Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Biologia Cellulare e Molecolare

Ciclo XXXI

**Settore Concorsuale: 05/E2**

**Settore Scientifico Disciplinare: BIO/11**

# Core Genome Multilocus Sequence Typing and Single Nucleotide Polymorphism Analysis in the Epidemiology of *Brucella melitensis* Infections

**Presentata da:** Claudio Patavino

**Coordinatore Dottorato**
Prof. Capranico Giovanni

**Supervisore**
Prof. Ferrè Fabrizio

**Esame finale anno 2019**

## Abstract

The use of whole-genome sequencing (WGS) using next-generation sequencing (NGS) technologies has become a widely accepted method for microbiology laboratories in the application of molecular typing, for outbreak tracing and genomic epidemiology. Several studies demonstrated the usefulness of WGS data analysis through single-nucleotide polymorphism (SNP) calling from a reference sequence analysis for *Brucella melitensis*, whereas gene-by-gene comparison through core-genome multilocus sequence typing (cgMLST) has not been explored so far. The current study developed an allele-based cgMLST method and compared its performance to that of the genome-wide SNP approach and the traditional multilocus variable-number tandem repeat analysis (MLVA) on a defined sample collection. The data set was comprised of 37 epidemiologically linked animal cases of brucellosis as well as 71 isolates with unknown epidemiological status, composed of human and animal samples collected in Italy. The cgMLST scheme generated in this study contained 2,704 targets of the *B. melitensis* 16M reference genome. We established the potential criteria necessary for inclusion of an isolate into a brucellosis outbreak cluster to be 6 *loci* in the cgMLST and 7 in WGS SNP analysis. Higher phylogenetic distance resolution was achieved with cgMLST and SNP analysis than with MLVA, particularly for strains belonging to the same lineage, thereby allowing diverse and unrelated genotypes to be identified with greater confidence. The application of a cgMLST scheme to the characterization of *B. melitensis* strains provided insights into the epidemiology of this pathogen, and it is a candidate to be a benchmark tool for outbreak investigations in human and animal brucellosis.

KEYWORDS:*Brucella melitensis*, MLVA, SNP analysis, cgMLST

**CONTENTS**

## LIST OF FIGURES

## LIST OF TABLES

## 1. INTRODUCTION

Brucellosis is an infectious disease caused by bacteria genus *Brucella*, currently is one of the world's most widespread zoonoses (diseases caused by infections that are shared between animals and people - Center for Diseases Control an Prevention - CDC, 2017), and it is a leading cause of economic losses in production of domestic ruminants (Seleem et al., 2010; Pappas et al., 2006). Humans can contract the disease by contact with infected animals or their products, with unpasteurized milk being the most common source of brucellosis in urban populations (Godfroid et al., 2005; Moreno 2014). *Brucella melitensis*, which infects primarily sheep and goats, is the most frequent agent of brucellosis in humans, and it leads to the most severe manifestation of the disease (Corbel 2006). Brucellosis can cause of range of signs and symptoms, some of which may present for prolonged periods of time, the initial symptoms are: fever, sweats, malaise, anorexia, headache, pain in muscles, joint, and/or back, fatigue.Others signs and symptoms may persist for longer periods of time or may never go away (recurrent fevers, arthritis, swelling of the testicle and scrotum area, swelling of the heart, (endocarditis), neurologic symptoms (in up to 5% of all cases), chronic fatigue, depression, swelling of the liver and/or spleen) (CDC, 2017).

Due to the high public health and economic burden of brucellosis, European countries have applied surveillance, control, and eradication programs for many years, and most of them have acquired the Officially *Brucella melitensis*-Free (OBF) status. The disease, however, still persists in several countries in the Mediterranean area.

Efficient and reliable surveillance methods are essential for detection and control of outbreaks and largely depend on collection and access to epidemiological data. Currently, epidemiological investigations rely on the availability of standardized and effective molecular typing methods and analysis tools that allow the public health laboratories to identify and trace an outbreak back to its source.

Identification and typing of *B. melitensis* are still traditionally performed with the use of biotyping techniques. This methodology, however, suffers from inconsistencies and requires handling of the live bacteria. For this reason, PCR-based typing is now commonly used as an alternative to the culture-dependent typing methods (Al Dahouk et al., 2007; Garofolo et al., 2013b) but do not provide sufficient resolution between the isolates. *B. melitensis* is a highly clonal, i.e., monomorphic pathogen, which renders its differentiation at the strain level very difficult (Waattam et al., 2014). Pattern-based techniques such as pulsed field gel electrophoresis and amplified fragment length polymorphism have been applied in the past, but these techniques were not able to differentiate *Brucella* at the subspecies level, which correlated with low intra- and interlaboratory reproducibility (Whatmore 2009).

In recent years, the typing methods have shifted toward genome-based approaches that finally allowed an accurate differentiation between *Brucella* isolates and the establishment of a common consensus for the subtyping schemes of this pathogen (Garofolo et al., 2013a; Le Fleche et al., 2006; Tan et al., 2015).

To date, Multiple-Locus Variable number tandem repeat Analysis (MLVA), is considered the most efficient typing method for *Brucella* spp. MLVA is a method used to perform molecular typing of particular microorganisms, and it utilizes the naturally occurring variation in the number of tandem repeated DNA sequences found in many different *loci* in the genome of a variety of organisms. The molecular typing profiles are uses to study transmission routes, to assess sources of infection and also to assess the impact of human intervention such as vaccination and use of antibiotics on the composition of bacterial populations. Several studies demonstrated that MLVA has a high discriminating resolution, in congruence with multilocus sequence typing (MLST), and is sufficient for in-depth study of either genome evolution or outbreak epidemiology (Carrico et al., 2013). According to MLVA schemes, the *B. melitensis* population can be divided into West Mediterranean, East Mediterranean, and American lineages (Al Dahouk et al., 2007; Whatmore et al., 2016).

Moreover, with the development of an international repository, the MLVA data can be stored on web servers and shared between research institutes, thereby increasing MLVA utility as a tool used for analysis of Brucella epidemiology in the world (http://microbesgenotyping.i2bc.parissaclay.fr/databases/view/-907) (Grissa et al., 2008).

However, this typing method has several weaknesses, related both to the nature of variable-number tandem repeats (VNTRs) as well as to laboratory demands of the technique itself (Garofolo et al., 2013b).

With advances in and decreased cost of whole-genome sequencing (WGS), new methods of pathogen typing, including gene-by-gene comparison using core genome multilocus sequence typing (cgMLST), as well as single-nucleotide polymorphism (SNP) calling based on a reference sequence analysis, are considered to be a suitable and more informative replacement of the gold standard typing schemes (Schurch et al., 2018; Hyden et al., 2016). cgMLST is performed by assigning specific alleles to a predefined set of core genes, i.e. genes present in all strains of a given bacterial species. Validated schemes for several pathogens are publicly available and can be shared to ensure reproducibility and comparability of the results across laboratories (Schurch et al., 2018).

The aim of this work is to report the results obtained by comparing these two methods of analysis used to describe a brucellosis outbreak.

# 2. REVIEW OF LITERATURE

## 2.1. *Brucella* spp. - taxonomy and classification

Brucellosis is a zoonotic bacterial disease caused by several species in the genus *Brucella*, which are gram-negative, non-spore-forming, intracellular facultative pathogens that cause chronic infection in mammalian hosts (WHO, 2006). *Brucella* is a member of the *Brucellaceae* family, in the order *Rhizobiales*, class *Alphaproteobacteria*. It shows close genetic relatedness to some plant pathogens and symbionts of the genera *Agrobacterium* and *Rhizobium*, as well as animal pathogens (*Bartonella*) and opportunistic or soil bacteria (*Ochrobactrum*) (OIE, 2016). Brucellosis is listed in the Terrestrial Animal Health Code of the World Organization for Animal Health (OIE) and must be compulsorily notified to the OIE. Reproductive losses are the most common syndrome in animals, while people may suffer from a debilitating nonspecific illness or localized involvement of various organs. Each species of *Brucella* tends to be associated with a specific animal host, but other species can be infected, especially when they are kept in close contact (OIE, 2017).

Until now twelve different *Brucella* species have been described (Scholz et al., 2016; Whatmore et al., 2014) (Table 1). The six species most commonly associated with domestic animals are *B. melitensis* and *B. abortus* (Meyer & Shaw, 1920), *B. suis* (Huddleson, 1929), *B. ovis* (Buddle, 1956), *B. neotomae* (Stoenner & Lackman, 1957) and *B. canis* (Carmichael & Bruner, 1968). *Brucella melitensis*, *B. abortus* and *B. suis* are further classified into biovars. In the 1990s, two new *Brucella* species were found in marine mammals (Ewalt et al., 1994; Ross et al., 1994) and these were subsequently categorized as *B. ceti* and *B. pinnipedialis* (Foster et al., 2007), both with zoonotic potential (Whatmore et al., 2008). Another newly described species, *B. microti* was isolated from common voles and red foxes (Scholz et al., 2008). Two additional novel strains have recently been isolated from humans and the first one was isolated from an infected human breast implant (Scholz et al., 2010). This strain was named *B. inopinata* and the second strain showed similarity to *B. inopinataand* and was isolated from a patient with chronic lung disease (Tiller et al., 2010). The two most recently described species are *B. papionis*, which was isolated from two baboons with retained placenta (Whatmore et al., 2014) and *B. vulpis* which was isolated in Austria from the mandibular lymph nodes of two red foxes (Scholz et al., 2016) (Table 1).

Inserir nota

**Table 2:** *Brucella* species and biovars

| Rough species* | Biovars | Preferred natural host | Main geographical área | Pathogenicityfor man |
|---|---|---|---|---|
| ***B. ovis*** | | Sheep (males) | Mediterranean coutries | No |
| ***B. canis*** | - | Dogs | USA, South America Central/Eastem Europe | Low |

| Smooth species* | Biovars | Preferred natural host | Main geographical area | Pathogenicity for man |
|---|---|---|---|---|
| ***B. melitensis*** | 1, 2, 3 | Sheep goats Wild ungulates | Mediterranean countries, Middle e Near East | High |
| ***B. abortus*** | 1, 2, 3, 4, 5, 6, (7), 9 | Bovines Wild ungulates | Europe, Americas, Africa, Asia | Moderate |
| ***B. suis*** | 1 | Suids | Americas, Asia, Oceania | High |
| | 2 | Suids, Hares | Central e Western Europe | Very Low |
| | 3 | Suids | USA, China | High |
| | 4 | Reindeer | USA, Canada, Russia | Moderate |
| | 5 | Wils rodents | Russia | High |
| ***B. neotomae*** | | Desert wood rat Neotoma lepida | USA | Unknown |

| | | | | |
|---|---|---|---|---|
| ***B. ceti*** | - | Cetaceans | - | High/Unknown |
| ***B. pinnipedialis*** | - | Pinnipeds | - | High/Unknown |
| ***B. microti*** | - | Common vole | Central Europe | Unknown |
| ***B. inopinata*** | - | Unknown | USA / Oceania | Unknown |
| ***B. papionis*** | - | Baboon | Unknown | Unknown |
| ***B. vulpis*** | - | Red fox | Unknown | Unknown |

*Colony morphology
From: Alton et al. (1988), Joint FAO/WHO Expert Committee on Brucellosis (1986), Whatmore (2009), Whatmore et al., (2014), Garin-Bastuji, 2014, OIE (2016), Rajada, 2016.

### 2.1.1 *Brucella melitensis*

It was suggested, because of the high homogeneity demonstrated by DNA-DNA hybridization studies (Table 2), that the entire genus should be a species (Al Dahouk & Nöckler, 2011), with *B. melitensis* as the only species and the other species should be considered as biovars (Verger et al., 1985, 1987). This was accepted by the Subcommittee on Taxonomy of *Brucella* in 1986 (Al Dahouk et al., 2007), but not yet by the *Brucella* research community.
The complete genome sequence of *B. melitensis*, *B. abortus* and *B. suis* is known, and the average genome size is 3.3 kilobases (Kb), with a GC content of 58-59%.

**Table 3:** Chromosomes statistics of *Brucella melitensis* (GenBank accession numbers NC_003317.1 and NC_003318.1).

| **Molecule Name** | **Type** | **Topology** | **Length** | **%A** | **%T** | **%C** | **%G** | **%AT** | **%GC** |
|---|---|---|---|---|---|---|---|---|---|
| *B. melitensis 16M* Chrom I | chromosome | circular | 2117144 | 21.3 | 21.4 | 28.4 | 28.6 | 42.7 | 57 |
| *B. melitensis 16M* Chrom II | chromosome | circular | 1177787 | 21.3 | 21.2 | 28.5 | 28.7 | 42.5 | 57.2 |

The *B.melitensis* strain 16M, (Table 3) primarily affects goats and sheep, and is the most virulent of the *Brucella* spp. in humans.

**Table 4:** Variants of Brucella melitensis, GenBank taxonomy No.: 224914.

***Brucella melitensis* biovar 1**

*B. melitensis* biovar 1 strain 16M, corresponding to ATCC 23456, is the type strain for this biovar. Strain REV-1 is the rough attenuated vaccine strain of this biovar. *B. melitensis* biovar 1 isolates 78, 87, 91, 113, 219, 256, 261, 376, 391, 392, 393, 400, 401, 402, 415, 450, 456, 457, 458, 461, 462, 485, LAR, and P217 were obtained from human blood and bone marrow samples. *B. melitensis* isolates 279, 280, and 371 were obtained from goat milk samples.

***Brucella melitensis* biovar 2**

Strain 63/9, corresponding to ATCC 23457, is the type strain for this biovar. *B. melitensis* isolate 84 was obtained from human blood and bone marrow samples.

***Brucella melitensis* biovar 3**

Ether strain, corresponding to ATCC 23458, is the type strain for this biovar. *B. melitensis* biovar 3 isolates 254, 255, 257, 258, 259 and 306 were obtained from human blood and bone marrow samples. *Brucella* isolates G914, G1024 and T64/40 also belong to *B. melitensis* biovar 3.

From: Morenoa et al., 2002, Gandara et al., 2001.

### 2.1.2 *Brucella* in mammals

The regions with the highest incidence rates of transmission between humans and animals are Central Asia and the Middle East, but a growing number of cases of human and animal brucellosis have been reported recently in the Balkan Peninsula and sub-Saharan Africa (Pappas, 2010).

The genus *Brucella* can infect a wide range of hosts, including humans, domestic animals and wild animals. In the last two decades, six new species have been discovered and the complexity of the *Brucella* genus has become evident (Scholz et al., 2016; Whatmore et al., 2014; Pappas, 2010).

Despite that each species of *Brucella* has a preferred host, cross-infection between animal species may occur (Corbel, 2006). *Brucella* can persist within macrophages for prolonged periods and can therefore produce chronic and sometimes life-long infections (Rajala, 2016).

#### 2.1.2.1 *Brucella in humans*

The incidence of human brucellosis is reported to be 500,000 new cases each year and is considered one of the most widespread zoonotic infections in the world (Rajala ,2016; Pappas et al., 2006). However, it is believed that the true number of human cases is much higher, since there are many cases that are not diagnosed or not reported to the OIE (Pappaset al., 2006, WHO, 2005). Published data in 2015 suggest that the incidence of human brucellosis exceeds 800,000 cases per year (Kirket al., 2015). It is estimated that about 50% of these cases are due to the ingestion of contaminated foods (Havelaar et al., 2015). In addition, it is estimated that 40% of *Brucella* cases results in chronic infection and 10% of cases results in schizitis in men (Kirket al., 2015).

*B. melitensis* is the most common species of *Brucella* in human diseases, with some estimates suggesting that it accounts for 70% of all infections (Rajada, 2016), is considered with the highest zoonotic potential (Blasco & Molina-Flores, 2011) followed by other species of *Brucella* with lower zoonotic potential *B. abortus* and *B. suis* (biovars 1, 3, 4 and 5) (Whatmore, 2009).

Human brucellosis is endemic in the Mediterranean region of Asia particularly in the Arabian Peninsula and Mongolia, and in North Africa (Hartigan,1997; Pappas et al., 2006). In countries where the disease was controlled or eradicated from domestic animals, there has been a marked decrease in cases of human brucellosis as in the case of the United States and European Union countries. However, there are cases of brucellosis in the United States that are directly associated with consumption of imported animal products, or areas where the disease is endemic (OIE, 2017).

Acccording to the European Centre for Disease Prevetion and Control-ECDC, in a study published in 2017, to improve 2014, 354 confirmed cases of brucellosis were reported by 18 EU/EEA countries, with an overall rate of 0.1 per 100,000 population. Eleven Member States reported zero cases. Greece, Spain and Portugal reported the highest numbers of cases (135, 60 and 50, respectively), corresponding to 69.2% of all cases reported in EU and EEA. Greece had the highest rate, 1.2 per 100,000 population. Figure 1, illustrates the country-specific rates per 100,000 population.

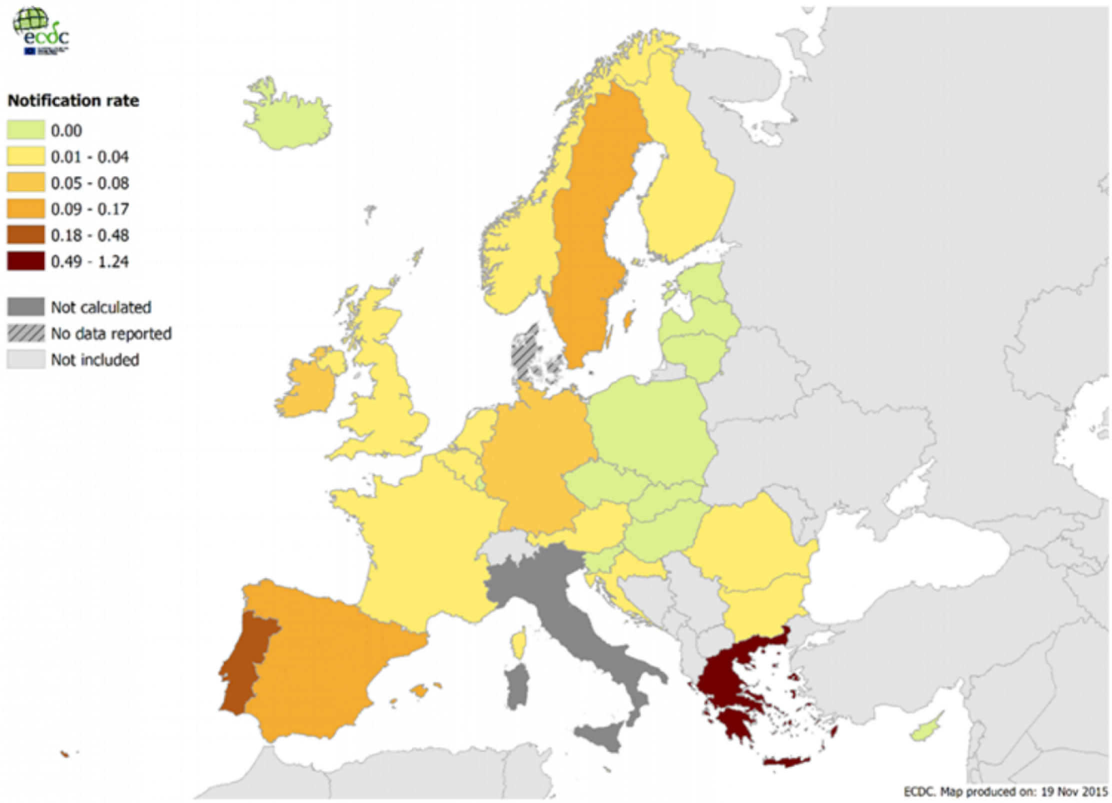


**Figure 1:** Reported confirmed brucellosis cases: rate per 100 000 population, EU/EEA, 2014. Country reports from Austria, Belgium, Bulgaria, Coatia, Cyprus, the Czech Republic, Estoni, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Norway, Poland, Portugal, Romania, Slovkia, Slovenia, Spain, Sweden, the United Kingdom.
From: European Centre for Disease Prevention and Control. Annual epidemiological report 2015. Brucellosis Stockholm: ECDC, 2016.

### 2.1.2.2 *Brucella* in other animals

The most common cause of bovine brucellosis is due to *B. abortus* infection (Godfroid et al., 2010; Whatmore, 2009; Corbel, 2006), but *B. suis* and *B. melitensis* infection may also occur in cattle (OIE, 2017; Corbel, 2006). In sheep and goats, the predominant cause of brucellosis is *B. melitensis* (Godfroid et al., 2010; Whatmore, 2009; Corbel, 2006), although *B. ovis* also infects sheep and has no zoonotic potential. *B. melitensis*, however, is also reported to be common in camels and cattle in some regions with extensive populations of small ruminants. The main cause of brucellosis in dogs is *B. canis* (Whatmore, 2009). However, dogs can also be infected with *B. abortus, B. melitensis* and *B. suis*, due to the consumption of placental or fetal material (Corbel, 2006). Therefore, grazing dogs may constitute a zoonotic risk as well as serve as a transmitter of diseases to livestock.

### 2.1.3 Transmission

The main form of transmission of brucellosis to humans is through direct contact with infected animals, eating or drinking contaminated animal products or inhaling agents transported through the air. Person-to-person transmission is rare (OIE, 2017). Indirect transmission through a contaminated environment can also play a significant role in transmission to humans (Corbel, 2006). The prevalent transmission mechanism in cattle is through direct contact between an infected animal and a susceptible animal (OIE, 2017; Whatmore, 2009; Corbel, 2006) and, as there is a large number of bacteria that is eliminated with fetuses and aborted discharges in pastures, barns contaminating the environment may be important transmission sites (Corbel, 2006).

There are also a number of doubts about the epidemiology of *Brucella* in wild animals, based on results of research on the wildlife infected by *Brucella*. One can raise disturbing questions about the importance of wildlife and how it can act as an important reservoir for transmission of the disease (Godfroid et al., 2010). The implementation of control strategies, once critical, has demonstrated *Brucella* infection in wild animals (Godfroidet al., 2010).

### 2.1.4 Epidemiology of *Brucella melitensis*

*B. melitensis* has been eradicated in some countries but continues to cause significant losses due to declining productivity and the loss of trade in much of the developing world. In the countries free of *B. melitensis,* the cost of surveillance to prevent its reintroduction is significant. There are also concerns that this organism could be used in a bioterrorist attack.

In endemic areas, human brucellosis has serious public health consequences. World wide, *B. melitensis* is the most prevalent species that causes human brucellosis, due in part to difficulties in immunizing goats and free-living sheep. In countries where animal eradication (through vaccination and / or elimination of infected animals) is not feasible, prevention of human infection is primarily based on awareness, food safety measures, occupational hygiene and laboratory safety. In most countries, brucellosis is a notifiable disease (OIE, 2017; CFSPH, 2009).

The most rational approach to preventing human brucellosis is the control and elimination of infection in animals. Milk pasteurization is another protection mechanism. Vaccination of cattle is recommended for the control of bovine brucellosis in enzootic areas with high prevalence. The same goes for caprine and ovine brucellosis. Eradication through testing and slaughter is the way to eliminate brucellosis in regions with low prevalence (OIE, 2016).

The identification of species and biovars of *Brucella* field strains isolated in outbreaks is essential to fully understand the epidemiology of the disease and to trace sources of infection, thereby improving the outcome of brucellosis eradication programmes. It is important to identify the presence of *Brucella* strains in livestock populations and to determine the presence of new strains that might previously have been considered exotic ( Di Giannatale et al., 2008).

*B.melitensis* occurs in the Middle East, some southern and eastern European countries, and parts of Asia and Latin America, including Mexico. It has been found in sub-Saharan Africa, particularly East Africa, but its distribution on that continent is still unclear. This organism is absent from domesticated animals in northern and central Europe, Canada, the U.S.A., Australia, New Zealand, Japan and some other countries. Sporadic cases are occasionally reported in travelers and immigrants in *B. melitensis* - free nations.

### 2.2 Diagnosis and typing methods

Distinction between species and biovars of *Brucella* spp. is currently based on differential tests. Historically, detection and biovar typing of *Brucella* spp.was based on culture and serological methods (Maio, et al., 2104).

### 2.2.1 Culture method: Isolation of *Brucella* spp.

Isolation is considered as the gold standard diagnostic method for brucellosis since it is specific and allows biotyping of the isolate, which is relevant for control of brucellosis using vaccination. In most processes where there is acute infection, after the incubation of the medium for 2-4 days, it is possible to observe small colonies in solid phase that slide through the agar. Isolation of *Brucella spp.* from blood culture is generally the first source of diagnosis of the disease in areas with low incidence. In the case of contaminated samples (abscesses, placental remains, etc.), selective culture media for isolation of *B. melitensis* may be cultured in a variety of selective media, such as Farrell, Thayer-Martin's or CITA media.

### 2.2.2 Serological methods for identifying positive animal/humans

The main serological tests recommended by the OIE (2017) for the diagnosis of *Brucella* spp, indicating specific brucellosis titres, such as serum agglutination test (SAT), Rose Bengal test (RBT), Coombs IgG and enzyme-linked immunosorbent assay (ELISA) are still frequently used.

#### 2.2.2.1 Rose Bengal test (RBT)

It is a dye used as an acidified and buffered antigen for the serological screening of brucellosis by direct agglutination of the serum with the undiluted serum of the patient.

Pink Bengal Antigen is used in screening and positive cases detected should be confirmed by a more specific serological test (OIE, 2017). It provides a diagnostic approach in a few minutes with a very high sensitivity and specificity. It has a high degree of correlation with sero-agglutination and, due to its simplicity, is very useful as initial test or screening test.

#### 2.2.2.2 Standard agglutination test (SAT)

In this test, serial dilutions of the serum to be tested are made for a constant amount of *B. abortus*. This antigen reacts with the two antibodies against *B. melitensis* and *B. suis* (OIE, 2017). The title of seroconversion from 1/160 is considered positive in countries where brucellosis is endemic, and it is possible that the cut-off point in the diagnosis of the disease will vary, with titres not exceeding 1/640 in the early stages of the disease. The results of this procedure require interpretation based on the patient's background and clinical evaluation, since the detection of antibodies varies at the onset of the disease or, in very advanced cases, resulting in a false negative. In this diagnostic test, the antibodies responsible for sero-agglutination, mainly of the IgM class, are usually detected in the course of 3-6 months, with or without cure of the disease.

### 2.2.2.3 ELISA IgG test

The enzyme immunoabsorption assay are techniques used to detect the presence of specific antibodies (IgG and IgM), with excellent sensitivity and specificity. In this technique, polystyrene plates previously treated with the antigen (*Brucella* lipopolysaccharide in the smooth phase) absorbed in the plates (OIE, 2017) are used. IgM antibodies are considered valuable due to their rapid disappearance after the acute phase of brucellosis, whereas IgG antibodies, as they may persist in cured individuals, are commonly detectable. Thus, the use of the ELISA technique, allows to know more accurately the profile of immunoglobulins in the course of the disease. However, the results do not offer the possibility of establishing a criterion to discern between cure and evolution for chronicity (OIE, 2017; CFSPH, 2009).

