DATABASE

Open Access

HelicoBase: a *Helicobacter* genomic resource and analysis platform

Siew Woh Choo[1,2*], Mia Yang Ang[1,2], Hanieh Fouladi[3], Shi Yang Tan[1,2], Cheuk Chuen Siow[1], Naresh VR Mutha[1], Hamed Heydari[1,4], Wei Yee Wee[1,2], Jamuna Vadivelu[5], Mun Fai Loke[5], Vellaya Rehvathy[5] and Guat Jah Wong[1,2]

Abstract

Background: *Helicobacter* is a genus of Gram-negative bacteria, possessing a characteristic helical shape that has been associated with a wide spectrum of human diseases. Although much research has been done on *Helicobacter* and many genomes have been sequenced, currently there is no specialized *Helicobacter* genomic resource and analysis platform to facilitate analysis of these genomes. With the increasing number of *Helicobacter* genomes being sequenced, comparative genomic analysis on members of this species will provide further insights on their taxonomy, phylogeny, pathogenicity and other information that may contribute to better management of diseases caused by *Helicobacter* pathogens.

Description: To facilitate the ongoing research on *Helicobacter*, a specialized central repository and analysis platform for the *Helicobacter* research community is needed to host the fast-growing amount of genomic data and facilitate the analysis of these data, particularly comparative analysis. Here we present HelicoBase, a user-friendly *Helicobacter* resource platform with diverse functionality for the analysis of *Helicobacter* genomic data for the *Helicobacter* research communities. HelicoBase hosts a total of 13 species and 166 genome sequences of *Helicobacter* spp. Genome annotations such as gene/protein sequences, protein function and sub-cellular localisation are also included. Our web implementation supports diverse query types and seamless searching of annotations using an AJAX-based real-time searching system. JBrowse is also incorporated to allow rapid and seamless browsing of *Helicobacter* genomes and annotations. Advanced bioinformatics analysis tools consisting of standard BLAST for similarity search, VFDB BLAST for sequence similarity search against the Virulence Factor Database (VFDB), Pairwise Genome Comparison (PGC) tool for comparative genomic analysis, and a newly designed Pathogenomics Profiling Tool (PathoProT) for comparative pathogenomic analysis are also included to facilitate the analysis of *Helicobacter* genomic data.

Conclusions: HelicoBase offers access to a range of genomic resources as well as tools for the analysis of *Helicobacter* genome data. HelicoBase can be accessed at http://helicobacter.um.edu.my.

Keywords: HelicoBase, *Helicobacter*, Genomic resources, Pairwise genome comparison tool, Pathogenomics profiling tool, Comparative analysis

Background

Helicobacter is a genus of Gram-negative bacteria possessing a characteristic spiral shape [1]. In the past, they were classified as members of the *Campylobacter* genus, but the *Helicobacter* genus has been recognized since 1989; currently with 29 known species (*H. acinonychis, H. anseris, H. aurati, H. bilis, H. bizzozeronii, H. brantae, H. canadensis, H. canis, H. cetorum, H. cholecystus, H. cinaedi, H. cynogastricus, H. felis, H. fennelliae, H. ganmani, H. hepaticus, H. mesocricetorum, H. marmotae, H. muridarum, H. mustelae, H. pametensis, H. pullorum, H. pylori, H. rappini, H. rodentium, H. salomonis, H. trogontum, H. typhlonius, H. winghamensis*) [2,3]. *Helicobacter* species have been found living in the lining of the upper gastrointestinal tract, as well as the liver of some birds and mammals [4]. *Helicobacter* bacteria can be isolated from feces, saliva and dental plaque of some infected people

* Correspondence: lchoo@um.edu.my
[1]Genome Informatics Research Laboratory, High Impact Research (HIR) Building, University of Malaya, 50603 Kuala Lumpur, Malaysia
[2]Department of Oral Biology and Biomedical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia
Full list of author information is available at the end of the article

which is consistent with known transmission routes [5-7]. The most widely known and well-studied species of the genus is *H. pylori*, a human pathogen which infects up to half of the human population [8]. In general, most patients with *H. pylori* infections do not have specific clinical symptoms or signs [9]. However, acute infection may appear as acute gastritis which may further develop into chronic gastritis if no treatment is given [10]. *H. pylori* is also strongly associated with peptic ulcers, duodenitis and stomach cancer [11]. *H. pylori* is a genetically diverse species that has co-evolved with the human race since their migration out of Africa 60,000 years ago [12], and subsequent geographic separation plus founder effects have resulted in distinct populations of bacterial strains that are specific for various geographical regions. In all, 7 populations and 3 subpopulations have been described: hpEurope (isolated from Europe, the Middle East, India and Iran), hpNEAfrica (isolated in Northeast Africa), hpAfrica1 (isolated from countries in Western Africa and South Africa), hpAfrica2 (so far only isolated from South Africa), hpAsia2 (isolated from Northern India and among isolates from Bangladesh, Thailand and Malaysia), hpSahul (from Australian Aboriginals and Papua New Guineans) and hpEastAsia with the subpopulations hspEAsia (from East Asians), hspMaori (from Taiwanese Aboriginals, Melanesians and Polynesians) and hspAmerind (Native Americans) [13-17].

With advances in next-generation sequencing technologies, many genomes of *Helicobacter* isolates have been sequenced by many laboratories [18-22]. The availability of these genome sequences from different sources has made it possible to get a deeper understanding of *Helicobacter* at the genomic level, for example through genome-wide comparative analyses. Such comparative analysis will have a profound impact on understanding the evolution, biology, diversity, evolution and pathogenicity of the *Helicobacter* spp. which may be useful in successfully managing *Helicobacter*-caused diseases.

Many specialized genomic databases or resources have been developed and published for well-studied human pathogens such as *Pseudomonas* Genome database [23,24], *Burkholderia* Genome Database [25], Cyanobacteria Gene Annotation Database (CYORF) [26] and *Mycobacterium abscessus* Genome and Annotation Database (MabsBase) [27]. But no such specialized genomic database is available for *Helicobacter* spp. despite the wealth of available data. Microbial Genome Database for Comparative Analysis (MBGD) [28] and the Integrated Microbial Genomes (IMG) system [29] do provide a wide array of microbial genomes including some *Helicobacter* strains for comparative genomics, but lack the virulence factor perspective for comparative pathogenomics. Another concern regarding most of the existing biological databases is their lack of user-friendly web interfaces, for example, allowing real-time and fast querying and browsing of genomic data.

To facilitate *Helicobacter* research, we have developed a specialized *Helicobacter* resource and analysis platform for the storage of the rapidly increasing genomic data of *Helicobacter*, which presents the data in a useful manner that is easy to access, and enables the analysis of these genomic data, particularly in the field of comparative genomics. The aims of HelicoBase are to provide a comprehensive set of genomic data and a set of useful analysis tools with diverse functionality for data analysis. For instance, HelicoBase is powered by two newly designed tools: PGC for pairwise genome comparison and PathoProT for comparative pathogenomics analysis. The AJAX-based real-time search feature and JBrowse [30] have also been integrated into HelicoBase to allow rapid and seamless searching and browsing of the *Helicobacter* genomic data and annotations. Here we provide an overview and describe some key features of HelicoBase.

Construction and content

HelicoBase has much useful functionality as shown in Figure 1. In the homepage, users can view the latest news & conferences, blogs & information, and the most recent papers related to *Helicobacter* spp. that we manually compiled from different sources. By clicking on the 'Browse' hyperlink on the homepage, users can browse general information on different *Helicobacter* species (Table 1), where each species is linked, e.g. through the "View Strains" button, to a table showing all available strains (either draft or complete genome) and associated strain information like genome size, GC content, number of contigs, CDSs, number of tRNAs and number of rRNAs. Each species has a 'Details' button which directs users to the list of all RAST-predicted Open Reading Frames (ORFs). Useful ORF information is provided including ORF ID, ORF type, functional classification, contig ID, start position and stop position. If users want more information about a specific ORF, they can click on the "Detail" button provided for the ORF. This will direct users to an ORF details page with information like subcellular localization, hydrophobicity, molecular weight, and amino acid and nucleotide sequences of the ORF of interest. JBrowse is integrated into the ORF details page, allowing users to visualize and browse around the genomic location of the ORF. All these annotation details and sequence data for the selected ORF can be downloaded in the same page as CSV and FASTA files, respectively. Furthermore, users can also download the whole-genome annotations and sequences through the provided 'Download' page.

