MICROBIAL GENOMICS

METHOD
Bertelli *et al.*, *Microbial Genomics* 2022;8:000818
DOI 10.1099/mgen.0.000818

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

Enabling genomic island prediction and comparison in multiple genomes to investigate bacterial evolution and outbreaks

Claire Bertelli[1,2]†, Kristen L. Gray[1]†, Nolan Woods[1], Adrian C. Lim[1], Keith E. Tilley[1], Geoffrey L. Winsor[1], Gemma R. Hoad[3], Ata Roudgar[3], Adam Spencer[3], James Peltier[3], Derek Warren[3], Amogelang R. Raphenya[4,5,6], Andrew G. McArthur[4,5,6] and Fiona S. L. Brinkman[1,*]

Abstract

Outbreaks of virulent and/or drug-resistant bacteria have a significant impact on human health and major economic consequences. Genomic islands (GIs; defined as clusters of genes of probable horizontal origin) are of high interest because they disproportionately encode virulence factors, some antimicrobial-resistance (AMR) genes, and other adaptations of medical or environmental interest. While microbial genome sequencing has become rapid and inexpensive, current computational methods for GI analysis are not amenable for rapid, accurate, user-friendly and scalable comparative analysis of sets of related genomes. To help fill this gap, we have developed IslandCompare, an open-source computational pipeline for GI prediction and comparison across several to hundreds of bacterial genomes. A dynamic and interactive visualization strategy displays a bacterial core-genome phylogeny, with bacterial genomes linearly displayed at the phylogenetic tree leaves. Genomes are overlaid with GI predictions and AMR determinants from the Comprehensive Antibiotic Resistance Database (CARD), and regions of similarity between the genomes are also displayed. GI predictions are performed using Sigi-HMM and IslandPath-DIMOB, the two most precise GI prediction tools based on nucleotide composition biases, as well as a novel BLAST-based consistency step to improve cross-genome prediction consistency. GIs across genomes sharing sequence similarity are grouped into clusters, further aiding comparative analysis and visualization of acquisition and loss of mobile GIs in specific sub-clades. IslandCompare is an open-source software that is containerized for local use, plus available via a user-friendly, web-based interface to allow direct use by bioinformaticians, biologists and clinicians (at https://islandcompare.ca).

DATA SUMMARY

(1) All code used in the implementation of IslandCompare can be accessed from GitHub at https://github.com/brinkmanlab/IslandCompare.

(2) The web interface can be accessed from https://islandcompare.ca.

(3) Reference genomes used for processing of draft genomes are from MicrobeDB [1].

(4) Genomes used for the BLAST consistency testing and development and clustering analysis are from Freschi *et al.* (2018) [2] and Hingston *et al.* (2017) [3].

(5) Genomes used to generate Fig. 2 were downloaded from pseudomonas.com [4].

Received 26 January 2022; Accepted 26 March 2022; Published 18 May 2022

Author affiliations: [1]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada; [2]Institute of Microbiology, Lausanne University Hospital and University of Lausanne, 1011 Lausanne, Switzerland; [3]Research Computing Group, Simon Fraser University, Burnaby, BC, Canada; [4]David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, ON, Canada; [5]Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada; [6]Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, ON, Canada.

***Correspondence:** Fiona S. L. Brinkman, brinkman@sfu.ca

Keywords: antimicrobial resistance; comparative genomics; genomic islands; interactive visualization; web server.

Abbreviations: AMR, antimicrobial resistance; BLAST, Basic Local Alignment Search Tool; CARD, Comprehensive Antibiotic Resistance Database; GFF, General Feature Format; GI, genomic island; LES, Liverpool epidemic strain; MCC, Matthews correlation coefficient; MCL, Markov cluster algorithm; NCBI, National Center for Biotechnology Information; RGI, Resistance Gene Identifier.

†These authors contributed equally to this work

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Four supplementary figures and one supplementary table are available with the online version of this article.

000818 © 2022 The Authors

This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

Impact Statement

Public-health microbiology is increasingly adopting a population-based approach to genomic epidemiology and characterization of bacterial outbreak strains, analysing many closely related isolates together instead of individual genomes. In this context, there is a need to rapidly compare the mobile genetic elements of bacterial genomes, particularly genomic islands (GIs), which are known to disproportionately encode genes involved in virulence and resistance to some antimicrobial drug classes. IslandCompare is a new web-based software application with a user-friendly interface that addresses this need by providing a platform for the prediction, clustering and visualization of GIs and associated antimicrobial-resistance genes across sets of microbial genomes.

INTRODUCTION

The acquisition of foreign genetic material from other microbial genomes, phages or environmental DNA is a major driver of bacterial and archaeal genome evolution [5, 6]. Clusters of genes of probable horizontal origin, commonly termed genomic islands (GIs), often provide adaptive traits that present a selective advantage and can eventually become fixed in the population. GIs disproportionately encode medically important adaptations, including virulence genes [7] and certain antimicrobial-resistance (AMR) determinants [8, 9]. Due to their highly dynamic nature, GIs and plasmids can represent one of the major sources of variation between strains, as was recently described for outbreak and non-outbreak strains of atypical enteropathogenic *Escherichia coli* [10] and for the *Pseudomonas aeruginosa* Liverpool epidemic strain (LES) [11]. Comparative GI analysis is becoming increasingly important as genomic epidemiology becomes a key investigative tool for pathogen outbreak analysis and characterization of microbial gene mobility. With the rapid decrease of sequencing costs and the increasing availability of dedicated databases and tools, whole-genome sequencing is progressively being implemented worldwide as a routine tool for outbreak analysis [12–14]. Furthermore, evolutionary analyses have moved towards larger-scale datasets, requiring adapted tools to track the integration and loss of larger genomic regions that may confer adaptive capabilities to their hosts.

