

Current state and future prospects of Horizontal Gene Transfer detection

Andre Jatmiko Wijaya^{1,2,3,*}, Aleksandar Anžel¹, Hugues Richard³, Georges Hattab^{1,2}

¹Center for Artificial Intelligent in Public Health Research (ZKI-PH), Robert Koch Institute, Nordufer 20, 13353 Berlin, Germany

²Department of Mathematics and Computer Science, Freie Universität, Arnimallee 14, 14195 Berlin, Germany

³Genome Competence Center (MF1), Robert Koch Institute, Nordufer 20, 13353 Berlin, Germany

*To whom correspondence should be addressed. Email: wijayaa@rki.de

Abstract

Artificial intelligence (AI) has been shown to be beneficial in a wide range of bioinformatics applications. Horizontal Gene Transfer (HGT) is a driving force of evolutionary changes in prokaryotes. It is widely recognized that it contributes to the emergence of antimicrobial resistance (AMR), which poses a particularly serious threat to public health. Many computational approaches have been developed to study and detect HGT. However, the application of AI in this field has not been investigated. In this work, we conducted a review to provide information on the current trend of existing computational approaches for detecting HGT and to decipher the use of AI in this field. Here, we show a growing interest in HGT detection, characterized by a surge in the number of computational approaches, including AI-based approaches, in recent years. We organize existing computational approaches into a hierarchical structure of computational groups based on their computational methods and show how each computational group evolved. We make recommendations and discuss the challenges of HGT detection in general and the adoption of AI in particular. Moreover, we provide future directions for the field of HGT detection.

Introduction

Artificial intelligence (AI) has revolutionized the world in various fields, including natural language processing (NLP), computer vision, and bioinformatics. With the rapid growth of available biological data, classical machine learning (ML) and deep learning (DL) as part of AI have been widely used to extract knowledge from data in genomics, proteomics, and other biological fields [1, 2]. Many bioinformatics tasks, such as protein function prediction [3], genome engineering [4], antibiotic discovery [5], and phylogenetic inference [6], have employed DL with major and minor successes [7]. The recent success of AlphaFold2 [8] capable of predicting protein structures that are on par with experimental measures has marked an important milestone of AI in bioinformatics. Due to its capacity to harness the massive amount of genomic data available, DL has gained popularity in population genetic inference where it can provide fast and accurate predictions [9]. An important feature of successful DL applications is its ability to recognize hidden patterns in large volume of data that traditional approaches are unable to uncover.

Horizontal Gene Transfer (HGT) is a major evolutionary force in bacteria and refers to the exchange of genetic material between a “donor” and a “recipient” organism [10, 11]. Bacteria evolve not only through vertical inheritance of genetic material but also through HGT [10]. The transferred genetic materials typically form syntenic blocks called genomic islands (GIs) [12]. Genetic material can be mobilized between organisms primarily through three mechanisms: Transformation, conjugation, and transduction, with additional facilitators

such as outer membrane vesicles (OMVs) [13, 14], virus-like particles [15], and phage-like particles [16]. The prevalence and high rate of HGT have been demonstrated by several studies. Examination of 7781 isolated genomes derived from the gut microbiomes of 48 individuals from 15 distinct populations reveals that 90% of these genomes participate in HGT [17] and the analysis of 827 isolate genomes of *Enterobacteriaceae* family from 14 livestock farms identified up to 2364 potential HGT events [18].

Among other things, HGT is of prime interest for its contribution to the widespread dissemination of antibiotic resistance genes (ARGs) [19, 20] and pathogenic determinants [21], which play a key role in the development of antimicrobial resistance (AMR) [22–25]. The emergence of AMR poses a major threat to global health and its wider implications present us with a growing public health crisis [26–29]. Although most HGT events are initially mostly neutral, with the transferred material then becoming domesticated [30], their adaptation can also confer multiple functionalities, such as secondary metabolism [31] or adaptation to an extreme environment [32, 33]. Deciphering HGT is therefore important to limit the future transfer of AMR and to address other issues related to public health, biotechnology, and environmental sustainability [34, 35].

Detection of HGT events is challenged by the mixing of different organisms and mechanisms at the genome level. Multiple organisms are involved in HGT, resulting in a mosaic structure [36]. In addition to the transfer mechanism, selective pressure, functional compatibility, and phylogenetic related-

Received: October 11, 2024. Revised: December 26, 2024. Editorial Decision: January 6, 2025. Accepted: February 4, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other

permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

ness contribute to the variation of HGT [37–40]. Moreover, genetic material has the propensity for other evolutionary processes such as single-nucleotide changes, recombination, duplications, and deletions [41]. A recent study also showed that bacteria often carry multiple mobile genetic elements (MGEs) of different types whose interactions affect the patterns of HGT [42]. These driving factors result in an extreme variation and a limited understanding of HGT, which impedes the discovery of all HGT events.

Over the past decades, numerous computational approaches have been developed to detect HGT and the number has been increasing as shown in Fig. 1; thanks to the surge of available sequencing data. Existing computational approaches have been diversified, focusing on particular facets of HGT, but most of them are powered by statistical analysis combined with scalable algorithms. Motivated by the progress made by AI in other bioinformatics applications and more available sequenced genomes, we believe that AI, especially DL, holds the power to recognize the hidden patterns of HGT as exemplified by recent studies [43–45], and therefore we fathom the adoption of AI in HGT detection. Several studies have already reviewed computational approaches for HGT detection [11,46–55]. However, the reviews did not present the trend of existing computational approaches and specifically investigated the use of AI in HGT detection. To this end, we conducted a comprehensive review built on previous reviews and extended them by collecting several computational approaches to HGT detection since the early 2000s. We then organized the computational approaches into four computational groups (AI-based, sequence composition, comparative genomics, and hybrid) and provided the trend of each computational group along with the adoption of AI. Finally, we discussed the challenges of HGT detection and potential limitations for adopting AI, as well as future directions of HGT detection.

Materials and methods

The papers were collected by referring to previous reviews and the list of papers was extended by performing literature searches on Google Scholar and Scopus databases with the following keywords: “Horizontal Gene Transfer Detection,” “Lateral Gene Transfer Detection,” and “Genomic Island Detection.” We selected papers from 1 January 2000 to 31 December 2023 and only papers developing a computational approach were included in this work. The papers were categorized into computational groups by listing the processes involved in each approach (refer to [Supplementary Table S1](#)) and following the categorization concepts introduced by previous reviews [11,46–55].

Results

The results are presented in three sections. First, we review existing computational approaches for the detection of HGT and illustrate the division of computational approaches into four computational groups. To make it easier for the reader to choose among all the tools, we select the most successful ones for each computational group (according to the number of citations) and briefly explain their methods. Second, we present the current trend of existing computational approaches. Lastly, we discuss the lack of reliable validation data sets

Computational approaches for HGT detection

Understanding different mechanisms of HGT is important because each mechanism can leave different signals on the sequence [56]. Extensive literature studies on HGT have shed light on demystifying the nature of HGT by providing a list of typical factors involved in HGT and the consequences of HGT [40]. These findings have driven researchers to develop computational approaches tailored to factors associated with HGT, such as plasmids and ARGs. Although plasmids are a very important source of HGT and ARGs are a special class of genes closely related with HGT, we decided to exclude them for two reasons. First, the identification of plasmids itself is insufficient to confirm HGT. While plasmids are well-known for facilitating HGT, they do not always carry and actively transfer genes [57]. Second, the presence of ARGs alone does not confirm HGT and a more comprehensive analysis is needed [58, 59]. Therefore, computational approaches that focus on classifying or annotating plasmids or ARGs are not included, e.g. MLPlasmids [60] and DeepARG [61]. This review focuses on computational approaches for the detection and localization of HGT events.

