Comparative Genome Analysis of Three *Brucella* spp. and a Data Model for Automated Multiple Genome Comparison.

David Matthew Sturgill

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Biology

Graduate Committee Members:
Dr. Cynthia Gibas, Chair
Dr. Stephen Boyle
Dr. Khidir Hilu
Dr. Stephen Melville
Dr. Jennifer Weller

September 12, 2003
Blacksburg, Virginia

Keywords: Comparative Genomics, Bioinformatics, *Brucella*, Host-pathogen interaction

Comparative Genome Analysis of Three *Brucella* spp. and a Data Model for Automated Multiple Genome Comparison.

David Matthew Sturgill

ABSTRACT

Comparative analysis of multiple genomes presents many challenges ranging from management of information about thousands of local similarities to definition of features by combination of evidence from multiple analyses and experiments. This research represents the development stage of a database-backed pipeline for comparative analysis of multiple genomes. The genomes of three recently sequenced species of *Brucella* were compared and a superset of known and hypothetical coding sequences was identified to be used in design of a discriminatory genomic cDNA array for comparative functional genomics experiments. Comparisons were made of coding regions from the public, annotated sequence of *B. melitensis* (GenBank) to the annotated sequence of *B. suis* (TIGR) and to the newly-sequenced *B. abortus* (personal communication, S. Halling, National Animal Disease Center, USDA).

A systematic approach to analysis of multiple genome sequences is described including a data model for storage of defined features is presented along with necessary descriptive information such as input parameters and scores from the methods used to define features. A collection of adjacency relationships between features is also stored, creating a unified database that can be mined for patterns of features which repeat among or within genomes.

The biological utility of the data model was demonstrated by a detailed analysis of the multiple genome comparison used to create the sample data set. This examination of genetic differences between three *Brucella* species with different virulence patterns and host preferences enabled investigation of the genomic basis of virulence. In the *B. suis* genome, seventy-one differentiating genes were found, including a contiguous 17.6 kb region unique to the species. Although only one unique species-specific gene was identified in the *B. melitensis* genome and none in the *B. abortus* genome, seventy-nine differentiating genes were found to be present in only two of the three *Brucella* species. These differentiating features may be significant in explaining differences in virulence or host specificity. RT-PCR analysis was performed to determine whether these genes are transcribed *in vitro*. Detailed comparisons were performed on a putative *B. suis* pathogenicity island (PAI). An overview of these genomic differences and discussion of their significance in the context of host preference and virulence is presented.

Acknowledgements

I would like to thank the following people, without whom this work would not have been possible:

My advisor Cynthia Gibas, for her patience and guidance and for answering innumerable questions. Thanks for unwavering dedication and genuine concern for her student's best interests and education.

The members of my review committee – Steve Melville, Stephen Boyle, Khidir Hilu for keeping me grounded in Biology; Jennifer Weller for not letting distance impede her invaluable help and guidance on the data model and database.

Everyone in the *Brucella* Microarray Research Group, especially to Oliver, Raju, Nathan, Amanda, and Stephen for their insightful comments and help in Biology, for lively Friday meetings, and for making the manuscript possible.

Special thanks to Vlada Ratushna for designing primers for RT-PCR, and to Sherry Poff and Sheela Ramamoorthy for performing the experiments.

Further gratitude to Stephen Boyle, for introducing me to the *Brucella* community and helping me see the big picture.

Thanks to Shirley Halling, for her expert assistance and access to the draft *Brucella abortus* genome.

Thanks to ISCB, for a travel fellowship that allowed me to gain experience presenting my research to the bioinformatics committee.

And finally Julieta and Smokey, whose love and support made it all possible.

Table of Contents

List of Figures

List of Tables

Abbreviations

API	Application Programming Interface
BLAST	Basic Local Alignment Search Tool
bp	base pair
CDS	Coding Sequence
COG	Clusters of Orthologous Groups
ER	Endoplasmic Reticulum
ERD	Entity Relationship Diagram
HSP	High Scoring Pair
JGI	The Joint Genome Institute
kb	kilobases (1kb = 1000 bases)
Mb	Megabases (1Mb = 1,000,000 bases)
MUM	Maximal Unique Matches
NCBI	National Center for Biotechnology Information
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
Pfam	Protein Families
RDBMS	Relational Database Management System
RT-PCR	Reverse-Transcription Polymerase Chain Reaction
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
TIGR	The Institute for Genome Research
USDA	United States Department of Agriculture

1 Introduction

1.1 Motivation

The bacterial genus *Brucella* includes several closely related species that have different virulence patterns and host ranges. To help identify mechanisms of infection and study host response and virulence in this important pathogen, the *Brucella* Microarray Research Group of Virginia Tech was formed to collaborate on producing microarrays and perform expression profile experiments.

Microarrays allow the study of expression patterns for thousands of genes simultaneously. To study virulence patterns in *Brucella*, we required a microarray that allowed both controlled comparison of similar genes (with probes matching sequences in each species equally well) and discrimination between species. Our goal was to design a single microarray chip that included probes for genes common to each *Brucella* species, as well as differentiating genes unique to a species. This array could be used for rapid species identification of a *Brucella* infection for diagnostic purposes, and also for examination of gene expression patterns during infection to identify potential vaccine targets.

To design these microarrays, a comparative genomics analysis was needed to identify common and differentiating features between the *Brucella* species. We performed a systematic comparison of three *Brucella* species and produced sets of probes consisting of common and differentiating or unique sequences. This comparison served as a basis for development of the GenoMosaic prototype.

We developed the GenoMosaic prototype to fill a need for tools for automated feature level analysis of multiple genome sequences. This prototype is based on generalized feature definitions and is built to be flexible and allow analysis of varying feature types and types of sequences. The application of this process to the *Brucella* comparison proved to be a valuable test case of the prototype, leading to new insights into the biology of *Brucella* and producing probe targets for further expression profile experiments.

1.2 Relevance to the Literature

1.2.1 Comparative genomics

Comparative genomics is a relatively new field of study that has arisen as a major tool to find meaning in newly sequenced genomes. Unfortunately, the development of adequate computational tools to perform such studies has not kept pace with the recent proliferation of newly available genomic data. This gap between available data and analysis tools will widen as sequencing of entire microbial genomes becomes routine. The rate at which data is proliferating is increasing rapidly. For example, the Joint Genome Institute (JGI) recently completed sequencing 15 genomes in one month [1].

In studying prokaryotic systems comparative genomics has proven especially useful, leading to a better understanding of systematics, bacterial lifestyle, virulence, and host-pathogen interactions. Much of the practical utility of comparative genomics comes from its use as an annotation tool. When a genome is newly sequenced, the next step is the laborious process of finding where the genes are and assigning function to them. Genes that are similar in sequence are likely to be similar in function, and since there is a great deal of synteny among related organisms, a significant amount of annotation information can be transferred from one genome to another [1].

One application of comparative genomics that shows a great deal of promise is the study of virulence. For example, JGI recently sequenced several species of *Xyella* (a bacterial plant pathogen spread by insects). Each species has a different pattern of virulence, one infecting only grapes, others infecting one of a broad spectrum of plants. By comparing these species on the sequence level, they hope to identify genetic features that account for these differences [1]. For this reason, groups of related microbes are now being sequenced rather than just single representatives of disparate groups. The sequencing of multiple *Brucella* spp. is a recent example of this new trend.

1.2.2 Biology of Brucella

Brucella is a facultative intracellular pathogen that causes abortion in cattle, goats and sheep and a febrile illness ("undulant fever") in humans. Animal brucellosis is a serious problem worldwide and is endemic globally, excluding countries such as the U.S., most of Western Europe and Canada, which have instituted strict eradication measures. In areas where it is endemic, human brucellosis is quite common but often not diagnosed. In these areas, poor diagnosis and lack of treatment can result in life-threatening complications [2]. Brucellosis causes major economic losses to the agriculture industry, jeopardizes wildlife populations, and the causative agent is classified as a category B pathogen by the Centers for Disease Control and Prevention [3]. *Brucella* does not naturally survive for long periods of time outside the host, although prolonged cold temperatures favor its survival. Although the pathogen can replicate on culture medium, it is adapted to a vertebrate niche as an intracellular bacterial pathogen. Transmission to humans occurs by ingestion of milk, milk products or by direct contact with tissues and fluids of infected animals [4].

There are six recognized *Brucella* species that differ in their preference for certain hosts. *B. abortus* preferentially infects cattle, *B. melitensis* infects sheep and goats, and *B. suis* infects pigs. All three of these species and *B. canis* can infect humans, although *B. melitensis* is associated with the most serious human infections. The *Brucellae* are grouped with the alpha-proteobacteria and are related to other cell-associated parasites of plants and animals [5]. The classical *Brucella* taxonomy consists of six species (*B. melitensis*, *B. abortus*, *B. suis*, *B. neotomae*, *B. ovis* and *B. canis*) differentiated by their host preferences. Later observations of high homology from DNA-DNA hybridization studies has lead some adopt a monospecific system. This classification was also accepted by the Subcommittee on the Taxonomy of *Brucella* in 1986, along with the caveat that the classical species names should be used "to avoid confusion." Most microbiologists still prefer to use the (biologically meaningful) species system, which recently has been given more credence by detailed biochemical and genetic studies [6].

Macrophages are the first target of *Brucella* invasion, and the bacteria can survive within this naturally hostile intracellular environment [7]. Macrophages are important in transporting *Brucella* to tissues throughout the host, where they can survive in a variety of cell types [8]. Several studies have suggested that *Brucella* delays phagolysosomal fusion as a survival mechanism in macrophages [9], while in non-professional phagocytes *Brucella* appears to modulate the interior of the phagosome and evades intracellular degradation by avoiding the endocytic/phagocytic cascade [10]. It is not known definitively where *Brucella* replicates within the vertebrate cell. Observations have suggested that *Brucella* replicates within the rough endoplasmic reticulum (ER) in several cell types, including trophoblasts [11] and Vero cells [12]. Studies identifying ER markers on *Brucella*-containing compartments have also supported the theory of the ER as the site of replication [10]. The basic mechanisms for intracellular survival and proliferation are not conclusively known, nor are the reasons for the different virulence patterns among *Brucella* species.

To identify differentiating features that may explain these patterns, we have carried out a three-way genome comparison of *B. abortus, B. suis* and *B. melitensis* at both the nucleotide and predicted coding sequence (CDS) levels. Genomic sequence features that appear to distinguish the three species were probed using PCR and RT-PCR, to verify their existence and uniqueness and to test for expression *in vitro*. Identification of the patterns of differently expressed genes is the first step in the development of species-specific diagnostic tests and will provide targets for the elucidation of differences in host preference and mechanisms of virulence among these closely related species.

1.2.3 Automated annotation

The paucity of an adequate standard analysis tool has lead to a proliferation of specialized, curated resources that support user queries on previously performed analyses. These resources do not allow on-demand comparative analyses nor the incorporation of additional sequences to the comparison. Methods for sequence-level whole genome comparison exist that provide for some degree of user-directed analysis, but such methods are designed only for pairwise alignment, and are ineffective for comparing highly diverged genomes [13]. Annotation tools such as Artemis [14] and Apollo [15] allow management of annotation information and visualization of features and comparative visualization, but do not perform any sequence analysis. Automated annotation packages such as Genotator [16] and DNannotator [17] perform automated sequence analysis to produce *de novo* annotation, but are designed for single sequences and provide limited visualization functionality. The motivation for this project is to fill this technology gap and create an application that integrates sequence-level comparison as the technique for identifying features and feature-level analysis into a flexible, scalable stand alone system.

1.3 Feature Mosaic Concept

The *feature mosaic* is an abstraction of a genomic sequence into a set of features. This allows the genome to be modeled as a string of computationally defined features that can be related to each other by adjacency and/or orientation. This simplifying abstraction establishes relationships between genomes and defines feature boundaries to allow for more efficient and useful set queries. For example, determining whether a gene exists in one genome but not another is relatively straightforward in a feature mosaic model, but is somewhat ambiguous in simple sequence to sequence comparisons without clear gene boundaries or frames of reference.

The abstraction also enables one to bridge the conceptual leap between annotation within genes and annotation describing larger-scale gene order and multi-genic features. By representing the genome as an abstract string of features, you can perform high level comparisons without working directly with sequences [18]. An example of this is comparing gene order between genomes. This comparison is much clearer with genes represented as an abstraction rather than looking at sequence strings. Once relationships and differences in order are identified, a relational database can easily present the corresponding sequence for additional analysis on demand.

Another key aspect of this model is that it is not limited to a definition of "features" as being coding sequence features. Instead, features are defined in a very generalized way with descriptive information linked to coordinate ranges, than can be applied to any aspect of the genome the user wishes to model. The corresponding string of features that results is a much more informative and useful description of the genome.

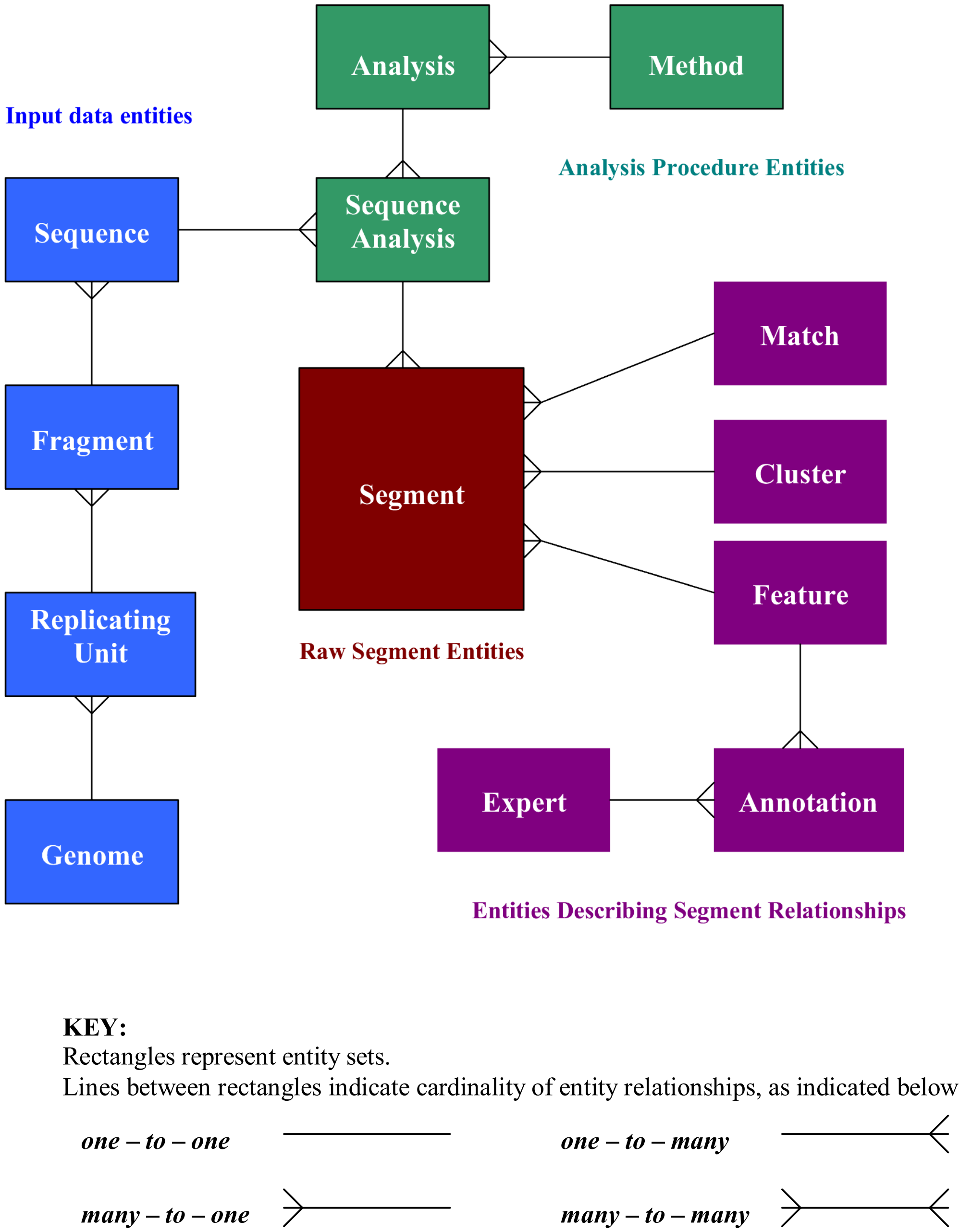
To construct a prototype, three closely related *Brucella* genomes were reduced to a feature mosaic representation by performing analyses using a representative set of comparative and content-based sequence analysis tools. The data model based on this generalized abstraction permits the incorporation of additional analyses, as well as the capability to store the results of operations on features as strings or lists.

2. Materials and Methods

2.1 Data Model Development

2.1.1 Development approach

The prototype of GenoMosaic is driven primarily by Perl scripts that communicate between the database and the sequence analysis applications. The most computationally intensive steps in the analysis pipeline are parsing large output text files that result from various analysis steps. Perl is very efficient at manipulating text and is well suited for this task. There is also a wealth of ready-made scripts available in Perl for manipulating output from standard bioinformatics applications; BioPerl [19] is an open source project that has offered useful sequence manipulation utilities in a Perl module format for several years. These utilities facilitate parsing of BLAST (Basic Local Alignment Search Tool) reports so that they can be entered into a database.



KEY:
Rectangles represent entity sets.
Lines between rectangles indicate cardinality of entity relationships, as indicated below:

Fig. 1: A simplified view of the GenoMosaic data model. Arrows signify the locations of many-to-one relationships. A complete data model can be found in the Appendix (Supplementary Fig. 1)

The open source Relational Database Management System (RDBMS) PostgreSQL was chosen to implement the database [20]. It is an SQL-based system that effectively handles large relational data sets and can communicate with Perl via the Perl DBI, the standard Application Programming Interface (API). This maximizes its capacity to accommodate additional analyses. The GenoMosaic project is designed to ultimately be an open source utility, which will benefit from the input and collaboration of disparate users.

2.1.2 Analysis types

We identified four generic types of sequence analysis that are used in genome annotation and comparison, and selected representative analyses for each of these types. The basic types of analysis are:

Sequence content analyses. These analyses define segments on a single sequence based on its nucleotide content. This includes *ab initio* gene prediction programs, ORF finders, tRNA finders, etc. Simple parsing of existing annotation is included under this heading, although no feature prediction analysis takes place. Since multiple conflicting annotations may exist for a single sequence, the database is also designed to allow the incorporation of information from more than one annotation by considering each annotation a separate instance of "analysis."

Pairwise matching. These analyses define segments based on pairwise matches to other segments. This includes pairwise alignment comparison using the local alignment tool BLAST [21] to define segments based on homology to another segment.

Cluster analysis. These analyses define groups (or clusters) of segments mutually related to another object. This can include mutual similarity to externally defined objects (e.g. COG cluster or Pfam protein family), or to other segments in the same or another genome, or by a shared relationship such as common experimental origin. Cluster relationships are helpful in defining repeating features and are also valuable in identification of important non-coding features [22].