Over the past decade, many algorithms have been developed to predict and visualize GIs, mainly in single genomes, by identifying hallmarks of GIs such as biased nucleotide composition, mobility genes, phage-related genes or direct repeats [15, 16]. However, most GI predictors are released as command-line tools, hampering their use by biologists, and only a few offer standalone graphical user interfaces or web services, often with limited data visualization [16]. IslandViewer was the first tool combining several GI predictors and offering an interactive and integrative data visualization with AMR genes and virulence factors [17–19]. As a result, it has rapidly become one of the most widely used and cited tools for GI prediction. However, IslandViewer does not allow comparative analysis, beyond side-by-side circular plots for a user-submitted and a reference genome.

To facilitate the comparative analysis and visualization of GIs, we have developed IslandCompare, a novel open-source user-friendly web service. IslandCompare offers GI prediction by two of the most accurate predictors, a novel BLAST-based module to improve cross-genome prediction consistency, GI clustering by sequence similarity and contextualized visualization with a phylogenetic tree for a few to hundreds of microbial genomes. It should aid investigations of bacterial and archaeal genome evolution for larger-scale datasets that are becoming more routinely obtained, including for investigations of pathogen outbreaks.

THEORY AND IMPLEMENTATION

IslandCompare workflow

To obtain a comparative view of GIs across several to hundreds of genomes, the IslandCompare workflow includes three parallel pipelines for (i) GI prediction and comparison, (ii) comparative genome visualization, and (iii) identification of AMR determinants (Fig. 1). The analysis workflow is hosted in Galaxy [20] and data processing is supported by a host of integrated tools [21–27]. IslandCompare takes as input GenBank (.gbk) or EMBL (.embl) files of draft or complete genomes with gene and protein annotations. IslandCompare supports draft genome submissions by allowing users to select an existing complete genome to be used as a reference for contig reordering. Contigs that could not be reordered by similarity to the reference genome, including repetitive sequences, will be placed towards the end of the pseudochromosome.

GIs are predicted using two of the most accurate tools available according to our software benchmarking in 2018 [16]: IslandPath-DIMOB [28] and Sigi-HMM [29]. Both tools rely on the identification of sequence composition biases (dinucleotides and codon usage, respectively) in coding regions and, hence, require that genome sequences submitted to IslandCompare are annotated. An additional BLAST-based consistency step was added to ensure prediction consistency across genomes in an analysis (discussed in greater depth in the following section). To visualize groups of GIs that are similar across genomes, GIs within 500 bp of one another are merged and considered as a single prediction and these GIs are clustered by sequence similarity. Mash [30] applies MinHash, reducing all GI sequences to representative sketches of *k*-mers for comparison, and produces a distance matrix estimated

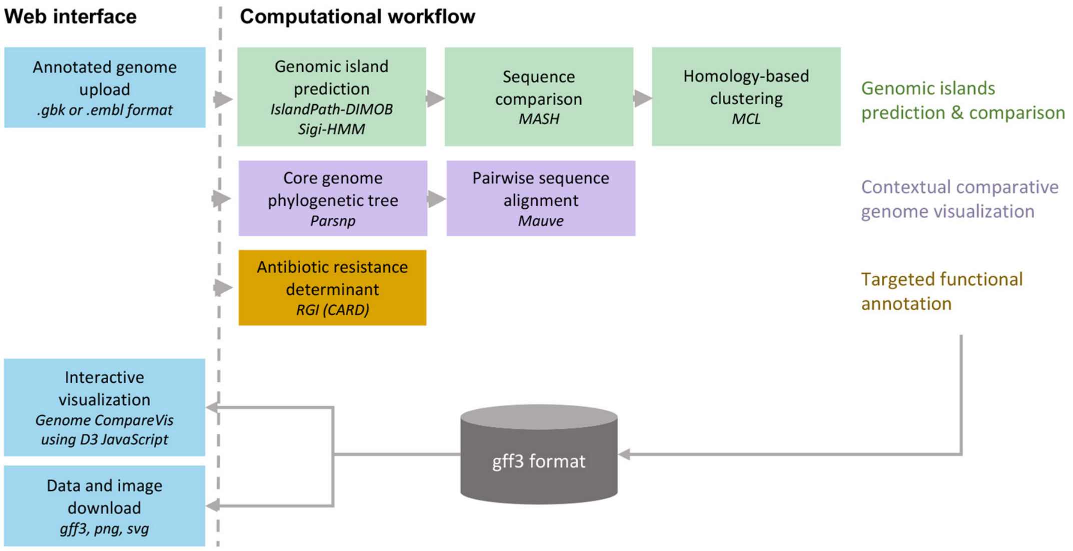
Web interface
Computational workflow
Annotated genome upload
.gbk or .embl format
Genomic island prediction
IslandPath-DIMOB
Sigi-HMM
Sequence comparison
MASH
Homology-based clustering
MCL
Genomic islands prediction & comparison
Core genome phylogenetic tree
Parsnp
Pairwise sequence alignment
Mauve
Contextual comparative genome visualization
Antibiotic resistance determinant
RGI (CARD)
Targeted functional annotation
Interactive visualization
Genome CompareVis using D3 JavaScript
Data and image download
gff3, png, svg
gff3 format

Fig. 1. The IslandCompare workflow. IslandCompare integrates three parallel workflows for the prediction and comparison of GIs, the phyletic visualization of genomes, and the annotation and highlighting of genes with potentially interesting functions such as AMR genes. All results are stored in a standard gff3 format that is used either for interactive visualization in the IslandCompare user-friendly web interface, with images available for download, or for the export of data to conduct further GI analyses.

using Jaccard index. This matrix is converted to a weighted graph that is resolved into clusters of similar sequences by the Markov cluster algorithm (MCL) [31].

