**Candidatus Mycoplasma girerdii** replicates, diversifies, and co-occurs with *Trichomonas vaginalis* in the oral cavity of a premature infant

Elizabeth K. Costello¹, Christine L. Sun², Erica M. Carlisle³, Michael J. Morowitz⁴, Jillian F. Banfield⁵ & David A. Relman¹,²,⁶

Genital mycoplasmas, which can be vertically transmitted, have been implicated in preterm birth, neonatal infections, and chronic lung disease of prematurity. Our prior work uncovered 16S rRNA genes belonging to a novel, as-yet-uncultivated mycoplasma (lineage ‘Mnola’) in the oral cavity of a premature neonate. Here, we characterize the organism’s associated community, growth status, metabolic potential, and population diversity. Sequencing of genomic DNA from the infant’s saliva yielded 1.44 Gbp of high-quality, non-human read data, from which we recovered three essentially complete (including ‘Mnola’) and three partial draft genomes (including *Trichomonas vaginalis*). The completed 629,409-bp ‘Mnola’ genome (*Candidatus* Mycoplasma girerdii str. UC-B3) was distinct at the strain level from its closest relative, vaginally-derived *Ca. M. girerdii* str. VCU-M1, which is also associated with *T. vaginalis*. Replication rate measurements indicated growth of str. UC-B3 within the infant. Genes encoding surface-associated proteins and restriction-modification systems were especially diverse within and between strains. In UC-B3, the population genetic underpinnings of phase variable expression were evident in vivo. Unique among mycoplasmas, *Ca. M. girerdii* encodes pyruvate-ferredoxin oxidoreductase and may be sensitive to metronidazole. This study reveals a metabolically unique mycoplasma colonizing a premature neonate, and establishes the value of genome-resolved metagenomics in tracking phase variation.

*Mycoplasma* species span three clades within the class Mollicutes, a group of obligate parasitic bacteria, affiliated with the phylum *Firmicutes*, that lack a cell wall. They establish relationships as commensals and pathogens of vertebrate hosts; in humans, colonization is typically limited to the mucosal surfaces of the respiratory and urogenital tracts. Invasive or disseminated infections do occur, but are more likely in immunocompromised or instrumented patients. Mycoplasmas are typically extracellular and surface-attached, but a few species can enter host cells and establish an intracellular niche. They evade the host immune system, in part by varying the expression of diverse surface proteins¹². Infections are often chronic, with untoward host immune responses contributing to the pathology¹³. Mycoplasmas are characterized by small cells, small genomes, and metabolic dependence on other organisms—qualities increasingly recognized as widespread across the tree of life¹⁻⁷. AT-biased base composition and the use of genetic code four⁸ further distinguish the genomes. *De novo* biosynthesis of nucleotides, amino acids and fatty acids is minimal to absent in mycoplasmas, and energy metabolism is restricted to glycolysis, arginine

¹Department of Medicine, Division of Infectious Diseases, Stanford University School of Medicine, Stanford, CA, 94305, USA. ²Department of Microbiology & Immunology, Stanford University School of Medicine, Stanford, CA, 94305, USA. ³Department of Surgery, Division of Pediatric Surgery, University of Iowa College of Medicine, Iowa City, IA, 52242, USA. ⁴Department of Surgery, Division of Pediatric Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA, 15224, USA. ⁵Department of Earth & Planetary Science, University of California, Berkeley, CA, 94720, USA. ⁶Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, 94304, USA. Correspondence and requests for materials should be addressed to D.A.R. (email: relman@stanford.edu)
hydrolysis or urea hydrolysis. Reputed as fastidious and slow-growing, generation times in culture range from
~1 hr for Ureaplasma spp., to 3 hr for M. mycoides, 6 hr for M. pneumoniae, and 16 hr for M. genitalium3, 4. Long
recognized as minimal self-replicating organisms, mycoplasmas and their study have sparked advances in genom-
ics, synthetic biology, and the concept of a minimal cell.

Intrauterine infection is a major cause of preterm birth and complications are the most frequently reported
organisms in the amniotic cavity10. Among spontaneous preterm births, the earlier the gestational age at delivery,
the more likely it is to be associated with intrauterine infection11. Vertical transmission is common, and
in newborns, mycoplasmas and ureaplasmas can cause bacteremia, meningitis and pneumonia. In utero expo-
sure to Ureaplasma spp. can also induce inflammatory responses leading to chronic lung injury in infants12.
Infection control in the neonatal intensive care unit (NICU) has traditionally relied on culture-based surveillance.
Given the fastidiousness of some mycoplasmas, and the expanding use of molecular surveillance methods, it is
perhaps surprising that until recently13 no new species had been discovered in this setting4. One caveat is that
cultivation-independent surveys of premature neonates have overwhelmingly prioritized the gut over sites more
likely to harbor mycoplasmas such as the respiratory tract.

Several recent studies have characterized the vaginal microbiota of women with trichomoniasis14–16. In one,
Martin et al.15 reported a nearly full-length 16S rRNA gene sequence from an uncultivated Mycoplasma sp. (lin-
eage ‘Mnola’) that was 85% identical to the closest human-associated species (M. genitalium, M. pneumoniae)
and 94% identical to the closest cloned sequences, which are from cattle rumen and termite gut. These authors
found ‘Mnola’ in 19/30 T. vaginalis-infected and 1/29 uninfected women attending a New Orleans, LA, sexually
transmitted infection (STI) clinic15. In another study, Fettweis et al.16 detected T. vaginalis in 49/51 women whose
vaginal microbiota comprised at least 1% ‘Mnola’. These authors also reported a genome sequence for ‘Mnola’
(618,983-bp Candidatus Mycoplasma girdneri strain VCU-M1), which they assembled from the vaginal metagenome
of a T. vaginalis-positive woman in preterm labor16. Although the number of observations is small, the presence
of Ca. M. girdneri appears to indicate the presence of T. vaginalis, the most prevalent non-viral, sexually trans-
mitted pathogen worldwide, and itself a risk factor for preterm birth10, 17. The presence of T. vaginalis does not
necessarily indicate the presence of Ca. M. girdneri.

