frontiers
in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 07 February 2017
doi: 10.3389/fcimb.2017.00031

Persistence of Functional Protein Domains in Mycoplasma Species and their Role in Host Specificity and Synthetic Minimal Life

Tjerko Kamminga[1,2], *Jasper J. Koehorst*[1], *Paul Vermeij*[3], *Simen-Jan Slagman*[2], *Vitor A. P. Martins dos Santos*[1], *Jetta J. E. Bijlsma*[3] *and Peter J. Schaap*[1]*

[1] *Laboratory of Systems and Synthetic Biology, Department of Agrotechnology and Food Sciences, Wageningen University and Research, Wageningen, Netherlands,* [2] *Bioprocess Technology and Support, MSD Animal Health, Boxmeer, Netherlands,* [3] *Discovery and Technology, MSD Animal Health, Boxmeer, Netherlands*

Mycoplasmas are the smallest self-replicating organisms and obligate parasites of a specific vertebrate host. An in-depth analysis of the functional capabilities of mycoplasma species is fundamental to understand how some of simplest forms of life on Earth succeeded in subverting complex hosts with highly sophisticated immune systems. In this study we present a genome-scale comparison, focused on identification of functional protein domains, of 80 publically available mycoplasma genomes which were consistently re-annotated using a standardized annotation pipeline embedded in a semantic framework to keep track of the data provenance. We examined the pan- and core-domainome and studied predicted functional capability in relation to host specificity and phylogenetic distance. We show that the pan- and core-domainome of mycoplasma species is closed. A comparison with the proteome of the "minimal" synthetic bacterium JCVI-Syn3.0 allowed us to classify domains and proteins essential for minimal life. Many of those essential protein domains, essential Domains of Unknown Function (DUFs) and essential hypothetical proteins are not persistent across mycoplasma genomes suggesting that mycoplasma species support alternative domain configurations that bypass their essentiality. Based on the protein domain composition, we could separate mycoplasma species infecting blood and tissue. For selected genomes of tissue infecting mycoplasmas, we could also predict whether the host is ruminant, pig or human. Functionally closely related mycoplasma species, which have a highly similar protein domain repertoire, but different hosts could not be separated. This study provides a concise overview of the functional capabilities of mycoplasma species, which can be used as a basis to further understand host-pathogen interaction or to design synthetic minimal life.

Keywords: mycoplasma, mollicutes, protein domains, genome comparison, host specificity, niche specificity, minimal genome, protein metabolism

OPEN ACCESS

Edited by:
Lorenza Putignani,
Bambino Gesù Children's Hospital
(IRCCS), Italy

Reviewed by:
Subramanian Dhandayuthapani,
Texas Tech University Health Sciences
Center, USA
Yang Zhang,
University of Pennsylvania, USA

***Correspondence:**
Peter J. Schaap
peter.schaap@wur.nl

Received: *08 November 2016*
Accepted: *23 January 2017*
Published: *07 February 2017*

Citation:
Kamminga T, Koehorst JJ, Vermeij P, Slagman SJ, Martins dos Santos VA, Bijlsma JJ and Schaap PJ (2017) Persistence of Functional Protein Domains in Mycoplasma Species and their Role in Host Specificity and Synthetic Minimal Life. Front. Cell. Infect. Microbiol. 7:31. doi: 10.3389/fcimb.2017.00031

INTRODUCTION

Mycoplasmas have evolved from a common gram-positive ancestor (Razin and Yogev, 1998) and the evolutionary path of genome reduction has led to an obligatory parasitic lifestyle which presumably has selected for those bacteria that best manipulate their hosts and make optimal use of their specific niche with a minimal set of genes. The mechanisms needed by these bacteria to survive in a vertebrate host, however, are not completely understood (Rosengarten et al., 2000). Research into infectious mechanisms used by mycoplasma species has been focused on identification of adhesive molecules (Rottem, 2003), lipoproteins (Browning et al., 2011), molecular mechanisms used to vary the composition of the surface of the bacterial membrane (Razin and Yogev, 1998) and production of oxidizing components (e.g., hydrogen peroxide and hydrogen sulfide) which cause damage to the host (Vilei and Frey, 2001; Großhennig et al., 2015). While these studies provide insight into the mechanisms used by mycoplasmas to infect the host they do not explain why a mycoplasma species is specific for its host. Besides being important pathogens, mycoplasma species have also been extensively studied because their gene set is expected to be close to the minimal amount of genes needed to sustain life (Gil et al., 2004). Recently, a major hallmark was achieved by publication of an engineered mycoplasma with a synthetic minimal genome of 473 genes based on the genome of *Mycoplasma mycoides* subsp. *capri* (Gibson et al., 2010; Hutchison et al., 2016) providing a benchmark for genome comparison studies aimed at determining gene essentiality.

Advancements in genome sequencing techniques led to the availability of a multitude of genomes from mycoplasma species. With this wealth of sequencing data, it is possible to study the complete repertoire of genes for a bacterial species, the pan-genome. Rouli et al. (2015) observed that bacterial species that have adopted an allopatric lifestyle in specific hosts, tend to have a closed pan-genome. In recent comparative genomics studies for mycoplasma and haemoplasma species, a sub-group within the mycoplasma genus, the pan-genome was reported to be open (Liu et al., 2012; Guimaraes et al., 2014). Here we present a genome-scale comparison of mycoplasma species at the functional level of protein domains. Proteins are the main working machinery of the cell and consist of functional domains, which are stable structurally independent and genetically mobile units. A protein function can thus be precisely described by taking into account the specific domain composition architecture (Koehorst et al., 2016b). Studying protein domain presence, instead of gene sequence similarity, allows for comparison of domain promiscuity, and expansion and domain architecture variability. In a recent study, this approach was used for comparison of 121 *Streptococcus* strains based on the protein domain composition of these strains (Saccenti et al., 2015) and the authors were able to capture metabolic flexibility within *Streptococcus* through the identification of differences in core metabolic pathways between pathogenic and non-pathogenic strains. By analyzing functional capability based on protein domains, we gain insight in functional flexibility of mycoplasma species and we hypothesized that this will allow us to capture functional differences between mycoplasma species explaining adaptation to a host or niche. This strategy is supported by the recent finding that for *Mycoplasma pneumonia* gene essentiality should be studied on the level of domains and not on the level of genes (Lluch-senar et al., 2015). All protein domains found in a species make up the pan-domainome of a species (Kuznetsov et al., 2006), containing core, accessory, and unique domains.

We performed a de novo annotation of 80 publically available mycoplasma genomes and included in this analysis the synthetic minimal genome variant of *M. mycoides* subsp. *capri* using a standardized pipeline for prokaryotic genomes focused on identification of protein domains. We determined the composition and size of the core- and pan-domainome of distinct mycoplasma species and of the complete mycoplasma genus. Incorporation of the synthetic minimal variant in the comparison allowed us to analyze the overlap between protein domains in the core domainome of mycoplasma species vs. the synthetic minimal bacterium to pinpoint functions essential for minimal life.

METHODS

Genome Retrieval and Data Handling

In total 65 complete and 15 draft mycoplasma genomes (Table S1) were obtained from the NCBI database on the 25th of August 2015 using the "rsync" interface. The dataset contained information from 34 mycoplasma species. For 20 species a single genome sequence was available while for the other 14 species multiple genomes were available (2–12 genomes per species). For 6 species only a draft genome sequence was available. Genome sizes range from 0.58 Mbp for *M. genitalium* to 1.36 Mbp for *M. penetrans*. Genome sequences were retrieved in FASTA format and were used as input for an in-house prokaryotic annotation platform (SAPP; Koehorst et al., 2016a). *Bacillus subtilis* strain 168 (NC_000964) (Weisburg et al., 1989) was used as outlier/common ancestor. Briefly, the SAPP platform consists of sets of modules required for genome annotation of prokaryotes. Different modules can be selected for analysis and results and metadata are directly stored in a graph-database using the RDF (Resource Description Framework) data model. Originally deposited genome annotations were obtained directly from the NCBI in GenBank format and converted into RDF. For three draft genomes no reference annotation was available (accession numbers: NZ_ANIV00000000, NZ_ANAB00000000, and NX_ANAA00000000).