To fully appreciate potential loss and acquisition of novel GIs, genome visualization must be contextualized with the phylogenetic relationships between the isolates of interest. Therefore, a phylogenetic tree is calculated using Parsnp v1.2 [32] based on SNPs in the core genome of all sequences submitted for analysis. Users wishing to use an existing species phylogeny can provide a tree in Newick format, with branch labels matching genome accession. Then, to parallelize and speed up the computations, pairs of genome sequences ordered according to the phylogeny are compared using Mauve v2015_02_13.0 [33], and regions sharing sequence similarity across the pairs of aligned genomes are displayed as grey areas.

AMR determinants are predicted using the Resistance Gene Identifier (RGI) v5.1.1 of the curated Comprehensive Antibiotic Resistance Database (CARD) v3.0.7 [34, 35]. RGI allows the identification of both protein variants due to gene mutations and protein homologues conferring resistance to antibiotics. IslandCompare only considers resistance determinants identified with the 'perfect' label, corresponding to a 100% amino acid identity match to a resistance determinant in the database, and 'strict' label that uses curated bitscore detection cut-offs. Loose labelled resistance determinants are not available in IslandCompare given the large number of spurious hits they may yield.

Web platform and visualization

IslandCompare provides a user-friendly web-based interface to allow direct use by bioinformaticians, biologists or clinicians. A drag-and-drop area allows one to easily upload genomes of interest and submit their analysis. A dynamic and interactive visualization strategy displays the bacterial core-genome phylogeny, regions of similarity between genomes, and bacterial genomes overlaid with GI predictions and AMR gene determinants (Fig. 2).

Data export

To facilitate downstream GI analyses, IslandCompare allows users to download their GI prediction results, as well as the predicted AMR genes and other annotated genome features, in General Feature Format (.GFF). Users can also elect to download only the GI predictions. Images of the phyletic comparative view can be exported to svg and png files for the preparation of publication-grade figures.

Software availability and web service implementation

IslandCompare version 1.0 can be accessed at https://islandcompare.ca. The web service allows users to submit archaeal and bacterial genomes of interest. Each account has a unique URL that can be bookmarked to access the results at any time during 3 months after data analysis. Older results will be deleted automatically. A separate command line interface tool (https://github.

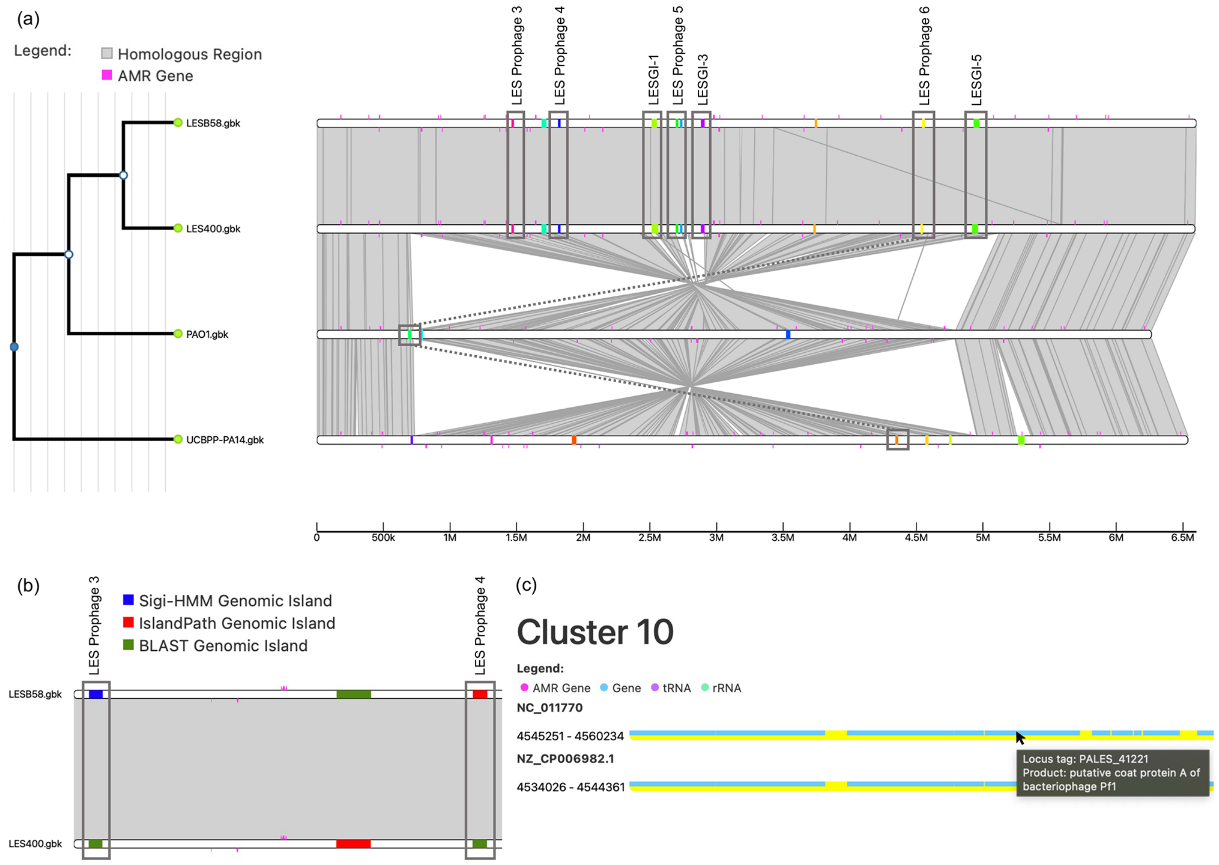

Fig. 2. Comparative visualization of four *P. aeruginosa* genomes highlighting GIs and AMR determinants. (a) A phylogeny (left) indicates the relationship between the isolates in the analysis, with zoom-in functionality available. (a – right, b) GIs are represented as coloured blocks placed on a linear representation of the genome (linear white bars indicate genomes, with alignments between genomes shown in grey), with GIs coloured by (a) cluster or (b) prediction method. (c) The cluster view allows users to explore gene content within a given GI.

com/brinkmanlab/islandcompare-cli) is available to enable analyses to be submitted and retrieved from the command line, an ideal option for those wishing to process larger numbers of genomes at once. For users wishing to install their own instance of IslandCompare, the source code is freely released under an MIT license (https://github.com/brinkmanlab/IslandCompare). Furthermore, IslandCompare can be deployed with Docker to a cloud computing cluster by following the instructions in the deployment subdirectory.