### 2.2.2.4 Coombs IgG

It is widely used for the diagnosis of chronic brucellosis. This technique is used to demonstrate the presence of binding and non-binding antibodies, primarily IgG (human immunoglobulin) which would be responsible for facilitating agglutination of non-ligand antibodies from the test serum, binding to the antigenic suspension of *B. abortus*. The obtained titre is proportional to the time of evolution of the disease, so high titres refer to a long period since the infection. Even in patients with adequate treatment and clinical evolution, the titre of these antibodies can be very high (CFSPH, 2009). This test may present cross-reactions with *Vibrio cholerae, Francisella tularensis* and *Yersinia serovar enterocolitica* 09.

### 2.2.3 Molecular methods

In addition to diagnosis by cultural and serological methods, brucellosis infection can be detected by specific molecular methods. Molecular techniques have shown accurate typing of *Brucella* spp. based on specific identification of *Brucella* nucleotide sequences associated with the genus, species, and biovars. Therefore, these methods are important tools for diagnosis in epidemiologic studies (Gopaul, KK et al., 2014). The commonly used methods are discussed below.

### 2.2.3.1 PCR and Real time PCR

PCR, including the real-time format, is an additional means of detection and identification of *Brucella* spp. In addition to the speed of the technique, PCR has higher bacterial detection capability of the insulation in various clinical forms (Paixão, 2009).

Despite the high degree of DNA homology within the genus *Brucella*, several molecular methods, including PCR, restriction fragment length polymorphism (RFLP) and Southern blot, were developed to allow the differentiation of *Brucella* species and some of their biovarsas (for a review, see Bricker, 2002, Moreno et al., 2002, Whatmore et al,.2014). Pulse field gel electrophoresis has been developed to allow the differentiation of several *Brucella* species. PCR can satisfactorily identify *Brucella* species and distinguish vaccine strains, but there was limited PCR validation for direct diagnosis.

Alternative approaches allowing identification of all *Brucella* species based on single nucleotide polymorphism (SNP) discrimination by either primer extension or real-time PCR or the ligase-chain-reaction have been described. These tests are rapid, simple, unambiguous, and based on a robust population genetic analysis that helps ensure the species/biovar specificity of the used markers (Whatmore et al., 2014).

### 2.2.3.2 Multiplex PCR

The first species-specific multiplex PCR for *Brucella* differentiation was described by Bricker & Halling, called AMOS-PCR, based on the polymorphism resulting from the specific localization of the IS711 insert sequence on the *Brucella* chromosome and comprised five oligonucleotide primers that could identify, without differentiating, *B. abortus* bv. 1, 2 and 4, but could not identify *B. abortus* bv. 3, 5, 6 and 9.

A new multiplex PCR assay (Suisladder) has been developed for the rapid and accurate identification of *B. suis* strains at the biovar level (Lopez-Goňi et al., 2011). Another advanced multiplex PCR is also able to discriminate between *B. suis* and *B. canis* and between *B. suis* and *B. microti* in only one step, and between the vaccine strains *B. abortus* S19, *B. abortus* RB51 and *B. melitensis* Rev. 1 (Kang et al., 2011). This test could also allow differentiation of the species infecting marine mammals, but this requires additional validation in field lineages. Other tests such as omp25, 2a and 2b PCR / RFLP are available and may be useful for identifying some species of *Brucella*.

### 2.2.3.3 Pulsed field gel electrophoresis (PFGE)

PFGE is a suitable technique to separate large DNA fragments by reorientation of DNA in agarose gel from the effects of alternating electric fields. This technique is considered as a gold standard in molecular typing for most bacteria and the possibility of discriminating genetic diversity, genetic distance between lineages, location of sources of contamination and distribution of organisms and epidemiological studies has been clearly demonstrated among *Brucella* species (Jun et al., 2013). This change in the electric fields causes the rearrangement of the conformational structure of the molecule allowing migration to the gel and the agarose keeps the DNA molecules intact and at the same time allows and diffuses the detergent and the protease (Bahmani, et al., 2017). This movement obeys the following principle: when an electric field is applied to the gel, the DNA molecules extend in the direction of the field and migrate in the gel. When the first field is removed and a second one is applied relative to the first, the DNA molecule needs to change its conformation and orientation before migrating toward the second electric field (Bahmani, et al., 2017; Patil et al., 2014). The time required for this reorientation to occur is proportional to the molecular weight of the fragment, the reorientation of the larger molecules being longer than that of the smaller molecules.

### 2.2.3.4 Multilocus sequence typing (MLST)

The development of DNA sequencing technologies has triggered the development of computational methods that can beuseful for the rapid detection of epidemiological information these methods include a multilocus sequencing scheme (Whatmore & Gopaul, 2014).

This method is based on the measurement of the DNA sequence variations in a set of highly conserved housekeeping genes and characterizes strains by their unique allelic profiles. Approximately 450-500 bp internal fragments of each gene are used, as these can be accurately sequenced on both strands using an automated DNA sequencer. For each house-keeping gene, the different sequences present within a bacterial species are assigned as distinct alleles and, for each isolate, the alleles at each of seven reference *loci* define the allelic profile or sequence type (Urwin and Maiden 2003Each isolate of a species is therefore unambiguously characterised by a series of seven integers which correspond to the alleles at the seven house-keeping *loci* (Urwin and Maiden, 2003; Maiden et al, 1998). In MLST, the number of nucleotide differences between alleles is ignored and sequences are given different allele numbers whether they differ at a single nucleotide site or at many sites.

According to Urwin and Maiden (2003), most bacterial species have sufficient variation within house-keeping genes to provide many alleles per *locus*, allowing billions of distinct allelic profiles to be distinguished using only seven house-keeping *loci* (Maiden et al.,1998). The MLST is based on the well-established principles of multilocus enzyme electrophoresis, however, it differs in that it assigns alleles at multiple home maintenance sites directly by DNA sequencing, rather than indirectly through the electrophoretic mobility of its gene products. The advantage of MLST is that sequence data are unambiguous and the allelic profiles of isolates can easily be compared to those in a large central database.

These sequence typing methods can be performed using both classical sequencing techniques (e.g. Sanger sequencing) and next-generation sequencing (NGS) but in the last decade, increasingly, MLST is being performed using NGS data due to the lower cost and it being much less time consuming.

### 2.2.3.5 Multi-locus variable number of tandem repeat analysis (MLVA)

The limitations of common typing techniques, for example failure to discriminate among biovars within the same species, stimulated the development of additional molecular typing techniques such as the Multiple Locus Variable Number Repeat Tandem (VNTR) Analysis (MLVA), which is a method used to perform molecular typing of specific microorganisms, utilizing variation occurring naturally in the number of tandem repeat DNA sequences found at many different *loci* in the genome of a variety of organisms (NIPHE - National Institute for Public Health and Environment – Ministry of Health, Welface and Sport of Netherlands, 2019). The method originates from forensic science, where it is used for fingerprints of DNA in samples of human origin. MLVA is also widely used to evaluate the molecular fingerprint of microorganisms, such as bacteria. Molecular typing profiles are used to study transmission pathways, to assess sources of infection, and to assess the impact of human intervention, such as vaccination and the use of antibiotics in the composition of bacterial populations (NIPHE, 2019).

MLVA measures the number of tandem repeats at a given *locus* and can differentiate between isolates within a given *Brucella* biovar **(**Maio, et al., 2104**)**. This method is based on PCR amplification of tandem repeats and boarding consensus regions, followed by amplicon separation and length measurement. The number of repeat units can be deduced from the measured amplicon size. Subsequently, genotypes can be assigned according to the gain or loss of discrete repeats, which leads to better insights into the genetic relationships between bacterial strains (BDN, 2018;EC, 2016). Multilocus VNTR analysis possesses the advantages of extreme resolving power, robustness, high throughput data port-ability, ease of data interpretation and concordance with epidemiological data (BDN, 2018). Another advantage is that, unlike pulsed-field gel electrophoresis (PFGE), MLVA is a PCR-based approach that requires only a small amount of DNA for the analysis.

## The role of WGS in the Epidemiological Investigation

The high degrees of genetic diversity within each species was already discovered by PFGE showing that itis possible to observe a significant size variability among genomes of the same species (Bergthorsson and Ochman, 1995; Thong et al., 1995). However, recent studies based on whole genome sequence comparison of isolates of the same species revealed an even much higher degree of intra-species variability than expected (Mira et al., 2010; Laing et al., 2011).

Events of mutations, insertions or deletions, genome rearrangements and transfer of exogenous DNA, in addition to extrachromosomal elements such as plasmids and phages, acts as a driving force for prokaryotic genome plasticity.

Furthermore, it is known that bacteria acquire a large percentage of their genetic diversity by gene acquisition through horizontal gene transfer (HGT); at the heart of HGT events there is a large variety of mobile genetic elements (MGEs) that confer to bacteria a rapid evolution and a high capacity for adaptation (Aminov, 2011).

The result of this complex evolutionary dynamics based on acquisition, loss or duplication of genomic elements, produce a high variability of sequences present in  members of a species and pose enormous challenges in the reconstruction of evolutionary relationships among bacterial isolates.

Due to this peculiar evolutionary process, many of the biological markers used in the traditional and molecular epidemiology, which are used to detect mainly those elements that are conserved and relatively stable among members of a given species, produce results not generally useful for finding differences among closely related strains, which are occurring mainly by events of horizontal transfer of genes.

This led to the development of different approaches able to allow a higher level of discriminatory power compared to conventional methods of molecular typing (Miller. 2016).

To date, the systems of epidemiological surveillance and outbreak detection are in constant evolution, directed to the development of new methods characterized by specific attributes such as high sensitivity and specificity, flexibility or timeliness, able to identify and characterize the pathogen responsible for an outbreak, but also useful to understand the origins and dynamics of the outbreak.

With the improvement of massively parallel DNA sequencing technologies, the real-time sequencing of entire pathogen genomes is now possible (Reuter et al., 2010; N.J. Loman et al., 2012). In contrast with genotyping, where only a small fraction of the pathogen genome is used to infer phylogenetic relationships, the whole genome of the pathogen can be used to resolve the transmission dynamics of an outbreak in much greater detail (N.J. Croucher et al., 2015).

These new sequencing technologies, along with the associated bioinformatics algorithms, have given rise to the field of genomic epidemiology where whole-genome analysis methods are integrated with traditional molecular diagnostics and genotyping methods to yield the ultimate resolution into outbreaks and epidemiologic investigations.

### 2.3.1 Application of WGS for bacterial typing and epidemiologic analysis

Recent studies have shown that WGS analysis is a powerful tool in molecular epidemiological research on infectious diseases, providing a better outcome compared to other techniques such as MLVA or MLST (Georgi et al., 2017; Pearce ME et al., 2018; Sun M et al,. 2017; (Janowicz et al., 2018). Thus, WGS emerges as an important tool to be widely used in the study of *Brucella* spp. because it provides excellent results in typing the entire genome of the bacterium (Garofolo et al., 2013).

In the genomic epidemiology, the variability between strains is translated mainly into measures of distance by determining single nucleotide polymorphisms in genome alignments (SNPs analysis) or by indexing allelic variation in hundreds to thousands of core genes, assigning types to unique allelic profiles (Gene-by-Gene analysis).

#### 2.3.1.1 Single Nucleotide Polymorphism (SNP) analysis

Because, in the short time-frame typical of an outbreak, the genomes of isolates within a cluster are expected to be highly related, the most common way to compare these genomes is to examine the differences in small and frequent changes, often focusing the variant detection on the identification of single nucleotide polymorphisms present between isolates.

This approach consists of finding/calling fixed nucleotide variants in the organisms in question and is based on the theory that random single mutations will happen independently over time throughout the genome of an organism. The amount of SNPs differences between two organisms will define the genomic distance between them and in turn define their relationship.

Unlike of gene-by-gene methods, approaches based on SNPs are more flexible as they do not require a predefined scheme.

Generally, in order to define differences, a common reference is used, to which these differences can be referred, so that the DNA sequence of each isolate is mapped to the reference genome to define the “SNP-calling”. Hence, a reference genome should be carefully selected or constructed. Ideally, often the analysed isolates are also very closely related, which is the case in an outbreak. In this setting, one genome out of a set of closely related samples may be sequenced and assembled, and then used as reference against which all others can be compared.

When a common reference is not available, two are the main strategies used to define a set of nucleotide variants present within a group of isolates, respectively called “core” and “whole-SNPs analysis”. In the first approach the SNPs panel is defined using only the nucleotides variants found inside genome regions shared among all investigated isolates, otherwise, in the whole-SNPs approach, it is required that at least two isolates have to share a region containing a *locus* hosting a nucleotide variant.

The SNP approach is much more discriminatory than approaches based on the study of a few gene sequences such as MLST because it relies on sequence differences in many more regions of the genome, thus offering a wider overview of the genomic distance between the investigated strains.

#### 2.3.1.2 Gene-by-Gene analysis: Core genome MLST (cgMLST)

In contrast to whole genome SNP analysis, the gene-by-gene comparison methods are based on the concept of allelic variation, meaning that recombinations and deletions or insertions of multiple positions are counted as single evolutionary events. This approach might be biologically more relevant than approaches that consider only point mutations especially in long-term epidemiological studies.

The cgMLST is a typing approach similar conceptually to the classic MLST methods, since both are based on the gene-by-gene comparison used to define the bacteria with a specific allelic profile. The main difference between these two genotyping techniques lies in the higher degree of discriminatory resolution given by the marked extension of the number of analyzed genes, from the restricted number of *loci* in a MLST scheme to several hundreds or even >1,000 genes used for the definition of the core genome profile.

The technique is based on the concept that a large proportion of the core genome consists of genes that play crucial roles in maintaining basic cellular functions, such as housekeeping and regulatory genes, and these genes can be regarded as relatively "stable" in comparison with accessory genes that are more frequently horizontally transferred (Hervé et al., 2005).

Furthermore, several studies have also demonstrated that signals of horizontal transfer in the core genome are present in different bacterial species (Michiel Vos and Xavier Didelot, 2009).

Identifying the core-genome from a collection of bacterial genomes typically relies on classifying genes into orthologous clusters based upon sequence similarity searches.Generally, the approach used to define the core genome panel is highly restrictive, and the size of the core genome depends on the number of genes per genome that are shared among all investigated isolates (strict core). Sometimes the stringency may need to be adjusted on the basis of different requirements, depending on the species diversity, input data quality and on the specific questions that the research is aiming at answering.

Based on these considerations, the delineation of the core genome can be relaxed from strict core to soft core, which correspondingly comprises genes shared by the majority of the strains, using a more soft definition.

Although much phenotypic variation can be explained by examining the accessory genome, however, many researchers feel that selectively neutral changes in the core genome, such as synonymous mutations in codons, represent a molecular clock that provides a more accurate record of strain evolution, useful for accurately inferring phylogenetic relationships (Foster et al., 2009).

## 3. OBJECTIVES

The overall aim of this thesis was to explore the potential of Whole Genome Sequencing (WGS) for the tracing of a pathogen outbreak. In particular, we developed a cgMLST scheme for *B. melitensis* and assessed the performance of cgMLST and a whole-genome SNP-based approach against the traditional MLVA-16 typing method using a set of animal outbreak-associated isolates and a set of isolates with unknown epidemiological status. The subject is complex, since several strains and biovars could be involved, and establishing clear phylogenetic relationships in order to trace the disease outbreak could be over the limits of the most common approaches. WGS holds much promises for this kind of analysis, providing a more complete and unbiased view of the relationships among the different isolates. We aim at implementing and comparing different approaches, test their reliability, evaluate the advantages and shortcomes, and extrapolate rules for their general applicability.

# 4 MATERIALS AND METHODS

## 4.1 Profile of *B. melitensis* strains analyzed

To evaluate the MLVA/WGS approach, based on epidemiological criteria, the isolates we analyzed were separated into two different groups and the obtained results were compared with those of MLVA-16.
The first group consisted of 37 strains of *B. melitensis* isolated during a single outbreak on 21 farms in the provinces of Frosinone, Rome, Isernia and Campobasso, central Italy (Fig. 2A).

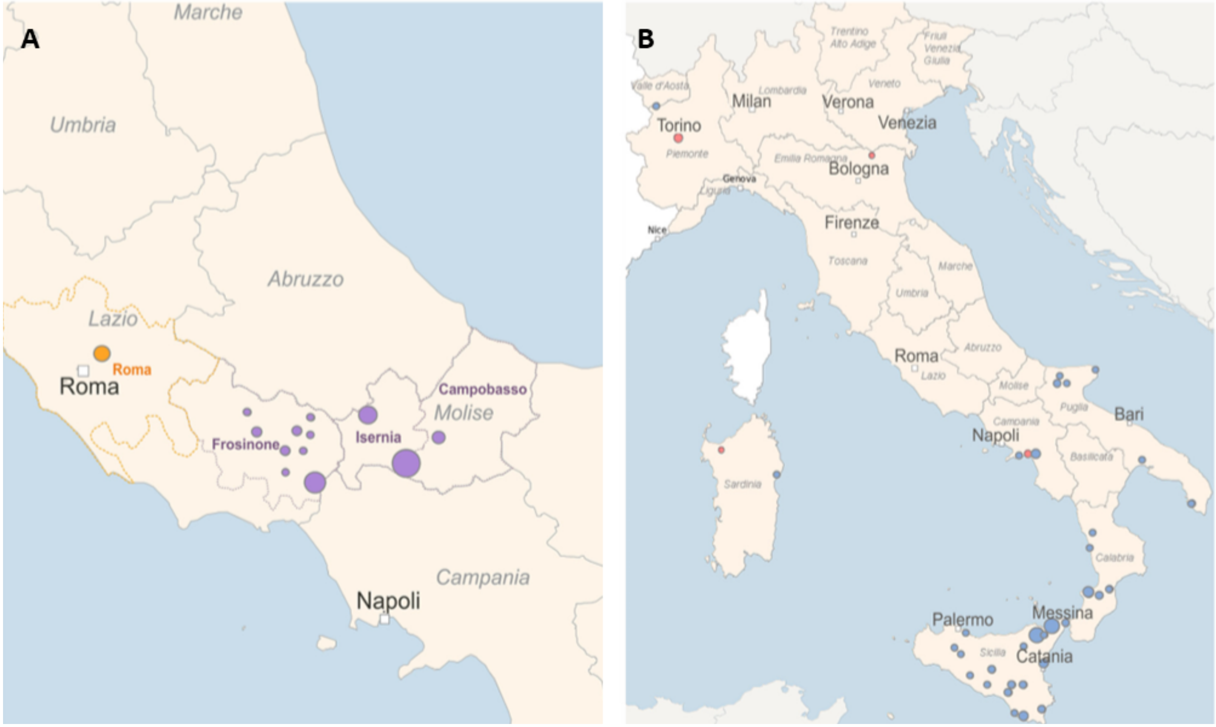


**Figure 2:** Geographical map for *B. melitensis* cases studied. **(A)** epidemiologically related isolates. Separate epidemiological clusters are marked with different colors respective to the provinces of isolation (purple, Frosinone, Isernia, and Campobasso; orange, Rome). **(B)** Isolates with unknown epidemiological status **(A)**. The red circles correspond to human isolates and the blue circles to animal isolates.
Janowicz et al., 2018.

The second group consisted of 64 *B.melitensis* isolates with unknown epidemiological status, collected in Italy from infected animals between 2011 and 2017 during the activities of the national eradication program, and two related and unrelated strains of *B. melitensis* isolated from humans cases. Figure 2B shows the geographical origin of these samples.

## 4.2 Isolation of *B. melitensis*

The samples from collected and inoculated animals were obtained from lymphatic glands (that is, mandibular, supramammary and genital lymph nodes), spleen, uterus or udder,whereas human isolates were obtained directly from blood culture.

Samples of *B. melitensis* were inoculated on sterile plates of *Brucella* selective agar containing serum agar dextrose, Hemin and Vitamin K1 media (Hi Media, India) and incubated at 37°C for 48 hours. The plates were observed at every 24 hours for the development of growth. After obtaining the growth, the colonies suspected for *Brucella* on the basis of cultural characteristics were selected and streaked again on plates containing *Brucella* selective agar with Hemin and Vitamin K1 and incubated at 37°C for 2 days to obtain the pure culture, following the standard protocol of the OIE (World Animal Health Organization Handbook - NB: Version adopted in May 2017).

### 4.2.1 Procedures for identifying colonies

Cultures showing typical *Brucella* characteristics were subjected to biotyping techniques such as H2S production, growth in the presence of thionin, and basic fuchsin (10–40 µg/mL) dye incorporated into tryptic soy Agar at different concentrations and $CO_2$ requirement immediately after the primary isolation, as previously described (Huddleson et al., 1931). Lead acetate strips were used to identify the production of $H_2S$ during growth, and the growth was evaluated on media containing streptomycin (2.5 µg/mL) to discriminate the isolates from vaccine strain Rev1, as previously described (OIE, 2017). Epidemiological data are presented in Table 5 (Annex 1).

## 4.3 Molecular identification

### 4.3.1 Extraction

DNA from the *B. melitensis* strains was extracted using the Maxwell 16 tissue DNA purification kit (Promega Corporation, Madison, WI) according to the manufacturer's instructions. All DNA samples extracted from the isolates were stored at - 80°C.

### 4.3.2 MLVA

Samples were genotyped using the MLVA-16 panel described by Le Flècheet et al. (2006) and Garofolo et al. (2013). Primers used for the MVLA reaction correspond to the 16 *loci* of the 16M genome of *Brucella melitensis* and are targeted to four markers - bruce04, bruce06, bruce16 and bruce21 - modified to provide longer amplicons, ensuring the absence of overlap with VNTR loci (Garofalo et al., 2013b). The MLVA primers for the 16 *loci* and fluorescent dyes used in capillary electrophoresis (EC) are given in Table 4.

**Table 5: MVLA primers used in each multiples reaction.**

| | | Locuss | Primersequences (5′ to 3′) | Primer [] | Allele size range (bp) |
|---|---|---|---|---|---|
| CE 1 | Multiplex1 | Bruce 30 | F: PET- TGACCGCAAAACCATATCCTTC<br>R:TATGTGCAGAGCTTCATGTTCG | 0.2µM | 119-199 🟥 |
| | | Bruce 08 | F: PET-ATTATTCGCAGGCTCGTGATTC<br>R: ACAGAAGGTTTTCCAGCTCGTC | 0.2µM | 312-384 🟥 |
| | | Bruce 11 | F: 6FAM-CTGTTGATCTGACCTTGCAACC<br>R: CCAGACAACAACCTACGTCCTG | 0.2µM | 257-1076 🟦 |
| | | Bruce 45 | F: 6FAM-ATCCTTGCCTCTCCCTACCAG<br>R: CGGGTAAATATCAATGGCTTGG | 0.2µM | 133-187 🟦 |
| | | Bruce 19 | F: NED-GACGACCCGGACCATGTCT<br>R: ACTTCACCGTAACGTCGTGGAT | 0.2µM | 79-205 🟨 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
|  | Multiplex2 | Bruce 06 | F: NED-GATTGCGGAACGTCTGAACT<br>R: TAACCGCCTTCCACATAATCG | 0.2µM | 312-714 🟨 |
|  |  | Bruce 42 | F: VIC-CATCGCCTCAACTATACCGTCA<br>R: ACCGCAAAATTTACGCATCG | 0.12µM | 164-914 🟩 |
| **CE 2** | Multiplex3 | Bruce 12 | F: NED-CGGTAAATCAATTGTCCCATGA<br>R: GCCCAAGTTCAACAGGAGTTTC | 0.2µM | 302-452 🟨 |
|  |  | Bruce 18 | F: PET-TATGTTAGGGCAATAGGGCAGT<br>R: GATGGTTGAGAGCATTGTGAAG | 0.2µM | 130-186 🟥 |
|  |  | Bruce 55 | F: PET-TCAGGCTGTTTCGTCATGTCTT<br>R: AATCTGGCGTTCGAGTTGTTCT | 0.2µM | 193-553 🟥 |
|  |  | Bruce 21 | F: 6FAM-CTCATGCGCAACCAAAACA<br>R: GTGGATACGCTCATTCTCGTTG | 0.2µM | 431-463 🟦 |
|  |  | Bruce 04 | F: VIC-CTGACGAAGGGAAGGCAATAAG<br>R: TGGTTTTCGCCAATATCAACAA | 0.2µM | 313-473 🟩 |
| **CE 3** | Multiplex4 | Bruce 07 | F: NED-GCTGACGGGGAAGAACATCTAT<br>R: ACCCTTTTTCAGTCAAGGCAAA | 0.2µM | 134-246 🟨 |
|  |  | Bruce 09 | F: VIC-GCGGATTCGTTCTTCAGTTATC<br>R: GGGAGTATGTTTTGGTTGTACATAG | 0.2µM | 124-292 🟩 |
|  |  | Bruce 43 | F: 6FAM-TCTCAAGCCCGATATGGAGAAT<br>R: TATTTTCCGCCTGCCCATAAAC | 0.2µM | 170-194 🟦 |
|  |  | Bruce 16 | F: 6FAM-ACGGGAGTTTTTGTTGCTCAAT<br>R: GGCCATATCCTTCCGCAATA | 0.2µM | 227-353 🟦 |

F: forward R: reverse CE: capillary electrophoresis. Expected allele size ranges are given in base pairs, each marked by its corresponding fluorescent dye. As can be seen, the test has been designed so that fragments would differ from one another by either size, fluorescence or both, to exclude the possibility of overlap in EC results. Source: Garofolo et al., 2013.

Briefly, to assign specific alleles, DNA extracted from each isolate was amplified by multiplex PCR using primers specific for each MLVA-16 *locus*. The PCR amplifications were performed in a total volume of 10 µl, containing 1x Type-it Multiplex PCR Master Mix (Qiagen), 0.5x solution buffer, and proper concentration of each fluorescent primer pairs according to Garofolo et al., (2013) (Table 4) and 5 to 10 ng DNA.

The thermocycling conditions were as follows: 96 °C for 5 min. followed by either 30 (for multiplex 1, 3 and 4) or 24 cycles (for multiplex 2) of: 95 °C for 30 s, 60 °C for 90 s and 72 °C for 30s; followed by 60 ° C for 30 min. Multiplex 2 was run for 24 cycles in order to contain VNTR amplification artifacts. For each strain, diluted MLVA PCR products (1: 225 in deionized water) with 0.25µl of LIZ 1200 size standard were subjected to CE on an ABI Prism 3500 Genetic Analyzer with POP-7 (Applied Biosystem Inc.). Reactions 1 and 2 were mixed in the EC, so that only three injections of EC were required to analyze the products of the four multiplex PCRs. The VNTR fragments were sized by Gene-mapper 4.1 (Applied Biosystems Inc.).

A phylogenetic tree was generated using the goeBURST algorithm in PHYLOViZ software to identify clonal complexes and founder MLVA types among the 71 isolates with unknown epidemiological status. MLVA-types were compared with each of the four VNTR *loci* and genetic relatedness between the strains was assessed using goeBURST version 1.2.1 (http://goeburst.phyloviz.net/) (Francisco et al., 2012).

The goeBURST algorithm identifies mutually exclusive groups of related MLVA types in a population. The algorithm also predicts the presumed founder(s) of each clonal complex and any single *locus* variant (SLV) and double *locus* variant (DLV) derivatives.

The primary founder of a group is defined as the MLVA type that has the greatest number of SLVs. goeBURST then constructed a spanning forest in which each MLVA type is a node and two MLVA types are connected if they are SLVs.