Evidence weighting or joining operations. These analyses examine multiple segments to define features. This includes automated methods for weighting different sources of evidence for a feature definition, and applying a confidence score. The most common use of these methods is to arrive at a consensus for a gene location when there are several conflicting predictions. User-entered evidence assessment and manual annotation also falls into this category, allowing the database to accept comments and annotation from several users.

Representative analyses from each of these categories were chosen to facilitate data modeling and creation of an Entity-Relationship Diagram (ERD) of the database structure.

2.1.3 ERD development

An entity relationship diagram (ERD) was produced as part of a data modeling process to assist in designing the database. Test runs of representative sequence analysis programs were performed and their output compiled to visualize and define the data entities to be stored.

Sequence content analyses was represented by a Glimmer run. Glimmer is an ab-initio gene prediction program developed by TIGR [23]. *Pairwise matching* was represented by a series of BLAST runs, which performed pairwise comparisons on both nucleotides and 6-frame translations of nucleotide sequence. *Cluster analyses* was represented by comparison of translated nucleotide sequence to the COG database [24]. *Evidence weighting* was represented by a rough method of evidence scoring in which one point was allotted for a pairwise segment match, two points if confirmed by a Glimmer prediction, and three points if the match also matched a COG cluster. This schema is only an example, and the prototype is scalable to allow more sophisticated systems.

The entity-relationship diagram was created using Allfusion ERwin Data Modeler [25]. Meaningful intuitive labels were assigned to each data element (e.g. *coord_start* is the starting coordinate for a segment), and appropriate relationships were defined (e.g. each single *analysis* can produce many *segments*). Fig. 1 shows a simplified version of the data model, while a complete data model can be found in the appendix (Supplementary Fig. 1). The GenoMosaic database structure is also in the appendix (Supplementary Fig. 2).

2.1.4 Entity definitions

The GenoMosaic data model is designed to incorporate complete results and analysis parameters for a comprehensive genome comparison. Fig. 1 shows the key entities designed to hold these data and the relationships between them. There are four basic groups of entities:

Input data entities. The basic unit of input for GenoMosaic is a genomic sequence. Since comparisons are sometimes done on incomplete genomes, the data model is designed so that one or more fragments from a genome can be entered. Thus each genome consists of replicating units (chromosomes or plasmids), and each replicating unit consists of *n* fragments, where *n* = 1 if the genome is complete or only one fragment is entered. For each species in a comparison, several replicating units may be entered. One important attribute that is not incorporated into the data model is sequencing confidence scores. These scores are commonly associated with each base in a sequence, and allow one to infer whether sequence level differences are the result of biology or sequencing error. Although they are commonly used, confidence scores are not usually included with publicly available GenBank data sets. They are not included in the feature definitions, but GenoMosaic does support the addition of this element if required.

Analysis procedure entities. GenoMosaic supports multiple sets of analyses performed on the same sequence, which can lead to differing predictions of the same feature. Evidence weighting presents the user with a ranking of possible predictions. For the user to make an informed choice among predictions, or to assess the validity of a feature prediction, search parameters used in the analysis need to be presented. However, adding fields for storing a complete set of analysis attributes would require some compromise of database efficiency and lead to a less streamlined table structure. To avoid this issue, analysis parameters are stored as a single parseable string, so that descriptions of analysis parameters can be stored in the same table structure with the results. Each analysis type has a standard format for this string, and the type of analysis determines how this string is handled.

Raw segment entities. Segment entities are the basic unit of information in the feature mosaic. A segment is a sequence range accompanied by orientation and location information, analysis parameters, and any results or score information from an analysis. Each segment entity is a discrete unit, and features can be composed of more than one segment entity. A key aspect of this entity is the generic *segment_score_string* field, which allows descriptive information from many different segment types to be stored as a parseable string in the same field. This allows a streamlined design and permits the generalized feature abstraction.

Entities describing segment relationships. These more complex entities are of several types: match, cluster, and composite entities. These entities describe relationships between individual or between groups of segments. Attributes of match and composite entities include the unique identifiers of all segments in the relationship, and score strings for the analysis that defines this relationship. These records will normally be generated as part of the analysis process, whereas composite entities will be generated later, defined by secondary analyses and evidence weighting of primary analysis results by user-defined rules. The collection of features resulting from the analysis represents the final step of the automated pipeline. Once this is accomplished, each genome can be represented as a series of segments, and can be presented for any arbitrary sequence range along with significance values and adjacency information. Manual annotation can be added to features defined by the automated analysis or added in *de novo*.

2.2 Sample Comparison – *Brucella* spp.

2.2.1 Genome sequence data and annotation

The complete, annotated sequences of *B. suis* and *B. melitensis* are available in GenBank (AE014291/AE014292 and AE008917/AE008918 respectively). *Brucella melitensis* has been annotated [4] using the ERGO bioinformatics suite and deposited in GenBank. The genome of *B. suis* was sequenced at TIGR and annotated using their standard procedures [5].

The published annotations of *B. melitensis* and *B. suis* were used in protein-to-protein comparisons based on known and predicted CDS, using protein-to-six frame translated nucleotide comparisons to the complete genome sequence to provide cross-validation. No complete annotation has been published for *B. abortus.* Draft *B. abortus* sequence and preliminary annotations (S. Halling, USDA, personal communication) were used to represent *B. abortus* in the comparison.

2.2.2 Standardization of data

Prior to sequence analysis and comparison, the genomic coordinates of *B. melitensis* and *B. abortus* were transformed to correspond to the coordinates of the *B. suis* genome. The location of bp #1 on each chromosome is specified as several hundred bases upstream of a replication initiation protein; 783 bases upstream of *dnaA* on Chromosome I and 154 bases upstream of *repC* on Chromosome II. This convention was not followed in the published *B. melitensis* genome, so coordinates were adjusted to provide a common frame of reference. Coordinates were shifted uniformly by the following values: *B. melitensis* Chromosome I, 2,003,350; Chromosome II, 92,117; *B. abortus* Chromosome I, 1,578,638; Chromosome II, 266,647.

2.2.3 Nucleotide composition

The percentage of guanine and cytosine nucleotides (G+C ratio) within a bacterial genome is consistent within species, and can be used as an indicator of gene origin [26]. The G+C ratio of differentiating regions was therefore calculated to examine whether they may have been acquired by horizontal transfer. The calculation was done using a locally developed Perl script. G+C content within segments at least 50 kb in length that varied by +/- 4.0% from the normal *Brucella* ratio of 57.2% was considered atypical.

2.2.4 Whole genome sequence comparison

Pairwise whole genome alignments for each combination of genomes were performed using MUMmer (v. 2.1) [13]. MUMmer finds all maximal unique matches (MUMs) between two input sequences. This analysis facilitates identification of regions of non-identity and single nucleotide polymorphisms between pairs of genomes with high sequence similarity. Although this analysis is useful, finding locations of SNPs and differentiating regions from MUMmer output is not straightforward. It is also not evident from the output which SNPs are within important coding regions. This process is amenable to automation by Perl scripts and incorporation into the GenoMosaic data model. SNPs and differential regions can themselves be defined as “features,” allowing one to identify them quickly and relate them back to other important features.

2.2.5 Sequence similarity comparison

Sequence based local alignments were performed using standalone BLAST [21]. A consensus of two BLAST programs (tblastx, blastn) was used to define regions of sequence match between genomes. BLAST was run pairwise with an e-value cutoff of 0.005 for each algorithm, and genes for which there were no hits either in coding sequence (CDS) or genomic DNA were considered absent. In addition, a post-BLAST cutoff was applied, in which hits of less than 60% identity were considered non-matches. High-Scoring Pairs (HSPs) covering less than 40% of the query sequence length were also considered non-matches. Differentiating genes were defined as genes that had no matches found by both tblastx and blastn by the cutoffs described above. Gene pairs identified as matches but having less than 95% sequence identity or < 80% full-length coverage were examined more closely and classified as secondary discriminating features. Since comparisons were done by local alignments, full-length matches of < 90% were not detected. Partial coverage HSPs for differentials were examined to determine if they could be combined to make a single match to meet our coverage cutoffs.

2.2.6 Experimental design for RT-PCR of differentiating regions

PCR and RT-PCR experiments were performed for all of the predicted ORFs from the differentiating islands of *B. suis, B. melitensis* and *B. abortus* to determine whether they are present in the genome as predicted, and whether they are transcribed. In addition to these differentiating ORFs, an approximately 6000 bp partial differential ORF from the *B. abortus* sequence was included in the RT-PCR experiment. This region contains about 1800 bp contiguous sequence unique to *B. abortus.* Two different primer pairs were designed for this ORF, with the first primer pair located inside the unique sequence. The second primer was designed to cross the *B. abortus* unique segment of the ORF, with the primers annealing to the parts of the ORFs common in all three species and predicted to produce specific fragments of different lengths.

When designing suitable primer pairs we tried to accommodate the maximum number of specific *B. suis* primers produced by TIGR for a genomic *Brucellae* microarray experiment, and designed our own primer pairs only for the ORF regions that the TIGR set did not cover. In some cases, mixed primer pairs included primers from both TIGR's set and our own. There were 105 primer pairs used to perform the RT-PCR reactions. This included eighty-one primers designed and synthesized by TIGR (I. Paulsen, personal communication) for a *B. suis* cDNA microarray, twenty-two primer pairs designed and synthesized by TIGR specifically for a *B. melitensis* miniarray experiment, and at Virginia Tech (VT) eighty-five primers designed using the Primer3 software [27] with a melting temperature of 60ºC, G+C content of 50% and primer length of close to 20 bp using default values for the rest of the parameters. Later, VT designed primers were checked using Nucleic Acid Quikfold (Mfold version 3.1 and the SantaLucia free energy parameters for DNA) to have the Tm of secondary structure formation less than 40ºC, and the 2-State Hybridization Server for DNA-DNA-hybrid formation [28-30].

2.2.7 PCR and RT-PCR protocols

B. suis, *B. melitensis* and *B. abortus* cultures were grown at 37°C for 36 hours in trypticase soy broth (Difco) and harvested at an OD_{550} = 0.8. The culture was quickly harvested by centrifugation and re-suspended in TE/Citrate/zwittergent 3-14/lysozyme lysing buffer [31]. RNA was extracted using an RNA extraction kit (Quiagen). Purity of the RNA was verified by spectroscopic analysis. Residual genomic DNA contamination was eliminated by treatment with five units of DNAse1(TaKaRa) for one hour at room temperature.

Reverse transcription was carried out using the Superscript first-strand synthesis system for RT-PCR (Invitrogen) following prescribed protocols. The cDNA from each *Brucella* species was used in a PCR reaction as the template with primers specific for each differentiating gene. Ready-to-go PCR beads (puRETaq, Amersham Biosciences) were used according to manufacturer's recommendations. Thermocycling was carried out in the gradient Mastercycler (Eppendorf). Cycling conditions were 90ºC for 5 minutes, 90ºC for 1 min of denaturation, 55ºC for 30 seconds annealing, 72ºC for 1 min extension for 45 cycles and 70ºC for 5 minutes of final extension. The RT-PCR products were electrophoretically separated by 1.5 % (TAE/TBE) agarose gels. Those primers that did not yield expected results were used to repeat the RT-PCR reactions. Those that were suspected of producing nonspecific bands were run at 57ºC annealing temperature. When the expected products were longer than 1 kb an increased extension time of 3 minutes was used in the second round of PCR reactions keeping all other conditions the same.

Sixty out of 111 primer pairs, which produced no amplicon for *B. suis*, *B. melitensis* and *B. abortus* in the reverse transcriptase reactions, were tested on the genomic DNA extracted from each of the three *Brucella* species. The genomic DNA for the PCR reactions was extracted by a phenol/chloroform protocol. The PCR reactions were performed simultaneously for all three *Brucella* species. The reactions were carried out in a final volume of 30 µl. Sterile water (26 µl) was added to the Amersham Biosciences puReTaq Ready-To-Go-PCR bead (each bead contains 2.5 units of PuReTaq DNA Polymerase) to give: 1.5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris-HCl, and 200 µM of each dNTP. The primer and genomic DNA concentrations were 10pmol and 50ng respectively. The DNA underwent denaturation for 5 min. at 95ºC, followed by 40 cycles of 1 min. of denaturation at 95 ºC, 1 min. for primer annealing at 55ºC and 3 min. extension time at 72ºC, and 72ºC for 10 min. of final extension. The PCR products were analyzed by 1% TBE agarose gel electrophoresis.

2.3 Detailed Analysis Methods

2.3.1 Pairwise comparison of genome fragments

For genome sequences to be accurately compared to one another, they need to be aligned with each other to compare the order of individual nucleotides. Alignment algorithms incorporate scoring matrices to calculate alignment quality scores, taking into account gaps in the sequence and nucleotide substitutions, and produce a text visualization of the optimal alignment.

Pairwise sequence comparison provides a common frame of reference between genomes, so that known information can be shared between them. When comparing an unknown sequence to a closely related and well annotated sequence, pairwise alignment can be used to identify probable locations of genes and make inferences about homology. Structural features of the genome independent of ORF locations can be determined by pairwise comparison of fragments to whole genomes. This can detect recurring patterns in non-coding regions of the genome and help identify regulatory regions. This analysis could also reveal patterns of genome rearrangement or gene duplication, which could be relevant to phylogeny. Inferences of this type would not be possible with a global alignment of two entire genome sequences.

BLAST is an algorithm that produces local alignments between a query sequence and a database of reference sequences. It searches for regions of local similarity between two sequences rather than optimal global alignments of whole sequences. The BLAST algorithm expedites local sequence alignment by breaking up the sequence into small "words," of 11 nucleotides, and first finding the occurrence of word matches. Word matches are extended into longer alignments without forming gaps, until the total alignment scores drop below a certain threshold. The top scoring alignments are then combined to form possible alignments covering the length of the total query sequence. The standard implementation of BLAST is maintained by NCBI and accessible through a Web interface or one can use standalone binaries.

BLAST output consists of a text file that can easily be processed using Perl scripts. Only selected data elements from the BLAST output was to be used, so a Perl script was used to parse it, fields for it will be added to the database, and the results imported.

2.3.1 Comparisons to Clusters of Orthologous Groups (COGs)

Putative genes of unknown function can be compared with a database of groups of proteins that share similar functions, to infer the functional group to which the unknown belongs. Proteins assigned to a group that have similar function across multiple species are considered orthologs. The COG database contains clusters of gene families that were determined by comparing protein sequences from 43 complete genomes representing 30 major phylogenetic lineages. Each COG represents at least three of these genomes and corresponds to a phylogenetically ancient conserved domain.

Orthologs are direct evolutionary counterparts related by vertical descent as opposed to paralogs which are genes within the same genome related by duplication [24]. With some exceptions, orthologous proteins typically have the same domain architecture and the same function.

The Clusters of Orthologous Groups of proteins (COGs) database has been designed as an attempt to classify proteins from completely sequenced genomes on the basis of the orthology concept. The COGs reflect one-to-many and many-to-many orthologous relationships as well as simple one-to-one relationships [24].

Coverage includes 56-83% of the gene products from each of the complete bacterial and archaeal genomes. The COG database can be searched several ways including by phylogenetic pattern and functional category using a search interface available from NCBI.

Impetus for development of this tool came from the fact that the pace of assigning functional properties to newly identified sequences has been slow. Analysis of complete microbial genomes has shown that prokaryotic proteins are in general highly conserved, with ~70% of them containing ancient conserved regions (ACRs) [24].

Fitting proteins into a COG is done using the COGNITOR program. This can be used through an NCBI Web interface, or in a standalone version. ORFs identified from previous analyses that have an undetermined function use this program to place them into broad functional categories. This category assignment is then linked to the sequence and stored in the database.

2.3.2 Computational gene finding

Several automated systems for gene finding exist that use a variety of predictive methods to determine the location of genes in the genome. Content-based methods infer gene location based on trends of nucleotide content within coding sequences compared with non-coding sequences. Pattern-recognition methods determine genes by the presence of characteristic sequence patterns such as start/stop codons. Some gene finding packages integrate both of these strategies, and focus differently on eukaryotic or prokaryotic genomes.

Glimmer (Gene Locator and Interpolated Markov Modeler) is maintained by the Institute for Genomic Research [23]. It is effective at finding genes in microbial genomes, and designed for use particularly with bacteria and archaea.

Glimmer uses a combination of interpolated Markov models to distinguish coding sequences from noncoding DNA. The first step it takes is to “train” the model to identify coding regions based on initial complete genes. For an unannotated genome, putative genes identified by strong homology can be used in this initial training set.

