Brucella genomics as we enter the multi-genome era

David O’Callaghan and Adrian M. Whatmore

Advance Access publication date 19 September 2011

Abstract
The genus Brucella includes species considered among the world’s most important zoonotic pathogens, with brucellosis remaining a significant problem in large parts of the world. Over the last decade a number of Brucella genomes have been fully sequenced providing new insights into this relatively poorly understood group of organisms. In the forthcoming months and years, the availability of many additional genomes should help in further understanding of the evolution, host specificity and pathogenicity of this group as well as providing a resource to further improve epidemiological typing tools. This article describes progress to date and looks forward to the opportunities that should be afforded as we enter an era of multiple, freely available, Brucella genome sequences.

Keywords: Brucella; genomics; genome sequence; zoonoses; virulence

BRUCELLA AND BRUCELLOSIS
The genus Brucella are members of the α-proteobacteria, a large and diverse group of Gram-negative microbes. Brucella is a good example of a subset of the α-proteobacteria that establishes long-term, often chronic, interactions with higher eukaryotes [1]. The bacterium is a facultative intracellular pathogen. It uses its virulence factors, including the VirB type IV secretion system, to block host cell defences and create a novel intracellular niche that allows it to survive and replicate in host macrophages. In animals, the bacterium causes infectious abortion and other reproductive problems while in man the disease is also known as Malta or Undulant Fever. Brucellosis is an important zoonosis and the disease in man is very severe and is contracted following direct contact with infected animals or following the ingestion dairy products derived from infected animals. Brucellosis has a worldwide impact in terms of its epidemiology, human health risks and its impact on trade. The greatest impact of brucellosis is in the poorer, rural areas of the world, without the funds, nor the infrastructure to establish surveillance and eradication programmes. For a recent comprehensive review of the biology of Brucella and brucellosis see Ref. [2].

PHYLOGENY AND CLASSIFICATION
The genus Brucella is classified within the family Brucellaceae of the order Rhizobiales in the alpha class of the proteobacteria. The Brucellaceae also include Mycoplana, Ochrobactrum, Pseudochrobactrum, Paenochrobactrum and Crabtreella. For many years, the genus was thought to be composed of six ‘classical’ species which could be differentiated by preferential mammalian host and a set of antigenic and metabolic phenotypes [3] (Table 1). The three major species in terms of disease and economic impact for man, Brucella abortus, Brucella melitensis and Brucella suis are further divided into biovars (bv) based on a range of phenotypic and serological characteristics [3]. In the early 1990s, Brucella-like strains were isolated from marine mammals and more recently novel strains have been isolated from rodents, non-human primates and from atypical human infections [4]. At present, there are 10 recognized species, with
several more new isolates, including a strain isolated from a Bull frog (H. Scholz, personal communication), under investigation likely to eventually be recognized as novel species.

**BRUCELLA GENOMICS**

**From the beginnings of Brucella genomics to physical maps**

The study of the *Brucella* genome began in the late 1960s, when DNA homology was used to show that the different strains were related [5, 6] followed by an attempt to generate mutants and map genes by Altenbern a few years later [7]. At the DNA level, the different species have over 90% identity in DNA/DNA hybridization assays, leading Verger et al. [8] to propose that *Brucella* was monospecific genus with a single species, *B. melitensis*. This decision was overturned in 2006 partly to ‘avoid potential harm’ that might result from grouping highly pathogenic organisms within the same species as organisms that appear non-pathogenic for man (see Refs [4, 9] for recent reviews). One of the first incentives to study the *Brucella* genome was to develop rapid and simple strain identification methods for epidemiology. The application of pulsed-field gel electrophoresis (PFGE) revolutionized the analysis of restriction fragment-length polymorphisms (RFLPs) of bacterial genomes. In 1988, Allardet Servent et al. [10] used PFGE to analyse the *XbaI* profiles of 23 different *Brucella* strains, including 19 reference strains. The RFLP profiles showed that each species has a conserved, unique profile. These data were recently extended to the recently discovered strains from marine mammals [11]. After a first imperfect physical map with a single circular chromosome [12], it became apparent that *B. melitensis* possesses two circular chromosomes, one of 1.1 Mb and the other of 2.2 Mb giving a genome of ~3.2 Mb, which can be visualized by PFGE if undigested genomic DNA is migrated [13]. Physical mapping showed just how conserved the *Brucella* genomes are [14, 15], however a fascinating observation was made when the undigested genomes of four of the *B. suis* bv were separated by PFGE. The genome of *B. suis* 1330 (bv1) was similar to those of *B. melitensis* and *B. abortus*, while those of strains Thomsen (bv2) and 40 (bv4) were different; the small chromosome was larger (1.35 Mb) and the large chromosome smaller (1.85 Mb) [15]. In bv3, strain 686, we observed a single circular replicon of 3.3 Mb. These three possible genomic structures appear to be the products of recombination events between the three *rrn* loci.