## 4.4 Whole-genome sequencing

### 4.4.1 Quantification of genomic DNA

Total genomic DNA was quantified with the Qubit Fluorometer (QubitTM DNA HS assay; Life Technologies, Thermo Fisher Scientific, Inc.), approximately 1 to 5µg of genomic DNA extracted from each isolate was sheared in a SonicMan microplate sonicator (Brooks Automation, Chelmsford, MA, USA) to produce fragments averaging 600 bp in length.

### 4.4.2 Library preparation

Library preparation was performed using the Nextera XT library Preparation kit (Illumina Inc., San Diego, CA) or Kapa high-throughput library preparation kit (KapaBiosystems, Wilmington, MA) according to the manufacturers' instructions. The fragment size distribution was also confirmed with an Agilent DNA high-sensitivity kit for the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).
The libraries were sequenced using the Illumina NextSeq 500 platform, producing 150-bp paired-end reads, or Illumina MiSeq, producing 300-bp paired-end reads.

## 4.5 Whole-genome sequencing data analysis

A bioinformatics pipeline has been implemented for the analysis of the data produced with the two Illumina platforms. The workflow for the pre-processing analysis has been set *ad hoc* for bacteria and we have fixed specific threshold values used for the quality evaluation of the samples.
During the implementation of our pipeline, one of the main aspects concerns the potential contamination of the data. A contamination in NGS data could be observed even when the sequencing was performed with stringent wet-lab protocols (Laurence M. et al., 2014) and this could have an impact on the results, mainly on the cluster analysis.
The steps taken and tools used during data analysis are outlined below (Figure 3).

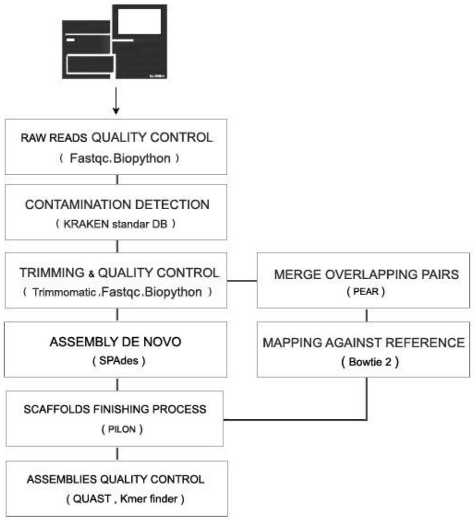


**Figure 3: Bioinformatic pipeline frameworks -The drawing below depicts the processes, tools, and data flows within the pipeline used for quality control and data refinement of raw reads obtained by sequencing. Details of the analyses are described later in individual paragraphs.**

### 4.5.1 Preliminary samples selection

This part of the work describes the control strategies that have been used for the evaluation of the data produced by the sequencing, with the objective of a preliminary selection of the samples subsequently used in the cgMLST and SNPs analysis.

#### 4.5.1.1 Quality check of raw-data

A preliminary control of the sequences dataset is an important step needed to evaluate the data obtained from the sequencing process and to perform a first step of samples selection.

With this aim, we have fixed three specific threshold values: theoretical coverage, Q30 and mean length of the raw reads.

The theoretical coverage (or depth), can be defined as the number of unique reads that include ("cover") a given nucleotide in the reconstructed genomic sequence. The threshold value for theoretical coverage was set as 40X, this depth of sequencing allows us to estimate an adequate final coverage threshold after the trimming process. Moreover, a relatively high depth value is useful to preserve a high enough real coverage in the presence of reads contamination. The coverage was calculated using the size of the reference genome (GenBank accession numbers NC_003317.1 and NC_003318.1).

Q30 is defined as the percentage of reads with mean quality at least of 30 Q-score (Phred scale). The Q30 threshold was set as at least 50. On the basis of our previous experiences, we estimated this value as sufficient to obtain a high Q30 after the process of trimming without a significant decrease of coverage.

Lastly, mean length was set as at least 100 nucleotides for the samples sequenced with the Illumina NextSeq 500 and at least 200 nucleotides for the samples sequenced with the Illumina MiSeq The reads length distribution control is useful to check for the presence of artifact sequences inside our dataset.

These preliminary analyses was performed with the bioinformatic tool FASTQC (version 0.11.6) and the Biopython library (version 1.72), and the results are shown in the Table 6 (Annex 2) and Table 7 (Annex 3).

#### 4.5.1.2 Contamination detection

The raw sequences obtained from the sequencing process may contain DNA from sources other than the sample. Those sequence contaminations are a serious concern to the quality of the data used for the downstream analysis, causing misassembly of sequence contigs and erroneous conclusions. Therefore, the detection of sequence contaminants is a necessary and required step for all sequencing projects.

For this process, each raw-reads dataset was classified with Kraken (version 2.0) (Wood & Salzberg, 2014), with the aim of obtaining a taxonomic classification of the samples. Raw-reads sequences were analyzed with the standard Kraken database, using default parameters as run conFiguretion.

With this aim, we defined as not contaminated all raw-reads datasets characterized by at least 80% of sequences classified as genus *Brucella.*

The results of the analysis are shown in the Table 7 (Annex 2) and Table 9 (Annex 4).

Each sample that has not passed these preliminary checks was discarded. When it was possible, the sample discarded was re-sequenced and checked again.

### 4.5.2 Data pre-processing: Trimming and merging of paired-end raw reads

Quality trimming at the beginning and, particularly, at the end of the reads, where the quality tends to drop, is necessary to obtain a high quality dataset.

The aim is to discard low quality portions while preserving the longest high quality part of a NGS read and to remove, when are present, residual portions of adapters used during the sequence process.

The paired-end raw reads were trimmed using Trimmomatic (version 0.33) (Bolger AM et al., 2014) and was performed with following parameters: SLIDINGWINDOW:4:15 MINLEN:50 AVGQUAL:28.

The samples with a theoretical coverage <40X were re-sequenced or discarded, and the results of this analysis are shown in Table 7 (Annex 3).

After the trimming, we performed a second finishing process starting by the trimmed reads with the aim of obtaining a further high quality dataset of reads used in the following step of *De novo* sequences refinement.

This finishing process, generally called merging, is based on the concept that the raw-reads datasets, produced by several current protocols able to generate sequences from both ends of a library of DNA fragments, are composed of a percentage of paired-end reads with sequence overlap. This occurs when the starting fragments are shorter than twice the read length and the resulting paired-end reads will be partially overlapped (Magoc et al,. 2011). One of the main effects

Nenhuma entrada de índice de ilustrações foi encontrada.obtained by the merging is to produce a further increase in the mean quality of the stitched reads.

The merging process was performed with the Pear software (Zhang J et al., 2013) using default parameters as run conFiguretion. The overlap degree of our dataset of reads is shown in Table 7 (Annex 3).

### 4.5.3 Genome reconstruction: *De novo* assembly and finishing process

After pre-processing, trimmed paired-end reads were used to make a scaffold-level assembly using SPAdes (version 3.11) (Bankevich et al., 2012).

We used the “careful” parameters set, recommended for the assembly of bacteria genomes, with the aim of reducing the number of mismatches and short indels. In addition, the analysis was set with a specific list of *k*-mer sizes according to the library preparation kit used in the sequencing phase, choosing *k*-mers with size 21, 33, 55, 77 for the data obtained with the Illumina NextSeq500 platform and 21, 33, 55, 77, 99, 127 for the data obtained with the Illumina MiSeq.

Scaffolds were further filtered, discarding sequences with length < 200 nucleotides.

Lastly, we performed a refinement of the sequences obtained by the *De novo* assembly using the PILON software (version 1.22) (Bruce J. et al., 2014). In this procedure, the dataset of reads obtained after the merging step are mapped against the scaffolds produced by the genome assembler, and filtered with Bowtie2 (version 2.3.4.1) (Langmead B et al., 2012); the obtained BAM files are then used to finish (i.e. error correction, gap filling, etc.) the *De novo* assemblies. The obtained results are showed in the Table 8 (Annex 5).

### 4.5.4 Assemblies quality control

A last check to evaluate the quality of the genomic assemblies obtained with SPAdes was attempted using two tools: QUAST (version 5.0.1) (Gurevich A et al., 2013) and KmerFinder (version 3.0) (Hasman H et al,. 2014) using default parameters as run conFiguretion.

The reference genome (NC_003317.1 and NC_003318.1) was the template in the KmerFinder database. The obtained results are showed in the Table 9 (Annex 6 ).

### 4.5.5 cgMLST analysis

To determine the cgMLST gene set, we performed a genome-wide gene-by-gene comparison using the cgMLST Target Definer (version 1.4) function of the SeqSphere + software (version 5.0.90) (Ridom GmbH, Münster, Germany). The sequences obtained by the scaffolds finishing process were used as input of this analysis.

The parameters used include the following filters to exclude certain genes of the *B. melitensis* bv. 1 strain 16M reference genome (NC_003317.1 and NC_003318.1) from the cgMLST scheme: a minimum length filter that discards all genes shorter than 50 bp; a start codon filter that discards all genes that contain no start codon at the beginning of the gene; a stop codon filter that discards all genes that contain no stop codon or more than one stop codon, or if the stop codon is not at the end of the gene; a homologous gene filter that discards all genes with fragments that occur in multiple copies within a genome (with identity of 90% and more than 100-bp overlap); a gene overlap filter that discards the shorter gene of a pair of genes from the cgMLST scheme if the two genes are overlapping by >4 bp. The remaining genes were then used in a pairwise comparison using BLAST, version 2.2.12 (the used parameters were the following: word size, 11; mismatch penalty, −1; match reward, 1; gap open costs, 5; gap extension costs, 2), with the query chromosomes of one representative for each of the other two *B. melitensis* biovars (*B. melitensis* bv. 2 strain 63/9 [NZ_CP007788.1 and NZ_CP007789.1] and *B. melitensis* bv. 3 strain Ether [NZ_CP007761.1 and NZ_CP007760.1]) (Camacho et al., 2009).

Using all genes of the reference genome that were common in all query genomes, with a sequence identity of ≥90% and 100% overlap and with the tart codon filter, stop codon filter, and stop codon percentage filter turned on, the final cgMLST scheme was formed. Therefore, all genes having no start or stop codon in one of the query genomes, as well as genes that had internal stop codons in more than 20% of the query genomes, were discarded.

### 4.5.6 SNP analysis

SNPs were identified using *In Silico* Genotyper (ISG), version 0.16.10-3 (Sahl et al., 2015). We used default filters to remove SNPs from duplicated regions, minimum quality was set to Phred 30, and the minimum allele frequency was set to 90% in all samples. We used the ISG pipeline with BWA-MEM (version 0.712-r1039) (Li et al., 2013) as the aligner.

GATK (version 3.9) (Auwera et al, 2013) was then used as the SNP caller and to determine ambiguity of SNPs from the BAM files. The SNPs were called based on the alignment of the trimmed reads against the reference *B. melitensis* bv. 1 strain 16M (NC_003317.1 and NC_003318.1).

### 4.5.7 Simpson’s Index of Diversity

To compare the performance of two WGS-based typing methods, SNP analysis and cgMLST with the gold standard MLVA-16, we used the Simpson’s Index of Diversity (SDI), a numerical index of being characterized as the same type using different analytical approaches (Hunter et al., 1988). This index is given by the following equation:

$$D=1-\frac{1}{N(N-1)}\sum_{j=1}^{S} n_{j}(n_{j}-1)$$

where $N$ is the total number of strains in the sample population, $s$ is the total number of types described, and $n_j$ is the number of strains belonging to the *j*th type. The probability that a single strain sampled at random will belong to the *j*th group is $n_j/N$. The probability that two strains sampled consecutively will belong to that group is n.(nj - 1)IN(N - 1). These probabilities can be summed for all the described types to give the probability that any two consecutively sampled strains will be the same type. This summation can be subtracted from 1 to give the equation above. The higher the SDI value we obtain for a tested method, the greater its discriminating power.

# 5. RESULTS

The aim of this Ph.D. project was to investigate and develop new approaches for epidemiological studies based on WGS data, at first focusing on the requirement for methods and metrics useful to improve the data obtained by next generation sequencing, but manly on evaluating the results obtained by the comparison of several analysis techniques.
Specifically, two subjects have been investigated and are discussed in the following chapters:

(i) Evaluation criteria and results obtained on WGS data by our bioinformatics pipelines.

(ii) Results and different discriminatory power observed using MLVA, cgMLST and SNPs analysis as phylogenetic approaches.

To compare whole genomes and identify differences with sufficient certainty and accuracy, the quality of the underlying data is crucial. Here, we describe the results of the quality metrics used to assess the quality control procedures. The results discussed in the following chapters concern only the samples used in the phylogenetic studies, which have passed the preliminary samples selection previously described.

## 5.1 Read libraries: Contamination and analysis of quality metrics

With the aim of identifying possible contaminations in our WGS data, a metagenomic approach was used for the rapid taxonomic characterization of raw read libraries obtained by the whole genome.
Considering the limits of this approach in resolving power at species-level, taxonomic classification was performed at genus-level.
The results shows that more than 90% of the total reads from each isolate can be assigned to the genus *Brucella*. Of the remaining 10%, most of these sequences were defined as unclassified while, on average, less than 1% was assigned to some other genus (as shown in Figure 4 and in Table 7 annex 3).
We have decided to not remove potential contaminant sequences because that, while the removal process of potential contaminants has the advantage of producing smaller and more homogenous data sets, this could nevertheless bear the risk of removing genomic sequences from the target organism such as, for example, repeated elements or regions resulting from horizontal gene transfers, thus leading to lower quality genomic assemblies. Overall, the results obtained from this analysis suggest that, if contaminations occurred, their entity can be negligible.

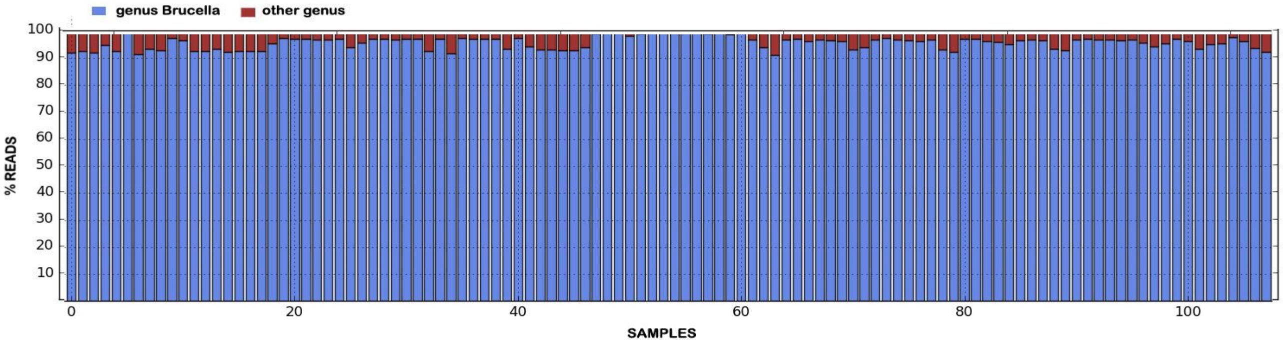


**Figure 4: Distribution of taxonomic profile obtained with the Kraken standard db for the 108 isolates analyzed. The color blue shows the percentage of reads assigned to genus *Brucella*.**

After the pre-processing analysis, we analyzed three specific metrics to define the quality of our dataset of reads:

Q30, sequencing depth and length of the reads.

Q30 is defined as the percentage of reads with mean quality at least of 30 Q-score (in the Phred scale for Illumina sequencing). The Q30 is therefore a simple metric to assess the overall quality of each library.

As showed in Table 8, we started from a datasets characterized by a high quality score, since Q30 lower than 70% was present in only in 19 dataset of raw reads, and no library has Q30 lower than 50%.

The trimming process allowed to further increase the mean quality, in fact, after removal of low-quality read ends no libraries with Q30 values lower than 70% remained, indicating that in each dataset at least 2/3 of the reads are composed by nucleotides with a base call mean accuracy around 99.9 % (which is the meaning of the value 30 in the Phred scale). Almost 90% of the libraries has Q30 greater than 80%. We concluded that raw read quality was already good, and the trimming procedure provided an improved quality set of reads for each library.

The depth of reads, or vertical Coverage, can be defined as the number of unique reads that include ("cover") a given nucleotide in the reconstructed genomic sequence, and was obtained as the ratio between the sum of all nucleotides sequenced and the size of genome reference (3.3Mb).

After trimming process, we have obtained a consequent decrease of vertical coverage (see Table 11), but this is the obvious cost of discarding low quality reads and read ends. Nevertheless, it should be noted that our datasets are however characterized by a theoretical coverage never less than 30X. Furthermore, a coverage greater than 150X was obtained for more than half of read libraries. These vertical coverage values are deemed more than sufficient for good quality genome assemblies and variants calling.

**Table 11: Q30 and vertical coverage observed before and after per-processing analysis.**

| **Q30 range values** | **Number of samples before pre-processing** | **Number of samples after pre-processing** |
|---|---|---|
| <50 | 0 | 0 |
| 50-60 | 13 | 0 |
| 61-70 | 6 | 0 |
| 71-80 | 13 | 4 |
| 81-90 | 75 | 16 |
| 91-100 | 0 | 89 |

| **V. COVERAGE range values** | **Number of samples before pre-processing** | **Number of samples after pre-processing** |
|---|---|---|
| < 30 | 0 | 0 |
| 30-60 | 5 | 9 |
| 61-90 | 16 | 21 |
| 91-120 | 12 | 13 |
| 121-150 | 16 | 18 |
| >150 | 59 | 47 |

Regarding the last quality metric analyzed, read length, after the pre-processing analysis we have obtained dataset of high quality reads with mean lengths never less than 80 nucleotides. The standard deviation of the fragment lengths was between 12.8 and 26.6 nucleotides, within the datasets of samples sequenced with the Illumina technology 2x150, and between 35.7 and 46.8 for the samples sequenced with the Illumina technology 2x300.

As showed in Figure 5 (B-C), also the comparison between mean and median showed evidence of a uniform distribution of read lengths in our samples. Having long reads is crucial in this kind of studies, because it reflects positively on read mappability on the reference genome (e.g. reducing the uncertainty on the read origin along the genome sequence), and facilitates genome assembly.

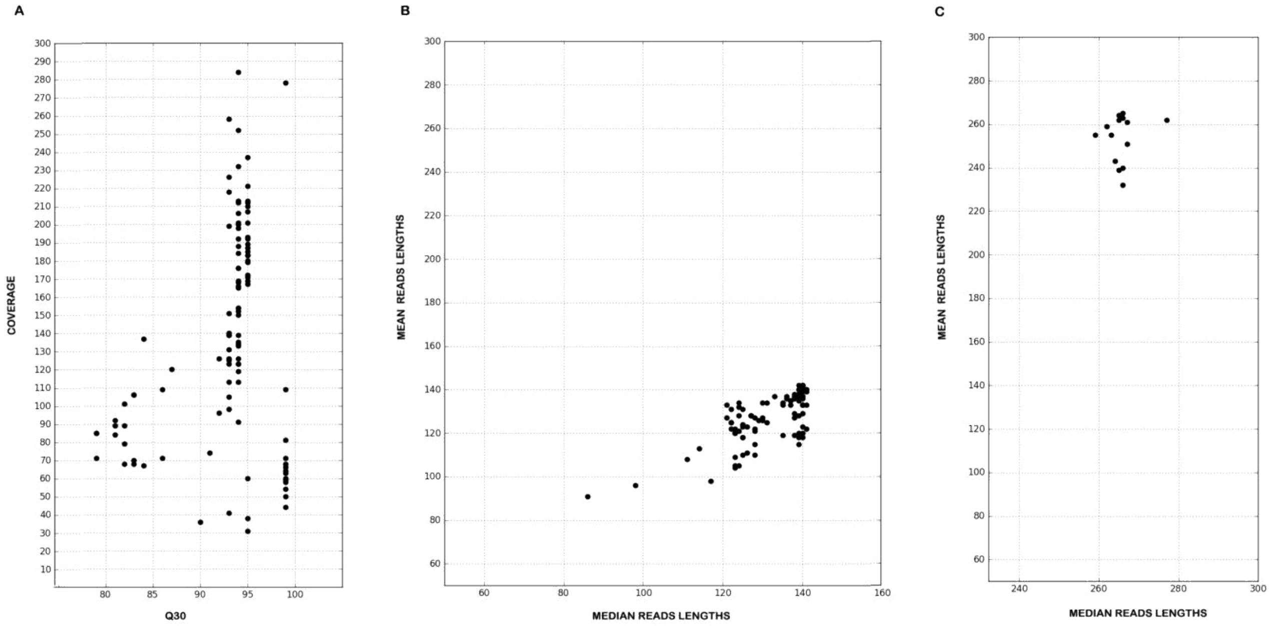


**Figure 5: Quality metrics values obtained after the trimming. (A) Comparison between Q30 and vertical coverage (B) Comparison between mean and median lengths of reads sequencing with Illumina NextSeq 500, (C) Comparison between mean and median lengths of reads sequencing with Illumina NextSeq 500, (C) Comparison between mean and median lengths of reads sequencing with Illumina MiSeq.**

## 5.2 *De novo* assemblies: Analysis of quality metrics

After quality control and cleaning, reads were considered of good enough quality for the reconstruction of each sample genomic sequence, through a *De novo* assembly strategy. The procedure can produce two kinds of constructs: i) contigs, defined as uninterrupted sequences built by read overlap; ii) scaffolds, defined as sets of ordered and oriented contigs, not necessarily contiguous along the genome (meaning that there could be gaps between contigs), built using pairing information from paired-end libraries. After the assembly procedure, described in detail in the Materials and Methods section, the quality of the reconstructed genomic sequences was evaluated. The analysis “completeness” of our sequences obtained after the *De novo* assembly. NG75 and LG75 are two of the metrics examined. NG75 is defined as the length of the smallest sequence in the set that contains the scaffolds whose combined length represents at least 75% of the length of the genome used as reference, while LG75 is calculated as the number of scaffolds used to obtain the NG75. The rationale is that a high NG75, as well as a low LG75, indicates that most of the genomic sequence in the analyzed assembly is contained in a small number of large scaffolds. On the other hand, low NG75 and high LG75 are a consequence of a very fragmented genome assembly. Fragmented genome reconstructions can impair the subsequent analysis, since they lead to uncertainty in the genomic fragments comparison and alignment, and because genes can be also fragmented, with gene parts included in a contig or scaffold, and other gene parts in different ones.

Genome assemblies having large NG75 are generally obtained when the read coverage is not only high, but also uniform along the genome, and when reads contain few errors.

All sets of assemblies were characterized by a NG75 value never less than 30,000 nucleotides and a LG75 never greater than 22 scaffolds and, in most cases, less than 10. The results, showed in Figure 6-A, indicate how our datasets of scaffolds are composed by a low number of sequences with large lengths, indicating that as the assemblies are not fragmented.

Furthermore, the results obtained by the alignment of each set of scaffolds against the reference genome, shows that at least 99.8% of the reference genome was covered (i.e. similar or identical to one or more scaffolds) by the sequences of each sample. As shows in Figure 6-B, through this large reconstruction degree of sequences, it was possible to assemble a high number of full-length genes, with a mean between 3,000 and 3,010 units for each sample.

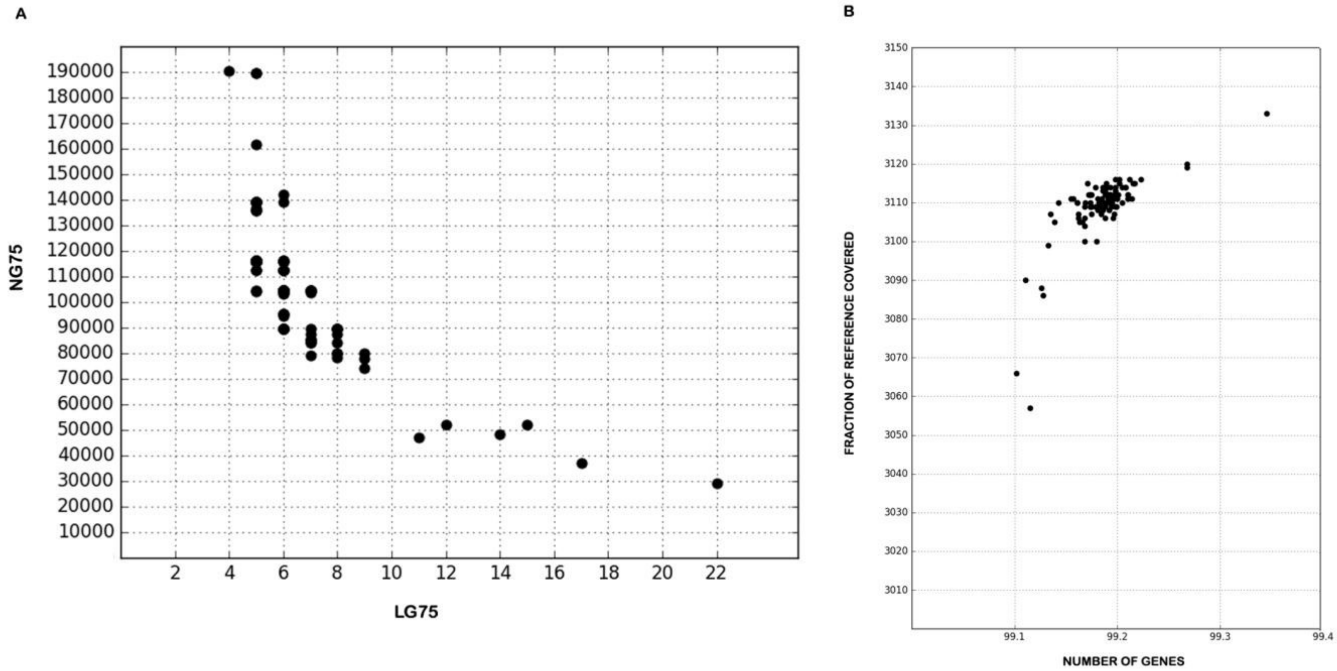


**Figure 6: Quality metrics values obtained after De *novo assembly* for each set of sequences assembled. (A) Comparison between NG75 and LG75 (B) Comparison between the fraction of reference genome covered and number of genes predicted.**

The last quality control of our data concerned the *k*-mers characterization (*k*-mers are read subsequences of fixed length equal to *k* nt) at species-level of each set of scaffolds assembled. The count of *k*-mers can highlight subsequences that are over-represented compared to the norm, indicating contamination or a sequencing bias introduced by the sequencing platform. The results of this analysis (**Annex 6 Table 10**) shows that, for each isolates, at least 99.8% of *k*-mers extracted was assigned to *B. melitensis.* The remaining 0.2% could be associated with genetic regions not present in the genome of *B. melitensis* used as reference in our database.

The following paragraphs shows the results obtained by the phylogenetic analysis; epidemiologically linked isolates and isolates with unknown epidemiological status are discussed.

## 5.3 Epidemiologically linked *B. melitensis* isolates

The outbreak-related isolates were detected and collected in 21 different farms in three Italian provinces over a period of 1.5 years. The culture-positive samples belonged to 37 animals that were analysed as a part of the within- and among-farm epidemiological investigation (Table 3 - Annex 1).