We uncovered a novel Ca. M. girdnerii-like 16S rRNA gene sequences in the oral cavity of a vaginally-delivered premature neonate18. To date and to our knowledge, this remains the sole report of ‘Mnola’ in
an infant; other than this case, this phylogenetically novel, as-yet-uncultivated mycoplasma has not been detected
outside of vaginal samples. Here, we applied genome-resolved metagenomics to further examine the oral samples
collected from the infant. Using this approach, we characterized the associated microbial community, growth
status, metabolic potential, and within-population diversity of the infant-associated strain. Furthermore, the
availability of the vaginally-derived genome (that of str. VCU-M1) provided a unique opportunity to compare the
genomes of two closely related populations sampled directly from different host individuals.

Methods

**Human subject, samples, and sequencing.** Clinical details about the subject, a premature male infant,
appear elsewhere (baby #3 therein)3, 12. In brief, the infant was delivered vaginally, weighing 750 grams, at 24 weeks
and 4 days of completed gestation. Preterm premature rupture of membranes (PPROM) complicated the delivery.
As noted in the infant’s chart, the mother received prenatal care throughout her pregnancy and was screened
for chlamydia, gonorrhea, hepatitis B, HIV, syphilis, and group B streptococcus. These tests were negative. The infant
was hospitalized in a level III NICU at the University of Chicago Comer Children’s Hospital. The Institutional
Review Board of the University of Chicago approved the study and the infant’s parents provided written informed
consent. All methods were performed in accordance with the relevant guidelines and regulations.

Stool, saliva, and skin swabs were collected on days of life (DOL) 8, 10, 12, 15, 18 and 213, 18. Before and dur-
ing this time, the infant was treated with intravenous antibiotics, specifically, ampicillin (DOL 1–7), gentamicin
(DOL 1–7 and 13–19), vancomycin (DOL 13–15) and cefotaxime (DOL 14–19). Antibiotics started on DOL 13
were given as empiric coverage for suspected sepsis; cultures of blood, urine, and cerebrospinal fluid were ulti-
mately negative. Respiratory specimens were not tested. The infant was intubated and mechanically ventilated
throughout the study period. As part of a previously published study, a portion of each sample was processed for
bacterial community characterization via the sequencing of amplified 16S rRNA genes13. The remainders were
archived at −80 °C.

**Ca. M. girdnerii-like 16S rRNA gene sequences were particularly abundant in the infant’s saliva on DOL 15
(26% of amplified reads), 18 (47%) and 21 (14%)3, 15. They were not detected in the infant’s stool or on his skin, or in
five other infants in the cohort3, 10. We reasoned that because the novel mycoplasma had been a dominant member
of a low-diversity bacterial community, namely, the oral microbiota of an antibiotic-treated premature infant,
its population genome could be assembled de novo from metagenomic read data. We recovered relatively little
genomic DNA from the infant’s saliva on DOL 15, 18 and 21; therefore, prior to sequencing, we pooled the three
days (see Supplementary Methods and Table S1 for details). Paired-end, 2 × 150 bp, Illumina HiSeq sequencing
yielded 19.41 million raw read-pairs.

**Assembly, preliminary annotation, and bin previews.** After trimming and filtering the raw reads,5.11 million high quality read-pairs (1.44 Gbp of data) were retained and advanced to de novo assembly, which, along
with read mapping, gene prediction and preliminary annotation, was performed as described in the
Supplementary Methods. Assembly using IDBA-UD produced 1,688 scaffolds ≥1 kbp in length, which together
summed to 11.88 Mbp. An overall alignment rate of 70.3% was achieved when the assembly-input reads were
mapped back to these scaffolds (improving modestly to 77.7% for all scaffolds ≥200bp). Prior to formal binning,
we found it useful to obtain an estimate of the number, types, and relative abundances of bins (genomes) present
in the dataset. We did this by creating an inventory of genes encoding ribosomal RNAs (Supplementary Table S2 and Data S1) and ribosomal proteins (Supplementary Fig. S1a).

Creating and assessing bins. Genome bins that were initially delineated using a tetranucleotide frequency-based emergent self-organizing map (ESOM) (Supplementary Fig. S2) were manually curated (i.e., validated and honed) using other genomic signatures (see Supplementary Methods for details). All scaffolds \( \geq 1 \) kbp in length were either binned \((n = 1413; \text{summing to } 11.45 \text{ Mbp})\) or removed \((n = 275 \text{ residual phiX or human scaffolds})\). Approximately 1.15 Gbp mapped back to the binned scaffolds. This estimate included both paired and unpaired reads, but only paired reads were used in the assembly. Bins were annotated in further detail (Supplementary Table S3) and assessed for coverage, abundance, completeness, and other characteristics as described in the Supplementary Methods. Replication indices were calculated using iRep software and the thresholds recommended by the authors\(^{18}\).

Finishing and characterizing the Mycoplasma genome. Minor local scaffolding errors were detected and corrected as described in the Supplementary Methods. The 618,983-bp \(\text{Ca. M. gierdii str. VCU-M1}\) genome was used to guide initial hypotheses about the order and orientation of our scaffolds. Subsequently, we were able to close (in silico) the 629,409-bp genome of the infant oral strain (str. UC-B3; see Supplementary Methods for details). The finished sequence (sensu Sharon and Banfield)\(^{19}\) was confirmed via careful read mapping and error checking; indeed, after depleting our reads of those mapping to str. UC-B3, none remained that mapped to str. VCU-M1. The genomes were aligned (Supplementary Fig. S3) and an average nucleotide identity (ANI) was calculated; for str. UC-B3, we also generated a set of gene predictions and annotations (Supplementary Table S4), inferred its metabolic functional potential (Supplementary Table S4), and compared its genes involved in energy metabolism to those of other Mollicutes (Supplementary Table S5) as described in the Supplementary Methods.