**Entering the genomic age**

In 2002, the complete genome sequences of *B. melitensis* 16 M and *B. suis* 1330 were released [16, 17] followed shortly afterwards by that of a *B. abortus* field isolate [18]. The comparison of these three genomes confirmed just how closely related the three species are. As previously shown by the physical maps, the genomes had extremely high levels of synteny (other than a large inversion in the *B. abortus* bv1 small chromosome [14, 19]) with only a small number indels [18]. At the sequence level, comparison of the three genomes revealed a very high level of conservation with a mean of 1 SNP every 463 nucleotides (for comparison with the other extreme; when two *Helicobacter pylori* isolates were compared only 8/1500 genes had ≥98% sequence identity) [18].

**Multiple genome analysis**

Over the past few years, several more genome sequences have been reported. The overriding conclusion was that the genomes are all highly conserved, but with a variety of indels and recombinations. Multiple alignments of nine closed genomes, illustrated in Figure 1, show the major difference in chromosome 1 in a *B. suis* bv2 genome previously identified by physical mapping as discussed above [15], where a 210 kb segment of chromosome I has been translocated to chromosome 2 [20]. In contrast, chromosome 2 was found to be somewhat more variable with more internal rearrangements including

<table>
<thead>
<tr>
<th>Table 1: Currently described Brucella species</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Species</strong></td>
</tr>
<tr>
<td><strong>‘Classical’ species</strong></td>
</tr>
<tr>
<td><em>B. melitensis</em></td>
</tr>
<tr>
<td><em>B. abortus</em></td>
</tr>
<tr>
<td><em>B. suis</em></td>
</tr>
<tr>
<td><em>B. canis</em></td>
</tr>
<tr>
<td><em>B. neotomae</em></td>
</tr>
<tr>
<td><em>B. ovis</em></td>
</tr>
<tr>
<td><strong>Newly described species</strong></td>
</tr>
<tr>
<td><em>B. pinnipedialis</em></td>
</tr>
<tr>
<td><em>B. ceti</em></td>
</tr>
<tr>
<td><em>B. microti</em></td>
</tr>
<tr>
<td><em>B. inopinata</em></td>
</tr>
</tbody>
</table>
the 700 kb inversion in \textit{B. abortus} genomes described above. Comparison of the genome of the live vaccine strain, \textit{B. abortus} S19 with two virulent strains (2308 and 9-941) identified a number of possible differences that may lead to the attenuation [21]. The most striking was a large deletion in a gene encoding an autotransporter protein. Over 96\% of the open reading frames were identical. Analysis of the \textit{B. ovis}, and \textit{B. canis} genomes, both representing naturally rough species, allowed the identification of the genetic lesions responsible. Interestingly, these events appear to be independent [20, 22] (see below for further discussion). Analysis of the \textit{B. ovis} data also suggested the genome of this species is in the early steps of reduction by gene inactivation and loss, a common theme in the evolution of the \alpha-proteobacteria as they adapt to their eukaryotic host [1, 23], suggesting that it is an ‘older’ species.

