AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE
Molecular Biology and Physiology

Inferring the Minimal Genome of *Mesoplasma florum* by Comparative Genomics and Transposon Mutagenesis

Vincent Baby,[a] Jean-Christophe Lachance,[a] Jules Gagnon,[a] Jean-François Lucier,[a] Dominick Matteau,[a] Tom Knight,[b] Sébastien Rodrigue[a]

[a]Département de Biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada
[b]Ginkgo Bioworks, Boston, Massachusetts, USA

ABSTRACT The creation and comparison of minimal genomes will help better define the most fundamental mechanisms supporting life. *Mesoplasma florum* is a near-minimal, fast-growing, nonpathogenic bacterium potentially amenable to genome reduction efforts. In a comparative genomic study of 13 *M. florum* strains, including 11 newly sequenced genomes, we have identified the core genome and open pan-genome of this species. Our results show that all of the strains have approximately 80% of their gene content in common. Of the remaining 20%, 17% of the genes were found in multiple strains and 3% were unique to any given strain. On the basis of random transposon mutagenesis, we also estimated that ~290 out of 720 genes are essential for *M. florum* L1 in rich medium. We next evaluated different genome reduction scenarios for *M. florum* L1 by using gene conservation and essentiality data, as well as comparisons with the first working approximation of a minimal organism, *Mycoplasma mycoides* JCVI-syn3.0. Our results suggest that 409 of the 473 *M. mycoides* JCVI-syn3.0 genes have orthologs in *M. florum* L1. Conversely, 57 putatively essential *M. florum* L1 genes have no homolog in *M. mycoides* JCVI-syn3.0. This suggests differences in minimal genome compositions, even for these evolutionarily closely related bacteria.

IMPORTANCE The last years have witnessed the development of whole-genome cloning and transplantation methods and the complete synthesis of entire chromosomes. Recently, the first minimal cell, *Mycoplasma mycoides* JCVI-syn3.0, was created. Despite these milestone achievements, several questions remain to be answered. For example, is the composition of minimal genomes virtually identical in phylogenetically related species? On the basis of comparative genomics and transposon mutagenesis, we investigated this question by using an alternative model, *Mesoplasma florum*, that is also amenable to genome reduction efforts. Our results suggest that the creation of additional minimal genomes could help reveal different gene compositions and strategies that can support life, even within closely related species.

KEYWORDS *Mesoplasma florum*, comparative genomics, minimal genome, transposon mutagenesis

Synthetic genomics is an emerging field of synthetic biology combining different approaches and technologies to chemically synthesize sections of chromosomes or even entire genomes (1, 2), thus enabling the generation of engineered organisms that significantly differ from those found in nature. Given sufficient knowledge and proper execution, this could lead to the rational design of organisms built to accomplish specific tasks (3). However, the complexity of current model organisms is overwhelming and outstrips our ability to understand how cells operate on a global scale (4). Minimal genomes, in addition to providing invaluable information about the essential genes

Received 5 December 2017 **Accepted** 9 March 2018 **Published** 10 April 2018

Citation Baby V, Lachance J-C, Gagnon J, Lucier J-F, Matteau D, Knight T, Rodrigue S. 2018. Inferring the minimal genome of *Mesoplasma florum* by comparative genomics and transposon mutagenesis. mSystems 3: e00198-17. https://doi.org/10.1128/mSystems.00198-17.

Editor David F. Savage, University of California, Berkeley

Address correspondence to Sébastien Rodrigue, Sebastien.Rodrigue@USherbrooke.ca.

and fundamental principles required to sustain life, would therefore facilitate systematic investigations toward a global understanding of cell functioning. Minimal cells could also become interesting platforms for rapid and affordable prototyping of engineered cells, further helping to uncover underlying genome design rules. So far, three main approaches have been used to determine the minimal gene set in various organisms: comparative genomic analyses, gene inactivation studies, and progressive genome reduction.

Comparative genomics uses sequence-based strategies to identify conserved genes, which are hypothesized to be maintained throughout evolution and shared across different organisms because of their contribution to cell fitness (5). The exact number and nature of conserved genes have been found to vary considerably between studies, depending on the phylogenetic distribution (6) and number of genomes analyzed (5, 7). For example, Land and colleagues reported that 3,188 genes were always detected in *Escherichia coli* (8) while only 38 genes were found to be shared by 147 different species of bacteria and archaea (6). Some conserved gene sets are thus certainly too large to reveal the minimal genome and rather correspond to important functions that are not necessarily essential but likely contribute to the fitness of an organism in its natural habitat (9). Other gene sets are simply too small to support basic functions like replication, transcription, and translation. The results obtained through comparative genomics approaches are thus highly dependent on the set of organisms analyzed.

Genes that are conserved within a species are thought to be important or essential in their natural environment. However, laboratory and environmental conditions can greatly differ, resulting in different genetic requirements. Experimental assessment of essential genes can be achieved through individual gene inactivation studies. In this regard, gene deletion (9, 10), transposon mutagenesis (11, 12), and transcriptional interference (13, 14) were used to identify dispensable genes. The results obtained from such experiments depend largely on growth conditions since, for example, cells may need certain metabolic pathways unless their products are already available in the medium (15). Certain genes are essential only in the presence of another gene, for example to balance or counteract the activity of another gene product. For example, the antitoxin of a toxin/antitoxin system is only essential while the toxin is also present (9). At lower insertion densities, transposon mutagenesis is likely to overestimate the number of essential genes because of the higher probability of missing genes simply by chance (16). On the other hand, gene inactivation strategies can overestimate the number of dispensable genes since duplicated sequences or alternate metabolic pathways may be interrupted individually but not simultaneously, a phenomenon called synthetic lethality (17). Overall, these phenomena can lead to biases and uncertainties in the estimation of the number of essential genes.

Cumulative gene deletions that result in genome reduction provide a more accurate picture of possible minimal genome compositions of a given organism. However, this approach involves considerable effort along with well-developed genetic tools. So far, this strategy has only been applied to a few organisms, including *E. coli* (18–21), *Bacillus subtilis* (22), *Streptomyces avermitilis* (23), *Pseudomonas putida* (24), and *Mycoplasma mycoides* subsp. *capri* (2). The latter organism has undergone the most drastic streamlining, with the removal of ~50% of its original genome, resulting in the creation of *M. mycoides* JCVI-syn3.0 containing a single chromosome of 531 kbp. *M. mycoides* JCVI-syn3.0 was described as the first "working approximation of a minimal cell" (2) and is currently the simplest organism capable of autonomous growth in axenic culture. Interestingly, minimal genome designs initially proposed for *M. mycoides* based on single gene inactivation by transposon mutagenesis and other literature-based knowledge were not viable (2). Many optimization and debugging steps were required to obtain *M. mycoides* JCVI-syn3.0, highlighting the difficulty of identifying and understanding the roles of essential genes, even in the simplest cells.

Mesoplasma florum is a bacterium first isolated from a lemon tree flower in 1984 (25). Unlike many other members of the class *Mollicutes*, *M. florum* shows a short doubling time of $<$40 min, requires no sterol for growth, and has no known pathogenic potential.

