Tan *et al. BMC Bioinformatics*
DOI 10.1186/s12859-014-0422-y

BMC Bioinformatics

DATABASE Open Access

YersiniaBase: a genomic resource and analysis platform for comparative analysis of *Yersinia*

Shi Yang Tan[1,2], Avirup Dutta[1], Nicholas S Jakubovics[3], Mia Yang Ang[1,2], Cheuk Chuen Siow[1], Naresh VR Mutha[1], Hamed Heydari[1,4], Wei Yee Wee[1,2], Guat Jah Wong[1,2] and Siew Woh Choo[1,2*]

Abstract

Background: *Yersinia* is a Gram-negative bacteria that includes serious pathogens such as the *Yersinia pestis*, which causes plague, *Yersinia pseudotuberculosis*, *Yersinia enterocolitica*. The remaining species are generally considered non-pathogenic to humans, although there is evidence that at least some of these species can cause occasional infections using distinct mechanisms from the more pathogenic species. With the advances in sequencing technologies, many genomes of *Yersinia* have been sequenced. However, there is currently no specialized platform to hold the rapidly-growing *Yersinia* genomic data and to provide analysis tools particularly for comparative analyses, which are required to provide improved insights into their biology, evolution and pathogenicity.

Description: To facilitate the ongoing and future research of *Yersinia*, especially those generally considered non-pathogenic species, a well-defined repository and analysis platform is needed to hold the *Yersinia* genomic data and analysis tools for the *Yersinia* research community. Hence, we have developed the YersiniaBase, a robust and user-friendly *Yersinia* resource and analysis platform for the analysis of *Yersinia* genomic data. YersiniaBase has a total of twelve species and 232 genome sequences, of which the majority are *Yersinia pestis*. In order to smooth the process of searching genomic data in a large database, we implemented an Asynchronous JavaScript and XML (AJAX)-based real-time searching system in YersiniaBase. Besides incorporating existing tools, which include JavaScript-based genome browser (JBrowse) and Basic Local Alignment Search Tool (BLAST), YersiniaBase also has in-house developed tools: (1) Pairwise Genome Comparison tool (PGC) for comparing two user-selected genomes; (2) Pathogenomics Profiling Tool (PathoProT) for comparative pathogenomics analysis of *Yersinia* genomes; (3) YersiniaTree for constructing phylogenetic tree of *Yersinia*. We ran analyses based on the tools and genomic data in YersiniaBase and the preliminary results showed differences in virulence genes found in *Yersinia pestis* and *Yersinia pseudotuberculosis* compared to other *Yersinia* species, and differences between *Yersinia enterocolitica* subsp. *enterocolitica* and *Yersinia enterocolitica* subsp. *palearctica*.

Conclusions: YersiniaBase offers free access to wide range of genomic data and analysis tools for the analysis of *Yersinia*. YersiniaBase can be accessed at http://yersinia.um.edu.my.

Keywords: YersiniaBase, *Yersinia*, Genomic resources, Comparative analysis

Background

Yersinia is a genus of Gram-negative rod shaped bacteria belonging to the Enterobacteriaceae family. The majority of *Yersinia* species are environmental and non-disease causing in mammals and can be isolated from many locations such as fresh water and soil. Some species can also be isolated from sources like patient's body or even dog faeces [1-4]. Three of the *Yersinia* species are known to cause disease in human: the plague bacillus *Y. pestis* and the enteropathogens *Y. enterocolitica* and *Y. pseudotuberculosis*. The most notorious and virulent species in this genus, *Y. pestis*, the causative agent of plague, is a recently emerged lineage from *Y. pseudotuberculosis* [3]. Plague, which has caused three pandemics in human history [5,6], is mainly a zoonotic infection caused by *Y. pestis*, a rodent pathogen that

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

can be transmitted to humans through the bite of an infected flea. Infection can lead to acute febrile lymphadenitis also called bubonic plague or may cause septicemia, pharyngeal, meningeal and fatal pneumonia plague [7]. The World Health Organization (WHO) has categorized this disease as re-emergent in some parts of the world as plague reappeared in Malawi, Mozambique, and India in 1994, in Algeria in 2003, and in Libya in 2009 [7-12]. The potential re-emergence of plague is a significant worldwide public health hazard. The problem has become even more severe with the emergence of multi-drug resistant strains and the potential use of plague as a bioweapon [13-15].

Yersinia enterocolitica, the other notorious human pathogen of the genus *Yersinia* is a zoonotic agent that causes gastrointestinal disease in humans. This species is also responsible for causing reactive arthritis and erythema nodosum. Enteropathogenic *Yersinia*, the etiological agents for yersiniosis, are generally acquired by consuming raw or inadequately thermally processed pork or milk as well as vegetable products and ready-to-eat meals [16,17]. As per the published annual e-reports of the European Food Safety Authority (EFSA) on the prevalence of zoonoses and foodborne pathogens in 2012, yersiniosis was listed as the third-most-common enteropathogenic disease after campylobacteriosis and salmonellosis [18].

All the three human pathogens share the similarity of carrying an extrachromosomal 70-kb pYV virulence plasmid that encodes the Ysc type III secretion system (TTSS), the protein microinjection apparatus and *Yersinia* outer proteins (Yops), a set of translocated effector proteins [19]. The TTSS is highly conserved among Gram-negative pathogens and is used by *Yersinia* pathogens to inject six different Yop effectors into the host cell [20]. These Yop effectors allow *Yersinia* to resist phagocytosis and also interfere with signalling pathways within the host cell [19].

Although the three pathogenic *Yersinia* species have been extensively studied for many years, the other *Yersinia* species, which include *Y. frederiksenii, Y. intermedia, Y. kristensenii, Y. bercovieri, Y. massiliensis, Y. mollaretii, Y. rohdei, Y. ruckeri,* and *Y. aldovae* have remained largely ignored and are not well studied especially at the genome level. These species are isolated from the environment as well as from infected humans and do not possess typical *Yersinia* virulence markers. Formerly these species were grouped under *Y. enterocolitica* as different biogroups and they have been considered as *Y. enterocolitica*-like species in some papers although they are clearly distinct from *Y. enterolitica* [2,21]. Although this group of species is generally considered non-pathogenic, some strains have been associated with infections in mankind [21,22]. Therefore, it is important for researchers to analyse them further and to determine whether they have the potential to become pathogenic, since the acquisition of virulence genes through horizontal gene transfer is very common in prokaryotes [23].

With *Y. pestis* being already declared as re-emergent, *Y. enterocolitica* being listed as the third-most-common enteropathogen and several ‘non-pathogenic’ strains of *Yersinia* showing signs of infections in humans, it is clear that the genus *Yersinia* is in a state of evolution and further study is required not only on its pathogenic strains but also on the non-pathogenic stains. In recent years, a new trend of collecting bacterial genomes into a single database has emerged as an effective way to analyse their genomes. Consequently many specialized genomic databases have been developed, especially for human disease pathogens. There are a number of databases, such the Microbial Genome Database for Comparative Analysis (MBGD) [24-26] and the Integrated Microbial Genomes (IMG) [27-30] system, which provide a wide array of microbial genomes including some *Yersinia* strains for comparative genomics. Another such database worth mentioning is PATRIC [31], which offers genomic and virulence factors information of some of the *Yersinia* strains, however does not provide functionalities for comparing, clustering and visualizing the virulence gene profiles of the user-selected *Yersinia* strains. These databases do not provide the option for comparative pathogenomics analysis based on the virulence factors of the strains. Moreover, most of the existing biological databases lack user-friendly web interfaces which allows real-time and fast querying and browsing of genomic data. In order to provide researchers with a specialized platform to access the genomic data of the pathogenic and the generally considered non-pathogenic *Yersinia* species, we have developed the genus specific database, YersiniaBase. YersiniaBase stores the *Yersinia* genomic resources along with various analysis tools particularly for comparative analysis of *Yersinia* strains. One of the features of this database is the PathoProT which has been developed in-house, with the intention of identifying the potential virulence markers in the *Yersinia* genus. This comparative analysis platform will help users gain a deeper insight into the different species of *Yersinia* especially in the area of their pathogenicity. Using the resources and the tools provided in YersiniaBase, we have carried out genomic analyses on several pathogenic and non-pathogenic strains of *Yersinia*. The preliminary results shown in this paper not only helped to identify some of the genome characteristics of *Yersinia* but also demonstrated the suitability and the usefulness of the database and the tools hosted by it for *Yersinia* research.