Genome Re-Annotation Using SAPP

Gene prediction was performed using Prodigal version 2.6.2 (Hyatt et al., 2010) with codon table 4 (The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code). Proteins were analyzed using InterProScan version 5.4-47.0 (Jones et al., 2014) with the complete set of applications enabled (TIGRFAM, PIRSF, ProDom, SMART, PROSITE Profiles&Pattern, HAMAP, PfamA, PRINTS, SUPERFAMILY, Coils and Gene3D). Protein domain information and other relevant information (GO terms, EC#'s) obtained from InterProScan were directly stored in the graph-database. For querying results a SPARQL endpoint was set-up on a local server using Blazegraph Workbench v2.1.0. The annotated genomes were uploaded in RDF format using the Blazegraph Webinterface and query results were obtained in R using RCurl (Temple Lang, 2015) and SPARQL (Van Hage, 2013).

Phylogenetic Analysis of Mycoplasma Genomes

16S rRNA sequences were obtained from the ARB-SILVA database (Quast et al., 2013) (Table S2). When available, sequences from the "all species living tree" project were used. For the synthetic JCVI-Syn3.0, the 16S rRNA sequence of the parental *M. mycoides* subsp. *capri* PG3 was used. 16S rRNA sequences were aligned with Clustal Omega (version 1.2.1). MEGA (version 7.0.14) was used to create a phylogenetic tree (maximum likelihood method with 500x bootstrapping). Archaeopteryx (version 0.9901) was used to visualize the tree and root the tree using *B. subtilis* as outlier. The phylogenetic tree was read into R and analyzed using the R package "ape" (Paradis et al., 2004). Comparison of the phylogenetic tree to the protein domain tree was done using the R package "dendextend" (Galili, 2015).

Analysis of Core- and Pan-Domainome

The total domain composition of each genome was obtained using SPARQL queries. Only domains which were assigned with an e-value of $<1E^{-07}$ were taken into account. In R, a matrix was created with all genomes and their domain composition in binary format, meaning that in this analysis only domain presence was considered. Clustering of species based on the presence/absence matrix was done using the function "hclust" in R; distances were calculated using the "Manhattan" distance. The R-package "micropan" (Snipen et al., 2009) was used to analyze the pan- and core-domainome of species from which five or more genomes were available and the same approach was used to analyze the complete mycoplasma database. To analyze how the amount of genomes sequenced affects the size of the pan- and core-domainome, a 10 times random sampling was done from the presence/absence domain matrix using sample sizes ranging from 1 to 80 genome sequences. The range of model complexities considered (k-range) was 3–5. Estimated core- and pan-domainome sizes were obtained using micropan; true core- and pan-domainome sizes were directly calculated from the sample set. Further analysis of differences between species was done using principal component analysis (PCA). Loading scores obtained with PCA were used to identify domains that contribute highly to group separation. To identify domains present in haemoplasma species that contribute highly to separation of this cluster from the other mycoplasma clusters a loadings score >0.02 was used. To identify domains that contribute highly to separation of the pneumoniae cluster and the spiroplasma/hominis cluster, cut-off values for the loading score of >0.05 and <-0.05 were used, respectively. Proteins with a metabolic function were extracted from the genome-scale metabolic model of *M. pneumonia* (Wodke et al., 2013) and extended with InterProScan domain annotations.

Analysis of Orthologous Proteins

A SPARQL query was used to generate a protein FASTA file using all mycoplasma genomes (JCVI-Syn3.0 was not taken into account). An all-against-all BLASTP (Wolf and Koonin, 2012) was performed of the mycoplasma proteins using an e-value cut-off of $1E^{-05}$ and a maximum target sequence of 10^5. The BLAST file created was used to find orthologous proteins with orthAgogue (Ekseth et al., 2014) excluding protein pairs with an overlap below 50%. Clustering of orthologous proteins was done using MCL (Enright et al., 2002) setting 1.5 as main inflation. With a SPARQL query the domain composition of all orthologous proteins was obtained based on InterPro identifiers.

Clustering of Hypothetical Mycoplasma Proteins

Hypothetical proteins (domain-less proteins) were obtained from the mycoplasma genomes using a SPARQL query. Orthologous protein clusters containing these hypothetical proteins were obtained from the list of orthologous protein clusters. Persistence of these orthologous clusters was determined in the complete set of genomes used, JCVI-Syn3.0 and *M. mycoides* subsp. *capri LC*. Haemoplasma species were excluded from this analysis.

Prediction of Host/Niche Specific Domains

K-nearest neighbor and random forest classification (Breiman, 2001) were used to classify mycoplasma species based on host or niche specificity and to identify domains important for classification. A binary domain presence/absence matrix was used as input. The R-package "class" was used to perform k-nearest neighbor (k-nn) classification (Venables and Ripley, 2002) and the R package "randomForest" (Liaw and Wiener, 2002) was used for random forest classification. 500 trees were built for each classification with random forest. Domains important for classification were found based on the mean decrease in node impurity (Gini index). Information from 26 mycoplasma genomes was used for the final niche classification and from 22 genomes for the final host classification (Tables S10, S11). K-nn classification for the niche dataset was done with a k-value of 5, 19 mycoplasma species in the training set and 7 mycoplasma species in the test set (4 infecting multiple tissue types and 3 infecting strictly respiratory tissue). For the host classification a k-value of 4 was used, 16 mycoplasma species in the training set and 6 mycoplasma species were used in the test set (3 species infecting ruminants, 1 species infecting pig and 2 species infecting humans).

RESULTS

Re-Annotation of Mycoplasma Genomes

The quality of the structural and functional annotation of publically available genomes can vary. In order to get for all studied genomes an up-to-date set of annotated genes and to minimize the risk of false discoveries resulting from methodological inconsistencies, the 80 publicly available mycoplasma genomes (Table S1) were re-annotated using a standardized set of algorithms (Koehorst et al., 2016a). Mycoplasma genomes have a low GC content and thus accuracies of the various original gene prediction methods applied were expected to be high (Tripp et al., 2015). Nevertheless, on average 3.7% more genes were found after re-annotation with the most recent version of prodigal (Hyatt et al., 2010). Consequently, the total number of proteins found in the re-annotated genomes was also higher (Table S3).

Mycoplasma Proteome and Predicted Pan- and Core-Domainome

Haemoplasma species, which specifically infect blood, have a higher number of predicted proteins relative to their genome size (**Figure 1A**), corresponding with a lower average CDS length (Guimaraes et al., 2011). This difference is caused by the presence of a large repertoire of proteins with a relatively short CDS length, which are part of paralogous gene families (do Nascimento et al., 2012). To survive in their specific niche, haemoplasma species can express these proteins and generate variability of proteins at the cell surface to prevent detection by the immune system of the host (Citti and Blanchard, 2013). Besides the haemoplasma species, a high amount of predicted proteins relative to the genome size was also found in *M. genitalium* G37 and JCVI-Syn3.0. Approximately 80% of the mycoplasma species proteins contained functional domains (**Figure 1B**). This percentage is similar to the average match percentage found if the whole UniProtKB is analyzed using InterProScan (Hunter et al., 2012). The *M. mycoides* based JCVI-syn3.0 synthetic genome contained the highest percentage of proteins with a recognizable domain (86%), ∼9% more than the parental template genome. Haemoplasma species were the notable exceptions, which despite their normal genome size, contained a significantly lower percentage of proteins with recognizable domains (22–54%). This difference occurs because the aforementioned variable surface proteins do not contain recognizable domains. As a direct result of the re-annotation strategy the total amount of unique functional domains per species increased with 0.8% on average.

The total pan-domainome consisted of 1737 domains, the core domainome consisted of 335 domains and the core-to-pan ratio was 19.3%. Analysis of the pan-domainome for species from which 5 or more genomes (*M. pneumoniae, M. gallisepticum, M. hyopneumoniae,* and *M. genitalium*) were available using Micropan (2- or 3-component system, Snipen et al., 2009) showed that the pan-domainome was closed (alpha > 1). A closed pan-domainome was also observed for the genus (9 component system, alpha > 1) taking into account all 80 mycoplasma genomes (**Figure 2**).

Functional Classification of Mycoplasma Species

To gain insight into a possible functional differentiation of mycoplasma species as a result of specific host co-evolution, we clustered mycoplasma species based on a presence/absence domain matrix and compared domain repertoire clustering with clustering based on 16S rRNA sequences (**Figure 3**). In the domain based functional tree, the monophyletic pneumonia cluster separated into three separate functional clusters. One of these separate clusters contains the haemoplasma species, which have a relatively low number of protein domains (Figure S1 and Table S4). *M. penetrans* and *M. iowae* form a second functional cluster; these species have a relatively high number of functional domains when compared to other species in the pneumonia 16S-phylogenetic group. The remaining species in this 16S-phylogenetic group are closely related to the spiroplasma cluster in the functional tree. The hominis 16S-phylogenetic cluster was completely maintained in the protein domain tree but compared to the 16S tree there were some re-arrangements, which can partly be explained by low significance in the assignment of branches in the 16S phylogenetic tree. Notable changes are: *M. hominis* and *M. arthritidis* clustered with *M. columbinum* and *M. pulmonis* clustered with *M. hyorhinis*. We did not observe a functional clustering based on host.