TABLE 1 *M. florum* strains and genome sequencing overview

Strain	GenBank accession no.	Genome size (bp)	% GC	No. of gaps	No. of protein-coding genes	No. of functional RNAs	No. of accessory genes	Genome coding percentage	Collection	Source	Reference or source
L1	AE017263.1	793,224	27.0	0	685	35	136	93.9	Florida	*Citrus limon*	25
W37	CP006778.1	825,824	27.0	0	731	35	179	93.3	Gibson City, IL	*Solidago* sp.	32
BARC 787	CP022514	738,512	27.1	0	651	35	102	94.0	Beltsville, MD	Unspecified insect	Unpublished[a,d]
MQ3	CP022512	793,277	27.0	0	698	35	146	94.2	Maryland	*Monobia quadriens*	34
CnuA-2	CP022513	813,801	27.0	0	710	35	158	93.9	Maryland	*Coleoptera*: *Cantharidae*	33
MouA-2	CP022508	781,099	27.0	0	685	35	134	93.5	Beltsville, MD	Vespid wasp	Unpublished[a,e]
W23	CP022505	773,885	27.1	0	688	35	137	94.1	North Platte, NE	*Helianthus annuus*	32
BARC 786	CP022510	765,660	27.4	1[b]	669	35	119	93.5	Beltsville, MD	Beetle	Unpublished[a,d]
BARC 781	CP022511	803,948	27.1	1[b]	691	35	139	92.9	Beltsville, MD	Beetle	Unpublished[a,d]
GF	CP022509	792,347	27.0	2[b,c]	699	35	147	94.0	Florida	*Citrus paradisi*	25
W17	CP022507	787,107	27.4	3[c]	693	35	140	92.6	Pawnee National Grassland, CO	*Aster* sp.	32
W20	CP022506	830,640	27.0	4[b,c]	740	35	187	92.5	Ogallala, NE	*Aster simplex*	32
W12	CP022432	829,202	27.0	4[b,c]	734	35	181	93.3	Kremmling, CO	*Chrysothamnus* sp.	32

[a]Gail E. Gasparich and Robert E. Davis, personal communication.
[b]Gaps containing tandem repeats.
[c]Gaps linked to gene duplications.
[d]Isolated by R. Whitcomb in 1986.
[e]Isolated by T. Clark and K. Hackett in 1986.

The genomes of two *M. florum* strains, L1 and W37 (26), have been completely sequenced, revealing a single circular chromosome of ~800 kbp and positioning this species among the simplest free-living organisms. Basic genetic manipulation tools comprising antibiotic resistance genes, plasmids, and transformation methods have recently been developed for *M. florum* (27). Furthermore, the complete genome of *M. florum* L1 has also been cloned in yeast and transplanted into a recipient *Mycoplasma capricolum* subsp. *capricolum* strain (28, 29), which will enable sophisticated modifications and reengineering of the *M. florum* chromosome. This combination of low cell complexity, ease of manipulation, and the availability of genome engineering methods makes *M. florum* an interesting model for systems biology and synthetic genomics.

Here, we report a comparative genomic analysis of 13 *M. florum* strains. These data were investigated in conjunction with transposon mutagenesis to identify conserved, accessory, and essential genes in this species. We also discuss different scenarios for eventual *M. florum* genome reduction efforts according to results presented here and using comparisons with the phylogenetically related strain *M. mycoides* JCVI-syn3.0.

RESULTS

Genome sequencing of 11 *M. florum* strains. Several *Mesoplasma* species have been isolated from plants or insects since the 1980s (25, 30–34). Of these, we have obtained 13 *M. florum* strains available from culture collections (Table 1). Two strains originated in Florida, and 11 were collected on a longitudinal transect from Maryland to Colorado (Fig. 1A). This variety of environments, host organisms, and climates was expected to result in diversity of gene content in the genomes analyzed. The complete

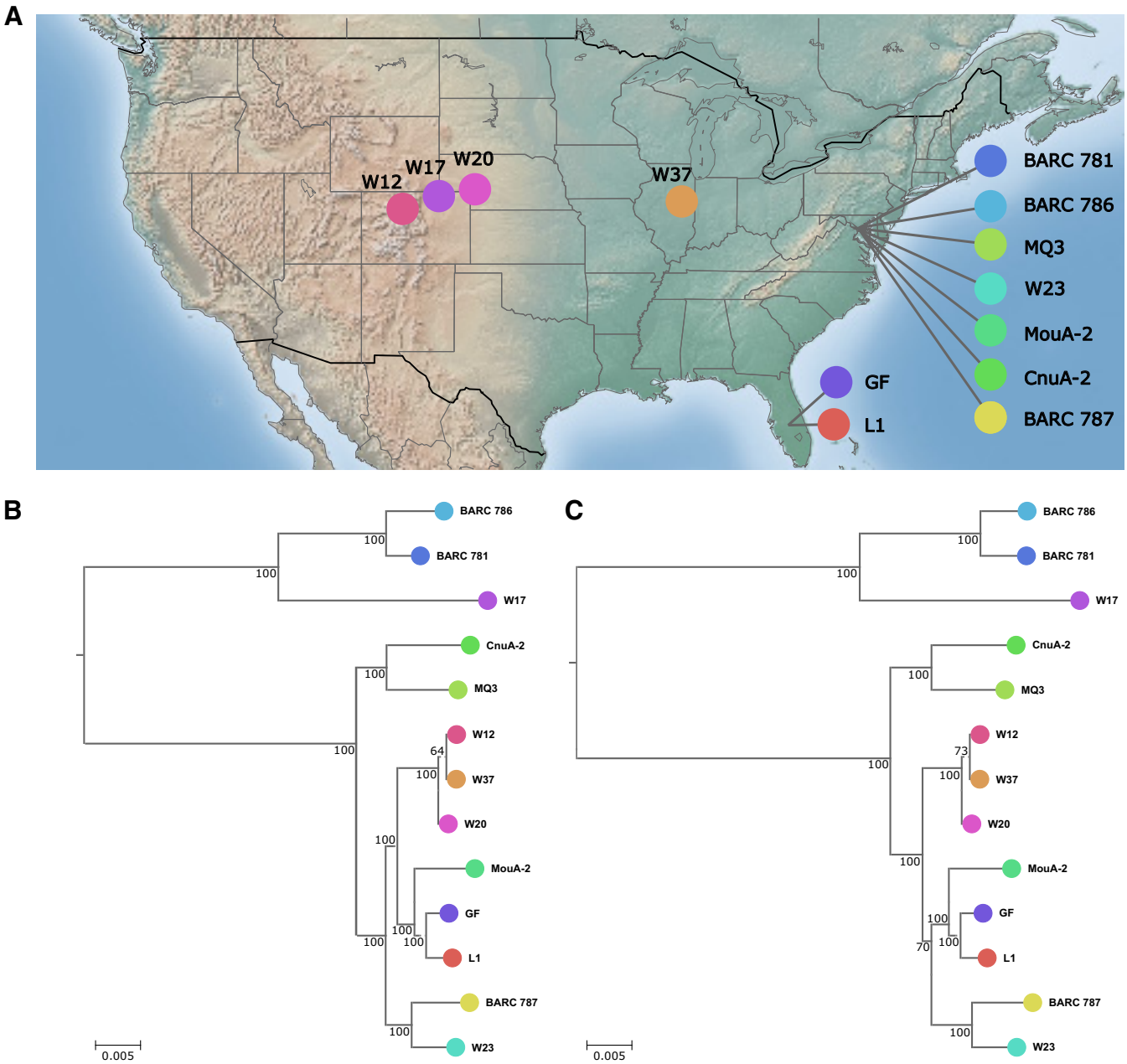

FIG 1 *M. florum* strain sampling and phylogeny. (A) Isolation sites of the 13 *M. florum* strains analyzed in this study. (B, C) *M. florum* phylogenetic trees constructed by using concatenated alignments of 412 conserved proteins and the Kimura distance model (B) or maximum likelihood (C). *M. capricolum* was used as the outgroup for both trees and is not shown because of the long branch length. Bootstrap values correspond to 100 repetitions. In both trees, branch length represents the substitution rate per site per unit of alignment length.

genome sequences of *M. florum* strains L1 and W37 have previously been reported (26). The genomes of the remaining 11 *M. florum* strains were sequenced by using a combination of the Illumina, Pacific Biosciences, and Sanger technologies. In total, seven genomes (L1, W37, BARC 787, MQ3, CnuA-2, MouA-2, and W23) were fully assembled, resulting in circular chromosomes without any ambiguous positions. The genomes of BARC 786, BARC 781, GF, W17, W20, and W12 were not unequivocally resolved during their assembly (containing one to four gaps) because of a total of five distinct repeated elements, either duplications or tandem repeats (Table 1). By keeping at least one copy of each problematic region in each final assembly, we estimate that virtually all genes are represented in these genomes and that no more than a few

kilobase pairs are missing from the chromosomes of these strains. The genome size of the 13 sequenced strains was found to vary between 738,512 bp (BARC 787) and 830,640 bp (20), with a mean of 794.5 ± 25.6 kbp. As expected from the genome sequences of other members of the class *Mollicutes* (35), the GC content of every strain analyzed was found to be relatively low, with an average of 27.1% ± 0.2% (Table 1).

