Research article

BBP: Brucella genome annotation with literature mining and curation
Zuoshuang Xiang¹, Wenjie Zheng² and Yongqun He*¹,³,⁴

Address: ¹Unit for Laboratory Animal Medicine, University of Michigan Medical School, Ann Arbor, MI 48109, USA, ²Webmail.us Inc., Blacksburg, VA 24060, USA, ³Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI 48109, USA and ⁴Bioinformatics Program, University of Michigan Medical School, Ann Arbor, MI 48109, USA

Email: Zuoshuang Xiang - zxiang@umich.edu; Wenjie Zheng - wzheng@webmail.us; Yongqun He* - yongqunh@umich.edu

* Corresponding author

Abstract

Background: Brucella species are Gram-negative, facultative intracellular bacteria that cause brucellosis in humans and animals. Sequences of four Brucella genomes have been published, and various Brucella gene and genome data and analysis resources exist. A web gateway to integrate these resources will greatly facilitate Brucella research. Brucella genome data in current databases is largely derived from computational analysis without experimental validation typically found in peer-reviewed publications. It is partially due to the lack of a literature mining and curation system able to efficiently incorporate the large amount of literature data into genome annotation. It is further hypothesized that literature-based Brucella gene annotation would increase understanding of complicated Brucella pathogenesis mechanisms.

Results: The Brucella Bioinformatics Portal (BBP) is developed to integrate existing Brucella genome data and analysis tools with literature mining and curation. The BBP InterBru database and Brucella Genome Browser allow users to search and analyze genes of 4 currently available Brucella genomes and link to more than 20 existing databases and analysis programs. Brucella literature publications in PubMed are extracted and can be searched by a TextPresso-powered natural language processing method, a MeSH browser, a keywords search, and an automatic literature update service. To efficiently annotate Brucella genes using the large amount of literature publications, a literature mining and curation system coined Limix is developed to integrate computational literature mining methods with a PubMed-powered manual curation and management system. The Limix system is used to quickly find and confirm 107 Brucella gene mutations including 75 genes shown to be essential for Brucella virulence. The 75 genes are further clustered using COG. In addition, 62 Brucella genetic interactions are extracted from literature publications. These results make possible more comprehensive investigation of Brucella pathogenesis. Other BBP features include publication email alert service, Brucella researchers’ contact database, and discussion forum.

Conclusion: BBP is a gateway for Brucella researchers to search, analyze, and curate Brucella genome data originated from public databases and literature. Brucella gene mutations and genetic interactions are annotated using Limix leading to better understanding of Brucella pathogenesis.
Background

*Brucella* is a Gram-negative, facultative intracellular cocco-bacillus which causes brucellosis in humans and animals [1]. *Brucella* are taxonomically placed in the alpha-2 subdivision of the class Proteobacteria. Traditionally there are six species of *Brucella* based on the preferential host specificity: *B. melitensis* (goats), *B. abortus* (cattle), *B. suis* (swine), *B. canis* (dogs), *B. ovis* (sheep) and *B. neotomae* (desert mice); two new species *B. cetaceae* (cetacean) and *B. pinnipediae* (seal) have recently been discovered [2]. The first four species are pathogenic to humans in decreasing order of severity making brucellosis a zoonotic disease. These *Brucella* species have been identified as priority agents amenable for use in biological warfare and bio-terrorism and listed as CDC/NIAID category B priority pathogens.

Complete genome sequences of 4 *Brucella* strains are currently available [3-6]. A typical *Brucella* genome usually has two circular chromosomes of approximately 2.1 MB and 1.2 MB. There are approximately 3,200 – 3,400 genes in each genome. The DNA sequences of different *Brucella* spp. share greater than 90% identity [4,6,7]. Genome sequences and annotated data are publicly available from existing databases such as RefSeq [8], Swissprot [9], and the TIGR Comprehensive Microbial Resource (CMR) [10]. These databases come from different sources and have different focuses. Different data visualization and analysis tools are also available in these database systems and other genome analysis systems. A web portal that integrates these data and analysis resources will greatly help *Brucella* gene research.

*Brucella* genome data in current databases is largely derived from computational analysis without literature support. It is partially due to the lack of a literature mining and curation system. The large amount of literature data can be used to not only validate the data obtained from computational analysis but also provide new insights not available from computational analysis. Literature mining techniques are being developed rapidly in the context of the genomic fields [11,12]. For example, Hu et al., [13] describe a rule-based system, RLIMS-P, for literature mining and database annotation of protein phosphorylation from MEDLINE abstracts. Stephens et al., [14] present an association and function discovery method to extract gene-gene interactions from co-occurring genes in MEDLINE abstracts. Hoffmann et al. [12] list more than 20 main text mining repositories and systems that are currently available. Compared to basic keyword search, many effective literature retrieval programs connect textual evidence to ontologies as main repository of formally represented knowledge. Ontologies are conceptual models that support consistent and unambiguous knowledge sharing and provide a framework for knowledge integration. TextPresso is a natural language processing (NLP) and ontology-based literature search engine with significant efficiency in biomedical literature retrieval [15]. Since computational literature mining techniques (e.g., TextPresso) still cannot guarantee precise retrieval, time consuming manual literature curation is required to obtain accurate results for database storage. It is possible for manual curation and computational text mining to work together for rapid retrieval and analysis of facts with standardization of the extracted information [16]. The PubSearch literature curation software is a literature curation management system with a powerful manual curation capability [17]. Our strategy of integrating different computational text mining tools including a TextPresso-powered program with a PubSearch-powered manual curation system has led to the development of a literature mining and curation system coined "Limirx" that is currently applied to *Brucella* genome annotation.