HelicoBase currently accumulates a total of 166 genome sequences from 13 *Helicobacter* species, which were downloaded and compiled from the National Center for

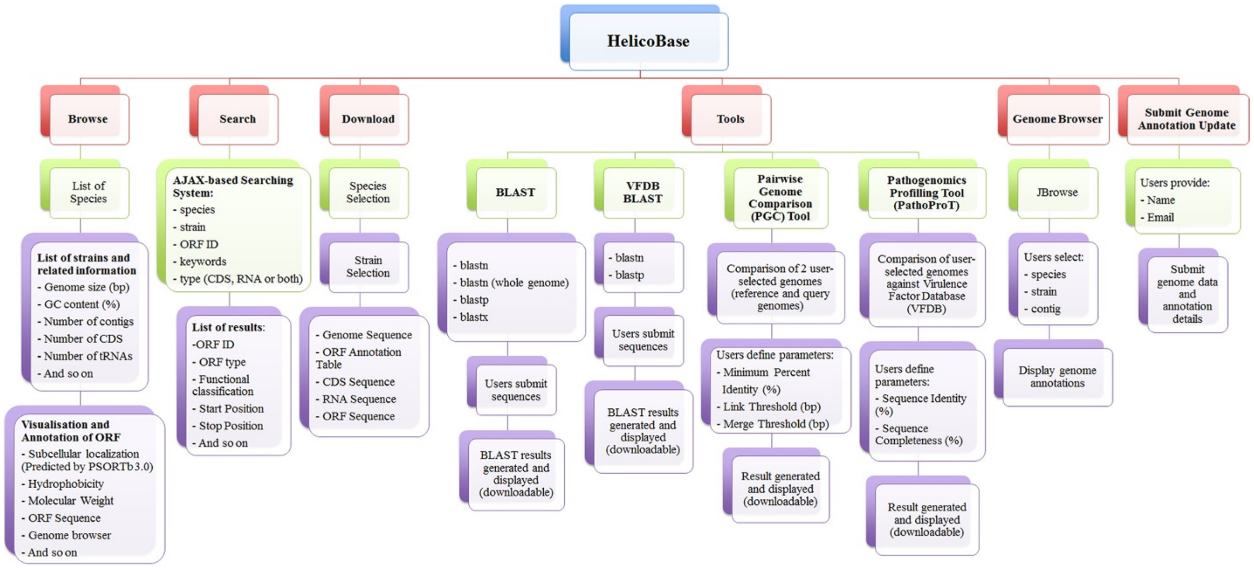

HelicoBase
Browse
Search
Download
Tools
Genome Browser
Submit Genome Annotation Update
List of Species
List of strains and related information
- Genome size (bp)
- GC content (%)
- Number of contigs
- Number of CDS
- Number of tRNAs
- And so on
Visualisation and Annotation of ORF
- Subcellular localization (Predicted by PSORTb 3.0)
- Hydrophobicity
- Molecular Weight
- ORF Sequence
- Genome browser
- And so on
AJAX-based Searching System:
- species
- strain
- ORF ID
- keywords
- type (CDS, RNA or both)
List of results:
-ORF ID
- ORF type
- Functional classification
- Start Position
- Stop Position
- And so on
Species Selection
Strain Selection
- Genome Sequence
- ORF Annotation Table
- CDS Sequence
- RNA Sequence
- ORF Sequence
BLAST
- blastn
- blastn (whole genome)
- blastp
- blastx
Users submit sequences
BLAST results generated and displayed (downloadable)
VFDB BLAST
- blastn
- blastp
Users submit sequences
BLAST results generated and displayed (downloadable)
Pairwise Genome Comparison (PGC) Tool
Comparison of 2 user-selected genomes (reference and query genomes)
Users define parameters:
- Minimum Percent Identity (%)
- Link Threshold (bp)
- Merge Threshold (bp)
Result generated and displayed (downloadable)
Pathogenomics Profilling Tool (PathoProT)
Comparison of user-selected genomes against Virulence Factor Database (VFDB)
Users define parameters:
- Sequence Identity (%)
- Sequence Completeness (%)
Result generated and displayed (downloadable)
JBrowse
Users select:
- species
- strain
- contig
Display genome annotations
Users provide:
- Name
- Email
Submit genome data and annotation details

Figure 1 Overview of HelicoBase. There are 5 main functionalities accessible from the navigation bar on the top of the webpages: Browse, Search, Download, Tools, and Genome Browser. There are four analysis tools incorporated in HelicoBase: standard BLAST, VFDB BLAST, Pairwise Genome Comparison (PGC), and Pathogenomics Profilling Tool (PathoProT).

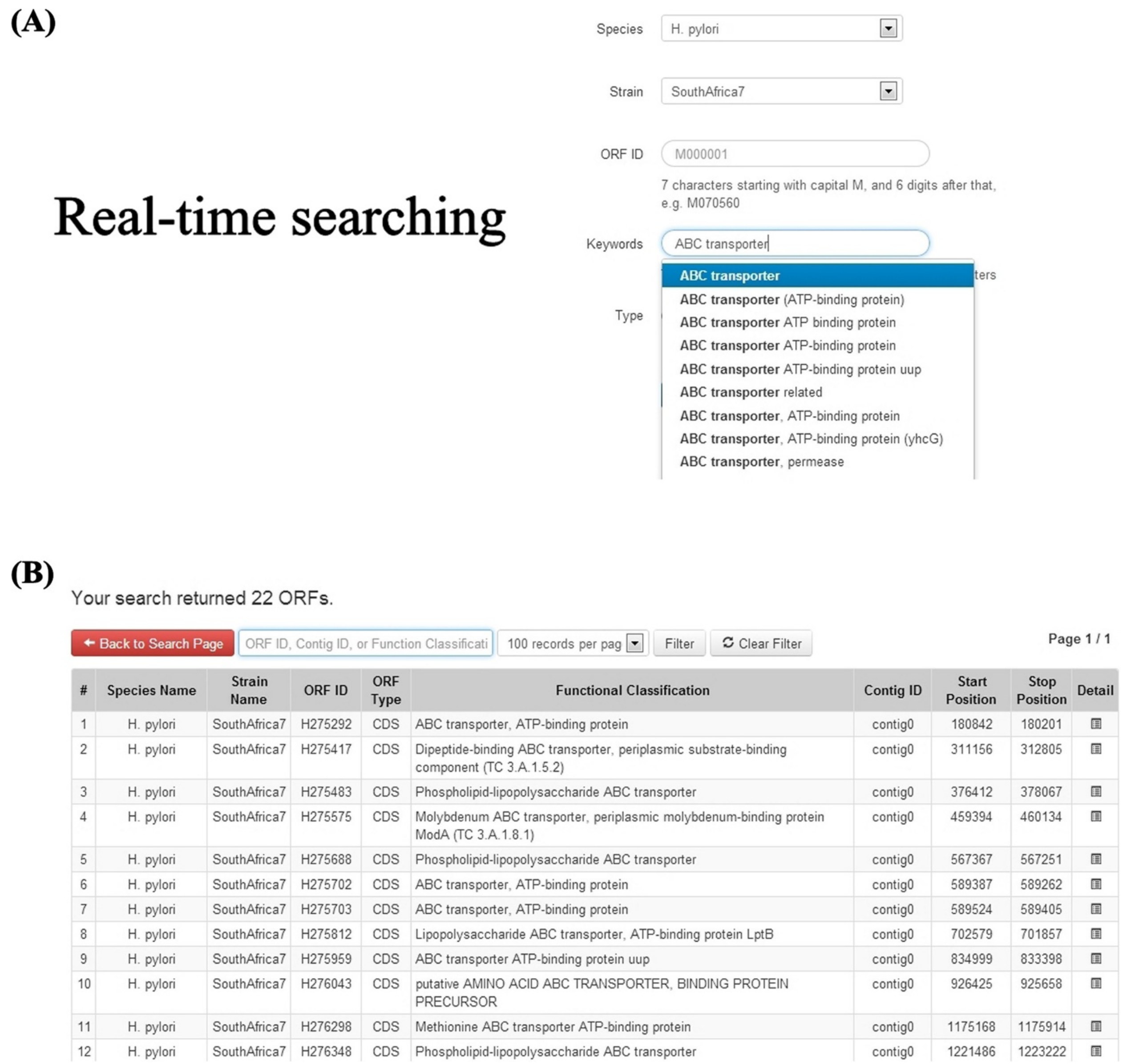
Biotechnology Information (NCBI) [31,32]. To have consistent annotations for comparative analyses, we re-annotate all genomes with the Rapid Annotation using Subsystem Technology (RAST) pipeline [33]. RAST has been successfully tested in annotating both complete and draft genomes of archaea and bacteria in the recent review by Liu et al. [34]. Using this well-established pipeline, functional elements like protein-encoding genes, rRNAs, tRNAs and pseudogenes can be predicted in each *Helicobacter* genome. All genome annotations were stored in our MySQL database. Currently HelicoBase has stored 280,550 coding sequences (CDSs), 6,683 rRNAs and 5,965 tRNA genes predicted in all 166 genomes of the 13 *Helicobacter* species. Among annotations generated by RAST include ORF type, functional classification, chromosomal position, nucleotide length, amino acid length and strand. Other annotations like subcellular localisation, hydrophobicity and molecular weight of the RAST-predicted proteins are also provided. For subcellular localization prediction, we used PSORTb version 3.0, a well-established software to determine the subcellular localization of putative proteins for prokaryotes [35]. In HelicoBase, the 280,550 RAST-predicted CDSs were categorised by PSORTb into 5 different categories such as cytoplasmic, cytoplasmic membrane, extracellular, outer membrane and periplasmic (Additional file 1: Figure S1).

