Infection, Genetics and Evolution 56 (2017) 92–98

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

Identification of genetic variants of *Brucella* spp. through genome-wide association studies

Jagadesan Sankarasubramanian[a], Udayakumar S. Vishnu[a], Paramasamy Gunasekaran[b], Jeyaprakash Rajendhran[a,*]

[a] Department of Genetics, School of Biological Sciences, Madurai Kamaraj University, Madurai 625021, Tamil Nadu, India
[b] Vice-Chancellor, VIT University, Bhopal,2016a, Madhya Pradesh, India

ARTICLE INFO

Keywords:
Brucella
SNP
Phylogeny
Alignment-free
Pan-GWAS
SNP-GWAS

ABSTRACT

Brucellosis is an important zoonotic disease caused by *Brucella* spp. We present a phylogeny of 552 strains based on genome-wide single nucleotide polymorphisms (SNPs) determined by an alignment-free k-mer approach. A total of 138,029 SNPs were identified from 552 *Brucella* genomes. Of these, 31,152 and 106,877 were core and non-core SNPs, respectively. Based on pan-genome analysis 11,937 and 972 genes were identified as pan and core genome, respectively. The pan-genome-wide analysis studies (Pan-GWAS) could not identify the group-specific variants in *Brucella* spp. Therefore, we focused on SNP based genome-wide association studies (SNP-GWAS) to identify the species-specific genetic determinants in *Brucella* spp. Phylogenetic tree representing eleven recognized *Brucella* spp. showed 16 major lineages. We identified 143 species-specific SNPs in *Brucella abortus* that are conserved in 311 *B. abortus* genomes. Of these, 141 species-specific SNPs were confined in the positively significant SNPs of *B. abortus* using SNP-GWAS. Since conserved in all the *B. abortus* genomes studied, these SNPs might have originated very early during the evolution of *B. abortus* and might be responsible for the evolution of *B. abortus* with cattle as the preferred host. Similarly, we identified 383 species-specific SNPs conserved in 132 *Brucella melitensis* genomes. Of these 379 species-specific SNPs were identified as positively associated using GWAS. Interestingly, > 98% of the SNPs that are significantly, positively associated with the traits showed 100% sensitivity and 100% specificity. These identified species-specific core-SNPs identified in *Brucella* genomes could be responsible for the speciation and their respective host adaptation.

1. Introduction

Brucellosis is a worldwide zoonotic disease that affects various mammals including humans. *Brucella* is a Gram-negative unsporulated and uncapsulated short bacillus that behaves as a facultative intracellular pathogen (Corbel and Brinley-Morgan, 1984). The frequent sources of human *Brucella* infections are farm animals such as cattle, sheep, goats and pigs. *Brucella* infections rarely occur in buffalo, camel, horse, reindeer and yak (Moreno, 2014). *Brucella* can also infect marine mammals such as dolphin, porpoise, and seal (Jensen et al., 1999). The genus *Brucella* has been classified into eleven species, designated on the basis of differences in pathogenicity and host preference such as *B. melitensis* (goats and sheep), *B. abortus* (cattle and bison), *B. suis* (swine, hares, rodents, and reindeer), *Brucella ovis* (sheep), *B. canis* (dogs), *Brucella neotomae* (wood rats), *Brucella ceti* (cetaceans) and *Brucella pinnipedialis* (pinnipeds), *Brucella microti* (common vole) and *Brucella inopinata* (human) (Audic et al., 2009; Lopez-Goni et al., 2008). For many years, the genus *Brucella* was composed of only six classical species. Later, several marine mammals isolates of *Brucella* spp., which do not fit into the classical species were reported (Cloeckaert et al., 2001; Jahans et al., 1997; Moreno et al., 2002). These marine mammal isolates formed separate genetic lineage and grouped along with *B. neotomae* and *B. suis* biovar 5 (Foster et al., 2002; Foster et al., 2007; Foster et al., 2009; Whatmore, 2009). These species were included in the genus *Brucella* since 2007 (Audic et al., 2011). *B. microti* isolated from common vole was described in 2008 (Audic et al., 2011; Scholz et al., 2008). Recently, *Brucella vulpis* was isolated from the lymph nodes of red foxes (Scholz et al., 2016). Phylogenetic analysis showed that *B. microti* is the closest relatives of *B. ovis* (Audic et al., 2011). *B. abortus* is classified into biovars 1–7; *B. melitensis* is classified into biovars 1–3 (Lopez-Goni et al., 2008; Seleem et al., 2010); *B. suis* is classified into biovars 1–5 (Garin-Bastuji and Hars, 2001). The *Brucella* genome is composed of two circular chromosomes of approximately 2.1 and 1.2 Mb (Michaux-Charachon et al., 1997). Both chromosomes share