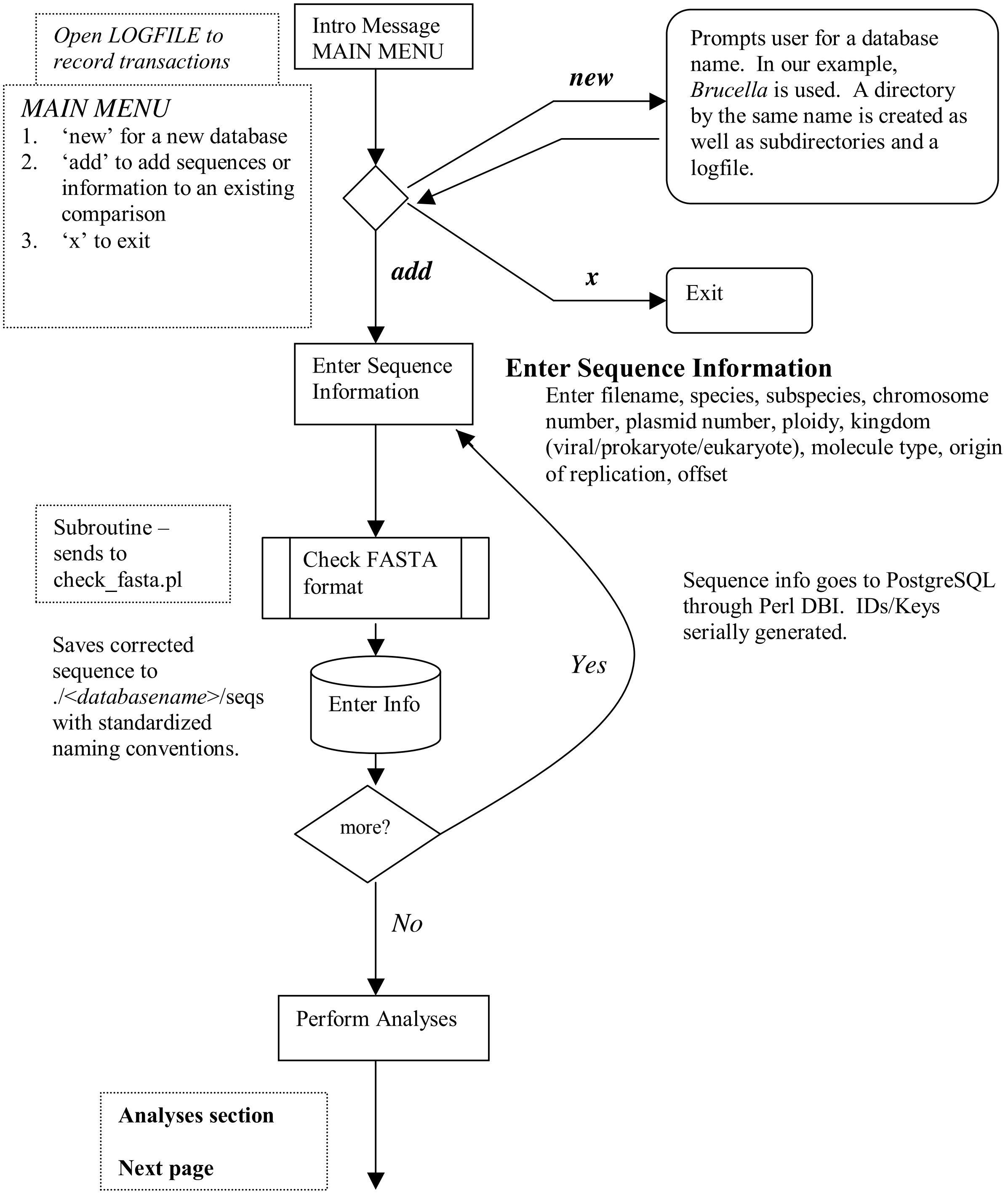

Open LOGFILE to record transactions
Intro Message MAIN MENU
MAIN MENU
1. 'new' for a new database
2. 'add' to add sequences or information to an existing comparison
3. 'x' to exit
new
Prompts user for a database name. In our example, Brucella is used. A directory by the same name is created as well as subdirectories and a logfile.
add
x
Exit
Enter Sequence Information
Enter Sequence Information
Enter filename, species, subspecies, chromosome number, plasmid number, ploidy, kingdom (viral/prokaryote/eukaryote), molecule type, origin of replication, offset
Subroutine – sends to check_fasta.pl
Check FASTA format
Sequence info goes to PostgreSQL through Perl DBI. IDs/Keys serially generated.
Saves corrected sequence to ./<databasename>/seqs with standardized naming conventions.
Enter Info
Yes
more?
No
Perform Analyses
Analyses section
Next page

Figure 2A: Flowchart of GenoMosaic processes – instantiation of input data entities

Analysis Section

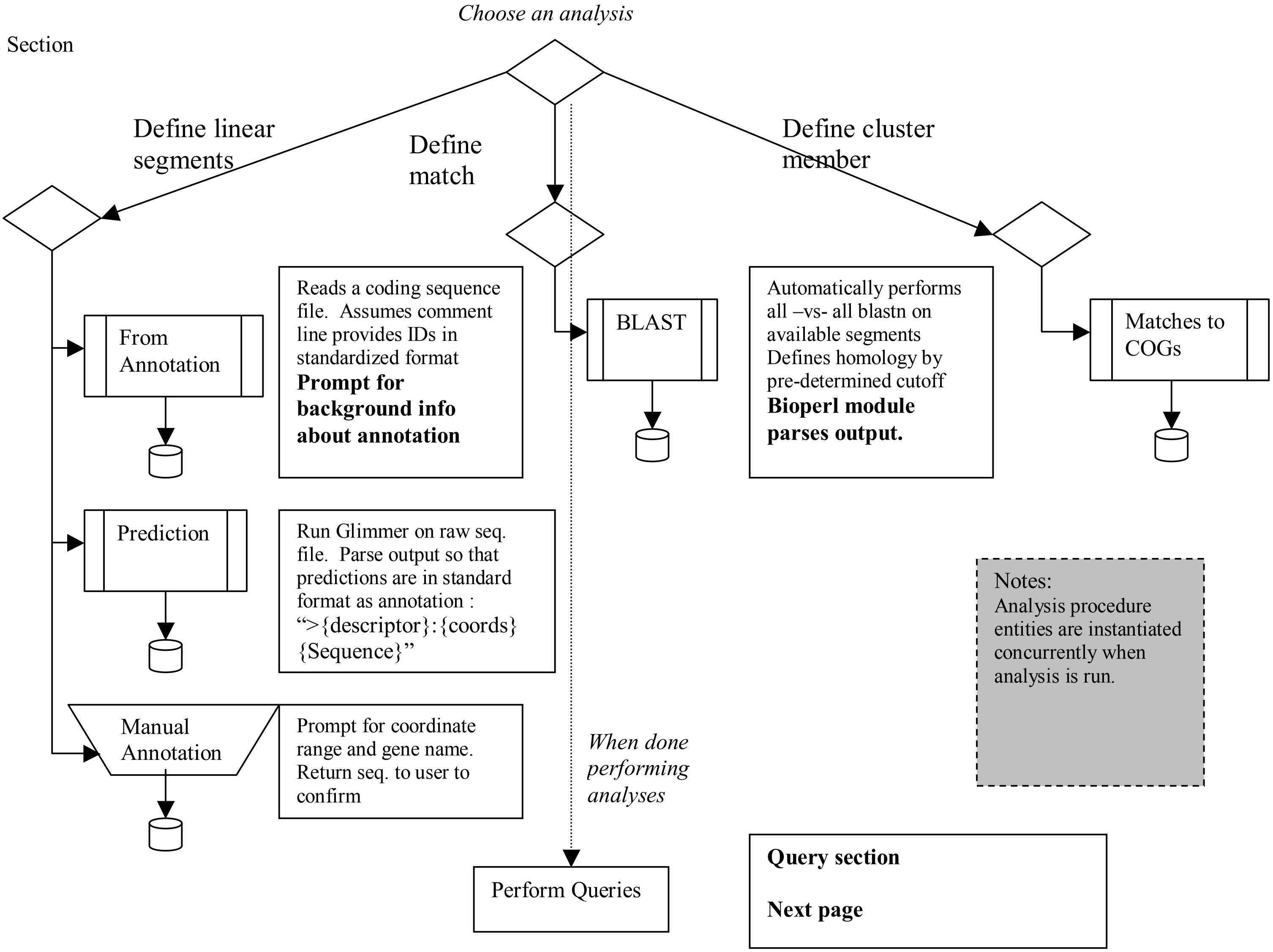

Choose an analysis
Define linear segments
Define match
Define cluster member
From Annotation
Reads a coding sequence file. Assumes comment line provides IDs in standardized format **Prompt for background info about annotation**
Prediction
Run Glimmer on raw seq. file. Parse output so that predictions are in standard format as annotation : “>{descriptor}:{coords} {Sequence}”
Manual Annotation
Prompt for coordinate range and gene name. Return seq. to user to confirm
BLAST
Automatically performs all –vs- all blastn on available segments Defines homology by pre-determined cutoff **Bioperl module parses output.**
Matches to COGs
Notes: Analysis procedure entities are instantiated concurrently when analysis is run.
When done performing analyses
Perform Queries
Query section
Next page

Figure 2B: Flowchart of GenoMosaic processes - analysis steps

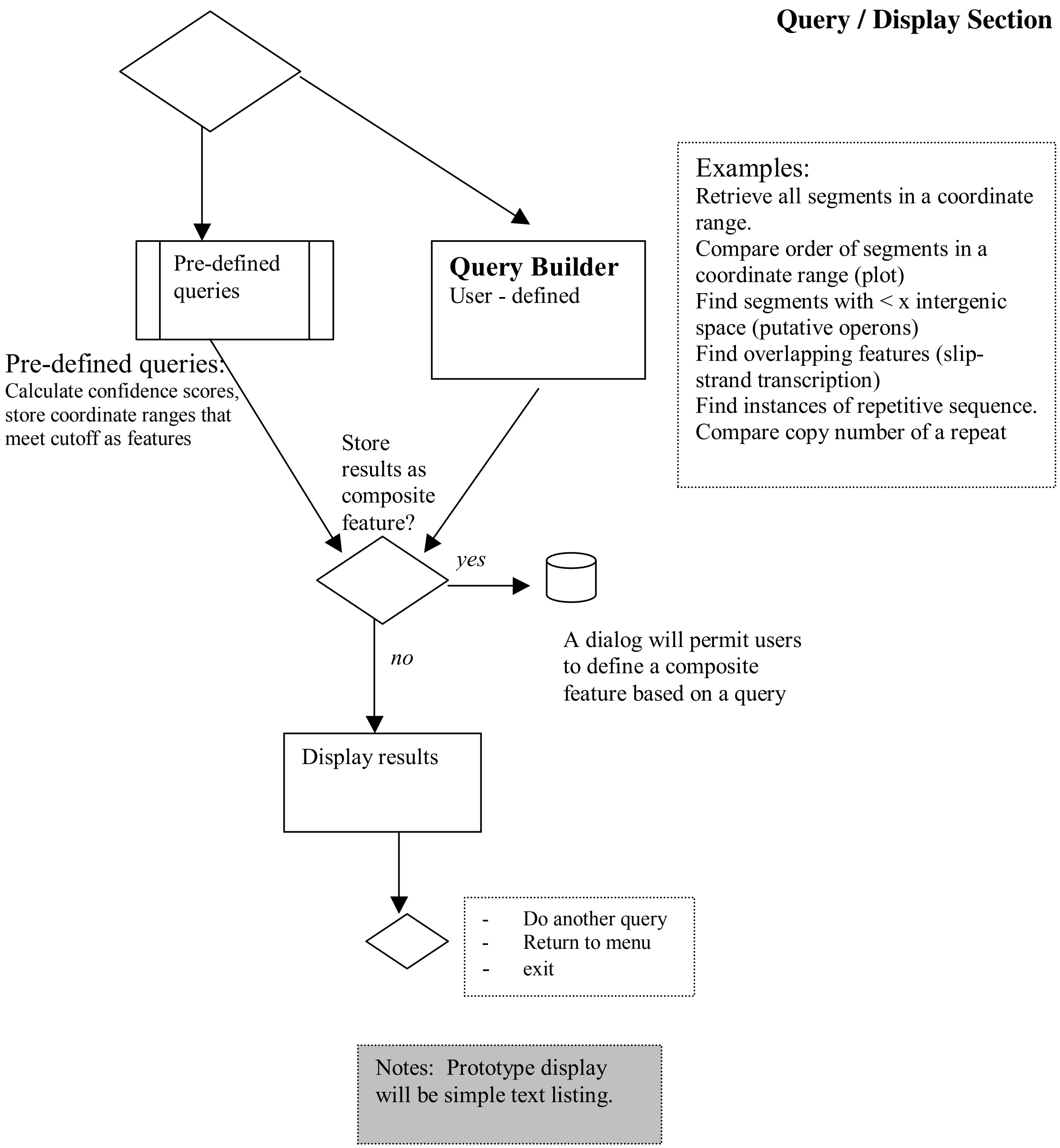

Query / Display Section
Pre-defined queries
Query Builder
User - defined
Examples:
Retrieve all segments in a coordinate range.
Compare order of segments in a coordinate range (plot)
Find segments with < x intergenic space (putative operons)
Find overlapping features (slip-strand transcription)
Find instances of repetitive sequence.
Compare copy number of a repeat
Pre-defined queries:
Calculate confidence scores, store coordinate ranges that meet cutoff as features
Store results as composite feature?
yes
no
A dialog will permit users to define a composite feature based on a query
Display results
- Do another query
- Return to menu
- exit
Notes: Prototype display will be simple text listing.

Fig. 2C: Flowchart of GenoMosaic processes - queries and composite feature definition

3. Results and Discussion

3.1 GenoMosaic Design

Representative sequence analyses were performed on *B. abortus*, *B. melitensis*, and *B. suis* to obtain a sample data set for data model development. Fig. 2 provides flowcharts of the process.

The creation and instantiation of the GenoMosaic database is done in a series of steps. Since the time it takes to complete some analyses is very long, it is not practical to perform this in one session. After some optimization, it is possible to automate all the steps described here to be performed in one session.

The basic steps to be performed are:

1. Creation of the database
The user is prompted to enter a database name. In our sample comparison, *Brucella* is used as the database name. The script will create a working directory of the same name and create the postgreSQL database structure.

2. Enter input data entities (sequences)
The user is prompted for the file names of sequences to enter, along with identifying information about the sequence. This information includes GenBank indentifiers, sequence length, chromosome number, etc.

3. Enter raw segment entities
In this step, any available annotations are parsed and entered into the database as segments. The script is designed to parse annotation files in standard GenBank format, in which a header line takes the format ">{descriptor}:{start}-{end} {additional descriptors}," with the sequence following on the next line. In addition to header line descriptors, the script will also parse a GenBank protein coding genes table (appendix), and associate in the database all available information with their respective segments.

In the GenoMosaic data model, annotations are considered a form of 'analysis.' As such, analysis procedure entities must be instantiated along with the annotation. These entities include information such as a reference and author for the annotation.

4. Perform analyses

The next step is to perform several *de novo* analyses to define new segments. As analyses are performed and their results entered into the database, their corresponding analysis procedure entities are also entered.

First, gene prediction analysis is performed on genomic sequences using Glimmer. The predictions generated by Glimmer are then put into standard GenBank format, and are parsed the same way as annotations.

The next set of analyses is performed to define entities describing segment relationships. These analyses are more complex than gene prediction with Glimmer, and require several steps to complete.

BLAST runs are performed to establish matches between segments. To perform automatic all – against – all BLAST runs, GenoMosaic writes all segment sequences for each unique *sequence_analysis_id* to files in a working directory, and calls *formatdb* and *blastall* commands on them. Each segment sequence contains its unique segment identifier, so that all matches can be associated with their respective segments. BLAST runs are performed using pre-determined specifications, which may be changed by the user. The Bioperl module Search:IO is used to parse the output into a format that can be entered into the database.

To define cluster entities, a local version of COGNITOR is run to compare individual segments against a database of COGs. To maximize search efficiency, this search is only conducted against prokaryotic COGs, but the script can easily be expanded to include other COG databases. The resulting output is parsed so that each COG that matches at least one segment is stored in the *cluster* table, linked with each segment ID that matches it.

5. Perform queries / composite feature definition

Now that the database is fairly complete it can be queried. GenoMosaic includes several pre-formed queries, such as presenting to the user all segments defined in a given coordinate range and finding unique or differentiating features. The process of defining composite features is much the same as performing a query. This is implemented as an SQL query that calculates confidence scores for a given coordinate range, and stores this information in the *feature_score_string* field of the *feature* table.

3.2 *Brucella* Comparison

3.2.1 Genome size and composition

Brucella melitensis, *B. suis*, and *B. abortus* each have approximately 3 million base pairs (Mb) of genomic DNA. In each species, the nucleotides are distributed over a larger chromosome of about 2 Mb and a smaller one of about 1 Mb. *Brucella suis* has a slightly larger C-value (total amount of genomic DNA) than the other two genomes, with a slightly smaller Chromosome I and slightly larger Chromosome II (Table 1). Other biovars of this species are variable in chromosome size and number [5]. Biovars 2 and 4 possess two chromosomes of 1.85 Mb and 1.35 Mb, while biovar 3 contains only one 3.1 Mb chromosome [32]. G+C content is equal in the three species at 57%.

Table 1: General features of the three *Brucella* genomes.

	Ch. I	Ch. II	Total	G+C
	(Mb)	(Mb)	(Mb)	%
B. abortus	2.13	1.16	3.29	57.3
B. melitensis	2.12	1.18	3.29	57.2
B. suis	2.11	1.21	3.32	57.2

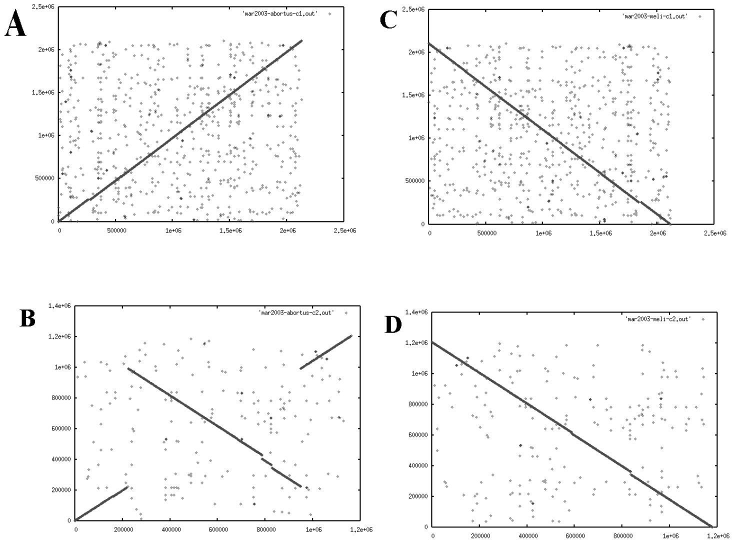


Fig. 3: Global alignment of *B. abortus* and *B. melitensis* genomes relative to *B. suis*. Percent identity plots of (A) *B. abortus* Chromosome I, (B) *B. abortus* Chromosome II, (C) *B. melitensis* Chromosome I, and (D) *B. melitensis* Chromosome II vs. *B. suis*. Alignment shows consistent colinearity between the genomes, the only exception being a large inversion in *B. abortus* Chromosome II starting at roughly bp #200,000. Prepared with Mummer v. 2.1 [13].

Table 2: Locations of *Brucella* differentiating islands

Island	Species	Chr.	Coordinates (bp)	G+C	Size (kb)	Genes
S1	*B. suis*	I	924993-926746	55.8%	1,753	BR0952-BR0954
S2	*B. suis*	II	343695-361343	55.6%	17,648	BRA0362-BRA0379
MA1	*B. melitensis*	I	1724077-1743975	52.3%	19,898	BMEI1674-BMEI1702
	B. abortus	I	305475-286341	52.2%	19,134	No annotation
SA1	*B. suis*	I	581316-584877	59.2%	3,561	BR0588-BR0593
	B. abortus	I	632376-635569	59.2%	3,193	No annotation
SA2	*B. suis*	II	610688-619976	56.9%	9,288	BRA0630-BRA0636
	B. abortus	II	871798-865719	56.9%	6,079	No annotation
MS2	*B. melitensis*	II	860832-885154	58.1%	24,322	BMEII0827-BMEII0848
	B. suis	II	402846-428212	58.2%	25,366	BRA0418-BRA0439

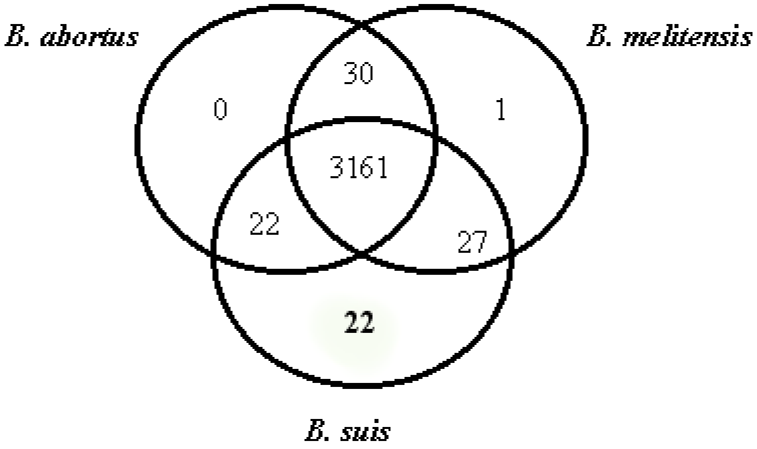


Fig. 4: *Brucella* gene content comparison by a Venn diagram.