Table 1 List of *Helicobacter* species and genomes in HelicoBase

#	Species	Number of draft genomes	Number of complete genomes
1	*H. acinonychis*	0	1
2	*H. bilis*	1	0
3	*H. bizzozeronii*	1	1
4	*H. canadensis*	2	0
5	*H. cetorum*	0	2
6	*H. cinaedi*	1	1
7	*H. felis*	0	1
8	*H. hepaticus*	0	1
9	*H. mustelae*	0	1
10	*H. pullorum*	1	0
11	*H. pylori*	106	43
12	*H. suis*	2	0
13	*H. winghamensis*	1	0

Real-time data searching feature

With advances in next-generation sequencing technologies and bioinformatics, it is anticipated that the data in HelicoBase will considerably increase as more genomes are sequenced in the future. Therefore, a user-friendly interface allowing users to rapidly search a massive amount of genomic data is vital. To give the *Helicobacter* research community a user-friendly and seamless search experience, we have implemented a powerful real-time AJAX-based search system in the “Search” page on the homepage. Users can search for an ORF by using different parameters including species name, strain, ORF ID, keywords of functional classification and type of sequence (Figure 2). Furthermore, when users type in the search keywords, the system will rapidly retrieve the matches



Species: H. pylori

Strain: SouthAfrica7

ORF ID: M000001

7 characters starting with capital M, and 6 digits after that, e.g. M070560

Keywords: ABC transporter

ABC transporter
ABC transporter (ATP-binding protein)
ABC transporter ATP binding protein
ABC transporter ATP-binding protein
ABC transporter ATP-binding protein uup
ABC transporter related
ABC transporter, ATP-binding protein
ABC transporter, ATP-binding protein (yhcG)
ABC transporter, permease

Your search returned 22 ORFs.

ORF ID, Contig ID, or Function Classificati

100 records per pag

Page 1 / 1

#	Species Name	Strain Name	ORF ID	ORF Type	Functional Classification	Contig ID	Start Position	Stop Position	Detail
1	H. pylori	SouthAfrica7	H275292	CDS	ABC transporter, ATP-binding protein	contig0	180842	180201	▤
2	H. pylori	SouthAfrica7	H275417	CDS	Dipeptide-binding ABC transporter, periplasmic substrate-binding component (TC 3.A.1.5.2)	contig0	311156	312805	▤
3	H. pylori	SouthAfrica7	H275483	CDS	Phospholipid-lipopolysaccharide ABC transporter	contig0	376412	378067	▤
4	H. pylori	SouthAfrica7	H275575	CDS	Molybdenum ABC transporter, periplasmic molybdenum-binding protein ModA (TC 3.A.1.8.1)	contig0	459394	460134	▤
5	H. pylori	SouthAfrica7	H275688	CDS	Phospholipid-lipopolysaccharide ABC transporter	contig0	567367	567251	▤
6	H. pylori	SouthAfrica7	H275702	CDS	ABC transporter, ATP-binding protein	contig0	589387	589262	▤
7	H. pylori	SouthAfrica7	H275703	CDS	ABC transporter, ATP-binding protein	contig0	589524	589405	▤
8	H. pylori	SouthAfrica7	H275812	CDS	Lipopolysaccharide ABC transporter, ATP-binding protein LptB	contig0	702579	701857	▤
9	H. pylori	SouthAfrica7	H275959	CDS	ABC transporter ATP-binding protein uup	contig0	834999	833398	▤
10	H. pylori	SouthAfrica7	H276043	CDS	putative AMINO ACID ABC TRANSPORTER, BINDING PROTEIN PRECURSOR	contig0	926425	925658	▤
11	H. pylori	SouthAfrica7	H276298	CDS	Methionine ABC transporter ATP-binding protein	contig0	1175168	1175914	▤
12	H. pylori	SouthAfrica7	H276348	CDS	Phospholipid-lipopolysaccharide ABC transporter	contig0	1221486	1223222	▤

Figure 2 Real-time search feature. (A) Example of Real-time searching with *H. pylori* strain SouthAfrica7 with "ABC transporter" as keyword. A list of matches with the typed keyword was retrieved from HelicoBase in a real-time manner. **(B)** Example of search output.

from HelicoBase in a real-time manner. This will help users to get the right keywords and will speed up their searching, which is vital in searching a huge database.

Utility

Pairwise genome comparison (PGC) tool: information aesthetic for comparative genomics

HelicoBase is not just designed as a genomic data repository, but also aims to be an analysis platform, particularly to facilitate comparative analysis of multiple *Helicobacter* genomes. The PGC tool is a newly designed in-house comparative analysis tool allowing users to compare two selected *Helicobacter* genomes and display the results in a circular layout on the fly. Through the provided web interface of PGC, users can choose two genomes of interest in HelicoBase for comparison. Alternatively, users can use an online custom web form to upload their own *Helicobacter* genome sequence for comparison with a *Helicobacter* genome in HelicoBase.

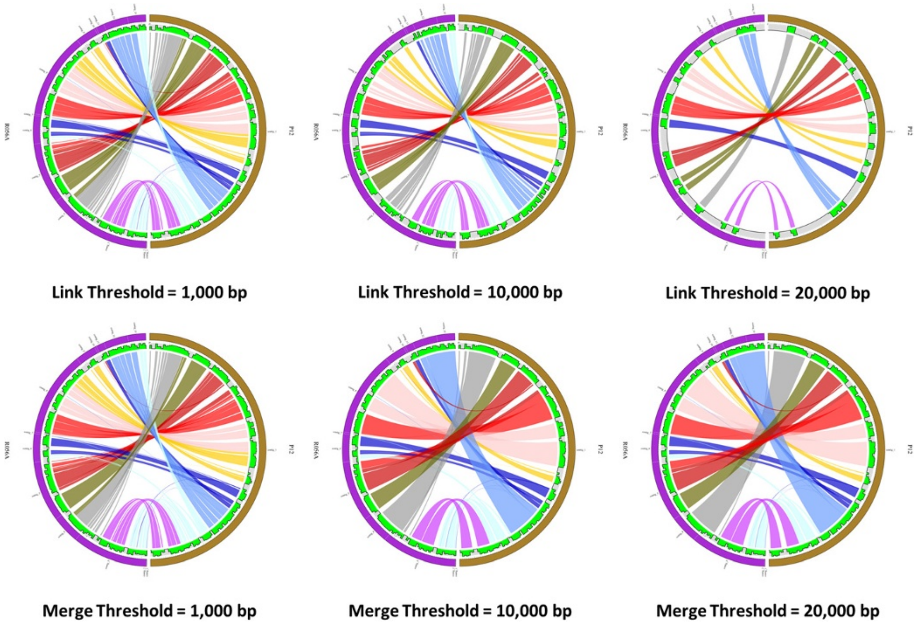
Three main parameters are provided: the minimum percent identity (%), merge threshold (bp) and link threshold (bp). By default, the thresholds of the PGC tool are set to be 95% minimum percent identity and 1,000 bp link threshold. But users may change the parameter freely to get different comparative results. The influences of different parameters on the display of the aligned genomes with

Circos are shown in Figure 3. The details of how the merge threshold works is shown in Additional file 2: Figure S2A.