Functional Differentiation of Haemoplasma Species

To determine which domains were important for separation of haemoplasma from mycoplasma species infecting tissue, we used principal component analysis (**Figure 4**). Based on the loading scores for the first and second principal component we could assess which domains contributed to group separation. Haemoplasma species were separated from the other mycoplasma species along the first principal component. We identified 30 domains in haemoplasma species that mainly contributed to separation of this cluster (**Table 1** and Table S5) and 400 domains present in the tissue infecting mycoplasma species that mainly contributed to separation of this cluster from the haemoplasma species cluster. Domains present in the haemoplasma species that contributed to group separation were ABC transporter domains for iron or vitamin B12. Multiple domains were found related to functional enzymes in purine metabolism (GMP synthase, IMP dehydrogenase, adenylosuccinate synthase) or L-aspartate metabolism (fumarate lyase family domains, part of adenylosuccinate lyase) which provides a precursor for purine metabolism (Santos et al., 2011). The presence of GMP synthase domains may provide the haemoplasma with the option to produce all purine bases from hypoxanthine which is present in blood (Guimaraes et al., 2011). An alternative function for these GMP synthase domains could be the production of glutamate which is present in a low concentration in blood (McMenamy et al., 1957). Three domains related to superoxide dismutase activity were also found, a function, which could provide protection when radicals are present in blood.

Functional Differentiation between the Hominis/Spiroplasma and Pneumoniae Groups

Along the second principal component the hominis and spiroplasma clusters were separated from the pneumoniae cluster. We found 43 domains present in the hominis/spiroplasma clusters that mainly contributed to separation from the pneumoniae cluster vs. 71 in the pneumoniae cluster that mainly contributed to separation from the hominis/spiroplasma cluster (**Table 1** and Table S5). In the hominis/spiroplasma cluster there was an increased presence of domains related to transport of magnesium and other divalent cations and also an increased capacity for chromate transport.

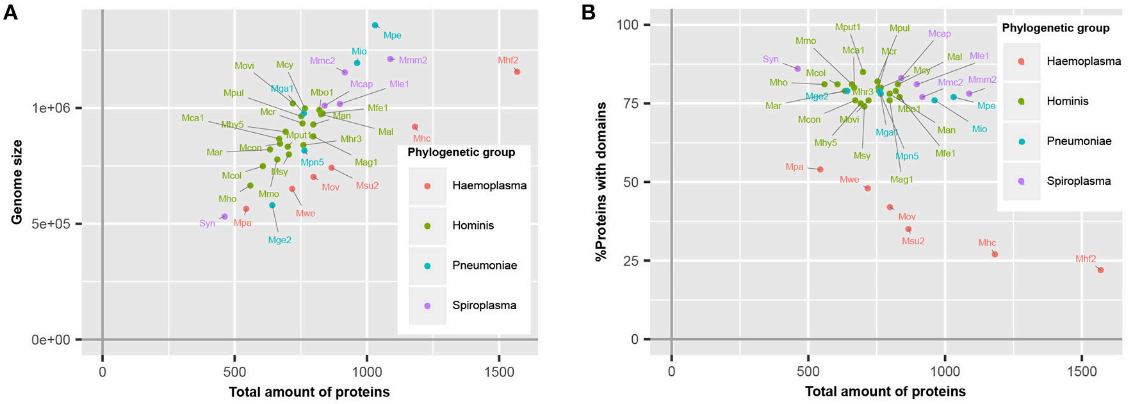

FIGURE 1 | Mycoplasma proteome specification. (A) Correlation between genome size and total amount of proteins. **(B)** Ratio of proteins covered by protein domains. Mag, *M. agalactiae*; Mal, *M. alligatoris*; Man, *M. anatis*; Mca, *M. canis*; Mcol, *M. columbinum*; Mge, *M. genitalium*; Mio, *M. iowae*; Mmc, *M. mycoides capri*; Movi, *M. ovipneumoniae*; Mpn, *M. pneumoniae*; Mar, *M. arthritidis*; Mbo, *M. bovis*; Mca, *M. capricolum* subsp. *capricolum*; Mcon, *M. conjunctivae*; Mcr, *M. crocodyli*; Mcy, *M. cynos*; Mfe, *M. fermentans*; Mga, *M. gallisepticum*; Mhc, *M. haemocanis*; Mhf, *M. haemofelis*; Mho, *M. hominis*; Mhy, *M. hyopneumoniae*; Mhr, *M. hyorhinis*; Mle, *M. leachii*; Mmo, *M. mobile*; Mmm, *M. mycoides* subsp. *mycoides*; Mov, *M. ovis*; Mpa, *M. parvum*; Mpe, *M. penetrans*; Mpul, *M. pulmonis*; Mput, *M. putrefaciens*; Msu, *M. suis*; Msy, *M. synoviae*; Mwe, *M. wenyonii*; Syn, JCVI-Syn3.0. Numbers relate to strains (Table S4).

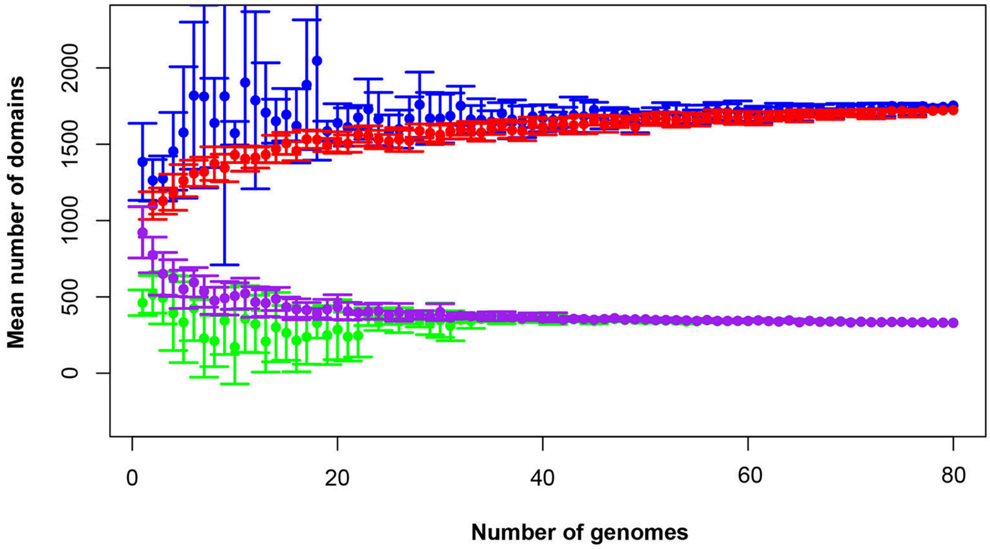

FIGURE 2 | The mycoplasma pan-domainome is closed. True and estimated core- and pan-domainome sizes were calculated using an iterative process ($n = 10$) in which a fixed number of genomes was randomly selected from the binary domain matrix. Estimated values were calculated using the R package MicroPan and true values were directly calculated from the input. Estimated pan-domainome (blue); true pan-domainome (red); estimated core-domainome (green) and true core-domainome (purple).

