast against CARD
	Novel gene	Percentage	Blast against COG
	Non-coding RNA	Count	Infernal against Rfam
Gene distribution	Gene density	Definition	Python scripts
	Intergenic distance	Definition	Python scripts
Boundary signature	tRNA	Binary	tRNAscan-SE
	Short repeat	Binary	Repseek

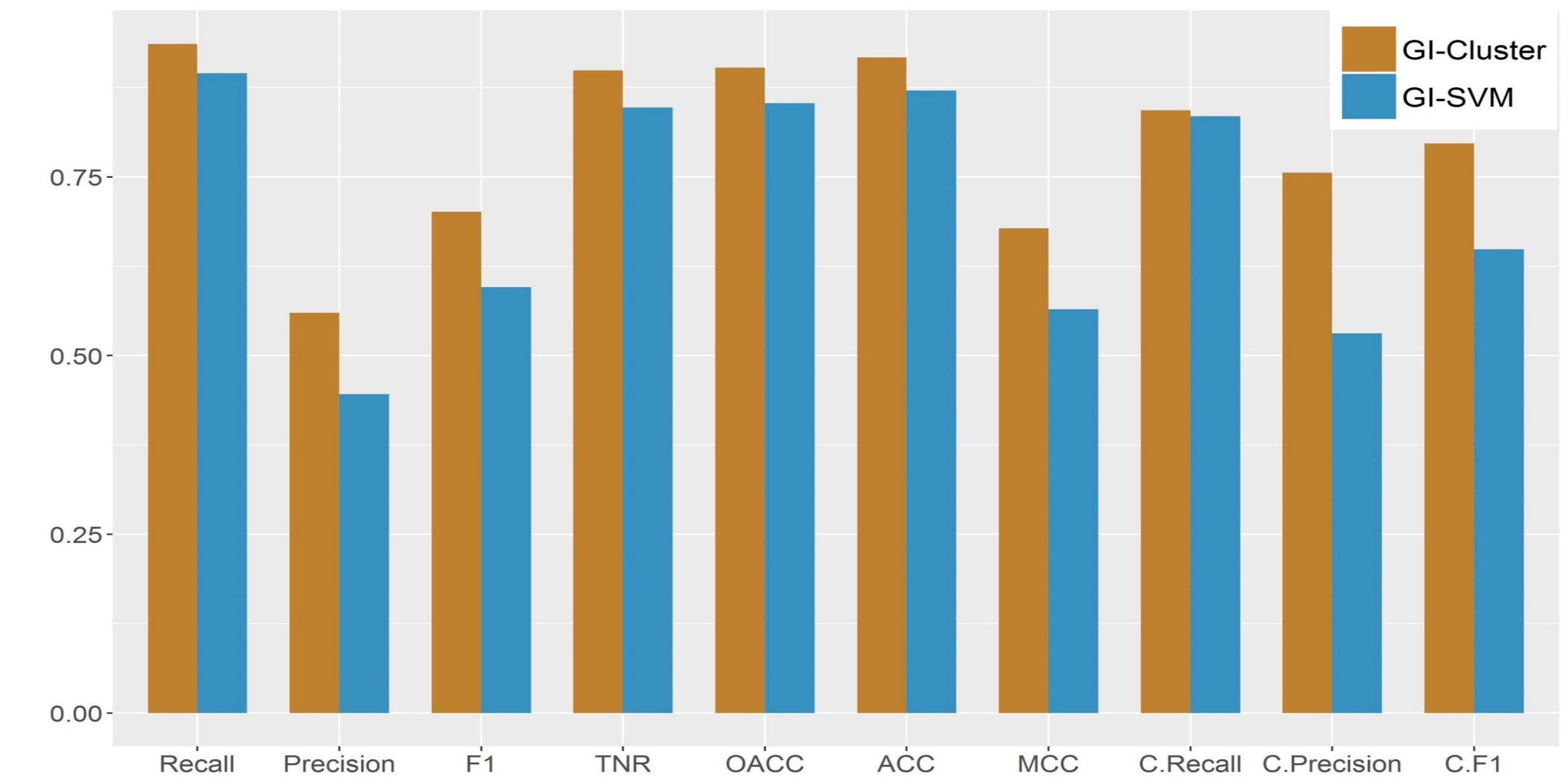
Results



We performed evaluations on two sets of GIs on 10 bacterial genomes from 6 orders [3]:
One is **L-dataset**, collected from literature;
The other is **C-dataset**, predicted by comparative genomics.

GI-Cluster had robust predictions on most of these genomes and sometimes outperformed a supervised method, GIHunter.

Figure 3. Performance comparison of GI-Cluster with GIHunter and IslandViewer.



GI-Cluster is also applicable to GI candidates predicted by programs with high recall but low precision.