Construction and content

Genome sequences

A total of 232 genome sequences of *Yersinia*, which include both draft and complete genomes in FASTA format were downloaded from National Center for Biotechnology Information (NCBI) [32]. Approximately 90% of the downloaded genome sequences belong to *Y. pestis*. All generally considered non-pathogenic *Yersinia* species were only available as draft genomes. A list of draft and complete genomes for each *Yersinia* species is shown in Table 1.

Annotation of genome sequences

For consistency, all of the downloaded genome sequences were annotated by using Rapid Annotation using Subsystem Technology (RAST) [33]. RAST can predict open reading frame (ORF), function of the ORF, nucleotide sequence and amino acid sequence in user-submitted genome. Besides RAST, PSORTb version 3.0 [34] was used to determine the subcellular localization of the proteins predicted by the RAST engine. We also computed the hydrophobicity and molecular weight of the RAST-predicted proteins using in-house developed Perl scripts. The summary of genome annotations of each strain is shown in Additional file 1: Table S1.

Database and web interface design

We implemented a relational database using MySQL version 14.12 (http://www.mysql.com). All of the biological data of *Yersinia* strains were rearranged to fit into the designed database schema and stored in the MySQL database (Additional file 2: Figure S1).

Table 1 Number of draft and complete genomes of each *Yersinia* species in YersiniaBase

Species	Number of draft genomes	Number of complete genomes
Y. aldovae	1	0
Y. bercovieri	1	0
Y. enterocolitica	8	3
Y. frederiksenii	1	0
Y. intermedia	1	0
Y. kristensennii	1	0
Y. massiliensis	1	0
Y. mollaretii	1	0
Y. pestis	196	12
Y. pseudotuberculosis	0	4
Y. rohdei	1	0
Y. ruckeri	1	0

Majority of the strains belonged to *Yersinia pestis*.

A web interface was built by using HyperText Markup Language (HTML), HyperText Preprocessor (PHP), JavaScript, jQuery, Cascading Style Sheets (CSS) and AJAX. CodeIgniter version 2.1.3, a popular PHP framework was used in order to a provide model-view-controller (MVC), which can separate application data, presentation and background logic and process into three distinct modules. This allowed the source codes and *Yersinia* biological data to be arranged in a clear and organized manner, indirectly allowing easier future updating of YersiniaBase.

Bioinformatics tools

We developed several bioinformatics tools and integrated them into YersiniaBase. We used Python, Perl, BioPerl [35] and R languages to develop the PGC for comparing between two genomes through global alignment, PathoProT for generating heat map to visualize presence and absence of virulence genes in selected genomes and YersiniaTree to generate phylogenetic tree of the *Yersinia* strains based on their housekeeping genes and 16S rRNA. The use of these three popular scripting languages allowed us to create complex pipelines, that perform back-end calculations, aided communications between the web server and the application server and also easier transfer of data from the web server to the application server and vice versa.

Utility and discussion

Overview of the database pipeline

All the *Yersinia* genomes after being downloaded from National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov) were annotated using the Rapid Annotation using Subsystem Technology (RAST) server [33] for consistency. All the CDS, ORF and RNA sequences predicted by the RAST server along with their annotations and predicted protein functions were downloaded. Then using the PSORTb version 3.0 [34] the subcellular localization of the RAST predicted proteins were determined. Further analyses of the protein-coding genes were performed using Bio-Perl [35] and also in-house Perl script to obtain information such as calculation of GC content (%), predicted hydrophobicity (pH), and molecular weight (Da) of the encoded proteins. All this information was then stored in the MySQL database server.

Web interface and functionalities of YersiniaBase

On entering the home page of YersiniaBase, visitors can view the news & conferences, blogs & information and the most recent published papers that are related to *Yersinia*, which we manually compile from various sources. The “Browse” menu allows the visitors to view the list of *Yersinia* species currently available in YersiniaBase, with each “View Strains” button leading the visitors to the

"Browse Strains" page, displaying all available strains of that respective species. In the "Browse Strains" page a general description of that particular species is given along with a table listing the strains of that species. Each strain is linked to their corresponding taxonomic classification page in NCBI and also to their page in Genome Online Database (GOLD). Furthermore, by clicking on the "Details" icon, visitors can obtain more comprehensive information of that particular strain such as their source and time of isolation, which we retrieved from NCBI, along with the list of ORFs, their respective function, start and stop positions in a tabular fashion in the "Browse ORF" page. Apart from that, each ORF is linked to its corresponding UniProt page along with their ORF ID being linked to its corresponding page in NCBI. By clicking on the Contig ID of each ORF, the corresponding contig information available in NCBI can be accessed. The Details button of each ORF leads the visitor to the "ORF Detail" page displaying the detailed information of that ORF such their type, start and stop positions, lengths of nucleotide as well as amino acid sequences. It further provides information on functional classification, subsystem (if available), strand, subcellular localization, hydrophobicity (pH) and molecular weight (Da). The page also displays the amino acid and the nucleotide sequence of the ORF along with the Genome Browser. The "Genome Browser" menu links user to JBrowse [36] in YersiniaBase. JBrowse allows users to view the position of each ORF at each genome graphically.

We also integrated in-house tools and as well as other tools into YersiniaBase to add functionalities to this *Yersinia* research platform. The "Tools" menu allows user to perform a BLAST search [37] against the *Yersinia* strains curated in YersiniaBase, as well as an exclusive BLAST search against the Virulence Factor Database (VFDB) [38-40]. Besides the BLAST search, users can also perform pairwise alignment of any two *Yersinia* genomes present in YersiniaBase of their choice by using the PGC, draw a heat map of the virulence genes profiles by using the PathoProT or construct phylogenetic tree by using YersiniaTree.

"Search" menu allows user to search the functional classification of a specific species and strain by providing keyword or ORF ID. Besides performing searching, user can also download the genome sequence, ORF annotation details in table format, ORF sequence, ribonucleic acid (RNA) and coding sequence (CDS) through "Download" menu. The overview of the functionalities of YersiniaBase is shown in Additional file 3: Figure S2.

Browsing *Yersinia* strains

The browse page of this database summarizes a list of 12 *Yersinia* spp., and their respective number of draft genomes and complete, as shown in Table 1. The "View Strains" button on the right of each species leads the user to a list of strains of the chosen species. In the table, the information available to the user are strain status (draft genome or complete genome), genome size in mega base pair (Mbp), percentage of GC content (%), number of contigs, number of predicted CDS, number of predicted tRNA and number of predicted rRNA. On the right of each strain, the user can find a small icon, which provides a hyperlink enabling the user to find a list of predicted ORFs of the selected strain. From there, the user can see the ORF type (CDS or RNA), functional classification, contig, start position and stop position associated with each predicted ORF of the strain. The small icon on the right of each ORF brings the user to another page which allows them to see additional information of the predicted ORF besides that described above, which includes nucleotide length (bp) and the sequence, predicted polypeptide length (amino acids) and the direction of transcription, subcellular localization, hydrophobicity (pH), molecular weight (Da) and SEED subsystem. The page is also equipped with JBrowse a fast and modern JavaScript-based genome browser [36] which will enable the user to navigate genome annotations and visualize the location of the ORF of the selected *Yersinia* strain. On the top of the page, the user can find a "Download" button to download annotation details, amino acid sequence and nucleotide sequence of the predicted ORF.

Advanced real-time searching

YersiniaBase stores large amounts of data related to *Yersinia*, most of them being ORFs which were predicted from the RAST engine, with more than one million genes and coding sequences. Hence, it is impractical to browse for the required information page by page as this will greatly slow down the progress. In order to solve the problem of searching for specific genes or functional classifications in such a large database, we implemented a real-time search engine using AJAX in YersiniaBase. We designed the real-time search engine in a way such that the communications between web interface and MySQL database is asynchronous; refreshing of web page is not needed to display the list of suggested functional classifications that match the entered keyword. As soon as the user enters one keyword, the real-time search engine will retrieve a list of functional classifications that contain the keyword entered by the user and display it to the user seamlessly. Once the user selects one item from the list, they will be presented with a table which contains a list of genes that are related to the entered keyword.

In-house developed pairwise genome comparison tool (PGC)

As pointed out earlier, *Y. pestis* has already been declared as re-emergent and *Y. enterocolitica* is listed as the third-most-common enteropathogen. Many of the previously considered non-pathogenic strains of *Yersinia* have also shown signs of infections in humans. It is evident that the genus *Yersinia* is constantly evolving and further study is required not only of the pathogenic strains but also of the non-pathogenic stains to gain a clear understanding of its biology. To have a clear idea of the lifestyle, the evolution and an extended view of the gene pool of a species, genomic information from a single *Yersinia* genome is unlikely to be sufficient. For detailed insights into the variations between different strains at the genetic level, the evolutionary changes among *Yersinia* species, the orthologous genes in species, as well as the genes that give each organism its unique characteristics, especially the potential regions associated with pathogenicity, a comparative study of multiple *Yersinia* genomes is required.