Cross-genome GI prediction consistency

The visual nature of IslandCompare has allowed us to identify for the first time cases where GIs are predicted inconsistently across closely related genomes. Fig. S1 (available with the online version of this article) illustrates such a case where a GI is predicted in a subset of genomes in the analysis, but not in others, despite the fact that a nearly identical sequence is present in these genomes as well. In order to evaluate the cause of these inconsistencies, cases of missed GIs were identified in a set of 40 *P. aeruginosa* genomes. After exploring a range of length coverage values (Fig. S2a), sequences that aligned with predicted GIs across a minimum of 95% of the length were retained in the analysis. By evaluating a subset of these cases, we determined that differences in the underlying annotations across genome files were impacting the downstream predictions made by the GI-prediction software. Of particular note, even seemingly trivial differences in a gene being annotated as a pseudogene in one genome and not as such in another genome could lead to inconsistencies in GI annotation. In one such example, outlined in Fig. S3, there are differences in which genes are labelled as pseudogenes, which in turn impacts the dinucleotide measurements in IslandPath, and subsequently the GI predictions. In 20% of the GIs considered missed by IslandPath, there was no mobility gene predicted by IslandPath in this region due to underlying annotation differences, which would have prevented a GI prediction from being made. Other differences in genes being predicted/not predicted could impact the oligonucleotide bias measures in the region and have downstream effects on GI prediction.

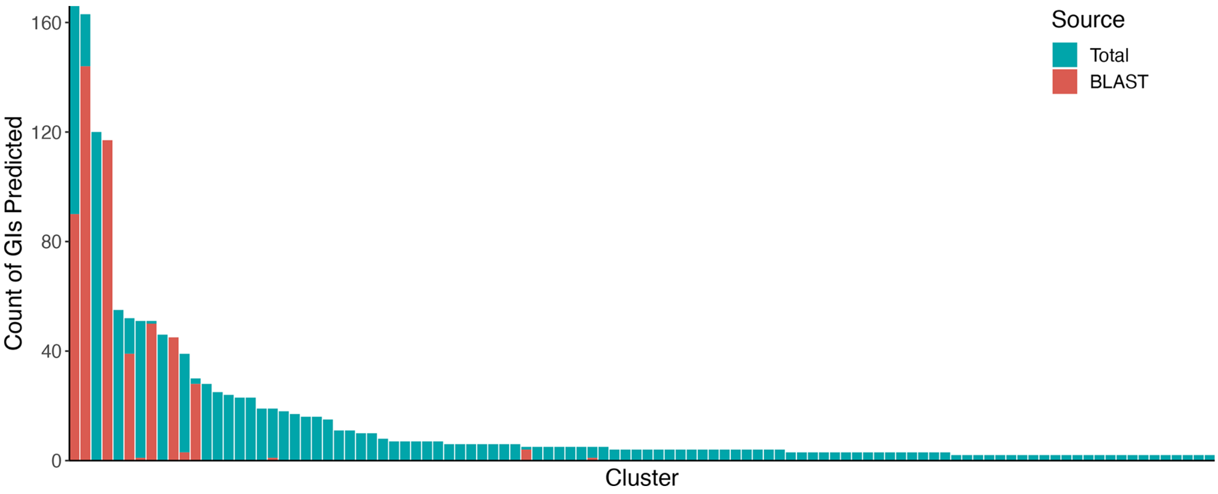

Fig. 3. Counts of predicted GIs for each cluster and proportion predicted by the BLAST-based consistency module for a dataset of 166 *L. monocytogenes* genomes. Only clusters with more than one GI sequence predicted in the dataset are represented here (see Fig. S4a, b for all clusters).

As a result of these findings, a BLAST-based consistency step is now incorporated into the IslandCompare workflow. The sequences of all GIs predicted by IslandPath-DIMOB and Sigi-HMM are aligned with the genomes in the analysis by nucleotide BLAST (BLASTN). BLAST hits are filtered (length≥400 bp; identity≥90%; *E* value ≤1.6e−7) and syntenous BLAST hits are considered as a single alignment. A given region is considered as aligning to a GI if the length of the total alignment is >5 kb and overall coverage to the GI query >95%. In cases where a region aligns to a predicted island and a GI is not already predicted in this region, the aligned region will be considered a GI prediction and will be labelled as a GI for all output files and the interactive visualization. This module was evaluated on an additional dataset of 166 *Listeria monocytogenes* genomes to verify selected cut-off values.