Metals are important co-factors and increased chromate transport capability possibly results in increased chromate resistance as observed in *B. subtilis* (Díaz-Magaña et al., 2009). Functionalities of other domains important to separate the hominis/spiroplasma cluster from the pneumoniae cluster were related to DNA/RNA modification, protein/peptide degradation and phosphopentomutase activity. The latter enzyme links nucleotide synthesis to the pentose phosphate pathway (PPP) (Pollack et al., 1997) and provides mycoplasma with the option to produce nucleotides from the purine/pyrimidine bases or alternatively to degrade nucleotides via the PPP and glycolysis. In the set of domains that mainly contributed to separation of the pneumonia cluster from the hominis/spiroplasma cluster, a functional domain related to NAD kinase activity, needed for the production of $NADP^+$, was found. Another domain was found linked to activity in the non-mevalonate pathway of isoprenoid synthesis: 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase. Activity of this pathway was shown for *M. penetrans* and *M. gallisepticum* (Eberl et al., 2004) and might reduce the need to obtain isoprenoid precursors from the host. There was an

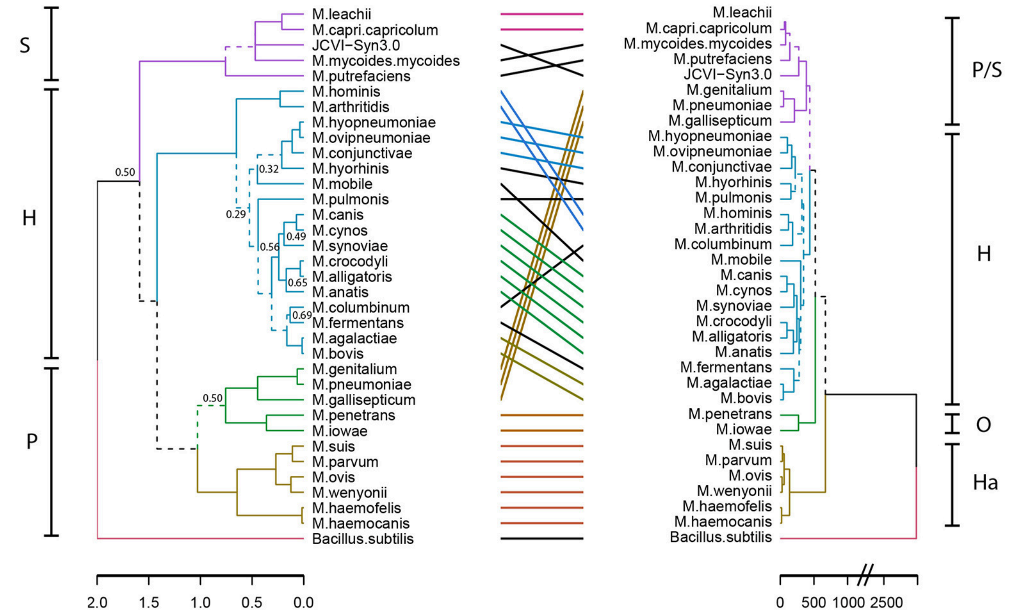

FIGURE 3 | Niche-driven functional evolution. Accelerated functional evolution causes separation of haemoplasma species and several other mycoplasma species when phylogenetic clusters are compared to functional clusters. Dashed lines indicate distinct branches. **Left:** Standard phylogenetic tree using 16S rRNA (maximum likelihood, 500x bootstrapped, see Table S2 for strains and sequences which were used). Only bootstrapping values <0.7 are shown, the phylogenetic tree with all bootstrapping values is shown in Figure S2. **Right:** functional clustering based on Manhattan distance calculated from the presence/absence matrix of domains. Groups indicated are: S, Spiroplasma; H, Hominis; P, Pneumoniae; Ha, Haemoplasma; and O, Other.

increased presence of a domain related to thioredoxin-disulfide reductase activity which produces reduced thioredoxin needed for the production of deoxyribonucleotides and is important for protection against oxidative stress (Ben-Menachem et al., 1997). The separation of mycoplasma species based on protein domain composition provided a concise overview of the functional differences between mycoplasma species.

Persistence of Protein Domains and of Orthologous Proteins

In order to compare the persistence of protein domains with the persistence of orthologous proteins, the complete set of orthologous proteins in the 80 mycoplasma genomes was determined using a standard bidirectional best hit approach (Wolf and Koonin, 2012) followed by orthology assessment with orthAgogue (Ekseth et al., 2014) and MCL clustering (Enright et al., 2002). We found >5000 clusters of orthologous proteins and examined in how many genomes these orthologous proteins are present (**Figure 5A**). Only 135 orthologous proteins are conserved amongst all mycoplasma species and we find an average persistence of orthologous proteins of 12.6%. The persistence of protein domains in the pan-domainome was much higher (average of 48.4%, **Figure 5A**).

Mycoplasma Pan- and Core-Domainome Analysis in Relation to JCVI-syn3.0

Clustering of mycoplasma species based on the pan-domainome did not show a correlation with their specific host. To further classify protein domains, we compared the pan- and core-domainome of the mycoplasma genus with the domainome of the minimal synthetic organism JCVI-Syn3.0 (Hutchison et al., 2016) which consisted of 869 domains (Table S7). For the synthetic organism we assumed that all protein domains in this organism were essential. The core domainome consisted of 335 domains, a relatively small amount. This can be explained because we took into account the haemoplasma species that grow in blood and cannot be cultured *ex vivo*. When the core-domainome was calculated for mycoplasma species infecting tissue, a larger size core was obtained of 479 protein domains. From this core 26 domains were not present in JCVI-Syn3.0 (Table S8). Apparently these domains are not essential for growth in a laboratory environment. The remaining 453 core domains show overlap with JCVI-Syn3.0 (Table S8), indicating that these persistent domains are essential for axenic growth in (complex) growth media. Interestingly, the remaining 416 domains in JCVI-Syn3.0 essential for minimal life are not persistent (**Figure 5B** and Figure S3) suggesting that within the mycoplasma domain landscape many alternative configurations exist that bypass their essentiality.

Metabolic Capability in Relation to Host Specificity

To assess if domains with a non-essential metabolic function determine host specificity, we obtained all domains with a metabolic function not present in the synthetic minimal organism. Domains with a metabolic function were derived

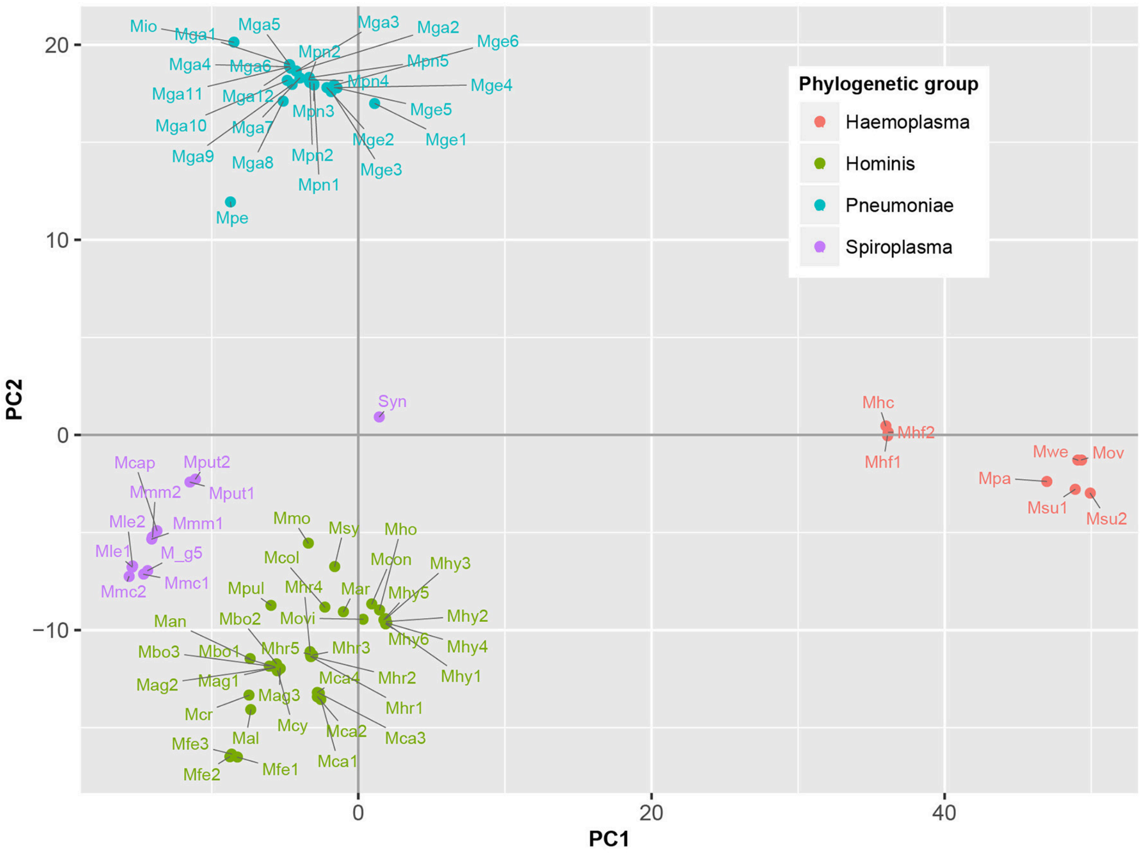

FIGURE 4 | Functional differentiation of mycoplasma species. Score plot is shown of principal component analysis done on the presence/absence matrix of the 80 mycoplasma strains and the synthetic bacterium JCVI-Syn3.0. Main phylogenetic groups are color coded. Note the separation of mycoplasma species infecting blood and tissue. Mag, *M. agalactiae*; Mal, *M. alligatoris*; Man, *M. anatis*; Mca, *M. canis*; Mcol, *M. columbinum*; M_g5, *M. g5847*, Mge, *M. genitalium*; Mio, *M. iowae*; Mmc, *M. mycoides capri*; Movi, *M. ovipneumoniae*; Mpn, *M. pneumoniae*; Mar, *M. arthritidis*; Mbo, *M. bovis*; Mca, *M. capricolum* subsp. *capricolum*; Mcon, *M. conjunctivae*; Mcr, *M. crocodyli*; Mcy, *M. cynos*; Mfe, *M. fermentans*; Mga, *M. gallisepticum*; Mhc, *M. haemocanis*; Mhf, *M. haemofelis*; Mho, *M. hominis*; Mhy, *M. hyopneumoniae*; Mhr, *M. hyorhinis*; Mle, *M. leachii*; Mmo, *M. mobile*; Mmm, *M. mycoides* subsp. *mycoides*; Mov, *M. ovis*; Mpa, *M. parvum*; Mpe, *M. penetrans*; Mpul, *M. pulmonis*; Mput, *M. putrefaciens*; Msu, *M. suis*; Msy, *M. synoviae*; Mwe, *M. wenyonii*; Syn, JCVI-Syn3.0. Numbers relate to strains (Table S4).