With that in mind, the PGC was developed and incorporated into YersiniaBase. This in-house developed online comparative genome analysis tool will allow the user to compare two selected *Yersinia* genomes. Unlike conventional methods which display alignments in a linear form, the PGC displays the alignments in a circular layout. On entering the web interface of PGC in YersiniaBase, the user can choose two *Yersinia* genomes of interest from the list for comparison. Alternatively, the user can upload their *Yersinia* genome sequence for comparison with the *Yersinia* genomes in YersiniaBase. Three parameters, which are minimum percent identity (%), merge threshold (bp) and link threshold (bp) can be defined manually by the user. Link threshold (LT) removes the links according to user-defined value, while merge threshold (MT) allows merging of links based on user-defined value. Link is where the region of query genome maps to region to reference genome. Merged means the merging of two links which is separated by gap into one wider link. Mapped region of query genome to reference genome will be shown if the region is higher than the link threshold; while gap will be displayed if it is higher than the merged threshold, else the links beside the gap will be merged into a wider link. If user set the merge threshold to 0Kbp, and link threshold to 1Kbp, then the result of PGC will be same as shown in Additional file 4: Figure S3(A) whereas if the merge threshold was changed to 2Kbp, the result will be same as Additional file 4: Figure S3(B).

In the PGC pipeline, after the sequences are aligned through NUCmer, the alignment results are immediately parsed to Circos [41], which then generates a circular ideogram layout to show the relationship between pairs of positions, with karyotypes and links encoding the position, size and orientation of the related genomic elements. Perl scripts were used to automate the multi-step process of this pipeline. The results generated by the PGC tool, both the NUCmer alignment results and the Circos plot can be downloaded using the "Download" button in the PGC result page. Additional file 5: Figure S4 briefly illustrates the work flow of PGC tools, describing the integration of both MUMmer and Circos, generating the required input files for the PGC to function.

At the time of writing this paper, a similar tool named Circoletto [42] already existed, which aligns two genomes by using BLAST, however PGC aligns two genomes by using NUCmer package in MUMmer 3.0 [43]. In comparison, the latter is more favourable and more suitable for whole-genome comparison as NUCmer uses global alignment which is more suitable for large-scale and rapid pairwise alignment between two large genomes while Circoletto uses BLAST, a local alignment program. PGC provides a user-friendly interface and requires no prior programming knowledge. PGC allows the user to adjust parameters such as minimum percent genome identity (%), merging of links/ribbons according to merge threshold and also the removal of links according to the user-defined link threshold through the provided online form. A histogram track showing the percentage of mapped regions along the genomes is added in the circular layout generated by PGC, making it helpful for the users to identify putative indels and repetitive regions in the compared genomes.

Pathogenomics profiling tool (PathoProT) for comparative pathogenomics analysis

Pathogenesis is due to the presence of virulence genes in bacteria [44], which is responsible for causing disease in the host. The acquisition of individual virulence factors may convert non-pathogenic bacteria into pathogens [45]. With the availability of genome sequences of different *Yersinia* species it is essential to do comparative analyses of the virulence factors in the *Yersinia* pathogen genomes to identify new potential virulence markers and to provide new insights into pathogenicity of this genus. In order to identify potential virulence genes in the *Yersinia* strains and to facilitate comparative analyses between different *Yersinia* strains, we have developed PathoProT, a unique comparative pathogenomics analysis tool to help users to get a deeper insight into the different species of *Yersinia* especially in an area of pathogenicity.

PathoProT was designed using in-house developed Perl and R scripts, where Perl handles the initial processes while R is used to generate hierarchical clustering and heat map visualization of multiple virulence gene profiles. The user can select a list of *Yersinia* strains for comparative analysis and set the cut-off for sequence identity and completeness through the online form on the PathoProT main page. PathoProT predicts virulence genes based on sequence homology of all protein sequences of user-selected strains against VFDB [38-40] using the well-established BLAST (Stand-alone) tool of NCBI [37,46-48], which is downloaded from NCBI and embedded in the PathoProT pipeline. The virulence factors retrieved from VFDB, a database consisting of the experimentally verified virulence factors from different pathogenic organisms (VFDB Version 2012 containing a total of 19775 proteins), for PathoProT are considered as the "known virulence factors". In the PathoProT pipeline, a BLAST search is performed with the default parameters of 50% sequence identity and 50% sequence completeness to identify the orthologs of these known virulence factors in the *Yersinia* genomes present in the YersiniaBase. However the user can change these default parameters for the BLAST search depending on their desired levels of stringency. In-house developed Perl script filters the BLAST results generated from a BLAST search against VFDB based on user-defined cut-off values for sequence identity and completeness to identify the virulence genes and selects only the user desired strains. Results from the filtering process will be used to tabulate the data matrix (strains versus virulence genes) This is followed by executing R scripts, to read the generated data matrix and perform hierarchical clustering (complete-linkage algorithm) of the *Yersinia* strains based on their virulence gene profiles. Finally, for the visualization of the virulence gene profiles, 'pheatmap' package is used to generate the heat map image which is the end-product of PathoProT. The use of heat map in PathoProT for visualization enables users to have a bird's eye view of the data in a graphical representation [49], allowing users to identify and visualize the virulence gene profile in their strains of interest, and enabling comparative analyses of virulence factors among different *Yersinia* strains. Additional file 6: Figure S5 shows the flow chart of PathoProT, briefly describing the pipeline which integrates both Perl script and R script, and the processes before generating the output file.

YersiniaTree: constructing *Yersenia* phylogenetic tree

Phylogenetic trees are important in studying the evolutionary relationships between different species, and their common ancestor. In YersiniaBase, we have designed and incorporated an automated pipeline written in Perl, YersiniaTree that will enable user to generate phylogenetic trees of the *Yersinia* strains based on their housekeeping genes and 16S rRNA. Our automated pipeline YersiniaTree, primarily requires two inputs from the user: (1) gene marker used to construct the phylogenetic tree (2) list of genomes in YersiniaBase which to be included in the tree. The automated pipeline also offers an optional feature where the users can input their nucleotide sequence in FASTA format along with the sequences which are retrieved from the database. Currently, this pipeline in YersiniaBase, allows users to choose from one out of five gene markers, which can be seen from the drop-down menu to construct the phylogenetic tree. Besides the common 16S rRNA, users can choose from any one of the following *gyr*B, *hsp*60, *rpo*B, or *sodA* for the construction of the phylogenetic tree. Previous studies have shown that phylogenetic trees based on these four housekeeping genes are more consistent with the biochemical profiles of *Yersinia* species than those based on the 16S rRNA [4]. After user provides all the necessary inputs, the front end PHP executes the backend Perl pipeline. The first step of our automated pipeline is to select target gene's nucleotide sequence of genomes chosen by user from FASTA file where complete listings of sequences are stored, into a temporary FASTA file. The Perl script then executes MAFFT [50], which is used to perform multiple sequence alignment across nucleotide sequences. Next, the output file from MAFFT is sent to FastTree [51] to construct phylogenetic tree in Newick format, followed by visualizing of the tree in SVG format by using Newick Utilities [52]. Finally, Perl script sends out the final image to PHP to display the phylogenetic tree in web browser.

Yersinia genome, gene and virulence factors database

Besides the in-house developed tools, we have also integrated BLAST as well as VFDB-BLAST into YersiniaBase. This will allow the users to perform their similarity search of their query sequences against *Yersinia* genome sequences, gene sequences and also against the virulence factors (downloaded from VFDB [38-40]) by using BLASTN, BLASTP or BLASTX. This specialized database for *Yersinia* provides faster searching against *Yersinia* sequences compared to NCBI NT or NR databases when the user is trying to find the closest *Yersinia* strains to their own query sequence.

Features of strains in YersiniaBase

Generally, the size of the *Yersinia* genomes was observed to be between 4.3 Mbp to 4.8 Mbp. However, there were a few exceptions. For example, the genome size of *Y. ruckeri* ATCC29473, a fish pathogen, was approximately 3.7 Mbp, which is smaller than the normal range. *Y. kristensennii* ATCC33638 and *Y. massiliensis* CCUG53443 had genome sizes larger than 5Mbp. RAST annotations of all the download *Yersinia* genomes, predicted a total of 1,009,094 ORFs. 997,893 of the ORFs encoded protein while 11,201 encoded functional RNA. The subcellular localization analysis of the RAST predicted proteins using PSORTb version 3.0, showed that the major portion of the predicted proteins produced by *Yersinia* were predicted to be located in cytoplasmic region. The results obtained from PSORTb are tabulated in Table 2.