Although there are no naturally occurring plasmids in any of the *Brucellae*, Chromosome II of *B. suis* contains plasmid-like replication genes, which is consistent with the theory that this 2nd chromosome was derived from a megaplasmid captured by an ancestral organism [5]. These replication genes were shown to differentiate *B. suis* from *B. melitensis* [5], and our analysis shows that they are also unique with respect to *B. abortus*.

3.2.2 Genome organization

As expected, a whole genome alignment displayed extensive synteny among the *Brucella* species (Fig. 3). The only exception is a large inversion found in *B. abortus*, beginning at approximately bp #200,000. This inversion corresponds to the 640 kb inversion identified by restriction mapping [33]. This inversion is not consistently characteristic of the species, as it is present in *B. abortus* biovars 2, 3, and 4 but not biovars 5, 6, and 9 [33]. The origin of this inversion is not known, but recombination was ruled out at *rrn* loci or insertion sequences, due to the lack of these sequences at the borders [33]. Significant gaps in the whole-genome alignments were found to correspond to the locations of differentiating islands identified through sequence similarity comparison (Table 2).

3.2.3 Gene content comparison

Reinforcing the conclusion that the three genomes are highly similar, the majority (>90%) of annotated genes were found to share 98-100% sequence identity with their apparent homologues in the other genomes. The differences in gene content are illustrated by a Venn diagram (Fig. 4). *Brucella suis* contains twenty-two genes which distinguish it from *B. melitensis* and *B. abortus*. *B. melitensis* contains one gene unique within the comparison and *B. abortus* contains none. Each species contains genes shared with a second species that distinguish both from the third species. The majority of differentiating genes are in large (~20 kb) islands, which partly account for differences in chromosome size. Most of these genes have functional assignments in existing annotation (Table 3). Locations of these genes in the genome are indicated in Fig. 5.

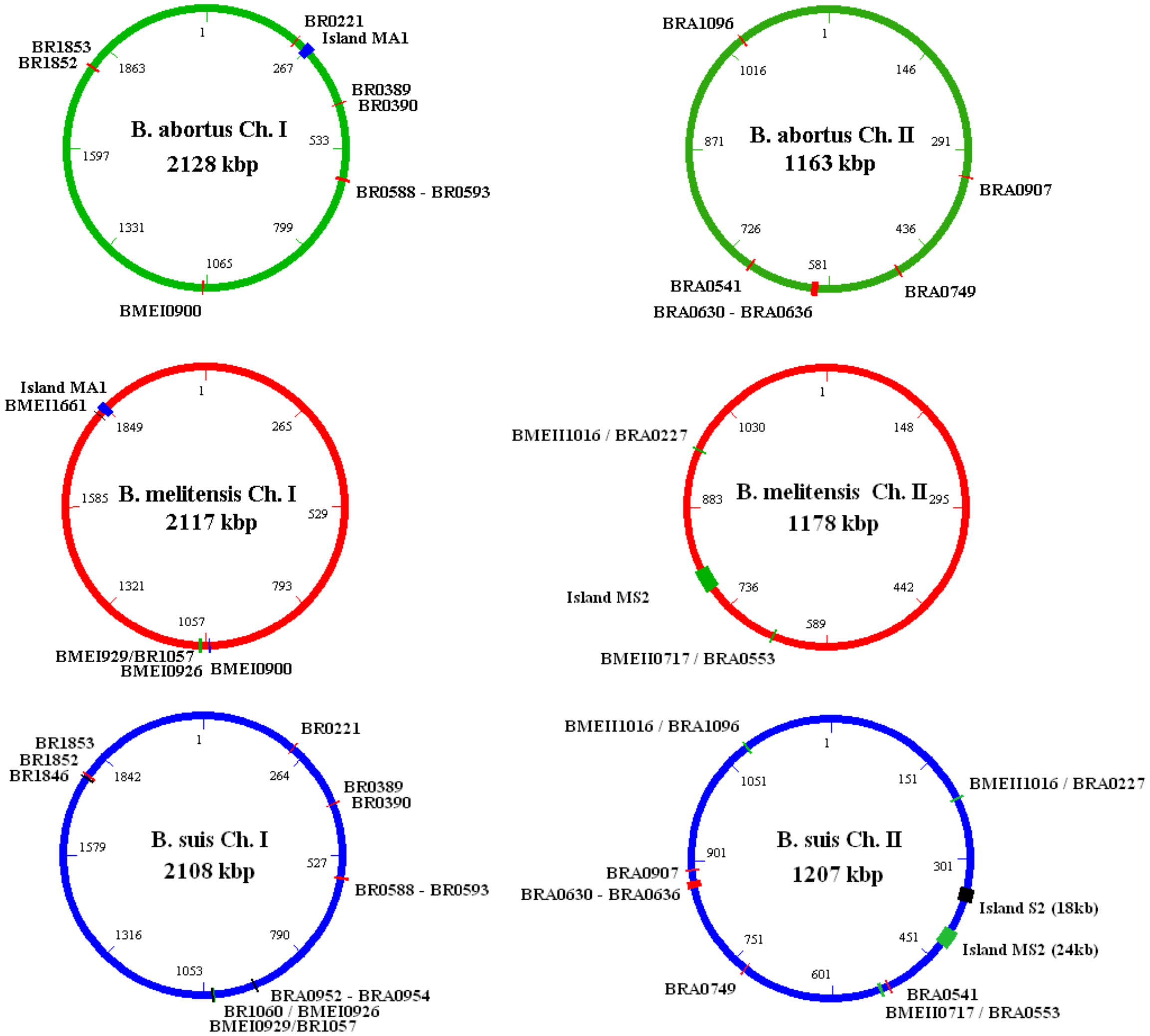


Fig. 5: Locations of *Brucella* gene differences. Gene color within chromosomes indicates the species it is absent in. Genes in black are unique. The majority of genes lie with one of several large islands. A complete list of gene differentials is given in Table 4.

Table 3: Detailed list of *Brucella* gene differentials. Coordinate ranges given reflect corrected (recalibrated) values. Differentials are listed in order of *B. suis* annotation (*B. melitensis* for genes absent in *B. suis*). Fig. 5 provides a visualization of the locations of these differentials within the genome.

	coordinates		coordinates		coordinates		
Chr	(*B. suis*)		(*B. melitensis*)		(*B. abortus*)		Gene Name
	start	*end*	*start*	*end*	*start*	*end*	
1	232957	233406			234317	234766	BR0221 transcriptional regulator, MerR family
1	397660	397932			419522	419794	BR0389 hypothetical protein
1	397922	398029			419784	419891	BR0390 hypothetical protein
1	581316	581984			606497	607165	BR0588 protease, putative
1	582006	583280			607187	608461	BR0589 major capsid protein, HK97 family
1	583445	584011			608626	609192	BR0590 conserved hypothetical protein
1	584008	584346			609189	609527	BR0591 conserved hypothetical protein
1	584343	584510			609524	609690	BR0592 hypothetical protein
1	584470	584877			609650	610057	BR0593 conserved hypothetical protein
1	924993	925826					BR0952 amino acid ABC transporter, permease protein
1	925829	926551					BR0953 amino acid ABC transporter, permease protein
1	926567	926746					BR0954 hypothetical protein
1	1025907	1026701	1071151	1071753			BMEI0926/BR1060 multidrug resistance protein A, HlyD family secretion protein
1	1031242	1031773	1075470	1076237			BMEI0929/BR1057 diguanylate cyclase/phosphodiesterase domain 1 (GGDEF)
1			1047752	1047994	1075458	1075700	BMEI0900 hypothetical protein
1	1777719	1778555					BR1846 hypothetical protein
1	1782718	1784646			1803257	1805185	BR1852 transcriptional regulator, Cro/CI family
1	1784643	1785317			1805182	1805856	BR1853 AzlC family protein
1			1826883	1827629			BMEI1661 recombinase
1			1837872	1838624	279596	280348	BMEI1674 hypothetical protein
1			1838745	1838981	279239	279475	BMEI1675 hypothetical protein
1			1839324	1839932	278288	278896	BMEI1676 hypothetical protein
1			1840203	1840667	277553	278017	BMEI1677 hypothetical protein
1			1840739	1841029	277191	277481	BMEI1678 hypothetical protein
1			1841095	1841340	276880	277125	BMEI1679 hypothetical protein
1			1841418	1841663	276557	276802	BMEI1680 hypothetical protein
1			1841727	1842203	276017	276493	BMEI1681 hypothetical protein
1			1842216	1842685	275535	276004	BMEI1682 hypothetical protein
1			1843070	1843618	274602	275150	BMEI1683 zinc-dependent metallopeptidase
1			1843640	1843804	274416	274580	BMEI1684 hypothetical protein
1			1844148	1844381	273839	274072	BMEI1685 hypothetical protein
1			1844465	1844818	273402	273755	BMEI1686 hypothetical protein
1			1844882	1845088	273132	273338	BMEI1687 hypothetical protein
1			1845085	1845675	272544	273135	BMEI1688 hypothetical protein

1			1845666	1846145	272074	272553	BMEI1689 hypothetical protein
1			1846187	1846492	271727	272032	BMEI1690 hypothetical protein
1			1846693	1848585	270406	271526	BMEI1691 hypothetical membrane spanning protein
1			1848735	1850651	267579	269495	BMEI1692 flagellar protein FlgJ
1			1850852	1851061	267169	267378	BMEI1693 hypothetical protein
1			1851163	1852179	266051	267067	BMEI1694 hypothetical protein
1			1852333	1852950	265281	265897	BMEI1695 hypothetical protein
1			1852920	1854449	263783	265311	BMEI1696 hypothetical membrane spanning protein
1			1854446	1855327	262905	263786	BMEI1697 virulence-associated protein E
1			1855324	1855638	262594	262908	BMEI1698 hypothetical protein
1			1855635	1855844	262388	262597	BMEI1699 hypothetical protein
1			1855845	1856063	262169	262387	BMEI1700 hypothetical protein
1			1856175	1856436	261796	262057	BMEI1701 hypothetical protein
1			1856574	1857770	260462	261658	BMEI1702 transposase
2	214967	215674	962770	963477			BMEII1016/BRA0227 protease I
2	343695	344903					BRA0362 site-specific recombinase, phage integrase family
2	345188	345418					BRA0363 DNA-binding protein, putative
2	345499	346557					BRA0364 RepA-related protein
2	347606	347932					BRA0365 hypothetical protein
2	347935	349578					BRA0366 TrbL protein
2	349581	349763					BRA0367 hypothetical protein
2	349766	350557					BRA0368 TrbJ protein
2	350655	350807					BRA0369 hypothetical protein
2	350825	351049					BRA0370 hypothetical protein
2	351052	351270					BRA0371 TraC protein
2	351940	352320					BRA0372 TraJ protein
2	352317	354269					BRA0373 TraI protein, putative
2	354676	356100					BRA0374 hypothetical protein
2	356254	357279					BRA0375 hypothetical protein
2	357313	358038					BRA0376 hypothetical protein
2	358217	359938					BRA0377 conserved hypothetical protein
2	360180	361004					BRA0378 hypothetical protein
2	361149	361343					BRA0379 DNA-damage-inducible protein J, putative
2	403085	403826	793022	794001			BMEII0849/BRA0418 GDP-4-dehydro-d-rhamnose reductase
2	403810	404880	791968	793038			BMEII0848 GDP-mannose 4,6-dehydratase
2	403810	404880	791968	793038			BRA0419 GDP-mannose 4,6-dehydratase
2	405078	406394	790454	791770			BMEII0847 glycosyltransferase
2	406415	407650	789198	790433			BRA0421 glycosyltransferase, group 1 family protein
2	406427	407650	789198	790421			BMEII0846 glycosyltransferase
2	407647	408843	788077	788828			BRA0422 glycosyltransferase, group 1 family protein
2	408092	408843	788077	789201			BMEII0845 lipopolysaccharide n-acetylglucosaminyltransferase

2	408914	409636	787284	788006			BRA0423 outer membrane protein, 31 kDa
2	408914	409573	787347	788006			BMEII0844 outer-membrane immunogenic protein precursor, 31 kDa
2	410033	410647	786273	786887			BRA0424 acetyltransferase, CysE/LacA/LpxA/NodL family
2	410033	410536	786384	786887			BMEII0843 putative colanic acid biosynthesis acetyltransferase WCAF
2	410659	411921	784999	786261			BMEII0842 hypothetical protein
2	411918	412390	784530	785002			BMEII0841 hypothetical protein
2	411918	412390	784530	785002			BRA0426 Bme2 protein
2	412532	413488	783432	784388			BMEII0840 glycosyltransferase involved in cell wall biogenesis
2	412532	413413	783507	784388			BRA0427 glycosyltransferase, group 2 family protein
2	413410	414537	782389	783510			BRA0428 undecaprenyl-phosphate alpha-n-acetylglucosaminyltransferase, putative
2	413410	414474	782446	783510			BMEII0839 putative undecaprenyl-phosphate alpha-n-acetylglucosaminyltransferase
2	414818	416323	780603	782108			BMEII0838 succinoglycan biosynthesis transport protein exot
2	414830	416323	780603	782096			BRA0429 polysaccharide biosynthesis protein
2	416339	417352	779574	780587			BMEII0837 glycosyltransferase
2	416339	417352	779574	780587			BRA0430 glycosyltransferase, group 2 family protein
2	417308	418549	778377	779618			BMEII0836 dTDP-4-dehydrorhamnose 3,5-epimerase
2	417308	418549	778377	779618			BRA0431 conserved hypothetical protein
2	418666	420045	776881	778260			BRA0432 glycosyltransferase, group 1 family protein
2	419122	420045	776881	777804			BMEII0835 glycosyltransferase
2	420083	421444	775482	776843			BMEII0834 glutamate-1-semialdehyde 2,1-aminomutase
2	420083	421444	775482	776843			BRA0433 glutamate-1-semialdehyde-2,1-aminomutase, putative
2	421423	422757	774169	775503			BMEII0833 hypothetical protein
2	421423	422757	774169	775503			BRA0434 conserved hypothetical protein
2	422872	423939	772987	774054			BMEII0832 UDP-glucose 4-epimerase
2	422878	423939	772987	774048			BRA0435 epimerase/dehydratase family protein, putative
2	423939	425291	771635	772987			BRA0436 conserved hypothetical protein
2	423978	425291	771635	772948			BMEII0831 hypothetical protein
2	425254	425809	771118	771672			BMEII0830 dTDP-4-dehydrorhamnose 3,5-epimerase
2	425254	425778	771149	771672			BRA0437 dTDP-4-dehydrorhamnose 3,5-epimerase
2	426099	427400	769528	770828			BRA0438 methyltransferase, putative
2	426099	426762	770166	770828			BMEII0829 possible S-adenosylmethionine-dependent methyltransferase
2	426834	427400	769528	770094			BMEII0828 possible S-adenosylmethionine-dependent methyltransferase
2	427325	428212	768716	769603			BMEII0827 glucose-1-phosphate cytidylyltransferase
2	427403	428212	768716	769525			BRA0439 nucleotidyltransferase family protein
2	521842	522066			692179	692403	BRA0541 hypothetical protein
2	610688	611938			605152	606402	BRA0630 amino acid dehydrogenase, putative
2	612027	612788			604302	605063	BRA0631 amino acid ABC transporter, periplasmic amino acid-binding protein
2	612944	613717			603373	604146	BRA0632 amino acid ABC transporter, periplasmic amino acid-binding protein
2	613902	615005			602085	603188	BRA0633 conserved hypothetical protein
2	615107	615556			601534	601982	BRA0634 transcriptional regulator, AsnC family
2	615836	617563			599527	601254	BRA0635 twin-arginine translocation signal domain protein

2	617674	618876			598292	599073	BRA0636 beta-ketoadipyl CoA thiolase
2	731323	732192			484958	485827	BRA0749 sugar ABC transporter, permease protein, putative
2	888804	890204			326910	328310	BRA0907 conserved hypothetical protein
2	1082617	1083330			1038414	1039127	BRA1096 transcriptional regulator, putative
2	1618077	1618117	663258	665282			BMEII0717/BRA0553 hemagglutinin, cell wall surface protein, putative

Table 4: Detailed results for RT-PCR analysis of proposed differential ORFs from *Brucella* species

#	**ORF Name**	**Function**		**Amplicon**	***B. suis***		***B. melitensis***		***B. abortus***	
				Size (bp)	**Pred.**	**Obs.**	**Pred.**	**Obs.**	**Pred.**	
	Obs.									
***B. suis* Chromosome I**										
1	BR0952	putative amino acid ABC transporter, permease protein		**396**	+	+	-	-	-	-
2	BR0953	putative amino acid ABC transporter, permease protein		**438**	+	+	-	-	-	-
3	BR0954	hypothetical protein		**153**	+	+	-	-	-	-
4	BR1846	hypothetical protein	*B. suis*:	**722**	+	+	+	+	-	-
			B. melitensis:	**469**						
***B. suis* Chromosome II**										
5	BRA0362	putative site-specific recombinase, phage integrase family		**722**	+	+	-	-	-	-
6	BRA0363	putative DNA-binding protein		**148**	+	+	-	-	-	-
7	BRA0364	putative RepA-related protein		**655**	+	+	-	-	-	-
8	BRA0365	hypothetical protein		**167**	+	+	-	-	-	-
9	BRA0366	putative TrbL protein		**170**	+	+	-	-	-	-
10	BRA0367	putative TrbL protein		**119**	+	+	-	-	-	-
11	BRA0368	putative TrbJ protein		**354**	+	+	-	-	-	-
12	BRA0369	hypothetical protein		**123**	+	+	-	-	-	-
13	BRA0370	hypothetical protein		**121**	+	+	-	-	-	-
14	BRA0371	putative TraC protein		**140**	+	+	-	-	-	-
15	BRA0372	putative TraJ protein		**218**	+	+	-	-	-	-
16	BRA0373	putative TraI protein		**173**	+	-	-	-	-	-
17	BRA0374	hypothetical protein		**768**	+	+	-	-	-	-
18	BRA0375	hypothetical protein		**648**	+	+	-	-	-	-
19	BRA0376	hypothetical protein		**532**	+	+	-	-	-	-
20	BRA0377	conserved hypothetical protein		**867**	+	+	-	-	-	-
21	BRA0378	hypothetical protein		**191**	+	+	-	-	-	-
22	BRA0379	putative DNA-damage-inducible protein J		**119**	+	+	-	-	-	-
***B. melitensis* Chromosome I**										
23	BMEI1661	recombinase		**218**	-	-	+	+	-	-
B. abortus										
24	6 kb Partial differential, primer pair 1			**782**	-	-	-	-	+	+
25	6 kb Partial differential, primer pair 2		*B. abortus:*	**4484**	+	-	+	-	+	-
			B. melitensis:	**1142**						
			B. suis:	**613**						