based on the genome-scale metabolic model of *M. pneumoniae* (Wodke et al., 2013) supplemented with InterPro annotations. This model contains 145 genes, coding for 145 proteins, and from this set of proteins 359 unique protein domains were obtained. Almost all proteins with a metabolic function were covered with domains (97%). Overall we found 162 domains (33.8% of the total core) with a metabolic function to be present in the core of the tissue infecting mycoplasma species and 197 accessory domains with a metabolic function present in the accessory domainome (pan minus core). In JCVI-Syn3 156 domains from the metabolic core domainome and 140 accessory domains with a metabolic function were present. Thus, 63 domains with a metabolic function were absent in JCVI-Syn3.0 and to assess whether these domains could be involved in host specificity we clustered mycoplasma species infecting tissue based on the presence/absence of these domains but we could not establish a correlation with host specificity (Figure S4).

Role of Hypothetical Proteins in Host Adaptation

Clustering based on the pan-domainome composition or on the metabolic domain complement absent in JCVI-Syn3.0 failed to show a direct link between specific domains and host specificity. We further analyzed if presence or absence of hypothetical proteins could explain host specificity. In our dataset, a protein was annotated as hypothetical when a protein did not contain a protein domain or when a protein contained a domain of unknown function (DUF). In total 58 DUFs were found in the mycoplasma genus, from which only 8 DUFs were present in JCVI-Syn3 (Table S9). There were no DUFs in the core domainome of the complete genus and only 2 DUFs in the core domainome of the tissue infecting mycoplasma species (DUF161 and DUF933). DUF161 is part of a membrane protein with unknown function; DUF933 is suggested to be part of a nucleoprotein complex and could function as a GTP-dependent translation factor. The total amount of DUFs found was too low to analyze a relation with the host and for further classification

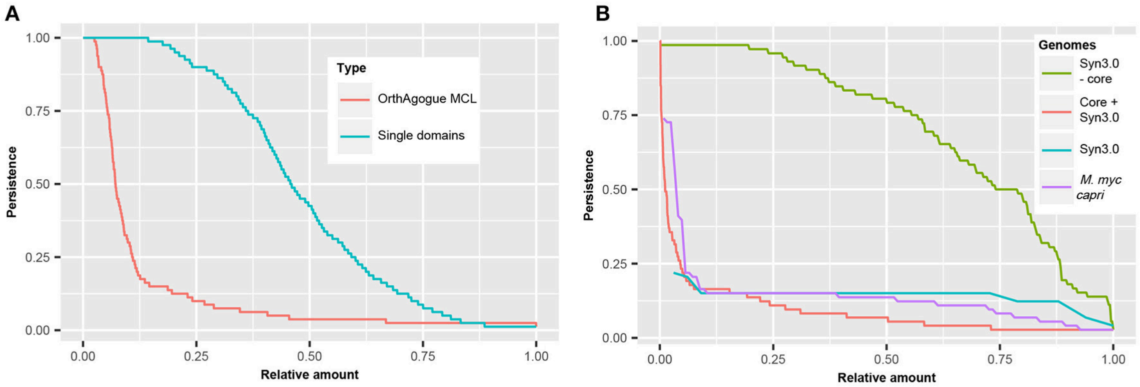

FIGURE 5 | Persistence of (essential) domains and (essential) hypothetical proteins. (A) Persistence of orthologous proteins based on sequence similarity (red) and persistence of single domains (blue). **(B)** Persistence of domains present in JCVI but absent from the core of the mycoplasma species infecting tissue (green). Persistence of hypothetical proteins for 72 mycoplasma species including JCVI-Syn3.0 (red), *M. mycoides capri* LC (purple), and JCVI-Syn3.0 (blue).

of hypothetical proteins we compared the complete set of hypothetical proteins in JCVI-Syn3.0 to the complete set of hypothetical proteins in the pan-genome of the mycoplasma species infecting tissue. In total 11,598 hypothetical proteins were found in the tissue infecting mycoplasma species which based on sequence similarity, could be clustered into 1766 orthologous protein clusters. The relative persistence of the hypothetical protein clusters showed a sharp decline with an average persistence of approximately 9% (**Figure 5B**). The total amount of genes with completely unknown functions in the genome of JCVI-Syn3.0 was only 65 (Hutchison et al., 2016) and we identified just 40 proteins to which no functional domains could be assigned. The persistence of orthologous protein clusters containing these hypothetical proteins was 14% (**Figure 5B**) which was higher than average. There was, however, conservation of clusters with hypothetical proteins from the spiroplasma phylogenetic group. In line with the finding that not all essential JCVI-Syn3.0 protein domains were persistent, essential hypothetical proteins were also not persistent suggesting that within the mycoplasma genus alternative solutions exist substituting these essential but currently unknown functions. We did not observe a relation with the host on the basis of the clustering of orthologous hypothetical proteins not present in JCVI-Syn3.0 (Figure S5).

Protein Domain Composition in Relation to Host or Niche

Clustering based on the complete pan-domainome of mycoplasma, the metabolic domains outside JCVI-syn3.0 as well as the hypothetical orthologous proteins outside JCVI-Syn3.0 did not show a relation with a mycoplasma species specific host. As a final effort, we applied two machine learning approaches: k-nearest neighbor (k-nn) and Random Forest (Chen and Ishwaran, 2012), to classify a mycoplasma species niche or host based on the pan-domainome composition. Both methods could predict with high accuracy whether the niche of a mycoplasma species was blood or tissue confirming the results already found using PCA (supplementary materials and Table S5). When the niche was specified in more detail (Table S6, Niche), the prediction accuracy decreased and species with a unique niche (e.g., *M. mobile* and *M. conjunctivae*) could not be assigned. Classification of mycoplasma growing in blood, strictly in the respiratory tract and in multiple tissue types including lung (Table S6, Niche 2) was possible using Random Forest with 95% prediction accuracy (5% out-of-bag error rate). The domain most important for classification was cell division protein *FtsZ* (IPR000158). This domain was present in many mycoplasma species but absent from *M. canis*, *M. gallisepticum,* and *M. hyopneumoniae*, which formed for a large part the species infecting the respiratory tract in our dataset. Absence of this specific domain does not mean that a species has no functional *FtsZ*, since there are alternative domain configurations possible (containing e.g., domain IPR003008 and IPR020805). To prevent prediction bias due to differences in the number of genomes available of a certain species, we decided to focus on the mycoplasma species infecting tissue for which we had at least two genomes and limited our search to two genomes per species. Using this smaller selection of genomes, prediction accuracy was higher (96% using the random forest classifier and 71% using k-nn classification) and we again identified the specific *FtsZ* domain (**Table 2** and Table S10) as an important domain for niche classification. We also identified a putative DNA-binding domain (IPR009061), present in phenylalanine-tRNA synthetases. In our database this domain was not present in the selected strains of *M. canis*, *M. hyopneumoniae,* and *M. pneumoniae* which are all present strictly in the respiratory tract. The domain was, however, present in other mycoplasma species identified as strictly present in the respiratory tract: *M. cynos*, *M. gallisepticum,* and *M. mycoides* subsp. *mycoides* SC. Also important for classification was restriction endonuclease, type I domain IPR000055, which was not present in *M. gallisepticum* strains used in our selection and was also absent from the *M. mycoides* subsp. *mycoides* SC strains used in our comparison. There was not a single domain uniquely present in all mycoplasma infecting the respiratory tract and absent from the mycoplasma infecting multiple tissue types.