Genome annotation and phylogenetic analysis. To apply a uniform procedure and minimize potential biases between annotations, gene prediction was performed for all of the *M. florum* genomes studied, including those of the previously sequenced L1 and W37 strains (see Data Sets S1 and S2 in the supplemental material, respectively). Every genome was predicted to contain 29 tRNA genes, as well as two rRNA gene loci, each encoding 5S, 16S, and 23S rRNAs. Between 651 and 740 protein-coding genes were predicted, depending on the *M. florum* strain analyzed, for a total of 9,074 putative proteins in the combined 13 genomes. The genomes showed a protein-coding gene to kilobase pair ratio of ~0.88, which is typical of bacterial genomes (36). All proteins were clustered on the basis of sequence homology (37). Genes conserved among all *M. florum* strains and *M. capricolum* as the outgroup were used to construct phylogenetic trees based on the Kimura distance model (38) (Fig. 1B) and maximum likelihood (39) (Fig. 1C). The two trees showed practically identical branch lengths and very similar topologies, diverging only slightly around the L1/GF/MouA-2 subgroup. Overall, the trees revealed two main groups that are, in turn, subdivided into smaller phylogenetic clusters. Although some closely related strains were isolated from nearby locations or similar organisms, the geographic origin and potential host were not sufficient to explain the phylogeny observed.

The predicted proteins in the *M. florum* strains were grouped into a total of 1,150 homologous gene clusters (Data Set S3). A core set of 546 clusters was observed in all representatives, resulting in the conservation of approximately 80% of the protein-coding genes in every strain (Fig. 2A and B). A majority (75.5%) of the 604 remaining gene clusters, also called the accessory genome or pangenome, was found in no more than three strains. Additionally, the number of gene cluster families discovered kept increasing as a function of the number of strains analyzed (Fig. 2A), suggesting that *M. florum* has an open pangenome and that gene diversity in this species was not fully explored by investigating these 13 genomes. Considering that an average of 23.5 gene families were found to be unique in each *M. florum* strain, several additional gene cluster families should be found simply by incorporating more genomes into the analysis.

Functional analysis. To obtain a functional overview of the *M. florum* genomes, the genes from the core genome and pangenome were classified into clusters of orthologous groups (COG) functional categories (40). Most (74.7%) of the core genes could be associated with a COG category, while most (66.7%) of the accessory genes could not. Overall, more than half (53.4%) of the COG associations belonged to the translation (J), carbohydrate metabolism (G), replication/recombination/repair (L), general function prediction only (R), and amino acid transport metabolism (E) categories. The core genome was significantly enriched (Fisher test with Bonferroni correction, $P < 0.05$) for functions related to translation (COG category J) compared to the frequency of this same category in the entire genome (Fig. 2C). The distribution of COG categories in the pangenome was more diversified, and no category was found to be statistically significantly enriched or deprived when all *M. florum* strains were considered. Genes linked to carbohydrate metabolism (mainly phosphotransferase system [PTS] components and β-glucosidases) and genome maintenance, mostly in restriction/modification systems (COG G and L categories, respectively) varied greatly in number, depending on the strains, although they were generally more abundant in the accessory genome (Fig. 2D).

Genome organization. We next analyzed the genomic organization in *M. florum* strains by characterizing the relative positions of the core genes. We found that although some rearrangements can be observed, the genomes are mostly syntenic

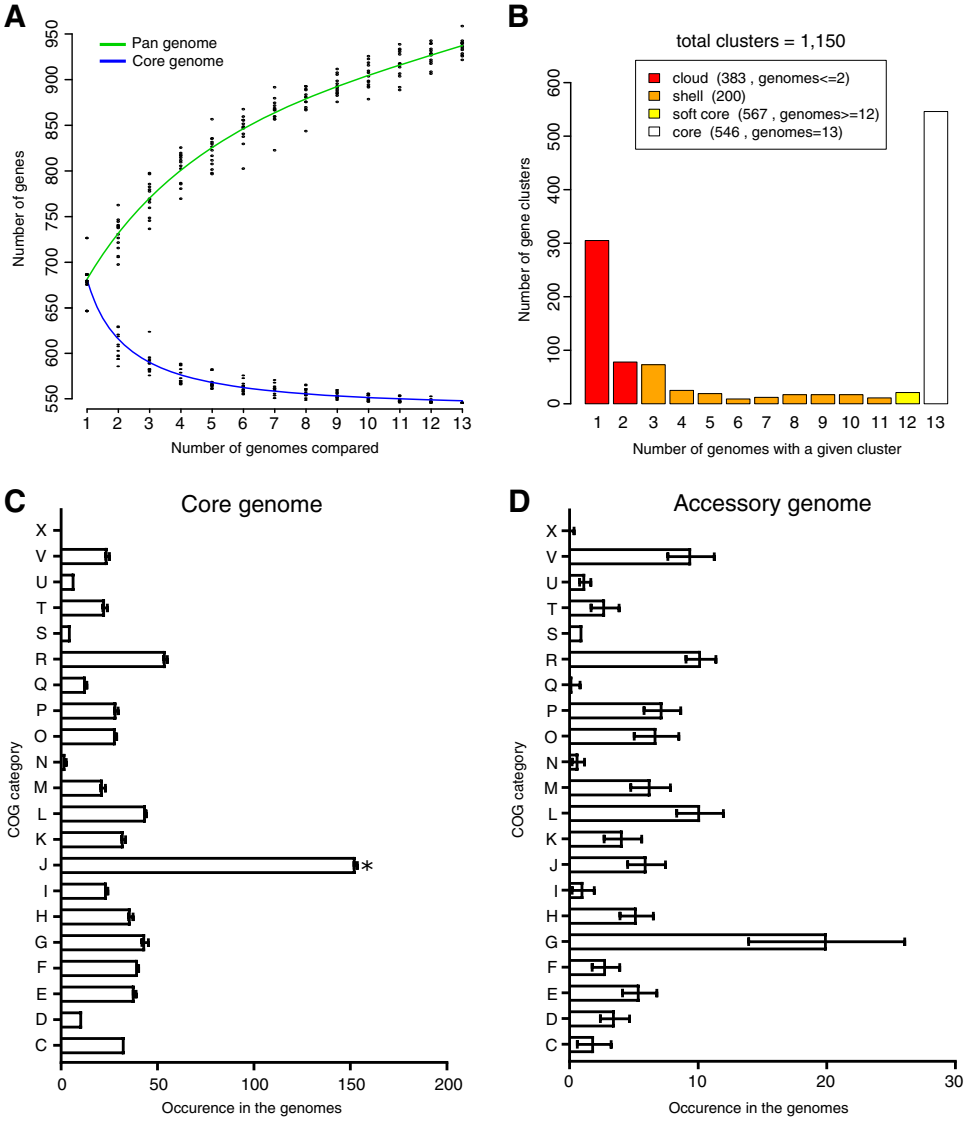

FIG 2 Pangenomes and core genomes of 13 *M. florum* strains. (A) Gene number estimation curves for the core genomes (blue, bottom curve) and pangenomes (green, top curve) were generated by the methods of Willenbrock et al. (60) and Tettelin et al. (61), respectively. (B) Prevalence of the different protein clusters across 13 strains. (C, D) Average number of protein groups in COG categories found in the core (C) and accessory (D) genomes of each strain. The COG categories are as follows: C, energy production and conversion; D, cell cycle control, cell division, and chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure, and biogenesis; K, transcription; L, replication, recombination, and repair; M, cell wall/membrane/envelope biogenesis; N, cell motility; O, posttranslational modification, protein turnover, chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolite biosynthesis, transport, and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms; U, intracellular trafficking, secretion, and vesicular transport; V, defense mechanisms; X, mobilome (prophages, transposons).