Analyzing length variation in DNA tandem repeats. We generated a list of candidate length-variable DNA tandem repeats, most of which were dinucleotide tandem repeats (DTRs) (Supplementary Table S6), for the finished genome of \(\text{Ca. M. gierdii str. UC-B3}\). For each candidate, we quantified the types and frequencies of length variants present in the population using a read alignment and counting procedure described in the Supplementary Methods. Each length variant’s potential consequence for phase variation was inferred by translating the affiliated open reading frame (ORF) and documenting premature stop codons downstream of the repeat. Where possible, we identified the corresponding homologous repeat in str. VCU-M1 and recorded its length (Supplementary Table S7).

Data availability. Datasets on which the conclusions of this manuscript rely have been deposited under NCBI BioProject PRJNA362245, BioSample SAMN06236793, Sequence Read Archive (SRA) accession code SRR5251628 (these are the raw reads), and Genomes accession code CP020122 (this is the complete genome sequence of \(\text{Ca. M. gierdii str. UC-B3}\)).

Results and Discussion

Gene inventories reveal \(\text{Trichomonas}\) and ample coverage of \(\text{Mycoplasma}\). Inventories of genes encoding rRNAs and ribosomal proteins suggested that the premature infant’s oral metagenome comprised the genomes of as few as six genera: \(\text{Pseudomonas, Mycoplasma, Streptococcus, Enterobacter, Staphylococcus}\) and \(\text{Trichomonas}\). The taxonomic identities and relative abundances of bacteria were similar to those observed in our earlier survey based on amplified 16S rRNA genes\(^{13}\) (Supplementary Tables S1, S2 and Fig. S1a). In addition to bacterial genomic DNA, the infant also harbored genomic DNA likely belonging to \(\text{T. vaginalis}\), the causative agent of human trichomoniasis and a microbial eukaryote ignored by bacterial broad range 16S rRNA gene PCRs. Vertical transmission of \(\text{T. vaginalis}\) has been documented, but is considered uncommon\(^{20–22}\). As expected, the depth of sequencing achieved for the targeted \(\text{Mycoplasma}\), estimated at \(\sim 80\)-fold average coverage (Supplementary Fig. S1a), was sufficient for genome reconstruction.

Binning recovers three essentially complete and three partial draft genomes. As predicted from our gene inventories, binning resulted in six genome bins (Supplementary Fig. S2). Table 1 describes them in detail and Supplementary Table S3 provides basic annotations. Bins 1–3, including the target \(\text{Mycoplasma}\) sp. ‘\(\text{Mnola}\)’ (bin 2), were most abundant; these draft genomes were essentially complete \(\text{(sensu Sharon and Banfield)}\)\(^{19}\) as indicated by their size and other estimates of completeness (Table 1). Bins 4–6 were incomplete, yet offer important clues about the mycoplasma’s microbial community context, e.g., bin 6 contains 149 scaffolds assigned to \(\text{Trichomonas}\) (Table 1).

Owing in part, we surmised, to recent antibiotic exposure, the microbial diversity of the infant’s oral cavity was extremely low. Given the paucity of infant oral metagenomic datasets available for comparison, it is difficult to say whether this level of diversity is typical of the upper respiratory tracts of similarly treated, premature infants.

\(\text{T. vaginalis}\) identification, coverage, and inferred route of transmission. This is the first report of a non-vaginal microbiome in which \(\text{Trichomonas}\) and \(\text{Mycoplasma}\) sp. ‘\(\text{Mnola}\)’-like sequences co-occur. We considered that (1) \(\text{T. vaginalis}\) is rarely found in neonates\(^{25}\); (2) other \(\text{Trichomonas}\) spp. (e.g., \(\text{T. tenax}\)) preferentially or opportunistically colonize the human oral and/or respiratory mucosa; (3) a sole species, \(\text{T. vaginalis}\), overwhelmingly dominates the database of \(\text{Trichomonas}\) genomic sequence; and (4) no prior study has relied on metagenomic data to characterize \(\text{Trichomonas}\) spp. \textit{in situ}. Therefore, we carefully evaluated our species-level assignment, observed patterns of assembly and coverage, and our assumptions about the mode of transmission.

Alignment of our \(\text{Trichomonas}\)-derived internal transcribed spacer 1 (ITS1)—5.8S rRNA-ITS2 sequence against those of other \(\text{Trichomonas}\) spp.\(^{24}\), followed by phylogenetic inference, indicated that \(\text{T. vaginalis}\) was the
To provide landmarks without overwhelming the map, we only included the five largest Tenericutes branches within the phylum Firmicutes.

Table 1. Genomes reconstructed from the oral metagenome of a 3-week-old premature infant. aEstimated using reference-based mapping; bAccession number of the vaginal strain, provided here for reference; cOur own finding, which differs from that of Fettweis et al; dAs described in the Supplementary Methods; n/a, not applicable; n/d, not determined; (Tenericutes) branches within the phylum Firmicutes.

closest species. Furthermore, our Trichomonas 18S rRNA gene sequence was 100% identical over 1,508 positions to that of a vaginally-derived T. vaginalis isolate25. We concluded that 'T. vaginalis' was indeed the best possible species-level assignment.

At > 176 Mbp, the draft genome of T. vaginalis is >26× larger than the largest bacterial genome in our dataset. To provide landmarks without overwhelming the map, we only included the five largest T. vaginalis genome fragments in our ESOM (Supplementary Fig. S2). Using this approach, our T. vaginalis-derived scaffolds were readily binned. However, the resulting scaffolds were small in size (maximum of ~6.5 kbp), inconsistent in coverage (range ~2–200×), and summed to only 0.15% of the length of the reference genome. We sought to improve our estimate of overall coverage of T. vaginalis, which we accomplished using reference-based mapping.