Table 2 The breakdown of subcellular compartments of predicted *Yersinia* proteins and their numbers in percentage format

Subcellular compartments	Percentage of localization (%)
Cytoplasmic	37.31
Cytoplasmic membrane	23.36
Extracellular	1.66
Outer membrane	2.52
Periplasmic	3.36
Unknown	31.80

Comparative genomic analysis

To demonstrate the usefulness of YersiniaBase, we carried out genomic analyses of 23 *Yersinia* strains available in YersiniaBase. Three phylogenetic trees were constructed, one using the MEGA6 [53], the second one using the phylogenetic tool of YersiniaBase, YersiniaTree and the other using the in-house developed tool PathoProT. We used the neighbour-joining method, Tamura-Nei model with 1,000 bootstrap replications to build phylogenetic tree with MEGA6 [53] based on the *gyr*B gene (Figure 1A). The tree generated by YersiniaTree was also based on *gyr*B (Figure 1B). The *gyr*B gene instead of 16S rRNA gene was used for phylogenetic analysis as 16S rRNA cannot resolve the phylogenetic relationships between closely related *Yersinia* species. Housekeeping genes such as *gyrB* produce phylogenetic trees which closely correlate with the biochemical species designations. Moreover studies have shown that *gyr*B is a better phylogenetic marker than 16S rRNA in the Enterobacteriaceae [4,54-56]. The other tree was constructed using the in-house developed tool PathoProT, with hierarchical clustering (complete-linkage) algorithm based on the virulence gene profiles of the selected *Yersinia* strains. The parameters for PathoProT

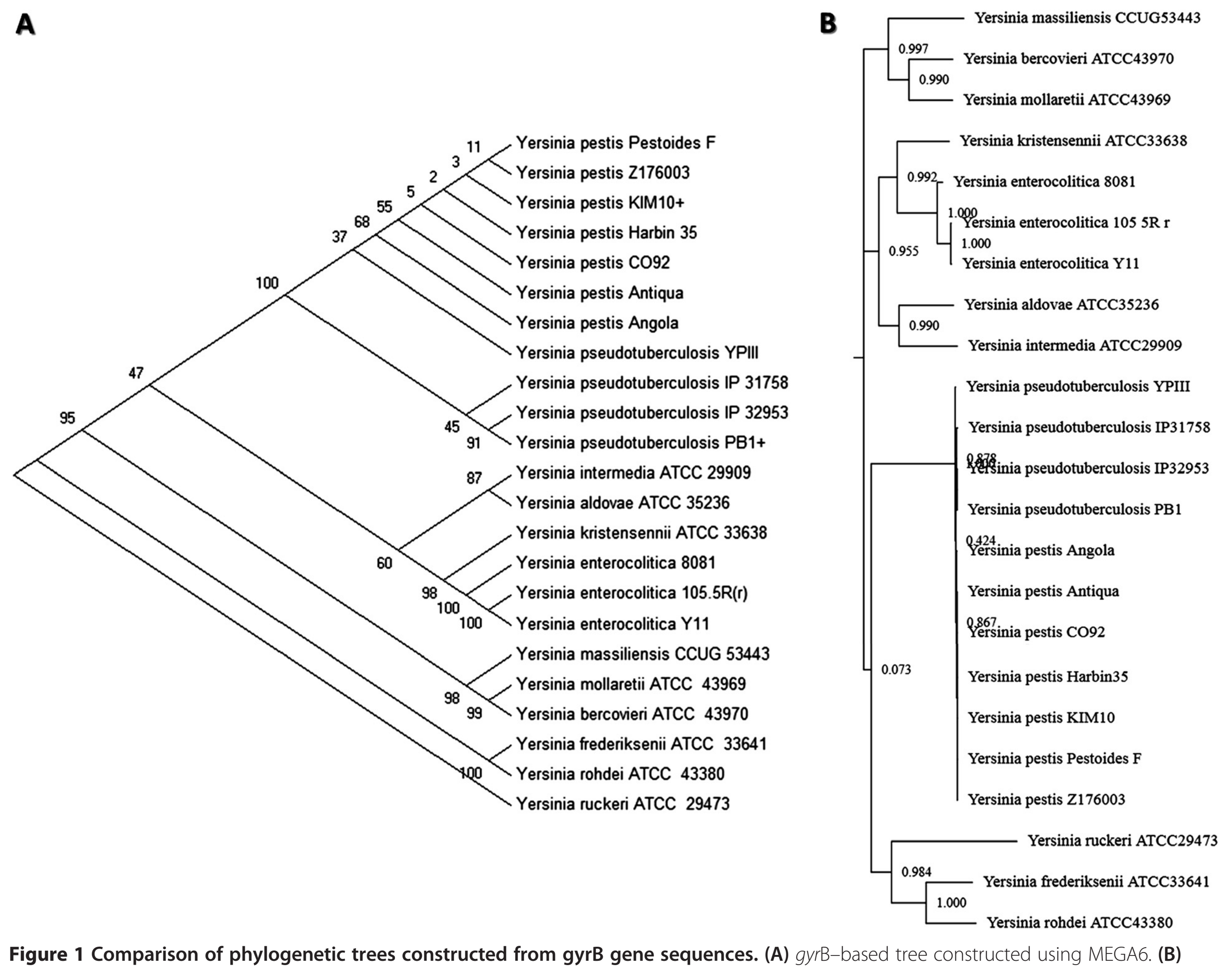
A
Yersinia pestis Pestoides F
Yersinia pestis Z176003
Yersinia pestis KIM10+
Yersinia pestis Harbin 35
Yersinia pestis CO92
Yersinia pestis Antiqua
Yersinia pestis Angola
Yersinia pseudotuberculosis YPIII
Yersinia pseudotuberculosis IP 31758
Yersinia pseudotuberculosis IP 32953
Yersinia pseudotuberculosis PB1+
Yersinia intermedia ATCC 29909
Yersinia aldovae ATCC 35236
Yersinia kristensennii ATCC 33638
Yersinia enterocolitica 8081
Yersinia enterocolitica 105.5R(r)
Yersinia enterocolitica Y11
Yersinia massiliensis CCUG 53443
Yersinia mollaretii ATCC 43969
Yersinia bercovieri ATCC 43970
Yersinia frederiksenii ATCC 33641
Yersinia rohdei ATCC 43380
Yersinia ruckeri ATCC 29473
11
3
2
5
55
68
37
100
47
95
45
91
87
60
98
100
100
98
99
100
B
Yersinia massiliensis CCUG53443
0.997
Yersinia bercovieri ATCC43970
0.990
Yersinia mollaretii ATCC43969
Yersinia kristensennii ATCC33638
0.992
Yersinia enterocolitica 8081
1.000
Yersinia enterocolitica 105 5R r
1.000
0.955
Yersinia enterocolitica Y11
Yersinia aldovae ATCC35236
0.990
Yersinia intermedia ATCC29909
Yersinia pseudotuberculosis YPIII
Yersinia pseudotuberculosis IP31758
0.878
1.000
Yersinia pseudotuberculosis IP32953
Yersinia pseudotuberculosis PB1
0.424
Yersinia pestis Angola
Yersinia pestis Antiqua
0.867
Yersinia pestis CO92
0.073
Yersinia pestis Harbin35
Yersinia pestis KIM10
Yersinia pestis Pestoides F
Yersinia pestis Z176003
Yersinia ruckeri ATCC29473
0.984
Yersinia frederiksenii ATCC33641
1.000
Yersinia rohdei ATCC43380

Figure 1 Comparison of phylogenetic trees constructed from gyrB gene sequences. (A) *gyr*B–based tree constructed using MEGA6. **(B)** *gyr*B–based tree constructed using in-house developed YersiniaTree.

were set to 50 for sequence identity (%) and 50 for sequence completeness (%). We also used the in-house developed PGC Tool for finding the differences between subspecies of *Y. enterocolitica*. The parameters for the PGC Tool were set to 90 for minimum percent identity, 2,000 for link threshold (bp) and 2,000 for merge threshold (bp). The list of *Yersinia* strains selected for the analysis is shown in Table 3.