Taking these initial predictions as input, GI-Cluster can help to reduce potential false positives and narrow down the search for true positives.

Figure 4. Improvement of GI-Cluster over the predictions of GI-SVM on *S. typhi* CT18 genome.

GI-Cluster provides scripts to visualize GIs predicted by different methods, which helps to show a more comprehensive picture of potential GIs.

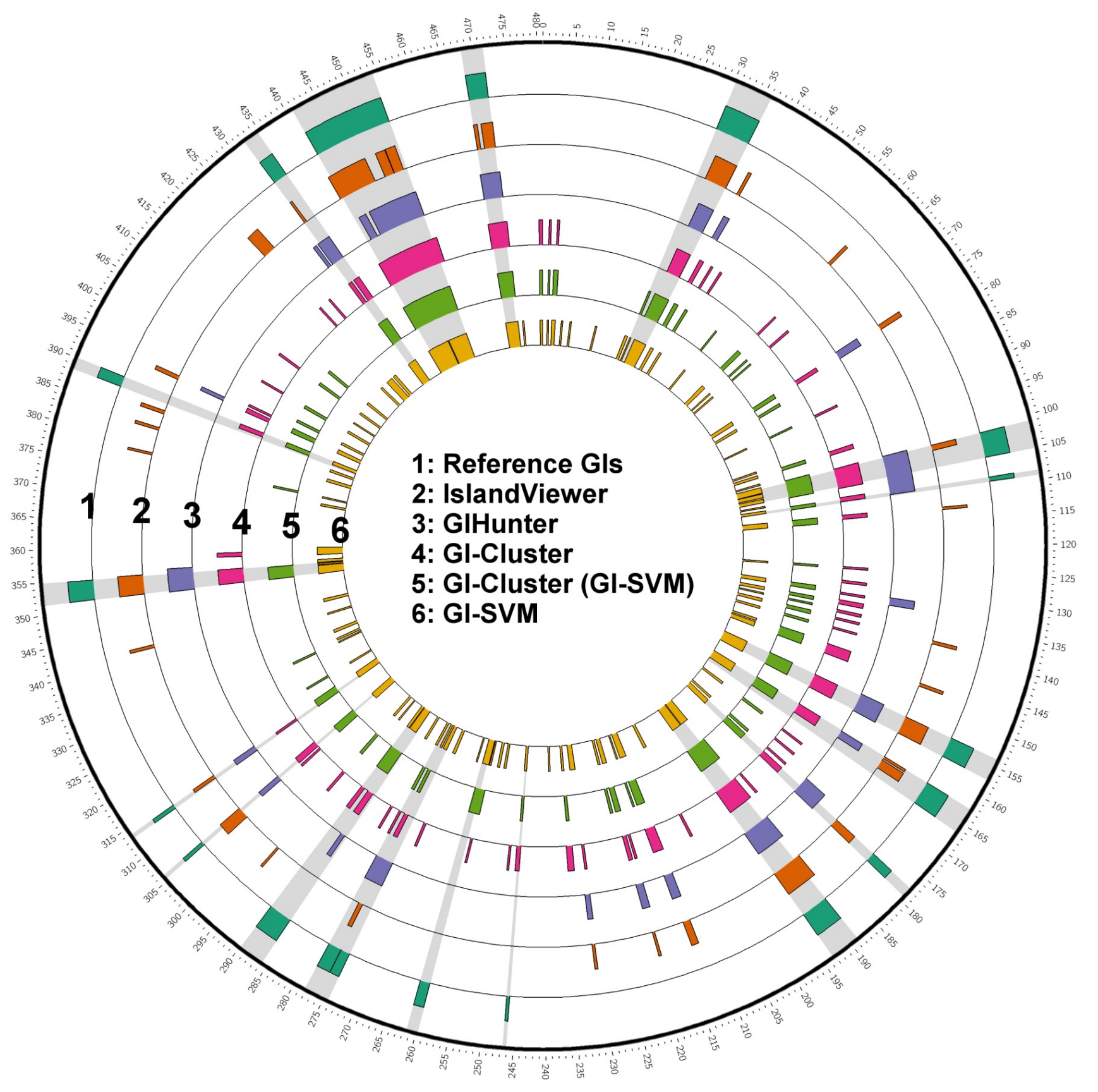


Figure 5. Comparisons of reference GIs and predicted GIs by multiple methods on *S. typhi* CT18 genome.

GI-Cluster provides scripts for visualizing the distribution of predicted GIs and GI-related features along a microbial genome.