The *brucellae* infect phagocytic macrophages and non-phagocytic epithelial cells (e.g., HeLa cells) in *vivo* and *in vitro* [18-20]. *Brucella* virulence relies on its ability of intracellular survival and replication. It is still unclear how many *Brucella* genes are essential for intracellular virulence and how virulent *Brucella* genes interact. It is hypothesized that mechanisms of *Brucella* pathogenesis can be better understood by systematically annotating *Brucella* gene mutations and genetic networks from all *Brucella* literature papers.

We have developed the *Brucella* Bioinformatics Portal (BBP) with focus on integrating *Brucella* genome data and analysis tools from existing resources and annotating *Brucella* genes and gene-gene interactions from literature publications. The updated information allows more comprehensive examination of *Brucella* pathogenesis. These genome annotation systems, together with other programs including publication email alert, *Brucella* researchers’ contact database, and discussion forum, makes BBP an ideal bioinformatics portal for the *Brucella* research community. The BBP website is publicly available [see Additional file 1].

Results and discussion

System architecture

A three-tier system architecture is implemented with two Linux servers (Figure 1). Users submit database or analysis queries using front-end web browsers via HTML forms. These requests are processed using PHP/Java/Perl/(middle-tier, application server) against the Oracle relational database (back-end, database server), or XML and MySQL databases in application server. The result of each query is then presented to the users through the web browser. The BBP Oracle database stores all the data schema and data for the programs developed in-house, including the liter-
nature MeSH data, ContactsDB, registration information, and Forum data. The Brucella Limix and BGBrowser databases are implemented in the application server using MySQL since both systems are modified from open sources with MySQL as the default database management system. The TextPresso XML database is also implemented in the application server. Table 1 shows all the data and analysis resources incorporated by BBP.

**Brucella genome data query, browsing, and analysis**

Two complementary programs, the InterBru database system and Brucella genome browser (BGBrowser), have been developed for Brucella genome data query, browsing, and analysis. Both programs allow query of *Brucella* gene data from all four complete genomes: *B. melitensis* 16 M [5], *B. suis* 1330 [3], and *B. abortus* strain 9–941 [4] and strain 2308 [6]. The InterBru web query interface allows users to search *Brucella* genes based on different gene features such as gene name, locus tag, protein molecular weight (MW) and isoelectric points (PI), RefSeq identifier, and Swissprot accession number (Figure 2A). The Generic Genome Browser, also known as GBrowse [21], is a popular genome browser tool due to its portability, simple installation, and convenient data input and easy integration with other software programs. Developed as a member of the GBrowse family, the BBP BGBrowser program provides web query interface and graphic representation of specific *Brucella* genes, proteins, and RNA features (Fig-
Table 1: Public databases and software programs linked or used in BBP. Unique database identifiers (e.g., RefSeq ID) are usually stored for linking to public database web pages. Brucella literature abstracts and full text PDF files are also extracted from PubMed. Software programs are integrated into BBP in different ways.

| Resources | Databases and analysis programs | Comments |
|-----------|---------------------------------|----------|
| Databases |                                 |          |
| NCBI      | PubMed                          | Biomedical publications |
|           | MeSH                            | Medical Subject Headings |
|           | RefSeq                          | Reference sequences |
|           | Genome                          | Genome summary |
|           | Gene                            | Gene information |
|           | Protein                          | Protein information |
|           | Nucleotide                      | Nucleotide information |
|           | CDD                             | conserved domains |
|           | Taxonomy                        | Brucella taxonomy information |
|           | 3D structure DB                  | 3D structures (typically of related proteins) |
|           | Feature tables                  | Protein coding genes, functional and structural RNAs, |
| EBI & SIB | Swissprot                       | Annotated protein data |
|           | TrEMBL                          | protein data |
|           | InterPro                        | protein families, domains and functions |
|           | PROSITE                         | protein families and domains |
| TIGR      | CMR                             | codon usage, EC numbers, condensed genome display, role category graph, GO terms, gene attribute, Role category gene list, RNAs, terminators |
| Others    | TIGRfam                         | TIGRfam assignments |
|           | PFam                            | protein domains and families |
|           | ProDom                          | protein domain families |
| Software programs integrated | BLAST                           | Blastn, blastp, blastx, tblastn, tblastx, PSI/PHI Blast, Mega Blast, Blast 2 sequences |
| NCBI      | GBrowse                         | genome browse and analysis |
|           | TextPresso                      | NLP text mining |
|           | PubSearch                       | literature curation management |
| GMOD      | Other                           | Programming tools |