***B. suis* and *B. melitensis* Chromosome I**									
26	BMEI0926 / BR1060	multidrug resistance protein A / putative HlyD family secretion protein	**207**	-	-	+	+	-	-
27	BMEI0929 / BR1057	putative GGDEF domain protein / Diguanylate cyclase/phosphodiesterase domain	**323**	+	+	+	+	-	-
***B. suis* and *B. melitensis* Chromosome II**									
28	BRA0227 / BMEII1016	putative ThiJ/PfpI family protein / protease I	**466**	+	+	+	-	-	-
29	BRA0418 / BMEII0849	putative fucose synthetase family protein / GDP-4-dehydro-D-rhamnose reductase	**363**	+	+	+	-	-	-
30	BRA0419 / BMEII0848	putative GDP-mannose 4,6-dehydratase Bme9 / GDP-mannose 4,6-dehydratase	**239**	+	-	+	+	-	-
31	BRA0420 / BMEII0847	putative glycosyltransferase / glycosyl transferase	**657**	+	+	+	-	-	-
32	BRA0421 / BMEII0846	putative glycosyltransferase, group 1 family protein / glycosyl transferase	**229**	+	-	+	-	-	-
33	BRA0422 / BMEII0845	putative glycosyltransferase, *B. suis* group 1 family protein /: *B. melitensis*: lipopolysaccharide N-acetylglucosaminyltransferase	**470** **398**	+	+	+	-	-	-
34	BRA0423 / BMEII0844	putative outer membrane protein, 31 kDa / 31 kDa outer-membrane immunogenic protein precursor	**317**	+	+	+	+	-	-
35	BRA0424 / BMEII0843	putative acetyltransferase, CysE/LacA/LpxA/NodL family / putative colanic acid biosynthesis acetyltransferase WCAF	**366**	+	-	+	-	-	-
36	BRA0425 / BMEII0842	putative membrane protein Bme3 / hypothetical protein	**774**	+	-	+	-	-	-
37	BRA0426 / BMEII0841	putative Bme2 protein / hypothetical protein	**286**	+	-	+	+	-	-
38	BRA0427 / BMEII0840	putative glycosyl transferase, group 2 family protein / glycosyltransferase involved in cell wall biogenesis	**279**	+	+	-	-	-	-
39	BRA0428 / BMEII0839	putative undecaprenyl-phosphate alpha-N-acetylglucosaminyltransferase	**672**	+	+	+	-	-	-
40	BRA0429 / BMEII0838	putative polysaccharide biosynthesis protein / succinoglycan biosynthesis transport protein exot	**306**	+	-	+	-	-	-
41	BRA0430 / BMEII0837	putative glycosyltransferase, group 2 family protein / glycosyltransferase	**488**	+	-	+	-	-	-
42	BRA0431 / BMEII0836	conserved hypothetical protein / dTDP-4-dehydrorhamnose 3,5-epimerase	**281**	+	-	+	-	-	-
43	BRA0432 / BMEII0835	putative glycosyltransferase, group 1 family protein / glycosyltransferase	**223 /708**	+	+	+	+	-	-
44	BRA0433 /	putative glutamate-1-semialdehyde-2,1-aminomutase /	**463**	+	+	+	-	-	-

	BMEII0834	glutamate-1-semialdehyde 2,1-aminomutase							
45	BRA0434 / BMEII0833	putative conserved hypothetical protein / hypothetical protein	**239**	+	+	+	-	-	-
46	BRA0435 / BMEII0832	putative epimerase/dehydratase family protein / UDP-glucose 4-epimerase	**642**	+	+	+	-	-	-
47	BRA0436 / BMEII0831	conserved hypothetical protein / hypothetical protein	**188**	+	+	+	+	-	-
48	BRA0437 / BMEII0830	putative dTDP-4-dehydrorhamnose 3,5-epimerase / dTDP -4-dehydrorhamnose 3,5-epimerase dTDP -4-dehydrorhamnose reductase	**285**	+	+	+	-	-	-
49	BRA0438 / BMEII0828	putative methyltransferase / possible s-adenosylmethionine-dependent methyltransferase	**452**	+	+	+	-	-	-
50	BRA0438 / BMEII0829	putative methyltransferase / possible s-adenosylmethionine-dependent methyltransferase	**155**	+	-	+	-	-	-
51	BRA0439 / BMEII0827	putative nucleotidyltransferase family protein / glucose-1-phosphate cytidylyltransferase	**525**	+	+	+	+	-	-
52	BRA0553 / BMEII0717	putative cell wall surface protein / hemagglutinin	**421**	+	+	+	-	-	-
***B. suis* and *B. abortus* Chromosome I**									
53	BR0221 / DI064	putative transcriptional regulator, MerR family	**91**	+	+	-	-	+	+
54	BR0389	hypothetical protein	**141**	+	-	-	-	+	+
55	BR0390 / DI073	hypothetical protein	**74**	+	-	-	-	+	-
56	BR0588	putative protease	**665**	+	-	-	-	+	+
57	BR0589 / DI066	major capsid protein, HK97 family / putative protein	**303**	+	-	-	-	+	+
58	BR0590 / DI067	conserved hypothetical protein	**71**	+	-	-	-	+	+
59	BR0591 / DI068	conserved hypothetical protein	**139**	+	-	-	-	+	-
60	BR0592	hypothetical protein	**91**	+	-	-	-	-	-
61	BR0593 / DI069	conserved hypothetical protein	**208**	+	+	-	-	-	-
62	BR1852 / DI071	transcriptional regulator, Cro/CI family,	**194**	+	-	-	-	+	+
63	BR1853 / DI072	putative AzlC family protein	**610**	+	+	-	-	+	+

***B. suis* and *B. abortus* Chromosome II**									
64	BRA0541 / DII007	hypothetical protein	**118**	+	-	-	-	+	-
65	BRA0630 / DII008	putative amino acid dehydrogenase	**736**	+	+	-	-	+	+
66	BRA0631 / DII001	putative amino acid ABC transporter, periplasmic amino acid-binding protein	**202**	+	-	-	-	+	-
67	BRA0632 / DII002	putative amino acid ABC transporter, periplasmic amino acid-binding protein	**321**	+	+	-	-	+	-
68	BRA0633 / DII003	conserved hypothetical protein	**591**	+	+	-	-	+	+
69	BRA0634 / DII005	putative transcriptional regulator, AsnC family	**276**	+	-	-	-	+	-
70	BRA0635 / DII006	putative twin-arginine translocation signal domain protein	**998**	+	+	-	-	+	+
71	BRA0636 / DII009	putative beta-ketoadipyl CoA thiolase	**635**	+	+	-	-	+	-
72	BRA0749 / DII010	putative sugar ABC transporter, permease protein	**310**	+	+	-	-	+	+
73	BRA0907 / DII011	conserved hypothetical protein	**825**	+	-	-	-	+	+
74	BRA1096 / DII012	putative transcriptional regulator	**393**	+	+	-	-	+	+
***B. melitensis* and *B. abortus* Chromosome I**									
75	BMEI0900	hypothetical protein	**212**	-	-	+	+	-	-
76	BMEI1674 / DI002	hypothetical protein	**597**	-	-	+	+	+	+
77	BMEI1675	hypothetical protein	**157**	-	-	+	+	+	+
78	BMEI1676 / DI006	hypothetical protein	**206**	-	-	+	-	+	+
79	BMEI1977 / DI008	hypothetical protein	**400**	-	-	+	+	+	+
80	BMEI1978 / DI010	hypothetical protein	**192**	-	-	+	+	+	+
81	BMEI1979	hypothetical protein	**201**	-	-	+	+	+	+
82	BMEI1980	hypothetical protein	**210**	-	-	+	+	+	+
83	BMEI1981 / DI014	hypothetical protein	**358**	-	-	+	+	+	+
84	BMEI1982 /	hypothetical protein	**418**	-	-	+	+	+	+

	DI015								
85	BMEI1683 / DI018	zinc-dependent metallopeptidase	**482**	-	-	+	+	+	+
86	BMEI1684 / DI019	hypothetical protein	**149**	-	-	+	-	+	-
87	BMEI1685	hypothetical protein	**163**	-	-	+	-	-	-
88	BMEI1686 / DI021	hypothetical protein	**265**	-	-	+	+	+	+
89	BMEI1687 / DI022	hypothetical protein	**167**	-	-	+	+	+	+
90	BMEI1688	hypothetical protein	**431**	-	-	+	-	-	-
91	BMEI1689 / DI025	hypothetical protein	**271**	-	-	+	+	+	+
92	BMEI1690 / DI026	hypothetical protein	**160**	-	-	+	-	+	+
93	BMEI1691	hypothetical membrane spanning protein	**206**	-	-	+	-	-	-
94	BMEI1692 / DI038	flagellar protein FlgJ	**201**	-	-	+	-	+	-
95	BMEI1693	hypothetical protein	**151**	-	-	+	+	+	+
96	BMEI1694 / DI042	hypothetical protein	**150**	-	-	+	-	+	+
97	BMEI1695	hypothetical protein	**239**	-	-	+	+	+	+
98	BMEI1696 / DI052	hypothetical membrane spanning protein	**526**	-	-	+	+	+	+
99	BMEI1697 / DI056	virulence-associated protein E	**857**	-	-	+	+	+	+
100	BMEI1698 / DI057	hypothetical protein	**245**	-	-	+	+	+	+
101	BMEI1699 / DI058	hypothetical protein	**183**	-	-	+	+	+	+
102	BMEI1700 / DI059	hypothetical protein	**207**	-	-	+	+	+	+
103	BMEI1701 / DI060	hypothetical protein	**221**	-	-	+	+	+	+
104	BMEI1702 / DI061	transposase	**169**	-	-	+	-	+	-

+ Obtained RT-PCR fragment of the expected length
- No band was observed in the RT-PCR experiment

Table 5: RT-PCR analysis of proposed differential ORFs from *Brucella* species

Location	***B. suis***			***B. melitensis***			***B. abortus***		
	Pred.[1]	**Obs.**[2]	**NB**[3]	**Pred.**	**Obs.**	**NB**	**Pred.**	**Obs.**	**NB**
B. suis Chr. I	4	4	-	1	1	-	-	-	-
B. suis Chr. II	18	17	1	-	-		-	-	-
B. melitensis Chr. I	-	-	-	1	1	-	-	-	-
B. abortus	-	-	-	-	-	-	1	1	-
B. suis + *B. melitensis* Chr. I	1	1	-	2	2	-	-	-	-
B. *suis* + *B. melitensis Chr. II*	25	16	9	24	6	18	-	-	-
B. suis + *B. abortus* Chr. I	11	3	8	-	-	-	9	7	2
B. suis + *B. abortus* Chr. II	11	7	4	-	-	-	11	6	5
B. melitensis + *B. abortus* Chr. I	-	-	-	30	21	9	26	23	3

[1] Predicted
[2] Observed
[3] No Band

3.2.4 Additional differentiating features

Gene matches identified using the methods and fixed cutoffs described in *Experimental Procedures* were assumed valid in cases where pairwise matches had greater than 90% sequence identity over their full length. Gene matches having lower sequence identity were classified and marked as possible secondary differentiating features, and may also be biologically significant. Only 4.6% of sequence matches between presumed homologues spanned less than 95% of the query sequence length. A higher proportion of homologues were full-length on Chromosome I than on Chromosome II. The highest proportion of non full-length homologues among pairwise comparisons was *B. abortus* Chromosome II relative to *B. melitensis*. These incomplete matches represent a broad range of gene types, including amino acid transport and metabolism genes.

3.2.5 RT-PCR of proposed differentiating regions

Reverse transcription PCR (RT-PCR) was performed on genes predicted to differentiate the three species in order to determine whether they are transcribed in culture and in the species-specific pattern expected. Table 4 details the results of the RT-PCR analysis. Table 5 summarizes the transcription detected and predicted vs. observed results for each differential gene. The RT-PCR analysis was performed using 106 pairs of primers for 102 differentiating regions of the three *Brucella* species. Sixty-one predicted differentiating genes did not appear to be transcribed under experimental conditions, as the predicted amplicon was not observed. No amplicons were detected by RT-PCR in control samples that were predicted to be missing the particular differentiating region being probed. Additional study is needed in these cases to determine if transcription occurs while *Brucella* resides in host cells. Standard PCR reactions were performed to confirm the presence of differentiating genes in genomic DNA when no amplicon was observed by RT-PCR.

Unique region on B. suis Chromosome I Four unique putative genes were identified for *B. suis* Chromosome I. Amplicons of the predicted sizes were obtained from *B. suis* in RT-PCR experiments for each of these ORFs. This suggests that each of these ORFs represents a true gene from Chromosome I of the *B. suis* genome. No amplicon was detected in the other two species for gene BR1060. Contrary to prediction, an amplicon was observed from *B. melitensis* for the *B. suis* unique gene BR1846, although its length was 253 bp shorter than the amplicon from *B. suis*.

Unique region on B. suis Chromosome II Eighteen unique ORFs coding for hypothetical proteins are located on *B. suis* Chromosome II. It was shown that seventeen of them are transcribed only in *B. suis*. Among the transcribed ORFs, five code for the family of TraA/B proteins. Only the putative TraI protein coding ORF produces no amplicon when analyzed by RT-PCR.

Unique region on B. melitensis Chromosome I A recombinase coding gene (BMEI1661) is the only unique gene predicted in *B. melitensis*. Transcription of this gene was detected in *B. melitensis*, but not in the other two *Brucella* species as measured by RT-PCR.

Partial differential region in B. abortus A partially unique 6 kb region identified in *B. abortus* was tested in two different RT-PCR reactions. This region, which was tentatively identified as a continuous gene in preliminary annotation (S. Halling, personal communication) contains two segments of sequence with homology to the other two species, separated by a 1800 bp segment unique to *B. abortus*. The PCR primer pair designed for amplification of the unique 1800 bp central segment of this region amplified a transcript unique to *B. abortus*. A second primer pair designed within the common segment and across the unique *B. abortus* region of the sequence was expected to yield amplicons of different length for all three *Brucella* species, but produced inconsistent results. A supplementary primer pair designed for a short 100 bp region in the high similarity region at the beginning of this ORF also detected no transcription. Therefore the 6 kb putative CDS from *B. abortus* requires further investigation and perhaps a re-examination of the annotated ORF borders.

Brucella suis and B. melitensis Chromosome I Two differential regions were identified for Chromosome I of *B. suis* and *B. melitensis*. The ORF (BMEI0929/BR1057) coding for a diguanylate cyclase/phosphodiesterase (GGDEF) domain appears to be transcribed in both *Brucella* spp. The multidrug resistance protein A ORF (BMEI0926/BR1060) produced the expected length amplicon in *B. melitensis.* In *B. suis,* no amplicon was produced because our primers did not have a site to anneal in this species.

Brucella suis and B. melitensis Chromosome II Twenty-five predicted CDS coding for hypothetical proteins in both *B. suis* and *B. melitensis* were analyzed by RT-PCR. Several transcription patterns were observed. Genes transcribed in both *Brucella* species included: the putative 31 kDa outer membrane protein, glycosyl transferase and glucose-1-phosphate cytidylyl transferase. The ORFs transcribed in *B. suis*, but not *B. melitensis* include putative ThiJ/PfpI protein, fucose synthetase, cell wall surface protein, undecaprenyl phosphate alpha-N-acetylglucosaminyltransferase, glutamate-1-semialdehyde-2,1-aminomutase, dTDP-4-dehydrorhamnose 3,5-epimerase, epimerase/dehydratase, methyltransferase, and three glycosyl transferases. Proteins that were transcribed in *B. melitensis*, but not *B. suis*, included putative GDP mannose 4,6-dehydratase Bme9 and Bme2 proteins. Seven predicted ORFs produced no transcript in either species. For one of the putative glycosyl transferases, RT-PCR was performed only in *B. suis* and the expected size amplicon was obtained.

Brucella suis and B. abortus Chromosome I Two out of eleven predicted ORFs common to *B. suis* and *B. abortus* Chromosome I produced the expected size amplicons in both species—A putative transcriptional regulator and an AzlC family protein. Seven of the ORFs predicted as common to the two species were transcribed in *B. abortus* but not in *B. suis*, including a putative protease, hypothetical transcriptional regulator from the Cro/CI family and a major capsid protein, HK97 family. One ORF produced no amplicon in either species. Another ORF was probed with primer pairs designed to detect transcription only in *B. suis*, and gave the predicted size amplicon.

Brucella suis and B. abortus Chromosome II Expected amplicons were detected for five of eleven *B. suis* and *B. abortus* Chromosome II predicted differential regions. These included: a putative sugar ABC transporter, permease protein, a transcriptional regulator, an amino acid dehydrogenase, and twin-arginine translocation signal domain protein. Two ORFs were transcribed in *B. suis* but not *B. abortus*: putative beta-ketoadipyl CoA thiolase and another putative amino acid ABC transporter. One ORF was transcribed in *B. abortus* but not *B. melitensis*, and no transcription was detected in either species for three ORFs.

Brucella melitensis and B. abortus Chromosome I Twenty out of thirty predicted ORFs common in *B. abortus* and *B. melitensis* were transcribed in both species. These included: a virulence-associated protein E, zinc-dependent metallopeptidase and twenty-eight other hypothetical protein encoding ORFs. All of the ORFs transcribed in *B. melitensis* were also transcribed in *B. abortus*. Three ORFs were transcribed only in *B. melitensis*. No transcription was detected in any species for three ORFs, including the transposase and flagellar protein FlgJ. Of four primer pairs designed to detect transcription only in *B. melitensis*, only two produced an amplicon.

3.2.6 Analysis of differentiating gene islands

We identified several multi-gene islands that contain the majority of differentiating genes (Table 2). These species-specific segments may be responsible for differences in virulence or host preferences, and may therefore be termed "islands" as an extension of the term "pathogenicity island," [34]. These six islands alone are sufficient to discriminate between the three *Brucella* species. In a pairwise comparison, thirty-three regions were described as unique to either *B. suis* or *B. melitensis* [5]. In our three-way comparison with *B. abortus*, we find that many of these differentiating features can no longer be considered unique for the purpose of discriminating among the three species. Fewer single-species specific genes remain: twenty-two unique genes in *B. suis* and one in *B. melitensis,* which demonstrates the homogeneity of the genus. A complete list of differentiating genes is given in Table 4 and their significance is described below.