Phylogeny of *Yersinia* based on *gyrB*

In this paper, we constructed two phylogenetic trees based on *gyrB*, one using MEGA6 (Figure 1A) while the other using our in-house developed tool YersiniaTree (Figure 1B). The tree generated by YersiniaTree was found to be consistent with that of the tree generated by the well-established tool MEGA. Both the trees showed similarity in the clustering pattern. In both the cases the strains of *Y. pestis* and *Y. pseudotuberculosis* grouped very close each other whereas *Y. enterocolitica* formed a separate cluster. In case of *Y. enterocolitica* strains 105.5R(r) and Y11 (both of which belong to the subspecies *palearctica*) grouped together while 8081 (subspecies *enterocolitica*) separated to form a different branch in the tree generated by YersiniaTree (Figure 1B). Similar type of clustering was also observed in the tree by MEGA (Figure 1A).

Table 3 List of *Yersinia* strains used in the analysis

Species name	Strain name
Y. aldovae	ATCC35236
Y. bercovieri	ATCC43970
Y. frederiksenii	ATCC33641
Y. intermedia	ATCC29909
Y. kristensennii	ATCC33638
Y. massiliensis	CCUG53443
Y. mollaretii	ATCC43969
Y. rohdei	ATCC43380
Y. ruckeri	ATCC29473
Y. enterocolitica	105.5R(r)
Y. enterocolitica	8081
Y. enterocolitica	Y11
Y. pestis	Angola
Y. pestis	Antiqua
Y. pestis	CO92
Y. pestis	Harbin35
Y. pestis	KIM10
Y. pestis	Pestoides F
Y. pestis	Z176003
Y. pseudotuberculosis	IP 31758
Y. pseudotuberculosis	IP 32953
Y. pseudotuberculosis	PB1+
Y. pseudotuberculosis	YPIII

Evolution of *Y. pestis* and *Y. pseudotuberculosis*

Apart from the trees generated based on *gyr*B a separate tree was constructed based on the virulence profiles of *Y. pestis* and *Y. pseudotuberculosis* strains using the PathoProT (Figure 2). Even in the tree by PathoProT (Figure 2), we found that *Y. pestis* and *Y. pseudotuberculosis* formed a cluster that was clearly separated from *Y. enterocolitica* and non-pathogenic *Yersinia* which was similar to that observed in both the trees based on *gyr*B. This similarity in the clustering pattern, suggests that the housekeeping genes and virulence genes of *Y. pestis* and *Y. pseudotuberculosis* evolved in a similar pathway and are distinct from the other *Yersinia* species. These results were in agreement with a recently published report suggesting that *Y. pestis* and *Y. pseudotuberculosis* are close relatives, however *Y. enterocolitica* shows diametric separation with most environmental species occupying intermediate branching positions between these two clusters [57,58]. The tree also indicates that *Y. pestis* Angola is the first among the *Y. pestis* strains to branch out from the closest common ancestor *Y. pseudotuberculosis* YPIII. This observation is supported by a recently

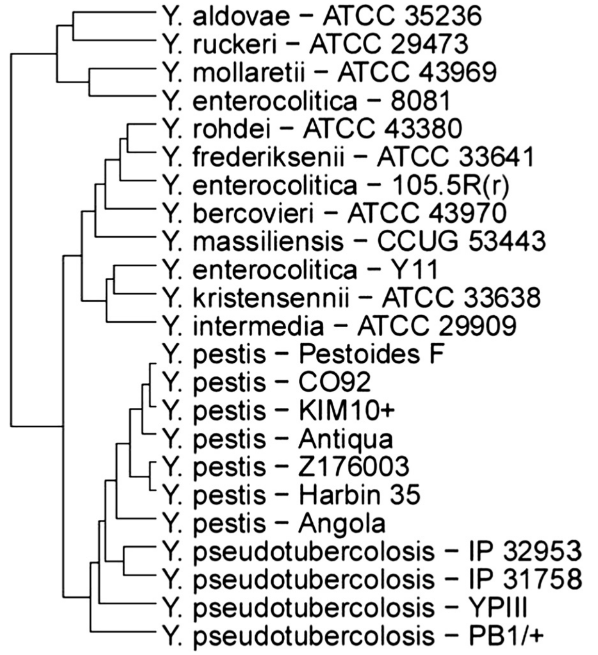
Y. aldovae – ATCC 35236
Y. ruckeri – ATCC 29473
Y. mollaretii – ATCC 43969
Y. enterocolitica – 8081
Y. rohdei – ATCC 43380
Y. frederiksenii – ATCC 33641
Y. enterocolitica – 105.5R(r)
Y. bercovieri – ATCC 43970
Y. massiliensis – CCUG 53443
Y. enterocolitica – Y11
Y. kristensennii – ATCC 33638
Y. intermedia – ATCC 29909
Y. pestis – Pestoides F
Y. pestis – CO92
Y. pestis – KIM10+
Y. pestis – Antiqua
Y. pestis – Z176003
Y. pestis – Harbin 35
Y. pestis – Angola
Y. pseudotubercolosis – IP 32953
Y. pseudotubercolosis – IP 31758
Y. pseudotubercolosis – YPIII
Y. pseudotubercolosis – PB1/+

Figure 2 Phylogenetic tree constructed by from virulence gene profiles with PathoProT.

published report, which stated that *Y. pestis* Angola belongs to one of the most ancient *Y. pestis* lineages sequenced to date, possessing genome characteristics which are intermediate between *Y. pseudotuberculosis* and the modern day *Y. pestis* [59]. These data suggest that *Y. pestis* is highly similar to *Y. pseudotuberculosis* genetically and the virulence genes that were lost or acquired did not cause it to diverge far from its ancestor. However, *Y. pestis* has undergone gene reduction and has accumulated pseudogenes since its emergence from *Y. pseudotuberculosis* [60,61].

Virulence genes in *Yersinia*

By calculating the number of virulence genes presents in every strain, we found that *Y. pestis* and *Y. pseudotuberculosis* generally have more virulence genes than other species and the number of virulence genes in different strains is more consistent within these species compared to *Y. enterocolitica* and non-pathogenic *Yersinia* (Figure 3). The number of virulence genes in non-pathogenic *Yersinia* varies from as low as 76 in *Y. ruckeri* ATCC29473, the fish pathogen to 389 in *Y. massiliensis* CCUG53443. However, it must be taken into consideration that all of the non-pathogenic *Yersinia* are only available in draft genomes and it is possible that some virulence genes have not been included in the sequences. Although *Y. enterocolitica* 8081 is a highly pathogenic strain [62], it has the third lowest number of virulence genes among *Y. enterocolitica* and non-pathogenic *Yersinia*. This suggests the other *Y. enterocolitica* and non-pathogenic *Yersinia* might only acquire virulence genes which do not play vital role in pathogenicity or the virulence genes were deactivated.

Another interesting observation from the study of the virulence genes of the *Yersinia* genomes is that *Y. pestis* and *Y. pseudotuberculosis*, which formed a cluster in both the trees, also had a consistency in the number of virulence genes present in their genomes, with an average of 436 and 422 identified virulence genes in case of *Y. pestis* strains and *Y. pseudotuberculosis* stains respectively, while on the other hand *Y. enterocolitica* had an average of 280 identified virulence genes. The average number of virulence genes from *Y. enterocolitica* was significantly lower as compared to the other two pathogens in *Yersinia*. This difference in the number of virulence genes in the three pathogenic species of *Yersinia* also indicates that *Y. pestis* and *Y. enterocolitica* are close relatives and might have followed a similar evolution pathway in acquiring or losing virulence genes whereas *Y. enterocolitica* clearly followed a different evolutionary path, independent from the other two. Surprisingly, *Y. enterocolitica* 8081, which is the highly virulent biotype, did not form the same cluster with the other two strains of *Y. enterocolitica*, but rather with 4 non-pathogenic *Yersinia*.