As showed in Figures 7-A, the MLVA-16 approach revealed the presence of 13 different genotypes, divided into two groups formed by single-locus variants and one double-locus variant. The Minimum spanning tree (MST) showed that the groups were split by mutations in the three hypervariable *loci* bruce04, bruce09, and bruce16.

One group included three genotypes of four isolates collected from farms located in the province of Rome, whereas in the remaining 33 strains from Isernia, Campobasso, and Frosinone provinces we identified 10 distinct genotypes.
In the cgMLST analysis, based on the *B. melitensis* 16M reference genome and using the assemblies previously obtained, we generated a gene panel of 2,704 targets.
The cgMLST clustering divided the isolates into two different genetic complexes, grouping the two farms from the province of Rome (complex 2) separately from the remaining 19 farms (complex 1). The genetic division measured with the cgMLST panel was for 164 different genes (Figure 7-B). The analysis using the *B. melitensis* panel found one prevalent genotype that was similar across the provinces of Frosinone, Campobasso, and Isernia, and that was found in 10 of the tested farms.

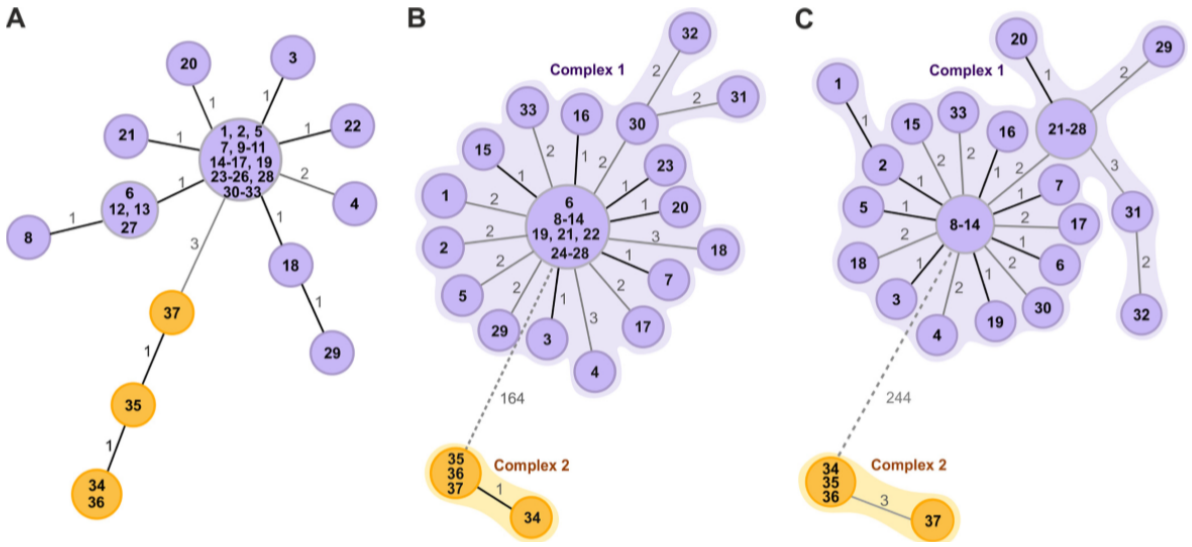


**Figure 7: Minimum spanning trees (MST) generated for 37 epidemiologically related isolates. Separate epidemiological clusters are marked with different colors indicating the provinces of isolation (purple, Frosinone, Isernia, and Campobasso; orange, Rome). (A) MST based on *B. melitensis* MLVA-16 typing. The distance labels correspond to the number of discriminating alleles. (B) MST generated using the gene-by-gene approach. cgMLST profiles were assigned using the *B.melitensis* task template with 2,704 target genes. The MST was created by cgMLST target pairwise comparison, ignoring missing values, with distance representing the number of diverse alleles. Separate complexes are highlighted. (C) MST based on SNP analysis using *B. melitensis* strain 16M as a reference. The distance labels correspond to the number of discriminating SNPs between neighboring genotypes. The prefix ItBM was omitted from the isolates' labels for simplicity.**

Sixteen isolates in complex 1 shared identical core genome profiles, and the largest distance between any two neighboring isolates was not greater than three genes. In complex 2, one isolate was separated from the other three by one gene difference.

Removing 50 targets from the analysis where any value was missing decreased the distances between the nodes even further, and classified all samples from Rome as identical (not shown). A within-farm genetic variation was also observed.
The SNP analysis identified 3,390 SNPs, of which 3,146 were classified as clean unique variants and included in further analysis. The tree split the samples into two genetic clusters with a distance of 244 SNPs between them (Figure 7-C). We observed a within-farm variation of 2 MLVA-16 *loci*, 3 cgMLST *loci*, and 4 SNPs. The maximum pairwise distance found in the two complexes was 6 cgMLST genes and 7 SNPs.

The comparison of discriminatory power of MLVA**,** cgMLST, and SNP typing showed that the SNP-based approach was superior to the other two methods, with a Simpson's Index of Diversity (SDI) of 0.922 and 95% confidence intervals (CI) of 0.866 to 0.978. SDI of cgMLST was calculated to be 0.815 (95% CI, 0.685 to 0.945), and SDI of MVLA-16 was 0.674 (95% CI, 0.505 to 0.843). SNP typing was a good predictor of cgMLST, with an Adjusted Wallace (AW) of 0.788 (95% CI, 0.546 to 1.000). The correspondence of the typing results, however, was not bidirectional, as the cgMLST to SNP AW was 0.295 (95% CI, 0.136 to 0.453). Comparison of the remaining pairs of typing schemes showed that there was no congruence between clusters they predicted (the AW of each pair did not exceed 0.03).

### 5.4 *B. melitensis* isolates with unknown epidemiological status

MST calculated using the MLVA-16 typing results showed a distance between directly linked nodes not exceeding 9 VNTR loci (Figure 8). Fifty-one MLVA-16 profiles were assigned to the 71 strains, and diverse allele variants were identified in all *loci* apart from bruce45. Eleven profiles were shared by more than one isolate, which, with the exception of one human isolate, corresponded to the samples originating from the same geographical location (Table 3 - Annex 1).

MLVA profiles tend to be conserved between epidemiologically linked strains; therefore, the strains from an outbreak are likely to have a similar MLVA profile.

Three MLVA-16 profiles, 10 (samples ItBM_41 to ItBM_44), 15 (samples ItBM_93 to ItBM_96 and ItBM_98), and 24 (ItBM_55 and ItBM_89 to ItBM_91), were identified in more than three strains, suggesting close relatedness of samples within these profiles. The method also allowed the identification of two clear outliers. Samples ItBM_38 and ItBM_39 showed a distance of 9 alleles from the nearest *B. melitensis* isolate and no relatedness to one another.

According to our MLVA-16 data, only three out of six human cases could be linked to a specific animal source analyzed in our study. Human samples ItBM_41 and ItBM_43, isolated from two patients in the city of Salerno, shared the same MLVA-16 profile as two animal isolates from a farm in Salerno province (samples ItBM_42 and ItBM_44), all collected in 2011. Human isolate ItBM_50 and two animal isolates (ItBM_51 and ItBM_52) were assigned to the MLVA-16 profile 42, but interestingly, ItBM_50 was isolated 4 years later than the animal strains. The other three human samples did not show sufficient relatedness to any of the animal isolates to reliably trace the source of infection. The number of variable *loci*, in these cases, ranged from 2 to 9 in relation to the closest neighboring MLVA-16 profile.

Thirteen complexes were assigned in the MST data analysis. Gene-by-gene analysis confirmed the relatedness of genotypes with MLVA-16 profiles 10 and 15; however, according to cgMLST two other isolates were at a distance from 0 to 1 gene away from the samples of the MLVA-16 profile 15, as was one other isolate of profile 10. ItBM_55, classified as MLVA-16 profile 24, was shown not to be closely linked to other isolates with the same MLVA-16 alleles when examined with a gene-by-gene approach.

Using cgMLST, four of the human isolates (ItBM_41, ItBM_43, ItBM_50, andItBM_108) were found at a distance not exceeding 2 alleles to the closest animal strain. Two of the human samples originating in Piedmont (ItBM_99 and ItBM_78) were genetically different from the animal samples, with 156 and 195 allele differences from the closest isolate, and could be identified as outliers, although they were distantly related to other Italian genotypes. Divergence of these two samples was not evident in MLVA-16 typing (distance of 2 to 3 alleles to other isolates).

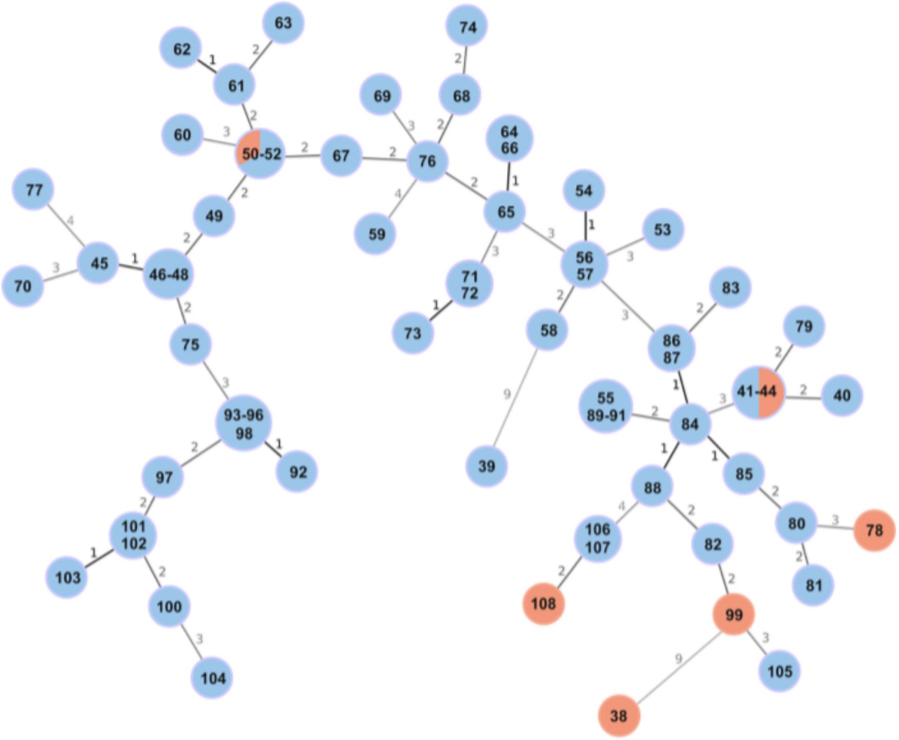


**Figure 8: Minimum spanning tree (MST) based on *B. melitensis* MLVA-16 typing results generated for 71 isolates with unknown epidemiological status. The tree was generated using the goeBURST algorithm in PHYLOViZ software The distance labels correspond to the number of discriminating alleles. The red nodes correspond to human isolates and the blue nodes to animal isolates. The prefix ItBM was omitted from the isolates' labels for simplicity**.

A total of 6,540 SNPs were discovered by mapping 71 genomes to the *B. melitensis* 16M reference strain. Out of these, 6,027 were considered high-quality discriminatory SNPs and were used to infer the relationship between the strains. We applied the threshold of 7 SNPs to detect the clusters of closely related cases, and in accordance with cgMLST analysis, we identified 13 complexes (Figure 9-B). The highest distances observed between two adjoining isolates were 2,616 and 2,235, belonging to the SNP profiles of ItBM_38 and ItBM_39, which also were marked as outliers by MLVA-16 and cgMLST analyses.

In agreement with cgMLST, two human cases (ItBM_78 and ItBM_99) could not be traced to any of the analyzed animal strains of *B. melitensis*, and both differed by more than 200 SNPs from the nearest SNP profile.

Close genetic relationship to at least one isolate from an animal host was confirmed for ItBM_41, ItBM_43, ItBM_108, andItBM_50.

SDI for the three typing schemes were calculated to be 0.986 (95% CI, 0.978 to 0.995) for MLVA-16, 0.988 (95% CI, 0.978 to 0.998) for cgMLST, and 0.992 (95% CI, of 0.985 to 1.000) for SNP typing. AW test showed the highest congruence between SNP and cgMLST-based clusters when the SNP method was used as a primary typing method (AW of 0.840; 95% CI, 0.753 to 0.927). When we used cgMLST as the primary method, however, the AW value dropped to 0.573 (95% CI, 0.290 to 0.856). MLVA-16 was a poor predictor of SNP (AW of 0.318; 95% CI, 0.112 to 0.524) and of cgMLST (AW of 0.494; 95% CI, 0.333 to 0.655).

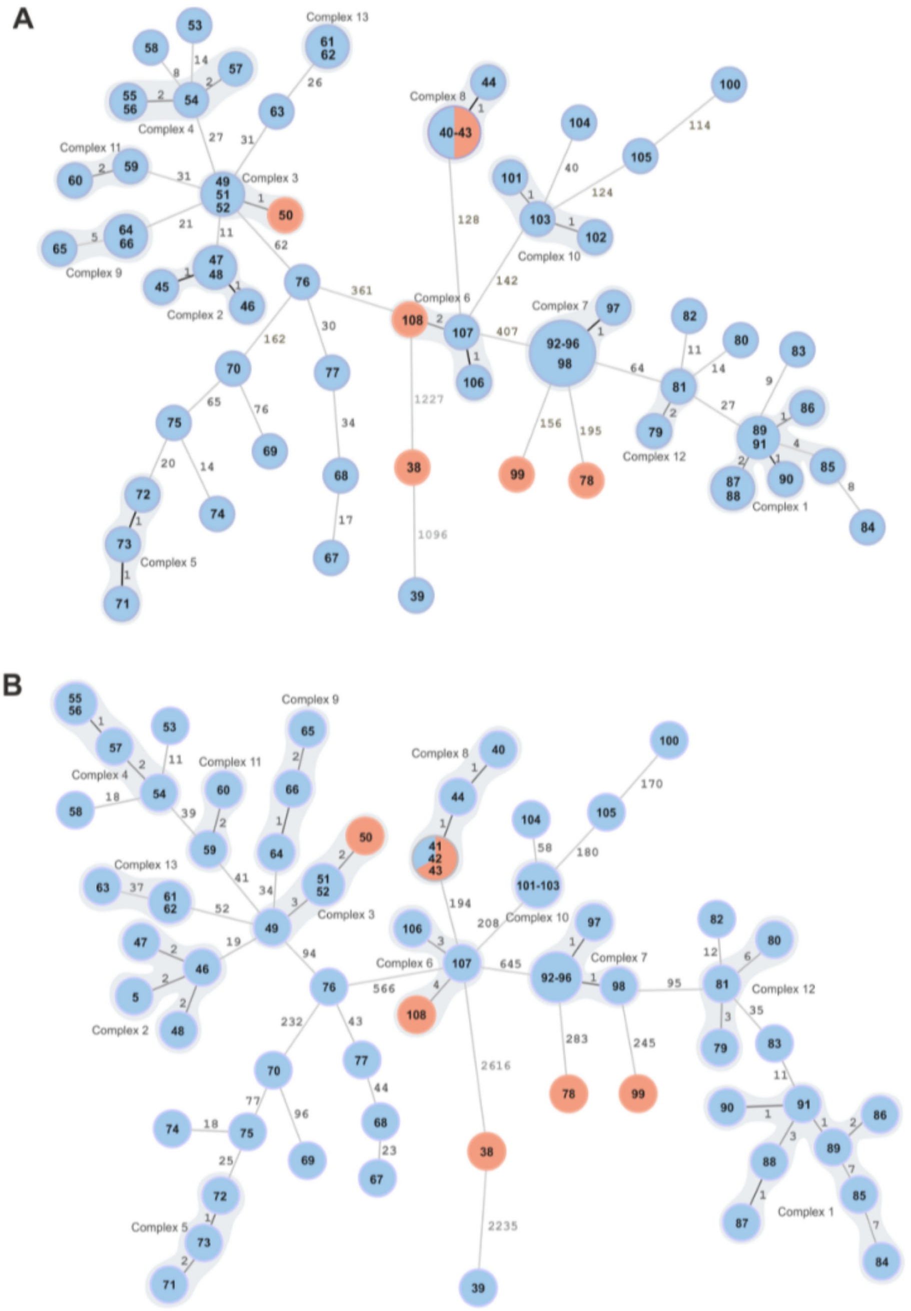


**Figure 9:** Minimum spanning trees (MST) based on WGS analysis results generated for 71 isolates with unknown epidemiological status. (A) MST generated using gene-by-gene approach. cgMLST profiles were assigned using B. melitensis task template with 2,704 target genes. The MST was created by cgMLST target pairwise comparison, ignoring missing values, with distance representing the number of diverse alleles. Separate complexes are highlighted. (B) MST based on SNP analysis using *B. melitensis* strain 16M as a reference. The distance labels correspond to the number of discriminating SNPs between neighboring genotypes. The red color nodes correspond to human isolates and the blue nodes to animal isolates. The prefix ItBM was omitted from the isolates' labels for simplicity.

## 6. DISCUSSION

Frequently, there is an obvious lack of data quality documentation within analysis based on sequencing experiments.Monitoring of sequencing data is a good starting point for analyzing the results. This stage should comprise the initial data control based on the raw data analysis with focus on reads quality, and further assessments, using different and specific metrics, should be done for each analysis step that requires succeeding manipulation of the data.

Previously, we have discussed how the possibility of cross contamination between biological samples from different species that have been processed or sequenced in parallel has the potential to be extremely deleterious for downstream analyses. In this project, contamination detection was performed in two different steps of our analysis, a first genus-level check within each set of raw reads and a further species-level check of sequences obtained with *De novo* assembly. These monitoring analyses allowed us to exclude, with a high probability, a significant contamination of our data.

To compare whole genomes and identify differences with sufficient certainty and accuracy the quality of the underlying data is crucial due to the error-rate of current sequencing technologies (Janowicz et al., 2018; Schurch et al., 2017; Georgi et al., 2017; Whatmore et al., 2016; Tan et al., 2015; Wattam et al., 2014; Garofolo et al., 2013; Al Dahouk et al., 2007).

The quality score (Phred scale) assigned by the sequencing platform is the first metric to define accuracy degree of NGS data and we have already see how, through per-processing analysis, it was possible to obtain dataset of reads composed by nucleotides with a mean base call accuracy around 99.9%, implying low error probabilities. Nevertheless, considering the high reliability of the sequencing technologies used in this project, the depth of the reads, or coverage, was the main metric used to evaluate the quality of our set of reads.

High coverage allows estimating the high “completeness” of whole genome sequenced, which is crucial since it allows for more sophisticated and downstream analysis, as well as since it can be used to correct wrong bases assignment through the comparison of multiple reads sequenced from the same genomic region. On the basis of our experiences, after per-processing analysis, a minimum coverage of 20x to 30x was assumed to give sufficient power to resolve ambiguous base assignments during assembly process. Our libraries have all coverage higher, in most cases remarkably, than these reference values.

Finally, is important to emphasize how an high reconstruction degree of each sample genomic sequence is essential in the epidemiological studies based on WGS data, mainly in the cluster techniques used in this work. The high number of genes assembled, for each isolates, allowed us to define a most wide panel in cgMLST analysis. Furthermore, having obtained datasets of scaffolds composed by a low number of large sequences and able to cover more than 99% of genome reference, allowed us to compare an high number of genomic regions. This is an important requirement to exploit the discriminatory power of SNPs analysis.

We were able to compare the performance of two WGS-based typing methods, SNP analysis and cgMLST with the gold standard MLVA-16 in an analysis of the phylogenetic relationship between *B. melitensis* isolates collected from human and animal samples in the context of a national surveillance program.

Based on the results, we found that all three typing schemes generally showed similar results among each other and, although the SNP analysis had the greatest resolving power in terms of detected differences among the isolates, the number of genotypes predicted in the surveillance was comparable (51 MLVA-16 types, 55 cgMLST types and 60 SNPs) and SDI were similar. However, the results obtained from the SDI test demonstrated that when the SDI test was applied to samples from epidemiologically linked sets, SNP analysis was superior in differentiating between closely related samples within the same epidemiological context. These data suggest that a change in the diagnostic approach can be beneficial, confirming the epidemiological profile to be analized, since although WGS-based approaches can be used as stand-alone tools in establishing phylogenetic relationships, MLVA-16 should ideally be supported by SNP or results from gene-gene typing as they provide more information.

As to the diagnostic accuracy, we could verify that all three typing methods predicted effectively the presence of two divergent genomes from the rest of the Italian strains. The distance between each of these and the nearest Italian isolates was found to be more than 1000 alleles, whereas no more than 412 difference alleles occurred between any of the local strains.

Most of the analyzed samples belonged to the western Mediterranean line of *B. melitensis*, while the outliers were members of the Eastern Mediterranean and United States strains. These results corroborate those of Garofolo et al. (2013), in which 206 isolates of *Brucella abortus* and *B. melitensis* from eight regions of southern Italy were genetically evaluated using Tandem Variable Numeric Replication (VNTRs), and verified the genetic diversity and geographical distribution of these VNTR genotypes in a fine-scale analysis using 16 Loci VNTR in a MLVA-16 methodology. The other two methods we used also confirmed this result, with SNP analysis identifying more than 2000 SNPs and 9 MLVA alleles for the closest Italian genotype for both samples.

The epidemiological investigation showed that ItBM_38 was isolated from a Syrian patient with a history of frequent trips to his / her country of origin, where the same lineage from the Eastern Mediterranean is believed to be prevalent (Georgi et al., 2017). The strain ItBM_39, on the other hand, was isolated from a goat imported from Spain to Italy.

In two human isolates, ItBM_50 and ItBM_108, we found the same SNP and cgMLST complexes as in animal strains. However, the samples presented variations in the epidemiological context, since they were collected with a few years difference and in different geographic locations. Therefore, the results suggest that animal isolates could be closely related (or ancestral) to the source of human infection, but not directly involved in the transmission event. In these cases, the observation based on WGS typing indicates that *B. melitensis* strains were circulating in the affected regions of Italy for many years, and that the surveillance program failed to eradicate them.

In the case of distantly related genomes of the same lineage, the analyses of cgMLST and SNP led to better results, because they provided higher resolution of the phylogenetic distance compared to MLVA-16, thus, the results of cgMLST and SNP allowed the identification of genotypes with more certainty, which are probably not connected to other circulating strains. This was particularly evident in the case of two clinical isolates (ItBM_99 and ItBM_78) and in the case of *B. melitensis* collected from an ibex (*Capra ibex ibex*) in the Gran Paradiso National Park, located in the Graian Alps in Italy (sample ItBM_100). The results showed that, although all the applied schemes could be used to identify genomic outliers very distant within the *Brucella* population, the WGS-based schemes were superior in the identification of unrelated cases belonging to the same lineage. In addition, we also found that, within groups of similar genotypes, the cgMLST was also performed for SNP analysis, but some discrepancies were observed in the MLVA-16 analysis. As an example, seven isolates from Sicily had profiles differing by a maximum of two SNPs or one gene (samples ItBM_92-ItBM_98), suggesting that they were closely related. However, while these variables are similar to MLVA-16 profile 15, one belongs to type 8 (1 distal allele, bruce19) and another to type 12 (2 distant alleles, bruce4 and bruce7). Therefore, the interpretation of the WGS results suggests that these were actually strains of the same complex, whereas the typing of MLVA-16 would not necessarily lead to the same conclusion. A similar observation was reported by Dallman et al., (2015), who demonstrated that the use of SNPs from *E. coli* O157 isolates was able to identify cases with twice the sensitivity of the MLVA-16 scheme, while Georgi et al. (2017) demonstrated that MLVA-16 presented lower sensitivity and analytical specificity in relation to the use of WGS, based on SNP typing, analyzing a set of 63 human isolates of *B. melitensis*. Interestingly, in our group of outbreak-related cases, we identified several genotypes that differed by one, two, or three hypervariable alleles and belonged to an outbreak caused by a single clone epidemic. When analyzed by WGS, we could observe that these strains were closely related (up to 6 genes or 7 different SNPs). All these epidemiological data show that the MLVA-16 test may not provide sufficient resolution to accurately predict the phylogenetic relationships between the isolates involved in a current outbreak or to obtain important information about the strains that circulated over the years without any direct connection to each other.

We can state from the results obtained that the SNP analysis was carried out successfully, having excellent applicability for the differentiation between *Brucella* species, as well as to map the geographical distribution, traceability and general dissemination of *B. melitensis*. Tan and collaborators (2015) were able to reconstruct the phylogeographic history of global dissemination of *B. melitensis* on a finer scale using SNP analyses of the whole genome of *B. melitensis* lineage collected worldwide. Georgi et al. (2017), which investigated through the analysis of SNPs based on complete genomes from an extensive collection of strains of *B. melitensis* isolated from human cases in Germany, also emphasize the importance of SNP analysis as a powerful tool in typing, as well as providing useful information on geographical origin and tracking analysis. A number of other scientific papers have been published, with appropriate genotyping approaches for rapid detection and diagnostic assays for epidemiological and clinical molecular studies, emphasizing the importance of a detailed knowledge of the *Brucella* phylogeny to have a better understanding of the ecology, evolutionary history and relations among hosts for this genre (Janowicz et al., 2018; Schurch et al., 2018; Pightlinget et al., 2014; Maio et al., 204; Garofolo et al., 2013; Jun et al., 2013; Kang et al., 2011; Maquart et al., 2009; Paixão, 2009). Despite the advances, to date there is no official cgMLST scheme validated for any of the *Brucella* species. Consequently, clustering types for specific data, and particularly for closely related lineages, can only be empirically evaluated and therefore subject to variation between laboratories (Janowicz et al., 2018). In order to reliably interpret the results, cutoff values first should be established based on the analysis of a significant number of closely related strains and unrelated strains sharing common or closely related profiles assigned using gold standard typing methods. Analyses of isolates related to outbreaks suggest that these outbreaks were caused by two independent epidemic clones circulating in central Italy during the same period. Since the maximum distance between pairs of isolates within complexes formed by these clones did not exceed 6 genes (cgMLST) or 7 SNPs, these results provided important findings as they highlight the possible criteria for inclusion of an isolate in a brucellosis outbreak, which allowed us to suggest a value of  6 loci in the analysis of cgMLST SNPs and  7 in the analysis of SNPs of the WGS.

Some studies on the implementation of molecular techniques (Jackson et al., 2016; Katz et al., 2015) argue that a general cutoff value applied in SNPs or cgMLST cannot always predict with confidence the relation of epidemiological proximity of the samples, yet according to Jackson et al. (2016), in a study carried out to enhance listeriosis outbreak detection and investigation, isolates with differences in SNPs ranging from 10 to 30 were frequently linked. In this way, we believe that the proposed cutoff values should be taken as guideline and interpreted in the context of available epidemiological information.

Using an approach that achieves maximum resolution is particularly important for tracing the spread of a disease during an outbreak (Janowicz, et al., 2018). SNP analysis potentially has the greatest discriminatory power among typing methods since nucleotide polymorphisms can be detected in both the coding and non-coding regions of the genome. However, the choice of a reference genome can significantly influence the number of SNPs identified, the accuracy of the alignments of cured read sequences, and reconstructed phylogenetic relationships (Pightling et al., 2014). CgMLST requires the availability of complete and accurately sequenced genomes in order to generate the typing schemes. The inclusion of coding sequences not only decreases the number of sites typed in the analysis, but at the same time facilitates the standardization and reproducibility of the analyses, since it focuses on a pre-defined set of genes. In WGS analysis, the quality of the readings, as well as the assembly, are fundamental to obtain results with cgMLST quality and mainly to obtain reliable results. Throughout our study, all samples reached at least 98% of good targets, since low quality assemblages probably have a small number of good targets and therefore lead to the generation of inaccurate results in the phylogenetic analysis. Therefore, we propose that data with less than 97%  of good targets should be taken with caution.