**Sequencing a representative of every \textit{Brucella} species and bv**

The advent of next generation sequencing has been a technological revolution. While sequencing the 16 M genome cost over one million dollars and took many months, it is now possible to sequence a genome overnight for a few hundred dollars. Working with the Broad Institute (Boston, MA, USA) we have recently sequenced the genomes of 25 strains, with the aim of giving us a representative genome of every species and bv to further understanding of diversity within the genus. The data are freely available at the Broad Institute website (http://www.broadinstitute.org/annotation/genome/brucella_group/MultiHome.html). Comparison of 42 available \textit{Brucella} genomes is under way in collaboration with Rebecca Wattam at VBI and Jeff Foster at NAU. Data analysis is available through the PATRIC website (http://www.patricbrc.org) and a manuscript describing a comparative analysis of these genomes is in preparation at the time of writing this review.

Although we have not yet published our analysis of the 42 genomes, recent studies have used different approaches to exploit genome sequence data from 10 strains. Wattam \textit{et al.} [20] used protein sequences to build phylogenetic trees, while Foster \textit{et al.} [24] used SNPs. Interestingly, the two studies produced virtually identical trees, which were also similar to those built on the basis of other approaches including data from multilocus sequence analysis (MLSA), variable number of tandem repeat (VNTR) analysis and PFGE RFLP analysis. Figure 2 shows a maximum likelihood phylogenetic analysis of 10 genomes based on concatenated protein families [20]. Division into
four clades corresponding to B. melitensis/B. abortus, B. ovis, B. suis/canis and B. ceti is consistent with trees based on other molecular approaches. Construction of a tree including closest phylogenetic neighbours confirms the homogeneity of the Brucella genomes described to date and confirms the position of Ochrobactrum as the closest relative of Brucella [20]. The trees fit almost perfectly with the classical typing systems, at least at the species level, confirming that, despite the extreme homogeneity of the classical members of the genus, virtually all the classical species correspond to phylogenetically distinct clusters. The one exception is B. canis which whole-genome phylogenies confirm arose as a clone from within the B. suis clade with estimates of the age of the split as in the range of 7500–22 500 years ago [24]. Interestingly, this fits with the timeline suggested for the domestication of dogs [25]. The very close relationship observed between B. canis and its closest B. suis relative explain difficulties described over many years in developing discriminatory diagnostic assays to separate these species. Since B. suis as currently defined is paraphyletic no single DNA-based assay can distinguish all B. suis isolates from all other Brucella species as the paraphyly of B. suis will inevitably mean such assays also identify B. canis [24]. Extension of this study using the data from over 40 genomes (discussed below) gives similar results. One recent study, however, using newly developed and unproven methods was completely at odds with all other reports [26]. Here, the traditional clades were split, with different conclusions depending on the method used. Some methods even separated two independent sequences of the same genome that has been sequenced by two different methods, implying that the same sequence is not its own closest relative.

**A second round of sequencing**

Although analysis of the first round of genome sequencing is still ongoing, rapid technological advances recently facilitated us obtaining funding for the sequencing of an additional 300 strains representing large collections of field isolates. Isolates have been selected to include as geographically, temporally and phenotypically diverse a collection as possible in order to further understanding of relationships between Brucella isolates. This will allow the development of new and more accurate epidemiological tools, at a level unimaginable only 5 years ago, driven by a detailed understanding of the population structure of the group.
Exploiting genomics data: molecular epidemiology and typing

Until very recently, the high levels of conservation and lack of variability in Brucella made it impossible to differentiate strains reliably to below the bv level. The availability of genome sequences has made high-resolution molecular epidemiology possible. Existing PCR and PCR RFLP assays were extended to encompass more strains. However, it was the development of MLSA, and particularly VNTR analysis, that have revolutionized Brucella epidemiology.