Absence of yersiniabactin in *Y. enterocolitica* subsp. *palearctica*

Besides having pYV virulence plasmid, highly pathogenic *Yersinia* strains also possess iron-regulated genes [63]. There have been reports stating that *Y. enterocolitica* serotype O:8 carries high pathogenicity island (HPI), which synthesizes yersiniabactin [64,65]. In order to get a better understanding of the pathogenicity of *Y. enterocolitica* subspecies, we used PathoProT to visualize the

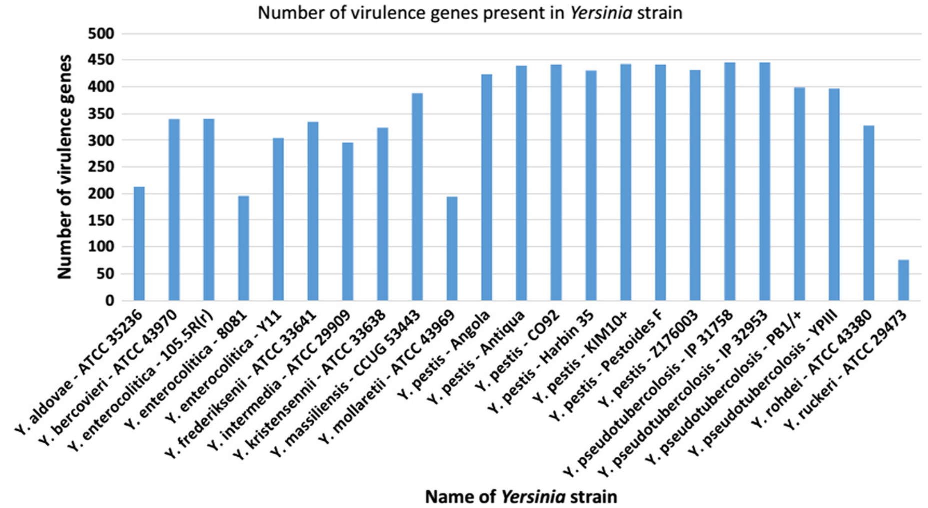

Figure 3 Number of virulence genes present in each *Yersinia* strain.

virulence gene profiles. From the heat map produced by PathoProT (part of the heat map is shown in Figure 4), we found that yersiniabactin was not present in *Y. enterocolitica* subsp. *palearctica*. The result is consistent with previous findings [62,66] and further suggests that *Y. enterocolitica* subsp. *enterocolitica* has followed a slightly different evolutionary pathway compared to subsp. *palearctica* and that yersiniabactin was acquired by *Y. enterocolitica* subsp. *enterocolitica* after it diverged from its common ancestor shared with subsp. *palearctica*.

We further visualized the difference between the two subspecies of *Y. enterocolitica* by using PGC Tool with *Y. enterocolitica* 8081 as reference genome (Figure 5). Another draft genome, *Y. enterocolitica* subsp. *enterocolitica* WA-314 was included. The start and stop position of the operon for yersiniabactin in 8081 are 2817014 and 2841569 respectively (red circle region in the Figure 5). From the three figures generated by PGC (Figure 5), we found only WA-314 has region mapped to the region encoding yersiniabactin in 8081.

Flagella genes in *Y. pestis* Angola

We found 9 flagella genes missing from *Y. pestis* Angola, but present in the rest of studied strains (Figure 6). The 9 flagella genes are *flh*A, *flh*B, *flh*E, *fli*L, *fli*M, *fli*N, *fli*O, *fli*Q, and *fli*R. In order to prevent false negative results, the 9 flagella genes present in other strains but absent in Angola were used to TBLASTN against the *Y. pestis* Angola genome. Using this approach, we found that the missing flagella genes matched to the Angola genome with query coverage more than 50% (except *fli*O and *fli*L), but with identity less than 50% (Table 4).

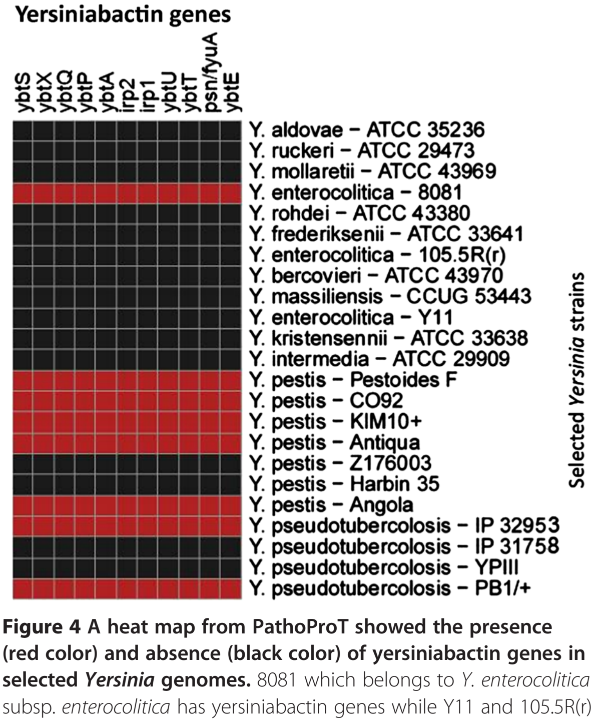

Figure 4 A heat map from PathoProT showed the presence (red color) and absence (black color) of yersiniabactin genes in selected *Yersinia* genomes. 8081 which belongs to *Y. enterocolitica* subsp. *enterocolitica* has yersiniabactin genes while Y11 and 105.5R(r) which belong to *Y. enterocolitica* subsp. *palearctica* do not have yersiniabactin genes.

Although *Yersinia* is non-motile, all the *Y. pestis* strains except for the strain Angola have a complete set of flagella associated genes, including flagella synthesis and flagella motor rotation [67]. There have been reports stating that mutations have occurred in the flagella and chemotaxis gene clusters [68], but sequence similarities of the *Y. pestis* flagella genes still can be found using BLAST, with the exception of *Y. pestis* Angola. We suggest that, after the emergence of other *Yersinia* strains, many mutations have taken place in the 9 flagella genes of the Angola genome and accumulated over time, which lead to low identity between Angola and other *Y. pestis*. Another possibility is that the flagella gene cluster might have been replaced in Angola with that of a different species by horizontal gene transfer.

Future work

With the advent of next generation sequencing (NGS), we expect more *Yersinia* strains will be published. We regularly keep track of the new developments related to *Yersinia* especially release of new *Yesinia* genomes in NCBI. We are already in the process of adding new *Yersinia* genomes to the YersiniaBase and will be updating the database as and when new genomes of *Yersinia* are available. At this moment, the database is optimized for whole-genome and comparative pathogenomic analysis. However we hope to incorporate other data types as well along with new features and functionalities in future. We are working on it and we also encourage other researchers or research groups to email us at girg@um.edu.my if they wish to share their annotations, opinions, and curated data with us. We welcome the suggestions for improving this database as it will help us make this database a comprehensive one.

Conclusions

With the advent of next-generation sequencing technologies, it is important that biological data can be stored and retrieved easily for analyses. YersiniaBase aims to provide a platform that stores genomic data and annotation details of *Yersinia* besides providing new bioinformatics tools that can assist researchers. In this paper, we have demonstrated how the tools can be used in analyses, including the identification of pathogenicity factors and genomic differences.

With the rapid advances and significant price drop in NGS technologies, we expect more *Yersinia* strains will be sequenced and published. Thus, we will continue updating the data in YersiniaBase once these genome sequences are available. In order to improve YersiniaBase,

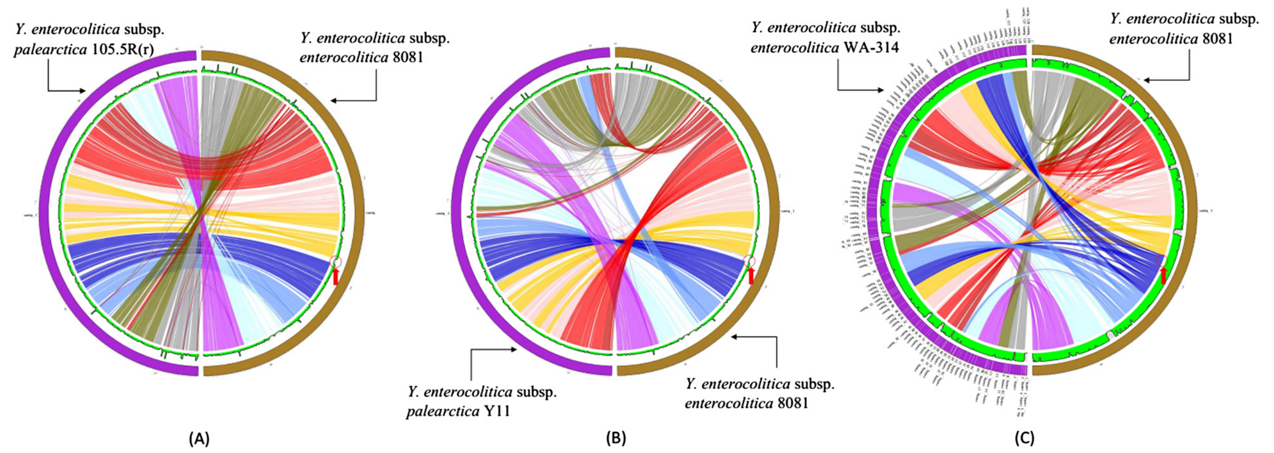
Y. enterocolitica subsp. *palearctica* 105.5R(r)
Y. enterocolitica subsp. *enterocolitica* 8081
Y. enterocolitica subsp. *palearctica* Y11
Y. enterocolitica subsp. *enterocolitica* 8081
Y. enterocolitica subsp. *enterocolitica* WA-314
Y. enterocolitica subsp. *enterocolitica* 8081
(A)
(B)
(C)

Figure 5 Genome comparisons between different subspecies of Yersinia enterocolitica by using PGC tool. PGC plots revealed yersiniabactin only present in Yersinia enterocolitica subsp. enterocolitica. **(A)** Comparison between 105.5R(r) and 8081 by using PGC showed that 105.5R(r) did not map to region of yersiniabactin in 8081. **(B)** Comparison between Y11 and 8081 by using PGC showed that Y11 did not map to region of yersiniabactin in 8081. **(C)** Comparison between WA-314 by using PGC showed that WA-314 mapped to region of yersiniabactin in 8081.

we welcome any comment on this database and the sharing of new *Yersinia* data. Users who want their data to be added into the database can do so by clicking the “Submit Annotation Update” link, at the bottom right corner of YersiniaBase. We hope that YersiniaBase will provide a comprehensive resource and analysis platform for of the *Yersinia* research community in the future.