\

## 7. CONCLUSIONS

Based on the results obtained, we can conclude that the WGS / NGS data can be effectively used to:

- obtain a better understanding of the epidemiology and dynamics of *Brucella* populations;

- collect detailed information that can be used to trace sources in case of outbreaks in animals, zoonotic or foodborne infections;

- facilitate the free transport and trade of animals and their by-products;

- facilitate the evaluation of the possible extension of an outbreak in progress and the reliable prediction of the routes of its spread;

According to the One Health approach, public health agencies can implement WGS to assist with disease control and eradication plans. In our study, both cgMLST and SNP analyses performed well on the genetic diversity of *B. melitensis*, and we demonstrated that the performance of the gene-for-gene approach was comparable to that of SNP analysis. Based on these results, we believe that the MLVA-16 typing of *B. melitensis* in Italy can now be successfully replaced by the more informative analysis provided by WGS.

# REFERENCES

Al Dahouk S and Nöckler K 2011 Implications of laboratory diagnosis on brucellosis therapy.Expert Rev. anti.infect. Ther.9 (7): 833–845

Al Dahouk S, Nockler K, Scholz HC, Pfeffer M, Neubauer H, Tomaso H. 2007. Evaluation of genus-specific and species-specific real-time PCR assays for the identification of *Brucella* spp. Clin Chem Lab Med 45: 1464–1470. https://doi.org/10.1515/CCLM.2007.305.

Al Dahouk S, Fleche PL, Nockler K, Jacques I, Grayon M, Scholz HC, Tomaso H, Vergnaud G, Neubauer H. 2007. Evaluation of *Brucella* MLVA typing for human brucellosis. J Microbiol Methods 69:137–145. https:// doi.org/10.1016/j.mimet.2006.12.015.

Alton G.G., Jones L.M., Angus R.D. & Verger J.M. (1988). Techniques for the Brucellosis Laboratory. Institut National de la Recherche Agronomique, Paris, France.

Aminov RI. Horizontal gene exchange in environmental microbiota. Front Microbiol. 2011;2:158. Published 2011 Jul 26. doi:10.3389/fmicb.2011.00158

Auwera GA, Carneiro MO, Hartl C, Poplin R, Levy-Moonshine A, Jordan T, Shakir K, Roazen

D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11.10.1–11.10.33.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. .

Bahmani, N., Naseri Z., M.Y., Alikhani M., Karami, Kalvandi R.2016. Evaluation of antibacterial effects of Withania coagulans and Cynara cardunculus extracts on clinical isolates of *Brucella*strainsInt. J. Med. Res. Health. Sci., 5 (9) (2016), pp. 90-9

BDN - Banca Dati Nazionale. 2018. Statistics on animal farming. Italian National Database for Identification and Registration of Animals. http://statistiche .izs.it/portal/page?_pageid73,12918&_dadportal&_schema PORTAL.

Bergthorsson U, Ochman, H.1995. Heterogeneity of genome sizes among natural isolates of *Escherichia coli*.J Bacteriol. 1995 Oct; 177(20): 5784–5789. PMCID: PMC177399

Blasco, J. M. & Molina-Flores, B. (2011). Control and eradication of *Brucella* melitensis infection in sheep and goats. Veterinary Clinics of North America: Food Animal Practice, 27, pp. 95-104.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST: architecture and applications. BMC Bioinformatics 10:421. .

Bricker B.J. PCR as a diagnostic tool for brucellosis. Vet. Microbiol. 2002;90:435–446. doi: 10.1016/S0378-1135(02)00228-6.

Bruce J. Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K. Young, Ashlee M. Earl (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE 9(11): e112963. doi:10.1371/journal.pone.0112963

Buddle, M. B. (1956). Studies on *Brucella* ovis(n. sp.), a cause of genital disease of sheep in New Zealand and Australia. Journal of Hygiene,54, pp. 351-364.

Carmichael, L. & Bruner, D. (1968). Characteristics of a newly-recognized species of *Brucella* responsible for infectious canine abortions. The Cornell Veterinarian, 48(4), 579-592.

Carrico JA, Sabat AJ, Friedrich AW, Ramirez M, ESCMID Study Group for Epidemiological Markers (ESGEM). 2013. Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the nextgeneration sequencing revolution. Euro Surveill 18:20382. https://doi .org/10.2807/ese.18.04.20382-en.

Carriço JA, Silva-Costa C, Melo-Cristino J, Pinto FR, de Lencastre H, Almeida JS, Ramirez M. 2006. Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant Streptococcus pyogenes. J Clin Microbiol 44:2524–2532. https://doi.org/ 10.1128/JCM.02536-05.

CDC. 2017. Onde Healt – Zoonotic Disease l animals. Centers for Disease Control and Prevention (CDC), available on:< https://www.cdc.gov/onehealth/basics/zoonotic-diseases.html>. Accessed 30 September 2018.

CFSPH – The Center for Food Security & Public Health.2009 (CFSPH), Animal Disease Information, Zoonotic, available on: <http://www.cfsph.iastate.edu/DiseaseInfo/disease.php?name=*Brucella*-melitensis&lang=en>.Accessed 19 Septem-ber 2018.

Corbel M. 2006. Brucellosis in humans and animals, WHO/CDS/EPR/ 2006.7. World Health Organization in collaboration with the Food and Agriculture Organization of the United Nations and World Organization for Animal Health. WHO Press, Geneva, Switzerland.

Davis S, Pettengill JB, Luo Y et al. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. Peer J Computer Science 2015; 1:e20.

Dallman TJ, Ashton PM, Byrne L, Perry NT, Petrovska L, Ellis R, Allison L, Hanson M, Holmes A, Gunn GJ, Chase-Topping ME, Woolhouse ME, Grant KA, Gally DL, Wain J, Jenkins C. 2015. Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia coli*O157:H7 strains causing severe human disease in the UK. Microb Genom 1:e000029.

/De Massis F, Ancora M, Atzeni M, Rolesu S, Bandino E, Danzetta ML, Zilli K, Di Giannatale E, Scacchia M. 2015. MLVA as an epidemiological tool to trace back *Brucella* melitensis biovar 1 re-emergence in Italy. Transbound Emerg Dis 62:463–469. https://doi.org/10.1111/tbed.12397.

Di Giannatale E, De Massis F, Ancora M, Zilli K, Alessiani A. 2008. Typing of *Brucella* field strains isolated from livestock populations in Italy between 2001 and 2006. Vet Ital 44:383–388.

Dittami SM, Corre E. Detection of bacterial contaminants and hybrid sequences in the genome of the kelp Saccharina japonica using Taxoblast. Morrison H, ed. PeerJ. 2017;5:e4073. doi:10.7717/peerj.4073.

European Centre for Disease Prevention and Control. Annual epidemiological report 2015. Brucellosis Stockholm: ECDC; 2016.

EC - European Commission. 2016. Commission Decision 2016/1811/UE Commission implementing decision (EU) 2016/1811 of 11 October 2016 amending annex II to Decision 93/52/EEC as regards the recognition of the Province of Brindisi in the region Puglia of Italy as officially free of brucellosis (B. melitensis). European Commission, Brussels, Belgium.

Ewalt, D. R., Payeur, J. B., Martin, B. M., Cummins, D. R. & Miller, W. G. (1994). Characteristics of a *Brucella* species from a bottlenose dolphin (Tursiops truncatus). Journal of Veterinary Diagnostic Investigation, 6, pp. 448-452.

Foster JT, Beckstrom-Sternberg SM, Pearson T, Beckstrom-Sternberg JS, Chain PS, Roberto FF, Hnath J, Brettin T, Keim P. 2009. Whole-genomebased phylogeny and divergence of the genus *Brucella*. J Bacteriol 191:2864–2870. https://doi.org/10.1128/JB.01581-08.

Foster, G., Osterman, B. S., Godfroid, J., Jacques, I. & Cloeckaert, A. (2007). *Brucella* cetisp. nov. and *Brucella* pinnipedialissp. nov. for *Brucella*strains with cetaceans and seals as their preferred hosts. International journal of systematic and evolutionary microbiology, 57, pp. 2688-2693.

Francisco AP, Bugalho M, Ramirez M, Carrico JA., Global Optimal eBURST analysis of Multilocus typing data using a graphic matroid approach,BMC Bioinformatics 2009, 10:152doi:10.1186/1471-2105-10-152

Gándara, B., Merino, A.L., Rogel, M.A., Martínez-Romero, E., 2001. Limited genetic diversity of Brucella spp. J. Clin. Microbiol. 39, 235–240.

Garofolo G, Di Giannatale E, De Massis F, Zilli K, Ancora M, Camma C, Calistri P, Foster JT. 2013a. Investigating genetic diversity of *Brucella* abortus and *Brucella* melitensis in Italy with MLVA-16. Infect Genet Evol 19:59–70. .

Garofolo G, Ancora M, Di Giannatale E. 2013b. MLVA-16 loci panel on *Brucella* spp. using multiplex PCR and multicolor capillary electrophoresis. J Microbiol Methods 92:103–107. https://doi.org/10.1016/j.mimet .2012.11.007.

Georgi E, Walter MC, Pfalzgraf MT, Northoff BH, Holdt LM, Scholz HC, Zoeller L, Zange S, Antwerpen MH. 2017. Whole genome sequencing of *Brucella* melitensis isolated from 57 patients in Germany reveals high diversity in strains from Middle East. PLoS One 12:e0175425. https://doi .org/10.1371/journal.pone.0175425.

Godfroid, J., Nielsen, K. & Saegerman, C. (2010). Diagnosis of brucellosis in livestock and wildlife. Croatian medical journal, 51, pp. 296-305.

Godfroid J, Cloeckaert A, Liautard JP, Kohler S, Fretin D, Walravens K, Garin-BastujiB,LetessonJJ.2005.FromthediscoveryoftheMaltafever's agent to the discovery of a marine mammal reservoir, brucellosis has continuously been a re-emerging zoonosis. Vet Res 36:313–326. https:// doi.org/10.1051/vetres:2005003.

Gopaul, KK., Dainty, AC., Muchowski, JK., Dawson, CE., Stack, JA., Whatmore, AM.
(2014) Direct molecular typing of *Brucella* strains in field material*Veterinary Record* 175, 282.

Gopaul KK, Sells J, Lee R, Beckstrom-Sternberg SM, Foster JT, Whatmore AM. Development and assessment of multiplex high resolution melting assay as a tool for rapid single-tube identification of five*Brucella* species. BMC Res Notes. 2014;7(1):903. doi: 10.1186/1756-0500-7-903.

Grissa I, Bouchon P, Pourcel C, Vergnaud G. 2008. On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. Biochimie 90:660–668. .

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-1075. doi:10.1093/bioinformatics/btt086.

Hanot Mambres D, Boarbi S, Michel P, Bouker N, Escobar-Calle L, Desqueper D, Fancello T, Van Esbroeck M, Godfroid J, Fretin D, Mori M. 2017. Imported human brucellosis in Belgium: bio and molecular typing of bacterial isolates, 1996-2015. PLoS One 12:e0174756. https://doi.org/ 10.1371/journal.pone.0174756.

Hasman H, Saputra D, Sicheritz-Ponten T, et al. Rapid Whole-Genome Sequencing for Detection and Characterization of Microorganisms Directly from Clinical Samples. Journal of Clinical Microbiology. 2014;52(8):3136. doi:10.1128/JCM.01369-14.

Havelaar, A.H., Kirk, M.D., Torgerson, P.R., Gibb, H.J., Hald, T., Lake, R.J., Praet, N., Bellinger, D.C., De Silva, N.R.,Gargouri, N., Speybroeck, N., Cawthorne, A., Mathers, C., Stein, C., Angulo, F.J. & Devleesschauwer, B. (2015). World Health Organization Global estimates and regional comparisons of the burden of foodborne disease in 2010. PLoS Medicine,12(12), e1001923.

Hervé Tettelin, Vega Masignani, Michael J Cieslewicz, Claudio Donati, Duccio Medini, Naomi L Ward, Samuel V Angiuoli, Jonathan Crabtree, Amanda L Jones, A Scott Durkin, et al. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences of the United States of America,102(39):13950–13955, 2005.

Henri C, Leekitcharoenphon P, Carleton HA, Radomski N, Kaas RS, Mariet JF, Felten A, Aarestrup FM, Gerner Smidt P, Roussel S, Guillier L, Mistou MY, Hendriksen RS. 2017. An assessment of different genomic approaches for inferring phylogeny of Listeria monocytogenes. Front Microbiol 8:2351. https://doi.org/10.3389/fmicb.2017.02351.

Huddleson, I. F. And Smith, L. H. 1931 A critical study of the *Brucella* agglutination. reaction and abortion rate in a herd of cattle under natural conditions. J. Am. Vet. Med. Assoc.79, 63-78.

Huddleson, I. F. (1929). Differentiation of the Species of the Genus *Brucella*. Michigan State College Agricultural Experimental Station Technical Bulletin, 100, pp. 1–16.

Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol. 1988;26(11):2465-6.

Kang SI, Heo EJ, Cho D, Kim JW, Kim JY, et al. (2011) Genetic comparison of *Brucella* canis isolates by the MLVA assay in South Korea. J Vet Med Sci 73: 779–786.

Hyden P, Pietzka A, Lennkh A, Murer A, Springer B, Blaschitz M, Indra A, Huhulescu S, Allerberger F, Ruppitsch W, Sensen CW. 2016. Whole genome sequence-based serogrouping of Listeria monocytogenes isolates. J Biotechnol 235:181–186. https://doi.org/10.1016/j.jbiotec.2016 .06.005.

Katz LS, Wagner DD, Petkau A et al. Lyve-SET: a high-quality SNP pipeline for aiding in bacterial pathogen outbreak investigations. In: Sequencing, Finishing, and Analysis in the Future Meeting, Santa Fe, NM, 2015.

Kirk, M.D., Pires, S.M., Black, R.E., Caipo, M., Crump, J.A., Devleesschauwer, B., Döpfer, D., Fazil, A., Fischer-Walker, C.L., Hald, T., Hall, A.J., Keddy, K.H., Lake, R.J., Lanata, C.F., Torgerson, P.R., Havelaar, A.H. & Angulo, F. (2015). World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. PLoS Medicine, 12,e1001921

Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, Katz LS, Stroika S, Gould LH, Mody RK, Silk BJ, Beal J, Chen Y, Timme R, Doyle M, Fields A, Wise M, Tillman G, Defibaugh-Chavez S, Kucerova Z, Sabol A, Roache K, Trees E, Simmons M, Wasilenko J, Kubota K, Pouseele H, Klimke W, Besser J, Brown E, Allard M, Gerner-Smidt P. 2016. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. Clin Infect Dis 63: 380–386. https://doi.org/10.1093/cid/ciw242.

Janowicz A, De Massis F, Ancora M, et al. Core Genome Multilocus Sequence Typing and Single Nucleotide Polymorphism Analysis in the Epidemiology of *Brucella* melitensis Infections. J Clin Microbiol. 2018;56(9):e00517-18. Published 2018 Aug 27. doi:10.1128/JCM.00517-18

Joint Food and Agriculture Organization of the United Nations (FAO)/World Health Organization (WHO) Expert Committee On BrucellosiS (1986). Technical Report Series 740, Sixth Report. WHO, Geneva, Switzerland.

Jun, LI Z.;Yun, C. B., HaiC., DiaoC. J., Yan Z. H., RiP.D,, Hai J., Li Z., Xu T., Wen K. C., Zhen Y, Guo T. Z. 2013. Molecular Typing of *Brucella* Suis Collected from 1960s to 2010s in China by MLVA and PFGE. Biomed Environ Sci, 2013; 26(6): 504-508

Junemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A, Goesmann A, von Haeseler A, Stoye J, Harmsen D. 2013. Updating benchtop sequencing performance comparison. Nat Biotechnol 31:294–296. https://doi.org/10.1038/nbt.2522.

Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

Laing C.R, Zhanga Y.,ThomasJ. E.,GannonaV. P.J. 2011.Everything at once: Comparative analysis of the genomes of bacterial pathogens. Veterinary Microbiology v. 153, e.1–2, 21, 13-26p.

Laurence M, Hatzis C, Brash DE. Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. Gilbert T, ed. PLoS ONE. 2014;9(5):e97876. doi:10.1371/journal.pone.0097876.

Le Fleche P, Jacques I, Grayon M, Al Dahouk S, Bouchon P, Denoeud F, Nockler K,

Li H. 2013. Aligning sequence reads, clone sequences, and assembly contigs with BWA-MEM. arXiv arXiv:13033997v1 [q-bio.GN]. https://arxiv .org/abs/1303.3997.

Lopez-Goni I, Garcia-Yoldi D, Marin CM, de Miguel MJ, Barquero-Calvo E, et al. (2011) New Bruce-ladder multiplex PCR assay for the biovar typing of *Brucella* suis and the discrimination of *Brucella* suis and *Brucella* canis . Vet Microbiol 154: 152–155.

Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27(21):2957-63

, Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA, 95, 3140-3145.

Maio E, Begeman L, Bisselink Y, van Tulden P, Wiersma L, Hiemstra S, Ruuls R, Grone A, Roest HI, Willemsen P, et al. 2014. Identification and typing of *Brucella* spp. in stranded harbour porpoises (Phocoena phocoena) on the Dutch coast. Vet Microbiol 173:118–124.

Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, RothbergJM,KarchH.2011.Prospectivegenomiccharacterizationofthe German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS One 6:e22751.

Meyer, K.F., and Shaw, E.B. "A comparison of the morphologic, cultural and biochemical characteristics of B. abortus and B. melitensis." J. Infect. Dis. (1920) 27:173-184.

Moreno E. 2014. Retrospective and prospective perspectives on zoonotic brucellosis. Front Microbiol 5:213. https://doi.org/10.3389/fmicb.2014 .00213.

Moreno E., Cloeckaert A., Moriyon I. *Brucella* evolution and taxonomy. Vet. Microbiol. 2002;90:209–227. doi: 10.1016/S0378-1135(02)00210-9

Maquart M, Le Fleche P, Foster G, Tryland M, Ramisse F, Djonne B, Al Dahouk S, Jacques I, Neubauer H, Walravens K, Godfroid J, Cloeckaert A, Vergnaud G. 2009. MLVA-16 typing of 295 marine mammal *Brucella* isolates from different animal and geographic origins identifies 7 major groups within *Brucella* ceti and *Brucella* pinnipedialis. BMC Microbiol 9:145. https://doi.org/10.1186/1471-2180-9-145.

Michiel Vos and Xavier Didelot. A comparison of homologous recombination rates in bacteria and archaea. The ISME journal, 3(2):199, 2009.

Mira N.P., Palma M., Guerreiro J.F., Sá-Correia I.2010.Genome-wide identification of Saccharomyces cerevisiae genes required for tolerance to acetic acid.Microb Cell Fact. 2010 Oct 25;9:79. doi: 10.1186/1475-2859-9-79.

Miller JM. Whole-genome mapping: a new paradigm in strain-typing technology. J Clin Microbiol. 2013;51(4):1066-70.

Nascimento M,SousaA,RamirezM,FranciscoAP,CarricoJA,VazC.2017. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics 33:128–129. https://doi.org/10.1093/bioinformatics/btw582.

Neubauer H, Guilloteau LA, Vergnaud G. 2006. Evaluation and selection of tandem repeat loci for a *Brucella* MLVA typing assay. BMC Microbiol 6:9. https://doi.org/10.1186/1471-2180-6-9.

NIPHE - National Institute for Public Health and Environment – Ministry of Health, Welface and Sport of Netherlands, 2019. Available < https://www.mlva.net/default.asp> , access in Feb.2019.

N.J. Croucher, X. DidelotThe application of genomics to tracing bacterial pathogen transmission Curr Opin Microbiol, 23 (2015), pp. 62-67.

N.J. Loman, C. Constantinidou, J.Z.M. Chan, M. Halachev, M. Sergeant, C.W. Penn, *et al.*High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunityNat Rev Microbiol, 10 (2012), pp. 599-606.

Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, Prjibelski AD, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, Clingenpeel SR, Woyke T, McLean JS, Lasken R, Tesler G, Alekseyev MA, Pevzner PA. 2013. Assembling single-cell genomes and minimetagenomes from chimeric MDA products. J Comput Biol 20: 714–737. https://doi.org/10.1089/cmb.2013.0084.

OIE. 2017. Terrestrial manual: brucellosis (*Brucella* abortus, B. melitensis and B. suis). Infection with B. abortus, B. melitensis and B. suis. Manual of diagnostic tests and vaccines for terrestrial animals. World Organization for Animal Health (OIE), Paris, France. Accessed 25 September 2018.

Paixão, T. A. 2009. Modelo de infecção gastrintestinal e o papel do LPS, urease e sistema de secreção do tipo 4 da *Brucella* melitensis em camundongos. Thesis presented to the Federal University of Minas Gerais, as a partial requirement to obtain the degree of Doctor of Animal Science with emphasis in Animal Pathology.Belo Horizontes, Brasil 2009.

Pappas, G. (2010). The changing *Brucella* ecology: novel reservoirs, new threats. International journal of antimicrobial agents, 36, pp. 8-11.

Pappas G, Panagopoulou P, Christou L, Akritidis N. 2006. *Brucella* as a biological weapon. Cell Mol Life Sci 63:2229–2236. https://doi.org/10 .1007/s00018-006-6311-4.

Patil DP, Bakthavachalu B, Schoenberg DR. Poly(A) polymerase-based poly(A) length assay. Methods Mol Biol. 2014;1125:13–23. Epub 2014/03/05. doi: 10.1007/978-1-62703-971-0_2 ; PubMed Central PMCID: PMC3951053.

Pearce ME, Alikhan NF, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. *Int J Food Microbiol*. 2018;274:1-11.

Pightling AW, Petronella N, Pagotto F. 2014. Choice of reference sequence and assembler for alignment of Listeria monocytogenes shortread sequence data greatly influences rates of error in SNP analyses.

Rajala, E.L. 2016. *Brucella* in Tajikistan - Zoonotic Risks of Urbanized Livestock in a Low-Income Country.Faculty of Veterinary Medicine and Animal Science Department of Clinical Sciences Uppsala.Doctoral Thesis Swedish University ofAgricultural Sciences Uppsala 2016

Reuter S., M.J. Ellington, E.J.P. Cartwright, C.U. Köser, M.E. Török, T. Gouliouris, *et al.*Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiologyJAMA Intern Med, 173 (2013), pp. 1397-1404

Sabat AJ, Hermelijn SM, Akkerboom V, Juliana A, Degener JE, Grundmann H, Friedrich AW. 2017. Complete-genome sequencing elucidates outbreak dynamics of CA-MRSA USA300 (ST8-spa t008) in an academic hospital of Paramaribo, Republic of Suriname. Sci Rep 7:41050. https:// doi.org/10.1038/srep41050.

Sahl JW, Beckstrom-Sternberg SM, Babic-Sternberg J, Gillece JD, Hepp CM, Auerbach RK, Tembe W, Wagner DM, Keim PS, Pearson T. 2015. The In Silico Genotyper (ISG): an open-source pipeline to rapidly identify and annotate nucleotide variants for comparative genomics applications. bioRxivhttps://doi.org/10.1101/015578.

Scholz,H.C., Revilla-Fernández, S., Al Dahouk, S., Hammerl J. A.,Zygmunt, M. S.,Cloeckaert A., Koylass M., Whatmore A. M., Blom J., Vergnaud G., Witte A., Aistleitner K., Hofer E.2016. *Brucella vulpis*sp. nov., isolated from mandibular lymph nodes of red foxes (*Vulpes vulpes*). International Journal of Systematic and Evolutionary Microbiology 66: 2090-2098, doi: 10.1099/ijsem.0.000998

Scholz H, Nöckler K, Göllner C, Bahn P, Vergnaud G, Tomaso H, Al Dahouk S, Kämpfer P, Cloeckaert A, Maquart M, Zygmunt M, Whatmore A, Pfeffer M, Huber B, Busse H, De B. Int J Syst Evol Microbiol 60(4):801-808 doi:10.1099/ijs.0.011148-0.

Scholz H C, Hubalek Z, Sedláček I, Vergnaud G, Tomaso H, Al Dahouk S andNöckler K 2008 *Brucella microti*sp. nov., isolated from the common vole Microtus arvalis. International Journal of Systematic and EvolutionaryMicrobiology,58 (2): 375-382

Severiano A, Pinto FR, Ramirez M, Carriço JA. 2011. Adjusted Wallace coefficient as a measure of congruence between typing methods. J Clin Microbiol 49:3997–4000. https://doi.org/10.1128/JCM.00624-11.

Schurch AC, Arredondo-Alonso S, Willems RJL, Goering RV. 2018. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus geneby-gene-based approaches. Clin Microbiol Infect 24:350–354. https://doi .org/10.1016/j.cmi.2017.12.016.

Seleem MN, Boyle SM, Sriranganathan N. 2010. Brucellosis: a reemerging zoonosis. Vet Microbiol 140:392–398. https://doi.org/10.1016/ j.vetmic.2009.06.021.

Stoenner, H. & Lackman, D. (1957). A new species of *Brucella* isolated from the desert wood rat, Neotoma lepidaThomas. American journal of veterinary research, 18, pp. 947-951.

Sun M, Jing Z, Di D, et al. Multiple Locus Variable-Number Tandem-Repeat and Single-Nucleotide Polymorphism-Based *Brucella* Typing Reveals Multiple Lineages in *Brucella melitensis* Currently Endemic in China. *Front Vet Sci*. 2017;4:215. Published 2017 Dec 14. doi:10.3389/fvets.2017.00215

Tan KK, Tan YC, Chang LY, Lee KW, Nore SS, Yee WY, Mat Isa MN, Jafar FL, Hoh CC, AbuBakar S. 2015. Full genome SNP-based phylogenetic analysis reveals the origin and global spread of *Brucella melitensis*. BMC Genomics 16:93. https://doi.org/10.1186/s12864-015-1294-x.

Thong K-L, Ngeow Y-F, Altwegg M, Navaratnam P, Pang T. Molecular analysis of *Salmonella enteritidis* by pulsed-field gel electrophoresis and ribotyping. J Clin Microbiol. 1995;33:1070–1074.

Tiller, R. V., Gee, J. E., Lonsway, D. R., Gribble, S., Bell, S. C., Jennison, A. V., Bates, J., Coulter, C., Hoffmaster, A. R. & De, B. K. (2010). Identification of an unusual Brucellastrain (BO2) from a lung biopsy in a 52 year-old patient with chronic destructive pneumonia. BMC microbiology,10(23), doi: 10.1186/1471-2180-10-23.

, Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol*., 11, 479-487.

Verger J M, Grimont F, Grimont P A D and Grayon M 1987 Taxonomy of the genus *Brucella*. Ann Inst Pasteur Microbiol 138: 235-238.

Verger J M, Grimont F, Grimont P A and Grayon M 1985 *Brucella*, a monospecific genus as shown by deoxyribonucleic acid hybridization. InternationalJournal of Systematic Bacteriology,35 (3): 292-295.

Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 2013;30(5):614-20.

Whatmore AM. 2009. Current understanding of the genetic diversity of *Brucella*, an expanding genus of zoonotic pathogens. Infect Genet Evol 9:1168–1184. https://doi.org/10.1016/j.meegid.2009.07.001.

Wattam AR, Foster JT, Mane SP, Beckstrom-Sternberg SM, BeckstromSternberg JM, Dickerman AW, Keim P, Pearson T, Shukla M, Ward DV, Williams KP, Sobral BW, Tsolis RM, Whatmore AM, O'Callaghan D. 2014. Comparative phylogenomics and evolution of the *Brucella*e reveal a path to virulence. J Bacteriol 196:920–930. https://doi.org/10.1128/JB .01091-13.

Whatmore AM, Koylass MS, Muchowski J, Edwards-Smallbone J, Gopaul KK, Perrett LL. 2016. Extended multilocus sequence analysis to describe the global population structure of the genus *Brucella*: phylogeography and relationship to biovars. Front Microbiol 7:2049. https://doi.org/10 .3389/fmicb.2016.02049.

Whatmore AM, Davison N, Cloeckaert A, Al Dahouk S, Zygmunt MS, Brew SD, Perrett LL, Koylass MS, Vergnaud G, Quance C, Scholz HC, Dick EJ Jr, Hubbard G, Schlabritz-Loutsevitch NE.2014.*Brucella* papionis sp. nov., isolated from baboons (Papio spp.).Int. J. Syst. Evol. Microbiol. 2014 Dec;64(Pt 12):4120-8. doi: 10.1099/ijs.0.065482-0. Epub 2014 Sep 21.

Whatmore A.M. & Gopaul K.K. (2011). Recent advances in molecular approaches to Brucella diagnostics and epidemiology. In: Brucella: Molecular Microbiology and Genomics, López-Goñi I. & O’Callaghan D., eds, Caister Academic Press, Norfolk, UK, 57–88.

Whatmore A.M. (2009). Current understanding of the genetic diversity of Brucella, an expanding genus of zoonotic pathogens. Infect. Genet. Evol., 9, 1168–1184.

Whatmore AM, Dawson CE, Groussaud P, Koylass MS, King AC, Shankster SJ, Sohn AH, Probert WS, McDonald WL. 2008.Marine mammal *Brucella* genotype associated with zoonotic infection.Emerg Infect Dis. 2008 Mar;14(3):517-8. doi: 10.3201/eid1403.070829.

WHO - World Health Organization 2006 Brucellosis in humans and animals WHO/CDS/EPR/2006.

Wood & Salzberg (2014) Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 2014;15:R46. doi: 10.1186/gb-2014-15-3-r46. [PMC free article] [PubMed] [Cross Ref]

# ANNEXES

## Annex 1

**Table 6:** *Brucella melitensis* isolates analyzed, according to epidemiological data.

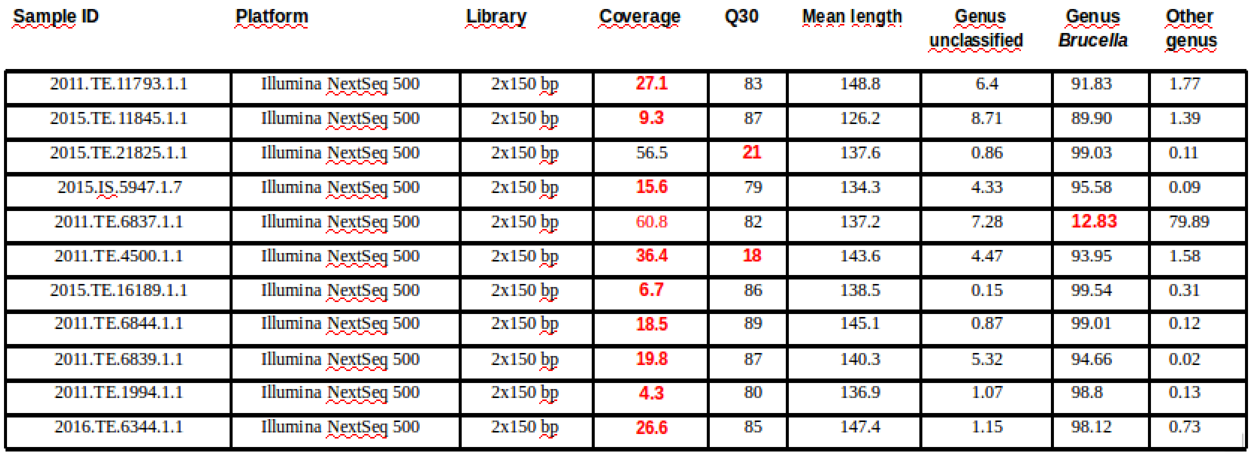
| Samplecode | Sample ID | % Good targets for cgMLST | MLVA profile ID | SNP profile ID | Collection date | Farmcode | Host species | Region | Province | City | SRA accessionNo. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Epidemiologicallylinkedisolates** | | | | | | | | | | | |
| ItBM_1 | 201 5.IS.2566.1.9 | 99.4 | 5 | 1 | 14.04.2015 | 1 | Sheep | Molise | Isernia | RioneroSannitico | SRR6958031 |
| ItBM_2 | 2015.IS.2547.1.11 | 99.4 | 5 | 2 | 14.04.2015 | 1 | Sheep | Molise | Isernia | RioneroSannitico | SRR6958032 |
| ItBM_3 | 2015.IS.3088.1.42 | 99.4 | 1 | 4 | 29.04.2015 | 2 | Sheep | Molise | Isernia | Roccamandolfi | SRR6958033 |
| ItBM_4 | 2015.IS.5088.1.8 | 99.2 | 13 | 5 | 06.05.2015 | 3 | Sheep | Molise | Campobasso | Bojano | SRR6958034 |
| ItBM_5 | 2015.TE.21824.1.1 | 99.3 | 5 | 6 | 16.06.2015 | 4 | Sheep | Lazio | Frosinone | Atina | SRR6958027 |
| ItBM_6 | 2015.CB.2220.1.19 | 99.4 | 3 | 7 | 22.03.2015 | 5 | Sheep | Molise | Campobasso | Castropignano | SRR6958028 |
| ItBM_7 | 2016.TE.17271.1.1 | 99.4 | 5 | 8 | 06.07.2016 | 6 | Cattle | Lazio | Frosinone | Terelle | SRR6958029 |
| ItBM_8 | 2015.CB.3742.1.20 | 99.4 | 2 | 9 | 21.05.2015 | 5 | Sheep | Molise | Campobasso | Castropignano | SRR6958030 |
| ItBM_9 | 2015.IS.2533.1.11 | 99.4 | 5 | 9 | 14.04.2015 | 1 | Sheep | Molise | Isernia | RioneroSannitico | SRR6958035 |
| ItBM_10 | 2015.IS.3088.1.36 | 99.4 | 5 | 9 | 29.04.2015 | 2 | Sheep | Molise | Isernia | Roccamandolfi | SRR6958036 |
| ItBM_11 | 2015.IS.3413.1.7 | 99.4 | 5 | 9 | 30.04.2015 | 2 | Sheep | Molise | Isernia | Roccamandolfi | SRR6957939 |
| ItBM_12 | 2015.TE.16173.1.1 | 99.4 | 3 | 9 | 24.04.2015 | 7 | Goat | Lazio | Frosinone | Sant'apollinare | SRR6957940 |
| ItBM_13 | 2015.TE.16200.1.1 | 99.3 | 3 | 9 | 01.06.2015 | 8 | Sheep | Lazio | Frosinone | Frosinone | SRR6957941 |
| ItBM_14 | 2016.TE.705.1.1 | 99.4 | 5 | 9 | 22.12.2015 | 9 | Sheep | Lazio | Frosinone | Monte San Giovanni Campano | SRR6957942 |
| ItBM_15 | 2015.TE.21825.1.1 | 99.4 | 5 | 10 | 07.07.2015 | 10 | Sheep | Lazio | Frosinone | Casalvieri | SRR6957943 |
| ItBM_16 | 2015.IS.2529.1.14 | 99.4 | 5 | 11 | 14.04.2015 | 1 | Goat | Molise | Isernia | RioneroSannitico | SRR6957944 |
| ItBM_17 | 2015.TE.16181.1.1 | 99.4 | 5 | 12 | 24.04.2015 | 7 | Goat | Lazio | Frosinone | Sant'apollinare | SRR6957945 |
| ItBM_18 | 2015.TE.16510.1.2 | 99.4 | 8 | 13 | 09.07.2014 | 11 | Goat | Lazio | Frosinone | Roccasecca | SRR6957946 |
| ItBM_19 | 2015.TE.16142.1.1 | 99.4 | 14 | 14 | 24.04.2015 | 12 | Sheep | Lazio | Frosinone | Sant'apollinare | SRR6957947 |
| ItBM_20 | 2015.TE.11849.1.3 | 99.4 | 15 | 15 | 28.04.2015 | 13 | Sheep | Lazio | Frosinone | Sant'apollinare | SRR6957948 |
| ItBM_21 | 2015.CB.3742.1.27 | 99.4 | 16 | 16 | 21.05.2015 | 5 | Sheep | Molise | Campobasso | Castropignano | SRR6957966 |
| ItBM_22 | 2015.IS.3088.1.30 | 99.3 | 16 | 16 | 29.04.2015 | 2 | Sheep | Molise | Isernia | Roccamandolfi | SRR6957965 |
| ItBM_23 | 2015.IS.3681.1.8 | 99.4 | 16 | 16 | 30.04.2015 | 2 | Sheep | Molise | Isernia | Roccamandolfi | SRR6957968 |
| ItBM_24 | 2015.TE.16142.1.2 | 99.4 | 16 | 16 | 24.04.2015 | 12 | Sheep | Lazio | Frosinone | Sant'apollinare | SRR69579567 |
| ItBM_25 | 2015.TE.16165.1.2 | 99.4 | 16 | 16 | 24.04.2015 | 7 | Sheep | Lazio | Frosinone | Sant'apollinare | SRR6957962 |
| ItBM_26 | 2015.TE.16189.1,1 | 99.4 | 16 | 16 | 05.05.2015 | 14 | Sheep | Lazio | Frosinone | San Donato Val Di Comino | SRR6957961 |
| ItBM_27 | 2015.TE.16194.1.1 | 99.4 | 16 | 16 | 05.05.2015 | 15 | Sheep | Lazio | Frosinone | Atina | SRR6957964 |
| ItBM_28 | 2016.TE.703.1.2 | 99.4 | 16 | 16 | 22.12.2015 | 16 | Sheep | Lazio | Frosinone | Monte San Giovanni Campano | SRR6957963 |
| ItBM_29 | 2014.TE.16510.1.7 | 99.4 | 17 | 17 | 09.07.2014 | 11 | Goat | Lazio | Frosinone | Roccasecca | SRR6957960 |
| ItBM_30 | 2016.CB.1265.1.7 | 99.4 | 18 | 18 | 23.02.2016 | 17 | Cattle | Molise | Campobasso | San Massimo | SRR6957959 |
| ItBM_31 | 2015.IS.6043.1.8 | 99.4 | 19 | 19 | 24.07.2015 | 18 | Cattle | Molise | Isernia | CantalupoNelSannio | SRR6957977 |
| ItBM_32 | 2015.IS.5947.1.7 | 99.4 | 20 | 20 | 22.07.2015 | 18 | Cattle | Molise | Isernia | CantalupoNelSannio | SRR6957978 |
| ItBM_33 | 2016.TE.17270.1.1 | 99.3 | 21 | 21 | 16.06.2016 | 19 | Sheep | Lazio | Frosinone | Pontecorvo | SRR6957975 |
| ItBM_34 | 2015.TE.11843.1.1 | 99.4 | 22 | 22 | 28.04.2015 | 20 | NA | Lazio | Rome | Rome | SRR6957976 |
| ItBM_35 | 2015.TE.11845.1.1 | 99.5 | 22 | 22 | 28.04.2015 | 21 | NA | Lazio | Rome | Rome | SRR6957973 |
| ItBM_36 | 2015.TE.11847.1.2 | 99.5 | 22 | 22 | 28.04.2015 | 20 | NA | Lazio | Rome | Rome | SRR6957974 |
| ItBM_37 | 2015.TE.11847.1.1 | 99.5 | 23 | 23 | 28.04.2015 | 20 | NA | Lazio | Rome | Rome | SRR6957971 |
| **Isolateswithunknownepidemiological status** | | | | | | | | | | | |
| ItBM_38 | 2011.TE.19513.1.1 | 99.4 | 1 | 1 | 2011 | NA | Human | EmiliaRomagna | Ferrara | Ferrara | SRR6957972 |
| ItBM_39 | 2011.TE.21031.1.1 | 99.9 | 4 | 2 | 2011 | 22 | Goat | Sardinia | Nuoro | Orosei | SRR6957969 |
| ItBM_40 | 2011.TE.3922.1.1 | 99.6 | 9 | 3 | 2011 | 23 | Goat | Campania | Salermo | MontecorvinoPugliano | SRR6957970 |
| ItBM_41 | 2011.TE.6299.1.1 | 99.5 | 10 | 4 | 2011 | NA | Human | Campania | Salermo | Salermo | SRR6957984 |
| ItBM_42 | 2011.TE.1994.1.1 | 99.6 | 10 | 4 | 2011 | 23 | Sheep | Campania | Salermo | MontecorvinoPugliano | SRR6957983 |
| ItBM_43 | 2011.TE.6299.1.2 | 99.6 | 10 | 4 | 2011 | NA | Human | Campania | Salermo | Salermo | SRR6957982 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ItBM_44 | 2011.TE.2461.1.1 | 99.6 | 10 | 5 | 2011 | 23 | Goat | Campania | Salermo | MontecorvinoPugliano | SRR6957981 |
| ItBM_45 | 2011.TE.12841.1.1 | 99.3 | 35 | 6 | 2011 | 24 | Sheep | Calabria | Vibo Valentia | Gerocarne | SRR6957988 |
| ItBM_46 | 2011.TE.12373.1.1 | 99.4 | 36 | 7 | 2011 | 25 | Sheep | Calabria | Vibo Valentia | Rombiolo | SRR6957987 |
| ItBM_47 | 2011.TE.12372.1.1 | 99.4 | 36 | 8 | 2011 | 26 | Sheep | Calabria | Vibo Valentia | Zungri | SRR6957986 |
| ItBM_48 | 2011.TE.12849.1.1 | 99.4 | 36 | 9 | 2011 | 27 | Sheep | Calabria | Vibo Valentia | Mileto | SRR6957985 |
| ItBM_49 | 2011.TE.13541.1.1 | 99.4 | 37 | 10 | 2011 | 28 | Coat | Sicily | Catania | Caltagirone | SRR6957980 |
| ItBM_50 | 2016.TE.6344.1.1 | 99.4 | 42 | 11 | 2016 | NA | Human | Sardinia | NA | NA | SRR6957979 |
| ItBM_51 | 2012.TE.24226.1.1 | 99.4 | 42 | 12 | 2012 | 29 | Sheep | Sicily | Catania | Mineo | SRR6957993 |
| ItBM_52 | 2012.TE.24240.1.1 | 99.4 | 42 | 12 | 2012 | 30 | Sheep | Sicily | Catania | Mineo | SRR6957994 |
| ItBM_53 | 2013.TE.15028.1.1 | 98.9 | 48 | 13 | 2013 | 31 | Sheep | Sicily | Caltanissetta | Niscemi | SRR6957995 |
| ItBM_54 | 2011.TE.4496.1.1 | 99.3 | 31 | 14 | 2011 | 32 | Sheep | Sicily | Ragusa | Scicli | SRR6957996 |
| ItBM_55 | 2011.TE.11814.1.1 | 99.3 | 24 | 15 | 2011 | 32 | Sheep | Sicily | Ragusa | Scicli | SRR6957989 |
| ItBM_56 | 2011.TE.11815.1.1 | 99.2 | 30 | 15 | 2011 | 33 | Sheep | Sicily | Messina | San PierNiceto | SRR6957990 |
| ItBM_57 | 2011.TE.11821.1.1 | 99.3 | 30 | 16 | 2011 | 34 | Sheep | Sicily | Messina | Santa Lucia Del Mela | SRR6957991 |
| ItBM_58 | 2011.TE.4484.1.1 | 99.4 | 32 | 17 | 2011 | 35 | Sheep | Sicily | Agrigento | Ravanusa | SRR6957992 |
| ItBM_59 | 2011.TE.744.1.1 | 99.3 | 50 | 18 | 2011 | 36 | Cattle | Puglia | Faggia | Apricena | SRR6957997 |
| ItBM_60 | 2011.TE.6840.1.1 | 99.3 | 43 | 19 | 2011 | 37 | Sheep | Puglia | Taranto | Massafra | SRR6957998 |
| ItBM_61 | 2011.TE.4500.1.1 | 99.3 | 38 | 20 | 2011 | 38 | Sheep | Sicily | Messina | Messina | SRR6958008 |
| ItBM_62 | 2011.TE.11798.1.1 | 99.3 | 51 | 20 | 2011 | 39 | Sheep | Sicily | Messina | Messina | SRR6958007 |
| ItBM_63 | 2013.TE.15003.1.1 | 99.3 | 44 | 21 | 2013 | 40 | Sheep | Sicily | Catania | San Michele Di Ganzaria | SRR6958010 |
| ItBM_64 | 2011.TE.11842.1.1 | 99.3 | 29 | 22 | 2011 | 41 | Sheep | Sicily | Messina | MontalbanoElicona | SRR6958009 |
| ItBM_65 | 2013.TE.15021.1.1 | 98.9 | 28 | 23 | 2013 | 42 | Sheep | Sicily | Messina | BarcellonaPozzo Di Gotto | SRR6958012 |
| ItBM_66 | 2011.TE.11802.1.1 | 99.4 | 29 | 24 | 2011 | 43 | Goat | Sicily | Messina | BarcellonaPozzo Di Gotto | SRR6958011 |
| ItBM_67 | 2011.TE.11782.1.1 | 99.4 | 41 | 25 | 2011 | 44 | Goat | Sicily | Catania | AciCatena | SRR6958014 |
| ItBM_68 | 2013.TE.15029.1.1 | 99.1 | 39 | 26 | 2013 | 45 | Cattle | Sicily | Messina | Briatico | SRR6958013 |
| ItBM_69 | 2011.TE.21687.1.1 | 99.4 | 33 | 27 | 2011 | 46 | Sheep | Calabria | Catanzaro | ChiaravalleCentrale | SRR6958016 |
| ItBM_70 | 2013.TE.15016.1.1 | 98.9 | 26 | 28 | 2013 | 47 | Sheep | Sicily | Palermo | Corleone | SRR6958015 |
| ItBM_71 | 2011.TE.1169.1.1 | 99.5 | 46 | 29 | 2011 | 48 | Goat | Calabria | Vibo Valentia | Pizzoni | SRR6958039 |
| ItBM_72 | 2011.TE.1171.1.1 | 99.5 | 46 | 30 | 2011 | 49 | Sheep | Calabria | Vibo Valentia | Briatico | SRR6958040 |
| ItBM_73 | 2011.TE.1164.1.1 | 99.5 | 47 | 31 | 2011 | 50 | Sheep | Calabria | Catanzaro | ChiaravalleCentrale | SRR6958037 |
| ItBM_74 | 2011.TE.7556.1.1 | 99.4 | 34 | 32 | 2011 | 51 | Sheep | Puglia | Lecce | Taviano | SRR6958038 |
| ItBM_75 | 2011.TE.2299.1.1 | 99.5 | 45 | 33 | 2011 | 52 | Sheep | Puglia | Lecce | Ugento | SRR6958043 |
| ItBM_76 | 2011.TE.11793.1.1 | 99.4 | 40 | 34 | 2011 | 53 | Goat | Sicily | Caltanissetta | Caltanissetta | SRR6958044 |
| ItBM_77 | 2011.TE.11791.1 | 99.4 | 49 | 35 | 2011 | 54 | Sheep | Sicily | Siracusa | Noto | SRR6958041 |
| ItBM_78 | 2015.TE.26270.1.1 | 99.4 | 5 | 36 | 2015 | NA | Human | Piedmont | Turin | Turin | SRR6958042 |
| ItBM_79 | 2011.TE.11789.1.1 | 99.4 | 6 | 37 | 2011 | 55 | Sheep | Sicily | Ragusa | Santa Croce Camerina | SRR6958045 |
| ItBM_80 | 2013.TE.13528.1.1 | 98.0 | 7 | 38 | 2013 | 56 | Sheep | Sicily | Messina | BarcellonaPozzo Di Gotto | SRR6958046 |
| ItBM_81 | 2013.TE.15005.1.1 | 99.5 | 27 | 39 | 2013 | 57 | Sheep | Sicily | Agrigento | Aragona | SRR6958026 |
| ItBM_82 | 2012.TE.18485.1.1 | 99.5 | 19 | 40 | 2012 | 58 | Sheep | Sicily | Caltanissetta | Caltanissetta | SRR6958025 |
| ItBM_83 | 2011.TE.11828.1.1 | 99.4 | 17 | 41 | 2011 | 59 | Cattle | Sicily | Messina | MontalbanoElicona | SRR6958024 |
| ItBM_84 | 2013.TE.15019.1.1 | 98.2 | 20 | 42 | 2013 | 60 | Sheep | Sicily | Messina | Santa Lucia Del Mela | SRR6958023 |
| ItBM_85 | 2011.TE.4491.1.1 | 99.5 | 23 | 43 | 2011 | 61 | Sheep | Sicily | Messina | San PierNiceto | SRR6958022 |
| ItBM_86 | 2011.TE.11844.1.1 | 99.4 | 22 | 44 | 2011 | 62 | Sheep | Sicily | Messina | MontalbanoElicona | SRR6958021 |
| ItBM_87 | 2011.TE.11805.1.1 | 99.4 | 22 | 45 | 2011 | 63 | Cattle | Sicily | Messina | Floresta | SRR6958020 |
| ItBM_88 | 2011.TE.11803.1.1 | 99.5 | 21 | 46 | 2011 | 64 | Goat | Sicily | Messina | MontalbanoElicna | SRR6958019 |
| ItBM_89 | 2011.TE.4488.1.1 | 99.5 | 24 | 47 | 2011 | 65 | Goat | Sicily | Messina | MontalbanoElicna | SRR6958018 |
| ItBM_90 | 2011.TE.4480.1.1 | 99.5 | 24 | 48 | 2011 | 65 | Sheep | Sicily | Messina | MontalbanoElicna | SRR6958017 |
| ItBM_91 | 2011.TE.11810.1.1 | 99.5 | 24 | 49 | 2011 | 66 | Goat | Sicily | Ragusa | Scicii | SRR6957951 |
| ItBM_92 | 2011.TE.4467.1.1 | 99.5 | 8 | 50 | 2011 | 67 | Sheep | Sicily | Siracusa | Noto | SRR6957952 |
| ItBM_93 | 2011.TE.4471.1.1 | 99.4 | 15 | 50 | 2011 | 68 | Sheep | Sicily | Palermo | Casteldaccia | SRR6957953 |
| ItBM_94 | 2011.TE.4474.1.1 | 99.5 | 15 | 50 | 2011 | 69 | Sheep | Sicily | Catania | AciCatena | SRR6957954 |
| ItBM_95 | 2011.TE.4479.1.1 | 99.5 | 15 | 50 | 2011 | 70 | Sheep | Sicily | Caltanissetta | Niscemi | SRR6957955 |
| ItBM_96 | 2011.TE.4486.1.1 | 99.4 | 15 | 50 | 2011 | 71 | Sheep | Sicily | Palermo | Prizzi | SRR6957956 |
| ItBM_97 | 2011.TE.11826.1.1 | 99.5 | 12 | 51 | 2011 | 69 | Sheep | Sicily | Catania | AciCatena | SRR6957957 |
| ItBM_98 | 2011.TE.4478.1.1 | 99.5 | 15 | 52 | 2011 | 72 | Sheep | Sicily | Messina | Novara Di Sicilia | SRR6957958 |
| ItBM_99 | 2017.TE.3072.1.1 | 99.3 | 18 | 53 | 2017 | NA | Human | Piedmont | Turin | Turin | SRR6957959 |
| ItBM_100 | 2016.TE.6008.1.1 | 99.5 | 11 | 54 | 2016 | NA | Ibex | Aosta Valley | Aosta | GranParadisoNational Park | SRR6957960 |

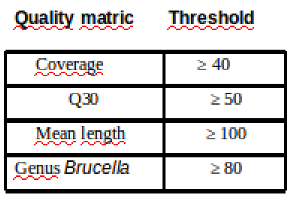
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ItBM_101 | 2011.TE.6837.1.1 | 99.7 | 13 | 55 | 2011 | 73 | Sheep | Puglia | Foggia | Vieste | SRR6958006 |
| ItBM_102 | 2011.TE.6838.1.1 | 99.7 | 13 | 55 | 2011 | 74 | Sheep | Puglia | Foggia | RignanoGarganico | SRR6958005 |
| ItBM_103 | 2011.TE.6839.1.1 | 99.7 | 14 | 55 | 2011 | 75 | Sheep | Puglia | Foggia | San Severo | SRR6958004 |
| ItBM_104 | 2011.TE.6844.1.1 | 99.7 | 16 | 56 | 2011 | 75 | Sheep | Puglia | Foggia | San Severo | SRR6958003 |
| ItBM_105 | 2011.TE.1995.1.1 | 99.6 | 25 | 57 | 2011 | 76 | Goat | Campania | Salermo | Ravello | SRR6958002 |
| ItBM_106 | 2011.TE.6057.1.1 | 99.8 | 3 | 58 | 2011 | 77 | Sheep | Calabria | Cosenza | San Lucido | SRR6958001 |
| ItBM_107 | 2011.TE.6076.1.1 | 98.7 | 3 | 59 | 2011 | 78 | Cattle | Calabria | Cosenza | Mongrassano | SRR6958000 |
| ItBM_108 | 2013.TE.2547.1.1 | 98.7 | 2 | 60 | 2013 | NA | Human | Piedmont | Turin | Turin | SRR6957999 |

## Annex 2:

**Tabela 7:** Quality metrics values of samples discarded during the preliminary data selection. Coverage: theoretical coverage calculated as LN/G (L=read length, N=number of reads, G=size of reference genome), Q30: the percentage of reads with mean quality at least of 30 Q-score (Phred scale), Mean quality: mean Q-score of reads, Genus unclassified: Percentage of reads  unclassified, Genus *Brucella*: Percentage of reads assigned to the genus *Brucella*,  Other Genus: Percentage of reads assigned to the other genus.
The sample discarded was re-sequenced and checked again.