Previous reviews have provided a wide array of computational approaches which fall into two widely recognized groups, namely composition-based and comparative genomics approaches [46,48–53]. Composition-based approaches capitalize on the alteration of the composition of a genome following the HGT, whereas comparative genomics approaches are driven by the consequences of HGT, such as phylogenetic incongruence and changes in synteny. Depending on the data processing strategy, computational approaches can also be categorized into window-based, windowless, bottom-up, and top-down approaches [50,53,62]. Window-based approaches use sliding window techniques on the genomic sequence to locate the alteration of composition, whereas windowless approaches apply statistical approaches, e.g. t-test, to detect composition bias. Bottom-up approaches typically identify only a few genes as sufficiently unusual to be considered foreign, leading to the prediction of many small fragments as part of GIs. To address the limitations of bottom-up methods, a top-down approach was introduced, which detects GIs by progressively dividing a genome into smaller regions through a recursive segmentation process [62].

Motivated by the preceding categorization concepts, we organized the computational approaches into a hierarchical structure of computational groups that encompasses previously established groups, as summarized in Table 1. This organization resulted in four primary groups, each of which contains its respective subgroups. These four primary groups are as follows:

- (1) Artificial intelligence-based approaches leverage classical ML and DL.
- (2) Sequence composition approaches aim at identifying composition bias within a sequence, which involves the characterization of components present in the sequence.
- (3) Comparative genomics approaches involve comparing sequences of different organisms to understand similarities, differences, and evolutionary relationships.
- (4) Hybrid approaches consist of any approaches that either combine various approaches into a series of computations or aggregate the outcomes from multiple approaches into the final result.

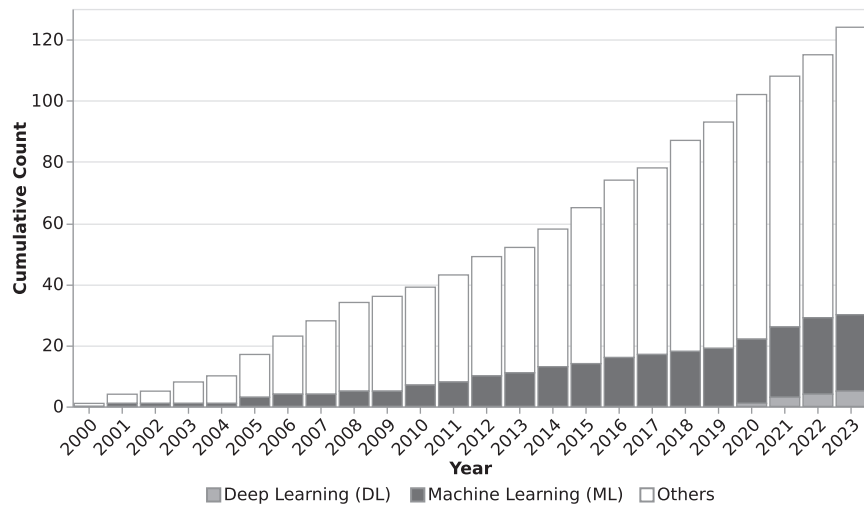


Figure 1. The trend of computational approaches for detection of HGT between 2000 and 2023, showing the application of ML as well as DL in this field.

Table 1. Description of the computational groups to identify HGT

Group	Sub-group	Objective
AI-based	Classical Machine Learning (ML)	leverage classical ML methods, without the use of deep neural networks, to analyze the data
	Deep Learning (DL)	utilize any deep neural networks to learn patterns of HGT
Sequence Composition	Window-based	detect composition bias within a sequence with sliding window techniques
	Windowless	detect composition bias within a sequence using top-down approaches to automatically split the sequence
Comparative Genomics	Alignment	alignment of sequences to infer shared homologous regions, which can be divided into Pairwise Sequence Alignment, Multiple Sequence Alignment, and Whole Genome Alignment, or perform similarity search against a database
	Alignment-free	quantify sequences similarity/dissimilarity based on the characteristics of the sequences
	Phylogenetic	identify incongruencies between gene trees and species trees or analyze phylogenetic profiles from closely and distantly related species
	Read Mapping	map reads (short sequences) to reference sequences to spot structural variants in the mapping coverage
	Synteny Analysis	identify conserved regions between sequences, e.g. gene order retention
Hybrid	Serial	cascade different approaches into a computational pipeline
	Parallel	merge results from multiple approaches into the final output

The sequence data used by the approaches as primary input can vary. Existing computational approaches predominantly require sequenced and assembled isolated genomes, whether draft or complete genomes, but the advent of next-generation sequencing (NGS) technologies facilitates analysis on NGS short reads without the need of assembly [63]; recently, metagenomic data sets were used to study HGT in a microbial community [64]. Figure 2 depicts a visual account of the proportion of different sequence data used across each computational group.

Isolated genomes provide the global genomic context allowing for a comprehensive assessment of HGT events, whereas NGS short reads offer the chance for rapid anal-

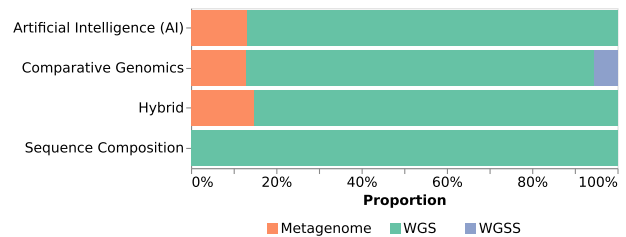


Figure 2. Proportion of different sequence data used in each computational group. Three sequence data are reported: metagenome, whole genome sequence (WGS), and WGSS. WGSS consists of non-assembled short reads of genomes. Metagenome refers to metagenomic data of non-assembled genomes of multiple organisms.

ysis to infer HGT events. Isolated genomes enable genomic signature analysis to identify composition bias, alignment to determine homologous regions between genomes, and phylogenetic analysis for tracing evolutionary relationships that may be indicative of HGT events. NGS short reads are often used in mapping based approaches to infer HGT events. Mapping based approaches aim at identifying structural variations based on alignment to reference genomes. Metagenomic data provide insights into community-level dynamics and allow for comprehensive analyses without prior isolation.

Artificial intelligence-based approaches

Artificial intelligence-based (AI-based) approaches cover any approaches that leverage either classical ML or DL to detect HGT. Table 2 provides a list of five most cited AI-based approaches that have the source code and the data set available. The leading AI-based approaches use supervised learning for a classification task to detect HGT. Although those methods are not aiming at explicitly detecting HGT, they can sort out prophages as the evidence of HGT.

VirFinder [65] trains a logistic regression (LR) to classify viral sequences from contigs of metagenomic data based on k -mer frequencies. It can be further used to identify prophages within large contigs of a genome. Phispy [66] uses a random forest (RF) to classify prophages in bacterial genomes, as a

Table 2. List of top five AI-based approaches for detecting HGT sorted by the number of citations according to Scopus (accessed on 14 March 2024)

Year	Approach	Methodology	Target
2017	VirFinder [65]	ML, classification; LR	Prophages
2012	Phispy [66]	ML, classification; RF	Prophages
2008	RVM [112]	ML, classification; RVM	GIs
2021	Zhou <i>et al.</i> [43]	DL, ML, classification, link prediction; GCN, LR, RF	Genes
2016	MSGIP [67]	ML, clustering; Mean Shift Clustering	GIs

Only approaches that provided data sets and their web or source code are listed for potential reproducibility. Methodology broadly covers tasks and algorithms of each approach.

Notes: genes: foreign genes; DL: Deep Learning; GCN: Graph Convolutional Neural Network; GIs: Genomic Islands; LR: Logistic Regression; ML: classical machine learning; MLP: Multi Layer Perceptron; RF: Random Forest; RVM: Relevance Vector Machine; SVM: Support Vector Machine.

suspected vector of HGT, based on multiple features: protein length, transcription strand direction, customized AT and GC skew, the abundance of unique phage words, phage insertion points and the similarity of phage proteins.