The following is a typical scenario when a Brucella researcher searches for more information about B. abortus sodC gene encoding Cu/Zn superoxide dismutase (SOD). The user starts with querying "sodC" gene in InterBru (Figure 2A). Four Brucella sodC genes from 4 Brucella genomes will be found, including one from B. abortus strain 2308 and one from B. abortus strain 9–941. The detail information about the sodC gene in strain 2308 is shown in the detailed Brucella gene information page (Figure 2C). This page includes basic gene information and through unique database identifiers links to many public databases, such as RefSeq, GenBank, Swissprot, InterPro, and PubMed. This page also contains sodC-specific gene annotation and genetic interaction data curated by the BBP team from literature using the Brucella Limix system. A link to the Brucella Limix is also available for users to annotate sodC gene. A direct link to PubMed allows users to access all Brucella sodC-related publications. Both DNA and protein sequences are provided with additional links to internal BLAST search services (regular Blast, Psi/Phi Blast, and Mega Blast) where different Brucella nucleotide and protein sequence libraries have been created for convenient use. For example, a simple Blastn search indicates that the sodC DNA sequence in B. abortus strain 2308 is 100% identical to that in B. abortus strain 9–941 but 99% identical to that in B. melitensis strain 16 M and B. suis strain 1330. The protein sequences in the four genomes are 100% identical to each other. The user is also directed to the BGBrowser to inspect the genes next to sodC in the genome, annotate restriction sites, or perform other analyses (Figure 2B). To get more information, the user can submit questions in the BBP discussion Forum or email to the Brucella listserv.
Brucella literature search

Four computational literature search methods have been developed to search Brucella literature: TextPresso for Brucella, MeSH browser, keyword search, and automatic Brucella publication update.

Textpresso is an information retrieval system available from the Generic Software Components for Model Organism Databases (GMOD) [22]. It splits papers into sentences and further to XML-tagged words or phrases, which are classified using categories of ontology. The specifically
designed ontology can be used to query information on specific classes of biological concepts (e.g., gene, mutant) and their relationships (e.g., association, regulation). It has been used in WormBase [23] and many other projects [24]. We have adopted and extended TextPresso for Brucella literature text mining. Currently it stores abstract information of 3930 Brucella publications. Among them 1083 papers have full-text contents. While it takes approximately 24 hours for TextPresso to preprocess these 3930 PubMed abstracts and 1083 full text PDF files in our server, the online query process is fast (~0.5 sec/query).

MeSH is the controlled vocabulary of medical and scientific terms assigned by experts and used for indexing articles in PubMed. MeSH terminology provides a consistent approach to retrieve information that may use different terminology for the same concepts. The BBP MeSH browser enables users to locate Brucella articles by the MeSH terms in the hierarchical MeSH tree structure. Figure 3 illustrates the detailed tree display for those who want to search for gene deletion.

A user can also search the locally built Brucella literature database by keywords such as author, journal, year, issue, and abstract. Although the Brucella literature database is updated periodically, it may miss the newest Brucella literature publications. In order to capture this portion of the literature, a BBP internal program has been developed to automatically extract the newly published Brucella papers from PubMed.

**Brucella literature mining and curation system (Limix)**

Although the text mining approaches efficiently provide queried articles and even sentences, the retrieved results are not precise and cannot be directly edited and stored in database. By contrast, a manual literature curation and management system usually allows edited literature data to be stored in database. The Brucella Limix system is developed through integrating literature text mining technologies (including TextPresso for Brucella, keywords search, and latest literature updates) and the PubSearch-powered manual literature curation and management program. Within one web page, a data curator is able to perform computational text mining, copy highlighted text from the computational search to an editable text field, edit, and further submit reviewed results to the backend database (Figure 4). Limix allows curators to conveniently search, update, validate and insert gene information. Figure 4 shows an example of using Limix to search and annotate phenotypes of a sodC mutation from Brucella literature. Limix is also a distributed curation system that is capable of involving external experts to support our curation efforts. Direct submissions from scientists will help keep the database as comprehensive, updated and accurate as possible.

**Literature-curated Brucella gene mutations and pathogenesis**

We have applied the Brucella Limix system for annotation of more than 900 Brucella genes. Out of more than 200 possible gene mutations from TextPresso-powered computational search, 107 mutations are manually confirmed, and 75 mutated genes are found to be attenuated inside macrophages or HeLa cells, or in an *in vivo* mouse model. It suggests that these 75 mutated *Brucella* genes are essential for *Brucella* virulence and pathogenesis. Although this list does not include those genes with attenuated mutation phenotype but without defined gene names, the number of attenuated mutations we have found is much more than any single research or review paper has discussed. The NCBI Clusters of Orthologous Groups (COGs) approach provides phylogenetic classification of proteins encoded in complete genomes [25]. The 75 *Brucella* genes are classified using the COG method for further analysis (Table 2). It first confirms the well-known pathogenesis mechanisms of *Brucella* type IV secretion system encoded by the virB operon [26], the BvrR-BvrS two-component regulatory system encoded by bvrR and bvrS [27], and the complete *Brucella* lipopolysaccharide [28]. Significant and stable attenuation are obtained in *Brucella* strains with mutations (e.g., *wboA*) resulting in the loss of normal lipopolysaccharide O-side-chain biosynthesis [29]. In addition, our curation clearly indicates the critical importance of transport and metabolism of various metabolites including amino acid, carbohydrate, lipid and inorganic ions (Table 2). Since the brucellae survive inside phagosomes of eukaryotic cells, bacterial attenuation after disruption of these genes suggests that the corresponding metabolites are not accessible to the bacteria inside the phagosomes, but they are essential for intracellular growth. Limix has also uncovered many gene mutations with important implications in understanding Brucella pathogenesis. For example, studies with a *B. abortus* sodC mutant suggest that Cu/Zn SOD protects *B. abortus* from respiratory burst of host macrophages [30]. The presence of an attenuated *fliF* mutant suggests a possible role for flagella in virulence [31], and it further leads to the recent discovery of a polar and sheathed flagellar structure in the early log phase of a growth curve in 2YT nutrient broth [32]. This finding has changed previous dogma that non-motile *Brucella* species do not have functional flagella.