| Sample ID | Platform | Library | Coverage | Q30 | Mean length | Genus unclassified | Genus *Brucella* | Other genus |
|---|---|---|---|---|---|---|---|---|
| 2011.TE.11793.1.1 | Illumina NextSeq 500 | 2x150 bp | **27.1** | 83 | 148.8 | 6.4 | 91.83 | 1.77 |
| 2015.TE. 11845.1.1 | Illumina NextSeq 500 | 2x150 bp | **9.3** | 87 | 126.2 | 8.71 | 89.90 | 1.39 |
| 2015.TE.21825.1.1 | Illumina NextSeq 500 | 2x150 bp | 56.5 | **21** | 137.6 | 0.86 | 99.03 | 0.11 |
| 2015.IS.5947.1.7 | Illumina NextSeq 500 | 2x150 bp | **15.6** | 79 | 134.3 | 4.33 | 95.58 | 0.09 |
| 2011.TE.6837.1.1 | Illumina NextSeq 500 | 2x150 bp | 60.8 | 82 | 137.2 | 7.28 | **12.83** | 79.89 |
| 2011.TE.4500.1.1 | Illumina NextSeq 500 | 2x150 bp | **36.4** | **18** | 143.6 | 4.47 | 93.95 | 1.58 |
| 2015.TE.16189.1.1 | Illumina NextSeq 500 | 2x150 bp | **6.7** | 86 | 138.5 | 0.15 | 99.54 | 0.31 |
| 2011.TE.6844.1.1 | Illumina NextSeq 500 | 2x150 bp | **18.5** | 89 | 145.1 | 0.87 | 99.01 | 0.12 |
| 2011.TE.6839.1.1 | Illumina NextSeq 500 | 2x150 bp | **19.8** | 87 | 140.3 | 5.32 | 94.66 | 0.02 |
| 2011.TE.1994.1.1 | Illumina NextSeq 500 | 2x150 bp | **4.3** | 80 | 136.9 | 1.07 | 98.8 | 0.13 |
| 2016.TE.6344.1.1 | Illumina NextSeq 500 | 2x150 bp | **26.6** | 85 | 147.4 | 1.15 | 98.12 | 0.73 |



| Quality matric | Threshold |
|---|---|
| Coverage | ≥ 40 |
| Q30 | ≥ 50 |
| Mean length | ≥ 100 |
| Genus *Brucella* | ≥ 80 |

## Annex 3

**Table 8:** Quality metrics values observed before (raw) **and after per-processing analysis (trimmed).**N. reads: Number of reads, Mbases: Number of base called (1 Mbase=**1.000.000 base), Mean length: mean length of reads, Q30: the percentage of reads with mean quality at least of 30 Q-score (Phred scale),** Mean quality: mean **Q-score of reads,** Coverage: **theoretical coverage calculated as LN/G (L=read length, N=number of reads, G=size of reference genome),** Overlap: **percentage of paired-end reads with sequence overlap.Reference genome: GenBank accession numbers NC_003317.1 and NC_003318.1.**

| Sample ID | N. reads (raw) | Mbases (raw) | Mean length (raw) | Q30 (raw) | Mean quality (raw) | Coverage (raw) | Overlap (raw) | N. reads (trimmed) | Mbases (trimmed) | Mean length (trimmed) | Q30 (trimmed) | Mean Quality (trimmed) | Coverage (trimmed) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2017.TE.3072.1.1 | 3970672 | 585.66 | 147.5 | 53.77 | 25.42 | 177 | 6.76 | 3058250 | 234.74 | 86.76 | 86.8 | 30.23 | 71 |
| 2016.TE.705.1.1 | 2346386 | 341.84 | 145.69 | 54.54 | 30.35 | 103 | 26.23 | 2095760 | 262.69 | 125.34 | 82.82 | 31.56 | 79 |
| 2016.TE.703.1.2 | 2477158 | 359.93 | 145.3 | 56.77 | 30.53 | 109 | 27.16 | 2239064 | 279.03 | 124.62 | 81.81 | 31.5 | 84 |
| 2016.TE.6344.1.1 | 3119482 | 450.81 | 144.51 | 58.51 | 30.66 | 136 | 28.08 | 2825974 | 351.91 | 124.53 | 83.14 | 31.59 | 106 |
| 2016.TE.6008.1.1 | 3150432 | 448.72 | 142.43 | 61.11 | 30.88 | 135 | 30.76 | 2840624 | 351.98 | 123.91 | 83.45 | 31.66 | 106 |
| 2016.TE.17271.1.1 | 3337942 | 443.33 | 132.82 | 85.73 | 33.48 | 134 | 35.42 | 3179546 | 407.1 | 128.04 | 93.17 | 33.65 | 123 |
| 2016.TE.17270.1.1 | 1025728 | 132.11 | 147.88 | 81.47 | 32.82 | 40 | 20.74 | 987388 | 120.88 | 140.82 | 90.84 | 33.03 | 36 |
| 2016.CB.1265.1.7 | 2113266 | 275.37 | 130.31 | 80.67 | 32.57 | 83 | 35.49 | 2014400 | 245.54 | 121.89 | 91.84 | 32.83 | 74 |
| 2015TE744_1 | 3310548 | 479.27 | 144.77 | 83.78 | 33.77 | 145 | 26.4 | 3128354 | 440.68 | 140.87 | 94.41 | 34.37 | 133 |
| 2015TE6844 | 3459380 | 498.78 | 144.18 | 85.23 | 33.96 | 151 | 31.66 | 3285742 | 460.81 | 140.25 | 94.25 | 34.43 | 139 |
| 2015TE6840 | 3053022 | 442.56 | 144.96 | 83.61 | 33.79 | 134 | 27.47 | 2878020 | 406.48 | 141.24 | 94.7 | 34.42 | 123 |
| 2015TE6839 | 4896028 | 656.2 | 134.03 | 86.66 | 34.33 | 198 | 37.29 | 4610660 | 603.92 | 130.98 | 95.15 | 34.75 | 183 |
| 2015TE6838_1 | 4422454 | 606.3 | 137.1 | 86.75 | 34.33 | 183 | 36.67 | 4181288 | 559.98 | 133.93 | 95.14 | 34.74 | 169 |
| 2015TE6837 | 4564150 | 657.2 | 143.99 | 85.44 | 34.0 | 199 | 31.08 | 4338126 | 608.05 | 140.16 | 94.41 | 34.47 | 184 |
| 2015TE6299_2 | 4196512 | 598.61 | 142.65 | 84.67 | 34.03 | 181 | 32.46 | 3954912 | 550.22 | 139.12 | 94.99 | 34.58 | 166 |
| 2015TE6057 | 2803470 | 407.79 | 145.46 | 82.3 | 33.57 | 123 | 25.34 | 2636888 | 372.9 | 141.42 | 94.38 | 34.26 | 113 |
| 2015TE4500 | 4775458 | 672.24 | 140.77 | 85.62 | 34.09 | 203 | 31.33 | 4522894 | 621.01 | 137.3 | 94.7 | 34.55 | 188 |
| 2015TE4496 | 4875488 | 690.13 | 141.55 | 85.39 | 34.11 | 209 | 33.04 | 4607226 | 637.08 | 138.28 | 95.16 | 34.64 | 193 |
| 2015TE4491 | 4803044 | 673.59 | 140.24 | 86.92 | 34.25 | 204 | 36.6 | 4568456 | 625.99 | 137.02 | 95.16 | 34.68 | 189 |
| 2015TE4488 | 1631636 | 218.61 | 133.98 | 85.4 | 34.19 | 66 | 34.89 | 1528150 | 200.1 | 130.95 | 95.24 | 34.7 | 60 |
| 2015TE4486 | 5619184 | 719.42 | 128.03 | 88.37 | 34.58 | 218 | 39.95 | 5311684 | 666.11 | 125.4 | 95.71 | 34.92 | 201 |
| 2015TE4484 | 5087224 | 685.28 | 134.71 | 87.41 | 34.4 | 207 | 39.75 | 4816890 | 634.01 | 131.62 | 95.34 | 34.79 | 192 |
| 2015TE4480 | 4210124 | 610.58 | 145.03 | 84.96 | 33.93 | 185 | 29.09 | 3997288 | 566.29 | 141.67 | 95.22 | 34.52 | 171 |
| 2015TE4479 | 4238542 | 607.53 | 143.33 | 84.83 | 33.99 | 184 | 32.83 | 4002894 | 559.49 | 139.77 | 94.87 | 34.54 | 169 |
| 2015TE4478 | 5537188 | 788.69 | 142.43 | 86.33 | 34.17 | 238 | 35.09 | 5263930 | 732.27 | 139.11 | 95.07 | 34.64 | 221 |
| 2015TE4474 | 5782474 | 825.7 | 142.79 | 86.45 | 34.19 | 250 | 35.01 | 5503142 | 766.33 | 139.25 | 94.75 | 34.61 | 232 |
| 2015TE4471 | 4492492 | 629.32 | 140.08 | 86.24 | 34.17 | 190 | 36.45 | 4256558 | 581.64 | 136.65 | 94.97 | 34.62 | 176 |
| 2015TE3922 | 4388566 | 616.7 | 140.52 | 86.04 | 34.18 | 186 | 34.63 | 4154112 | 570.58 | 137.35 | 95.27 | 34.67 | 172 |
| 2015TE2461 | 3743000 | 539.48 | 144.13 | 83.44 | 33.77 | 163 | 27.98 | 3524306 | 495.39 | 140.56 | 94.92 | 34.46 | 150 |
| 2015TE2299 | 5096416 | 721.48 | 141.57 | 85.48 | 34.08 | 218 | 33.42 | 4819920 | 665.95 | 138.17 | 94.94 | 34.59 | 201 |
| 2015TE1994 | 4618304 | 653.95 | 141.6 | 85.7 | 34.16 | 198 | 34.24 | 4369656 | 604.37 | 138.31 | 95.23 | 34.67 | 183 |
| 2015TE12849 | 5203116 | 666.65 | 128.12 | 88.29 | 34.58 | 202 | 39.34 | 4915398 | 617.27 | 125.58 | 95.87 | 34.93 | 187 |
| 2015TE12373_1 | 4112772 | 593.07 | 144.2 | 85.85 | 34.0 | 179 | 31.77 | 3920692 | 552.01 | 140.79 | 95.07 | 34.51 | 167 |
| 2015TE12372_1 | 3132928 | 454.83 | 145.18 | 83.42 | 33.76 | 137 | 27.93 | 2955484 | 417.97 | 141.42 | 94.69 | 34.4 | 126 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2015TE11828 | 3270820 | 473.5 | 144.76 | 83.12 | 33.54 | 143 | 28.47 | 3091060 | 433.5 | 140.24 | 93.52 | 34.11 | 131 |
| 2015TE11826 | 5964826 | 760.75 | 127.54 | 87.26 | 34.45 | 230 | 38.09 | 5602628 | 699.63 | 124.88 | 95.67 | 34.87 | 212 |
| 2015TE11821 | 5004090 | 648.7 | 129.63 | 86.91 | 34.38 | 196 | 37.09 | 4706686 | 595.77 | 126.58 | 95.11 | 34.77 | 180 |
| 2015TE11814 | 4382106 | 513.3 | 148.18 | 87.38 | 35.0 | 155 | 39.34 | 2978648 | 406.3 | 128 | 94.51 | 35.0 | 123 |
| 2015TE11810 | 3853966 | 552.7 | 143.41 | 85.41 | 33.94 | 167 | 31.7 | 3666478 | 511.06 | 139.39 | 94.06 | 34.39 | 154 |
| 2015TE11805 | 5732044 | 740.5 | 129.19 | 88.24 | 34.55 | 224 | 39.54 | 5419366 | 684.74 | 126.35 | 95.32 | 34.87 | 207 |
| 2015TE11803 | 4490872 | 642.08 | 142.97 | 85.52 | 34.09 | 194 | 34.07 | 4252154 | 593.65 | 139.61 | 95.07 | 34.61 | 179 |
| 2015TE11802 | 3898134 | 552.05 | 141.62 | 85.53 | 34.11 | 167 | 35.31 | 3686854 | 509.51 | 138.2 | 94.95 | 34.62 | 154 |
| 2015TE11793 | 6163192 | 769.29 | 124.82 | 87.17 | 34.43 | 233 | 38.59 | 5759946 | 703.18 | 122.08 | 95.43 | 34.82 | 213 |
| 2015TE11791 | 6528132 | 853.09 | 130.68 | 87.59 | 34.47 | 258 | 40.76 | 6150602 | 785.33 | 127.68 | 95.5 | 34.84 | 237 |
| 2015TE11782 | 4214726 | 601.98 | 142.83 | 85.33 | 34.01 | 182 | 33.29 | 3994330 | 555.03 | 138.95 | 94.34 | 34.48 | 168 |
| 2015TE1169 | 5230422 | 747.07 | 142.83 | 86.16 | 34.15 | 226 | 34.52 | 4972368 | 693.84 | 139.54 | 95.16 | 34.64 | 210 |
| 2015TE1164 | 4067212 | 587.66 | 144.49 | 85.52 | 33.98 | 178 | 29.23 | 3872702 | 545.57 | 140.87 | 94.81 | 34.49 | 165 |
| 2015.TE.26270.1.1 | 4552956 | 522.74 | 114.81 | 67.67 | 31.64 | 158 | 36.62 | 4031710 | 396.16 | 98.26 | 87.27 | 32.12 | 120 |
| 2015.TE.21825.1.1 | 5409828 | 783.79 | 144.88 | 84.67 | 33.79 | 237 | 31.66 | 5138870 | 720.75 | 140.26 | 93.91 | 34.11 | 218 |
| 2015.TE.21824.1.1 | 4331008 | 633.32 | 146.23 | 84.32 | 33.74 | 191 | 28.92 | 4110048 | 582.24 | 141.66 | 94.08 | 34.09 | 176 |
| 2015.TE.16200.1.1 | 4304332 | 475.5 | 145.56 | 85.21 | 34.7 | 144 | 38..2 | 2665788 | 372.9 | 135 | 93.38 | 35.0 | 113 |
| 2015.TE.16194.1.1 | 5259910 | 660.68 | 125.61 | 88.58 | 34.49 | 200 | 37.75 | 4976754 | 610.95 | 122.76 | 95.43 | 34.66 | 185 |
| 2015.TE.16189.1.1 | 4984232 | 721.25 | 144.71 | 84.59 | 33.79 | 218 | 32.91 | 4719614 | 661.92 | 140.25 | 94.08 | 34.14 | 200 |
| 2015.TE.16181.1.1 | 5363288 | 692.03 | 129.03 | 87.3 | 34.3 | 209 | 36.91 | 5061054 | 635.84 | 125.63 | 94.87 | 34.5 | 192 |
| 2015.TE.16173.1.1 | 5131812 | 737.71 | 143.75 | 85.75 | 33.91 | 223 | 32.74 | 4896612 | 681.32 | 139.14 | 94.12 | 34.17 | 206 |
| 2015.TE.16165.1.2 | 2692506 | 345.63 | 128.37 | 87.14 | 33.6 | 104 | 36.44 | 2575492 | 318.95 | 123.84 | 92.97 | 33.68 | 96 |
| 2015.TE.16142.1.2 | 2754832 | 348.93 | 126.66 | 88.2 | 33.75 | 105 | 37.99 | 2641224 | 324.04 | 122.68 | 93.74 | 33.83 | 98 |
| 2015.TE.16142.1.1 | 4968788 | 711.24 | 143.14 | 85.7 | 33.99 | 215 | 34.05 | 4728916 | 655.87 | 138.69 | 94.39 | 34.26 | 198 |
| 2015.TE.11849.1.3 | 2016494 | 294.6 | 146.1 | 54.51 | 30.35 | 89 | 22.8 | 1816738 | 227.64 | 125.3 | 82.38 | 31.5 | 68 |
| 2015.TE.11847.1.2 | 2707746 | 395.47 | 146.05 | 53.46 | 30.27 | 119 | 23.15 | 2441026 | 304.67 | 124.81 | 81.72 | 31.46 | 92 |
| 2015.TE.11847.1.1 | 4081086 | 575.63 | 141.05 | 63.05 | 31.03 | 174 | 33.03 | 3668508 | 452.45 | 123.33 | 84.07 | 31.72 | 137 |
| 2015.TE.11845.1.1 | 2610484 | 382.15 | 146.39 | 54.17 | 30.32 | 115 | 22.6 | 2356324 | 296.17 | 125.69 | 82.46 | 31.5 | 89 |
| 2015.TE.11843.1.1 | 2641882 | 385.01 | 145.73 | 53.8 | 30.29 | 116 | 24.58 | 2384074 | 295.7 | 124.03 | 81.83 | 31.47 | 89 |
| 2015.IS.6043.1.8 | 2303568 | 285.27 | 123.84 | 65.44 | 31.4 | 86 | 35.96 | 1983644 | 221.18 | 111.5 | 84.18 | 32.22 | 67 |
| 2015.IS.5947.1.7 | 5291286 | 755.62 | 142.8 | 86.39 | 34.03 | 228 | 34.12 | 5050074 | 699.97 | 138.61 | 94.35 | 34.26 | 212 |
| 2015.IS.5088.1.8 | 2314376 | 295.54 | 127.7 | 62.93 | 31.15 | 89 | 35.09 | 1984642 | 226.52 | 114.13 | 83.06 | 32.05 | 68 |
| 2015.IS.3681.1.8 | 5285354 | 763.01 | 144.36 | 85.21 | 33.85 | 231 | 32.28 | 5026224 | 703.56 | 139.98 | 94.23 | 34.16 | 213 |
| 2015.IS.3413.1.7 | 3446562 | 503.24 | 146.01 | 82.96 | 33.53 | 152 | 24.31 | 3260042 | 459.54 | 140.96 | 93.71 | 33.95 | 139 |
| 2015.IS.3088.1.42 | 3083012 | 447.04 | 145.0 | 86.31 | 33.96 | 135 | 32.46 | 2952806 | 414.47 | 140.37 | 93.85 | 34.16 | 125 |
| 2015.IS.3088.1.36 | 3210326 | 450.99 | 140.48 | 86.78 | 34.05 | 136 | 37.57 | 3065768 | 416.2 | 135.76 | 93.57 | 34.21 | 126 |
| 2015.IS.3088.1.30 | 2442758 | 329.19 | 134.76 | 86.35 | 34.14 | 99 | 36.25 | 2307296 | 301.48 | 130.66 | 94.61 | 34.38 | 91 |
| 2015.IS.2566.1.9 | 5673580 | 817.11 | 144.02 | 84.59 | 33.77 | 247 | 33.52 | 5381142 | 748.15 | 139.03 | 93.55 | 34.06 | 226 |
| 2015.IS.2547.1.11 | 4969386 | 716.57 | 144.2 | 85.0 | 33.77 | 217 | 32.55 | 4736844 | 659.62 | 139.25 | 93.67 | 34.03 | 199 |
| 2015.IS.2533.1.11 | 6389166 | 922.38 | 144.37 | 86.14 | 33.83 | 279 | 33.41 | 6122206 | 853.75 | 139.45 | 93.74 | 34.03 | 258 |
| 2015.IS.2529.1.14 | 3446302 | 503.25 | 146.03 | 84.04 | 33.67 | 152 | 25.68 | 3274274 | 461.59 | 140.97 | 93.5 | 33.99 | 139 |
| 2015.CB.3742.1.27 | 3760012 | 545.14 | 144.98 | 84.22 | 33.7 | 165 | 30.0 | 3568660 | 500.16 | 140.15 | 93.7 | 34.02 | 151 |
| 2015.CB.3742.1.20 | 6193750 | 895.6 | 144.6 | 86.29 | 33.94 | 271 | 32.02 | 5923698 | 831.76 | 140.41 | 94.51 | 34.2 | 252 |
| 2015.CB.2220.1.19 | 3137088 | 456.62 | 145.56 | 84.41 | 33.69 | 138 | 30.3 | 2987032 | 418.72 | 140.18 | 92.97 | 33.94 | 126 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2014.TE.16510.1.7 | 3001416 | 430.81 | 143.54 | 59.26 | 30.72 | 130 | 29.3 | 2708838 | 335.35 | 123.8 | 82.95 | 31.6 | 101 |
| 2014.TE.16510.1.2 | 2039900 | 295.32 | 144.77 | 59.16 | 30.71 | 89 | 26.8 | 1855704 | 233.44 | 125.8 | 83.61 | 31.63 | 70 |
| 2013.TE.2547.1.1 | 1233338 | 152.99 | 124.05 | 84.47 | 33.08 | 46 | 38.11 | 1173664 | 138.38 | 117.91 | 93.53 | 33.24 | 41 |
| 2013.TE.15029.1.1 | 1099044 | 149.56 | 127.89 | 86.73 | 33.43 | 45 | 39.45 | 1043602 | 128.28 | 123.34 | 95.02 | 33.61 | 38 |
| 2013.TE.15028.1.1 | 966836 | 136.74 | 123.14 | 88.33 | 33.41 | 41 | 43.28 | 916086 | 104.93 | 128.54 | 95.31 | 33.6 | 31 |
| 2013.TE.15021.1.1 | 4406858 | 557.0 | 149.76 | 85.89 | 34.56 | 168 | 32.91 | 3366678 | 448.4 | 128.84 | 94.66 | 35.7 | 135 |
| 2013.TE.15019.1.1 | 4672188 | 582.9 | 144.54 | 88.14 | 34.1 | 176 | 33.52 | 3473106 | 465.2 | 131.55 | 93.18 | 35.67 | 140 |
| 2013.TE.15016.1.1 | 4105338 | 496.4 | 147.82 | 89.5 | 34.95 | 150 | 35.09 | 2929296 | 395.5 | 123.19 | 94.28 | 34.58 | 119 |
| 2013.TE.15005.1.1 | 4517580 | 553.4 | 147.92 | 89.29 | 34.22 | 167 | 34.05 | 3320110 | 444.2 | 129.97 | 94.62 | 34.4 | 134 |
| 2013.TE.15003.1.1 | 4945346 | 624.3 | 135.14 | 84.1 | 34.6 | 189 | 33.12 | 3811146 | 502.3 | 135.14 | 94.66 | 34.23 | 152 |
| 2013.TE.13528.1.1 | 3452176 | 434.6 | 141.66 | 86.36 | 34.42 | 131 | 36.91 | 2592388 | 347.5 | 127.33 | 93.70 | 35.17 | 105 |
| 2012.TE.24240.1.1 | 2548870 | 368.51 | 144.58 | 55.52 | 30.42 | 111 | 25.57 | 2305444 | 282.24 | 122.42 | 79.87 | 31.36 | 85 |
| 2012.TE.24226.1.1 | 2521078 | 366.58 | 145.41 | 55.34 | 30.41 | 111 | 25.96 | 2283602 | 281.87 | 123.43 | 79.96 | 31.36 | 85 |
| 2012.TE.18485.1.1 | 3065262 | 442.14 | 144.24 | 59.1 | 30.73 | 133 | 27.49 | 2810876 | 360.44 | 128.23 | 86.36 | 31.91 | 109 |
| 2011TE1171 | 7236512 | 1014.84 | 140.24 | 87.03 | 34.17 | 307 | 37.18 | 6894160 | 938.4 | 136.11 | 94.24 | 34.34 | 284 |
| 2011.TE.7556.1.1 | 904522 | 272.26 | 301.0 | 76.03 | 33.21 | 82 | 31.36 | 826542 | 219.89 | 266.04 | 99.4 | 35.76 | 66 |
| 2011.TE.6299.1.1 | 688528 | 207.25 | 301.0 | 71.95 | 32.75 | 62 | 25.85 | 637596 | 167.31 | 262.4 | 99.44 | 35.67 | 50 |
| 2011.TE.6076.1.1 | 921962 | 277.51 | 301.0 | 78.14 | 33.43 | 84 | 39.05 | 792074 | 211.59 | 267.14 | 99.39 | 35.8 | 64 |
| 2011.TE.4467.1.1 | 631386 | 190.05 | 301.0 | 69.89 | 32.52 | 57 | 29.46 | 566592 | 146.78 | 259.05 | 99.31 | 35.56 | 44 |
| 2011.TE.21687.1.1 | 3786988 | 1139.88 | 301.0 | 76.12 | 33.22 | 345 | 32.26 | 3448860 | 917.5 | 266.03 | 99.42 | 35.76 | 278 |
| 2011.TE.21031.1.1 | 927502 | 279.18 | 301.0 | 74.33 | 33.05 | 84 | 29.15 | 856932 | 226.77 | 264.63 | 99.39 | 35.7 | 68 |
| 2011.TE.1995.1.1 | 807702 | 243.12 | 301.0 | 76.76 | 33.27 | 73 | 33.2 | 736528 | 196.42 | 266.69 | 99.47 | 35.78 | 59 |
| 2011.TE.19513.1.1 | 742036 | 223.35 | 301.0 | 78.21 | 33.44 | 67 | 36.58 | 669496 | 179.03 | 267.4 | 99.42 | 35.8 | 54 |
| 2011.TE.13541.1.1 | 2158182 | 310.08 | 143.68 | 55.94 | 30.46 | 93 | 27.04 | 1950038 | 237.25 | 121.66 | 79.81 | 31.36 | 71 |
| 2011.TE.12841.1.1 | 1578496 | 475.13 | 301.0 | 73.88 | 33.42 | 143 | 39.15 | 1309442 | 362.84 | 277.1 | 99.56 | 36.12 | 109 |
| 2011.TE.11844.1.1 | 960812 | 289.2 | 301.0 | 74.96 | 33.08 | 87 | 27.63 | 894098 | 237.09 | 265.18 | 99.45 | 35.72 | 71 |
| 2011.TE.11842.1.1 | 801882 | 241.37 | 301.0 | 75.94 | 33.2 | 73 | 32.77 | 731414 | 194.5 | 265.92 | 99.48 | 35.77 | 58 |
| 2011.TE.11815.1.1 | 854380 | 257.17 | 301.0 | 76.24 | 33.28 | 77 | 36.37 | 756910 | 199.81 | 263.98 | 99.38 | 35.78 | 60 |
| 2011.TE.11798.1.1 | 1101148 | 331.45 | 301.0 | 77.05 | 33.32 | 100 | 34.32 | 1002804 | 267.35 | 266.6 | 99.42 | 35.77 | 81 |
| 2011.TE.11789.1.1 | 860540 | 259.02 | 301.0 | 76.73 | 33.29 | 78 | 31.72 | 790012 | 210.06 | 265.9 | 99.42 | 35.76 | 63 |

## Annex 4

**Table 9:** Results of taxonomic classification at *genus*-level. Genus unclassified: Percentage of reads unclassified, Genus *Brucella*: Percentage of reads assigned to the genus *Brucella*, Other Genus: Percentage of reads assigned to the other genus.