TABLE 1 | Top 10 domains responsible for separation of mycoplasma functional clusters.

	Enriched in haemoplasma[a]		Enriched in hominis/spiroplasma[b]		Enriched in pneumoniae[c]
ID[d]	**InterPro description[e]**	**ID[d]**	**InterPro description[e]**	**ID[d]**	**InterPro description[e]**
IPR026023	Ribonucleotide reductase small subunit, prokaryotic	IPR029048	Heat shock protein 70kD, C-terminal domain	IPR002606	Riboflavin kinase, bacterial
IPR029022	ABC transporter, BtuC-like	IPR013826	DNA topoisomerase, type IA, central region, subdomain 3	IPR003526	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
IPR000522	ABC transporter, permease protein	IPR004398	RNA methyltransferase, RsmD	IPR006660	Arsenate reductase-like
IPR001674	GMP synthase, C-terminal	IPR003442	tRNA threonylcarbamoyl adenosine modification protein TsaE	IPR011631	Protein of unknown function DUF1600
IPR004837	Sodium/calcium exchanger membrane region	IPR006667	SLC41 divalent cation transporters, integral membrane domain	IPR023344	Uncharacterized domain MG237, C-terminal
IPR001670	Alcohol dehydrogenase, iron-type	IPR006668	Magnesium transporter, MgtE intracellular domain	IPR015271	Protein of unknown function DUF1951
IPR001093	IMP dehydrogenase/GMP reductase	IPR016947	Bacteriophage gamma, gammalsu0035	IPR013825	DNA topoisomerase, type IA, central region, subdomain 2
IPR019065	Restriction endonuclease, type II, NgoFVII	IPR000748	Pseudouridine synthase, RsuA/RluB/E/F	IPR012760	RNA polymerase sigma factor RpoD, C-terminal
IPR020471	Aldo/keto reductase subgroup	IPR001525	C-5 cytosine methyltransferase	IPR001844	Chaperonin Cpn60
IPR023210	NADP-dependent oxidoreductase domain	IPR003370	Chromate transporter	IPR002423	Chaperonin Cpn60/TCP-1

[a] *Domains enriched in haemoplasma functional cluster.*
[b] *Domains enriched in hominis/spiroplasma functional cluster.*
[c] *Domains enriched in pneumoniae functional cluster.*
[d] *InterPro Identifier.*
[e] *Domain description obtained from InterProScan.*

TABLE 2 | Top 10 domains relevant for niche classification: Strictly respiratory or multiple tissue types.

	Domain information	Abundance (%)[a]	
ID[d]	**InterPro description[e]**	**Respiratory system[b]**	**Multiple[c]**
IPR009061	DNA binding domain, putative	40	100
IPR000055	Restriction endonuclease, type I, HsdS	50	94
IPR022749	N6 adenine-specific DNA methyltransferase, N12 class, N-terminal	20	81
IPR000158	Cell division protein FtsZ	40	100
IPR011701	Major facilitator superfamily	50	88
IPR003798	DNA recombination RmuC	20	75
IPR008280	Tubulin/FtsZ, C-terminal	40	88
IPR002198	Short-chain dehydrogenase/reductase SDR	20	75
IPR011089	Domain of unknown function DUF1524	0	50
IPR005864	ATPase, F0 complex, subunit B, bacterial	60	100

[a] *Abundance of a protein domain in the specific niche.*
[b] *Abundance in mycoplasma species with a strictly respiratory niche.*
[c] *Abundance in mycoplasma species with multiple niches including respiratory.*
[d] *InterPro Identifier.*
[e] *Domain description obtained from InterProScan.*

For identification of domains important to classify mycoplasma hosts, we first used the complete diversity in hosts mentioned in Table S6 and obtained a prediction accuracy of <80% using random forest. We decided to use a more focused approach and selected only mycoplasma species growing in tissue for which we had two species per host and two genomes per species. Genomes for cows and goats were pooled into a ruminants group. With this grouping, we could accurately predict (83% accuracy with k-nn classification and 100% with random forest) if a mycoplasma species from the selected genomes infects a pig, ruminant, or human. The most discriminatory domains identified from the random forest analysis (**Table 3** and Table S11) related to peptidase functions (IPR000668, IPR005151, and IPR029045). A phosphodiesterase domain (IPR024654 and related family IPR000979) was found to be important for host differentiation, this domain only occurs in the human pathogens taken into account. A *RmlC*-like jelly roll fold domain (IPR014710), which is related to mannose/myo-inositol metabolism, was identified in the pig and ruminant species but was absent from species that infect humans. Two domains of unknown function were found: DUF2714 and DUF285 (IPR021222 and IPR005046). The DUF285 domain has probably been exchanged between ruminant species via horizontal gene transfer (Nouvel et al., 2010). Several domains related to proteins expressed at the bacterial surface were found (IPR011889 and IPR027593). A glycine cleavage domain was found (IPR002930) which was absent from the selected mycoplasma species infecting humans. Using the Random Forest prediction, on specific species groups, we have identified a number of protein domains which could relate to host specificity.

DISCUSSION

All mycoplasma species have reduced genomes and could be considered minimal organisms. Arguably, the most studied minimal organism to date is *M. pneumoniae*, a human pathogen causing inflammation of lung tissue in humans. From this bacterium we have knowledge of the genome (Dandekar et al., 2000), transcriptome (Güell et al., 2009), proteome (Batisse et al., 2009), metabolome (Yus et al., 2009; Maier et al., 2013) and several regulatory mechanisms including the role of non-coding RNA's (Lloréns-Rico et al., 2016). Interactions with the host have also been extensively studied (Rottem, 2003) but despite of the wealth of information on this minimal organism we still cannot explain why there is a preference for colonization of human lung tissue. Knowing that it is not a simple case of adhesion properties (Krause, 1996), we hypothesized that there is a complex combination of functions that determines a bacterial host or niche. To find these functions, we clustered species based on domain presence to find direct leads and ultimately used a random forest classification algorithm on the complete mycoplasma pan-domainome to find sets of domains that predict the specific host or niche of a mycoplasma species. By considering presence or absence of proteins domains we deviate from the classical approach in which bacterial genomes are compared based on orthologous proteins. We found that the persistence of single domains is higher which indicates that conservation of the structural information in the protein domains is more important than maintaining the gene sequences in which the domains are present. A similar result was recently found in a comparative genomics study of 432 *Pseudomonas* species (Koehorst et al., 2016a), indicating that this could be a trend amongst bacterial species. By using Random Forest classification, we could predict with high accuracy whether a mycoplasma species infects tissue or blood and found metabolic properties in the haemoplasma cluster that could explain why this organism successfully infects blood. Zooming into functional species clusters, the prediction accuracy decreases and it is not possible to predict a host or niche of closely related species such as *M. haemofelis* and *M. haemocanis*, within the haemoplasma group, or *M. agalactiae* and *M. bovis*, within the hominis group. Despite the lower prediction accuracy we were still able to identify differences between mycoplasma species in relation with its specific host or niche if we used larger clusters as was shown for the differentiation of mycoplasma colonizing ruminants, pigs, or humans. To determine the specific role of a signifying protein function (e.g., one of the peptidase functions) in host-pathogen interaction would require additional laboratory studies.

To understand in greater detail which factors determine host or niche specificity, more mycoplasma genomes of species of specific interest could be sequenced. This will provide more detailed information on the variation in the domain composition of this species, increasing the accuracy of host prediction. Further information needed to understand host or niche specificity could also follow from functional annotation of proteins without a protein domain, which make up ∼20% of the total proteome of a mycoplasma species. The machine learning approaches applied did not take domain abundance into account as we used the binary domain matrix as input to avoid overfitting. (Dual-)Trancriptomics studies might provide the additional insight needed to explain the interplay between host and pathogen. For example, a recent study on the chicken pathogen *M. gallisepticum* (Pflaum et al., 2015) showed temporal phase variation in the expression of *vlhA* genes during infection. Finally, the strict host specificity for mycoplasma species can be challenged since several mycoplasma species infect a broad range of hosts (e.g., *M. bovis* and *M. mycoides* subsp. *mycoides*) and mycoplasmas normally isolated from animals are sometimes found in humans and vice versa (Huang et al., 2001; Pitcher and Nicholas, 2005). The assumption of strict host specificity for mycoplasma species could be incorrect and mycoplasma may be able to infect a wider range of hosts and ecosystems than previously anticipated (Citti and Blanchard, 2013).