* Corresponding author.
E-mail address: jrajendhran@nrcbsmku.org (J. Rajendhran).

https://doi.org/10.1016/j.meegid.2017.11.016
Received 11 October 2017; Received in revised form 8 November 2017; Accepted 14 November 2017
Available online 15 November 2017

similar GC content, a similar proportion of coding regions and equivalent housekeeping gene distribution (Ficht, 2010).

Rather than constructing phylogenetic inferences from a small portion of the genome (like MLST, MLVA), entire genomes can now be compared. Whole-genome comparisons and phylogenetic analysis of *Brucella* have only been done on a limited scale (Foster et al., 2009; Wattam et al., 2014). Among the genomes with conserved genetic loci, genome-wide SNPs appear to be a better choice for a phylogenetic analysis because of their coverage of the whole genome and the inclusion of intergenic regions (Morin et al., 2004). Recently, we reported the biovar-specific differentiation of *B. suis* by phylogenetic analysis based on genome-wide SNPs (Sankarasubramanian et al., 2016a). Genome-wide association studies (GWAS) are employed for the identification of genome-wide genetic variants associated with a trait (Klein et al., 2005). The GWAS has identified several variants related to human diseases such as type 2 diabetes (McCarthy and Zeggini, 2009), rheumatoid arthritis (Bossini-Castillo et al., 2014), Crohn's disease (Lee and Parkes, 2011), etc. Recently, GWAS has been extended to microorganisms also. The GWAS in bacteria has been focused on the virulence and antibiotic resistance in clinical pathogens (Read and Massey, 2014). We have applied GWAS to identify the *Brucella* species-specific variants. We have used two independent approaches namely, pan-GWAS and SNP-GWAS. Also, we present an alignment-free SNP-based phylogenetic analysis of 552 *Brucella* strains from various geographical origins and report the identification of core species-specific SNPs using SNP-GWAS.

2. Materials and methods

2.1. Data retrieval

All the *Brucella* genome sequences available in NCBI and BrucellaBase (Sankarasubramanian et al., 2016a, 2016b) were used in this study. Whole genome sequences (complete and draft) of 552 *Brucella* strains belonging to 11 different *Brucella* species (311 *B. abortus*, 20 *B. canis*, 7 *B. ceti*, 2 *B. inopinata*, 132 *B. melitensis*, 1 *B. microti*, 1 *B. neotomae*, 16 *B. ovis*, 4 *B. pinnipedialis*, 56 *B. suis* and 2 *B. vulpis*).

2.2. Pan-genome inference of Brucella spp.

The pan-genome of 552 *Brucella* strains was inferred with rapid, large-scale prokaryote pan-genome analysis (Roary) version 3.6.1 (Page et al., 2015). The GenBank annotations were converted to gff using BioPerl script (bp_genbank2gff3.pl), which was fed to Roary as input; in turn, Roary generated a gene presence/absence matrix. A multi-FASTA alignment was made for core genes using PRANK version 0. 130, PRANK version 130410 and a phylogenetic tree was constructed based on the presence and absence of core and accessory genes among the *Brucella* strains using FastTree (parameter –nt –gtr) version 2.1.9 (Price et al., 2010). A gene present in at least 99% of the tested strains was considered as a core gene. A non-core gene present in at least two *Brucella* strains was considered as an accessory gene, and a gene found in only one *Brucella* strain was considered as a unique gene. The synteny plot was generated (roary_plot.py) based on the core, accessory and unique genes present in 552 *Brucella* genomes using gene presence/absence matrix and phylogenetic tree.