with a large number of genes conserved between the different strains and variable regions located within the same relative genomic loci in every strain (Fig. 3). Furthermore, the gene order conservation (GOC) scores (41) of the core genes calculated for every combination of two strains averaged 0.98 ± 0.06, where a GOC value of 1 indicates that all core genes are found in the same order between two strains. We also noted the presence of large weakly conserved or nonconserved regions, consistent with the presence of genomic islands. Regions with at least three consecutive noncore genes were further studied to determine their potential status as genomic islands

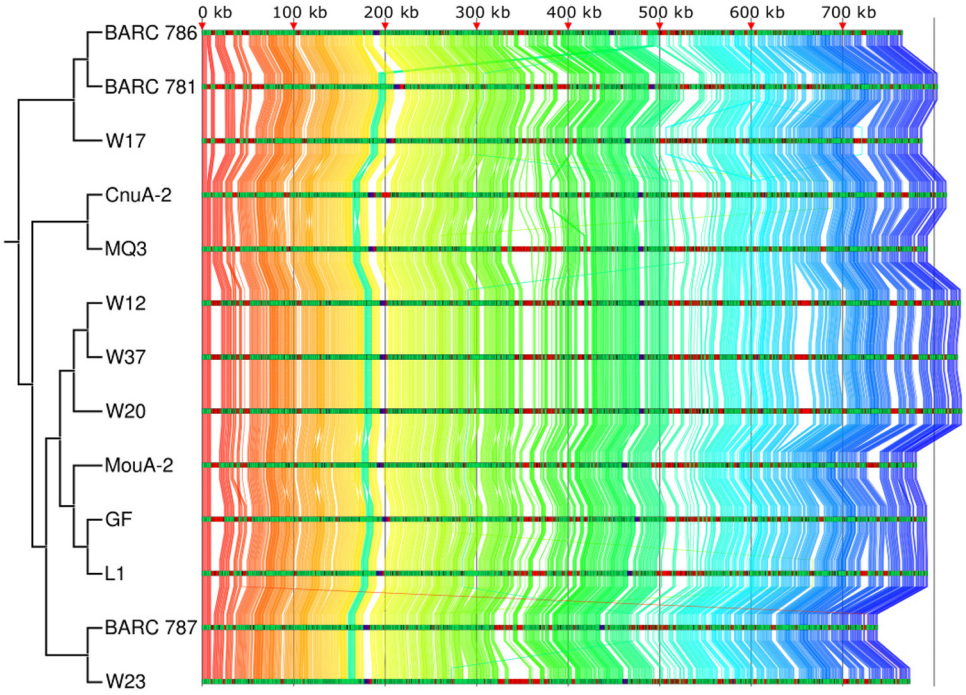

FIG 3 Core genome synteny of the 13 *M. florum* strains. Protein-encoding genes of the core genome are linked across all strains by using a color gradient based on the gene order observed in *M. florum* BARC 786. Each genome track is colored on the basis of core proteins (green), noncore proteins (red), and functional RNAs (purple). The topology of the distance tree is shown on the left.

resulting from horizontal gene transfer. An average of 13.9 ± 3.4 such regions were identified in each genome, representing 19.4% of all noncore gene clusters. One such cluster, detected only in strains W37, W12, and W20, was located near the 660-kbp position and contained 16 hypothetical proteins, as well as 1 protein predicted to be part of a phage tail fiber. Using BLASTP, we searched for protein homologs in other organisms. Of those, 12 had hypothetical protein homologs in spiroplasmas and entomoplasmas, which are the genera closest to mesoplasmas among the members of the class *Mollicutes*. We also detected an 11-gene cluster ranging from position 14,058 to position 24,472 in strain BARC 781. According to the BARC 781 genome annotation, this cluster contained genes related to type IV secretion systems and a predicted mobile element protein, whereas in the other strains this region was enriched in PTS component proteins. Interestingly, the genomic location between translation initiation factor 2 and *dnaJ* was also found to be highly variable, even among closely related strains. Depending on the strain, this locus contained between 1 and 21 genes encoding proteins annotated as PTS components, hypothetical proteins, phage-related proteins, restriction-modification systems, and transcription regulators. These results suggest that horizontal gene transfer events have occurred within the *M. florum* species, shaping its genomic landscape.

Transposon mutagenesis in *M. florum* L1. In addition to comparative genomic data analysis, we also performed transposon mutagenesis to identify essential genes in *M. florum* L1. A collection of 2,806 mutants in which insertions occurred, on average, every ~280 bp across the genome resulted in the interruption of 430 of 720 genes (Data Set S4). No transposon was observed in the remaining 290 genes, which could be essential genes or simply have been missed given the transposon insertion density. The probability of observing no transposon insertions within a gene was calculated on the basis of the transposon insertion density and the length of each individual coding sequence (CDS) as previously described (16, 42), assuming that the probability of having N insertions in a gene of length L follows a Poisson distribution (Data Set S1).

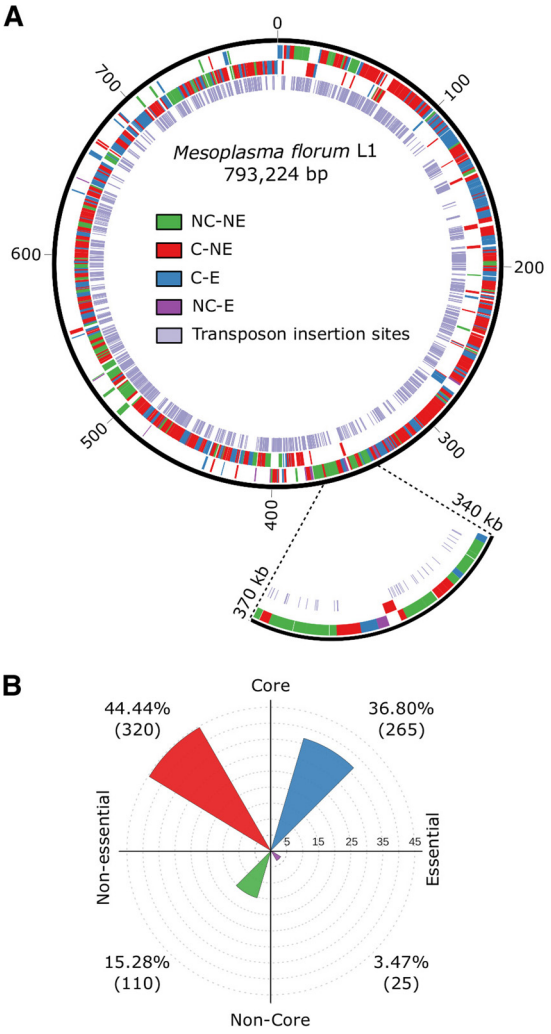

FIG 4 Overview of *M. florum* L1 genomic landscape based on gene conservation and essentiality. (A) Classification of *M. florum* L1 genes shown by a color representing core (C), noncore (NC), essential (E), or nonessential (NE) genes with plus strand genes in the outermost layer and minus strand genes in the middle layer. Transposon insertion sites are also in the innermost layer. The 340- to 370-kbp region is enlarged to show an example of a locus containing all types of gene categories. (B) Gene distribution across the different categories.

Although the transposon insertion density in our work is superior to what was reported in other gene inactivation studies involving *Mollicutes* (12, 43), the average probability that a gene could have been missed by chance in our experiment is ~10%, corresponding to ~69 genes for the entire *M. florum* L1 genome. However, this distribution is strongly skewed toward small genes, with more than half (37/69) of the potentially missed genes having a length of <400 bp. For simplicity, all 290 genes that were not interrupted by a transposon were nonetheless considered putatively essential despite these limitations.

The gene interruption and conservation data for the L1 strain were combined to identify genes that are most likely to be important for *M. florum* (Fig. 4A). Most of the genes presumed to be essential in the L1 strain on the basis of transposon mutagenesis were also conserved across all *M. florum* strains, with only 25 that were not associated with the core genome (Fig. 4B). A putative function could be attributed to only 5 of these 25 noncore but essential genes, of which 4 could be associated with a COG category and the remaining 20 were annotated as hypothetical proteins (Data Set S1). The 110 genes interrupted by transposons and absent from the core genome represent interesting first-step candidates for genome streamlining.