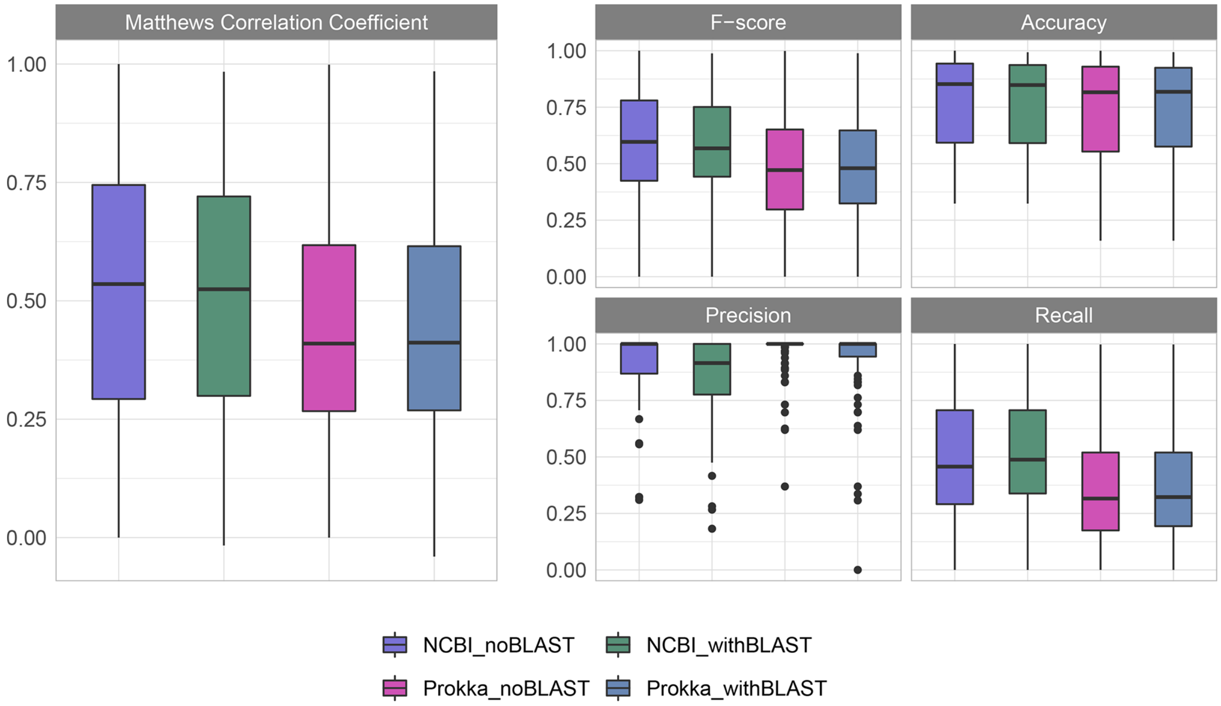

Fig. 4. GI prediction metrics across 86 genomes with known positive and negative GI regions. Predictions were made on the same set of genomes annotated by either NCBI or Prokka. All analyses were run for the GI results both with and without the BLAST-based consistency module results included.

Table 1. Mean GI prediction metrics across 86 genomes annotated with either NCBI or Prokka shown with or without the new BLAST-based consistency module

Predictor	MCC	F-score	Accuracy	Precision	Recall
NCBI_noBLAST	0.524	0.592	0.771	0.906	0.497
NCBI_withBLAST	0.514	0.597	0.767	0.849	0.521
Prokka_noBLAST	0.439	0.466	0.744	0.963	0.355
Prokka_withBLAST	0.438	0.473	0.746	0.925	0.366

Evaluation of GI predictions

A dataset of 86 genomes with known positive and negative GI regions developed by Bertelli *et al.* [16] was used for computing GI prediction performance metrics. As IslandCompare is intended to be used with sets of related genomes, four reference genomes were selected to run with each test genome; reference genomes were randomly selected from a list of available genomes within the same species. A full list of test genomes and associated references used in the analysis can be found in Table S1; all test and reference genomes were downloaded from RefSeq. Given the impact of annotations on downstream GI predictions, GI prediction metrics were computed for the dataset both with National Center for Biotechnology Information (NCBI) annotations (NCBI Annotation Pipeline versions 3.0–4.12) and Prokka (version 1.13). In addition, all metrics were computed for the results both with and without the new BLAST-based consistency module included. True positives (TP) and false positives (FP) were identified on a per nucleotide basis for nucleotides predicted as being within GIs that overlapped with the positive and negative datasets, respectively. Nucleotides that were not labelled as a GI were categorized as true negatives (TN) or false negatives (FN) if they overlapped with the negative or positive datasets, respectively. Accuracy, recall, precision, F-score and Matthews correlation coefficient (MCC) were calculated according to the following formulas:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{F1 score} = \frac{\text{2TP}}{\text{2TP+FP+FN}}$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN+FP}\times\text{FN}}{\text{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Evaluation of GI clustering method

A range of Mash *k*-mer sizes and cluster granularity parameters for the MCL step were evaluated to ensure optimal clustering. GI predictions and clusters were generated in IslandCompare for datasets of 166 *L. monocytogenes* genomes and 40 *P. aeruginosa* genomes. BLASTN [36] was used to determine which GIs aligned to one another. Syntenous BLAST alignments separated by less than 6 kb were merged and GIs with a total alignment length spanning >50% of the sequence length were considered as cluster pairs (all cut-offs were determined empirically by assessing a range of values). These cluster pairs were used to determine which GI cluster pairs predicted by IslandCompare were true/false predictions. Accuracy, recall and precision were calculated according to the same formulas used for the GI prediction evaluation.

RESULTS AND DISCUSSION

IslandCompare has been developed to enable direct GI prediction, comparison and explorative visualization across many genomes. The visual output (Fig. 2) features an interactive and linear representation of each genome overlaid with GI predictions. The tree can be displayed as a phylogram or a cladogram, by toggling branches, and a simple click on internal nodes allows the user to select sub-clades of interest for visualization. A zoom-in functionality enables visualization of genes and their annotations for a selected genomic region. GIs are coloured uniformly according to their sequence cluster and hovering over one allows the user to highlight similar GIs across all genomes. Selecting a GI brings in a specific view of all members of the cluster with its respective position in each genome, gene annotations and flagged AMR determinants, which will be expanded with further information in the future.