Availability and system requirements

YersiniaBase is available online at http://yersinia.um.edu.my. All the sequences and annotations described in this paper could be viewed and downloaded from the website. YersiniaBase is best viewed by Internet Explorer 8.x or higher, Mozilla Firefox® 10.x or higher, Safari 5.1 or higher, Chrome 18 or higher and any other equivalent browser software. If your browser is older, you may have trouble viewing many of our web site features properly. This web site is best viewed at a screen resolution of 1024 × 768 pixels or higher.

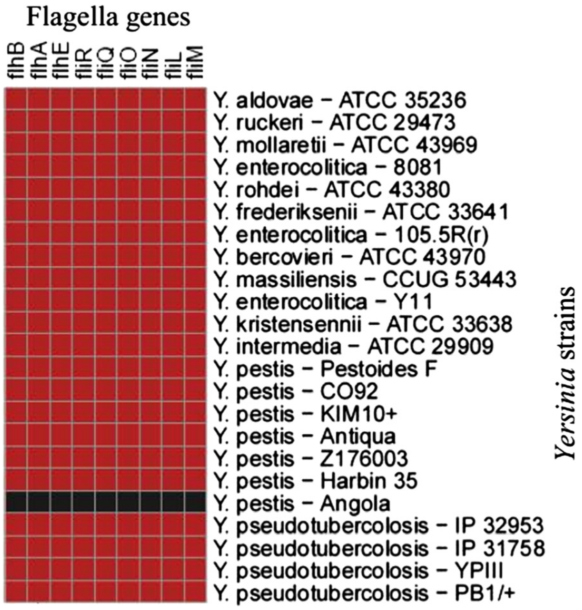

Figure 6 A heat map showed the absence (black color) of nine flagella genes in Angola, but present (red color) in other selected *Yersinia* strains.

Table 4 The query sequence coverage and percentage of protein identity of nine missing flagella genes in Angola but present in other *Yersinia* strains TBLASTN against genome of Angola

Gene	Contig	Sequence identity (%)	Sequence coverage (%)
*flh*B	Angola_contig0	36.6	97
*flh*B	Angola_contig0	29.38	97
*flh*A	Angola_contig0	47.44	88
*flh*A	Angola_contig0	46.59	88
*flh*E	Angola_contig0	40	67
*flh*E	Angola_contig0	30.56	67
*fli*R	Angola_contig0	32.13	84
*fli*R	Angola_contig0	27.27	84
*fli*R	Angola_contig0	29.52	84
*fli*Q	Angola_contig0	41.38	65
*fli*Q	Angola_contig0	44	65
*fli*O	Angola_contig0	31.25	27
*fli*N	Angola_contig0	40.78	69
*fli*N	Angola_contig0	26.56	69
*fli*L	Angola_contig0	34.21	20

In general, these flagella genes showed very low identity, suggesting that these genes might be highly mutated over evolution time.

Additional files

Additional file 1: Table S1. List of strains available in YersiniaBase and their corresponding strain name, genome status, genome size (bp), total number of contig, number of open reading frame (ORF), number of tRNA, number of rRNA and percentage of GC content.

Additional file 2: Figure S1. Underlying relational database schema of YersiniaBase, which consists of four tables, with function of each table, is depicted in the figure. The relationship between tables which is defined as followed: (1) each species has one or many strains; (2) each strain has one or many open reading frames; (3) each open reading frame has only one nucleotide sequence, none or one amino acid sequence. Strain table has foreign key that links to primary key in species table; ORF table has foreign key that links to primary key in strain table; sequence table has foreign key that links to primary key in ORF table.

Additional file 3: Figure S2. Overall functionalities in YersiniaBase.

Additional file 4: Figure S3. (A) Green and blue link were displayed as the mapped region, because the mapped region is higher than the link threshold, while the gap is present between green and blue link because the gap is wider than the value of merge threshold (0Kbp in this case). (B) Since the gap (1Kbp) is smaller than 2Kbp (merge threshold in this case), the green and blue link beside the gap were merged into a wider link of 8Kbp (2Kbp Green Link + 1Kbp Gap + 5Kbp Blue link).

Additional file 5: Figure S4. Brief description of processes taken in PGC pipeline after user submits the job to our server.

Additional file 6: Figure S5. Brief description of processes taken in PathoProT pipeline after user submits the job to our server.

Abbreviations
AJAX: Asynchronous JavaScript and XML; BLAST: Basic Local Alignment Search Tool; bp: Base pair; CDS: Coding DNA sequence; CSS: Cascading Style Sheets; EFSA: European Food Safety Authority; HPI: High pathogenicity island; HTML: HyperText Markup Language; Jbrowse: JavaScript-based genome browser; MVC: Model-view-controller; NCBI: National Center for Biotechnology Information; NGS: Next generation sequencing; ORF: Open reading frame; PathoProT: Pathogenomics Profiling Tool; PGC: Pairwise Genome Comparison tool; PHP: HyperText Preprocessor; RAST: Rapid Annotation using Subsystem Technology; RNA: Ribonucleic acid; TTSS: Type III secretion system; VFDB: Virulence factors database; WHO: World Health Organization; Yop: Yersinia outer protein.

Competing interests
All authors declare that they have no competing interests.

Authors' contributions
SWC, SYT, CCS, NVRM, MYA, WYW and HH designed and developed the database system. SYT, AD and CCS generated annotations and analyzed data. SYT, AD, SWC, NSJ and GJW wrote the manuscript. All authors have read and approved the final manuscript.

Acknowledgements
We would like to thank all members of Genome Informatics Research Group (GIRG) in contributing to this research.

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/08and UM Research Grant (UMRG) [Account No. UMRG: RG541-13HTM].

Author details
[1]Genome Informatics Research Laboratory, High Impact Research Building (HIR) Building, University of Malaya, 50603 Kuala Lumpur, Malaysia. [2]Department of Oral Biology and Biomedical Sciences, Faculty of Dentistry, University of Malaya, 50603 Kuala Lumpur, Malaysia. [3]Center for Oral Health Research, School of Dental Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne, United Kingdom. [4]Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia.

Received: 12 September 2014 Accepted: 11 December 2014
Published online: 16 January 2015