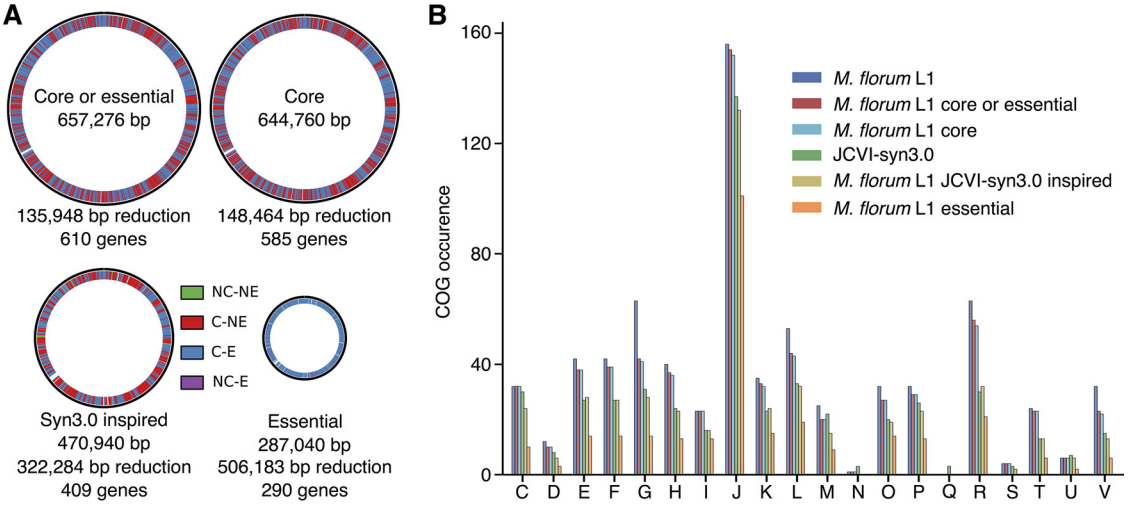

FIG 5 Genome reduction designs for *M. florum* L1. (A) Representation of four different versions of reduced *M. florum* L1 genome based on gene conservation, function, and essentiality. Genes are shown by a color representing core (C), noncore (NC), essential (E), or nonessential (NE) genes. In each case, the number of deleted bases is shown and corresponds to the sum of the lengths of the CDSs of the deleted genes. (B) Number of protein-encoding genes in each COG categories found in the different designs. The COG categories are as described in the legend to Fig. 2.

Potential scenarios for *M. florum* genome reduction. Three reduced versions of the *M. florum* L1 genome were designed on the basis of the gene conservation and transposon mutagenesis data (Fig. 5A). The first scenario excluded all noncore genes that were also interrupted by a transposon (610 genes remaining), the second contained only the core genes (585 genes from the 546 conserved gene cluster families in *M. florum* L1), and the third included only the putatively essential genes (290 genes). Both rRNA loci were included in these genome configurations, although only a single copy could be sufficient to sustain growth (2, 44). In each scenario, the approximate genome size was estimated by removing the CDSs of the candidate genes, while all non-CDSs were kept since most of the promoters and regulatory sequences in *M. florum* remain to be identified. A fourth genome reduction strategy was prepared by including all of the *M. florum* L1 protein-coding genes having an ortholog in *M. mycoides* JCVI-syn3.0, currently the closest approximation of a minimal genome. Interestingly, orthologs were identified for 401 of the 585 *M. florum* core genes. The *M. mycoides* JCVI-syn3.0-inspired genome reduction approach also included 5 genes identified as nonessential and noncore, while 57 genes marked as essential in *M. florum* L1 were absent.

We compared the COG category occurrences in the genome reduction strategies, as well as in the original *M. florum* L1 and *M. mycoides* JCVI-syn3.0 genomes (Fig. 5B). Globally, the proportions of the COG categories were similar in the genome reduction scenarios. Genes related to translation (COG category J) were more represented than any other. Genes related to lipid metabolism (I) made up the second most conserved category, although no gene predicted to perform lipid degradation or elongation was detected in *M. florum*. This category instead contained multiple acyltransferases, which could be used to anchor lipoproteins to the membrane (35), as well as a number of NADH and short-chain dehydrogenases. Overall, the carbohydrate metabolism (G) category contained the largest number of accessory genes since even in the most conservative proposition, 33.8% of these genes were identified as dispensable. Of those, 72.7% were PTS transport components or 6-phospho-β-glucosidases whose functions were duplicated.

DISCUSSION

Minimal cells constitute powerful tools to better understand the fundamental components and the basic mechanisms that support life. The first approximation of a

minimal gene set was recently provided with the creation of *M. mycoides* JCVI-syn3.0 (2). Technical advances now also enable the exploration of the *M. florum* minimal genome. The development of *oriC*-based plasmids and antibiotic selection markers (27) constituted the basic steps that led to the whole-genome cloning of *M. florum* in yeast (29). This was followed by the establishment of a genome transplantation protocol for *M. florum* and by the investigation of the impact of phylogenetic distance on this procedure (28). *M. florum* is therefore a bona fide candidate for genome reduction. However, this raises a few questions. Which genes should be removed to obtain a minimal *M. florum* genome? Given their phylogenetic proximity, would a minimal *M. florum* genome differ from or be equivalent to the minimal *M. mycoides* JCVI-syn3.0 genome? What could be learned by creating minimal genomes based on different cell chassis?

Two different approaches, comparative genomics and random transposon mutagenesis, were used to determine the gene composition of a putative minimal *M. florum* genome. The former exposed genes important for the survival of *M. florum* in its natural habitat, whereas the latter revealed the genes likely to be essential under laboratory conditions. Through the analysis of 13 different strains (Fig. 1), we determined the composition of the *M. florum* core genome and explored the diversity of its pangenome (Fig. 2). Although some strains were isolated from distant sites (Fig. 1A) and from different plants or insects (see Table 1), a total of 546 different protein-coding gene clusters, out of an average of 688 $\pm$ 23 per strain, were found to compose the core *M. florum* genome. Random transposon mutagenesis of strain L1 predicted a total of ~430 dispensable and ~290 putatively essential genes under laboratory conditions. It is possible that the relatively low transposon insertion density (on average, one insertion every ~280 bp) spared a small number of genes simply by chance, which would result in the inclusion of a few dispensable genes in the minimal genome. However, this is unlikely to significantly affect our general conclusions about which genes should be deleted first during an eventual reduction of the *M. florum* genome. Generating additional transposon insertion mutants would, however, increase the precision and confidence level of these predictions, especially for small genes.

Combining comparative genomics and transposon mutagenesis data can provide contrasting perspectives on which genes should be included in a minimal *M. florum* genome. While the 585 core genes could be expected to be sufficient for the survival of *M. florum* L1, ~25 noncore genes are expected to be essential according to our transposon mutagenesis of *M. florum* L1 (Fig. 4B). A minimal genome design based on conserved genes only is thus highly unlikely to produce a viable cell. This can be explained by the differences in the growth conditions and evolutionary pressures experienced by *M. florum* in the environment compared to laboratory settings. In fact, a majority, 320 (55%), of the of the 585 *M. florum* core genes are not essential in rich medium (Fig. 4B). An alternative scenario that includes only the 290 putatively essential genes is also questionable, as synthetic lethality is likely to occur and result in a nonviable minimal *M. florum* genome. This interpretation is supported by the fact that initially proposed minimal *M. mycoides* genome designs based on transposon mutagenesis and other literature-based knowledge were not viable (2). Preservation of both the core and essential genes would remove a total of 110 genes, which has a reasonable chance of success but would most probably remain far from the minimal genome composition.