Relevance vector machine (RVM) is one of the pioneers ML approaches for HGT detection. It learns the characteristics of GIs based on features such as interpolated variable order motifs (IVOMs), GC content, gene density, size of GIs, presence of insertion sites, integrase, phage, noncoding RNA, and repeats. Zhou *et al.* [43] employ ML and DL approaches, including LR, RF, and graph convolutional neural network (GCN), to understand the effect of functional gene content, phylogenetic distance, and co-occurrence on HGT. To test the effect of these features, they construct a HGT network from publicly available genome databases. While phylogenetic distance and co-occurrence between organisms could be used to predict HGT events, functional similarity turns out to be the strongest determinant of HGT. Moreover, it is shown that AI-based approaches can be used to extract HGT patterns from large-scale HGT networks. MSGIP [67] utilizes mean shift clustering algorithm to cluster genomic regions with similar nucleotides composition. Clusters exhibiting distinct features from the rest of the genome are considered GIs. By varying windows between 10 and 200 kb, MSGIP can capture GIs of various lengths.

An up-and-coming approach worthy of consideration is geNomad, a recent DL approach that classifies MGEs [68]. It combines two classifiers based on raw nucleotide sequences and their gene content, respectively, for identifying sequences of plasmids and viruses and has been tested to detect prophages with high precision. The classifier for raw nucleotide sequences employs an encoder containing convolutional neural networks (CNNs) and self-attention, whereas the classifier for the gene content utilizes a decision tree with ensemble learning.

Sequence composition approaches

Sequence composition approaches use the fact that the sequence composition varies across species as a result of different environmental factors. Thus, the presence of a compositional bias within a genome can be indicative of transfer from an other organism. Compositional bias is assessed by segmenting a genome into multiple regions, which may consist of either nucleotides or contiguous genes; any segments with compositional bias are labeled as GIs or foreign genes. A genome

Table 3. List of top five sequence composition approaches for detecting HGT sorted by the number of citations according to Scopus (accessed on 14 March 2024)

Year	Approach	Methodology	Target
2006	AlienHunter [73]	Window-based; IVOM	GIs
2006	SIGI-HMM [74]	Window-based; HMM	GIs
2018	MTGIpick [62]	Window-based; <i>t</i> -test, MJSD, MSA	GIs
2017	Zisland-Explorer [77]	Windowless; GC-Profile	GIs
2020	2SigFinder [75]	Window-based; <i>t</i> -test, MJSD	GIs

Only approaches that provided data set and their web or source code are listed for potential reproducibility. Methodology broadly covers tasks and algorithms of each approach.

Note: GIs: Genomic Islands; HMM: Hidden Markov Model; IVOM: Interpolated Variable Order Motifs; MJSD: Markovian Jensen-Shannon Divergence; MSA: Multiscale Segmentation Algorithm.

can be segmented by applying a fixed-length sliding window or any windowless approach to find the breakpoints. Various compositional attributes—GC content, codon usage, or *k*-mer frequencies—are used to analyze compositional bias [69–72]. Table 3 lists five most cited approaches that provide source code and data sets.

Alien Hunter [73] introduced a novel computational approach, interpolated variable order motifs (IVOMs), to handle *k*-mers differently, where high-order *k*-mers are prioritized more than low-order *k*-mers because the former are considered to contain more information than the latter. SIGI-HMM [74] uses codon usage bias with Hidden Markov Model (HMM) to identify GIs.

MTGIpick [62] and 2SigFinder [75] rely on Markovian Jensen–Shannon Divergence (MJSD) to determine the boundaries of GIs [76]. It is a windowless approach that recursively divides a genome into smaller genomic regions based on the score of each nucleotide. MTGIpick and 2SigFinder combine small- and large-scale statistical testing to identify GIs. ZislandExplorer [77] exploits GC divergence to separate potential GIs from the core genome and assigns a score, based on codon usage bias, to the potential GIs to determine primary GIs candidates.

In addition to statistical analysis, classical ML approaches have been used to identify genomic regions with compositional bias. For instance, Centroid [78] and Wn-SVM [79] use clustering and one-class SVM, respectively, to find GIs. However, due to their use of ML, our classification methodology places them under the umbrella of AI-based approaches.

Comparative genomics approaches

Comparative genomics approaches detect HGT based on the sporadic phylogenetic distribution of transferred genomic regions [50,53]. Comparative genomics approaches essentially involve comparing genomes to study their relationships. Depending on the goal of a study, comparative genomics approaches offer a wide range of tasks, such as alignment, alignment-free, phylogenetic, read mapping, and synteny analysis. A list of comparative genomics approaches with available source code and data sets is provided in Table 4.

Alignment is carried out to determine homologous regions between genomes that may indicate functional, structural, or evolutionary relationships [80]. Alignment tools, such as BLAST [81] for pairwise alignment and MAUVE [82] for mul-

Table 4. List of top five comparative genomics approaches for detecting HGT sorted by the number of citations according to Scopus (14 March 2024)

Year	Approach	Methodology	Target
2016	Phaster [119]	Alignment; BLAST	Prophages
2006	Phage_Finder [120]	Alignment; HMM, BLASTP, tRNA/tmRNA detection*	Prophages
2008	IslandPick [84]	Alignment, implicit phylogenetic; CVTree, Mauve, BLAST	GIs
2021	MobileElementFinder [83]	Alignment; BLAST	MGEs
2018	RANGER-DTL 2.0 [89]	Explicit phylogenetic; MPR	Genes

Only approaches provided data set and their web or source code are listed for potential reproducibility. Methodology broadly covers tasks and algorithms of each approach.

Notes: genes: foreign genes; DTL: Duplication, Transfer, and Loss; GIs: Genomic Islands; HMM: Hidden Markov Model; MPR: Maximum Parsimony Reconciliation.

*Detection tools used: tRNAscan-SE and Aragorn.

multiple sequence alignment, are often used in comparative genomics approaches. MobileElementFinder [83] uses BLAST search against a reference database of MGEs to find MGEs in *Salmonella enterica* genomes. IslandPick finds GIs by using MAUVE on a database of reference genomes with negative and positive examples [84].

Alignment-free approaches emerged to overcome the limitation of alignment approaches, which is genetic recombination and shuffling [85]. The other advantage is that those methods are much faster than alignment approaches. Alignment-free approaches can be broadly classified into word count based and match length based [86]. Term Frequency-Inverse Document Frequency (TF-IDF), a widely used method in NLP for text document analysis, was adopted in an alignment-free approach to measure the relevance of a genomic region to a genome based on word count [87].

Phylogeny based approaches are divided into two types: implicit, by building phylogenetic profiles from closely and distantly related species, or explicit, by detecting inconsistencies between gene trees and species trees [49]. The lineage probability index (LPI) was introduced to measure the likelihood of a gene coming from HGT based on the gene distribution of closely and distantly related species [88]. Explicit phylogenetic analysis involves tree reconstruction. MetaCHIP [64] integrates Ranger-DTL 2.0 [89], an efficient algorithm for tree reconciliation, to refine the results from its alignment approach and to provide information on the direction of gene flow.

Synteny analysis measures the conservation of genomic regions between genomes and typically works well to identify HGT between closely related species. Closely related species share most of their genomic regions; differences in the genomic regions between them may indicate that HGT has occurred. The synteny index (SI), a score to measure the evolutionary distance between a pair of genomes, is used to detect HGT [90–92].

A related problem is the detection of HGT directly from whole genome shotgun sequence (WGSS) data. Under the assumption that there is only one donor and one recipient genome, most short reads of a recipient genome will align to the reference of the recipient genome and the remaining short

Table 5. List of top five hybrid approaches for detecting HGT sorted by the number of citations according to Scopus (accessed on 14 March 2024)

Year	Approach	Methodology	Target
2017	IslandViewer4 [101]	Window-based, alignment, implicit phylogenetic; Mauve, BLAST, CVTree, HMM	GIs
2018	IslandPath-DIMOB [102]	Window-based, alignment; HMM, BLAST	GIs
2015	PAIDB v2.0 [121]	Window-based, alignment; SIGI-HMM, IslandPath-DIMOB	GIs
2022	VRprofile2 [104]	Window-based, alignment, classification; SIGI-HMM, IslandPath-DIMOB, BLAST	GIs, prophages
2016	GIPsy [106]	Window-based, alignment; SIGI-HMM, BLAST, HMM, tRNAscan-SE	GIs

Only approaches that provided data set and their web or source code are listed for potential reproducibility. Methodology broadly covers tasks and algorithms of each approach.