Once a user submits their job to our server, PGC will align both genome sequences using NUCmer, from the MUMmer package [36]. Our pipeline will process the output files generated by NUCmer and generate a few input files (configuration file, karyotype file and so on) to be used to generate a circular graphic plot using Circos, which is a powerful tool to display the relationship between the two aligned genomes [37]. The circular representation of the two aligned genomes provides a clear view of the similarities and differences (e.g. indels and rearrangements) in genome structure of the selected *Helicobacter* strains. The detailed workflow on how PGC works after users submit their jobs to our server is shown in Figure 4.

As a case study, we compared the genomes of two closely related *Helicobacter* strains, *H. pylori* J99 and *H. pylori* India7 using PGC (Figure 5). In general, both genomes are conserved. However, we still can observe differences e.g. indels between the genomes. Further analysis on one of the large indels in the genome of *H. pylori* India7 revealed a putative intact prophage as predicted by PHAge Search Tool (PHAST) [38]. The observation of the intact prophage suggests that it might be recently inserted into the genome of *H. pylori* India7 through Horizontal Gene Transfer (HGT). The introduction of the intact prophage in the genome of *H. pylori* India7 strain has probably conferred pathogenicity to the bacterial host and may represent an adaptation to different environments. This example demonstrates the usefullness of PGC for viewing and interpreting the genetic differences between two genomes.

Although a similar tool, called Circoletto [27] is available, PGC has some advantages over this online tool. Firstly, while Circoletto aligns sequences using BLAST (local alignment) the alignment algorithm used in PGC is based on the NUCmer (global alignment) package in MUMmer, which is suitable for large-scale and rapid genome alignment. Secondly, PGC provides many useful options for users. For instance, a user can adjust settings such as minimum percent genome identity (%), merging of links/ribbons according to MT, and the removal of links according to the user-defined LT through the provided online form. Thirdly, in the circular layout generated by PGC, a histogram track showing the percentage of mapped regions along the genome is provided. This


Link Threshold = 1,000 bp

Link Threshold = 10,000 bp

Link Threshold = 20,000 bp

Merge Threshold = 1,000 bp

Merge Threshold = 10,000 bp

Merge Threshold = 20,000 bp

Figure 3 Output of different cut-offs for the Link Threshold (LT) and Merge Threshold (MT) when comparing *H. pylori* R056A and *H. pylori* P12. Different user-defined cut-offs affect the output display of the two aligned genomes. The top three plots were generated at genome identity of 95% and MT of 0 bp, but at different LT cut-offs. The three plots at the bottom were generated at genome identity of 95% and LT of 1,000 bp, but at different MT cut-offs. Each half circle (either left or right) represents each separate genome/assembly. The coloured links show the homologous regions in the two selected genomes. The green track is the alignment histogram; each 10 Kbp window in the diagram is represented by a histogram bar and the height of each bar illustrates the total number of bases of the opposite genome aligned to this 10 Kbp window region. The upper border of the grey area delineates 10 Kbp height. If the height is higher than the 10 Kbp, it may indicate non-specific alignment or windows containing repetitive regions. A trough may indicate an unmapped region which could be an insertion e.g. prophage insertion. We can clearly observe how different user-defined thresholds affect the display of the two aligned genomes.

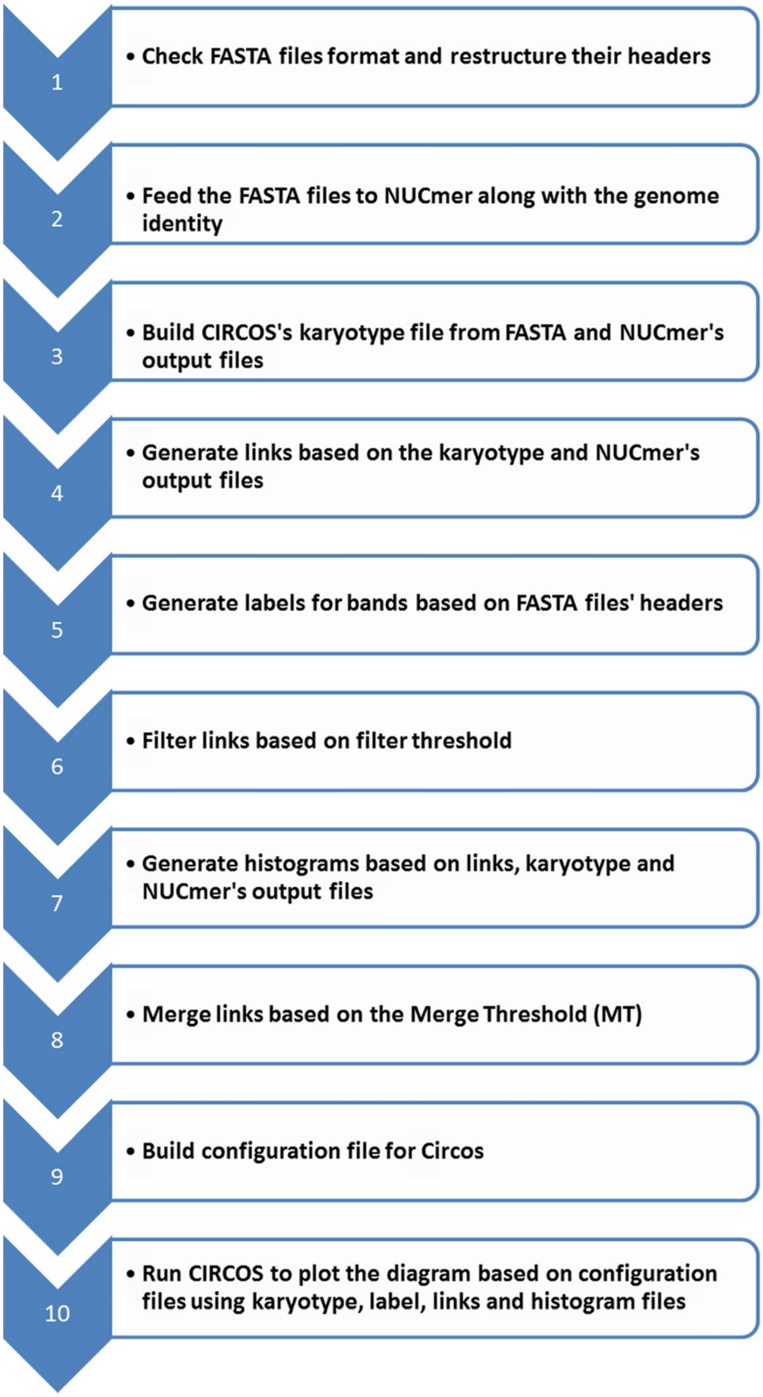

1
• Check FASTA files format and restructure their headers
2
• Feed the FASTA files to NUCmer along with the genome identity
3
• Build CIRCOS's karyotype file from FASTA and NUCmer's output files
4
• Generate links based on the karyotype and NUCmer's output files
5
• Generate labels for bands based on FASTA files' headers
6
• Filter links based on filter threshold
7
• Generate histograms based on links, karyotype and NUCmer's output files
8
• Merge links based on the Merge Threshold (MT)
9
• Build configuration file for Circos
10
• Run CIRCOS to plot the diagram based on configuration files using karyotype, label, links and histogram files

Figure 4 A flow chart that briefly illustrates the processes involved in PGC Pipeline after a job is submitted to our server.

track is very useful and helps users to identify putative indels and repetitive regions in the *Helicobacter* genomes. Additional file 2: Figure S2B shows how the data in the histogram track is calculated. Besides the Circoletto, RCircos is another tool with a similar function to PGC, which was developed by Zhang et al. [39]. RCircos was developed using R packages that come with R base installation. The package supports Circos 2D data track plots such as scatter, line, histogram, heat map, tile, connectors, links, and text labels [40]. However, unlike PGC which has a user-friendly interface and is easy to use without knowledge in programming, users need to have knowledge in the R programming language in order to run the RCircos and no user-friendly interface is provided.