An evaluation of the BLAST-based consistency module, which was integrated to ensure that GIs are predicted consistently across genomes within the analysis, in a dataset of *L. monocytogenes* genomes indicated that the selected length coverage threshold of 95% is effective at retaining only the spike of nearly identical sequences targeted by this module (Fig. S2b). For this dataset, the BLAST-based consistency module predictions contributed to 600/1595 merged GI results. The proportion of GI predictions made by BLASTN within a single cluster were highly variable (Fig. 3), indicating that some GIs were more prone to being missed while

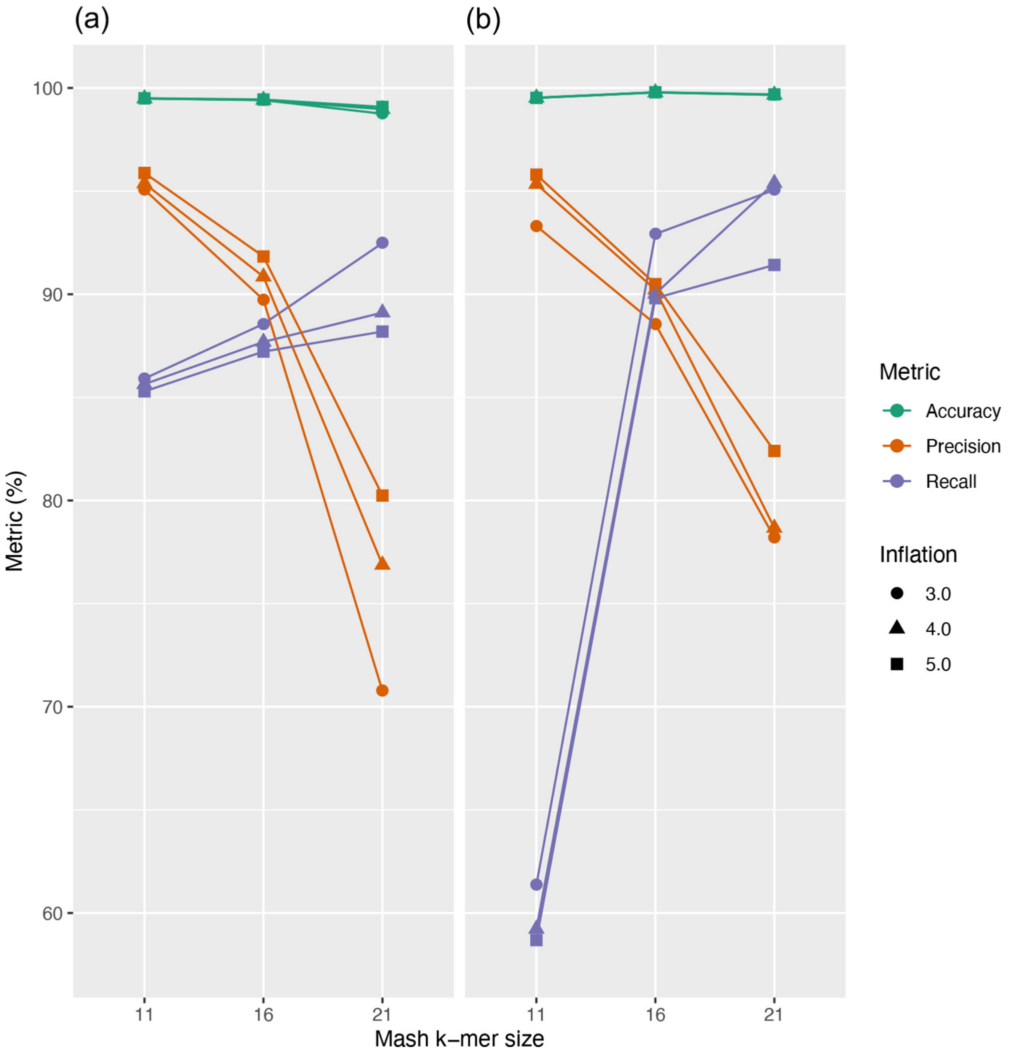

Fig. 5. Evaluation of a range of parameters for GI clustering in IslandCompare for datasets of (a) 166 *L. monocytogenes* genomes and (b) 40 *P. aeruginosa* genomes. Based on this analysis, a *k*-mer size of 16 and inflation value of 5.0 were selected for the Mash and MCL steps in the IslandCompare clustering pipeline, respectively.

others were predicted very consistently by the composition-based GI prediction tools (see Fig. S4a for the proportion of BLAST-based consistency module predictions by cluster for the *P. aeruginosa* dataset). While the new BLAST-based consistency module will help improve GI prediction consistency across genomes in IslandCompare, these results allude to the general importance of ensuring annotation consistency for downstream analyses (again, see Fig. S3). Similar issues would be expected to arise for other gene-annotation-dependent analyses, including prediction of AMR genes and microbial typing (depending on the software used).

Overall, the GI predictions made on the NCBI-annotated genomes produced slightly higher MCC values and F-scores, although the Prokka-annotated genome predictions had moderately higher precision, and accuracy was comparable across all sets (Fig. 4, Table 1). This analysis was performed with the same genomes annotated by two different platforms, NCBI and Prokka, due to the aforementioned importance of gene annotations in the downstream prediction of GIs. There were more predictions made when the genomes were annotated with NCBI than with Prokka. With a greater number of predictions made, it would be expected that the number of true positives recalled when using NCBI is higher (hence, the higher recall). However, this is at the expense of some additional false positives, impacting the precision. The results with just the IslandPath-DIMOB and Sigi-HMM predictions (noBLAST) were also compared to the complete results with the BLAST-based consistency module added. Generally speaking, the addition of this module afforded a slight increase in recall at the expense of a small dip in precision, although the difference was moderate. Compared to the previous analysis of IslandPath-DIMOB and Sigi-HMM [16], the MCC and F-score values were higher in the NCBI_noBLAST results (combination of IslandPath-DIMOB and Sigi-HMM than for either individual tool. For all other metrics (accuracy, recall, precision) the results for the NCBI_noBLAST most closely resembled the best result for either IslandPath-DIMOB or Sigi-HMM from this previous study, despite the fact that an intermediate result would have been expected. This could be due to the updated NCBI annotations in all genomes used in the analysis, the slight variation in the dataset (which contains fewer genomes than the previous study due to reference availability), or some combination thereof. The decision of which annotation platform to use and whether to include the BLAST-based consistency module results will depend upon the application and priorities of a given user. For example, if a user wishes to obtain as many GI predictions as possible and confidently compare GI content across genomes, then it would be advisable for them to annotate with NCBI and include the BLAST consistency results in their analysis, but another user wishing to explore the high confidence GI contents of their population as a whole with less regard to differences across individual genomes ought to consider annotating with Prokka and disregarding the BLAST consistency results.