**Literature-curated Brucella genetic interactions and pathogenesis**

*Brucella* pathogenesis relies on interactions between individual *Brucella* genes. Besides individual *Brucella* gene mutations, we have also analyzed *Brucella* genetic interactions using all accessible *Brucella* literature publications. As defined in the original TextPresso paper [15], *Brucella* genetic interactions are retrieved using a TextPresso-pow-
MeSH Browser. All the Brucella literature publications can be visualized by the interactive MeSH-tree browser. The two clickable numbers in each line links to all publications with the term as a MeSH term or a major MeSH term, respectively. This figure shows the hierarchical MeSH tree structure leading to Mutagenesis and Gene Deletion.
Limix also allows curators to add matches from at least one paper (Table 3). If the cutoff becomes 2, the positive rate becomes 23.5% (50/213) if the cutoff becomes 2. Full text contents are used. For example, the true positive number of matches in case both abstracts and inclusion of only abstracts leads to higher true positive rate (44.7% vs. 4.4%). It is possible to significantly increase true positive rate by raising the searching threshold (1330) are confirmed to be true interactions if both abstracts and full text contents are searched for in addition to abstracts. On the condition that only one match is required for positive hits during computational text mining, 58 out of 1330 possible genetic interactions (true positive rate is 4.4% (58/1330)) are confirmed to be true interactions if both abstracts and full text contents are used, and only 17 out of 38 genetic interactions are confirmed to be true (true positive rate is 44.7% (17/38)) if only abstracts are considered (Table 3). This indicates that inclusion of full text contents results in more confirmed results (58 vs. 17), while inclusion of only abstracts leads to higher true positive rate (44.7% vs. 4.4%). It is possible to significantly increase true positive rate by raising the searching threshold of the number of matches in case both abstracts and full text contents are used. For example, the true positive rate becomes 23.5% (50/213) if the cutoff becomes 2 matches from at least one paper (Table 3).

Limix also allows curators to add Brucella genetic interactions that are not detected by the TextPresso-based text mining approach. Currently 62 genetic interactions are available in the Limix databases. There are 48 genes involved in these interactions, and 28 of them are shared with the attenuated Brucella gene mutation list as discussed above. The finding of these genetic interactions has provided more comprehensive investigation of Brucella pathogenesis. For example, it not only confirms the importance of type IV secretion system and the BvrR-BvrS two-component regulatory system in Brucella pathogenesis. For example, it not only confirms the importance of type IV secretion system and the BvrR-BvrS two-component regulatory system in Brucella pathogenesis. For example, it not only confirms the importance of type IV secretion system and the BvrR-BvrS two-component regulatory system in Brucella pathogenesis.

Table 2: Clustering of 75 attenuated Brucella genes found from literature search using the COG classification method.

| COG clusters                                                                 | Mutated Brucella genes                                      |
|------------------------------------------------------------------------------|-------------------------------------------------------------|
| C: Energy production and conversion                                          | 4: cydB, enyB, narG, pyc                                        |
| D: Cell cycle control, mitosis and meiosis                                   | 1: virB5                                                       |
| E: Amino acid transport and metabolism                                       | 15: arC, ascC, gcvF, gcvT, glnA, glyA, hisC, hisD, hisF, leuA, leuc, lysA, pepN, pheA, serB |
| F: Nucleotide transport and metabolism                                       | 10: dut, purD, purE, purF, purH, purI, purM, purN, pyrB, pyrD |
| G: Carbohydrate transport and metabolism                                     | 7: eryC, gnd, hvd, pgi, pgm, upgA, zwf                           |
| H: Coenzyme transport and metabolism                                         | 3: cobB, hemH, hivC                                             |
| I: Lipid transport and metabolism                                            | 2: bocA, upp5                                                  |
| J: Translation                                                               | 3: pth, rpsS, rpsA                                             |
| K: Transcription                                                             | 3: miaA, oxyR, rpoA                                            |
| L: Replication, recombination and repair                                      | 2: recA, xseA                                                 |
| M: Cell wall/membrane biogenesis                                             | 6: gmd, omp10, omp19, omp25, virB1, wboA                        |
| O: Posttranslational modification, protein turnover, chaperones              | 5: cydD, degP, dnaK, lon, tig                                  |
| P: Inorganic ion transport and metabolism                                    | 2: sodC, znuA                                                  |
| Q: Secondary metabolites biosynthesis, transport and catabolism              | 1: pncA                                                       |
| R: General function prediction only                                          | 2: gltD, hfq                                                   |
| T: Signal transduction mechanisms                                            | 4: bvrR, bvrS, ntrC, ntrY                                       |
| U: Intracellular trafficking and secretion                                   | 9: hif, virB2, virB10, virB1 I, virB3, virB4, virB6, virB8, virB9 |
| -: Not in COGs                                                                | 6: cowB, glnD, gltP, gpt, norM, wbbB                           |

A software program based on Graphviz [37] is developed to display all the genetic interactions in the Scalable Vector Graphics (SVG) format [38] (Figure 5). SVG is a language for describing two-dimensional graphics and
Integrated computational text mining and manual curation in Limix

Figure 4

Integrated computational text mining and manual curation in Limix. The computational text mining frame shows a typical TextPresso-type result after query for the sodC keyword and "mutant" category. All sodC words and words under mutant category are clearly labeled in colors. One sentence containing both sodC and mutant words is highlighted in bold and considered as one match. A curator can easily highlight and copy text from this frame to an editable text field below the frame within the same page. The data can be further edited and submitted to a backend database by clicking an 'update' button. Other literature retrieval approaches (e.g., keywords search) are also available in the computational text mining frame.
graphical applications in XML and is currently supported by many internet browsers. A click on each node in the map will link to the detailed gene information page in InterBru search. Once an edge (straight line) is clicked, the detail on the specific gene-gene interaction is shown. Figure 4 demonstrates the interaction between two Brucella genes sodC and hfq. A future direction is to integrate our curated genetic interaction data with known interaction and pathway knowledge from existing databases, such as KEGG [39], BIND [40], and DIP [41].