Our finding that the pan-domainome of the mycoplasma genus is closed supports the general expectation that species with an allopatric lifestyle have a lower chance of gaining genes by horizontal gene transfer (HGT). This finding, however, seems to contradict the recent comparative genomics reports on an open pan-genome for mycoplasma species (Liu et al., 2012; Guimaraes et al., 2014). Possible mechanisms that could contribute to the increase of the pan-genome have been described to be: (1) variation in expression and structure of surface antigens, (2) horizontal gene transfer (HGT), (3) genetic drift, and (4) phage attack (Citti and Blanchard, 2013). HGT events between species outside the mycoplasma genus are rare (Sirand-Pugnet et al., 2007) and phage attacks are not common in mycoplasma species (Tu et al., 2001). Thus, we expect that genetic drift and sequence variations in the regions coding for variable surface proteins

TABLE 3 | Top 10 domains relevant for host classification: Ruminants, pigs or humans.

Domain information		Abundance (%)[a]		
ID[e]	InterPro description[f]	Ruminants[b]	Pigs[c]	Humans[d]
IPR000668	Peptidase C1A, papain C-terminal	100	0	0
IPR005151	Tail specific protease	100	0	0
IPR000979	Phosphodiesterase MJ0936/Vps29	0	0	100
IPR014710	RmlC-like jelly roll fold	100	100	0
IPR021222	Protein of unknown function DUF2714	100	100	0
IPR005046	Protein of unknown function DUF285	100	0	0
IPR002931	Transglutaminase-like	100	0	0
IPR002930	Glycine cleavage H-protein	100	100	0
IPR011889	Bacterial surface protein 26-residue repeat	92	0	0
IPR029045	ClpP/crotonase-like domain	100	0	0

[a] *Abundance of a protein domain in the specific host.*
[b] *Abundance in ruminant species.*
[c] *Abundance in pig species.*
[d] *Abundance in humans.*
[e] *InterPro Identifier.*
[f] *Domain description obtained from InterProScan.*

contribute to an increase in the pan-genome size but that this increase is mainly related to genes encoding proteins without characterized domains.

Because the pan-domainome of mycoplasma species is closed, sequencing additional strains will not add to the overall systems level understanding of mycoplasma physiology and focus should be on further understanding of the mycoplasma strains for which the genome sequence is known. In this study we incorporated the minimal JCVI-Syn3.0, which is based on a *M. mycoides* template. We considered a protein domain essential when it was present in the minimal synthetic bacterium meaning that the protein domain is needed for growth in a complex cultivation medium under laboratory conditions. We also consider it likely that none of the domains in the minimal synthetic bacterium are needed to maintain growth in the specific host since the genome has been minimized for growth outside the host. By comparing the core domainome of the mycoplasma genus with JCVI-Syn3.0 we found that almost all domains present in the mycoplasma core are also present in the minimal synthetic organism and are likely needed to support growth in axenic media under laboratory conditions. The synthetic bacterial genome still contains 17% of essential protein coding genes with an unknown function. We found that conserved hypothetical proteins in the spiroplasma functional group are conserved in JCVI-Syn3.0. This finding is in line with the general notion that conserved hypothetical proteins are more likely to be essential (Galperin and Koonin, 2004) but in the case of mycoplasma this conservation is limited to mainly the functional cluster, and not to the complete genus. Both findings can provide a guideline for the design of minimal bacterial synthetic genomes. We expect that when mycoplasma species from other functional groups are taken as a template, alternative configurations will emerge showing flexibility in the composition of the pan-domainome of minimal synthetic bacteria designed from mycoplasma ancestors.

AUTHOR CONTRIBUTIONS

All authors contributed to study design and interpretation. TK, JB, and PS drafted the manuscript. JK provided scripts and methods used in this research. All authors revised the manuscript and approved the final version. All authors take responsibility for accuracy and integrity of the work.

FUNDING

This work was financially supported by MSD Animal Health, Bioprocess Technology & Support, Boxmeer, Netherlands. This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 634942.

ACKNOWLEDGMENTS

We thank Pascal Sirand-Pugnet for critically reviewing the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fcimb.2017.00031/full#supplementary-material

REFERENCES

Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., et al. (2009). Proteome organization in a genome-reduced bacterium. *Science* 326, 1235–1240. doi: 10.1126/science.1176343

Ben-Menachem, G., Himmelreich, R., Herrmann, R., Aharonowitz, Y., and Rottem, S. (1997). The thioredoxin reductase system of mycoplasmas. *Microbiology* 143, 1933–1940. doi: 10.1099/00221287-143-6-1933

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Browning, G. F., Marenda, M. S., Noormohammadi, A. H., and Markham, P. F. (2011). The central role of lipoproteins in the pathogenesis of mycoplasmoses. *Vet. Microbiol.* 153, 44–50. doi: 10.1016/j.vetmic.2011.05.031

Chen, X., and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics* 99, 323–329. doi: 10.1016/j.ygeno.2012.04.003

Citti, C., and Blanchard, A. (2013). Mycoplasmas and their host: emerging and re-emerging minimal pathogens. *Trends Microbiol.* 21, 196–203. doi: 10.1016/j.tim.2013.01.003

Dandekar, T., Huynen, M., Regula, J. T., Ueberle, B., Zimmermann, C. U., Andrade, M. A., et al. (2000). Re-annotating the *Mycoplasma pneumoniae* genome sequence: adding value, function and reading frames. *Nucleic Acids Res.* 28, 3278–3288. doi: 10.1093/nar/28.17.3278

Díaz-Magaña, A., Aguilar-Barajas, E., Moreno-Sánchez, R., Ramírez-Díaz, M. I., Riveros-Rosas, H., Vargas, E., et al. (2009). Short-chain chromate ion transporter proteins from *Bacillus subtilis* confer chromate resistance in *Escherichia coli. J. Bacteriol.* 191, 5441–5445. doi: 10.1128/JB.00625-09

do Nascimento, N. C., Santos, A. P., Guimaraes, A. M. S., Sanmiguel, P. J., and Messick, J. B. (2012). Mycoplasma haemocanis - the canine hemoplasma and its feline counterpart in the genomic era. *Vet. Res.* 43:66. doi: 10.1186/1297-9716-43-66

Eberl, M., Hintz, M., Jamba, Z., Beck, E., and Jomaa, H. (2004). *Mycoplasma penetrans* is capable of activating Vγ9/ Vδ2 T cells while other human pathogenic mycoplasmas fail to do so. *Infect. Immun.* 72, 4881–4883. doi: 10.1128/iai.72.8.4881-4883.2004

Ekseth, O. K., Kuiper, M., and Mironov, V. (2014). OrthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics* 30, 734–736. doi: 10.1093/bioinformatics/btt582

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Galili, T. (2015). Dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31, 3718–3720. doi: 10.1093/bioinformatics/btv428

Galperin, M. Y., and Koonin, E. V. (2004). "Conserved hypothetical" proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* 32, 5452–5463. doi: 10.1093/nar/gkh885

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. A., et al. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329, 52–56. doi: 10.1126/science.1190719

Gil, R., Silva, F. J., Peretó, J., and Pereto, J. (2004). Determination of the core of a minimal bacterial gene set determination of the core of a minimal bacterial gene set[†]. *Microbiol. Mol. Biol. Rev.* 68, 518–537. doi: 10.1128/MMBR.68.3.518-537.2004

Großhennig, S., Ischebeck, T., Gibhardt, J., Busse, J., Feussner, I., and Stülke, J. (2015). Hydrogen sulfide is a novel potential virulence factor of *Mycoplasma pneumoniae*: characterization of the unusual cysteine desulfurase/desulfhydrase hape. *Mol. Microbiol.* 100, 42–54. doi: 10.1111/mmi.13300

Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., et al. (2009). Transcriptome complexity in a genome-reduced bacterium. *Science* 326, 1268–1271. doi: 10.1126/science.1176951

Guimaraes, A. M. S., Santos, A. P., Do Nascimento, N. C., Timenetsky, J., and Messick, J. B. (2014). Comparative genomics and phylogenomics of hemotrophic mycoplasmas. *PLoS ONE* 9:e91445. doi: 10.1371/journal.pone.0091445

Guimaraes, A. M. S., Santos, A. P., SanMiguel, P., Walter, T., Timenetsky, J., and Messick, J. B. (2011). Complete genome sequence of Mycoplasma suis and insights into its biology and adaption to an erythrocyte niche. *PLoS ONE* 6:e19574. doi: 10.1371/journal.pone.0019574

Huang, S., Li, J. Y., Wu, J., Meng, L., and Shou, C. C. (2001). Mycoplasma infections and different human carcinomas. *World J. Gastroenterol.* 7, 266–269. doi: 10.3748/wjg.v7.i2.266