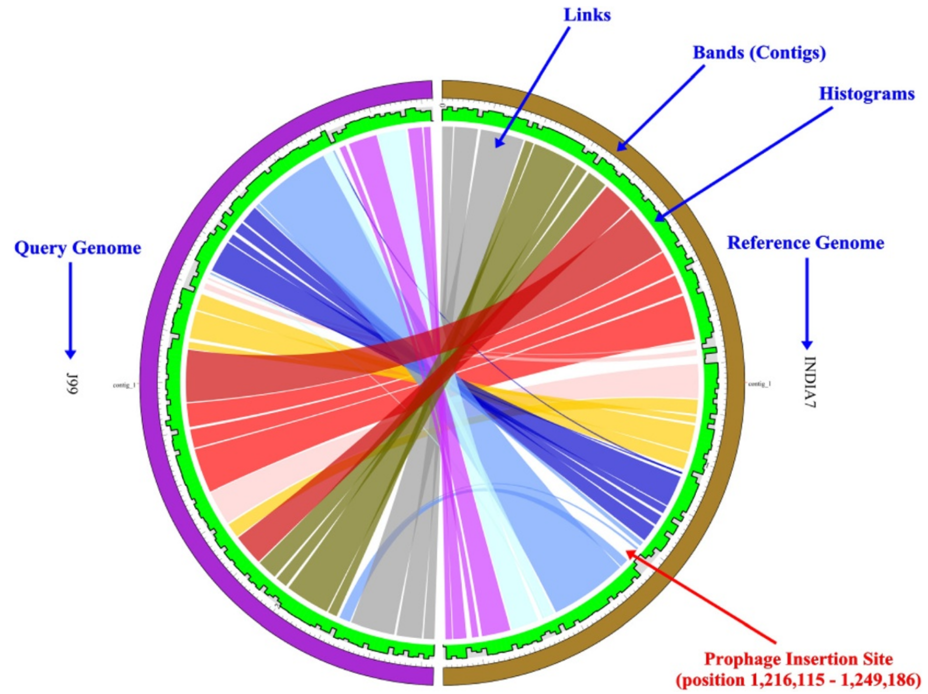


Figure 5 Analysis of two closely related *Helicobacter* strains using PGC tool. Genome comparison are performed between *H. pylori* J99 and *H. pylori* India7. The "flat" pattern in the histogram track indicates the two genomes are generally conserved/similar, whereas the gaps may indicate unaligned genomic regions e.g. indels. A large indel was predicted as prophage sequence.

A newly designed pathogenomics profiling tool (PathoProT) for comparative pathogenomics analysis

Virulence factors are molecules present in bacteria, which are responsible for causing disease in the host or converting non-pathogenic bacteria into pathogens [41,42]. The availability of sequenced genomes of different *Helicobacter* species makes the comparative analyses of virulence factors in the *Helicobacter* pathogen genomes feasible and may provide new insights into pathogen evolution and the diverse virulence strategies employed. Understanding the pathogenic mechanisms of these pathogens would aid in the treatment and prevention of *Helicobacter*-caused diseases.

To identify virulence genes and facilitate the comparative pathogenomics analysis of multiple bacterial species/strains, we have developed a unique Pathogenomics Profiling Tool (PathoProT). PathoProT predicts virulence genes based on sequence similarity by BLASTing all RAST-predicted protein sequences in user-selected strains against the VFDB [43-45]. A gene will be defined as a virulence gene if it has a BLAST hit that meets defined cut-offs e.g. 50% sequence identity and 50% sequence completeness set by the users. Once the putative virulence genes are identified in each user-defined strain, PathoProT will cluster (agglomerative hierarchical cluster analysis) the virulence genes and strains based on their virulence gene profiles and visualize them as a heat map with dendrograms. Through the heat map, users can examine the similarities and differences of the virulence gene profiles between different groups of strains e.g. non-pathogenic versus pathogenic strains (Figure 6). The detailed steps involved in our PathoProT pipeline after the BLAST searches are completed are shown in Figure 7. For more details on the usage of PathoProT tool, we included a "Help" page in the PathoProT tool, aimed to provide definitions and support to users.

Figure 6, gives a bird eye's view of the virulence genes that are present and widely distributed across all *Helicobacter* species. In general, it is clearly shows that *H. pylori* strains have more virulence genes compared to other species, which may explain their high virulence [46,47].

Interestingly, *H. hepaticus* ATCC51449 harbours three unique virulence genes, which are *cdtA*, *cdtB* and *cdtC*. The cytolethal distending toxins (CDTs) constitute the most recently discovered family of bacterial protein toxins. CDTs are unique among bacterial toxins as they have the ability to induce DNA double strand breaks in both proliferating and non-proliferating cells, thereby causing irreversible cell cycle arrest or death of the target cells [48]. It has been shown that CDTs encoded by the three genes, *cdtA*, *cdtB*, and *cdtC* are required for cytotoxicity [49]. When cdtA, *cdtB*, and *cdtC* are present together, the CDTs interact with one another to form an active tripartite holotoxin. The presence of these three genes in

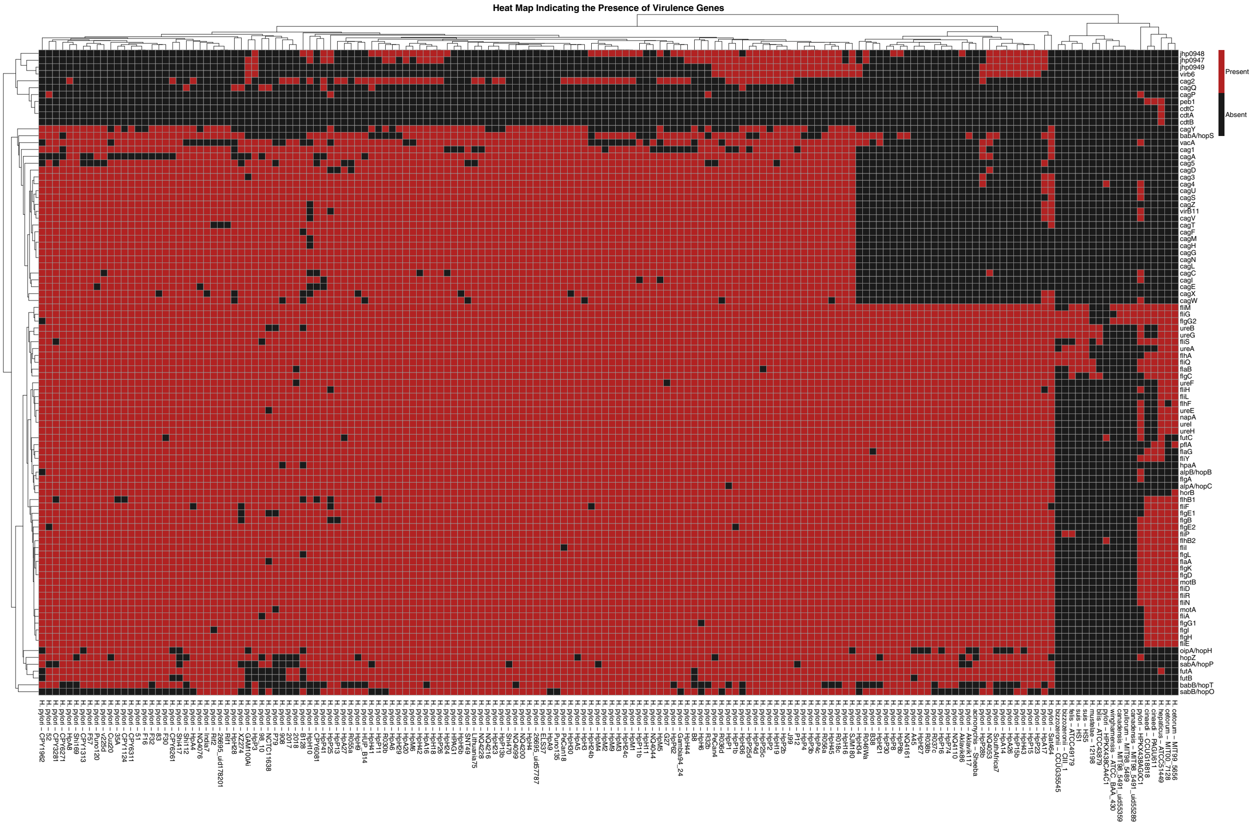


Figure 6 A heat map generated by PathoProT. All strains in HelicoBase were used to generate this heat map.