| Sample ID | Genus unclassified | Genus *Brucella* | Other genus | Sample ID | Genus unclassified | Genus *Brucella* | Other genus | Sample ID | Genus unclassified | Genus *Brucella* | Other genus |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 201 5.IS.2566.1.9 | 2.58 | 96.76 | 0.66 | 2011.TE.744.1.1 | 2.54 | 96.88 | 0.58 | 2015.TE.11847.1.1 | 5.67 | 93.35 | 0.98 |
| 2015.IS.2547.1.11 | 2.41 | 96.95 | 0.64 | 2011.TE.6840.1.1 | 2.42 | 97.02 | 0.56 | 2011.TE.19513.1.1 | 0.14 | 99.48 | 0.38 |
| 2015.IS.3088.1.42 | 2.29 | 97.10 | 0.61 | 2011.TE.4500.1.1 | 3.54 | 95.74 | 0.72 | 2011.TE.21031.1.1 | 0.30 | 99.32 | 0.38 |
| 2015.IS.5088.1.8 | 11.50 | 87.35 | 1.15 | 2011.TE.11798.1.1 | 0.19 | 99.33 | 0.48 | 2011.TE.3922.1.1 | 3.20 | 96.08 | 0.72 |
| 2015.TE.21824.1.1 | 1.93 | 97.51 | 0.56 | 2013.TE.15003.1.1 | 3.80 | 95.79 | 0.41 | 2011.TE.6299.1.1 | 0.27 | 98.71 | 1.02 |
| 2015.CB.2220.1.19 | 2.24 | 97.14 | 0.62 | 2011.TE.11842.1.1 | 0.20 | 99.00 | 0.8 | 2011.TE.1994.1.1 | 2.80 | 96.49 | 0.71 |
| 2016.TE.17271.1.1 | 4.85 | 94.24 | 0.91 | 2013.TE.15021.1.1 | 1.02 | 98.34 | 0.64 | 2011.TE.6299.1.2 | 0.27 | 98.71 | 1.02 |
| 2015.CB.3742.1.20 | 2.10 | 97.28 | 0.62 | 2011.TE.11802.1.1 | 2.82 | 96.50 | 0.68 | 2011.TE.2461.1.1 | 2.61 | 96.80 | 0.59 |
| 2015.IS.2533.1.11 | 2.21 | 97.20 | 0.59 | 2011.TE.11782.1.1 | 2.78 | 96.57 | 0.65 | 2011.TE.12841.1.1 | 0.29 | 99.16 | 0.55 |
| 2015.IS.3088.1.36 | 3.44 | 95.81 | 0.75 | 2013.TE.15029.1.1 | 5.41 | 93.69 | 0.9 | 2011.TE.12373.1.1 | 2.19 | 97.20 | 0.61 |
| 2015.IS.3413.1.7 | 2.26 | 97.18 | 0.56 | 2011.TE.21687.1.1 | 0.20 | 99.45 | 0.35 | 2011.TE.12372.1.1 | 2.33 | 97.10 | 0.57 |
| 2015.TE.16173.1.1 | 2.37 | 97.00 | 0.63 | 2013.TE.15016.1.1 | 1.62 | 97.92 | 0.46 | 2011.TE.12849.1.1 | 5.99 | 93.02 | 0.99 |
| 2015.TE.16200.1.1 | 5.20 | 93.88 | 0.92 | 2011.TE.1169.1.1 | 2.54 | 96.82 | 0.64 | 2011.TE.13541.1.1 | 6.98 | 92.11 | 0.91 |
| 2016.TE.705.1.1 | 6.21 | 92.89 | 0.9 | 2011.TE.1171.1.1 | 3.04 | 96.21 | 0.75 | 2016.TE.6344.1.1 | 5.79 | 93.25 | 0.96 |
| 2015.TE.21825.1.1 | 2.25 | 97.14 | 0.61 | 2011.TE.1164.1.1 | 2.32 | 97.08 | 0.6 | 2011.TE.4479.1.1 | 2.62 | 96.74 | 0.64 |
| 2015.IS.2529.1.14 | 2.21 | 97.22 | 0.57 | 2011.TE.7556.1.1 | 0.22 | 99.38 | 0.4 | 2011.TE.4486.1.1 | 5.86 | 93.15 | 0.99 |
| 2015.TE.16181.1.1 | 6.39 | 92.70 | 0.91 | 2011.TE.2299.1.1 | 3.19 | 96.13 | 0.68 | 2011.TE.11826.1.1 | 6.69 | 92.32 | 0.99 |
| 2015.TE.16510.1.2 | 5.74 | 93.46 | 0.8 | 2011.TE.11793.1.1 | 7.85 | 91.14 | 1.01 | 2011.TE.4478.1.1 | 2.58 | 96.76 | 0.66 |
| 2015.TE.16142.1.1 | 2.35 | 97.03 | 0.62 | 2011.TE.11791.1 | 4.99 | 94.08 | 0.93 | 2017.TE.3072.1.1 | 4.44 | 94.10 | 1.46 |
| 2015.TE.11849.1.3 | 6.48 | 92.63 | 0.89 | 2015.TE.26270.1.1 | 8.34 | 90.34 | 1.32 | 2016.TE.6008.1.1 | 6.01 | 93.05 | 0.94 |
| 2015.CB.3742.1.27 | 2.29 | 97.08 | 0.63 | 2011.TE.11789.1.1 | 0.20 | 99.32 | 0.48 | 2011.TE.6837.1.1 | 2.53 | 96.87 | 0.6 |
| 2015.IS.3088.1.30 | 5.13 | 93.95 | 0.92 | 2013.TE.13528.1.1 | 4.89 | 94.27 | 0.84 | 2011.TE.6838.1.1 | 3.80 | 95.40 | 0.8 |
| 2015.IS.3681.1.8 | 2.27 | 97.11 | 0.62 | 2013.TE.15005.1.1 | 1.32 | 98.16 | 0.52 | 2011.TE.6839.1.1 | 4.85 | 94.26 | 0.89 |
| 2015.TE.16142.1.2 | 6.48 | 92.59 | 0.93 | 2012.TE.18485.1.1 | 4.24 | 94.98 | 0.78 | 2011.TE.6844.1.1 | 2.47 | 96.90 | 0.63 |
| 2015.TE.16165.1.2 | 6.45 | 92.64 | 0.91 | 2011.TE.11828.1.1 | 2.52 | 96.90 | 0.58 | 2011.TE.1995.1.1 | 0.17 | 99.46 | 0.37 |
| 2015.TE.16189.1 | 2.26 | 97.15 | 0.59 | 2013.TE.15019.1.1 | 3.87 | 95.56 | 0.57 | 2011.TE.6057.1.1 | 2.42 | 97.01 | 0.57 |
| 2015.TE.16194.1.1 | 7.28 | 91.73 | 0.99 | 2011.TE.4491.1.1 | 2.97 | 96.32 | 0.71 | 2011.TE.6076.1.1 | 0.18 | 99.41 | 0.41 |
| 2016.TE.703.1.2 | 6.14 | 92.96 | 0.9 | 2011.TE.11844.1.1 | 0.20 | 99.32 | 0.48 | 2013.TE.2547.1.1 | 7.51 | 91.43 | 1.06 |
| 2014.TE.16510.1.7 | 6.07 | 92.99 | 0.94 | 2011.TE.11805.1.1 | 5.75 | 93.29 | 0.96 | 2015.TE.11843.1.1 | 6.40 | 92.69 | 0.91 |
| 2016.CB.1265.1.7 | 5.63 | 93.42 | 0.95 | 2011.TE.11803.1.1 | 2.52 | 96.82 | 0.66 | 2015.TE.11845.1.1 | 6.46 | 92.65 | 0.89 |
| 2015.IS.6043.1.8 | 2.26 | 96.49 | 1.25 | 2011.TE.4488.1.1 | 5.16 | 94.02 | 0.82 | 2015.TE.11847.1.2 | 6.69 | 92.41 | 0.9 |
| 2015.IS.5947.1.7 | 2.45 | 96.88 | 0.67 | 2011.TE.4480.1.1 | 2.11 | 97.31 | 0.58 | 2011.TE.4467.1.1 | 0.30 | 99.20 | 0.5 |
| 2016.TE.17270.1.1 | 2.02 | 97.35 | 0.63 | 2011.TE.11810.1.1 | 2.67 | 96.72 | 0.61 | 2011.TE.4471.1.1 | 3.12 | 96.16 | 0.72 |
| 2012.TE.24226.1.1 | 6.65 | 92.46 | 0.89 | 2011.TE.11814.1.1 | 2.83 | 96.48 | 0.69 | 2011.TE.4474.1.1 | 2.54 | 96.80 | 0.66 |
| 2012.TE.24240.1.1 | 6.92 | 92.15 | 0.93 | 2011.TE.11815.1.1 | 0.25 | 98.13 | 1.62 | 2011.TE.4496.1.1 | 3.03 | 96.30 | 0.67 |
| 2013.TE.15028.1.1 | 6.49 | 92.53 | 0.98 | 2011.TE.11821.1.1 | 6.15 | 92.92 | 0.93 | 2011.TE.4484.1.1 | 3.90 | 95.25 | 0.85 |

## Annex 5

**Table 10**: Quality metrics values obtained after De novo assembly. Quality metrics values obtained after *De novo assembly*.
NA75: Length of the smallest sequence in the set that contains the scaffolds whose combined length represents at least 75% of the length of the genome used as reference, LA75: number of scaffolds used to obtain the NG75, Reference %: percentage of the reference genome covered by the alignment of each set of scaffolds, Genes: predicted number of genes.
Reference genome: GenBank accession numbers NC_003317.1 and NC_003318.1

| Sample ID | NA75 | LA75 | Reference % | Genes |
|---|---|---|---|---|
| 2013.TE.15019.1.1 | 36935 | 17 | 99.102 | 3066 |
| 2013.TE.15021.1.1 | 52023 | 12 | 99.111 | 3090 |
| 2013.TE.13528.1.1 | 31287 | 22 | 99.115 | 3057 |
| 2013.TE.15028.1.1 | 52094 | 15 | 99.126 | 3088 |
| 2013.TE.15016.1.1 | 48424 | 14 | 99.128 | 3086 |
| 2015.IS.3088.1.30 | 78135 | 8 | 99.133 | 3099 |
| 2015.TE.26270.1.1 | 84353 | 7 | 99.135 | 3107 |
| 2017.TE.3072.1.1 | 46919 | 11 | 99.139 | 3105 |
| 2011.TE.6844.1.1 | 116098 | 6 | 99.143 | 3110 |
| 2011.TE.6839.1.1 | 116098 | 5 | 99.155 | 3111 |
| 2012.TE.18485.1.1 | 95560 | 6 | 99.157 | 3111 |
| 2011.TE.6840.1.1 | 112517 | 6 | 99.161 | 3110 |
| 2015.TE.11847.1.2 | 116116 | 5 | 99.161 | 3110 |
| 2015.TE.11849.1.3 | 89381 | 8 | 99.162 | 3107 |
| 2016.TE.705.1.1 | 115952 | 6 | 99.162 | 3106 |
| 2013.TE.15003.1.1 | 79300 | 7 | 99.163 | 3105 |
| 2016.TE.17270.1.1 | 79839 | 8 | 99.166 | 3105 |
| 2015.IS.5088.1.8 | 78123 | 9 | 99.168 | 3100 |
| 2012.TE.24240.1.1 | 89381 | 8 | 99.168 | 3106 |
| 2013.TE.15029.1.1 | 89381 | 6 | 99.168 | 3104 |
| 2015.CB.2220.1.19 | 112447 | 6 | 99.168 | 3109 |
| 2011.TE.11815.1.1 | 135779 | 5 | 99.169 | 3110 |
| 2011.TE.6057.1.1 | 104150 | 5 | 99.171 | 3115 |
| 2011.TE.6837.1.1 | 112488 | 6 | 99.172 | 3112 |
| 2016.TE.6008.1.1 | 95529 | 6 | 99.173 | 3112 |
| 2011.TE.6838.1.1 | 95569 | 6 | 99.174 | 3109 |
| 2015.TE.11843.1.1 | 103493 | 6 | 99.174 | 3109 |
| 2011.TE.4484.1.1 | 112465 | 5 | 99.174 | 3110 |
| 2011.TE.2299.1.1 | 116116 | 6 | 99.174 | 3112 |
| 2013.TE.2547.1.1 | 85482 | 7 | 99.175 | 3107 |
| 2013.TE.15005.1.1 | 87377 | 8 | 99.175 | 3109 |
| 2011.TE.4491.1.1 | 104525 | 7 | 99.175 | 3112 |
| 2011.TE.744.1.1 | 104441 | 6 | 99.179 | 3109 |
| 2015.TE.16181.1.1 | 115952 | 5 | 99.193 | 3112 |
| 201 5.IS.2566.1.9 | 115952 | 5 | 99.193 | 3112 |
| 2014.TE.16510.1.7 | 103494 | 6 | 99.194 | 3109 |
| 2015.TE.16165.1.2 | 104525 | 6 | 99.195 | 3110 |
| 2015.TE.16510.1.2 | 89381 | 8 | 99.196 | 3106 |
| 2015.TE.16142.1.1 | 87444 | 7 | 99.197 | 3111 |
| 2012.TE.24226.1.1 | 89381 | 8 | 99.197 | 3107 |
| 2011.TE.11842.1.1 | 139012 | 5 | 99.197 | 3112 |
| 2015.TE.16200.1.1 | 79821 | 8 | 99.198 | 3109 |
| 2011.TE.4480.1.1 | 104525 | 6 | 99.198 | 3113 |
| 2011.TE.4471.1.1 | 139050 | 5 | 99.198 | 3114 |
| 2011.TE.6076.1.1 | 189699 | 5 | 99.198 | 3116 |
| 2016.CB.1265.1.7 | 85559 | 7 | 99.199 | 3109 |
| 2015.IS.5947.1.7 | 104525 | 6 | 99.201 | 3112 |
| 2011.TE.4474.1.1 | 112496 | 5 | 99.202 | 3115 |
| 2011.TE.6299.1.1 | 136525 | 5 | 99.202 | 3116 |
| 2011.TE.11805.1.1 | 104525 | 6 | 99.205 | 3114 |
| 2015.CB.3742.1.27 | 115952 | 6 | 99.205 | 3110 |
| 2011.TE.4467.1.1 | 94406 | 6 | 99.208 | 3114 |
| 2011.TE.6299.1.2 | 116098 | 5 | 99.208 | 3114 |
| 2011.TE.2461.1.1 | 116098 | 6 | 99.212 | 3116 |
| 2011.TE.11798.1.1 | 135771 | 5 | 99.214 | 3111 |
| 2011.TE.19513.1.1 | 104698 | 6 | 99.215 | 3115 |
| 2011.TE.21687.1.1 | 142045 | 6 | 99.217 | 3115 |
| 2011.TE.1995.1.1 | 116098 | 5 | 99.223 | 3116 |
| 2011.TE.11844.1.1 | 161547 | 5 | 99.268 | 3120 |
| 2011.TE.11789.1.1 | 189698 | 5 | 99.268 | 3119 |
| 2011.TE.21031.1.1 | 190425 | 4 | 99.345 | 3133 |
| 2015.IS.6043.1.8 | 74264 | 9 | 99.18 | 3100 |
| 2011.TE.12372.1.1 | 84352 | 7 | 99.19 | 3112 |
| 2011.TE.11810.1.1 | 89380 | 6 | 99.19 | 3112 |
| 2011.TE.11803.1.1 | 116097 | 5 | 99.19 | 3112 |
| 2011.TE.11793.1.1 | 89381 | 6 | 99.2 | 3112 |
| 2011.TE.11821.1.1 | 104441 | 6 | 99.182 | 3110 |
| 2015.IS.3088.1.42 | 112447 | 6 | 99.182 | 3109 |
| 2015.TE.11845.1.1 | 103680 | 7 | 99.184 | 3107 |
| 2011.TE.4488.1.1 | 79811 | 9 | 99.185 | 3108 |
| 2015.TE.16142.1.2 | 104525 | 7 | 99.185 | 3109 |
| 2011.TE.11814.1.1 | 115952 | 6 | 99.185 | 3110 |
| 2011.TE.1994.1.1 | 139049 | 5 | 99.185 | 3111 |
| 2016.TE.6344.1.1 | 89381 | 7 | 99.186 | 3109 |
| 2015.IS.3088.1.36 | 104525 | 7 | 99.186 | 3108 |
| 2011.TE.1169.1.1 | 116116 | 5 | 99.186 | 3113 |
| 2011.TE.7556.1.1 | 139050 | 5 | 99.186 | 3114 |
| 2015.TE.16173.1.1 | 115952 | 5 | 99.187 | 3110 |
| 2011.TE.13541.1.1 | 84360 | 8 | 99.188 | 3106 |
| 2011.TE.1164.1.1 | 89380 | 6 | 99.188 | 3113 |
| 2016.TE.703.1.2 | 104525 | 7 | 99.188 | 3109 |
| 2011.TE.12841.1.1 | 139012 | 5 | 99.188 | 3112 |
| 2016.TE.17271.1.1 | 104525 | 6 | 99.189 | 3110 |
| 2011.TE.11826.1.1 | 112583 | 6 | 99.189 | 3115 |
| 2011.TE.3922.1.1 | 116098 | 5 | 99.189 | 3112 |
| 2011.TE.4479.1.1 | 139050 | 6 | 99.189 | 3114 |
| 2015.IS.3413.1.7 | 89381 | 6 | 99.191 | 3112 |
| 2015.TE.21824.1.1 | 89381 | 6 | 99.191 | 3110 |
| 2011.TE.4496.1.1 | 115952 | 5 | 99.191 | 3111 |
| 2015.IS.2529.1.14 | 89381 | 6 | 99.192 | 3108 |
| 2011.TE.4500.1.1 | 115952 | 5 | 99.192 | 3112 |
| 2011.TE.1171.1.1 | 139050 | 5 | 99.192 | 3112 |
| 2015.TE.21825.1.1 | 104525 | 6 | 99.193 | 3111 |
| 2011.TE.12373.1.1 | 104525 | 6 | 99.193 | 3112 |
| 2011.TE.11802.1.1 | 104525 | 6 | 99.193 | 3111 |
| 2011.TE.4486.1.1 | 104525 | 6 | 99.193 | 3114 |
| 2011.TE.11828.1.1 | 116097 | 5 | 99.189 | 3114 |
| 2015.IS.2533.1.11 | 112448 | 6 | 99.21 | 3111 |
| 2015.CB.3742.1.20 | 115952 | 5 | 99.21 | 3112 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2011.TE.4478.1.1 | 139050 | 5 | 99.179 | 3114 | | 2015.IS.2547.1.11 | 104525 | 6 | 99.2 | 3112 | | 2015.IS.3681.1.8 | 115952 | 5 | 99.21 | 3111 |
| 2015.TE.11847.1.1 | 104525 | 5 | 99.181 | 3108 | | 2011.TE.11791.1 | 104525 | 6 | 99.2 | 3111 | | 2015.TE.16194.1.1 | 104525 | 6 | 99.21 | 3112 |
| 2011.TE.11782.1.1 | 115952 | 5 | 99.181 | 3111 | | 2015.TE.16189.1.1 | 115952 | 5 | 99.2 | 3111 | | 2011.TE.12849.1.1 | 104525 | 7 | 99.21 | 3112 |

## Annex 6

**Table 11:** Results of taxonomic classification at species-level. Results of taxonomic classification at species-level.% k-mers : Percentage of all k-mers assigned to template, Template: Reference genome used as template of a species. The reference genome (NC_003317.1 and NC_003318.1) was the *Brucella melitensis* template in the KmerFinder database and labeled as GCF_000007125.1.

| Sample ID | % k-mers | Template | Sample ID | % k-mers | Template | Sample ID | % k-mers | Template |
|---|---|---|---|---|---|---|---|---|
| 201 5.IS.2566.1.9 | 99.99 | GCF_000007125.1 | 2013.TE.15016.1.1 | 99.90 | GCF_000007125.1 | 2015.TE.11845.1.1 | 99.93 | GCF_000007125.1 |
| 2015.IS.2547.1.11 | 99.99 | GCF_000007125.1 | 2011.TE.1169.1.1 | 99.97 | GCF_000007125.1 | 2015.TE.11847.1.2 | 99.93 | GCF_000007125.1 |
| 2015.IS.3088.1.42 | 99.96 | GCF_000007125.1 | 2011.TE.1171.1.1 | 99.97 | GCF_000007125.1 | 2015.TE.11847.1.1 | 99.93 | GCF_000007125.1 |
| 2015.IS.5088.1.8 | 99.99 | GCF_000007125.1 | 2011.TE.1164.1.1 | 99.97 | GCF_000007125.1 | 2011.TE.19513.1.1 | 99.95 | GCF_000007125.1 |
| 2015.TE.21824.1.1 | 99.97 | GCF_000007125.1 | 2011.TE.7556.1.1 | 99.97 | GCF_000007125.1 | 2011.TE.21031.1.1 | 100.0 | GCF_000007125.1 |
| 2015.CB.2220.1.19 | 99.98 | GCF_000007125.1 | 2011.TE.2299.1.1 | 99.97 | GCF_000007125.1 | 2011.TE.3922.1.1 | 99.89 | GCF_000007125.1 |
| 2016.TE.17271.1.1 | 99.99 | GCF_000007125.1 | 2011.TE.11793.1.1 | 99.93 | GCF_000007125.1 | 2011.TE.6299.1.1 | 99.86 | GCF_000007125.1 |
| 2015.CB.3742.1.20 | 99.99 | GCF_000007125.1 | 2011.TE.11791.1 | 99.95 | GCF_000007125.1 | 2011.TE.1994.1.1 | 99.86 | GCF_000007125.1 |
| 2015.IS.2533.1.11 | 99.99 | GCF_000007125.1 | 2015.TE.26270.1.1 | 99.99 | GCF_000007125.1 | 2011.TE.6299.1.2 | 99.86 | GCF_000007125.1 |
| 2015.IS.3088.1.36 | 99.95 | GCF_000007125.1 | 2011.TE.11789.1.1 | 100.0 | GCF_000007125.1 | 2011.TE.2461.1.1 | 99.86 | GCF_000007125.1 |
| 2015.IS.3413.1.7 | 99.99 | GCF_000007125.1 | 2013.TE.13528.1.1 | 99.96 | GCF_000007125.1 | 2011.TE.12841.1.1 | 99.97 | GCF_000007125.1 |
| 2015.TE.16173.1.1 | 99.99 | GCF_000007125.1 | 2013.TE.15005.1.1 | 99.95 | GCF_000007125.1 | 2011.TE.12373.1.1 | 99.97 | GCF_000007125.1 |
| 2015.TE.16200.1.1 | 99.98 | GCF_000007125.1 | 2012.TE.18485.1.1 | 99.88 | GCF_000007125.1 | 2011.TE.12372.1.1 | 99.97 | GCF_000007125.1 |
| 2016.TE.705.1.1 | 99.99 | GCF_000007125.1 | 2011.TE.11828.1.1 | 100.0 | GCF_000007125.1 | 2011.TE.12849.1.1 | 99.97 | GCF_000007125.1 |
| 2015.TE.21825.1.1 | 99.99 | GCF_000007125.1 | 2013.TE.15019.1.1 | 99.99 | GCF_000007125.1 | 2011.TE.13541.1.1 | 99.99 | GCF_000007125.1 |
| 2015.IS.2529.1.14 | 99.97 | GCF_000007125.1 | 2011.TE.4491.1.1 | 99.99 | GCF_000007125.1 | 2016.TE.6344.1.1 | 99.99 | GCF_000007125.1 |
| 2015.TE.16181.1.1 | 99.99 | GCF_000007125.1 | 2011.TE.11844.1.1 | 99.99 | GCF_000007125.1 | 2012.TE.24226.1.1 | 100.0 | GCF_000007125.1 |
| 2015.TE.16510.1.2 | 99.99 | GCF_000007125.1 | 2011.TE.11805.1.1 | 99.99 | GCF_000007125.1 | 2012.TE.24240.1.1 | 99.99 | GCF_000007125.1 |
| 2015.TE.16142.1.1 | 99.99 | GCF_000007125.1 | 2011.TE.11803.1.1 | 99.99 | GCF_000007125.1 | 2013.TE.15028.1.1 | 99.93 | GCF_000007125.1 |
| 2015.TE.11849.1.3 | 99.99 | GCF_000007125.1 | 2011.TE.4488.1.1 | 99.99 | GCF_000007125.1 | 2011.TE.4496.1.1 | 100.0 | GCF_000007125.1 |
| 2015.CB.3742.1.27 | 99.99 | GCF_000007125.1 | 2011.TE.4480.1.1 | 99.97 | GCF_000007125.1 | 2011.TE.6844.1.1 | 99.91 | GCF_000007125.1 |
| 2015.IS.3088.1.30 | 99.95 | GCF_000007125.1 | 2011.TE.11810.1.1 | 99.99 | GCF_000007125.1 | 2011.TE.1995.1.1 | 99.88 | GCF_000007125.1 |
| 2015.IS.3681.1.8 | 99.99 | GCF_000007125.1 | 2011.TE.4467.1.1 | 99.95 | GCF_000007125.1 | 2011.TE.6057.1.1 | 99.89 | GCF_000007125.1 |
| 2015.TE.16142.1.2 | 99.99 | GCF_000007125.1 | 2011.TE.4471.1.1 | 99.95 | GCF_000007125.1 | 2011.TE.6076.1.1 | 99.91 | GCF_000007125.1 |
| 2015.TE.16165.1.2 | 99.99 | GCF_000007125.1 | 2011.TE.4474.1.1 | 99.95 | GCF_000007125.1 | 2013.TE.2547.1.1 | 99.91 | GCF_000007125.1 |
| 2015.TE.16189.1 | 99.99 | GCF_000007125.1 | 2011.TE.4479.1.1 | 99.95 | GCF_000007125.1 | 2011.TE.11821.1.1 | 99.96 | GCF_000007125.1 |
| 2015.TE.16194.1.1 | 99.99 | GCF_000007125.1 | 2011.TE.4486.1.1 | 99.93 | GCF_000007125.1 | 2011.TE.4484.1.1 | 99.96 | GCF_000007125.1 |
| 2016.TE.703.1.2 | 99.96 | GCF_000007125.1 | 2011.TE.11826.1.1 | 99.96 | GCF_000007125.1 | 2011.TE.744.1.1 | 99.99 | GCF_000007125.1 |
| 2014.TE.16510.1.7 | 99.99 | GCF_000007125.1 | 2011.TE.4478.1.1 | 99.95 | GCF_000007125.1 | 2011.TE.6840.1.1 | 100.0 | GCF_000007125.1 |
| 2016.CB.1265.1.7 | 99.99 | GCF_000007125.1 | 2017.TE.3072.1.1 | 99.99 | GCF_000007125.1 | 2011.TE.4500.1.1 | 99.97 | GCF_000007125.1 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2015.IS.6043.1.8 | 99.96 | GCF_000007125.1 | | 2016.TE.6008.1.1 | 99.86 | GCF_000007125.1 | | 2011.TE.11798.1.1 | 99.97 | GCF_000007125.1 |
| 2015.IS.5947.1.7 | 99.99 | GCF_000007125.1 | | 2011.TE.6837.1.1 | 99.99 | GCF_000007125.1 | | 2013.TE.15003.1.1 | 99.99 | GCF_000007125.1 |
| 2016.TE.17270.1.1 | 99.97 | GCF_000007125.1 | | 2011.TE.6838.1.1 | 99.91 | GCF_000007125.1 | | 2011.TE.11842.1.1 | 99.92 | GCF_000007125.1 |
| 2015.TE.11843.1.1 | 99.93 | GCF_000007125.1 | | 2011.TE.6839.1.1 | 99.91 | GCF_000007125.1 | | 2013.TE.15021.1.1 | 99.99 | GCF_000007125.1 |
| 2013.TE.15029.1.1 | 99.95 | GCF_000007125.1 | | 2011.TE.11814.1.1 | 99.99 | GCF_000007125.1 | | 2011.TE.11802.1.1 | 99.97 | GCF_000007125.1 |
| 2011.TE.21687.1.1 | 99.95 | GCF_000007125.1 | | 2011.TE.11815.1.1 | 99.97 | GCF_000007125.1 | | 2011.TE.11782.1.1 | 99.96 | GCF_000007125.1 |