To illustrate the utility of IslandCompare, we performed an analysis of a set of *P. aeruginosa* genomes (Fig. 2a, b, c). This analysis includes the LES B58 strain, whose GI contents were characterized by Winstanley *et al.* [11], as well as another LES and the reference strains PAO1 and PA14, in order to illustrate how IslandCompare can facilitate the identification of unique GIs. A few key GIs are labelled in the figure, as named in the Winstanley *et al.* publication; IslandCompare identified 7/11 GIs discussed in this study, consistent with the focus of GI predictions on precision (islands identified are highly likely to be true islands) at the expense of recall. Most of these GIs (5/7) were only predicted in one of the two LES isolates by the IslandPath-DIMOB and Sigi-HMM modules (subset of these cases shown in Fig. 2b), but were confirmed to be present in both with the BLAST-based consistency module included. Two GIs inserted in tandem – LES prophage 5 (broken into two predictions in IslandCompare) and LES GI-3 – are easily identifiable with the IslandCompare visual; it can be seen from the Mauve alignment that these islands are unique to the LES isolates. LES prophage 6, a PF1-like phage, can be seen in both LES isolates and the cluster view of this island is shown in Fig. 2(c). Similar Pf1 islands can be seen in PA14 and PAO1 as well, although for this GI only the LES sequences cluster together. LES prophages 3 and 4, as well as GI-1 and GI-5 can also be easily identified as GIs only present in the LES isolates from this figure. This example analysis demonstrates the effectiveness of IslandCompare for rapidly identifying differences in GI content across closely related genomes.

As IslandCompare integrates a complex comparative pipeline building upon existing tools, the time-to-result can range from minutes to several hours and depends on the number of genomes submitted. Datasets of 20, 100 and 1000 draft *Enterococcus faecium* genomes [37, 38] ran in 22 min, 59 min and 14 h, respectively. Therefore, users who plan to regularly run large analyses (eg >200 genomes) are encouraged to set up IslandCompare on a local server using the containerized version that we provide (instructions under the deployment subdirectory of https://github.com/brinkmanlab/IslandCompare). To favour parallelization and decrease computation time, multiple sequence alignments with Mauve that are time-consuming have been replaced by pairwise alignments, while retaining the sensitivity of the tool. The core genome SNP-based reconstruction of a phylogenetic tree using Parsnp is also a time-consuming step. Furthermore, Parsnp requires that closely related genomes are analysed to run successfully and the inclusion of distant genomes may lead to pipeline abortion. Hence, the user can bypass Parsnp by providing a Newick tree as an additional input file, allowing one to then proceed directly with the pairwise Mauve sequence alignment that handles adequately more distant genome alignments.

The clustering of GIs into groups sharing sequence similarity allows users to rapidly identify similar integrated elements shared across genomes. However, GIs are well known to be highly dynamic regions where rapid evolution by mutation or gene loss is often observed [39, 40]. Multiple GI integration in the same location or further gene integration have also been observed [41–43], leading to further sequence diversity. Hence, GIs may rapidly differ and sequence similarity-based clustering may not perfectly capture the complex evolution of these elements. In this first IslandCompare release, the display of region-wide similarity (grey shaded areas) with high genomic synteny and similarity with GIs belonging to different clusters allows one to identify and visualize these cases, before more advanced strategies can be implemented in future releases. Default clustering parameters for IslandCompare were selected to provide a balance between recall and precision, while prioritizing precision (Fig. 5). Additional work is underway to provide targeted prediction of a curated set of GIs described in previous studies for a selection of target pathogens, with 404 GI sequences and associated metadata already collected.

IslandCompare provides users with a user-friendly interface for comparative GI analysis that does not require advanced command-line bioinformatics skills. It combines well-accepted, widely used GI predictors with a novel BLAST-based component to improve cross-genome prediction consistency that is not available in any other GI prediction tool. This resource should enable more robust comparison of GIs to gain further insights into pathogen evolution.

Funding information
This work was supported by a Swiss National Science Foundation fellowship [P300-PA_164673 to C.B.], grants from the Société Académique Vaudoise [to C.B.], grants from Genome Canada and Canadian Institutes of Health Research (CIHR) [to F.S.L.B. and A.G.M.], scholarship support from NSERC-CREATE (Natural Sciences and Engineering Research Council of Canada), CIHR and Simon Fraser University [to K.L.G.], and supported by a Cisco Research Chair in Bioinformatics and a David Braley Chair in Computational Biology [to A.G.M.].

Acknowledgements

We would like to thank other members of the Brinkman lab for their helpful comments, researchers who developed the Sigi-HMM tool for providing GI predictions, and additional researchers involved in RGI and CARD for providing curated AMR determinant data.