We estimated an overall T. vaginalis genome coverage of ~0.2×, based on read-mapping across the entire T. vaginalis genome or on mapping to a set of T. vaginalis-specific single-copy genes (n = 16)26. To validate this estimate, we divided the coverage of our most highly represented (207×) scaffold, encoding the 1.3 kbp Mariner transposase (repeat R8)27 (Supplementary Table S3), by the overall coverage (0.2×), which yielded a predicted copy number of 1035. This is close to the reported copy number of 98227. Repetitive sequences compose an estimated 50% of women with over 65% of the T. vaginalis genome in the infant’s pooled saliva sample (see Methods), there were roughly 385 Mycoplasma sp. ‘Mnola’ genomes.

T. vaginalis is associated with adverse pregnancy outcomes including preterm delivery and the delivery of a low-birthweight infant, and may increase the risk of PPROM28, 29. Rates of infection vary among populations, but an estimated 50% of women with T. vaginalis are asymptomatic30–32. We strongly suspect that the infant acquired T. vaginalis (and the uncultivated mycoplasma) at delivery via passage through the birth canal, even though to our knowledge, the mother had not been screened or treated for T. vaginalis and was presumably asymptomatic or had symptoms that did not prompt reporting, screening or treatment. We can only speculate as to whether the presence of one or both organisms played a causal role in PPROM and preterm delivery in this case.
Evidence for in vivo growth of Ca. M. girerdii strain UC-B3. We closed the 629,409-bp genome of the infant oral Mycoplasma sp. ‘Mnola’ and aligned it to the closest reference, the 618,983-bp genome of Ca. M. girerdii str. VCU-M1 (Table 1; alignment dotplot shown in Supplementary Fig. S3). The alignment covered 96.67% and 98.18% of bases, respectively, and the average nucleotide identity (ANI) was 99.44%. The Mash distance, 0.0052, corresponded well with the ANI. Given the high degree of alignment coverage and overall nucleotide sequence similarity, we concluded that the genomes were distinct at the level of strains 33. We applied the name Ca. M. girerdii str. UC-B3 (University of Chicago baby #3) to the infant oral strain. The genome-wide alignment revealed a large-scale inversion spanning str. UC-B3 positions 96–244 kbp, several smaller insertions and deletions, and numerous repetitive elements (Supplementary Fig. S3).

Using the closed genome sequence, we sought evidence in support of in vivo growth of Ca. M. girerdii str. UC-B3 by examining the pattern of read coverage in relation to the origin of replication 34. The origin was designated as described in the Supplementary Methods. Coverage decreased smoothly from origin to terminus, with coverage near the origin exceeding that near the terminus by ~1.4 fold (iRep = 1.42) (Fig. 1a). Excess copy number near the origin (i.e., a coverage ratio >1) arises from active replication forks within cells. This ratio has been shown to be directly proportional to growth rate 34. Our data suggest that Ca. M. girerdii str. UC-B3 was replicating within the oral cavity of the 3-week-old premature infant. Because samples from days 15, 18 and 21 were pooled (see Methods), we cannot determine on which day(s) this occurred.

Genome closure is not required for the calculation of iRep values 18. Therefore, we can infer from our draft genomes that Pseudomonas aeruginosa and Streptococcus parasanguinis were also replicating within the oral cavity of the 3-week-old premature infant. Because samples from days 15, 18 and 21 were pooled (see Methods), we cannot determine on which day(s) this occurred.

Finally, coverage patterns can highlight genomic rearrangements 35. The large-scale inversion revealed by our genome-wide alignment of str. UC-B3 to str. VCU-M1 (Supplementary Fig. S3) was also apparent as an inversion.

Figure 1. Pattern of sequencing read coverage suggests in vivo growth of Ca. M. girerdii strain UC-B3 and large-scale genome inversion with respect to strain VCU-M1. For each window, the average number of reads aligned to each base is plotted against the genome position corresponding to the first base. Windows are non-overlapping and sized at 10 kbp (red squares) or 1 kbp (black circles). For both panels, the aligned reads are those generated for the present study (i.e., the infant oral metagenomic reads). As a proxy for the origin of replication, the first base of each genome was set to the first base of dnaA. (a) Read alignment to the strain UC-B3 genome (this study) reveals a peak-to-trough ratio of ~1.4 (iRep = 1.42), suggesting active replication forks at the time of sampling. (b) Read alignment to the strain VCU-M1 genome 16 confirms a large-scale inversion initially revealed by the whole genome alignment (Supplementary Fig. S3). Arrows indicate putative inversion break points.
Ca. M. girerdii encodes PFOR and is predicted to be sensitive to metronidazole. Ca. M. girerdii strains UC-B3 and VCU-M1 are predicted to be metabolically alike (see the Supplementary Results and Discussion for an overview of Ca. M. girerdii’s metabolic potential). Their predicted metabolism appears to represent a variation on the theme of mycoplasma metabolic streamlining that is particularly novel at the pyruvate locus.

Ca. M. girerdii is unusual among glycolytic (sugar-fermenting) mycoplasmas in lacking pyruvate dehydrogenase (PDH), the enzyme responsible for oxidatively decarboxylating pyruvate to acetyl-CoA. Instead of using PDH, Ca. M. girerdii likely relies on pyruvate formate-lyase (PFL) and/or pyruvate-ferredoxin oxidoreductase (PFOR), both of which were present in the two genomes (for str. UC-B3, see Supplementary Table S4). These enzymes are characteristic of anaerobically growing bacteria and have not been detected in other mycoplasmas (Supplementary Table S5).