Notes: GIs: Genomic Islands; HMM: Hidden Markov Model;

reads will map to the donor genome. By combining those information about the mapping coverage and reads overlapping recipient and donor genome, one can identify the HGT region [63,93].

Hybrid approaches

Although sequence composition approaches can detect HGT, they are prone to a high rate of false positive and false negative in their results [94]. It is necessary that the transferred genomic region be distinct and long enough to show compositional bias [91]. Special sequences in an organism [88], intra-genomic variations [95], and amelioration [96] would reduce the performance of sequence composition approaches. Moreover, sequence composition approaches alone cannot identify the HGT donor because they are reference-free.

Analogously, comparative genomics approaches also come with drawbacks. These approaches are highly dependent on the availability and quality of the reference genomes. Alignment assumes that similar sequences must be collinear, which is not always true [86], and aligning multiple sequences is time-consuming [97]. Furthermore, inaccuracies in alignment and alignment-free approaches may lead to different results [98]. Other evolutionary processes, such as recombination, gene duplication, and gene loss, may obfuscate phylogenetic analysis [56, 99, 100].

Hybrid approaches combine computational approaches from different groups to compensate for the drawbacks. Combining approaches can be done serially, by cascading them in a pipeline, or in parallel, by aggregating results from different approaches. The performance of hybrid approaches relies on individual approaches and decision rules for integrating predictions [53]. Table 5 lists five approaches with most citations that provide source code and data sets.

IslandViewer4 [101] combines IslandPath-DIMOB [102], SIGI-HMM [74], and two comparative genomics approaches, IslandPick [84] and Islander [103], into an integrated interface for GIs detection. IslandPath-DIMOB predicts GIs based on dinucleotide bias and the presence of a MGE using similarity search against a database of known MGEs. VRprofile2 [104]

incorporates SIGI-HMM and IslandPath-DIMOB with other computational methods in a parallel workflow to identify integrons, prophages, ICEs, gene-coding proteins, and known virulence factors.

SIGI-HMM and IslandPath-DIMOB are also integrated by IslandCompare [105] in a computational workflow followed by a sequence comparison to ensure the consistency of the analysis. For a more comprehensive analysis, IslandCompare provides contextual comparative genome visualization, including functional annotation of antibiotic resistance determinants. SIGI-HMM is also part of GIPsy [106] pipeline to identify GIs. GIPsy [106] detects GIs based on known GI features such as composition bias using SIGI-HMM, presence of transposase genes, flanking transfer RNA (tRNA) genes, and absence in other organisms of the same genus or closely related species. It can also detect GIs with specific functionalities by identifying factors for virulence, metabolism, antibiotic resistance, or symbiosis.

ShadowCaster [107] is a clear example of a hybrid approach. It sequentially combines an AI-based approach and a comparative genomic approach. It uses a one-class SVM classifier to determine genes with biases in codon usage and tetranucleotide frequencies, and then performs an alignment to estimate a Bayesian probability, called the phylogenetic shadow, based on the gene distribution.

Current trend of computational approaches

The trend presented in Fig. 3 shows that the number of computational approaches increases at a rate compatible with an exponential distribution ($R^2 = 0.998$, refer to [Supplementary Fig. S1](#)). We also provided the count of each computational group per year (refer to [Supplementary Table S2](#)) and visualized them with a non-cumulative stacked bar chart (refer to [Supplementary Fig. S2](#)).

The growth in comparative genomics approaches outpaces the other computational groups. Between 2010 and 2023, the number of comparative genomic approaches has increased by over 200%, while the number of sequence composition approaches has increased by ~67%, with a sign of plateauing between 2016 and 2023. In 2010, the number of sequence composition approaches ranked second just below comparative genomics, but it was outnumbered by the number of hybrid and AI-based approaches in 2022. The greatest increase in hybrid approaches occurred in 2022 with five additional approaches. Meanwhile, AI-based approaches, dominated by classical ML algorithms, experienced a major increase in 2021 with four additional approaches. To date, comparative genomics approaches hold the biggest portion of available computational approaches to HGT detection followed by hybrid, AI-based, and sequence composition approaches.

Validation data sets

Many computational approaches are available, but comparing them in a proper benchmark has been challenging due to the lack of reliable validation data sets [44, 50, 84]. Almost every computational approach introduces its own validation data set, making the approaches difficult to compare with each other. Three categories of data sets can be considered: literature, curated, and simulated data sets.

A literature data set is a collection of HGT events discovered in published literature, for example, 51 horizontally transferred genes related to heavy metal resistance were found

in the analysis of the complete genome of *Rhodanobacter denitrificans* 2APBS1 [108], and literature studies identified six prophage gene clusters and five annotated pathogenicity islands in *Pseudomonas aeruginosa* LESB58 genome [109]. A curated data set is constructed by comparative analysis on genomes in reference databases, such as GenBank [110] and RefSeq [111], with the assumption that genes both present in closely and distantly related species are considered horizontally transferred [73, 84]. Vernikos *et al.* created a data set for ML approaches from 37 strains of 3 genera with 331 GIs and 337 non-GIs [112]. Langille *et al.* constructed a data set from 117 strains of 22 genera with 771 GIs and 3700 non-GIs [84]. These GIs contained a total of 11 404 annotated genes with an average of 14.8 genes per GI and 97.5 genes per strain. Finally, a simulated data set is crafted by artificially inserting foreign genes into genomes under study where the inserted genes have no orthology with the recipient genome [76, 79, 107, 113–116]. Sanchez *et al.* created three simulated data sets by randomly transferring ten genes from the donor species that had no orthology with the recipient genome [107]. Jani *et al.* simulated HGT by selecting one organism as a recipient and 2 organisms as donors, then inserting 12 segments of size 30, 50, and 80 kbp from each donor into the recipient genomes [116].

Baneerjee *et al.* collected available validation data sets, including curated and literature data sets, to train and evaluate their AI-based approach HGT detection [45]. The proposed AI-based approach compared to five other HGT detection approaches, namely IslandViewer4 [101], IslandPath-DIMOB [102], SIGI-HMM [74], Islander [103], and AlienHunter [73]. Zhou *et al.* conducted a large-scale analysis of highly conserved HGT events between distantly related organisms, resulting in a network of 147 889 HGT events among 6566 genomes. The resulting network is sparse, suggesting the role of selective pressure on conserved HGT and that only certain genetic material can be transferred horizontally between two given organisms [43].

Discussion

In this work, we show that there has been a remarkable growth in the number of computational approaches to HGT detection. There are a variety of computational approaches with different computational methods, which can be categorized into the computational groups mentioned above. Grouping computational approaches is not an easy task, but it helps researchers and young scientists navigate the sea of information in HGT detection. With the exception of the hybrid approaches, the proposed groups are self-contained. In the hybrid group, one could argue that some of them could be placed in the other groups, but as long as an approach incorporates multiple approaches, it belongs to the hybrid approaches.