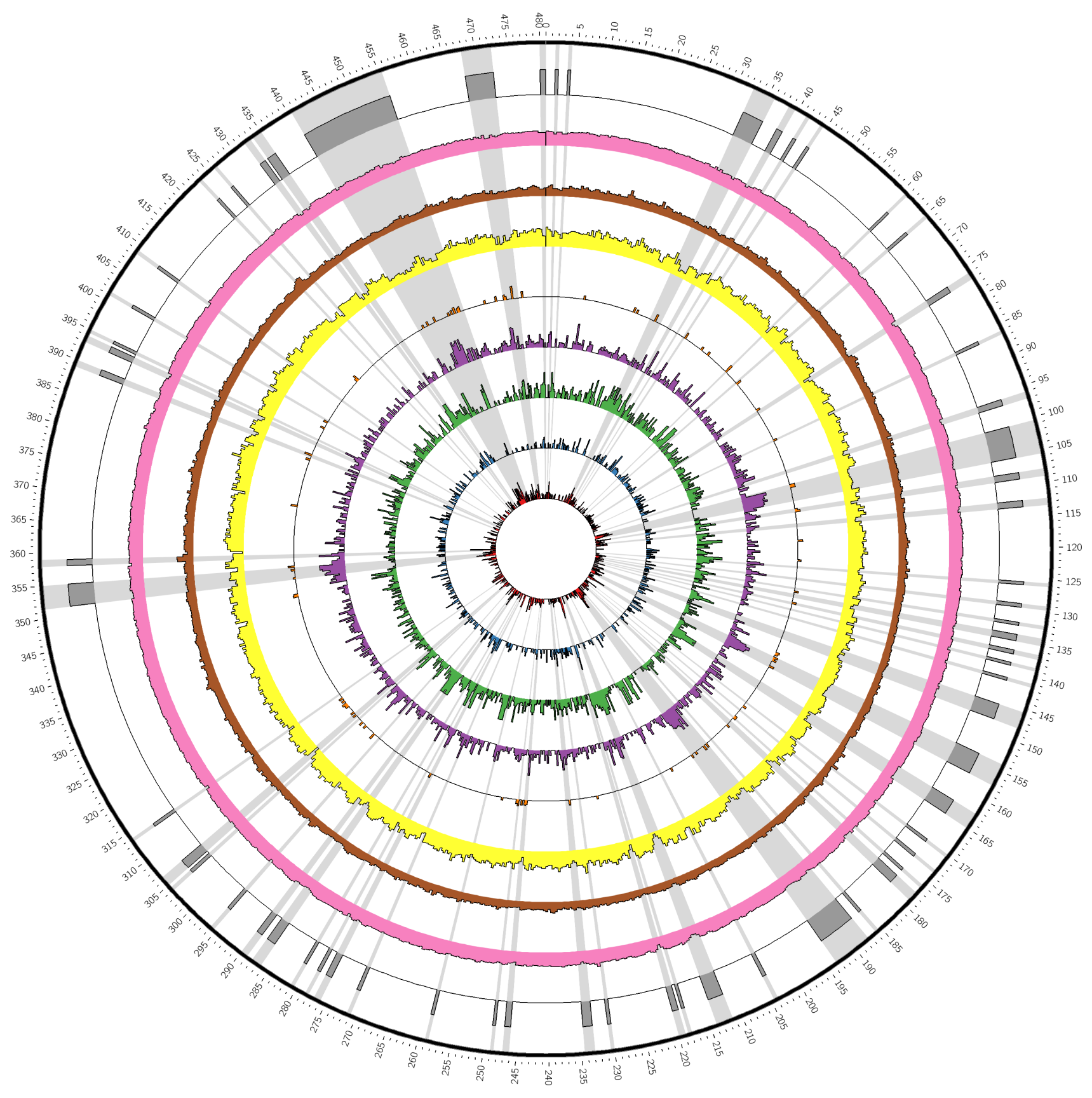


Figure 6. GIs and GI-related features computed by GI-Cluster on *S. typhi* CT18 genome.

Conclusion

- We develop a novel method for GI detection, GI-Cluster, which takes advantage of consensus clustering to identify GIs by effectively integrating multiple GI-related features.
- GI-Cluster takes genome sequence and related databases as input. It has comparable performance as supervised methods and is widely applicable.
- GI-Cluster is also a pipeline that annotates a newly sequenced microbial genome from multiple aspects. The extensive annotations and visualizations are helpful for manual analysis of GIs.
- As a stand-alone tool, GI-Cluster is easy to use and adapt to specific requirements.

Reference

Lu, B., & Leong, H. W. (2016). Computational methods for predicting genomic islands in microbial genomes. Comput Struct Biotechnol J, *14*, 200-206.
Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. Intern J Pattern Recognit Artif Intell, 25(03), 337-372.
Wei, W., Gao, F., Du, M. Z., Hua, H. L., Wang, J., & Guo, F. B. (2016). Zisland Explorer: detect genomic islands by combining homogeneity and heterogeneity properties. Brief. Bioinform, bbw019.