Another possibility is to infer the minimal *M. florum* L1 genome on the basis of *M. mycoides* JCVI-syn3.0. A total of 409 *M. florum* L1 genes have homologs in *M. mycoides* JCVI-syn3.0. Of these, 404 are part of the *M. florum* L1 "core or putatively essential" gene set. Since all of the genes present in *M. mycoides* JCVI-syn3.0 are essential or have a strong impact on cell fitness, this reveals interesting differences between these organisms. Despite their phylogenetic relatedness, 69 gene families are found only in *M. mycoides* JCVI-syn3.0. Conversely, 57 putatively essential *M. florum* L1 genes have no homolog in *M. mycoides* JCVI-syn3.0 (Data Set S1). It is possible that some of these genes perform equivalent functions although their sequences differ

significantly. However, a majority of these *M. florum* L1 (~61%) and *M. mycoides* JCVI-syn3.0 (~54%) genes are annotated as encoding putative or hypothetical proteins with no clear function, making further investigations more difficult. This highlights our current inability to unambiguously assign functions to a large number of genes and to analyze cell physiology by using a truly functional perspective, which constitutes a major challenge for biology. Genome scale *in silico* models (45) would constitute an attractive tool to help organize, refine, and compare the available information on minimal genomes. Nevertheless, a scenario emerging from this comparison would be to combine the 57 putatively essential genes found only in *M. florum* L1 to the 409 genes that have a homolog in *M. mycoides* JCVI-syn3.0. This would likely represent a better approximation of a minimal *M. florum* genome, given the data currently available. This also implies that the genome-reduced versions of these two organisms would, in large part, be similar but still differ despite their phylogenetic relatedness.

What could be the conceptual nature of the differences observed between *M. mycoides* JCVI-syn3.0 and the proposed *M. florum* minimal genome? In principle, the minimal genome can be divided into three categories, a hard, a semihard, and a soft minimal genome. The hard minimal genome includes genes encoding functions that are essential and performed in a similar fashion across different strains or species (i.e., genome replication, protein synthesis, etc.). The semihard category contains functions essential for any organism but for which alternative genes or strategies are possible to fulfill the same requirement. For instance, different gene families can ensure the same functions, as exemplified by nonorthologous gene displacement (46). The soft minimal genome is, on the other hand, composed of genes that are crucial in a given organism or environment but not necessarily in others. The availability of particular nutrients in the environment or the presence of a particular gene that affects the essentiality of other genes represents a possible factor affecting the soft minimal genome. The differences between *M. florum* L1 and *M. mycoides* JCVI-syn3.0 should vastly reside in either the semihard or soft minimal genome category. Since the semihard minimal genome of phylogenetically closely related bacteria is expected to be relatively small, the soft minimal genome is more likely to explain the distinctions between minimal *M. florum* L1 and *M. mycoides* JCVI-syn3.0 genomes. Indeed, the gene composition of these strains derives from data obtained in rich but slightly different media. Transposon mutagenesis of both strains in a set of different media would presumably lead to the identification of many environment-specific essential genes.

In conclusion, although the technology needed to build entire genomes is now accessible, synthetic genomics is increasingly limited by our understanding of cell functioning. A significant fraction of genetic components are still poorly characterized, even in the most thoroughly studied organisms. Because of their lower complexity, minimal genomes offer a remarkable opportunity to investigate the most fundamental cellular functions that support life. Furthermore, the construction of minimal synthetic chromosomes will facilitate the generation of several genome versions that could help better define the rules governing genome organization. The use of minimal cells will also facilitate the establishment of comprehensive whole-cell models, which is currently hindered by excessive biological complexity. These models could become powerful tools to predict cell behavior and to create synthetic genomes (47). Overcoming these important challenges will constitute a stepping stone toward the rational design and programming of complete genomes.

MATERIALS AND METHODS

Culture conditions and molecular biology methods. *M. florum* strains were grown at 34°C in ATCC 1161 medium (27). Genomic DNA (gDNA) extraction was performed with the Quick-gDNA MiniPrep kit (Zymo Research) for the preparation of Illumina libraries and with Puregene Yeast/Bact. kit B (Qiagen) for PCR-free Sanger sequencing and Pacific Biosciences libraries. Extractions were made in accordance with the manufacturer's instructions, except that *M. florum* cells were washed in a resuspension buffer (8 mM HEPES, 272 mM sucrose; pH 7.4) prior to gDNA extraction.

Genome assemblies and annotation. Genomes were assembled as previously described for the *M. florum* W37 genome (26). For each strain, two Illumina libraries were prepared, one with 200- to 250-bp inserts and the second with 450- to 750-bp inserts, both sequenced in paired-end reads of 144

and 100 bp, respectively. Pacific Biosciences RS libraries (obtained by C_2 chemistry) with inserts of >5 kbp were also prepared. Error correction of the Pacific Biosciences reads was performed by using the Illumina reads, and all sequences were subsequently assembled with Roche gsAssembler version 2.6 and Ray version 2.1.0 (48). The two assemblies were merged and manually inspected before being completed (or completely scaffolded; Table 1) by Sanger sequencing of selected PCR products (the primers used are available on request). PCRs were performed with VeraSeq DNA polymerase (Enzymatics). The routine PCR conditions used were 30 s at 95°C; 30 cycles of 10 s at 95°C, 30 s at the appropriate annealing temperature, and 30 s/kbp at 72°C; and 2 min at 72°C. PCR products were purified by solid-phase reversible immobilization bead capture with Agencourt AMPure XP magnetic beads (Beckman Coulter, Inc.) (49). Sanger sequencing reads were generated by the Plateforme de Séquençage et de Génotypage of the Centre de Recherche du Centre Hospitalier de l'Université Laval. Pacific Biosciences sequencing was performed by the Yale Center for Genome Analysis and at the Centre d'innovation Génome Québec et Université McGill. All of the genomes were annotated with the RAST server (50) and FIGfam Release 70.

Comparative genomics and functional analysis. The annotated genomes were analyzed with get_homologues (51) (v.2.0) to identify the core genome and pangenome by the COGtriangle (37) method for bidirectional best-hit determination. The comparison of the *M. florum* strains and *M. capricolum* subsp. *capricolum* ATCC 2734 (NCBI RefSeq accession no. NC_007633.1) to identify the protein set used for the phylogenetic tree creation was performed by the same method, as was the comparison of *M. florum* L1 and the synthetic bacterium Syn3.0 (GenBank accession no. CP014940.1). Genome synteny was determined by using the core protein cluster coordinates and visualizing their locations across all of the genomes by using the GMV genome browser (v.1e-93) (52). GOC scores were determined by using the core genes as previously described (41). Genes separated by fewer than five core genes were considered contiguous (41). For the functional analysis, the latest COG database (53) was downloaded from the NCBI ftp website (ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/data/). The proteins predicted in the 13 *M. florum* genomes were compared to the database with BLAST (54). The results were filtered on the basis of the E value with a stringent threshold of 1e-10. Available conversion tables from the NCBI ftp website were used to convert matching proteins to COG identification numbers to functional categories. Enriched categories were determined with the Fisher exact test and Bonferroni correction for multiple tests by comparing the frequency of genes in a COG category in the *M. florum* core or accessory genome with that in the entire genome.

Phylogenetic tree construction. An alignment of the amino acid sequence of 412 conserved proteins of the 13 sequenced *M. florum* strains and *M. capricolum* was made with ClustalO (v.1.2.1) (55). Unaligned and low-confidence regions were removed with Gblocks (v.0.91b) (56) to produce a sequence matrix of 138,476 amino acid sites. Both phylogenetic trees (Fig. 1B and C) were made from this alignment with SeaView (v.4.6.1) (57). The distance tree was generated by neighbor joining with BIONJ (58) and the Kimura distance model (38). The maximum-likelihood tree was generated with PhyML (v3.0) (39) and the LG evolutionary model (59). Bootstrap values were calculated by using 100 regular bootstrap replicates.

Transposon mutagenesis. Tn*5* transposomes were assembled *in vitro* with EZ-Tn*5* transposase (Epicentre) without Mg^{2+} as recommended by the manufacturer. The transposon DNA was obtained from the digestion of the pTT01 plasmid with restriction enzyme PvuII and conferred tetracycline resistance. The transposomes were transformed via electroporation as described by Matteau et al. (27) with a voltage of 2.5 kV. Insertion mutants were selected on ATCC 1161 solid medium supplemented with 15 μg/ml tetracycline. Colonies were picked as they became visible, with certain mutants growing more slowly than the parental strain. The transposon insertion site was determined by Sanger sequencing of gDNA. Genes that contained at least one transposon insertion were considered nonessential.