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 40, 1–7. doi: 10.1093/nar/gks456

Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., et al. (2016). Design and synthesis of a minimal bacterial genome. *Science* 351:aad6253. doi: 10.1126/science.aad6253

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Koehorst, J. J., Saccenti, E., Schaap, P. J., Martins dos Santos, V. A. P., and Suarez-Diez, M. (2016b). Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics. *F1000Research* 5, 1987. doi: 10.12688/f1000research.9416.2

Koehorst, J. J., van Dam, J. C. J., van Heck, R. G. A., Saccenti, E., dos Santos, V. A. P. M., Suarez-Diez, M., et al. (2016a). Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data. *Sci. Rep.* 6:38699. doi: 10.1038/srep38699

Krause, D. C. (1996). *Mycoplasma pneumoniae* cytadherence: unravelling the tie that binds. *Mol. Microbiol.* 20, 247–253. doi: 10.1111/j.1365-2958.1996.tb02613.x

Kuznetsov, V., Pickalov, V., and Kanapin, A. (2006). Proteome complexity measures based on counting of domain-to-protein links for replicative and non-replicative domains. *Bioinform. Genome Regul. Struct. II* 329–341. doi: 10.1007/0-387-29455-4_32

Liaw, A., and Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2, 18–22. Available online at: http://cran.r-project.org/doc/Rnews/

Liu, W., Fang, L., Li, M., Li, S., Guo, S., Luo, R., et al. (2012). Comparative genomics of Mycoplasma: analysis of conserved essential genes and diversity of the pan-genome. *PLoS ONE* 7:e35698. doi: 10.1371/journal.pone.0035698

Lloréns-Rico, V., Cano, J., Kamminga, T., Gil, R., Latorre, A., Chen, W.-H., et al. (2016). Bacterial antisense RNAs are mainly the product of transcriptional noise. *Sci. Adv.* 2, 1–10. doi: 10.1126/sciadv.1501363

Lluch-Senar, M., Delgado, J., Chen, W., Lloréns-Rico, V., O'Reilly, F. J., Wodke, J. A., et al. (2015). Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol. Syst. Biol.* 11:780. doi: 10.15252/msb.20145558

Maier, T., Marcos, J., Wodke, J. A. H., Paetzold, B., Liebeke, M., Gutiérrez-Gallego, R., et al. (2013). Large-scale metabolome analysis and quantitative integration with genomics and proteomics data in *Mycoplasma pneumoniae*. *Mol. Biosyst.* 9, 1743–1755. doi: 10.1039/c3mb70113a

McMenamy, R. H., Lund, C. C., and Oncley, J. L. (1957). Unbound amino acid concentrations in human blood plasmas. *J. Clin. Invest.* 1, 1672–1679. doi: 10.1172/jci103568

Nouvel, L. X., Sirand-Pugnet, P., Marenda, M. S., Sagné, E., Barbe, V., Mangenot, S., et al. (2010). Comparative genomic and proteomic analyses of two *Mycoplasma agalactiae* strains: clues to the macro- and micro-events that are shaping mycoplasma diversity. *BMC Genomics* 11:86. doi: 10.1186/1471-2164-11-86

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Pflaum, K., Tulman, E. R., Beaudet, J., Liao, X., and Geary, S. J. (2015). Global changes in *Mycoplasma gallisepticum* phase-variable lipoprotein gene vlhA expression during *in vivo* infection of the natural chicken host. *Infect. Immun.* 84, 351–355. doi: 10.1128/IAI.01092-15

Pitcher, D. G., and Nicholas, R. A. J. (2005). Mycoplasma host specificity: fact or fiction? *Vet. J.* 170, 300–306. doi: 10.1016/j.tvjl.2004.08.011

Pollack, J. D., Williams, M. V., and McElhaney, R. N. (1997). The comparative metabolism of the mollicutes (Mycoplasmas): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. *Crit. Rev. Microbiol.* 23, 269–354. doi: 10.3109/10408419709115140

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, 590–596. doi: 10.1093/nar/gks1219

Razin, S., and Yogev, D. (1998). Molecular biology and pathogenicity of mycoplasmas. *Microbiol. Mol. Biol. Rev.* 62, 1094–1156.

Rosengarten, R., Citti, C., Glew, M., Lischewski, A., Droesse, M., Much, P., et al. (2000). Host-pathogen interactions in mycoplasma pathogenesis: virulence and survival strategies of minimalist prokaryotes. *Int. J. Med. Microbiol.* 290, 15–25. doi: 10.1016/S1438-4221(00)80099-5

Rottem, S. (2003). Interaction of mycoplasmas with host cells. *Physiol. Rev.* 83, 417–432. doi: 10.1152/physrev.00030.2002

Rouli, L., Merhej, V., Fournier, P.-E., and Raoult, D. (2015). The bacterial pangenome as a new tool for analyzing pathogenic bacteria. *New Microbes New Infect.* 7, 72–85. doi: 10.1016/j.nmni.2015.06.005

Saccenti, E., Nieuwenhuijse, D., Koehorst, J. J., Dos Santos, V. A. P. M., and Schaap, P. J. (2015). Assessing the metabolic diversity of streptococcus from a protein domain point of view. *PLoS ONE* 10:e0137908. doi: 10.1371/journal.pone.0137908

Santos, A. P., Guimaraes, A. M. S., do Nascimento, N. C., Sanmiguel, P. J., Martin, S. W., and Messick, J. B. (2011). Genome of *Mycoplasma haemofelis*, unraveling its strategies for survival and persistence. *Vet. Res.* 42, 102. doi: 10.1186/1297-9716-42-102

Sirand-Pugnet, P., Lartigue, C., Marenda, M., Jacob, D., Barré, A., Barbe, V., et al. (2007). Being pathogenic, plastic, and sexual while living with a nearly minimal bacterial genome. *PLoS Genet.* 3:e75. doi: 10.1371/journal.pgen.0030075

Snipen, L., Almøy, T., and Ussery, D. W. (2009). Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics* 10:385. doi: 10.1186/1471-2164-10-385

Temple Lang, D. (2015). *RCurl: General Network (HTTP/FTP/...) Client Interface for R*. Available online at: https://cran.r-project.org/package=RCurl

Tripp, H. J., Sutton, G., White, O., Wortman, J., Pati, A., Mikhailova, N., et al. (2015). Toward a standard in structural genome annotation for prokaryotes. *Stand. Genomic Sci.* 10, 45. doi: 10.1186/s40793-015-0034-9

Tu, A, H., Voelker, L. L., Shen, X., and Dybvig, K. (2001). Complete nucleotide sequence of the mycoplasma virus P1 genome. *Plasmid* 45, 122–126. doi: 10.1006/plas.2000.1501

Van Hage, W. R. (2013). *SPARQL: SPARQL Client*. Available online at: https://cran.r-project.org/package=SPARQL

Venables, W. N., and Ripley, B. D. (2002). "Random and mixed effects," in *Modern Applied Statistics with S*, eds J. Chambers, W. Eddy, W. Härdle, S. Sheather, and L. Tierney (New York, NY: Springer New York), 271–300. doi: 10.1007/978-0-387-21706-2_10

Vilei, E. M., and Frey, J. (2001). Genetic and biochemical characterization of glycerol uptake in *mycoplasma mycoides* subsp. mycoides SC: its impact on H(2)O(2) production and virulence. *Clin. Diagn. Lab. Immunol.* 8, 85–92. doi: 10.1128/cdli.8.1.85-92.2001

Weisburg, W. G., Tully, J. G., Rose, D. L., Petzel, J. P., Oyaizu, H., Yang, D., et al. (1989). A phylogenetic analysis of the mycoplasmas: basis for their classification. *J. Bacteriol*. 171, 6455–6467. doi: 10.1128/jb.171.12.6455-6467.1989

Wodke, J. A. H., Puchałka, J., Lluch-Senar, M., Marcos, J., and Yus, E., Godinho, M., et al. (2013). Dissecting the energy metabolism in *Mycoplasma pneumoniae* through genome-scale metabolic modeling. *Mol. Syst. Biol.* 9, 653. doi: 10.1038/msb.2013.6

Wolf, Y. I., and Koonin, E. V. (2012). A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol. Evol.* 4, 1286–1294. doi: 10.1093/gbe/evs100

Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.-H., et al. (2009). Impact of genome reduction on bacterial metabolism and its regulation. *Science* 326, 1263–1268. doi: 10.1126/science.1177263

Conflict of Interest Statement: TK, PV, SJS, and JB are employed by MSD-AH, a pharmaceutical company producing veterinary vaccines.

The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.