AI-based approaches have gained attention for their potential to reveal the patterns of HGT, although they are now limited to sequence classification. The recent progress of AI in modeling complex evolutionary processes shows its potential for HGT detection [9]. Sequence composition approaches may provide rapid analysis because they are reference-free but they alone are prone to errors and unable to capture the intricacy of HGT. Therefore, sequence composition approaches are better coupled with other approaches as hybrid approaches. The increase in hybrid approaches, especially in the last two years, is in accordance with a previous review [53] that stated that a combination of methods is often more suitable and desirable

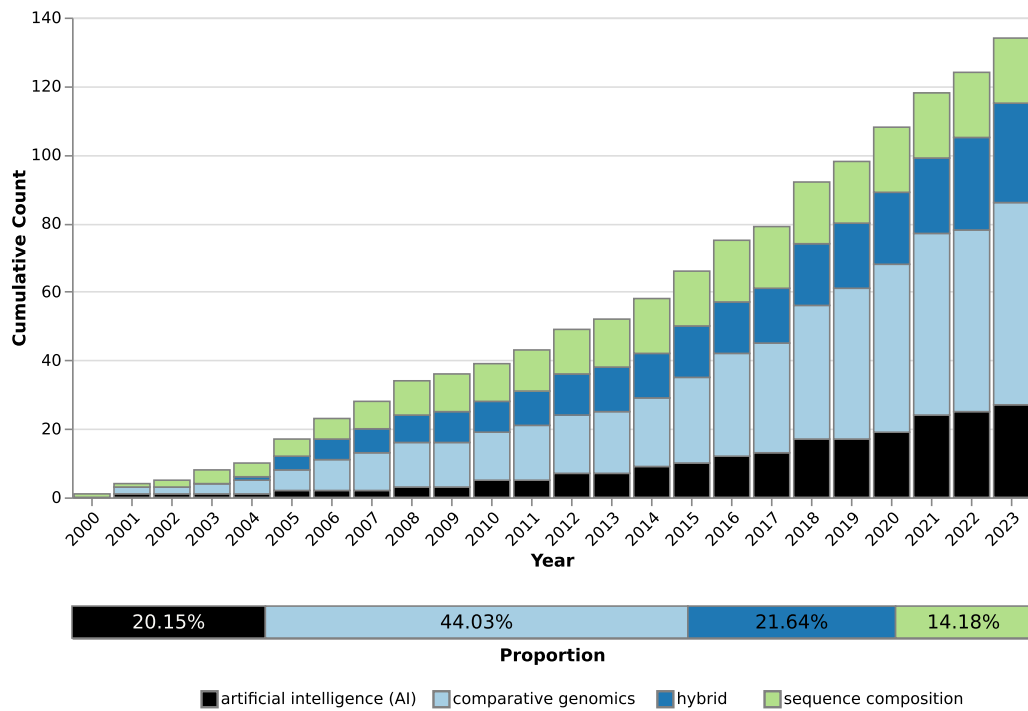


Figure 3. The trend of computational approaches between 2000 and 2023. In general, the number of computational approaches shows an exponential increasing trend between 2000 and 2023 with comparative genomics outpacing the growth rate of other computational groups. AI-based and hybrid approaches show a progressive increase over the years outnumbering sequence composition approaches.

for HGT detection. As a recent example, De *et al.* combined different approaches into an integrative approach, DICEP, to identify GIs [117]. In conjunction with AI-based approaches, hybrid approaches could provide a comprehensive and robust means of detecting HGT by leveraging the strength of different computational groups. Nevertheless, comparative genomics approaches are still more prevalent owing to the proliferation of the registered genomes in reference databases. Despite the drawbacks, they can provide a more comprehensive analysis involving multiple genomes.

With the exponential growth in computational approaches and the proliferation of sequencing data, many important questions related to HGT can be addressed. However, most available computational approaches focus on detecting HGT in WGS and only accept single genomes for analysis. Computational approaches capable of analyzing multiple genomes and genomes sequenced using different technologies simultaneously are needed as the amount of available sequenced genomes continues to accumulate exponentially [53].

Despite the potential of AI in this field, its widespread adoption remains constrained, presenting both opportunities and challenges that warrant discussion. A critical barrier to the broader use of AI in bioinformatics, including HGT detection, is the lack of training and validation data sets. To the best of our knowledge, no standardized training data sets exist for AI implementation in HGT detection. The reliability of the available validation data sets has not been verified by convincing biological evidence [50]; moreover, the validation data sets are often limited to a few genomes, which do not cover as much microbial diversity [53]. Accurate evaluation and validation of computational approaches are based heavily on high-quality data sets. Nevertheless, there is a large collection of genomes

curated by Banerjee *et al.* [45] for training an AI-based approach that is worth considering.

To address these challenges and advance the field, concerted efforts are imperative. Collaborative initiatives within the scientific community are needed to develop and curate comprehensive benchmark data sets that accurately reflect the complexity of HGT events. More precisely, Brito pointed out potential pitfalls that may have irretrievably affected the understanding of a complete or perfect picture of HGT in natural communities [11]. This collaborative approach will facilitate the development of AI-based approaches, thereby promoting innovative computational approaches for HGT detection.

While this review provides useful findings, there are some limitations. First, we selected approaches that directly answer the question of whether or where HGT has occurred which may leave out approaches related to HGT detection, for instance, approaches that identify factors associated with HGT. These factors may indicate HGT but further analysis is required for confirmation, except the presence of prophages and MGEs as evidence of HGT. Due to the focus of prophages and GIs, not all HGT events are detected by the presented approaches, *e.g.* HGT events via plasmids are missed. Second, there might be arguments over the correlation between the presence of MGEs and HGT. Some studies consider MGEs as the mediator of HGT [118], whereas some other studies see MGEs as parts of GIs [12]. In this review, we decided to consider MGEs as parts of GIs meaning if MGEs are present in a genome, then HGT has most likely occurred. Third, we only discussed 5 examples per computational group based on the number of citations. This might introduce bias to older approaches. However, we only selected those that are well-maintained and also noticed that the older approaches listed

in this review are still used in newer approaches, for instance, IslandPath-DIMOB [102] and SIGI-HMM [74] are used in a newer approach developed in 2022 [105]. More approaches are listed in the [Supplementary Table S1](#).

Conclusion

In the study, we reviewed over 100 computational approaches to HGT detection. We organized these approaches into a hierarchical structure with four main groups (AI-based, sequence composition, comparative genomics, and hybrid) and showed how the number of computational approaches has increased since the early 2000s. Of the many computational approaches available, only a handful withstand the test of time. The enduring relevance of SIGI-HMM [74], published in 2006, and MJSD [76], first introduced in 2009, is evidenced by their continued use in recent approaches, which may explain the stagnation experienced by sequence composition approaches. Moving forward, integrated interfaces for detecting GIs incorporating different approaches and features like visualization, for example, IslandViewer4 [101], are preferable for comprehensive results and easier analysis. We also discuss the adoption of AI in HGT detection and how the scientific community can remove the barrier to exploring AI. While the significant growth of computational approaches for HGT detection highlights remarkable progress, the limited amount of validation data sets impedes the adoption of AI, and under-utilization of DL methods that directly tackle HGT detection problems underscores the need for collaborative efforts, innovation, and resource development to overcome these barriers. Embracing AI and leveraging its capabilities in conjunction with comprehensive data sets could unlock opportunities to advance our understanding of HGT dynamics.

Acknowledgements

The authors would like to express their gratitude to Jens Stoye for his helpful comments and constructive feedback.

Author contributions: A.J.W.: Conceptualization, data curation, investigation, validation, visualization, writing—original draft. A.A.: Visualization, writing—review and editing. H.R.: Supervision, writing—review and editing. G.H.: Supervision, writing—review and editing

Supplementary data

[Supplementary data](#) is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

No external funding.

Data availability

All the data and code to produce the graphs are available in <https://github.com/andrejw27/Current-State-and-Future-Prospects-of-Horizontal-Gene-Transfer-Detection> and <https://doi.org/10.5281/zenodo.14625185>.