3.2.7 Metabolism

Three-way genome comparison revealed a potential unique amino-acid utilization ability in two species. Several components of an amino acid ABC transport system were found in *B. abortus* and *B. suis* but were absent in *B. melitensis*. This may indicate that *B. abortus* and *B. suis* have the

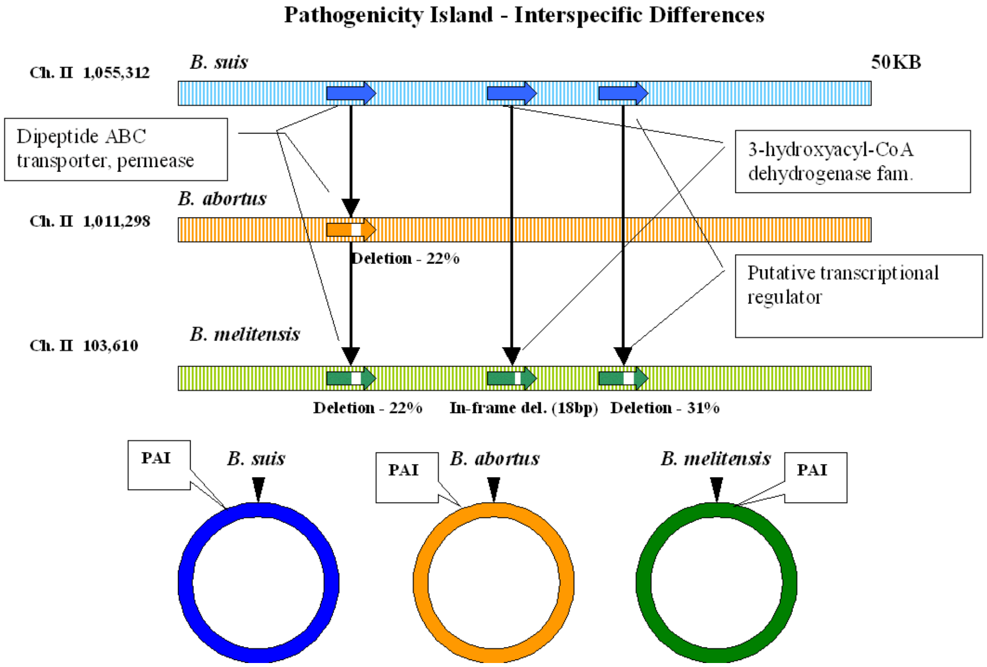


Fig. 6: Putative pathogenicity islands in *Brucella* spp. Detailed analysis of a putative pathogenicity island from *B. suis* reveals potentially significant differences in *B. abortus* and *B. melitensis*. Both *B. abortus* and *B. melitensis* have an in-frame deletion in one gene, while *B. melitensis* has in-frame deletions in two other genes.

ability to utilize a nutrient that *B. melitensis* does not. Most of these genes are present on the differentiating island SA2 (Table 2), suggesting that the acquisition or loss of this island was related to a change in environment or nutrient availability for the ancestral species. Two ABC transporter permeases (BR0952/BR0953) unique to *B. suis* were also identified which may confer for this species a metabolic activity unique among the *Brucellae.* Transcription of these genes in *B. suis* was detected by RT-PCR (Table 5).

3.2.8 Virulence

A detailed analysis of a 50 kb putative pathogenicity island [5] (BRA1072-1116/BMEII0183-227) was performed to complement our general comparison of gene content. This 50 kb region resides on Chromosome II of each *Brucella* species and may represent a composite transposon (Fig. 6) [5]. It is flanked with insertion sequences that suggest a foreign origin, and has a slightly atypical G+C content (56.8%). Although this island does not contain obvious virulence genes, it includes a large number of peptide ABC transporter genes which may encode a metabolic function relevant to pathogenicity. Comparison with *B. suis* shows that this region is also present in *B. melitensis* and *B. abortus* but with deletions in the dipeptide ABC transporter permease protein gene, the 3-hydroxyacyl-CoA dehydrogenase family protein gene, and a transcriptional regulator. Each of these small deletions is in-frame, but result in missing amino acids and altered function, leading to important metabolic differences between the three species.

A 25 kb island present in *B. suis* and *B. melitensis* was revealed by three-way comparison to be a potentially important differentiating feature. This island, absent only in *B. abortus* (island MS2,

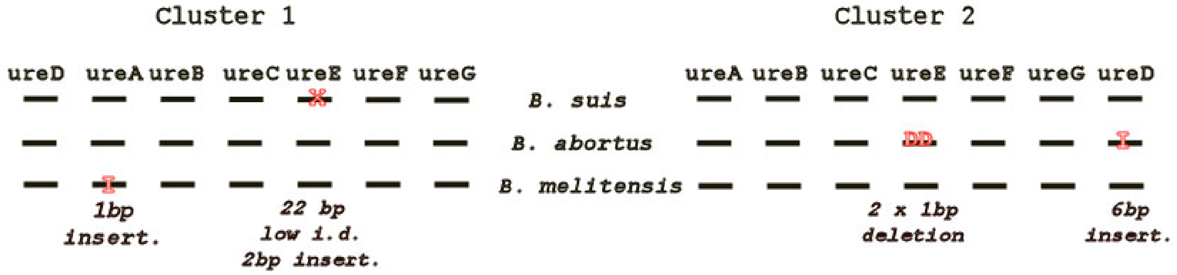


Fig. 7: Urease cluster comparison in *Brucella* spp. Comparison of two urease clusters present in all three *Brucella* species reveals differences among individual genes. Within each species, the clusters are paralogous and located on opposite ends of the chromosome. Insertions are marked with "I," deletions with "D," and regions of low identity with "X." *B. suis* annotation was used as reference for comparison. Consensus used to determine insertions vs. deletions.

Table 2), contains five glycosyl transferases (BMEII0835/0837/0840/0845-0847; BRA0420-0422/0427/0430/0432) and a succinoglycan biosynthesis transport protein (BMEII0838/BRA0429). However, no transcription of succinoglycan biosynthesis transport protein was detected by RT-PCR for either species. In *B. melitensis*, transcription of four out of five glycosyl transferases was detected by RT-PCR, while in *B. suis* transcription of only one of these genes was observed. These genes may be important in O- side chain biosynthesis - one of the known virulence determinants of *Brucella* [36]. This island also contains several uncharacterized genes that may be novel virulence factors of unknown function, including a putative outer membrane protein and several conserved hypothetical proteins. This island was shown to be present in *B. melitensis*, *B. suis*, *B. ovis*, *B. canis*, and *B. neotomae*; but not in *B. abortus* [37]. Vizcaino *et al.*, conjecture that this region is absent due to a deletion event before the differentiation of this species and its biovars, since none of the *B. abortus* biovars possess this region. The deletion of this island may have impacted the host range of *B. abortus* and driven its divergence from the *Brucella* ancestor.

A three-way comparison reveals species-specific differences in two gene clusters of urease subunits present on Chromosome II of *B. suis*, *B. abortus*, and *B. melitensis* (*ure*A-G-1 BR0267-BR0273 and *ure*A-G-2 BR1356-BR1362 in *B. suis*). Some subunits of these clusters are conserved among other bacterial species, and ureases have been shown to be important to virulence in several animal models of bacterial infection [5]. *B. melitensis* has a 1 bp insertion in *ure*A-1 (BR0268), representing a potential frameshift. A 6 bp insertion in the *ure*D-2 (BR1362) gene of *B. abortus* was identified, within overlapping segments of a highly repetitive region of the gene. In the *ure*E-2 gene (BR1359) of *B. abortus* two separate single base deletions are present, possibly shifting the frame of translation. Finally, the last 22 bp of *ure*E-1 (BR0271) were shown to be 100% identical in *B. abortus* and *B. melitensis* but significantly diverged in *B. suis*, including a 2 bp deletion. This variation predicts a frameshift insertions or deletion in at least one urease cluster gene in each species, which could prove to be significant to virulence differences (Fig. 7). Additional

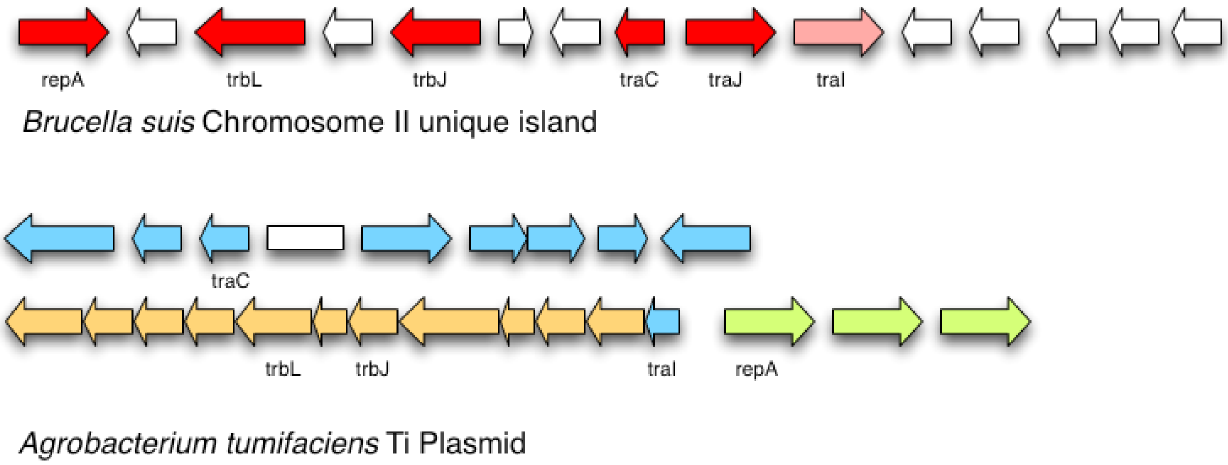

repA
trbL
trbJ
traC
traJ
traI
Brucella suis Chromosome II unique island
traC
trbL
trbJ
traI
repA
Agrobacterium tumifaciens Ti Plasmid

Fig. 8: Presence of conjugal transfer genes in *B. suis* unique region. Six members of *Agrobacterium tumefaciens* Ti plasmid *tra* and *trb* clusters of the conjugal transfer system are present in a *B. suis* unique region. Structure of the *A. tumefaciens* tra/trb region is from [35].

biochemical and genetic tests are needed to test the impact each gene has on urease activity.

3.2.9 Secretion Systems

The *vir*B region of *Brucella* encodes components of a type IV secretion system essential to intracellular trafficking and virulence [38]. Type IV secretion systems are macromolecular secretion pathways composed of multi-protein complexes, ancestral to bacterial conjugation systems [38, 39]. The type IV system of *Brucella* is homologous to the T-DNA transfer system of the closely related *Agrobacterium tumefaciens*. Transcription of the *virB* operon in *Brucella* is specifically induced within macrophages, and phagosome acidification is a key intracellular signal inducing VirB expression. Although the exact role of the VirB system is unclear, it is hypothesized that the type IV secretion system exports effector proteins from the phagosomal compartment into host cells. The identity and function of these effectors is unknown [38]. A comparison of the *virB* operon among the *Brucellae* has been described previously [40]. Out of twelve ORFs on *B. melitensis* Chromosome II, eleven were shown to have homologues in the *B. suis*, *B. abortus*, and *A. tumefaciens* genomes [40]. Our comparison confirms the conclusion that the *B. abortus virB* operon shares 97% identity with *B. suis*.

Our analysis also revealed a cluster of transfer genes (*tra/trb*) unique to *B. suis* and potentially significant to secretion (island S2, Fig. 8). Transcription of all but one gene in this island was observed by RT-PCR. Several genes in this region (*trbL*, *trbJ*, *traC*, *traJ*, *traI*, and *repA*) are homologous to genes involved in mating pair formation described for *Escherichia coli* plasmid RP4 [41], to receptor complex formation in bacteriophage-host gene transfer systems [42], and to genes of type IV secretion systems of other species of bacteria. *Agrobacterium* contains both a *vir*B type IV secretion system and a *tra/trb* bacterial conjugation system. These systems are homologous and share common ancestral origins, but they are functionally independent and physically separate [35, 43]. *Brucella* spp. lacks a bacterial conjugation system, which suggests

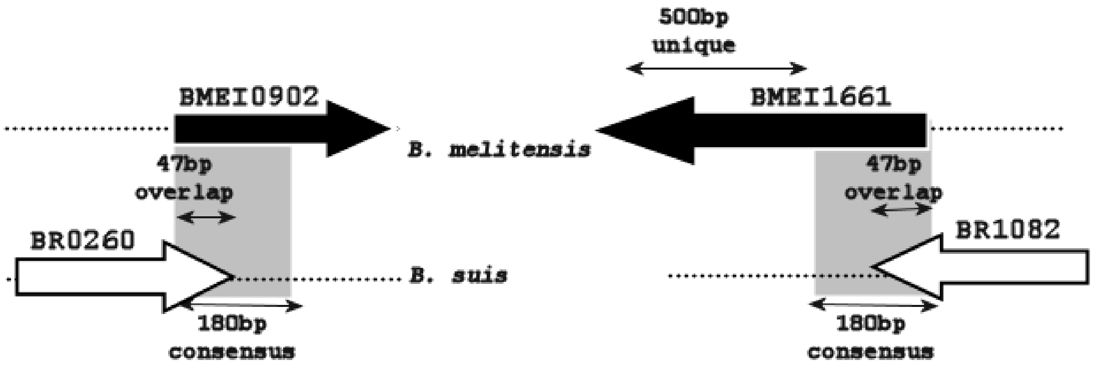


Fig. 9: Recombinase genes in *B. melitensis* and *B. suis.* Each genome shares a 180 bp consensus, but BMEI1661 contains ~500 bp unique to the species.

that the genes in this region play a role in type IV secretion, or are part of an uncharacterized macromolecule or gene transfer system. The majority of genes in this region are of unknown function. The *tra*J gene of the IncN plasmid pKM101 is homologous to the *virD4* gene of the *Agrobacterium vir*B operon [39]. This is the only gene of the 12-gene *Agrobacterium vir*B operon that has not been previously identified in *Brucella*. Although *Brucella* has not been observed to form a pilus, the *tra*C gene in *E. coli* is involved in pilus assembly [44]. Our results from RT-PCR experiments (Table 5) indicate that these *tra/trb* genes are expressed in *B. suis*. Additional studies are needed to determine if these genes have an important function in *B. suis*, or whether they are simply an artifact of some prior gene transfer event. The organization of this unique island (island S2) suggests a pattern of co-expression. The short intergenic region between the ORFs may indicate that these genes are organized as operons and are co-transcribed. In the case of the BRA0372-BRA0373 operon, the start codon of BRA0373 lies within BRA0372 that may indicate a –1 or –2 frameshift mechanism for expression of BRA0373. Examples of this type of gene/operon organization have primarily been identified in viruses [45, 46]. It has also been identified in prokaryotes [47], although in some cases it can be an artifact of annotation error [48]. Additional study is needed to confirm the annotation in this case.

Type III secretion systems are assembled from components of flagellar machinery (Christie, 2000). Although *Brucella* does not produce flagella, our analysis reveals a flagellar gene (FlgJ – BMEI1692) present in differentiating island MA1. This gene is on Chromosome I, instead of within one of three flagellar gene clusters on Chromosome II. It is also more than twice (~640 aa) the normal size (~313 aa) for this protein. In *B. melitensis*, all the structural genes for flagellum formation are present but genes for the chemotactic receptors or transducers are absent [49]. Based on the presence of several flagellar genes and a homolog of the LcrD virulence superfamily in *B. abortus*, it has been suggested that *Brucella* has the potential for motility and type III secretion [50]. However, a recent study detected no expression by RT-PCR in *B. melitensis* grown in Albimi broth in four flagellar genes (*flhB*,*flhP*,*fliR*,*fliF*) that are present in *B. melitensis*, *B. suis*, *B. abortus*, and *B. ovis* [51]. Our RT-PCR results revealed no expression of the flagellar differential FlgJ in *Brucella* grown in trypticase soy broth. Expression was detected in ten genes within the same island MA1 that are defined as hypothetical proteins [49]. Recent studies suggest that a flagellar gene promoter (*fliF*) is induced when *B. suis* is replicating in macrophages; additional studies on flagellar gene expression are being performed [49]. Thus it is likely that flagellar gene expression occurs when *Brucella* is replicating in an intracellular environment such as macrophages but not when grown in pure culture. The intriguing questions remaining to be answered are what product(s) are being excreted and for what purposes.

3.2.10 Site-specific recombinases

A recombinase gene (BMEI1661) was identified as the sole unique gene for *B. melitensis*, and our RT-PCR results indicate that it was transcribed. There are two resolvase family genes (BME1661/BMEI0902) in the *B. melitensis* annotation for Chromosome I located in opposite orientations. These two genes share homology over a 180 bp consensus sequence. However, one recombinase (BMEI1661) is much larger than the other (747 bp vs. 231 bp). They may be considered paralogous, but BME1661 contains more than 500 bp not present in any other species (Fig. 9).

In the *B. suis* annotation, there are also two resolvase recombinases of equal size (617 bp) and almost identical, and also in opposite orientations. These only have small matches to BME1661/BMEI0902 (~40 bp). However, both *B. abortus* and *B. suis* contain 2 copies of ~180 bp BME1661/BMEI0902, mostly within intergenic sequence.

Overall, a 180 bp consensus is present in two copies on all three species, but ~500 bp of the BMEI1661 gene in *B. melitensis* is unique to this species (Fig. 9). Site-specific recombination has been shown to be involved with acquisition of drug resistance genes and with alteration of gene expression [52], suggesting that this unique gene may play an important role in virulence.

3.2.11 Evolutionary implications

Our analysis reinforces the view that the *Brucellae* are highly similar genetically. It has been suggested that the low rate of genetic exchange between *Brucella* spp. and other species is due to their niches within cells as intracellular parasites (Boschiroli *et al.*, 2002). However, several multi-gene differentiating islands identified in our comparison (Table 2) contain atypical G+C contents that is consistent with gene acquisition via horizontal transfer. Island MA1 exhibits a G+C content of 52% and contains a putative phage integrase family transposase at the end of the gene cluster in both *B. abortus* and *B. melitensis*. *Escherichia coli* has a G+C content of 51.4%, and has been demonstrated to transfer a broad host range plasmid to *Brucella* under laboratory conditions [53]. Other islands have base compositions close to the average *Brucella* G+C content. Island MS2 exhibits a G+C content of 58% in both *B. melitensis* and *B. suis.* The presence of phage genes suggests that lysogenic conversion may have occurred (Boyd and Brussow, 2002). The island S2 that is unique to *B. suis* and containing 5 *tra/trb* genes has a G+C content of 55.6% and is flanked by a phage integrase homologue. Two phage gene homologues (a HK97 family phage major capsid protein and putative phage head-tail adaptor) are present within island SA1 and two phage gene

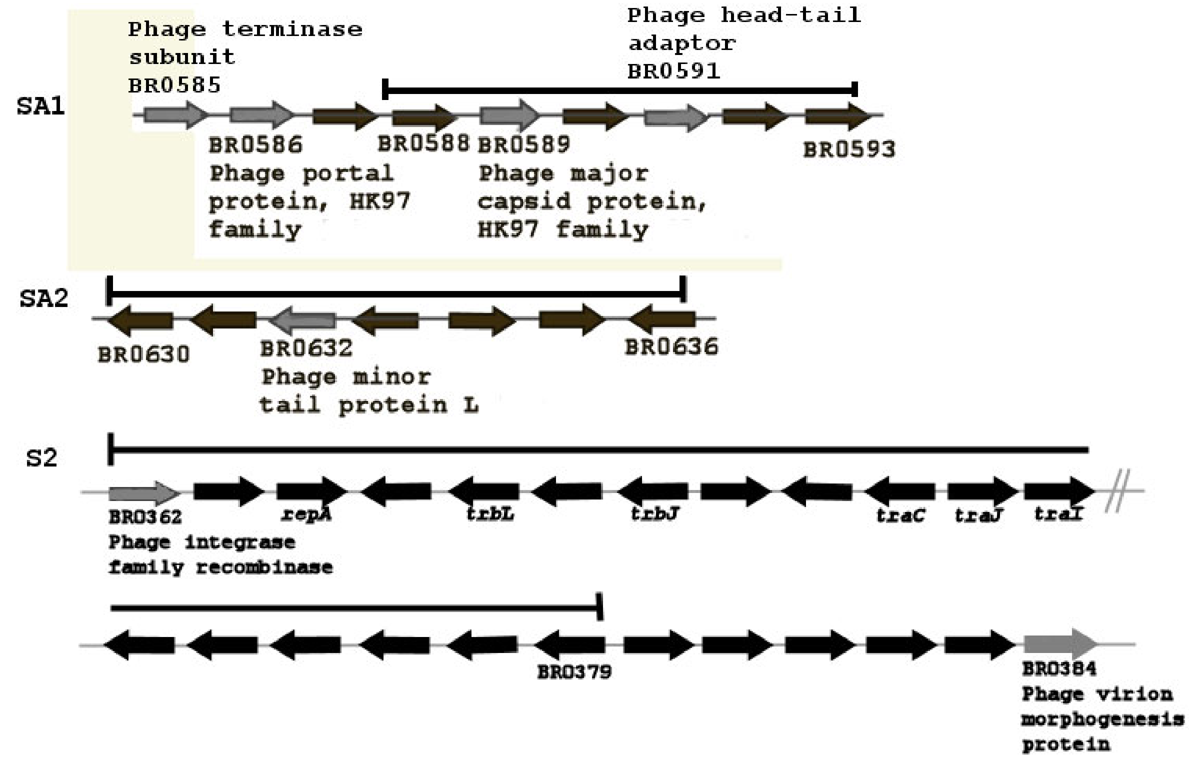

Phage terminase subunit BR0585
Phage head-tail adaptor BR0591
SA1
BR0586 Phage portal protein, HK97 family
BR0588
BR0589 Phage major capsid protein, HK97 family
BR0593
SA2
BR0630
BR0632 Phage minor tail protein L
BR0636
S2
BR0362 Phage integrase family recombinase
repA
trbL
trbJ
traC
traJ
traI
BR0379
BR0384 Phage virion morphogenesis protein

Fig. 10: Homologues of phage genes within and flanking *Brucella* differentiating islands. Figure shows islands as annotated in *B. suis*. Putative phage gene homologues not in annotation were identified using TIGRFAM (http://www.tigr.org/TIGRFAMs/index.shtml).

homologues (a HK97 family portal protein and a phage terminase subunit) flank the island (Fig. 10). Island SA2 contains a phage minor tail protein L homologue. This evidence is consistent with phage-mediated transduction and suggests that phages may have helped the *Brucellae* adapt to their intracellular niches.