H. hepaticus ATCC51449 is supported by a recent study by Vincent et al., who identified these virulence genes in *H. hepaticus* species [50].

In summary, we have demonstrated that PathoProT can be used to identify virulence genes in *Helicobacter* strains by sequence homology. Moreover, comparative pathogenomics analysis can be easily performed to compare strains/groups of strains e.g. non-pathogenic strains versus pathogenic strains, which can give better insights into the biology, evolution and virulence of the *Helicobacter* strains of interest. In other words, PathoProT can be used to answer interesting biological questions including what are the conserved virulence genes in a group of *Helicobacter* strains and enables strain-specific/group-specific virulence genes to be easily viewed in the generated heat map.

Other tools

BLAST is included in HelicoBase to allow for easy similarity searching for sequences of interest [51]. The built-in BLAST in HelicoBase provides two main functions: (1) standard BLAST which will search the provided query sequence against genome or ORF sequences (either nucleotide or protein) in HelicoBase; (2) VFDB BLAST which searches the provided query sequence against the Virulence Factor Database (VFDB) [43,44,52]. Users can use VFDB BLAST if they want to determine whether their sequence of interest is a virulence gene based on sequence homology.

JBrowse is another tool that we have integrated into HelicoBase to give users a seamless browsing experience [30] (Figure 8). This next generation AJAX-based genome browser built with JavaScript and HTML5 enables the user to explore the genome of interest with unparalleled speed and scales easily to multi-gigabase genomes and deep-coverage sequencing [45]. JBrowse preserves the user's sense of location by avoiding discontinuous transitions, offering smooth and fast animated panning, zooming, navigation and track selection. With the advances in next-generation sequencing technologies and bioinformatics tools, we anticipate that many more *Helicobacter* genomes will be sequenced and annotated. Therefore, a user-friendly JBrowse that allows rapid and seamless browsing of high volumes of genomic data will be a major advantage.

HelicoBase development and implementation

HelicoBase rests on MySQL version 14.12 (http://www.mysql.com) and was hosted using Ubuntu Lucid 10.04 Web server application (http://www.ubuntu.com).

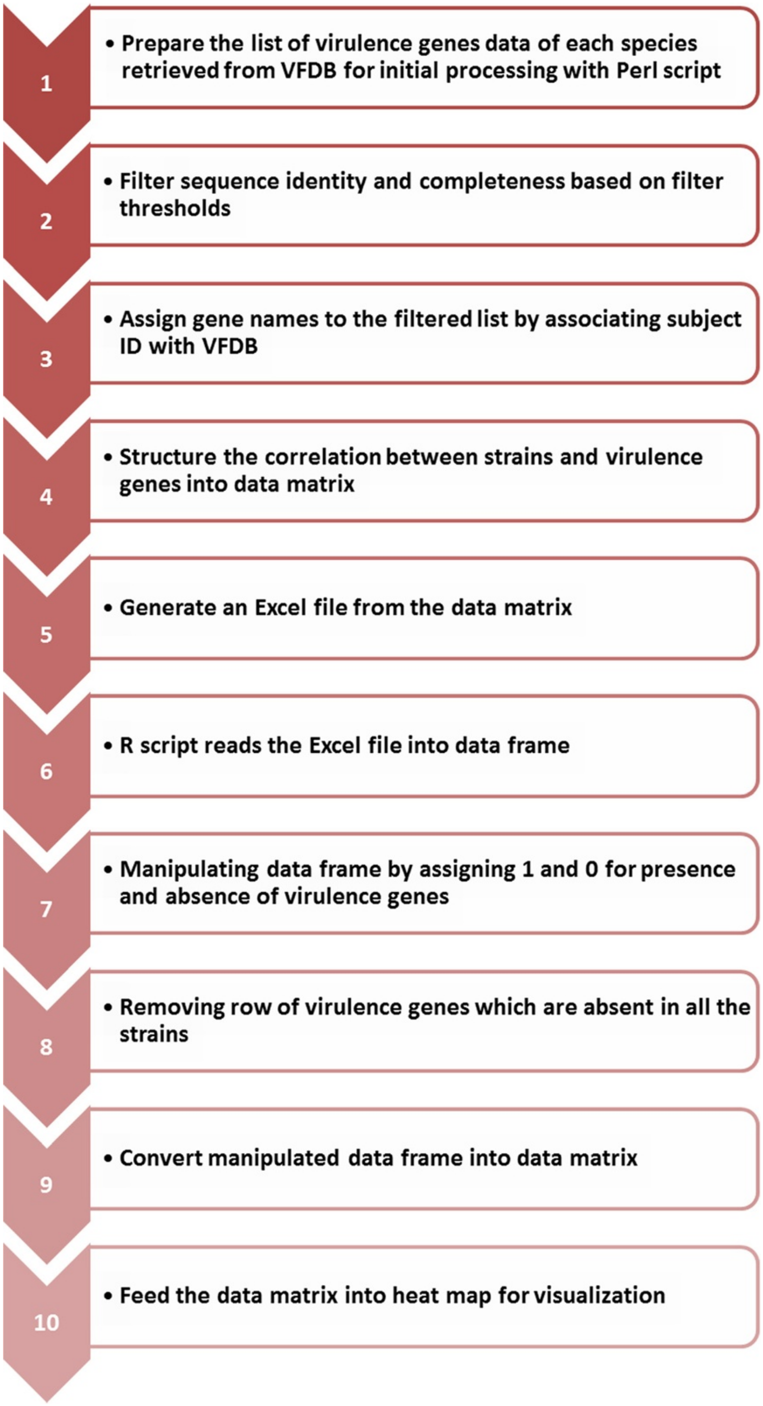
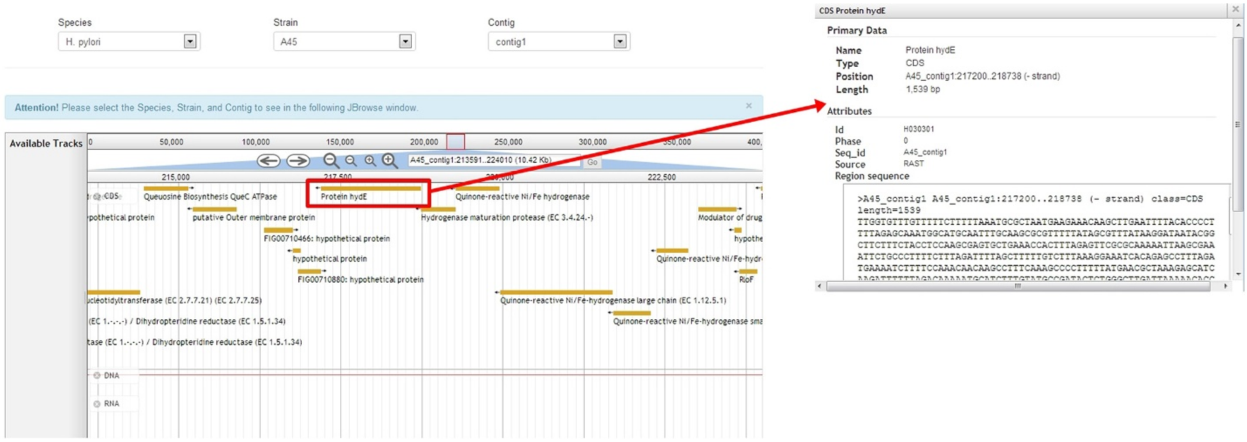

1
• Prepare the list of virulence genes data of each species retrieved from VFDB for initial processing with Perl script
2
• Filter sequence identity and completeness based on filter thresholds
3
• Assign gene names to the filtered list by associating subject ID with VFDB
4
• Structure the correlation between strains and virulence genes into data matrix
5
• Generate an Excel file from the data matrix
6
• R script reads the Excel file into data frame
7
• Manipulating data frame by assigning 1 and 0 for presence and absence of virulence genes
8
• Removing row of virulence genes which are absent in all the strains
9
• Convert manipulated data frame into data matrix
10
• Feed the data matrix into heat map for visualization

Figure 7 Flow chart that briefly illustrates the processes involved in PathoProT Pipeline.