2.3. Prediction of genome-wide SNPs and phylogenetic analysis

Rather than using one genome as the reference genome to predict the SNPs, we opted to predict the genome-wide SNPs using an alignment free method. Genome-wide SNPs were predicted using kSNP v.3 (Gardner et al., 2015), a tool that identifies SNPs without genome alignment and annotates the identified SNPs. Concatenations of *Brucella* SNPs were gathered in a data matrix (95% majority SNP matrix) using kSNP v.3. The SNP locations were determined using *B. abortus* 2308 genome (Accession no: NC_007618, NC_007624) as the reference strain. The k-mer values used in kSNP range from 13 to 31; for bacteria, either 19 or 21 is widely used as the k-mer value (Gardner and Hall, 2013). We have screened various k-mer values using Kchooser tool in the kSNP software package and found that the optimum value of k-mer for *Brucella* genomes as 21. Therefore, we used kSNP k-mer size as 21, which determine SNPs on a codon flanked by ten bases on each side that are conserved (matching completely) among at least two genomes. The kSNP tool can build a phylogenetic tree using Maximum Likelihood, Neighbor Joining, or parsimony approaches based on either all SNPs or only core SNPs. Also, it reports the number of SNPs unique to each node. SNPs predicted on a k-mer conserved in all the genome sequences (552 genomes) are considered as core-SNPs, which was extracted using kSNP script (core_SNP3). An SNP identified in a genomic region, which is missing in one or more genomes is referred as a noncore SNP. Based on the distributions of core-SNPs among 552 genomes, a phylogenetic tree was constructed using Maximum Likelihood method under kSNP tool (FastTreeMP).

2.4. Annotation of SNPs

kSNP3.0 annotates SNPs by downloading the GenBank annotation files (.gbk format). SNPs were annotated using *B. abortus* 2308 (Accession no: NC_007618, NC_007624) as the stranded reference strains using kSNP script (annotate_SNPS_from_genebank_files3). Based on the SNPs annotation, types of SNPs such as non-synonymous, synonymous, or in the intergenic regions were identified. Synonymous SNPs do not change the amino acid sequence, whereas non-synonymous SNPs change the amino acid sequence of the protein.

2.5. Tree visualization

The phylogenetic tree was visualized using Interactive Tree of Life (iTOL) (Letunic and Bork, 2016).

2.6. Genotype-phenotype association

Scoary version 1.6.15 (Brynildsrud et al., 2016) was used to associate patterns of presence/absence of gene or SNP with particular phenotypes (traits) through pan-GWAS and SNP-GWAS analyses, respectively. Scoary needs two basic input files: the gene presence/absence matrix, augmented with gene presence/absence information and a binary trait matrix created using the particular strain. Based on the whole genome- and SNP-based phylogenetic analyses, the *Brucella* spp. were clustered into 16 major groups, and each group was considered as an individual trait. For the gene-based analysis, output file generated from Roary (gene presence/absence) was used to find out the genes positively associated with each group. Likewise, the predicted SNPs from the 552 *Brucella* genomes were further converted to Scoary format using (vcf2scoary), which was then used to generate the presence/absence matrix file. Scoary analysis was performed with 1000 permutation replicates, and the genes were reported as significantly associated with a trait if they attained a naive P-value < 0.05, a Benjamini-Hochberg-corrected P-value < 0.05, and an empirical P-value < 0.05. From the pan-GWAS and SNP-GWAS, the positively and negatively associated genes and SNPs were identified by the Odds ratio of > 1 and < 1, respectively.

3. Result

A total of 552 *Brucella* genome sequences including 109 complete and 443 draft genomes, representing eleven species of *Brucella* (Supplementary Table S1) was taken for the analysis. We performed pan-GWAS and SNP-GWAS to find out the significantly associated genes and SNPs with 16 phylogenetic groups of *Brucella*. Of these, nine are species-specific, five are *B. suis* biovar-specific, and two *B. ceti* host-