3.2.12 Single nucleotide polymorphism (SNPs)

Genome comparison based on identification of homologues gives an incomplete picture of genetic differences. To complement our comparison approach, we quantified relative numbers of SNPs, which can lead to functional differences not detectable by homolog comparison. When comparing *B. abortus* to the other *Brucella* species, we identified over three times more SNPs within genes relative to *B. suis* annotation (3,721) than to *B. melitensis* (1,052). Also significant to gene expression are insertions/deletions within genes. We identified 182 insertions/deletions in *B. abortus* relative to *B. suis* and 110 relative to *B. melitensis*. We also detected 128 hypervariable regions of 5 mismatches or more in *B. abortus* relative to *B. suis*, and 58 relative to *B. melitensis*. These data suggest that *B. abortus* and *B. melitensis* may have diverged more recently than *B. suis*.

3.2.13 Taxonomic implications

The high degree of similarity our analysis demonstrates between these three genomes at both the gene and nucleotide levels lends weight to the hypothesis that the *Brucella* spp. should be grouped as biovars of the same species [5]. However, the biological differences between them warrant the retention of the classical species names for clinical and diagnostic reasons and practical convenience. Discrimination between these species is important for host-pathogen studies and for diagnostic purposes. Our analysis reveals sufficient genomic differences to discriminate between the three species.

4. Conclusions

4.1 *Brucella* Comparison

Rather than providing easy answers to questions of host preference or virulence determinants, our results provide a launching point for other studies. In the case of *B. abortus*, we do not find a "smoking gun" – a unique gene that has obvious implication for host preference patterns--but we have a better inventory of suspect genes to investigate.

Brucella is closely related to the soil bacterium *Ochrobactrum anthropi,* whose genome sequence will also be published [54]. This will allow for another dimension of comparison. *O. anthropi*, while very similar genetically, has a very different lifestyle than *Brucella* [54].

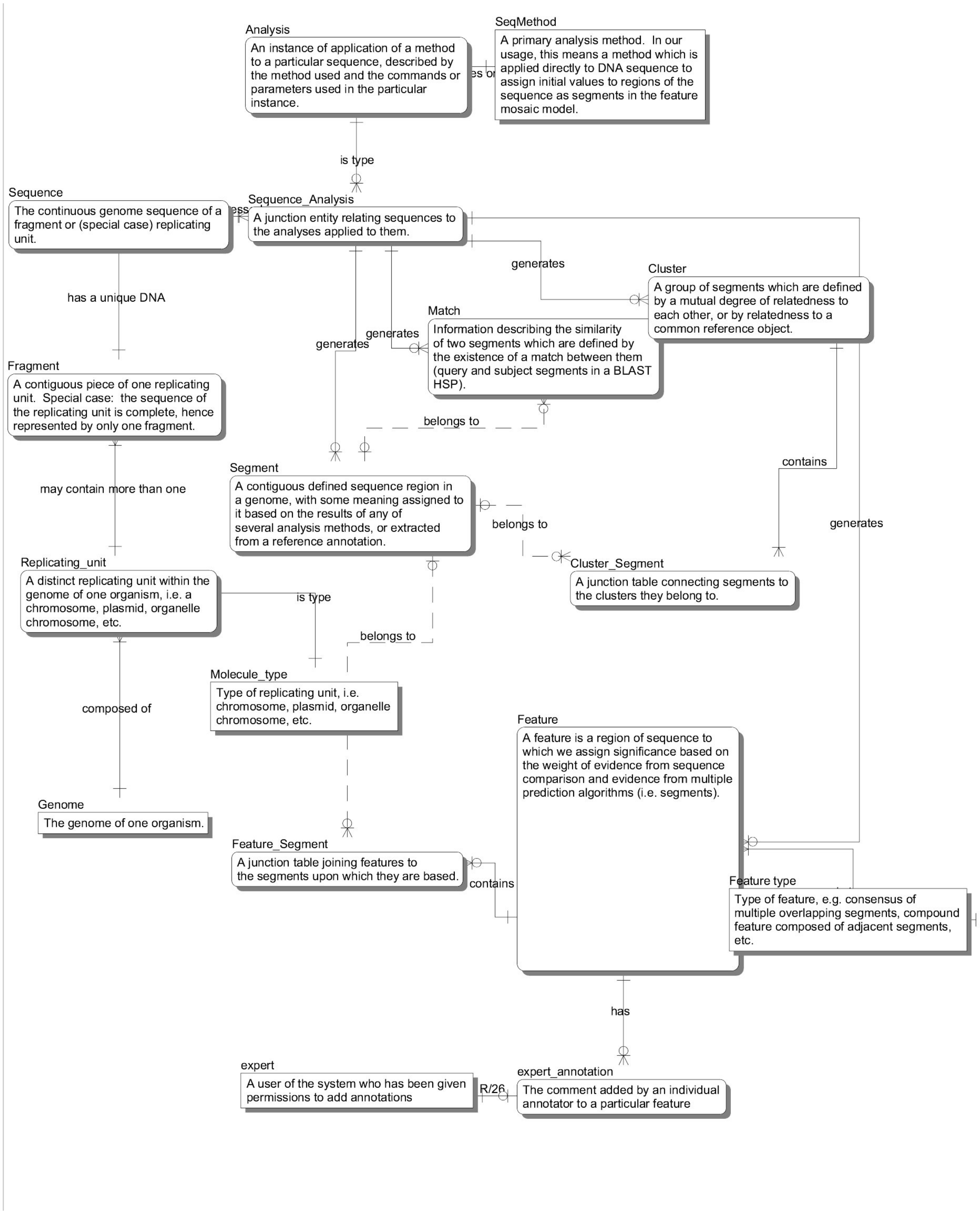
Results of these *Brucella* comparisons are currently being used to design discriminatory DNA oligonucleotide arrays for differential diagnosis of *Brucella* infections, as well as to examine differential gene expression during host-pathogen interactions. With these experiments, we hope to determine whether differences in virulence or host preferences between *Brucella* spp. are due to unique genes or differences in expression. We anticipate that the answers will lie in the results from a combination of the two approaches.

4.2 GenoMosaic Development

The utility of the GenoMosaic prototype was demonstrated by its ability to store and query sequence analysis results. The amount of information generated by a genome sequence comparison of just three species was immense, and was very difficult to manage without the benefit of database-backed query tools. The next step in the development of GenoMosaic will be to test its ability to handle additional sequences in the *Brucella* example, and also to test its flexibility by analyzing a much different set of sequences (such as chloroplast genomes).

APPENDIX A
Supplementary Figures

Supplementary Fig. 1A – GenoMosaic entity-relationship diagram. For a key to data modeling symbols, see Fig. 1



Analysis
An instance of application of a method to a particular sequence, described by the method used and the commands or parameters used in the particular instance.

SeqMethod
A primary analysis method. In our usage, this means a method which is applied directly to DNA sequence to assign initial values to regions of the sequence as segments in the feature mosaic model.

Sequence
The continuous genome sequence of a fragment or (special case) replicating unit.

Sequence_Analysis
A junction entity relating sequences to the analyses applied to them.

Cluster
A group of segments which are defined by a mutual degree of relatedness to each other, or by relatedness to a common reference object.

Match
Information describing the similarity of two segments which are defined by the existence of a match between them (query and subject segments in a BLAST HSP).

Fragment
A contiguous piece of one replicating unit. Special case: the sequence of the replicating unit is complete, hence represented by only one fragment.

Segment
A contiguous defined sequence region in a genome, with some meaning assigned to it based on the results of any of several analysis methods, or extracted from a reference annotation.

Cluster_Segment
A junction table connecting segments to the clusters they belong to.

Replicating_unit
A distinct replicating unit within the genome of one organism, i.e. a chromosome, plasmid, organelle chromosome, etc.

Molecule_type
Type of replicating unit, i.e. chromosome, plasmid, organelle chromosome, etc.

Feature
A feature is a region of sequence to which we assign significance based on the weight of evidence from sequence comparison and evidence from multiple prediction algorithms (i.e. segments).

Genome
The genome of one organism.

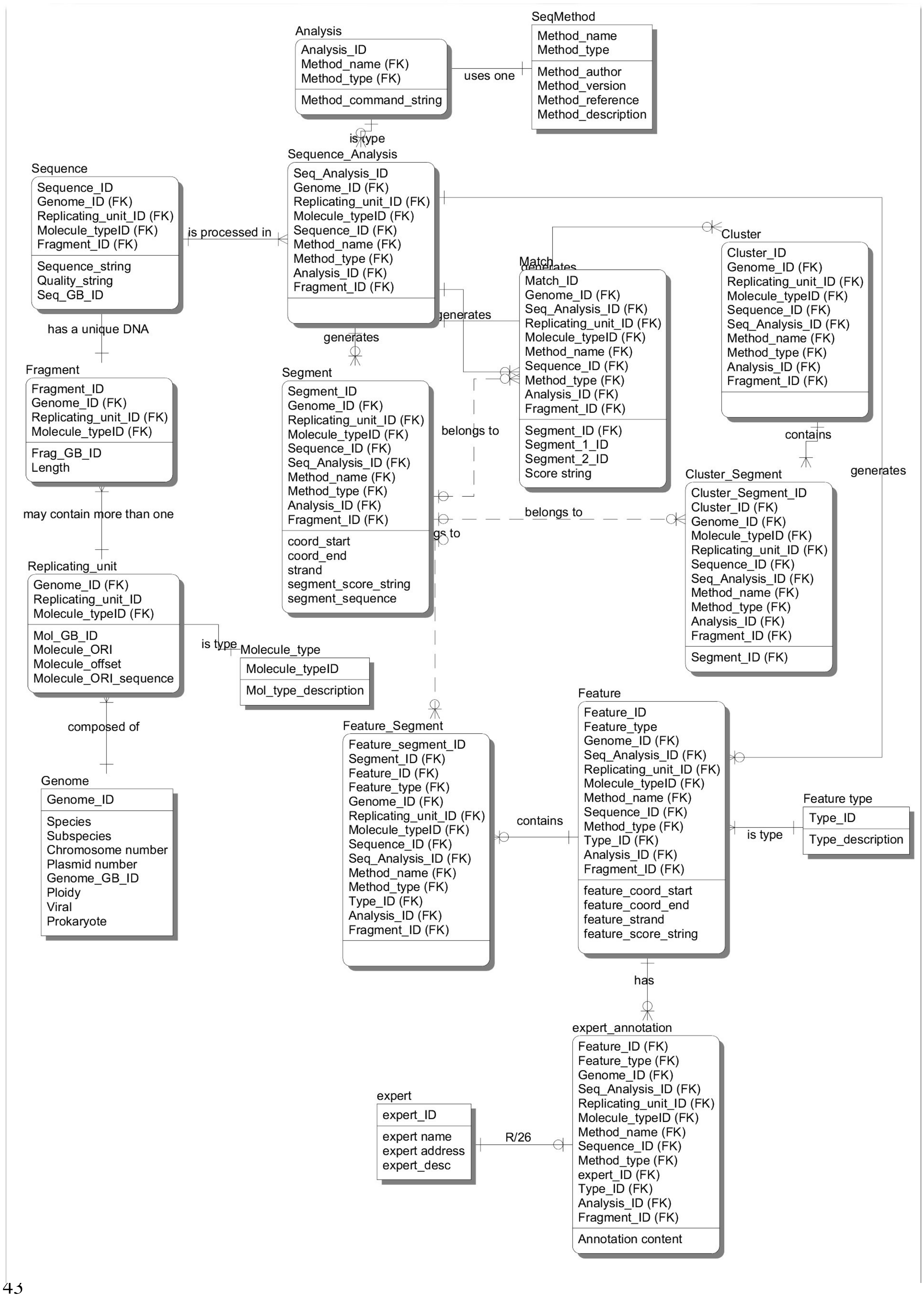
Feature_Segment
A junction table joining features to the segments upon which they are based.

Feature type
Type of feature, e.g. consensus of multiple overlapping segments, compound feature composed of adjacent segments, etc.

expert
A user of the system who has been given permissions to add annotations

expert_annotation
The comment added by an individual annotator to a particular feature

Supplementary Fig. 1B – GenoMosaic entity-relationship diagram. For a key to data modeling symbols, see Fig. 1