Reference
1. Aleksic S, Steigerwalt AG, Bockemuhl J, Huntleycarter GP, Brenner DJ. Yersinia-rohdei Sp-Nov isolated from human and Dog feces and surface-water. Int J Syst Bacteriol. 1987;37(4):327–32.
2. Wauters G, Janssens M, Steigerwalt AG, Brenner DJ. Yersinia mollaretii sp. nov. and Yersinia bercovieri sp. nov., Formerly Called Yersinia enterocolitica Biogroups 3A and 3B. Int J Syst Bacteriol. 1988;38(4):424–9.
3. Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, Carniel E. Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc Natl Acad Sci U S A. 1999;96(24):14043–8.
4. Merhej V, Adekambi T, Pagnier I, Raoult D, Drancourt M. Yersinia massiliensis sp. nov., isolated from fresh water. Int J Syst Evol Microbiol. 2008;58(Pt 4):779–84.
5. Drancourt M, Raoult D. Past plague. In: Raoult D, Drancourt M, editors. Paleomicrobiology: Past Human Infections. Berlin Heidelberg: Springer-Verlag; 2008. p. 145.
6. Prentice MB, Rahalison L. Plague. Lancet. 2007;369(9568):1196–207.
7. Perry RD, Fetherston JD. Yersinia pestis–etiologic agent of plague. Clin Microbiol Rev. 1997;10(1):35–66.
8. Butler T. Yersinia infections: centennial of the discovery of the plague bacillus. Clin Infect Dis Off Publ Infect Dis Soc Am. 1994;19(4):655–61. quiz 662–653.
9. Campbell GL, Hughes JM. Plague in India: a new warning from an old nemesis. Ann Intern Med. 1995;122(2):151–3.
10. Human plague in 1994. Releve epidemiologique hebdomadaire/Section d'hygiene du Secretariat de la Societe des Nations = Weekly epidemiological record/Health Section of the Secretariat of the League of Nations. 1996;71(22):165–168.
11. Bertherat E, Bekhoucha S, Chougrani S, Razik F, Duchemin JB, Houti L, et al. Plague reappearance in Algeria after 50 years, 2003. Emerg Infect Dis. 2007;13(10):1459–62.
12. Cabanel N, Leclercq A, Chenal-Francisque V, Annajar B, Rajerison M, Bekkhoucha S, et al. Plague outbreak in Libya, 2009, unrelated to plague in Algeria. Emerg Infect Dis. 2013;19(2):230–6.
13. Guiyoule A, Gerbaud G, Buchrieser C, Galimand M, Rahalison L, Chanteau S, et al. Transferable plasmid-mediated resistance to streptomycin in a clinical isolate of Yersinia pestis. Emerg Infect Dis. 2001;7(1):43–8.
14. Josko D. Yersinia pestis: still a plague in the 21st century. Clin Lab Sci J Am Soc Med Technol. 2004;17(1):25–9.
15. Pradel E, Lemaitre N, Merchez M, Ricard I, Reboul A, Dewitte A, et al. New insights into how Yersinia pestis adapts to its mammalian host during bubonic plague. PLoS Pathog. 2014;10(3):e1004029.
16. Drummond N, Murphy BP, Ringwood T, Prentice MB, Buckley JF, Fanning S. Yersinia enterocolitica: a brief review of the issues relating to the zoonotic pathogen, public health challenges, and the pork production chain. Foodborne Pathog Dis. 2012;9(3):179–89.
17. Zadernowska A, Chajecka-Wierzchowska W, Laniewska-Trokenheim L. Yersinia enterocolitica: a dangerous, but often ignored. Foodborne Pathogen Food Rev Int. 2014;30(1):53–70.
18. Eurosurveillance editorial t: The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2010. Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2012;17(10).
19. Cornelis GR. The Yersinia Ysc-Yop virulence apparatus. Int J Med Microbiol IJMM. 2002;291(6–7):455–62.
20. Hueck CJ. Type III protein secretion systems in bacterial pathogens of animals and plants. Microbiol Mol Biol Rev MMBR. 1998;62(2):379–433.
21. Sulakvelidze A. Yersiniae other than Y. enterocolitica, Y. pseudotuberculosis, and Y. pestis: the ignored species. Microbes Infec Inst Pasteur. 2000;2(5):497–513.
22. Noble MA, Barteluk RL, Freeman HJ, Subramaniam R, Hudson JB. Clinical significance of virulence-related assay of Yersinia species. J Clin Microbiol. 1987;25(5):802–7.
23. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A. 1999;96(7):3801–6.
24. Uchiyama I. MBGD: microbial genome database for comparative analysis. Nucleic Acids Res. 2003;31(1):58–62.
25. Uchiyama I. MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. Nucleic Acids Res. 2007;35(Database issue):D343–6.

26. Uchiyama I, Higuchi T, Kawai M. MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. Nucleic Acids Res. 2010;38(Database issue):D361–5.
27. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Res. 2010;38(Database issue):D382–90.
28. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. 2012;40(Database issue):D115–22.
29. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, et al. The integrated microbial genomes (IMG) system. Nucleic Acids Res. 2006;34(Database issue):D344–8.
30. Markowitz VM, Szeto E, Palaniappan K, Grechkin Y, Chu K, Chen IM, et al. The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res. 2008;36(Database issue):D528–33.
31. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 2014;42(Database issue):D581–91.
32. 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(Database issue):D61–5.
33. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75.
34. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics. 2010;26(13):1608–15.
35. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, et al. The bioperl toolkit: Perl modules for the life sciences. Genome Res. 2002;12 (10):1611–8.
36. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. 2009;19(9):1630–8.
37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
38. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33(Database issue):D325–8.
39. 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(Database issue):D539–42.
40. 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(Database issue):D641–5.
41. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45.
42. Darzentas N. Circoletto: visualizing sequence similarity with Circos. Bioinformatics. 2010;26(20):2620–1.
43. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
44. Peterson JW: Bacterial Pathogenesis. In: Medical Microbiology. Edited by Baron S, 4th edn. Galveston (TX); 1996.
45. Hensel M, Shea JE, Gleeson C, Jones MD, Dalton E, Holden DW. Simultaneous identification of bacterial virulence genes by negative selection. Science. 1995;269(5222):400–3.
46. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinforma. 2009;10:421.
47. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004;32(Web Server issue):W20–5.
48. Tao T: Standalone BLAST Setup for Windows PC. In: BLAST® Help [Internet]. Updated 2014 Apr 10 edn. Bethesda (MD): National Center for Biotechnology Information (US); 2010.
49. Wilkinson L, Friendly M. The history of the cluster heat Map. Am Stat. 2009;63(2):179–84.
50. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
51. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol. 2011;7:539.
52. Junier T, Zdobnov EM. The newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics. 2010;26(13):1669–70.
53. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30(12):2725–9.
54. Dauga C. Evolution of the gyrB gene and the molecular phylogeny of Enterobacteriaceae: a model molecule for molecular systematic studies. Int J Syst Evol Microbiol. 2002;52(Pt 2):531–47.
55. Kotetishvili M, Kreger A, Wauters G, Morris Jr JG, Sulakvelidze A, Stine OC. Multilocus sequence typing for studying genetic relationships among Yersinia species. J Clin Microbiol. 2005;43(6):2674–84.
56. Sihvonen LM, Haukka K, Kuusi M, Virtanen MJ, Siitonen A, group YEs. Yersinia enterocolitica and Y. enterocolitica-like species in clinical stool specimens of humans: identification and prevalence of bio/serotypes in Finland. Eur J Clin Microbiol Inf Dis Off Publ Eur Soc Clin Microbiol. 2009;28 (7):757–65.
57. Nunes-Alves C. Bacterial evolution: parting of the ways for Yersinia. Nat Rev Microbiol. 2014;12(6):394–5.
58. Reuter S, Connor TR, Barquist L, Walker D, Feltwell T, Harris SR, et al. Parallel independent evolution of pathogenicity within the genus Yersinia. Proc Natl Acad Sci U S A. 2014;111(18):6768–73.
59. Eppinger M, Worsham PL, Nikolich MP, Riley DR, Sebastian Y, Mou S, et al. Genome sequence of the deep-rooted Yersinia pestis strain Angola reveals new insights into the evolution and pangenome of the plague bacterium. J Bacteriol. 2010;192(6):1685–99.
60. Chain PS, Hu P, Malfatti SA, Radnedge L, Larimer F, Vergez LM, et al. Complete genome sequence of Yersinia pestis strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. J Bacteriol. 2006;188(12):4453–63.
61. Chen PE, Cook C, Stewart AC, Nagarajan N, Sommer DD, Pop M, et al. Genomic characterization of the Yersinia genus. Genome Biol. 2010;11(1):R1.
62. Wang X, Li Y, Jing H, Ren Y, Zhou Z, Wang S, et al. Complete genome sequence of a Yersinia enterocolitica "Old World" (3/O:9) strain and comparison with the "New World" (1B/O:8) strain. J Clin Microbiol. 2011;49(4):1251–9.
63. Carniel E, Mazigh D, Mollaret HH. Expression of iron-regulated proteins in Yersinia species and their relation to virulence. Infect Immun. 1987;55 (1):277–80.
64. Pelludat C, Rakin A, Jacobi CA, Schubert S, Heesemann J. The yersiniabactin biosynthetic gene cluster of Yersinia enterocolitica: organization and siderophore-dependent regulation. J Bacteriol. 1998;180(3):538–46.
65. Rakin A, Noelting C, Schubert S, Heesemann J. Common and specific characteristics of the high-pathogenicity island of Yersinia enterocolitica. Infect Immun. 1999;67(10):5265–74.
66. Batzilla J, Antonenka U, Hoper D, Heesemann J, Rakin A. Yersinia enterocolitica palearctica serobiotype O:3/4–a successful group of emerging zoonotic pathogens. BMC Genomics. 2011;12:348.
67. Macnab RM. How bacteria assemble flagella. Annu Rev Microbiol. 2003;57:77–100.
68. Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, et al. Genome sequence of Yersinia pestis, the causative agent of plague. Nature. 2001;413(6855):523–7.