### Brucella Genetic Interaction Map

To view our curated Brucella genetic interaction map, you need SVG support, which is available in Firefox version >= 1.5. For Microsoft Internet Explorer, you can install Adobe SVG Viewer from [here](#). Note: Click the blue edges for detailed interaction description below the map.

![Brucella Genetic Interaction Map](image)

### Table 3: TextPresso-predicted and manually curated Brucella genetic interactions

One match means one highlighted sentence containing at least 2 genes and at least one word under “association” or “regulation” category. Each match represents for one predicted genetic interaction. The results are shown by manually verified vs. TextPresso-predicted interactions. The number of verified vs. predicted interactions varies depending on the numbers (#) of matches and papers to use as the cutoffs and whether or not to use full text contents besides paper abstracts.

| # of papers | # of matches | # of papers | # of matches |
|-------------|--------------|-------------|--------------|
| 1           | 17/38        | 2           | 58/1330      |
| 2           | 3/5          | 3           | 33/71        |
| 3           | 1/1          | 4           | 28/50        |
| 4           | 1/1          | 8           | 8/10         |
| 8           | 0/0          |             | 6/7          |

Abstract Full text (containing abstract)

| # of papers | # of matches | # of papers | # of matches |
|-------------|--------------|-------------|--------------|
| 1           | 17/38        | 2           | 58/1330      |
| 2           | 3/5          | 3           | 33/71        |
| 3           | 1/1          | 4           | 28/50        |
| 4           | 1/1          | 8           | 8/10         |
| 8           | 0/0          |             | 6/7          |

**Figure 5**

**Brucella genetic interaction map and description.** Limix is used to find and confirm 62 Brucella genetic interactions. In the Brucella genetic interaction map displayed in a SVG form, any node can be clicked for detailed gene information, and any edge can be clicked to show description of the specific interaction.
BBP is designed to link international Brucella scientists and researchers. BBP contains a ContactsDB database that currently provides contact information for more than 100 Brucella researchers in the world. The ContactsDB can be queried based on first name, last name, address, city, institute, state, zip code, and country. Any Brucella researcher can also enter new contact information or update existing information using an interactive web page. The BBP discussion Forum has been created to facilitate discussion between scientists. Only registered BBP members can initiate a topic, reply to a message, or edit their own messages. Unregistered users can view all discussions. Up to now more than 50 Brucella researchers from 18 different countries have registered in BBP. Another BBP feature is the Publication Email Alert Service. This service automatically notifies users of newly published papers within a user-defined time interval. Those users who have not registered for this service can view new publications by visiting our automatic new Brucella paper updating website.

Conclusion
Many different databases related to Brucella genomes and genes exist. A variety of computational tools are also available for functional genomic analysis. The Brucella Bioinformatics Portal is a gateway to provide or link functional Brucella gene information and analysis tools useful for the Brucella researchers. Besides summarizing Brucella genomics related databases and analysis tools in HTML formats, we have also developed the InterBru database and the Brucella Genome Browser (BGBrowser). InterBru allows users to search for specific Brucella gene information and provide links to existing databases. BGBrowser provides graphic visualization and analysis tools. Since most of current Brucella genes and gene-gene interaction data are derived from computational analysis and often lack literature support, we further developed several computational Brucella literature search tools for efficient retrieval of Brucella articles. The Brucella Limix system is also developed to allow retrieved data from text mining tools to be directly copied, edited and submitted to a backend relational database. The Brucella Limix system has been used to annotate a large number of Brucella genes and to find 62 Brucella genetic interactions and 75 attenuated gene mutations from literature publications in PubMed. These annotated results provide more comprehensive understanding of Brucella pathogenesis. These programs, together with other portal features including the ContactsDB and Forum, facilitate the Brucella research community to obtain and annotate Brucella genome sequences in one website. BBP is the first integrated system for Brucella genome analysis.

BBP adopts and extends many open-source software programs for Brucella genome annotation including three GMOD open-source software programs, GBrowse, TextPresso, and PubSearch (Table 1). Many interactive graphical interfaces (e.g., MeSH browser and genetic interaction map) have also been developed for efficient literature mining and database curation. While many NLP-based text mining tools (e.g., TextPresso) significantly improve the capability of biomedical text mining, an automatic literature retrieval tool that can be as accurate as manual literature curation still does not exist [12]. As far as we know, among existing web-based dedicated genome databases, BBP is the first to strongly integrate a literature manual curation and management system (e.g., PubSearch) with NLP-based computational literature mining techniques (e.g., TextPresso for Brucella) into an efficient literature mining and curation system (Limix). The BBP Limix system also provides a genetic solution for annotating other genomes and genes based on published literature data.