Development used a combination of PHP 5.3 and Perl 5 languages, with CodeIgniter 2.1.3 framework for web tier and Twitter Bootstrap front-end framework for the presentation layer.

The web server architecture was designed to be scalable and combined with a flexible PHP coding interface enables users with different devices to connect to HelicoBase in a fast and readable manner. For tools such as BLAST, VFDB BLAST, PathoProT, and PGC users can submit their analysis jobs through the provided interfaces and these jobs will be submitted to the cluster server in a firewall-enabled secure process. The job scheduler in turn makes it possible for the application server to process the submitted jobs in a fair and parallel manner with fast



Species: H. pylori
Strain: A45
Contig: contig1

Attention! Please select the Species, Strain, and Contig to see in the following JBrowse window.

A45_contig1:213591..224010 (10.42 Kb)

CDS Protein hydE

Primary Data

Name: Protein hydE
Type: CDS
Position: A45_contig1:217200..218738 (- strand)
Length: 1,539 bp

Attributes

Id: H030301
Phase: 0
Seq_id: A45_contig1
Source: RAST
Region sequence

```
>A45_contig1 A45_contig1:217200..218738 (- strand) class=CDS
length=1539
TTGGTGTTTGTTTTTCTTTTTAAATGCGCTAATGAAGAAACAAGCTTGAATTTTACACCCCT
TTTAGAGCAAATGGCATGCAATTTGCAAGCGCGTTTTTATAGCGTTTATAAGGATAATACGG
CTTCTTTCTACCTCCAAGCGAGTGCTGAAACCACTTTAGAGTTCGCGCAAAAATTAAGCGAA
ATTCTGCCCTTTTCTTTAGATTTTAGCTTTTTGTCTTTAAAGGAAATCACAGAGCCTTTAGA
TGAAAATCTTTTCCAAACAACAAGCCTTTCAAAGCCCCTTTTTATGAACGCTAAAGAGCATC
```

Available Tracks
0
50,000
100,000
150,000
200,000
250,000
300,000
350,000
400,
Go
215,000
217,500
220,000
222,500
CDS
Queuosine Biosynthesis QueC ATPase
Protein hydE
Quinone-reactive Ni/Fe hydrogenase
pothetical protein
putative Outer membrane protein
Hydrogenase maturation protease (EC 3.4.24.-)
Modulator of drug
FIG00710466: hypothetical protein
hypothe
hypothetical protein
Quinone-reactive Ni/Fe-hydr
FIG00710880: hypothetical protein
RloF
cleotidyltransferase (EC 2.7.7.21) (EC 2.7.7.25)
Quinone-reactive Ni/Fe-hydrogenase large chain (EC 1.12.5.1)
(EC 1.-.-.-) / Dihydropteridine reductase (EC 1.5.1.34)
Quinone-reactive Ni/Fe-hydrogenase sma
tase (EC 1.-.-.-) / Dihydropteridine reductase (EC 1.5.1.34)
DNA
RNA

Figure 8 A sample output of JBrowse in HelicoBase. A genomic region of the contig 1 of *H. pylori* A45 was visualised in JBrowse. Clicking on gene encoding for protein hydE displays a pop-up window with the useful information associated with the gene.

processing speed. Our cluster computer (5 nodes, 12 CPUs for each node and 625GB of RAM in total) prompts the database server to run in a fast fibre-optic internal network to retrieve necessary information needed to process the submitted jobs. Meanwhile, our MySQL data structures were formed in a manner that supports fast and localized searching which makes the user-website interactions fast and also user-friendly. To construct HelicoBase, we used the several software components: RAST [33], BLAST [51], MUMmer [53], PSORTb [54], Circos [37] and JBrowse [30].

Discussion and conclusion

With advances in high-throughput sequencing technologies, it is imperative that the abundant data generated can be easily accessible for analysis. With HelicoBase we aim to provide a one-stop resource platform that will make it easy to access and analyse whole-genome genomic data and information for *Helicobacter* spp. through an organised and user-friendly interface. PGC and PathoProT are some of the bioinformatics tools for comparative analysis implemented in HelicoBase which allow researchers to conveniently assimilate and explore the data in an intuitive manner.

HelicoBase will be updated from time to time as more genome sequences of *Helicobacter* spp. become available. To accelerate the development of HelicoBase, we encourage researchers to email us at girg@um.edu.my if they would like to share their annotations and related data with us. Suggestions on improving HelicoBase are most welcome.

Availability and requirements

HelicoBase is available online at http://helicobacter.um.edu.my. All sequences and annotations described in this paper can be downloaded from that site.

Additional files

Additional file 1: Figure S1. Classification of CDS Subcellular Localization in HelicoBase. Protein-coding genes with ambiguous and low subcellular scores were classified into unknown category.

Additional file 2: Figure S2. (A) A diagram showing how the merge threshold works with the merging process by PGC tool. Merge Threshold provides users with the ability to ignore minimal spaces between adjacent links. Adjacent links are those which are adjacent in their position in both of the genomes. (B) A diagram showing how the data in histogram track was calculated based on different scenarios. Basically histogram bars delineate the total length of links (bp) that are mapped to a particular window. The window denotes 10 kbp slices of genomes. Note that having bar with the height equal to borderline does not necessarily mean that the whole window is covered with links. All this information is available on the 'Help' icon provided in PGC tool.

Competing interest

All authors declare that they have no competing interest.

Authors' contributions

HH, MYA, WYW, NVRM, CCS and SWC designed and developed the database system. MYA, HF, SYT, VR generated annotations and analysed data. HH, MYA, SWC, JV, MFL, and GJW wrote the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

We would like to thank Professor Dr Robert White (Department of Physiology, Development and Neuroscience, University of Cambridge) for his assistance in proofreading the manuscript. We also thank Amir Hessam Yazdi for providing IT support in this project. Special thanks to Aini Mohamed Yusoff for helping us to compile the genome sequences.

Funding

This project was funded by University of Malaya and Ministry of Education (MOHE), Malaysia under the High Impact Research (HIR) grant UM.C/HIR/MOHE/08.

Author details

[1]Genome Informatics Research Laboratory, High Impact Research (HIR) Building, University of Malaya, 50603 Kuala Lumpur, Malaysia. [2]Department of Oral Biology and Biomedical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia. [3]Faculty of Medicine and Health Sciences, University Putra Malaysia, 43400 Serdang, Selangor Darul Ehsan, Malaysia. [4]Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia. [5]Department of Medical Microbiology, Faculty of Medicine Building, University of Malaya, 50603 Kuala Lumpur, Malaysia.