MLSA is a powerful tool to study the global epidemiology and phylogenetic relationships in bacterial populations. Combining data from sets of multiple gene fragments can be highly discriminatory while retaining signatures of longer-term evolutionary relationships [27]. Since the first application of MLSA to Brucella in 2007 [28], it has been used to show that the marine mammal isolates would more parsimoniously be divided into at least three species [29] agreeing with physical mapping data and VNTR data [11, 29], and to confirm that new isolates of B. microti and B. inopinata as well as novel strains isolated from a baboon, and from Australian rodents, are truly Brucella [30–35]. In all these cases, the novel species appear to be associated with distinct mammalian populations or species.

The identification of hypervariable regions of tandem repeats caused by slipped strand mispairing or replication slippage [36] or by recombination between homologous repeat sequences [37] has allowed the development of high-resolution tests to discriminate between strains. These markers, known as VNTRs, are considered high-speed molecular clocks [38]. Comparison of the genome sequences of B. melitensis 16 M, B. suis 1330 and B. abortus 9–941 allowed Bricker [39] to identify eight base pair tandem repeat sequences at nine different loci and develop a typing scheme called ‘HOOF-Prints’ (Hypervariable Octomeric Oligonucleotide Finger-Prints). This allowed discrimination between B. abortus bv1 strains from unrelated outbreaks and was a breakthrough in molecular epidemiology. At least two other sets of VNTR markers have been identified, notably a 21-locus scheme [40] and a 15-locus scheme [41]. Although there is much overlap between the three schemes the latter two schemes differ from ‘HOOF-Prints’ in that they include loci with a much broader range of evolutionary speeds allowing reliable clustering of isolates at higher ‘taxonomic’ levels while retaining the exquisite discriminatory power of ‘HOOF-Prints’.

Genomic islands and variability

Genome sequence data has allowed the identification and characterization of the genomic islands (GIs) that differ amongst the different Brucella species. The main interest in genomic islands is due to their potential role in the acquisition of virulence or fitness determinants via horizontal transfer.

Certain Brucella GIs carry genes encoding virulence factors, these include the ubo and wbk clusters on chromosome 1 encoding LPS biosynthesis genes, genes encoding polysaccharide synthesis and key outer membrane proteins [42–44]. The TIR domain containing protein Btp1 [45] is carried on the 21 kb GI-3 [42]. The MgtC protein, a key virulence determinant involved in magnesium uptake [46] is found on a region which has clearly been acquired horizontally [47] as have the two loci carrying the tatABC and tatD genes encoding the twin argentine transporter (Lavigne and O’Callaghan, unpublished data).

Several studies have been unable to show a role for certain GI-encoded genes in virulence [48, 49], however this may be because the virulence assay used was inappropriate to observe host-specific virulence factors. Several of the Brucella GIs contain phage and plasmid-related genes again attesting to their acquisition by horizontal transfer. This leads to the question of when did these events occur? In nature, brucellae are physically and genetically isolated as they multiply principally within their preferential host and have no obvious period of life in the environment. This has led to the suggestion that the different strains of Brucella are clones that have co-evolved independently with their mammalian hosts [14, 50]. A constant feature of the Brucella genome is the absence of plasmids and temperate phage and there is no evidence of transfer of genetic material in nature, although transduction at very low frequency has been shown (Koulakov and O’Callaghan, unpublished data). Recently it has been suggested that genetic exchanges can occur between intracellular Rickettsia by conjugation [51], however Brucella does not have a functional conjugation system. Many Brucella GIs carry integrases belonging to the tyrosine recombinase family suggesting that these regions are potentially mobile. Until recently, there have been very few reports of genomic instability in Brucella other than the...
observation that the IS711 insertion sequence had jumped into the \textit{wboA} gene of the vaccine strain RB51 resulting in its rough phenotype as well as a report demonstrating transposition of IS711 [52]. Lavigne et al. [48] showed that the IncP region in \textit{B. suis} could excise and exist as circular intermediates, however no stable excision events were identified. More recently, Mancilla et al. [53] reported that the spontaneous dissociation from smooth to rough LPS seen in \textit{B. abortus} was often due to a recombinase-mediated excision of GI-2 which carries the \textit{wboA} and \textit{wboB} genes. A similar event was most probably the cause of the rough phenotype in \textit{B. ovis}, but not in \textit{B. canis}. More work is needed to investigate the stability of GI in \textit{Brucella}.