Supplementary Fig. 2 – GenoMosaic database structure

```
                 Table "public.genome"
      Column       |         Type          | Modifiers
-------------------+-----------------------+-----------
 genome_id         | integer               | not null
 species           | character varying(60) |
 subspecies        | character varying(60) |
 chromosome_number | integer               |
 plasmid_number    | integer               |
 genome_gb_id      | character varying(60) |
 ploidy            | character varying(3)  |
 viral             | boolean               |
 prokaryote        | boolean               |
Indexes: genome_pkey primary key btree (genome_id)

              Table "public.replicating_unit"
        Column         |         Type          | Modifiers
-----------------------+-----------------------+-----------
 replicating_unit_id   | integer               | not null
 genome_id             | integer               |
 molecule_type_id      | integer               |
 molecule_gb_id        | character varying(60) |
 molecule_ori          | character varying(60) |
 molecule_offset       | integer               |
 molecule_ori_sequence | character varying(60) |
Indexes: replicating_unit_pkey primary key btree (replicating_unit_id)
Foreign Key constraints: genome_id_fk FOREIGN KEY (genome_id) REFERENCES
genome(genome_id) ON UPDATE NO ACTION ON DELETE NO ACTION,
                         molecule_type_id_fk FOREIGN KEY (molecule_type_id) REFERENCES
molecule_type(molecule_type_id) ON UPDATE NO ACTION ON DELETE NO ACTION

                 Table "public.molecule_type"
          Column           |         Type          | Modifiers
---------------------------+-----------------------+-----------
 molecule_type_id          | integer               | not null
 molecule_type_description | character varying(60) |
Indexes: molecule_type_pkey primary key btree (molecule_type_id)

                 Table "public.fragment"
       Column        |         Type          | Modifiers
---------------------+-----------------------+-----------
 fragment_id         | integer               | not null
 replicating_unit_id | integer               |
 fragment_gb_id      | character varying(60) |
 length              | integer               |
Indexes: fragment_pkey primary key btree (fragment_id)
Foreign Key constraints: replicating_unit_id_fk FOREIGN KEY (replicating_unit_id)
REFERENCES replicating_unit(replicating_unit_id) ON UPDATE NO ACTION ON DELETE NO
ACTION
```

```
                Table "public.sequence"
     Column      |          Type           | Modifiers
-----------------+--------------------------+-----------
 sequence_id     | integer                  | not null
 fragment_id     | integer                  |
 sequence_string | character varying(10000) |
 quality_string  | character varying(60)    |
 sequence_gb_id  | character varying(60)    |
Indexes: sequence_pkey primary key btree (sequence_id)
Foreign Key constraints: fragment_id_fk FOREIGN KEY (fragment_id) REFERENCES
fragment(fragment_id) ON UPDATE NO ACTION ON DELETE NO ACTION

                     Table "public.sequence_analysis"
        Column        |  Type   |      Modifiers
----------------------+---------+-----------------------------------------------------
 sequence_analysis_id | integer | not null default
nextval('public.sequence_analysis_sequence_analysis_id_seq'::text)
 sequence_id          | integer |
 analysis_id          | integer |
Indexes: sequence_analysis_pkey primary key btree (sequence_analysis_id)
Foreign Key constraints: sequence_id_fk FOREIGN KEY (sequence_id) REFERENCES
"sequence"(sequence_id) ON UPDATE NO ACTION ON DELETE NO ACTION,
                         analysis_id_fk FOREIGN KEY (analysis_id) REFERENCES
analysis(analysis_id) ON UPDATE NO ACTION ON DELETE NO ACTION

                      Table "public.analysis"
        Column         |          Type          |   Modifiers
-----------------------+------------------------+-------------------------------------
 analysis_id           | integer                | not null default
nextval('public.analysis_analysis_id_seq'::text)
 method_name           | character varying(60)  |
 method_type           | character varying(60)  |
 method_command_string | character varying(200) |
Indexes: analysis_pkey primary key btree (analysis_id)

                    Table "public.seqmethod"
       Column       |          Type          |    Modifiers
--------------------+------------------------+----------------------------------------
 seqmethod_pk       | integer                | not null default
nextval('public.seqmethod_seqmethod_pk_seq'::text)
 method_name        | character varying(60)  |
 method_type        | character varying(60)  |
 method_author      | character varying(60)  |
 method_version     | character varying(60)  |
 method_reference   | character varying(400) |
 method_description | character varying(200) |
Indexes: seqmethod_pkey primary key btree (seqmethod_pk)

                     Table "public.segment"
        Column        |           Type           |  Modifiers
----------------------+--------------------------+------------------------------------
 segment_id           | integer                  | not null default
nextval('public.segment_segment_id_seq'::text)
 sequence_analysis_id | integer                  |
 segment_name         | character varying(100)   |
 coord_start          | integer                  |
 coord_end            | integer                  |
 strand               | character(2)             |
 segment_score_string | character varying(100)   |
 segment_sequence     | character varying(10000) |
Indexes: segment_pkey primary key btree (segment_id)
Foreign Key constraints: sequence_analysis_id_fk FOREIGN KEY (sequence_analysis_id)
REFERENCES sequence_analysis(sequence_analysis_id) ON UPDATE NO ACTION ON DELETE NO
ACTION

                      Table "public.match"
        Column        |          Type          |  Modifiers
----------------------+------------------------+--------------------------------------
 match_id             | integer                | not null default
nextval('public.match_match_id_seq'::text)
 segment_id           | integer                |
 sequence_analysis_id | integer                |
 segment_1_id         | integer                |
 segment_2_id         | integer                |
 score_string         | character varying(100) |
Indexes: match_pkey primary key btree (match_id)
Foreign Key constraints: segment_id_fk FOREIGN KEY (segment_id) REFERENCES
segment(segment_id) ON UPDATE NO ACTION ON DELETE NO ACTION,
                         sequence_analysis_id_fk FOREIGN KEY (sequence_analysis_id)
REFERENCES sequence_analysis(sequence_analysis_id) ON UPDATE NO ACTION ON DELETE NO
ACTION

                     Table "public.cluster"
   Column   |  Type   |    Modifiers
------------+---------+--------------------------------------------------------------
 cluster_id | integer | not null default
nextval('public.cluster_cluster_id_seq'::text)
 match_id   | integer |
Indexes: cluster_pkey primary key btree (cluster_id)
Foreign Key constraints: match_id_fk FOREIGN KEY (match_id) REFERENCES
"match"(match_id) ON UPDATE NO ACTION ON DELETE NO ACTION

                   Table "public.cluster_segment"
       Column       |  Type   |   Modifiers
--------------------+---------+--------------------------------------------------------
 cluster_segment_id | integer | not null default
nextval('public.cluster_segment_cluster_segment_id_seq'::text)
 cluster_id         | integer |
 segment_id         | integer |
Indexes: cluster_segment_pkey primary key btree (cluster_segment_id)
Foreign Key constraints: cluster_id_fk FOREIGN KEY (cluster_id) REFERENCES
"cluster"(cluster_id) ON UPDATE NO ACTION ON DELETE NO ACTION,
                         segment_id_fk FOREIGN KEY (segment_id) REFERENCES
segment(segment_id) ON UPDATE NO ACTION ON DELETE NO ACTION

                   Table "public.feature_segment"
        Column        |  Type   |  Modifiers
----------------------+---------+-----------------------------------------------------
 feature_segment_id   | integer | not null default
nextval('public.feature_segment_feature_segment_id_seq'::text)
 sequence_analysis_id | integer |
 segment_id           | integer |
 feature_id           | integer |
Indexes: feature_segment_pkey primary key btree (feature_segment_id)
Foreign Key constraints: feature_id_fk FOREIGN KEY (feature_id) REFERENCES
feature(feature_id) ON UPDATE NO ACTION ON DELETE NO ACTION,
                         sequence_analysis_id_fk FOREIGN KEY (sequence_analysis_id)
REFERENCES sequence_analysis(sequence_analysis_id) ON UPDATE NO ACTION ON DELETE NO
ACTION

                    Table "public.feature"
        Column        |           Type           | Modifiers
----------------------+--------------------------+-----------------------------------
 feature_id           | integer                  | not null default
nextval('public.feature_feature_id_seq'::text)
 sequence_analysis_id | integer                  |
 type_id              | integer                  |
 feature_coord_start  | integer                  |
 feature_coord_end    | integer                  |
 feature_strand       | character(3)             |
 feature_score_string | character varying(100)   |
 feature_sequence     | character varying(10000) |
Indexes: feature_pkey primary key btree (feature_id)
Foreign Key constraints: type_id_fk FOREIGN KEY (type_id) REFERENCES
feature_type(type_id) ON UPDATE NO ACTION ON DELETE NO ACTION,
                         sequence_analysis_id_fk FOREIGN KEY (sequence_analysis_id)
REFERENCES sequence_analysis(sequence_analysis_id) ON UPDATE NO ACTION ON DELETE NO
ACTION

                 Table "public.feature_type"
      Column      |          Type          | Modifiers
------------------+------------------------+------------------------------------------
 type_id          | integer                | not null default
nextval('public.feature_type_type_id_seq'::text)
 type_description | character varying(200) |
Indexes: feature_type_pkey primary key btree (type_id)

                  Table "public.expert"
       Column       |          Type          | Modifiers
--------------------+------------------------+-----------
 expert_id          | integer                | not null
 expert_name        | character varying(100) |
 expert_address     | character varying(100) |
 expert_description | character varying(200) |
Indexes: expert_pkey primary key btree (expert_id)

                     Table "public.expert_annotation"
        Column        |          Type           |  Modifiers
----------------------+-------------------------+-------------------------------------
 expert_annotation_id | integer                 | not null default
nextval('public.expert_annotation_expert_annotation_id_seq'::text)
 feature_id           | integer                 |
 expert_id            | integer                 |
 annotation_content   | character varying(1000) |
Indexes: expert_annotation_pkey primary key btree (expert_annotation_id)
Foreign Key constraints: feature_id_fk FOREIGN KEY (feature_id) REFERENCES
feature(feature_id) ON UPDATE NO ACTION ON DELETE NO ACTION
```

References

1. Preuss, P., *Berkeley Lab Science Beat: Comparative Genomics at the Joint Genome Institute: an Interview.* 2002.
2. Boschiroli, M.L., V. Foulongne, and D. O'Callaghan, *Brucellosis: a worldwide zoonosis.* Curr Opin Microbiol, 2001. 4(1): p. 58-64.
3. CDC, *CDC-PHEPR Biological Diseases/Agents*. 2003, CDC.
4. Franz, D.R., *Foreign animal disease agents as weapons in biological warfare.* Ann N Y Acad Sci, 1999. 894: p. 100-4.
5. Paulsen, I.T., et al., *The Brucella suis genome reveals fundamental similarities between animal and plant pathogens and symbionts.* Proc Natl Acad Sci U S A, 2002. 99(20): p. 13148-53.
6. Moreno, E., A. Cloeckaert, and I. Moriyon, *Brucella evolution and taxonomy.* Vet Microbiol, 2002. 90(1-4): p. 209-27.
7. Pizarro-Cerda, J., et al., *Virulent Brucella abortus prevents lysosome fusion and is distributed within autophagosome-like compartments.* Infect Immun, 1998. 66(5): p. 2387-92.
8. Arenas, G.N., et al., *Intracellular trafficking of Brucella abortus in J774 macrophages.* Infect Immun, 2000. 68(7): p. 4255-63.
9. Frenchick, P.J., R.J. Markham, and A.H. Cochrane, *Inhibition of phagosome-lysosome fusion in macrophages by soluble extracts of virulent Brucella abortus.* Am J Vet Res, 1985. 46(2): p. 332-5.
10. Pizarro-Cerda, J., E. Moreno, and J.P. Gorvel, *Invasion and intracellular trafficking of Brucella abortus in nonphagocytic cells.* Microbes Infect, 2000. 2(7): p. 829-35.
11. Anderson, T.D. and N.F. Cheville, *Ultrastructural morphometric analysis of Brucella abortus-infected trophoblasts in experimental placentitis. Bacterial replication occurs in rough endoplasmic reticulum.* Am J Pathol, 1986. 124(2): p. 226-37.
12. Detilleux, P.G., B.L. Deyoe, and N.F. Cheville, *Entry and intracellular localization of Brucella spp. in Vero cells: fluorescence and electron microscopy.* Vet Pathol, 1990. 27(5): p. 317-28.
13. Delcher, A.L., et al., *Fast algorithms for large-scale genome alignment and comparison.* Nucleic Acids Research, 2002. 30(11): p. 2478-2843.
14. Berriman, M. and K. Rutherford, *Viewing and annotating sequence data with Artemis.* Brief Bioinform, 2003. 4(2): p. 124-32.
15. Lewis, S.E., et al., *Apollo: a sequence annotation editor.* Genome Biol, 2002. 3(12): p. RESEARCH0082.
16. Harris, N.L., *Annotating sequence data using Genotator.* Mol Biotechnol, 2000. 16(3): p. 221-32.
17. Liu, C., et al., *DNannotator: Annotation software tool kit for regional genomic sequences.* Nucleic Acids Res, 2003. 31(13): p. 3729-35.
18. Zafar, N., R. Mazumder, and D. Seto, *CoreGenes: A computational tool for identifying and cataloging "core" genes in a set of small genomes.* BMC Bioinformatics, 2002. 3(1): p. 12.
19. BioPerl, *The BioPerl project.* 2003.
20. PostgreSQL, *PostgreSQL open source RDMS.* 2003.
21. Altschul, S.F., et al., *Basic local alignment search tool.* J. Mol. Biol. 215:403-410, 1990.
22. Hertz, G.Z., G.W. Hartzell, 3rd, and G.D. Stormo, *Identification of consensus patterns in unaligned DNA sequences known to be functionally related.* Comput Appl Biosci, 1990. 6(2): p. 81-92.

23. Delcher, A.L., et al., *Improved microbial gene identification with GLIMMER.* Nucleic Acids Res, 1999. 27(23): p. 4636-41.
24. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Res, 2000. 28(1): p. 33-6.
25. Associates, C., *ERwin Data Modeler*. 2003.
26. Busse, H.J., E.B. Denner, and W. Lubitz, *Classification and identification of bacteria: current approaches to an old problem. Overview of methods used in bacterial systematics.* J Biotechnol, 1996. 47(1): p. 3-38.
27. Rozen, S. and H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers.* Methods Mol Biol, 2000. 132: p. 365-86.
28. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. 31(13): p. 1-10.
29. Zuker, M., *Nucleic acid quickfold.* 2003.
30. Zuker, M., *Nucleic acid 2-state hybridization server.* 2003.
31. Halling, S.M. and N.A. Koster, *Use of detergent extracts of Brucella abortus RB51 to detect serologic responses in RB51-vaccinated cattle.* J Vet Diagn Invest, 2001. 13(5): p. 408-12.
32. Jumas-Bilak, E., et al., *Differences in chromosome number and genome rearrangements in the genus Brucella.* Mol Microbiol, 1998. 27(1): p. 99-106.
33. Michaux-Charachon, S., et al., *Genome structure and phylogeny in the genus Brucella.* J Bacteriol, 1997. 179(10): p. 3244-9.
34. Perna, N.T., et al., *Genome sequence of enterohaemorrhagic Escherichia coli O157:H7.* Nature, 2001. 409(6819): p. 529-33.
35. Alt-Morbe, J., et al., *The conjugal transfer system of Agrobacterium tumefaciens octopine-type Ti plasmids is closely related to the transfer system of an IncP plasmid and distantly related to Ti plasmid vir genes.* J Bacteriol, 1996. 178(14): p. 4248-57.
36. Fernandez-Prada, C.M., et al., *Interactions between Brucella melitensis and human phagocytes: bacterial surface O-Polysaccharide inhibits phagocytosis, bacterial killing, and subsequent host cell apoptosis.* Infect Immun, 2003. 71(4): p. 2110-9.
37. Vizcaino, N., et al., *Characterization of a Brucella species 25-kilobase DNA fragment deleted from Brucella abortus reveals a large gene cluster related to the synthesis of a polysaccharide.* Infect Immun, 2001. 69(11): p. 6738-48.
38. Boschiroli, M.L., et al., *The Brucella suis virB operon is induced intracellularly in macrophages.* Proc Natl Acad Sci U S A, 2002. 99(3): p. 1544-9.
39. Christie, P.J., *Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines.* Mol Microbiol, 2001. 40(2): p. 294-305.
40. DelVecchio, V.G., et al., *The genome sequence of the facultative intracellular pathogen Brucella melitensis.* Proc Natl Acad Sci U S A, 2002. 99(1): p. 443-8.
41. Haase, J., et al., *Bacterial conjugation mediated by plasmid RP4: RSF1010 mobilization, donor-specific phage propagation, and pilus production require the same Tra2 core components of a proposed DNA transport complex.* J Bacteriol, 1995. 177(16): p. 4779-91.
42. Grahn, A.M., et al., *Assembly of a functional phage PRD1 receptor depends on 11 genes of the IncP plasmid mating pair formation complex.* J Bacteriol, 1997. 179(15): p. 4733-40.
43. Cook, D.M., et al., *Ti plasmid conjugation is independent of vir: reconstitution of the tra functions from pTiC58 as a binary system.* J Bacteriol, 1997. 179(4): p. 1291-7.

44. **Schmidt-Eisenlohr, H., N. Domke, and C. Baron, *TraC of IncN plasmid pKM101 associates with membranes and extracellular high-molecular-weight structures in Escherichia coli.* J Bacteriol, 1999. 181(18): p. 5563-71.**
45. **Choi, J., Z. Xu, and J.H. Ou, *Triple decoding of hepatitis C virus RNA by programmed translational frameshifting.* Mol Cell Biol, 2003. 23(5): p. 1489-97.**
46. **van Eyll, O. and T. Michiels, *Non-AUG-initiated internal translation of the L* protein of Theiler's virus and importance of this protein for viral persistence.* J Virol, 2002. 76(21): p. 10665-73.**
47. **Rogozin, I.B., et al., *Purifying and directional selection in overlapping prokaryotic genes.* Trends Genet, 2002. 18(5): p. 228-32.**
48. **Szymanski, M. and J. Barciszewski, *Lessons from sequenced genomes. Overlapping genes in Methanococcus jannaschii?* IUBMB Life, 2000. 49(2): p. 121-3.**
49. **Letesson, J.J., et al., *Fun stories about Brucella: the "furtive nasty bug".* Vet Microbiol, 2002. 90(1-4): p. 317-28.**
50. **Halling, S.M., *On the presence and organization of open reading frames of the nonmotile pathogen Brucella abortus similar to class II, III, and IV flagellar genes and to LcrD virulence superfamily.* Microb Comp Genomics, 1998. 3(1): p. 21-9.**
51. **Abdallah, A.I., et al., *Type III Secretion Homologues Are Present in Brucella melitensis, B. ovis, and B. suis biovars 1, 2, and 3.* Curr Microbiol, 2003. 46(4): p. 241-5.**
52. **Grindley, N.D., *Site-specific recombination: synapsis and strand exchange revealed.* Curr Biol, 1997. 7(10): p. R608-12.**
53. **Verger, J.M., et al., *Conjugative transfer and in vitro/in vivo stability of the broad-host-range IncP R751 plasmid in Brucella spp.* Plasmid, 1993. 29(2): p. 142-6.**
54. **Tsolis, R.M., *Comparative genome analysis of the alpha -proteobacteria: relationships between plant and animal pathogens and host specificity.* Proc Natl Acad Sci U S A, 2002. 99(20): p. 12503-5.**

date revised September 15, 2003

CURRICULUM VITAE

DAVID M. STURGILL

Personal Data

Address:	3737 Legation St., NW Apt. 201 Washington, DC 20015
Phone:	(202) 244-5999
Email:	dsturgil@vt.edu
Office:	4107 Derring Hall

Education **Virginia Tech** Blacksburg, VA

Master of Science in Biology, expected September 2003
Bioinformatics Option

Virginia Tech Blacksburg, VA

Bachelor of Science in Biology, 1990

Additional Graduate Studies

Foundation of Advanced ES, National Institutes of Health
Essentials of Toxicology, Spring, 1995.
Fundamentals of Epidemiology, Fall, 2000.

Professional Experience **AMERICAN INTERNATIONAL HEALTH ALLIANCE** Washington, DC

1997-2001 **Associate, Information and Communications Technology**

- Organized and conducted information technology workshops, and developed curricula for health professionals in the NIS (Newly Independent States of the former Soviet Union) and CEE (Central and Eastern Europe)
- Collaborated on company website design, designed and maintained interactive features including internet-accessible database of over 800 electronic documents
- Managed initiative to train NIS physicians in development of evidence-based clinical practice guidelines

TECHNICAL ASSESSMENT SYSTEMS, INC. Washington, DC

1992-1996 **Associate Scientist**

- Performed statistical analyses, managed and presented data for submission to regulatory agencies
- Attended and reported on congressional hearings and tracked legislation for corporate clients

BIOSIS, INC. Philadelphia, PA

1990-1992 **Editor/Analyst**

- Edited and indexed scientific journals for incorporation into commercial database
- Organized taxonomic and other scientific reference materials for group use

Teaching Experience

Virginia Tech

Graduate Teaching Assistant

General Biology Lab / Principles of Biology Lab

Fall 2001, Spring 2002, Fall 2002

American International Health Alliance

New Independent States Information Coordinator Training Workshops
Training in medical informatics for medical professionals
L'viv, Ukraine; July 14-18, 1997
St. Petersburg, Russia; July 19-24, 1999
L'viv, Ukraine; July 26-30, 1999
Tbilisi, Georgia; August 4-6, 1999
Almaty, Kazakstan; October 11-16, 1999
Tbilisi, Georgia; April 22-29, 2000
Almaty, Kazakstan; June 5-10, 2000

Central and Eastern Europe Information Coordinator Training Workshops
Training in medical informatics for medical professionals
Krk, Croatia; July 21-25, 1997
Kosice, Slovakia; July 12-16, 1998

Clinical Practice Guidelines and Continuous Quality Improvement Workshops
Caucasus Region: Tbilisi, Georgia. October 19-25, 2000
Russian Federation: Moscow, Russia. January 24-26, 2001
West NIS: Kiev, Ukraine. January 29-31, 2001
Central Asia: Almaty Kazakstan. June 18-22, 2001

Presentations

Systematic Multiple Comparisons Reveal Significant Differences in Genomes of Brucella abortus, B. melitensis, and B. suis.
CRWAD - 54th Annual Brucellosis Satellite Meeting
St. Louis, MO. November, 2002

Systematic Genomic Comparison of Three Brucella Spp. and a Data Model for Feature-Based Multiple Genome Analysis
Intelligent Systems for Molecular Biology - ISMB 2002
Poster Session
Edmonton, Canada. August, 2002

Introduction to Continuous Quality Improvement and Clinical Practice Guidelines
AIHA Annual Conference - Plenary Session
Budapest, Hungary. July, 2000

Publications

GenoMosaic: On-Demand Multiple Genome Comparison and Comparative Annotation
Cynthia Gibas [1] *, David Sturgill [1] , and Jennifer Weller [2] . [1] Department of Biology, Virginia Polytechnic Institute and State University and [2] School of Computational Science, George Mason University. *Corresponding author.
Published in *Proceedings of the IEEE BIBE Conference, 2003*

Multiple Comparisons Reveal Significant Differences in Genomes of *Brucella abortus*, *B. melitensis*, and *B. suis.*
David Sturgill[1], Shirley Halling[2] and Cynthia Gibas[1].
[1]Department of Biology, Virginia Tech, Blacksburg, VA; [2]National Animal Disease Center, U.S. Dept. of Agriculture, Ames, IA.
Submitted to *Molecular Microbiology* for publication, 2003

Language Skills: Intermediate Spanish and Russian