PFL produces acetyl-CoA and formate from pyruvate and coenzyme-A. PFL requires a ‘pyruvate formate-lyase activase’ and co-occurs with a replacement part, an ‘autonomous glycol radical cofactor’, both of which were present in the two genomes. The radical cofactor is thought to quickly restore activity in stress-damaged PFL. Among Mollicutes, PFL has been detected in a few Spiroplasma spp. and in Acholeplasma spp. (Supplementary Table S5), the latter occupying the taxonomically ambiguous base of the class Mollicutes, which branches within the phylum Firmicutes. PFOR produces acetyl-CoA, CO₂, H⁺ and reduced ferredoxin from pyruvate, coenzyme-A and oxidized ferredoxin. PFOR is typically coupled to a hydrogenase that is responsible for re-oxidizing the reduced ferredoxin. We identified genes encoding ferredoxins in the Ca. M. girerdii genomes, but were unable to identify genes encoding hydrogenases. Thus, alternatively, PFOR may operate in reverse and couple with PFL to reduce CO₂ to formate, which may then feed into reductive one-carbon metabolism. PFOR is found in Acholeplasma brassicaceae, Acholeplasma morum, Ca. Izimaplasma spp., and on several contigs assembled from human gut metagenomes (so-called ‘CAG’ entities) (Supplementary Table S5)—all of which are predicted to be ‘basal’ Mollicutes in the manner noted above. Ca. M. girerdii’s PFOR is highly novel and its presence is not likely to be explained by recent lateral transfer. Its 1192-aa sequence is in the range of 53–55% identical to PFORs from a diverse set of Firmicutes species.

Finally, if PFOR were active, we would expect Ca. M. girerdii to be sensitive to metronidazole, the antibiotic commonly used to treat infections caused by anaerobes including T. vaginalis, but not used in this infant. PFOR reduces metronidazole to its active form. Because Ca. M. girerdii str. UC-B3 persisted and possibly bloomed in the face of antibiotic exposure, we also looked for putative resistance mechanisms (see Supplementary Methods). On balance, it seemed likely that intrinsic resistance mechanisms (e.g., lack of a cell wall), which could be inferred for all mycoplasmas, played a larger role than acquired ones in str. UC-B3 (see Supplementary Results and Discussion for details).

Genomic features distinguishing Ca. M. girerdii str. UC-B3 from str. VCU-M1. The Ca. M. girerdii genomes are AT-rich, small, and use genetic code four (UGA codes for Trp instead of Stop) (Table 1). These traits are common among the Mycoplasmataceae. The genomes also reflect the phylogenetic novelty first noted in their 16S rRNA genes (see Introduction). The genomes of the most closely related organisms, e.g., M. hominis, M. penetrans (of the M. pneumoniae group sensu Davis et al. (Supplementary Fig. S4), are not broadly alignable to those of Ca. M. girerdii at the level of nucleotide sequences.

Only 63% of str. UC-B3 proteins had a valid hit in a non-VCU-M1 genome; among these hits, the median level of sequence identity was 50% (Fig. 2). Most non-VCU-M1 hits (~70%) were to other Mollicutes, with the rest mainly to Firmicutes. These findings confirm the designation of a new species made by Fettweis et al., and are notable because uncultivated organisms with a high degree of phylogenetic novelty are rarely encountered in human body habitats (although see refs 40, 41).

Most str. UC-B3 protein-coding genes (n = 529 of 574) had an easily-identified ortholog in str. VCU-M1, defined here as a syntetic, high-identity (≥97% nucleotide identity), reciprocal best hit (see Supplementary Results and Discussion for details). We defined those without easily-identified orthologs (n = 45) as UC-B3’s ‘variable set’ (Supplementary Table S4). Most ‘variable set’ genes (n = 40) could be grouped into one of four distinct classes, representing a first glimpse into the major themes of strain-level differentiation in Ca. M. girerdii. These four classes, and the features composing them (Fig. 3a,b), are described in detail in the Supplementary Results and Discussion.

The ‘variable set’ included genes encoding a diverse array of BspA-like surface proteins (named for B. forsythus surface protein A), of which str. UC-B3 contains at least 28 distinct copies, as well as those encoding restriction-modification (R-M) systems (str. UC-B3 contains 18 genes involved in R-M systems) (Fig. 3a). Notably, T. vaginalis also encodes a diverse array of BspA-like surface proteins. BspA was identified in Tannerella (formerly Bacteroides) forsythus, contains a structural motif known as a Treponema pallidum Leucine-rich repeat (TplLR), and is thought to be involved in host cell attachment, co-aggregation, and the stimulation of immune responses. Genomic regions exclusive to str. UC-B3, including an 8.6-kbp genomic island carrying cytosine methyltransferases (Fig. 3b), tend to reflect recent lateral exchange with Mycoplasma hominis (see Supplementary Results and Discussion). Given Ca. M. girerdii’s highly reduced genome size and metabolic repertoire, expansion and elaboration within these classes of genes almost certainly highlights ecological strategies that are important to the organism, which may include attachment, immune evasion, and defense against foreign DNA.
Length variation in DNA tandem repeats within population UC-B3. Natural populations are rarely isogenic. During the sampling period, str. UC-B3 was likely replicating in vivo (Fig. 1a). Therefore we investigated within-population variation in str. UC-B3, as a window onto recent diversification. We began with routine variant calling—mapping reads to the UC-B3 genome and identifying polymorphic sites (mainly SNPs and short indels). We found relatively few such sites: ~50, depending on the stringency of read mapping and variant calling. Salient among them were indels of dinucleotides at dinucleotide tandem repeats (DTRs). We also noticed that many of UC-B3’s polymorphic DTRs appeared within or near genes belonging to classes likely to exhibit phase variation (e.g., BspA-like surface proteins; HsdS specificity proteins of type I R-M systems). We therefore undertook a systematic analysis in which we (1) located all simple sequence repeats in the str. UC-B3 genome, (2) identified those exhibiting length variation, (3) determined the frequencies of length variants, and (4) predicted possible consequences for associated gene expression (i.e., phase variation) as described in the Supplementary Methods.