Data availability. The complete genome sequences and annotations of the 11 newly sequenced *M. florum* strains are available in GenBank under the following accession numbers: BARC 781, CP022511; BARC 786, CP022510; BARC 787, CP022514; CnuA-2, CP022513; GF, CP022509; MouA-2, CP022508; MQ3, CP022512; W17, CP022507; W20, CP022506; W23, CP022505; W12, CP022432. The comparison of NCBI annotations and gene predictions used in our study for strain L1 is presented in Data Set S1. Gene predictions for the 12 other *M. florum* strains can be found in Data Set S2. The composition of the *M. florum* COG can be found in Data Set S3. The transposon insertion sites observed in *M. florum* L1 are provided in Data Set S4.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSystems.00198-17.

DATA SET S1, XLSX file, 0.1 MB.

DATA SET S2, XLSX file, 4.2 MB.

DATA SET S3, XLSX file, 0.1 MB.

DATA SET S4, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

We thank members of the Rodrigue and Jacques laboratory and Alain Lavigueur at the Université de Sherbrooke for helpful discussions and comments on the manuscript. We also thank Gail E. Gasparich and Robert E. Davis for their help in retracing some

M. florum strain origins. We are also grateful for technical assistance provided by Samantha L. Burke during the generation of the *M. florum* L1 Tn*5* transposon mutagenesis library and by the team of the Centre de Calcul Scientifique at the Université de Sherbrooke. Access to computational resources was provided in part by Calcul Québec (http://www.calculquebec.ca) and Compute Canada (http://www.computecanada.ca).

This research was supported by the Natural Sciences and Engineering Research Council of Canada, the Fonds de Recherche du Québec Nature et technologies, and the Defense Advanced Research Projects Agency under research contract HR0011-12-C-0010. The views expressed here are those of the authors and do not reflect the official policy or position of the Department of Defense of the United States Government. Tom F. Knight is a founder of Ginkgo BioWorks Inc. (Boston, MA) and is currently employed by that company.

V.B. and S.R. wrote the manuscript; V.B. and J.-C.L. prepared the figures; V.B. and D.M. performed the experiments; V.B., J.-C.L., J.G., and J.-F.L. performed the analysis; and V.B., T.K., and S.R. designed the project.

REFERENCES

1. Richardson SM, Mitchell LA, Stracquadanio G, Yang K, Dymond JS, DiCarlo JE, Lee D, Huang CLV, Chandrasegaran S, Cai Y, Boeke JD, Bader JS. 2017. Design of a synthetic yeast genome. Science 355:1040–1044. https://doi.org/10.1126/science.aaf4557.
2. Hutchison CA, III, Chuang RY, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L, Pelletier JF, Qi ZQ, Richter RA, Strychalski EA, Sun L, Suzuki Y, Tsvetanova B, Wise KS, Smith HO, Glass JI, Merryman C, Gibson DG, Venter JC. 2016. Design and synthesis of a minimal bacterial genome. Science 351:aad6253. https://doi.org/10.1126/science.aad6253.
3. Annaluru N, Ramalingam S, Chandrasegaran S. 2015. Rewriting the blueprint of life by synthetic genomics and genome engineering. Genome Biol 16:125. https://doi.org/10.1186/s13059-015-0689-y.
4. Csörgő B, Nyerges Á, Pósfai G, Fehér T. 2016. System-level genome editing in microbes. Curr Opin Microbiol 33:113–122. https://doi.org/10.1016/j.mib.2016.07.005.
5. Koonin EV. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 1:127–136. https://doi.org/10.1038/nrmicro751.
6. Charlebois RL, Doolittle WF. 2004. Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res 14:2469–2477. https://doi.org/10.1101/gr.3024704.
7. Harris JK, Kelley ST, Spiegelman GB, Pace NR. 2003. The genetic core of the universal ancestor. Genome Res 13:407–412. https://doi.org/10.1101/gr.652803.
8. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. 2015. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141–161. https://doi.org/10.1007/s10142-015-0433-4.
9. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2:2006.0008. https://doi.org/10.1038/msb4100050.
10. Datsenko KA, Wanner BL. 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc Natl Acad Sci U S A 97:6640–6645. https://doi.org/10.1073/pnas.120163297.
11. Christen B, Abeliuk E, Collier JM, Kalogeraki VS, Passarelli B, Coller JA, Fero MJ, McAdams HH, Shapiro L. 2011. The essential genome of a bacterium. Mol Syst Biol 7:528. https://doi.org/10.1038/msb.2011.58.
12. Hutchison CA, III, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter JC. 1999. Global transposon mutagenesis and a minimal *Mycoplasma* genome. Science 286:2165–2169. https://doi.org/10.1126/science.286.5447.2165.
13. Forsyth RA, Haselbeck RJ, Ohlsen KL, Yamamoto RT, Xu H, Trawick JD, Wall D, Wang L, Brown-Driver V, Froelich JM, C KG, King P, McCarthy M, Malone C, Misiner B, Robbins D, Tan Z, Zhu Zy ZY, Carr G, Mosca DA, Zamudio C, Foulkes JG, Zyskind JW. 2002. A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. Mol Microbiol 43:1387–1400. https://doi.org/10.1046/j.1365-2958.2002.02832.x.
14. Peters JM, Colavin A, Shi H, Czarny TL, Larson MH, Wong S, Hawkins JS, Lu CHS, Koo BM, Marta E, Shiver AL, Whitehead EH, Weissman JS, Brown ED, Qi LS, Huang KC, Gross CA. 2016. A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. Cell 165:1493–1506. https://doi.org/10.1016/j.cell.2016.05.003.
15. Nichols RJ, Sen S, Choo YJ, Beltrao P, Zietek M, Chaba R, Lee S, Kazmierczak KM, Lee KJ, Wong A, Shales M, Lovett S, Winkler ME, Krogan NJ, Typas A, Gross CA. 2011. Phenotypic landscape of a bacterial cell. Cell 144:143–156. https://doi.org/10.1016/j.cell.2010.11.052.
16. Balázsi G. 2008. Statistical evaluation of genetic footprinting data, p 355–359. *In* Ostermam AL, Gerdes SY (ed), Microbial gene essentiality: protocols and bioinformatics. Humana Press, Totowa, NJ.
17. Butland G, Babu M, Díaz-Mejía JJ, Bohdana F, Phanse S, Gold B, Yang W, Li J, Gagarinova AG, Pogoutse O, Mori H, Wanner BL, Lo H, Wasniewski J, Christopolous C, Ali M, Venn P, Safavi-Naini A, Sourour N, Caron S, Choi JY, Laigle L, Nazarians-Armavil A, Deshpande A, Joe S, Datsenko KA, Yamamoto N, Andrews BJ, Boone C, Ding H, Sheikh B, Moreno-Hagelseib G, Greenblatt JF, Emili A. 2008. eSGA: *E. coli* synthetic genetic array analysis. Nat Methods 5:789–795. https://doi.org/10.1038/nmeth.1239.
18. Pósfai G, III, Plunkett G, Fehér T, Frisch D, Keil GM, Umenhoffer K, Kolisnychenko V, Stahl B, Sharma SS, de Arruda M, Burland V, Harcum SW, Blattner FR. 2006. Emergent properties of reduced-genome *Escherichia coli*. Science 312:1044–1046. https://doi.org/10.1126/science.1126439.
19. Park MK, Lee SH, Yang KS, Jung SC, Lee JH, Kim SC. 2014. Enhancing recombinant protein production with an *Escherichia coli* host strain lacking insertion sequences. Appl Microbiol Biotechnol 98:6701–6713. https://doi.org/10.1007/s00253-014-5739-y.
20. Hashimoto M, Ichimura T, Mizoguchi H, Tanaka K, Fujimitsu K, Keyamura K, Ote T, Yamakawa T, Yamazaki Y, Mori H, Katayama T, Kato J. 2005. Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. Mol Microbiol 55:137–149. https://doi.org/10.1111/j.1365-2958.2004.04386.x.
21. Mizoguchi H, Mori H, Fujio T. 2007. *Escherichia coli* minimum genome factory. Biotechnol Appl Biochem 46:157–167. https://doi.org/10.1042/BA20060107.
22. Reuß DR, Altenbuchner J, Mäder U, Rath H, Ischebeck T, Sappa PK, Thürmer A, Guérin C, Nicolas P, Steil L, Zhu B, Feussner I, Klumpp S, Daniel R, Commichau FM, Völker U, Stülke J. 2017. Large-scale reduction of the *Bacillus subtilis* genome: consequences for the transcriptional network, resource allocation, and metabolism. Genome Res 27:289–299. https://doi.org/10.1101/gr.215293.116.
23. Komatsu M, Uchiyama T, Omura S, Cane DE, Ikeda H. 2010. Genome-minimized *Streptomyces* host for the heterologous expression of secondary metabolism. Proc Natl Acad Sci U S A 107:2646–2651. https://doi.org/10.1073/pnas.0914833107.
24. Leprince A, de Lorenzo V, Völler P, van Passel MW, Martins dos Santos VA. 2012. Random and cyclical deletion of large DNA segments in the genome of *Pseudomonas putida*. Environ Microbiol 14:1444–1453. https://doi.org/10.1111/j.1462-2920.2012.02730.x.
25. McCoy RE, Basham HG, Tully JG, Rose DL, Carle P, Bove JM. 1984.