**FUTURE PROSPECTS FOR BRUCELLA GENOMICS**

Genomic studies have already contributed substantially to our understanding of the biology of \textit{Brucella} and have facilitated development of new tools to identify and characterize members of the group. As we move into an era of availability of multiples of genomes within individual species progress should accelerate in a number of areas where genomic analyses offer huge promise. These include both directly understanding relationships between isolates and providing a framework for the generation of hypotheses for further biological investigation.

Exploitation of existing and forthcoming genome sequences will potentiate our understanding across multiple areas of \textit{Brucella} biology. Virulence and pathogenic processes associated with \textit{Brucella} remain relatively poorly understood and multiple genome comparisons should facilitate the identification of additional regions of difference that may be associated with virulence. These may range from the type of large GIs described earlier through to SNPs such as that recently identified as affecting promoter function and gene expression of a potential virulence factor [54]. Comparison of genomes of attenuated strains with virulent strains may also provide further clues. In addition to the previously completed genome sequence of the live vaccine strain \textit{B. abortus} S19, genome sequences of a number of Chinese vaccine strains [55, 56] have recently been released and projects ongoing at the Broad Institute will generate complete sequences of vaccine strains \textit{B. abortus} RB51 and \textit{B. melitensis} Rev1. One fundamental issue that arises repeatedly when trying to control brucellosis is the difficulty in distinguishing, by serological approaches, naturally infected animals from those that have received one of the live attenuated vaccines. If comparative genomics can offer clues towards development of a reliable assay to do this it could hugely impact on ability to tackle the worldwide brucellosis problem.

One of the enduring mysteries of \textit{Brucella} remains the basis of their host specificity in the face of the relatively minor genetic variation between species. For example genome sequences have shown that \textit{B. canis} is separated from its nearest \textit{B. suis} neighbour by only 253 SNPs [24] despite the apparently exquisite host specificity of the former. Identification of these genetic differences provides a realistic framework for examining the biological basis of host specificity.

A controversial area in \textit{Brucella} biology remains the taxonomy of the group. A sensible and rational taxonomy should ideally reflect phylogenetic relationships and the availability of multiple whole genome sequences and a robust and extensive phylogenetic framework will be crucial in informing ongoing discussions in this area. Genome sequence data should also further understand the distribution of markers that may be used to define the species. Traditionally, an insertion sequence, IS711, is considered a robust marker for the \textit{Brucella} genus. Comparative analysis of \textit{Brucella} genomes with those of related organisms will clarify such issues as well as potentially allowing the identification of additional useful molecular markers.

The promised availability of hundreds of genome sequences in the near future will undoubtedly further understanding of the evolutionary history of the group and its emergence from, and relationship with, other members of the Brucellaceae. The reliability of an evolutionary hypothesis is crucially dependent on inclusion of as representative and extensive a data set as possible and some earlier hypotheses, based on limited numbers of genomes, are likely to need to be reassessed given the emerging data. Understanding the evolutionary framework and population structure of an organism is also crucial to the design of rational diagnostic assays based on robust and stable markers such as canonical SNPs [24]. Such SNPs, that define particular clades, have already been identified on the basis of MLSA and exploited in reliable and robust diagnostic assays [57–59]. At the moment, such assays have been largely confined to identification at the level of
FUNDING
INSERM, Universite de Montpellier 1 (Work in INSERM U1047); Agence National pour La Recherche; Foundation Pour la Recherche Medical; DEFRA (Work on brucellosis at AHVLA).

References