Author contributions

Conceptualization: C.B., F.S.L.B. Methodology: C.B., F.S.L.B., K.L.G., A.C.L. Software development and maintenance: A.C.L., N.W., C.B., K.L.G., K.E.T., G.L.W., G.R.H., A.R., A.S., J.P., D.W., A.R.R., A.G.M., F.S.L.B. Validation: C.B., K.L.G. Writing – original draft preparation: C.B., K.L.G. Writing – review and editing: C.B., K.L.G., N.W., A.C.L., K.E.T., G.L.W., G.R.H., A.R., A.S., J.P., D.W., A.R.R., A.G.M., F.S.L.B. Supervision: C.B., K.L.G., F.S.L.B. Funding: C.B., F.S.L.B., K.L.G., A.G.M.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. Langille MGI, Laird MR, Hsiao WWL, Chiu TA, Eisen JA, *et al*. MicrobeDB: a locally maintainable database of microbial genomic sequences. *Bioinformatics* 2012;28:1947–1948.
2. Freschi L, Bertelli C, Jeukens J, Moore MP, Kukavica-Ibrulj I, *et al*. Genomic characterisation of an international *Pseudomonas aeruginosa* reference panel indicates that the two major groups draw upon distinct mobile gene pools. *FEMS Microbiol Lett* 2018;365:fny120.
3. Hingston P, Chen J, Dhillon BK, Laing C, Bertelli C, *et al*. Genotypes associated with *Listeria monocytogenes* isolates displaying impaired or enhanced tolerances to cold, salt, acid, or desiccation stress. *Front Microbiol* 2017;8:369.
4. Winsor GL, Griffiths EJ, Lo R, Dhillon BK, Shay JA, *et al*. Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the Pseudomonas genome database. *Nucleic Acids Res* 2016;44:D646–D653.
5. Dobrindt U, Hochhut B, Hentschel U, Hacker J. Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* 2004;2:414–424.
6. Aminov RI. Horizontal gene exchange in environmental microbiota. *Front Microbiol* 2011;2:158.
7. Ho Sui SJ, Fedynak A, Hsiao WWL, Langille MGI, Brinkman FSL. The association of virulence factors with genomic islands. *PLoS One* 2009;4:e8094.
8. Hall RM. *Salmonella* genomic islands and antibiotic resistance in *Salmonella enterica*. *Future Microbiol* 2010;5:1525–1538.
9. Gilmore MS, Lebreton F, van Schaik W. Genomic transition of enterococci from gut commensals to leading causes of multidrug-resistant hospital infection in the antibiotic era. *Curr Opin Microbiol* 2013;16:10–16.
10. Ingle DJ, Tauschek M, Edwards DJ, Hocking DM, Pickard DJ, *et al*. Evolution of atypical enteropathogenic *E. coli* by repeated acquisition of LEE pathogenicity island variants. *Nat Microbiol* 2016;1:15010.
11. Winstanley C, Langille MGI, Fothergill JL, Kukavica-Ibrulj I, Paradis-Bleau C, *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.
12. Ladner JT, Grubaugh ND, Pybus OG, Andersen KG. Precision epidemiology for infectious disease control. *Nat Med* 2019;25:206–211.
13. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet* 2018;19:9–20.
14. Bertelli C, Greub G. Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clin Microbiol Infect* 2013;19:803–813.
15. Langille MGI, Hsiao WWL, Brinkman FSL. Detecting genomic islands using bioinformatics approaches. *Nat Rev Microbiol* 2010;8:373–382.
16. Bertelli C, Tilley KE, Brinkman FSL. Microbial genomic island discovery, visualization and analysis. *Brief Bioinform* 2019;20:1685–1698.
17. Langille MGI, Brinkman FSL. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 2009;25:664–665.
18. Dhillon BK, Chiu TA, Laird MR, Langille MGI, Brinkman FSL. IslandViewer update: improved genomic island discovery and visualization. *Nucleic Acids Res* 2013;41:W129–W132.
19. Bertelli C, Laird MR, Williams KP, Simon Fraser University Research Computing Group, Lau BY, *et al*. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res* 2017;45:W30–W35.
20. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 2018;46:W537–W544.
21. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, *et al*. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;25:1422–1423.
22. Woods N, Brinkman FSL. Brinkman galaxy tools; 2019. https://zenodo.org/record/3364789
23. Woods N. Brinkmanlab/biopython-convert: update biopython to v1.79; 2021. https://zenodo.org/record/5502644#.YnJWTtrMJPY
24. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
25. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.
26. Cock PJA, Chilton JM, Grüning B, Johnson JE, Soranzo N. NCBI BLAST+ integrated into Galaxy. *Gigascience* 2015;4:39.
27. Woods N. Brinkmanlab/feature_merge: ignore strand option; 2020. https://doi.org/10.5281/ZENODO.3364784
28. Bertelli C, Brinkman FSL. Improved genomic island predictions with IslandPath-DIMOB. *Bioinformatics* 2018;34:2161–2167.
29. Waack S, Keller O, Asper R, Brodag T, Damm C, *et al*. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* 2006;7:142.
30. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, *et al*. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:132.
31. Enright AJ, Van Dongen SA, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;30:1575–1584.
32. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 2014;15:524.
33. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 2010;5:e11147.
34. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, *et al*. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2020;48:D517–D525.
35. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, *et al*. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2017;45:D566–D573.
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.

37. Gouliouris T, Raven KE, Ludden C, Blane B, Corander J, *et al*. Genomic surveillance of *Enterococcus faecium* reveals limited sharing of strains and resistance genes between livestock and humans in the United Kingdom. *mBio* 2018;9:e01780-18.

38. Zaheer R, Cook SR, Barbieri R, Goji N, Cameron A, *et al*. Surveillance of *Enterococcus* spp. reveals distinct species and antimicrobial resistance diversity across a one-health continuum. *Sci Rep* 2020;10:3937.

39. Beutlich J, Jahn S, Malorny B, Hauser E, Hühn S, *et al*. Antimicrobial resistance and virulence determinants in European *Salmonella* genomic island 1-positive *Salmonella enterica* isolates from different origins. *Appl Environ Microbiol* 2011;77:5655–5664.

40. Mukhopadhyay AK, Chakraborty S, Takeda Y, Nair GB, Berg DE. Characterization of VPI pathogenicity island and CTXphi prophage in environmental strains of *Vibrio cholerae*. *J Bacteriol* 2001;183:4737–4746.

41. Williams KP. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res* 2002;30:866–875.

42. Reiter W-D, Palm P, Yeats S. Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. *Nucleic Acids Res* 1989;17:1907–1914.

43. Pavlovic G, Burrus V, Gintz B, Decaris B, Guédon G. Evolution of genomic islands by deletion and tandem accretion by site-specific recombination: ICESt1-related elements from *Streptococcus thermophilus*. *Microbiology* 2004;150:759–774.