In population UC-B3, all observed indels of iterations of simple sequence repeats appeared in repeat tracts ≥5 iterations in length, most of which were DTRs. The UC-B3 genome contains 20 DTRs ≥5 and ranging up to 12 iterations in length (Fig. 3a and Supplementary Table S6). Eight different motifs are represented (all but ‘tg’, ‘gt’, ‘gc’ and ‘cg’) (Supplementary Table S6). Of the 20 UC-B3 DTRs, 19 have a clear homolog in str. VCU-M1, and 12 of these differ in the number of iterations (Supplementary Table S7). Of the 20 UC-B3 DTRs, 13 exhibited differences in the number of iterations within the UC-B3 population (Supplementary Table S7). In all cases, as expected, the DTR in the UC-B3 genome (i.e., the wild-type) was most frequent. We also noted that all length variants featured indels of one or more dinucleotide iterations, and not of odd numbers of nucleotides, or of differing bases, as might occur by chance or error (see Supplementary Results and Discussion). In addition to the DTRs, we identified two, length-variable homopolymeric tracts (Supplementary Table S7).

The frequencies of length variants at repeat sites (average 0.082; range 0.017–0.446) (Supplementary Table S7) far exceeded those of dinucleotide indels at non-repeat sites (see Supplementary Methods) and those expected for indel sequencing errors (see Supplementary Results and Discussion). Affected genes included eight BspA-like surface proteins; non-solitary HsdS proteins associated with the type I R-M loci; and methyltransferases associated with type III R-M loci (see Supplementary Results and Discussion). With respect to phase variation, for a number of affected genes, the wild-type is predicted to be ‘off’ (i.e., not expressed or severely truncated) owing to frameshift mutations that induce premature stop codons (i.e., early termination), while one or more low-frequency variants are predicted to be ‘on’ (Supplementary Table S7). An example of this is shown in Fig. 3, in which only the +1 DTR variant produces an intact BspA-like protein (Fig. 3c,d). For the two HsdS loci, each containing two length-variable DTRs, the consequences for phase variation are more difficult to predict, as outlined in Supplementary Table S7. Both type III R-M methyltransferases (see Supplementary Results and Discussion) are phase ‘off’ in wild-type UC-B3, with the M. hominis-like mod controlled by a homopolymer tract G(8), rather than by a DTR (Supplementary Table S7).
Figure 3. Genomic features distinguishing Ca. M. girerdii strain UC-B3 from strain VCU-M1 include those with evidence of phase variation within population UC-B3. (a) Circular genome map of Ca. M. girerdii strain UC-B3. Inner track shows RNA genes (black; staggered for clarity). Center and outer tracks show open reading frames (ORFs; color graded on percent amino acid identity to reciprocal best hit in VCU-M1). ORFs in gray have a close nucleotide hit in strain VCU-M1 but were not predicted there by the authors. Purple lines, located between center and outer tracks, represent dinucleotide tandem repeats (DTRs) of five or more iterations (Supplementary Table S6). ORFs from variable classes were manually placed in the outer track. Those with asterisks display evidence of phase variation (Supplementary Table S7); those without labels (n = 5) encode fructose-bisphosphate aldolase; and those composing restriction modification systems are labeled clockwise as follows: R, restriction endonuclease; M, modification methylase; S, specificity subunit. The subjects of panels b–d are labeled in bold. SSR, simple sequence repeat. (b) Linear map detailing genomic island in UC-B3, which is absent from VCU-M1. ORFs drawn as arrows; those without labels encode hypothetical proteins. (c) Evidence of phase variation for selected feature no. 11 in Supplementary Table S7. Plotted are length variant frequencies for DTR ‘CA’ associated with BspA-like protein labeled in bold in panel a. Homologous ORF in VCU-M1 contains CA(10) (Supplementary Table S7). TRUNC, frameshift-induced premature stop codon resulting in truncated BspA-like protein; INTACT, likely produces intact BspA-like protein shown in panel (d). The intact protein was modeled against the highest scoring template (pneumococcal vaccine antigen PcpA) using Phyre2. The subjects of panels b–d are labeled in bold. SSR, simple sequence repeat. (b) Linear map detailing genomic island in UC-B3, which is absent from VCU-M1. ORFs drawn as arrows; those without labels encode hypothetical proteins.
Phase variation randomly and reversibly generates phenotypic diversity within clonal populations via heritable genetic or epigenetic means, typically manifesting in ‘on/off’ patterns of gene expression (i.e., switching). This allows small populations of bacteria to rapidly explore new niches, or to evade the host immune system when colonizing a new host individual. Phase variation is well-documented among the types of genes in which it was observed here. Phase variation of components of R-M systems has been proposed to either facilitate the acquisition of potentially beneficial foreign DNA or to regulate expression of suites of genes. Although sequence read data have been used to assess the frequency of trc tract variants in a laboratory-grown population of Helicobacter canadensis, to our knowledge, ours is the first study to do so using metagenomic read data and a de novo assembled genome from a natural population. Although our data are short reads and therefore the variants remain unlinked, we suggest that the Ca. M. girerdii str. UC-B3 population comprised multiple subpopulations expressing unique arrays of e.g., BspA-like surface proteins.

Conclusions

Vast, unexplored regions of the tree of life have been uncovered by sequence-based surveys of the microbial world. This phylogenetic terra incognita encompasses countless diverse lineages, many of which have been repeatedly detected in the environment but otherwise remain mysterious. This so-called ‘microbial dark matter’ is encountered most places microbes naturally reside, including the human body. The current study stands as an example of how genome-resolved metagenomics provide a powerful and reproducible means of shedding light on the ecology and evolution of such lineages—here, an emerging uncultivated mycoplasma, Ca. M. girerdii, associated with trichomoniasis in women.

By uncovering all forms of variation, these approaches offer unparalleled access to the dynamics of the human microbiome. We found that the oral cavity of a vaginally-delivered, three-week-old premature infant harbored both Ca. M. girerdii and T. vaginalis. Vertical transmission of these organisms from an asymptomatic mother is surmised; however, a weakness of our study is its lack of clinical data from the mother. For example, we do not have direct evidence that she carried T. vaginalis. Nevertheless, chance findings such as these remind us of the need for careful screening prior to the deliberate transfer of maternal vaginal microbes to Cesarean-delivered neonates. Because our work suggests that the association of Ca. M. girerdii with T. vaginalis could extend to the premature infant upper respiratory tract, follow-up studies would be warranted in which the microbiomes of pregnant women and their newborns were monitored over time in patient populations at high risk of trichomoniasis and preterm birth, using methods capable of detecting T. vaginalis.