Methods

Server and programming tools
This BBP system is built on two Dell Poweredge 2580 servers, one serving as database server and another as application server. Both servers are running the Redhat Linux operating system (Redhat Enterprise Linux ES 4). The database server is powered by Oracle 10g database management system. Two open source software programs, Apache HTTP Server and Apache Tomcat, are installed as the HTTP application server and the servlet container respectively. Different programming languages including PHP, Perl, and Java are implemented for development of a variety of BBP modules. The two servers also back each other regularly to secure the data.

InterBru and BGBrowser
InterBru is a web-based relational database system that contains various Brucella data and links to public databases. The protein MW and PI are calculated from the protein sequences using the modules (Bio::Tools::pI Calculator and Bio::Tools::SeqStats) from Bioperl [42]. The InterBru data can be searched by different features and sorted for proper display (Figure 2A). The Brucella Genome Browser (BGBrowser) (Figure 2B) is developed based on the GBrowse [21], one of the GMOD software programs [22]. In order to speed up the query process, all Brucella sequence and annotation information for BGBrowser are stored in the database server instead of flat files. Both InterBru and BGBrowser share the same output page of detailed gene information (Figure 2C).

Blast@BBP
The Blast module in BBP uses the latest web server version of BLAST obtained from NCBI [43]. It includes regular BLAST services (blastn, blastp, blastx, tblastn, tblastx),
PSI/PHI BLAST, Mega BLAST, and BLAST 2 sequences. These services are implemented in the BBP application server and can be used to search nucleotide or protein BLAST libraries containing sequences from individual or combined Brucella genomes. The sequence libraries are updated periodically to reflect newly curated annotations and when new genomes are added.

**TextPresso-powered Brucella literature search**

As a software program from the GMOD, Textpresso uses a modified GNU license and is free for academic purposes [24]. The Textpresso package is downloaded from the Textpresso website [24]. An automatic download program is first used to download and extract from PubMed all Brucella-related article information, including titles, authors, publish years, volumes, pages, journal names and abstracts. A BBP script is also developed to extract all Brucella-related full-text PDF files from PubMed. The PDF files are converted into plain text files using the open source XPDF [44]. The converted full text together with abstracts and titles are tokenized into sentences and then to XML-tagged words or phrases representing different ontology categories according to a pre-defined ontology format. All the processed information, including fully annotated abstracts, titles, full texts, citation information, authors, years, keywords and categories, is indexed for efficient query. A web query interface is installed and modified for users to search against the indices and check detailed matching records.

**Brucella Limix**

To develop the literature mining and curation system, the PubSearch version 0.81 is first adopted and extended from GMOD [22]. PubSearch is originally designed for Arabidopsis in the TAIR project [17]. We replace the TAIR data from the software download with new Brucella genomic data from NCBI, Swiss-Prot and other repositories. Currently Limix stores information for 6033 Brucella-related articles downloaded from PubMed, including those without abstract content. The 20346 GO ontology terms downloaded from the GO database [45] allow users to associate Brucella gene names with specific GO terms. Limix also includes batch mode loading of data from other databases (e.g., PubMed and GO databases), and data indexing. We have also modified many PubSearch features to make it fit in with bacterial genome annotation. The PubSearch-powered page programmed in Java is used as the primary Limix web page specifically for manual curation and management. Since Textpresso uses Perl CGI instead of Java, we use a HTML frame inside the primary page to hold the Textpresso-powered computational text mining program (Figure 4). The text mining HTML frame also contains literature keywords search and automatic Brucella literature update programs. A JavaScript program is developed to copy highlighted sentences from the text mining frame to an editable text field in the primary curation page.

**MeSH Browser**

The Brucella literature MeSH browser is developed by utilizing the hierarchy tree structure of MeSH terms downloaded from PubMed and stored in the BBP Oracle database. MeSH Browser allows users to search associated articles to a specific MeSH term in the MeSH tree by clicking and expanding the MeSH nodes. The nodes in the MeSH tree can be dynamically expanded with no waiting for pages to reload by using the Asynchronous JavaScript and XML (Ajax) technique [46].

**Publication email alert service and automatic updates**

The BBP publication email alert service is initiated by a subscribed user to specify the notification frequency (daily, weekly, bimonthly or monthly) and the keywords to be searched against the PubMed database. A daily Linux cron job checks the subscription database, searches for updates in PubMed, and sends the updated paper notification to the users through email. The automatic literature update program allows all Brucella-related publications from the current and previous months automatically updated in the BBP website. It is implemented by dynamically querying PubMed for updated publications during a certain month when the page is opened. This program is also integrated into the Brucella Limix for data curators to obtain the newest publications not stored in local publication database.

**ContactsDB and forum**

The ContactsDB database stores contact information of individual Brucella researchers in an Oracle database. A PHP program is developed for the users to query, submit, and update contact information in the BBP ContactsDB web page. The discussion forum program is also implemented with PHP and Oracle database.