References

- Larrañaga P, Calvo B, Santana R *et al.* Machine learning in bioinformatics. *Brief Bioinform* 2006;7:86–112. <https://doi.org/10.1093/bib/bbk007>
- Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform* 2017;18:851–69.
- Kulmanov M, Khan MA, Hoehndorf R *et al.* DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* 2018;34:660–8. <https://doi.org/10.1093/bioinformatics/btx624>
- Chuai G, Ma H, Yan J *et al.* DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol* 2018;19:80. <https://doi.org/10.1186/s13059-018-1459-4>
- Wong F, Zheng EJ, Valeri JA *et al.* Discovery of a structural class of antibiotics with explainable deep learning. *Nature* 2024; 626:177–85. <https://doi.org/10.1038/s41586-023-06887-8>
- Suvorov A, Hochuli J, Schrider DR. Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Syst Biol* 2020;69:221–3. <https://doi.org/10.1093/sysbio/syz060>
- Sapoval N, Aghazadeh A, Nute MG *et al.* Current progress and open challenges for applying deep learning across the biosciences. *Nat Commun* 2022;13:1728. <https://doi.org/10.1038/s41467-022-29268-7>
- Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583–9. <https://doi.org/10.1038/s41586-021-03819-2>
- Huang X, Rymbekova A, Dolgova O *et al.* Harnessing deep learning for population genetic inference. *Nat Rev Genet* 2024; 25:61–78. <https://doi.org/10.1038/s41576-023-00636-3>
- Thomas CM, Nielsen KM. Mechanisms of, and barriers to, Horizontal Gene Transfer between bacteria. *Nat Rev Microbiol* 2005;3:711–21. <https://doi.org/10.1038/nrmicro1234>
- Brito IL. Examining Horizontal Gene Transfer in microbial communities. *Nat Rev Microbiol* 2021;19:442–53. <https://doi.org/10.1038/s41579-021-00534-7>
- Juhas M, Van Der Meer JR, Gaillard M *et al.* Genomic islands: tools of bacterial Horizontal Gene Transfer and evolution. *FEMS Microbiol Rev* 2009;33:376–93. <https://doi.org/10.1111/j.1574-6976.2008.00136.x>
- Fulsundar S, Harms K, Flaten GE *et al.* Gene transfer potential of outer membrane vesicles of *Acinetobacter baylyi* and effects of stress on vesiculation. *Appl Environ Microbiol* 2014;80:3469–83. <https://doi.org/10.1128/AEM.04248-13>
- Fischer S, Cornils K, Speiseder T *et al.* Indication of horizontal DNA gene transfer by extracellular vesicles. *PLoS One* 2016;11:e0163665. <https://doi.org/10.1371/journal.pone.0163665>
- McDaniel LD, Young E, Delaney J *et al.* High frequency of Horizontal Gene Transfer in the oceans. *Science* 2010;330:50. <https://doi.org/10.1126/science.1192243>
- Bárdy P, Füzik T, Hrebík D *et al.* Structure and mechanism of DNA delivery of a gene transfer agent. *Nat Commun* 2020;11:3034. <https://doi.org/10.1038/s41467-020-16669-9>
- Groussin M, Poyet M, Sistiaga A *et al.* Elevated rates of Horizontal Gene Transfer in the industrialized human microbiome. *Cell* 2021;184:2053–67. <https://doi.org/10.1016/j.cell.2021.02.052>
- Shaw LP, Chau KK, Kavanagh J *et al.* Niche and local geography shape the pangenome of wastewater-and livestock-associated Enterobacteriaceae. *Sci Adv* 2021;7:eabe3868. <https://doi.org/10.1126/sciadv.abe3868>
- Powell M. Antimicrobial resistance in *Haemophilus influenzae*. *J Med Microbiol* 1988;27:81–7. <https://doi.org/10.1099/00222615-27-2-81>
- Maree M, Thi Nguyen LT, Ohniwa RL *et al.* Natural transformation allows transfer of SCC mec-mediated methicillin resistance in *Staphylococcus aureus* biofilms. *Nat Commun* 2022;13:2477. <https://doi.org/10.1038/s41467-022-29877-2>

21. Hacker J, Bender L, Ott M *et al.* Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various extra intestinal *Escherichia coli* isolates. *Microb Pathog* 1990;8:213–25. [https://doi.org/10.1016/0882-4010\(90\)90048-U](https://doi.org/10.1016/0882-4010(90)90048-U)
22. De la Cruz F, Davies J. Horizontal Gene Transfer and the origin of species: lessons from bacteria. *Trends Microbiol* 2000;8:128–33. [https://doi.org/10.1016/S0966-842X\(00\)01703-0](https://doi.org/10.1016/S0966-842X(00)01703-0)
23. Botelho J, Schulenburg H. The role of integrative and conjugative elements in antibiotic resistance evolution. *Trends Microbiol* 2021;29:8–18. <https://doi.org/10.1016/j.tim.2020.05.011>
24. Tao S, Chen H, Li N *et al.* The spread of antibiotic resistance genes *in vivo* model. *Can J Infect Dis Med Microbiol* 2022;2022:3348695. <https://doi.org/10.1155/2022/3348695>
25. Djordjevic SP, Jarocki VM, Seemann T *et al.* Genomic surveillance for antimicrobial resistance—a One Health perspective. *Nat Rev Genet* 2024;25:142–57. <https://doi.org/10.1038/s41576-023-00649-y>
26. Murray CJL, Ikuta KS, Sharara F *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* 2022;399:629–55. [https://doi.org/10.1016/S0140-6736\(21\)02724-0](https://doi.org/10.1016/S0140-6736(21)02724-0)
27. Munk P, Brinch C, Møller FD *et al.* Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. *Nat Commun* 2022;13:7251. <https://doi.org/10.1038/s41467-022-34312-7>
28. O'Neill J. Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. In: *Review on Antimicrobial Resistance*. London: Wellcome Trust, 2016.
29. Bengtsson-Palme J, Kristiansson E, Larsson DGJ. Environmental factors influencing the development and spread of antibiotic resistance. *FEMS Microbiol Rev* 2018;42:fux053. <https://doi.org/10.1093/femsre/fux053>
30. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. *Nat Rev Genet* 2015;16:472–482. <https://doi.org/10.1038/nrg3962>
31. Sullivan JT, Trzebiatowski JR, Cruickshank RW *et al.* Comparative sequence analysis of the symbiosis island of *Mesorhizobium loti* strain R7A. *J Bacteriol* 2002;184:3086–95. <https://doi.org/10.1128/JB.184.11.3086-3095.2002>
32. Larbig KD, Christmann A, Johann A *et al.* Gene islands integrated into tRNAGly genes confer genome diversity on a *Pseudomonas aeruginosa* clone. *J Bacteriol* 2002;184:6665–80. <https://doi.org/10.1128/JB.184.23.6665-6680.2002>
33. Navarro CA, Von Bernath D, Jerez CA. Heavy metal resistance strategies of acidophilic bacteria and their acquisition: importance for biomining and bioremediation. *Biol Res* 2013;46:363–71. <https://doi.org/10.4067/S0716-97602013000400008>
34. Sun D. Pull in and push out: mechanisms of Horizontal Gene Transfer in bacteria. *Front Microbiol* 2018;9:2154. <https://doi.org/10.3389/fmicb.2018.02154>
35. Ebmeyer S, Kristiansson E, Larsson DGJ. A framework for identifying the recent origins of mobile antibiotic resistance genes. *Commun Biol* 2021;4:8. <https://doi.org/10.1038/s42003-020-01545-5>
36. Smith JM. Analyzing the mosaic structure of genes. *J Mol Evol* 1992;34:126–9. <https://doi.org/10.1007/BF00182389>
37. Oliveira PH, Touchon M, Cury J *et al.* The chromosomal organization of Horizontal Gene Transfer in bacteria. *Nat Commun* 2017;8:841. <https://doi.org/10.1038/s41467-017-00808-w>
38. Porse A, Schou TS, Munck C *et al.* Biochemical mechanisms determine the functional compatibility of heterologous genes. *Nat Commun* 2018;9:522. <https://doi.org/10.1038/s41467-018-02944-3>
39. Kintsés B, Méhi O, Ari E *et al.* Phylogenetic barriers to horizontal transfer of antimicrobial peptide resistance genes in the human gut microbiota. *Nat Microbiol* 2019;4:447–58. <https://doi.org/10.1038/s41564-018-0313-5>
40. Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol* 2022;20:206–18. <https://doi.org/10.1038/s41579-021-00650-4>
41. Szpara ML, Van Doorslaer K. Mechanisms of DNA virus evolution. *Enc Virol* 2021;1-5:71–8. <https://doi.org/10.1016/B978-0-12-809633-8.20993-X>
42. Horne T, Orr VT, Hall JP. How do interactions between mobile genetic elements affect Horizontal Gene Transfer?. *Curr Opin Microbiol* 2023;73:102282. <https://doi.org/10.1016/j.mib.2023.102282>
43. Zhou H, Beltrán JF, Brito IL. Functions predict Horizontal Gene Transfer and the emergence of antibiotic resistance. *Sci Adv* 2021;7:eabj5056. <https://doi.org/10.1126/sciadv.abj5056>
44. Assaf R, Xia F, Stevens R. Identifying genomic islands with deep neural networks. *BMC Genomics* 2021;22:281. <https://doi.org/10.1186/s12864-021-07575-5>
45. Banerjee P, Eulenstein O, Friedberg I. Discovering genomic islands in unannotated bacterial genomes using sequence embedding. *Bioinform Adv* 2024;4:vbae089. <https://doi.org/10.1093/bioadv/vbae089>
46. Langille MGI, Hsiao WWL, Brinkman FSL. Detecting genomic islands using bioinformatics approaches. *Nat Rev Microbiol* 2010;8:373–82. <https://doi.org/10.1038/nrmicro2350>
47. Carvalho MOd, Loreto ELS. Methods for detection of horizontal transfer of transposable elements in complete genomes. *Genet Mol Biol* 2012;35:1078–84. <https://doi.org/10.1590/S1415-47572012000600024>
48. Che D, Hasan MS, Chen B. Identifying pathogenicity islands in bacterial pathogenomics using computational approaches. *Pathogens* 2014;3:36–56. <https://doi.org/10.3390/pathogens3010036>
49. Ravenhall M, Škunca N, Lassalle F *et al.* Inferring Horizontal Gene Transfer. *PLoS Comput Biol* 2015;11:e1004095. <https://doi.org/10.1371/journal.pcbi.1004095>
50. Lu B, Leong HW. Computational methods for predicting genomic islands in microbial genomes. *Comput Struct Biotechnol J* 2016;14:200–6. <https://doi.org/10.1016/j.csbj.2016.05.001>
51. Soares SdC, Oliveira LdC, Jaiswal AK *et al.* Genomic islands: an overview of current software tools and future improvements. *J Integr Bioinform* 2016;13:82–9. <https://doi.org/10.1515/jib-2016-301>
52. da Silva Filho AC, Raittz RT, Guizelini D *et al.* Comparative analysis of genomic island prediction tools. *Front Genet* 2018;9:619. <https://doi.org/10.3389/fgene.2018.00619>
53. Bertelli C, Tilley KE, Brinkman FSL. Microbial genomic island discovery, visualization and analysis. *Brief Bioinform* 2019;20:1685–98. <https://doi.org/10.1093/bib/bby042>
54. Douglas GM, Langille MGI. Current and promising approaches to identify Horizontal Gene Transfer events in metagenomes. *Genome Biol Evol* 2019;11:2750–66. <https://doi.org/10.1093/gbe/evz184>
55. Shikov AE, Malovichko YV, Nizhnikov AA *et al.* Current methods for recombination detection in bacteria. *Int J Mol Sci* 2022;23:6257. <https://doi.org/10.3390/ijms23116257>
56. Zaneveld JR, Nemergut DR, Knight R. Are all Horizontal Gene Transfers created equal? Prospects for mechanism-based studies of HGT patterns. *Microbiology* 2008;154:1–15. <https://doi.org/10.1099/mic.0.2007/011833-0>
57. Dewan I, Uecker H. A mathematician's guide to plasmids: an introduction to plasmid biology for modellers. *Microbiology* 2023;169:001362. <https://doi.org/10.1099/mic.0.001362>
58. Liu F, Luo Y, Xu T *et al.* Current examining methods and mathematical models of horizontal transfer of antibiotic resistance genes in the environment. *Front Microbiol* 2024;15:1371388. <https://doi.org/10.3389/fmicb.2024.1371388>
59. Brown CL, Maile-Moskowitz A, Lopatkin AJ *et al.* Selection and Horizontal Gene Transfer underlie microdiversity-level