We found that the assembled genome of the infant oral strain, Ca. M. girerdii str. UC-B3, exhibited excess copy number near the origin of replication. From this we inferred a positive growth rate—possibly a relatively fast growth rate, especially for a mycoplasma—at some point during the infant’s third week of life. We also found that this strain. UC-B3 demonstrated high-frequency variation in the lengths of DNA tandem repeats, likely capable of inducing the ‘on/off’ switching of gene expression (i.e., phase variation). Given the types of genes affected, we infer that this strain. UC-B3 exhibited cell-to-cell variation in its expressed repertoires of BspA-like surface proteins and restriction-modification systems. Well-known among mycoplasmas and other, mainly pathogenic bacteria, phase variation is a strategy that facilitates colonization, evasion of host defenses, and protection of cells from phage and restriction-modification systems. This allows small populations of bacteria to rapidly explore new niches, or to evade the host immune system when colonizing a new host individual. Phase variation is well-documented among the types of genes in which it was observed here. Phase variation of components of R-M systems has been proposed to either facilitate the acquisition of potentially beneficial foreign DNA or to regulate expression of suites of genes. Although sequence read data have been used to assess the frequency of trc tract variants in a laboratory-grown population of Helicobacter canadensis, to our knowledge, ours is the first study to do so using metagenomic read data and a de novo assembled genome from a natural population. Although our data are short reads and therefore the variants remain unlinked, we suggest that the Ca. M. girerdii str. UC-B3 population comprised multiple subpopulations expressing unique arrays of e.g., BspA-like surface proteins.