**Abbreviations**

Ajax Asynchronous JavaScript and XML

BBP Brucella Bioinformatics Portal

CDD The Conserved Domain Database

CMR TIGR Comprehensive Microbial Resource

COG The Clusters of Orthologous Groups

EC Enzyme Commission

GMOD Generic Software Components for Model Organism Databases
GO Gene Ontology
Limix Literature Mining and Curation System
MeSH Medical Subject Headings
MW Molecular weight
NCBI The National Center for Biotechnology Information
NIAID The National Institute of Allergy and Infectious Diseases
NIH National Institutes of Health
NLP Natural Language Processing
SOD Superoxide Dismutase
PI Isoelectric points
TIGR The Institute for Genomic Research

Authors' contributions
ZX: Current webmaster, software programmer, and database administrator.
WZ: Previous webmaster, software programmer, and database administrator.
YH: Project initiator, designer, software programmer, and manager. As an active Brucella researcher, YH also curated all Brucella genetic interactions and mutations available in BBP using the Brucella Limix system.

Additional material

Additional File 1
BBP website screenshot. The image provided is the screenshot of the Brucella Bioinformatics Portal (BBP) website home page. The BBP URL is: [http://helab.bioinformatics.med.umich.edu/bbp/](http://helab.bioinformatics.med.umich.edu/bbp/). Click here for file [http://www.biomedcentral.com/content-supplementary/1471-2105-7-347-S1.tiff](http://www.biomedcentral.com/content-supplementary/1471-2105-7-347-S1.tiff)

Acknowledgements
WZ was supported by the NIH-NIAID R21 grant (#1R21AI057875-01) to Dr. He. ZX was supported by Dr. He’s startup funding from the University of Michigan. We appreciate the encouragement and suggestions provided by Drs. Stephen Boyle, Nammalwar Srinaganathan, Gerhardt Schurig, and Bruno Sobral at the Virginia Polytechnic Institute and State University, and Dr. Raju Lathigra at the Walter Reed Army Institute of Research. We thank Dr. Hans-Michael Muller from the California Institute of Technology for his technical help in our development of Textpresso-powered Brucella literature search engine. We also appreciate the reviews and suggestions provided by our colleague Dr. Lesley Colby at the University of Michigan Medical School.