- heterogeneity in resistance gene fate during wastewater treatment. *Nat Commun* 2024;15:5412.
60. Arredondo-Alonso S, Rogers MRC, Braat JC *et al.* mlplasmids: a user-friendly tool to predict plasmid-and chromosome-derived sequences for single species. *Microb Genom* 2018;4:e000224. <https://doi.org/10.1099/mgen.0.000224>
 61. Arango-Argoty G, Garner E, Pruden A *et al.* DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 2018;6:23. <https://doi.org/10.1186/s40168-018-0401-z>
 62. Dai Q, Bao C, Hai Y *et al.* MTGipick allows robust identification of genomic islands from a single genome. *Brief Bioinform* 2018;19:361–73.
 63. Trappe K, Marschall T, Renard BY. Detecting Horizontal Gene Transfer by mapping sequencing reads across species boundaries. *Bioinformatics* 2016;32:i595–604. <https://doi.org/10.1093/bioinformatics/btw423>
 64. Song W, Wemheuer B, Zhang S *et al.* MetaCHIP: community-level Horizontal Gene Transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome* 2019;7:36. <https://doi.org/10.1186/s40168-019-0649-y>
 65. Ren J, Ahlgren NA, Lu YY *et al.* VirFinder: a novel *k*-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 2017;5:69. <https://doi.org/10.1186/s40168-017-0283-5>
 66. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic Acids Res* 2012;40:e126. <https://doi.org/10.1093/nar/gks406>
 67. de Brito DM, Maracaja-Coutinho V, de Farias ST *et al.* A novel method to predict genomic islands based on mean shift clustering algorithm. *PLoS One* 2016;11:e0146352. <https://doi.org/10.1371/journal.pone.0146352>
 68. Camargo AP, Roux S, Schulz F *et al.* Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 2024;42:1303–12.
 69. Garcia-Vallvé S, Romeu A, Palau J. Horizontal Gene Transfer in bacterial and archaeal complete genomes. *Genome Res* 2000;10:1719–25. <https://doi.org/10.1101/gr.130000>
 70. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. *Nature* 2000;405:299–304. <https://doi.org/10.1038/35012500>
 71. Nakamura Y, Itoh T, Matsuda H *et al.* Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* 2004;36:760–6. <https://doi.org/10.1038/ng1381>
 72. Hooper SD, Berg OG. Detection of genes with atypical nucleotide sequence in microbial genomes. *J Mol Evol* 2002;54:365–75. <https://doi.org/10.1007/s00239-001-0051-8>
 73. Vernikos GS, Parkhill J. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* 2006;22:2196–203. <https://doi.org/10.1093/bioinformatics/btl369>
 74. Waack S, Keller O, Asper R *et al.* Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinform* 2006;7:142. <https://doi.org/10.1186/1471-2105-7-142>
 75. Kong R, Xu X, Liu X *et al.* 2SigFinder: the combined use of small-scale and large-scale statistical testing for genomic island detection from a single genome. *BMC Bioinform* 2020;21:159. <https://doi.org/10.1186/s12859-020-3501-2>
 76. Arvey AJ, Azad RK, Raval A *et al.* Detection of genomic islands via segmental genome heterogeneity. *Nucleic Acids Res* 2009;37:5255–66. <https://doi.org/10.1093/nar/gkp576>
 77. Wei W, Gao F, Du MZ *et al.* Zisland Explorer: detect genomic islands by combining homogeneity and heterogeneity properties. *Brief Bioinform* 2017;18:357–66.
 78. Rajan I, Aravamuthan S, Mande SS. Identification of compositionally distinct regions in genomes using the centroid method. *Bioinformatics* 2007;23:2672–7. <https://doi.org/10.1093/bioinformatics/btm405>
 79. Tsigirig A, Rigoutsos I. A sensitive, support-vector-machine method for the detection of Horizontal Gene Transfers in viral, archaeal and bacterial genomes. *Nucleic Acids Res* 2005;33:3699–707. <https://doi.org/10.1093/nar/gki660>
 80. Fang X, Wang F, Liu L *et al.* A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nat Mach Intell* 2023;5:1087–96. <https://doi.org/10.1038/s42256-023-00721-6>
 81. Altschul SF, Madden TL, Schäffer AA *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–402. <https://doi.org/10.1093/nar/25.17.3389>
 82. Darling ACE, Mau B, Blattner FR *et al.* Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 2004;14:1394–403. <https://doi.org/10.1101/gr.2289704>
 83. Johansson MHK, Bortolaia V, Tansirichaiya S *et al.* Detection of mobile genetic elements associated with antibiotic resistance in *Salmonella enterica* using a newly developed web tool: MobileElementFinder. *J Antimicrob Chemother* 2021;76:101–9. <https://doi.org/10.1093/jac/dkaa390>
 84. Langille MGI, Hsiao WWL, Brinkman FSL. Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinform* 2008;9:329. <https://doi.org/10.1186/1471-2105-9-329>
 85. Vinga S, Almeida J. Alignment-free sequence comparison—a review. *Bioinformatics* 2003;19:513–23. <https://doi.org/10.1093/bioinformatics/btg005>
 86. Bernard G, Chan CX, Chan YB *et al.* Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Brief Bioinform* 2019;20:426–35. <https://doi.org/10.1093/bib/bbx067>
 87. Cong Y, Chan YB, Ragan MA. A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF. *Sci Rep* 2016;6:30308. <https://doi.org/10.1038/srep30308>
 88. Podell S, Gaasterland T. DarkHorse: a method for genome-wide prediction of Horizontal Gene Transfer. *Genome Biol* 2007;8:R16. <https://doi.org/10.1186/gb-2007-8-2-r16>
 89. Bansal MS, Kellis M, Kordi M *et al.* RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics* 2018;34:3214–6. <https://doi.org/10.1093/bioinformatics/bty314>
 90. Shifman A, Ninyo N, Gophna U *et al.* Phylo SI: a new genome-wide approach for prokaryotic phylogeny. *Nucleic Acids Res* 2014;42:2391–404. <https://doi.org/10.1093/nar/gkt1138>
 91. Adato O, Ninyo N, Gophna U *et al.* Detecting Horizontal Gene Transfer between closely related taxa. *PLoS Comput Biol* 2015;11:e1004408. <https://doi.org/10.1371/journal.pcbi.1004408>
 92. Sevillya G, Adato O, Snir S. Detecting Horizontal Gene Transfer: a probabilistic approach. *BMC Genomics* 2020;21:1–11. <https://doi.org/10.1186/s12864-019-6395-5>
 93. Seiler E, Trappe K, Renard BY. Where did you come from, where did you go: refining metagenomic analysis tools for Horizontal Gene Transfer characterisation. *PLoS Comput Biol* 2019;15:e1007208. <https://doi.org/10.1371/journal.pcbi.1007208>
 94. Friedman R, Ely B. Codon usage methods for Horizontal Gene Transfer Detection generate an abundance of false positive and false negative results. *Curr Microbiol* 2012;65:639–42. <https://doi.org/10.1007/s00284-012-0205-5>
 95. Guindon S, Perrière G. Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. *Mol Biol Evol* 2001;18:1838–40. <https://doi.org/10.1093/oxfordjournals.molbev.a003972>

96. Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 1997;44:383–97. <https://doi.org/10.1007/PL00006158>
97. Zieleszinski A, Vinga S, Almeida J *et al.* Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol* 2017;18:186. <https://doi.org/10.1186/s13059-017-1319-7>
98. Sicheritz-Pontén T, Andersson SGE. A phylogenomic approach to microbial evolution. *Nucleic Acids Res* 2001;29:545–52. <https://doi.org/10.1093/nar/29.2.545>
99. Poptsova MS, Gogarten JP. The power of phylogenetic approaches to detect horizontally transferred genes. *BMC Evol Biol* 2007;7:45. <https://doi.org/10.1186/1471-2148-7-45>
100. Steenwyk JL, Li Y, Zhou X *et al.* Incongruence in the phylogenomics era. *Nat Rev Genet* 2023;24:834–50. <https://doi.org/10.1038/s41576-023-00620-x>
101. Bertelli C, Laird MR, Williams KP *et al.* IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res* 2017;45:W30–5. <https://doi.org/10.1093/nar/gkx343>
102. Bertelli C, Brinkman FSL. Improved genomic island predictions with IslandPath-DIMOB. *Bioinformatics* 2018;34:2161–7. <https://doi.org/10.1093/bioinformatics/bty095>
103. Hudson CM, Lau BY, Williams KP. Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes. *Nucleic Acids Res* 2015;43:D48–53. <https://doi.org/10.1093/nar/gku1072>
104. Wang M, Goh YX, Tai C *et al.* VRprofile2: detection of antibiotic resistance-associated mobilome in bacterial pathogens. *Nucleic Acids Res* 2022;50:W768–73. <https://doi.org/10.1093/nar/gkac321>
105. Bertelli C, Gray KL, Woods N *et al.* Enabling genomic island prediction and comparison in multiple genomes to investigate bacterial evolution and outbreaks. *Microb Genom* 2022;8:mgen000818. <https://doi.org/10.1099/mgen.0.000818>
106. Soares SC, Geyik H, Ramos RTJ *et al.* GIPSy: genomic island prediction software. *J Biotechnol* 2016;232:2–11. <https://doi.org/10.1016/j.jbiotec.2015.09.008>
107. Sánchez-Soto D, Agüero-Chapin G, Armijos-Jaramillo V *et al.* ShadowCaster: compositional methods under the shadow of phylogenetic models to detect Horizontal Gene Transfers in prokaryotes. *Genes (Basel)* 2020;11:756. <https://doi.org/10.3390/genes11070756>
108. Hemme CL, Green SJ, Rishishwar L *et al.* Lateral gene transfer in a heavy metal-contaminated-groundwater microbial community. *MBio* 2016;7:10–1128. <https://doi.org/10.1128/mBio.02234-15>
109. Winstanley C, Langille MGI, Fothergill JL *et al.* Newly introduced genomic prophage islands are critical determinants of *in vivo* competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res* 2009;19:12–23. <https://doi.org/10.1101/gr.086082.108>
110. Benson DA, Cavanaugh M, Clark K *et al.* GenBank. *Nucleic Acids Res* 2012;41:D36–42. <https://doi.org/10.1093/nar/gks1195>
111. O’Leary NA, Wright MW, Brister JR *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;44:D733–45. <https://doi.org/10.1093/nar/gkv1189>
112. Vernikos GS, Parkhill J. Resolving the structural features of genomic islands: a machine learning approach. *Genome Res* 2008;18:331–42. <https://doi.org/10.1101/gr.7004508>
113. Shrivastava S, Siva Kumar Reddy CV, Mande SS. INDeGenIUS, a new method for high-throughput identification of specialized functional islands in completely sequenced organisms. *J Biosci* 2010;35:351–64. <https://doi.org/10.1007/s12038-010-0040-4>
114. Jaron KS, Moravec JC, Martinková N. SigHunt: Horizontal Gene Transfer finder optimized for eukaryotic genomes. *Bioinformatics* 2014;30:1081–6. <https://doi.org/10.1093/bioinformatics/btt727>
115. Metzler S, Kalinina OV. Detection of atypical genes in virus families using a one-class SVM. *BMC Genomics* 2014;15:913. <https://doi.org/10.1186/1471-2164-15-913>
116. Jani M, Azad RK. IslandCafe: compositional anomaly and feature enrichment assessment for delineation of genomic islands. *G3: Genes, Genomes, Genetics* 2019;9:3273–85. <https://doi.org/10.1534/g3.119.400562>
117. De R, Jani M, Azad RK. DICEP: an integrative approach to augmenting genomic island detection. *J Biotechnol* 2024;388:49–58. <https://doi.org/10.1016/j.jbiotec.2024.04.011>
118. Sobczyk PA, Hazen TH. Horizontal Gene Transfer and mobile genetic elements in marine systems. *Methods Mol Biol* 2009;435–53. https://doi.org/10.1007/978-1-60327-853-9_25
119. Arndt D, Grant JR, Marcu A *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16–21. <https://doi.org/10.1093/nar/gkw387>
120. Fouts DE. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* 2006;34:5839–51. <https://doi.org/10.1093/nar/gkl732>
121. Yoon SH, Park YK, Kim JF. PAIDB v2. 0: exploration and analysis of pathogenicity and resistance islands. *Nucleic Acids Res* 2015;43:D624–30. <https://doi.org/10.1093/nar/gku985>