Acholeplasma florum, a new species isolated from plants. Int J Syst Bacteriol 34:11–15. https://doi.org/10.1099/00207713-34-1-11.
26. Baby V, Matteau D, Knight TF, Rodrigue S. 2013. Complete genome sequence of the *Mesoplasma florum* W37 strain. Genome Announc 1:e00879-13. https://doi.org/10.1128/genomeA.00879-13.
27. Matteau D, Pepin ME, Baby V, Gauthier S, Arango Giraldo M, Knight TF, Rodrigue S. 2017. Development of *oriC*-based plasmids for *Mesoplasma florum*. Appl Environ Microbiol 83:e03374-16. https://doi.org/10.1128/AEM.03374-16.
28. Labroussaa F, Lebaudy A, Baby V, Gourgues G, Matteau D, Vashee S, Sirand-Pugnet P, Rodrigue S, Lartigue C. 2016. Impact of donor–recipient phylogenetic distance on bacterial genome transplantation. Nucleic Acids Res 44:8501–8511. https://doi.org/10.1093/nar/gkw688.
29. Baby V, Labroussaa F, Brodeur J, Matteau D, Gourgues G, Lartigue C, Rodrigue S. 2018. Cloning and transplantation of the *Mesoplasma florum* genome. ACS Synth Biol 7:209–217. https://doi.org/10.1021/acssynbio.7b00279.
30. Tully JG, Bove JM, Laigret F, Whitcomb RF. 1993. Revised taxonomy of the class *Mollicutes*: proposed elevation of a monophyletic cluster of arthropod-associated *Mollicutes* to ordinal rank (*Entomoplasmatales* ord. nov.), with provision for familial rank to separate species with nonhelical morphology. Int J Syst Bacteriol 43:630–630. https://doi.org/10.1099/00207713-43-3-630a.
31. Tully JG, Whitcomb RF, Hackett KJ, Rose DL, Henegar RB, Bové JM, Carle P, Williamson DL, Clark TB. 1994. Taxonomic descriptions of eight new non-sterol-requiring *Mollicutes* assigned to the genus *Mesoplasma*. Int J Syst Bacteriol 44:685–693. https://doi.org/10.1099/00207713-44-4-685.
32. Whitcomb RF, Tully JG, Rose DL, Stephens EB, Smith A, McCoy RE, Barile MF. 1982. Wall-less prokaryotes from fall flowers in central United States and Maryland. Curr Microbiol 7:285–290.
33. Tully JG, Rose DL, Whitcomb RF, Hackett KJ, Clark TB, Henegar RB, Clark E, Carle P, Bove JM. 1987. Characterization of some new insect-derived acholeplasmas. Isr J Med Sci 23:699–703.
34. Clark TB, Tully JG, Rose DL, Henegar R, Whitcomb RF. 1986. Acholeplasmas and similar nonsterol-requiring Mollicutes from insects: missing link in microbial ecology. Curr Microbiol 13:11–16. https://doi.org/10.1007/BF01568152.
35. Razin S, Yogev D, Naot Y. 1998. Molecular biology and pathogenicity of mycoplasmas. Microbiol Mol Biol Rev 62:1094–1156.
36. Koonin EV, Wolf YI. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res 36:6688–6719. https://doi.org/10.1093/nar/gkn668.
37. Kristensen DM, Kannan L, Coleman MK, Wolf YI, Sorokin A, Koonin EV, Mushegian A. 2010. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics 26:1481–1487. https://doi.org/10.1093/bioinformatics/btq229.
38. Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120. https://doi.org/10.1007/BF01731581.
39. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321. https://doi.org/10.1093/sysbio/syq010.
40. Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637. https://doi.org/10.1126/science.278.5338.631.
41. Rocha EPC. 2006. Inference and analysis of the relative stability of bacterial chromosomes. Mol Biol Evol 23:513–522. https://doi.org/10.1093/molbev/msj052.
42. Gerdes SY, Scholle MD, Campbell JW, Balázsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabási AL, Oltvai ZN, Osterman AL. 2003. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. J Bacteriol 185:5673–5684. https://doi.org/10.1128/JB.185.19.5673-5684.2003.
43. French CT, Lao P, Loraine AE, Matthews BT, Yu H, Dybvig K. 2008. Large-scale transposon mutagenesis of *Mycoplasma pulmonis*. Mol Microbiol 69:67–76. https://doi.org/10.1111/j.1365-2958.2008.06262.x.
44. Asai T, Zaporojets D, Squires C, Squires CL. 1999. An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. Proc Natl Acad Sci U S A 96:1971–1976. https://doi.org/10.1073/pnas.96.5.1971.
45. Danchin A, Fang G. 2016. Unknown unknowns: essential genes in quest for function. Microb Biotechnol 9:530–540. https://doi.org/10.1111/1751-7915.12384.
46. Koonin EV, Mushegian AR, Bork P. 1996. Non-orthologous gene displacement. Trends Genet 12:334–336. https://doi.org/10.1016/0168-9525(96)20010-1.
47. Karr JR, Sanghvi JC, MacKlin DN, Gutschow MV, Jacobs JM, Bolival B, Assad-Garcia N, Glass JI, Covert MW. 2012. A whole-cell computational model predicts phenotype from genotype. Cell 150:389–401. https://doi.org/10.1016/j.cell.2012.05.044.
48. Boisvert S, Laviolette F, Corbeil J. 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J Comput Biol 17:1519–1533. https://doi.org/10.1089/cmb.2009.0238.
49. Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ, Chisholm SW. 2010. Unlocking short read sequencing for metagenomics. PLoS One 5:e11840. https://doi.org/10.1371/journal.pone.0011840.
50. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. 2008. The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75. https://doi.org/10.1186/1471-2164-9-75.
51. Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol 79:7696–7701. https://doi.org/10.1128/AEM.02411-13.
52. Popendorf K, Tsuyoshi H, Osana Y, Sakakibara Y. 2010. Murasaki: a fast, parallelizable algorithm to find anchors from multiple genomes. PLoS One 5:e12651. https://doi.org/10.1371/journal.pone.0012651.
53. Galperin MY, Makarova KS, Wolf YI, Koonin EV. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43:D261–D269. https://doi.org/10.1093/nar/gku1223.
54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
55. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75.
56. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334.
57. Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221–224. https://doi.org/10.1093/molbev/msp259.
58. Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685–695. https://doi.org/10.1093/oxfordjournals.molbev.a025808.
59. Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol 25:1307–1320. https://doi.org/10.1093/molbev/msn067.
60. Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW. 2007. Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. Genome Biol 8:R267. https://doi.org/10.1186/gb-2007-8-12-r267.
61. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." Proc Natl Acad Sci U S A 102:13950–13955. https://doi.org/10.1073/pnas.0506758102.