Received: 9 January 2014 Accepted: 9 July 2014
Published: 16 July 2014

References

1. Sycuro LK, Pincus Z, Gutierrez KD, Biboy J, Stern CA, Vollmer W, Salama NR: **Peptidoglycan crosslinking relaxation promotes< i> Helicobacter pylori </i>'s helical shape and stomach colonization.** *Cell* 2010, **141:**822–833.
2. Dick JD: **Helicobacter (Campylobacter) pylori: a new twist to an old disease.** *Annu Rev Microbiol* 1990, **44:**249–269.
3. Owen R: **Helicobacter-species classification and identification.** *Br Med Bull* 1998, **54:**17–30.
4. Fox JG, Lee A: **The role of Helicobacter species in newly recognized gastrointestinal tract diseases of animals.** *Lab Anim Sci* 1997, **47:**222–255.
5. Goodman KJ, Correa P: **The transmission of Helicobacter pylori. A critical review of the evidence.** *Int J Epidemiol* 1995, **24:**875–887.
6. Axon A: **Review article is Helicobacter pylori transmitted by the gastro-oral route?** *Aliment Pharmacol Ther* 1995, **9:**585–588.
7. Dowsett S, Kowolik M: **Oral Helicobacter pylori: can we stomach it?** *Crit Rev Oral Biol Med* 2003, **14:**226–233.
8. Brown LM: **Helicobacter pylori: epidemiology and routes of transmission.** *Epidemiol Rev* 2000, **22:**283.
9. Kuipers E, Nelis G, Klinkenberg-Knol E, Snel P, Goldfain D, Kolkman J, Festen H, Dent J, Zeitoun P, Havu N: **Cure of Helicobacter pylori infection in patients with reflux oesophagitis treated with long term omeprazole reverses gastritis without exacerbation of reflux disease: results of a randomised controlled trial.** *Gut* 2004, **53:**12–20.
10. Sipponen P, Hyvärinen H: **Role of Helicobacter pylori in the pathogenesis of gastritis, peptic ulcer and gastric cancer.** *Scand J Gastroenterol* 1993, **28:**3–6.
11. Kuipers E: **Helicobacter pylori and the risk and management of associated diseases: gastritis, ulcer disease, atrophic gastritis and gastric cancer.** *Aliment Pharmacol Ther* 1997, **11:**71–88.
12. Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, van der Merwe SW: **An African origin for the intimate association between humans and Helicobacter pylori.** *Nature* 2007, **445:**915–918.
13. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, Blaser MJ, Graham DY, Vacher S, Perez-Perez GI: **Traces of human migrations in Helicobacter pylori populations.** *Science* 2003, **299:**1582–1585.
14. Achtman M, Azuma T, Berg DE, Ito Y, Morelli G, Pan ZJ, Suerbaum S, Thompson SA, Van Der Ende A, Van Doorn LJ: **Recombination and clonal groupings within Helicobacter pylori from different geographical regions.** *Mol Microbiol* 1999, **32:**459–470.
15. Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, Wu J-Y, Maady A, Bernhöft S, Thiberge J-M, Phuanukoonnon S: **The peopling of the Pacific from a bacterial perspective.** *Science* 2009, **323:**527–530.
16. Devi SM, Ahmed I, Francalacci P, Hussain MA, Akhter Y, Alvi A, Sechi LA, Mégraud F, Ahmed N: **Ancestral European roots of Helicobacter pylori in India.** *BMC Genomics* 2007, **8:**184.
17. Tay CY, Mitchell H, Dong Q, Goh K-L, Dawes IW, Lan R: **Population structure of Helicobacter pylori among ethnic groups in Malaysia: recent acquisition of the bacterium by the Malay population.** *BMC Microbiol* 2009, **9:**126.
18. Duncan SS, Bertoli MT, Kersulyte D, Valk PL, Tamma S, Segal I, McClain MS, Cover TL, Berg DE: **Genome sequences of three hpAfrica2 strains of Helicobacter pylori.** *Genome Announc* 2013, **1:**e00729-00713.
19. Behrens W, Bönig T, Suerbaum S, Josenhans C: **Genome sequence of Helicobacter pylori hpEurope strain N6.** *J Bacteriol* 2012, **194:**3725–3726.
20. Uchiyama J, Takeuchi H, Kato S-i, Takemura-Uchiyama I, Ujihara T, Daibata M, Matsuzaki S: **Complete genome sequences of two Helicobacter pylori bacteriophages isolated from Japanese patients.** *J Virol* 2012, **86:**11400–11401.
21. Clancy CD, Forde BM, Moore SA, O'Toole PW: **Draft genome sequences of Helicobacter pylori strains 17874 and P79.** *J Bacteriol* 2012, **194:**2402–2402.
22. Goto T, Ogura Y, Hirakawa H, Tomida J, Morita Y, Akaike T, Hayashi T, Kawamura Y: **Complete genome sequence of Helicobacter cinaedi strain PAGU611, isolated in a case of human bacteremia.** *J Bacteriol* 2012, **194:**3744–3745.
23. Winsor GL, Van Rossum T, Lo R, Khaira B, Whiteside MD, Hancock RE, Brinkman FS: **Pseudomonas genome database: facilitating user-friendly, comprehensive comparisons of microbial genomes.** *Nucleic Acids Res* 2009, **37:**D483–D488.
24. Winsor GL, Lam DK, Fleming L, Lo R, Whiteside MD, Nancy YY, Hancock RE, Brinkman FS: **Pseudomonas genome database: improved comparative analysis and population genomics capability for Pseudomonas genomes.** *Nucleic Acids Res* 2011, **39:**D596–D600.
25. Winsor GL, Khaira B, Van Rossum T, Lo R, Whiteside MD, Brinkman FS: **The Burkholderia genome database: facilitating flexible queries and comparative analyses.** *Bioinformatics* 2008, **24:**2803–2804.
26. Furumichi M, Sato Y, Omata T, Ikeuchi M, Kanehisa M: **CYORF: community annotation of cyanobacteria genes.** *Genome Inform Ser* 2002, **13:**402–403.
27. Heydari H, Wee WY, Lokanathan N, Hari R, Yusoff AM, Beh CY, Yazdi AH, Wong GJ, Ngeow YF, Choo SW: **MabsBase: a mycobacterium abscessus genome and annotation database.** *PLoS ONE* 2013, **8:**e62443.
28. Uchiyama I: **MBGD: microbial genome database for comparative analysis.** *Nucleic Acids Res* 2003, **31:**58–62.
29. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P: **IMG: the integrated microbial genomes database and comparative analysis system.** *Nucleic Acids Res* 2012, **40:**D115–D122.
30. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH: **JBrowse: a next-generation genome browser.** *Genome Res* 2009, **19:**1630–1638.
31. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35:**D61–D65.
32. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2005, **33:**D501–D504.
33. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M: **The RAST server: rapid annotations using subsystems technology.** *BMC Genomics* 2008, **9:**75.
34. Liu Z, Ma H, Goryanin I: **A semi-automated genome annotation comparison and integration scheme.** *BMC Bioinformatics* 2013, **14:**172.
35. Nancy YY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.** *Bioinformatics* 2010, **26:**1608–1615.
36. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5:**R12.
37. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: an information aesthetic for comparative genomics.** *Genome Res* 2009, **19:**1639–1645.
38. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS: **PHAST: a fast phage search tool.** *Nucleic Acids Res* 2011, **39:**W347–W352.
39. Zhang H, Meltzer P, Davis S: **RCircos: an R package for Circos 2D track plots.** *BMC Bioinformatics* 2013, **14:**1–5.
40. Zhang H, Zhang MH, Plot RCI, Point RD, Plot RGC: *Package 'RCircos'.* ; 2013.
41. Litwin CM, Calderwood S: **Role of iron in regulation of virulence genes.** *Clin Microbiol Rev* 1993, **6:**137–149.
42. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405:**299–304.
43. Yang J, Chen L, Sun L, Yu J, Jin Q: **VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics.** *Nucleic Acids Res* 2008, **36:**D539–D542.
44. Chen L, Xiong Z, Sun L, Yang J, Jin Q: **VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors.** *Nucleic Acids Res* 2012, **40:**D641–D645.
45. Westesson O, Skinner M, Holmes I: **Visualizing next-generation sequencing data with JBrowse.** *Brief Bioinform* 2013, **14:**172–177.
46. Covacci A, Telford JL, Del Giudice G, Parsonnet J, Rappuoli R: **Helicobacter pylori virulence and genetic geography.** *Science* 1999, **284:**1328–1333.
47. Kusters JG, van Vliet AH, Kuipers EJ: **Pathogenesis of Helicobacter pylori infection.** *Clin Microbiol Rev* 2006, **19:**449–490.
48. Thelestam M: **Cytolethal distending toxins.** In *Reviews of physiology, biochemistry and pharmacology.* Springer; 2005:111–133.
49. Lara-Tejero M, Galan JE: **CdtA, CdtB, and CdtC form a tripartite complex that is required for cytolethal distending toxin activity.** *Infect Immun* 2001, **69:**4358–4365.

50. Young VB, Knox KA, Schauer DB: **Cytolethal distending toxin sequence and activity in the enterohepatic pathogen Helicobacter hepaticus.** *Infect Immun* 2000, **68:**184–191.
51. Ye J, McGinnis S, Madden TL: **BLAST: improvements for better sequence analysis.** *Nucleic Acids Res* 2006, **34:**W6–W9.
52. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q: **VFDB: a reference database for bacterial virulence factors.** *Nucleic Acids Res* 2005, **33:**D325–D328.
53. Delcher AL, Salzberg SL, Phillippy AM: **Using MUMmer to identify similar regions in large sequence sets.** *Curr Protoc Bioinformatics* 2003, **Chapter 10:**Unit 10 13.
54. Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, Lambert C, Nakai K, Brinkman FS: **PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria.** *Nucleic Acids Res* 2003, **31:**3613–3617.

doi:10.1186/1471-2164-15-600
Cite this article as: Choo *et al.*: HelicoBase: a *Helicobacter* genomic resource and analysis platform. *BMC Genomics* 2014 **15**:600.