References

1. Razin, S., Yogev, D. & Naot, Y. Molecular biology and pathogenicity of mycoplasmas. Microbiol. Mol. Biol. Rev. 62, 1094–1156 (1998).
2. van der Woude, M. W. Phase variation: how to create and coordinate population diversity. Curr. Opin. Microbiol. 14, 205–211 (2011).
3. May, M., Balish, M. F. & Blanchard, A. The order Mycoplasmatales in The Prokaryotes - Firmicutes and Tenericutes (eds Rosenberg, E. et al.) 515–550 (Springer-Verlag, 2014).
4. Waites, K. B. & Taylor-Robinson, D. Mycoplasma and Ureaplasma in Manual of clinical microbiology (eds Jorgensen, J. H. et al.) 1088–1105 (ASM Press, 2015).
5. Kantor, R. S. et al. Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. mBio. 4, e00708–13 (2013).
6. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. The ISME Journal. 8, 1553–1565 (2014).
7. Hug, L. A. et al. A new view of the tree of life. Nat Microbiol. 1, 16048 (2016).
8. Bove, J. M. Molecular characters of mollicutes. Clin. Infect. Dis. 17(Suppl 1), S10–31 (1993).
9. Razin, S. Physiology of mycoplasmas. Adv. Microb. Physiol. 10, 1–80 (1973).
10. Goldemberg, R. L., Culhane, J. F., Iams, J. D. & Romero, R. Epidemiology and causes of preterm birth. Lancet. 371, 75–84 (2008).
11. Waites, D. H., Krohn, M. A., Hillier, S. L. & Eschenbach, D. A. The association of occult amniotic fluid infection with gestational age and neonatal outcome among women in preterm labor. Obstet. Gynecol. 97, 351–357 (1996).
12. Waites, K. B., Katz, B. & Schelonka, R. L. Mycoplasmas and ureaplasmas as neonatal pathogens. Clin. Microbiol. Rev. 18, 757–789 (2005).
13. Costello, E. K., Carlisle, E. M., Bik, E. M., Morowitz, M. J. & Relman, D. A. Microbiome assembly across multiple body sites in low-birthweight infants. mBio. 4, e00782–13 (2013).
14. Brotman, R. M. et al. Association between Trichomonas vaginalis and vaginal bacterial community composition among reproductive-age women. Sex. Transm. Dis. 39, 807–812 (2012).
15. Martin, D. H. et al. Unique vaginal microbiota that includes an unknown Mycoplasma-like organism is associated with Trichomonas vaginalis infection. J. Infect. Dis. 207, 1922–1931 (2013).
16. Fettweis, J. M. et al. An emerging mycoplasma associated with trichomoniasis, vaginal infection and disease. PLoS ONE 9, e10943 (2014).
17. Edwards, T., Burke, P., Smalley, H. & Hobbs, G. Trichomonas vaginalis: Clinical relevance, pathogenicity and diagnosis. Crit. Rev. Microbiol. 42, 406–417 (2016).
18. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. Nat. Biotechnol. 34, 1256–1263 (2016).
19. Sharon, I. & Banfield, J. F. Genomes from metagenomics. Science 342, 1057–1058 (2013).
20. Smith, J. M., Wang, M., Zhang, W. & Xie, S. Trichomonas vaginalis infection in a premature newborn. J. Perinatol. 22, 502–503 (2002).
21. Carter, J. E. & Whithaus, K. C. Neonatal respiratory tract involvement by Trichomonas vaginalis: a case report and review of the literature. Am. J. Trop. Med. Hyg. 78, 17–19 (2008).
22. Bruns, M. J., van Straaten, I. L. M. & Ruijs, G. J. M. Respiratory disease and Trichomonas vaginalis in premature newborn twins. Pediatr. Infect. Dis. J. 32, 1029–1030 (2013).
23. Hirt, R. P. & Sherrard, J. Trichomonas vaginalis origins, molecular pathobiology and clinical considerations. Curr. Opin. Infect. Dis. 28, 72–79 (2015).
24. Duboucher, C. et al. Molecular identification of Tritrichomonas foetus-like organisms as coinfecting agents of human Pneumocystis pneumonia. J. Clin. Microbiol. 44, 1165–1168 (2006).
25. Dumasay, K. G. B. & Rivera, W. L. First report of Trichomonas tenax infections in the Philippines. Parasitol. Int. 63, 400–402 (2014).
26. Cornelius, D. C. et al. Genetic characterization of Trichomonas vaginalis isolates by use of multilocus sequence typing. J. Clin. Microbiol. 50, 3293–3300 (2012).
27. Carlton, J. M. et al. Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. Science 315, 207–212 (2007).
28. Minkoff, H. et al. Risk factors for prematurity and premature rupture of membranes: a prospective study of the vaginal flora in pregnancy. Am. J. Obstet. Gynecol. 150, 965–972 (1984).
29. Choi, M. et al. Trichomonas vaginalis associated with low birth weight and preterm delivery. Sex. Transm. Dis. 24, 353–360 (1997).
30. Choi, M. M. D. et al. Demographic and behavioral predictors of Trichomonas vaginalis infection among pregnant women. Obstet. Gynecol. 78, 1087–1092 (1991).
31. Schwicker, J. R. & Burgess, D. Trichomoniasis. Clin. Microbiol. Rev. 17, 794–803 (2004).
32. Schwebke, J. R. & Burgess, D. Trichomoniasis. Clin. Microbiol. Rev. 17, 794–803 (2004).
33. Skvagard, O., Bak, M., Løbner-Olesen, A. & Tonnerup, N. Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. Genome Research 21, 1388–1393 (2011).
34. White, D. The physiology and biochemistry of prokaryotes. (Oxford University Press, 1995).
35. Davis, J. J., Xia, E., Overbeek, R. A. & Olsen, G. J. Genomes of the class Erysipellovibrio clarify the firmicute origin of the class. Mol. Ecol. 15, 2727–2741 (2013).
36. Xiong, W. et al. CO2-fixing one-carbon metabolism in a cellulose-degrading bacterium Clostridium thermocellum. Proc. Natl. Acad. Sci. USA 102, 2567–2572 (2005).
37. Korem, T. et al. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science 349, 1101–1106 (2015).
38. Skovgaard, O., Bak, M., Løbner-Olesen, A. & Tonnerup, N. Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. Genome Research 21, 1388–1393 (2011).
39. Dong, X. et al. Simple sequence repeats in Erysipelotrichia clarifying the firmicute origin of the class. Int. J. Syst. Evol. Microbiol. 63, 2727–2741 (2013).
40. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. USA 102, 2567–2572 (2005).
41. Korem, T. et al. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science 349, 1101–1106 (2015).
42. Skovgaard, O., Bak, M., Løbner-Olesen, A. & Tonnerup, N. Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. Genome Research 21, 1388–1393 (2011).
43. Dwyer, J. J. & Willens, J. H. Clinical and molecular recognition of the phylum Erysipelotrichia. Clin. Microbiol. Rev. 20, 353–360 (2007).
44. Inagaki, S., Onishi, S., Kuramitsu, H. K. & Sharma, A. Multiple functions of the leucine-rich repeat protein LrrA of Bacteroides thetaiotaomicron. Mol. Microbiol. 33, 5727–5737 (1999).
45. Schirmr, M., D'Amore, R., Iyaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics 17, 125 (2016).
46. Henderson, I. R., Owen, P. & Nataro, J. P. Molecular switches - the ON and OFF of bacterial phase variation. Mol. Microbiol. 33, 919–932 (1999).
47. van der Woude, P. M. W. & Bäumler, A. J. Phase and antigenic variation in bacteria. Clin. Microbiol. Rev. 17, 581–611 (2004).
48. Skrkihanta, Y. N., Maguire, T. L., Stacey, K. J., Grimmond, S. M. & Jennings, M. P. The phasevarion: a genetic system controlling coordinated, random switching of expression of multiple genes. Proc. Natl. Acad. Sci. USA 102, 5547–5551 (2005).
49. Aras, R. A., Takata, T., Ando, T., van der Ende, A. & Blaser, M. J. Regulation of the HpyII restriction-modification system of Helicobacter pylori by gene deletion and horizontal reconstitution. Mol. Microbiol. 42, 369–382 (2001).
50. Hoskisson, P. A. & Smith, M. C. M. Hypervariation and phase variation in the bacteriophage “resistome.” Curr. Opin. Microbiol. 10, 396–400 (2007).
51. Skrkihanta, Y. N., Fox, K. L. & Jennings, M. P. The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. Nat. Rev. Micro. 8, 196–206 (2010).
52. Snyder, L. A. et al. Simple sequence repeats in Helicobacter canadensis and their role in phase variable expression and C-terminal sequence switching. BMC Genomics 11, 87 (2010).
53. Dominguez-Bello, M. G. et al. Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. Nat. Med. 22, 250–253 (2016).
54. Dwyer, K. & Voelker, L. L. Molecular biology of mycoplasmas. Annu. Rev. Microbiol. 50, 25–57 (1996).
55. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 10, 845–858 (2015).
Acknowledgements
This work was supported by NIH Grants R01 AI092531 (J.F.B., M.J.M., D.A.R.) and K99 HD074743 (E.K.C.), a Walter V. and Idun Berry Postdoctoral Fellowship (E.K.C.), the March of Dimes Prematurity Research Center at Stanford University School of Medicine, and the Thomas C. and Joan M. Merigan Endowment at Stanford University (D.A.R.).

Author Contributions
Conceived and designed study (E.K.C., M.J.M., J.F.B., D.A.R.); Oversaw and coordinated clinical aspects of study (M.J.M., E.M.C.); Acquired, analyzed and interpreted sequence data (E.K.C.); Contributed analysis scripts, analyzed and interpreted data (C.L.S.); Oversaw key aspects of data analysis (J.F.B.); Drafted manuscript (E.K.C.); Revised manuscript critically for intellectual content (all authors); Read and approved final manuscript (all authors).

Additional Information
Supplementary information accompanies this paper at doi:10.1038/s41598-017-03821-7

Competing Interests: The authors declare that they have no competing interests.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017