References
1. Corbel MJ: Brucellosis: an overview. Emerg Infect Dis 1997, 3(2):213-221.
2. Cloeckaert A, Verger JM, Grayon M, Paquet JV, Garin-Bastuji B, Foster G, Godfroid J: Classification of Brucella spp. isolated from marine mammals by DNA polymorphism at the omp2 locus. Microbes Infect 2001, 3(9):729-738.
3. Paulsen IT, Seshadri R, Nelson KE, Eisen JA, Heidelberg JF, Read TD, Dodson RJ, Lymayam L, Brinkac LM, Beaman MJ, Daugherty SC, Deboy RT, Durkin AS, Kolonya JF, Madupu R, Nelson WC, Ayodeji B, Kraul M, Shetty J, Malek J, Van Aken SE, Riedmuller S, Tettelin H, Gill SR, White O, Salzberg SL, Hoover DL, Lindler LE, Halling SM, Boyle SM, Fraser CM: The Brucella suis genome reveals fundamental similarities between animal and plant pathogens and symbionts. Proc Natl Acad Sci U S A 2002, 99(20):13148-13153.
4. Halling SM, Peterson-Burch BD, Bricker BJ, Zierner RL, Qing Z, Li LL, Kapur V, Alt DP, Olsen SC: Completion of the genome sequence of Brucella abortus and comparison to the highly similar genomes of Brucella melitensis and Brucella suis. J Bacteriol 2005, 187(8):2715-2726.
5. DelVecchio VG, Kapralov V, Redkar RJ, Patra G, Mujer C, Los T, Ivanova N, Anderson I, Bhattacharya A, Lykidis A, Reznik G, Jablonski L, Larsen N, D’Souza M, Bernal A, Mazur M, Gotsman E, Selkov E, Eizer PH, Hagius S, O’Callaghan D, Leteszeny JJ, Haselkorn R, Kyrpides N, Overbeek R: The genome sequence of the facultative intracellular pathogen Brucella melitensis. Proc Natl Acad Sci U S A 2002, 99(1):443-448.
6. Chain PS, Comerci DJ, Tolmasque MY, Larimer FW, Malfatti SA, Vergez LM, Aguero F, Land ML, Ugalde RA, Garcia E: Whole-genome analyses of speciation events in pathogenic Brucellae. Infect Immun 2005, 73(12):8353-8361.
7. Ratushna VG, Sturgill DM, Ramamoorthy S, Reichow SA, He Y, Lathiagra R, Srinaganathan N, Halling SM, Boyle SM, Gibas CJ: Molecular targets for rapid identification of Brucella spp. Microb Biotechnol 2006, 6:113.
8. Wheeler DL, Barrett T, Benson DA, Bryant SH, Cambillau C, Choudhuri R,.e309. Chisholm GM, Church DM, DiCuccio M, Edgar R, Federhen S, Hettich RL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JJ, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Shlipak M, Starchenko G, Suzek TO, Tatusova T, Tatusova TA, Wagner L, Yaschenko E: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2005, 33(Database issue):D39-45.
9. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O’Donovan C, Phan I, Pilbout S, Schneider M: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003, 31(1):365-370.
10. Peterson JD, Umayam LA, Dickinson T, Hickey EW, White O: The Comprehensive Microbial Resource. Nucleic Acids Res 2001, 29(1):123-125.
11. Scherf M, Apple A, Werner T: The next generation of literature analysis: integration of genomic analysis into text mining. Brief Bioinform 2005, 6(3):287-297.
12. Hoffmann R, Krallinger M, Andres E, Tamames J, Blaschke C, Valencia A: Text mining for metabolic pathways, signaling cascades, and protein networks. Sci STKE 2005, 2005(283):pe21.
13. Hu ZZ, Narayanaswamy M, Ravikumar KE, Vijay-Shanker K, Wu CH: Literature mining and database annotation of protein phosphorylation using a rule-based system. Bioinformatics 2005, 21(11):2759-2765.
14. Stephens M, Palakia M, Mukhopadhyay S, Raje R, Mostafa J: Detecting gene relations from Medline abstracts. Pac Symp Biocomput 2001:483-495.
15. Muller HM, Kenny EE, Sternberg PW: Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2004, 2(11):e309.
16. Rebolz-Schuhmann D, Kirsch H, Couto F: Facts from text—is text mining ready to deliver? PLoS Biol 2005, 3(2):e65.
tional annotation of the Arabidopsis genome using control-
led vocabularies. Plant Physiol 2004, 135(2):745-755.
18. Ko J, Splittger GA: Molecular host-pathogen interaction in brucellosis: current understanding and future approaches to vaccine development for mice and humans. Clin Microbiol Rev 2003, 16(1):65-78.
19. Kohler S, Michaux-Charachon S, Porte F, Ramuz M, Liuard JP: What is the nature of the replicative niche of a stealthy bug named Brucella? Trends Microbiol 2003, 11(5):215-219.
20. Roopt RM, Bellaire BH, Valderas MW, Cardelli JA: Adaptation of the Brucellae to their intracellular niche. Mol Microbiol 2004, 52(3):621-630.
21. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: The generic genome browser: a building block for a model organism system database. Genome Res 2002, 12(10):1599-1610.
22. GMOD: The Generic Model Organism Database Project. [http://www.gmod.org].
23. Schwarz EM, Antoshechkin I, Bastani C, Bieri T, Blasiar D, Canaran P, Cervone F, Chan J, Chen N, Chen WJ, Davis F, Fiedler TJ, Girard L, Harris TW, Kenny EE, Kishore R, Kashower D, Lee R, Muller HM, Nakamura C, Ozersky P, Petcherski A, Rogers A, Spooner W, Tuli MA, Van Auken K, Wang L, Deimann R, Spieth J, Stein LD, Sternberg PW: WormBase: better software, richer content. Nucleic Acids Res 2006, 34(Database issue):D475-8.
24. Textpresso: Textpresso. [http://www.textpresso.org/].
25. Tatusov RL, Galperin MY, Dianov LG, Miroshnichenko AP, Koonin EV: The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 2000, 28(1):33-36.
26. O’Callaghan D, Cazevieille C, A K, Wang D, Durbin R, Spieth J, Stein LD, Sternberg PW: ACMG: automated single nucleotide polymorphism annotation for large-scale genotyping. Nucleic Acids Res 2003, 31(10):2811-2812.
27. Sola-Landa A, Pizarro-Cerda J, Grismbeck K, Wang D, Durbin R, Spieth J, Stein LD, Sternberg PW: ACMG: automated single nucleotide polymorphism annotation for large-scale genotyping. Nucleic Acids Res 2003, 31(10):2811-2812.
28. Allen CA, Adams LG, Ficht TA: Transposon-derived Brucella abortus mutant rough mutant are attenuated and exhibit reduced intracellular survival. Infect Immun 1998, 66(3):1008-1016.
29. McQuiston JR, Vemulapalli R, Inzana TJ, Schurig GG, Srinagananthan N, Fritzinger D, Hadfield TL, Warren RA, Smellings N, Hoover D, Hal- ing S, Boyle JM: Genetic characterization of a Tn5-disrupted glycosyltransferase gene homolog in Brucella abortus and its effect on lipopolysaccharide composition and virulence. Infect Immun 1999, 67(8):3830-3835.
30. Gee JM, Valderas MW, Kovach ME, Gripek VK, Robertson GT, Ng WL, Richardson JM, Winkler ME, Roopt RM: The Brucella abortus Ccu gene product is required for optimal resistance to oxidative killing by murine macrophages and wild-type virulence in experimentally infected mice. Infect Immun 2005, 73(5):2873-2880.
31. Lesrato P, Dricot A, Delrus RM, Lambert C, Martineili V, De Bolle X, Letesson JJ, Tibor A: Attenuated signature-tagged mutagenesis mutants of Brucella melitensis identified during the acute phase of infection in mice. Infect Immun 2003, 71(12):7035-7040.
32. Fretin D, Faconnier A, Kohler S, Halling S, Leonard S, Nijenhuis C, Pernoz J, Lesrato P, Delrus RM, Danese I, Vanderhauze J, Tibor A, DeBolle X, Letesson JJ: The sheathed flagellum of Brucella melitensis is involved in persistence in a murine model of infection. Cell Microbiol 2005, 7(5):687-698.
33. Marchesini M, Ugalde JE, Csizmene C, Comerci DJ, Ugalde RA: N-terminal-capturing screening system for the isolation of Brucella abortus genes encoding surface exposed and secreted proteins. Microb Pathog 2004, 37(2):95-105.
34. Robertson GT, Roopt RM: The Brucella abortus host factor I (HF-I) protein contributes to stress resistance during stationary phase and is a major determinant of virulence in mice. Mol Microbiol 1999, 34(4):690-700.
35. Roopt RM, Gee JM, Robertson GT, Richardson JM, Ng WL, Winkler ME: Brucella stationary-phase gene expression and virulence. Annu Rev Microbiol 2003, 57:57-76.