

## **Stable isotope informed genome-resolved metagenomics reveals that Saccharibacteria utilize microbially processed plant derived carbon**

Evan P. Starr (starr.evan@berkeley.edu)<sup>1</sup>, Shengjing Shi<sup>2</sup>, Steven J. Blazewicz<sup>3</sup>, Alexander J. Probst<sup>4</sup>, Donald J. Herman<sup>5</sup>, Mary A. Firestone<sup>5</sup> and Jillian F. Banfield (jbanfield@berkeley.edu)\*<sup>5,6,7</sup>

<sup>1</sup>Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA

<sup>2</sup>Lincoln Science Centre, AgResearch Ltd, Christchurch 8140, New Zealand

<sup>3</sup>Nuclear and Chemical Sciences Division, Lawrence Livermore National Laboratory, Livermore, USA

<sup>4</sup>Biofilm Center, University of Duisburg-Essen, Essen 45141, Germany

<sup>5</sup>Department of Environmental Science, Policy, and Management, University of California, Berkeley, California 94720, USA

<sup>6</sup>Department of Earth and Planetary Science, University of California, Berkeley, California 94709, USA;

<sup>7</sup>Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

## **Abstract:**

## **Background:**

The transformation of plant photosynthate into soil organic carbon and its recycling to CO<sub>2</sub> by soil microorganisms is one of the central components of the terrestrial carbon cycle. There are currently large knowledge gaps related to which soil-associated microorganisms take up plant carbon in the rhizosphere and the fate of that carbon.

## **Results:**

We conducted an experiment in which common wild oats (*Avena fatua*) were grown in a <sup>13</sup>CO<sub>2</sub> atmosphere and the rhizosphere and non-rhizosphere soil was sampled for genomic analyses. Density gradient centrifugation of DNA extracted from soil samples enabled distinction of microbes that did and did not incorporate the <sup>13</sup>C into their DNA. A 1.45 Mbp genome of a Saccharibacteria (TM7) was identified and, despite the microbial complexity of rhizosphere soil, curated to completion. The genome lacks many biosynthetic pathways, including genes required to synthesize DNA *de novo*. Rather, it requires externally-derived nucleotides for DNA and RNA synthesis. Given this, we conclude that rhizosphere-associated Saccharibacteria recycle DNA from bacteria that live off plant exudates and/or phage that acquired <sup>13</sup>C because they preyed upon these bacteria and/or directly from the labelled plant DNA. Isotopic labeling indicates that the population was replicating during the six-week period of plant growth. Interestingly, the genome is ~30% larger than other complete Saccharibacteria genomes from non-soil environments, largely due to more genes for complex carbon utilization and amino acid metabolism. Given the ability to degrade cellulose, hemicellulose, pectin, starch and 1,3-β-glucan, we predict that this Saccharibacteria generates energy by fermentation of soil necromass and plant root exudates to acetate or lactate. The genome encodes a linear electron transport chain featuring a terminal oxidase, suggesting that this Saccharibacteria may respire aerobically. The genome encodes a hydrolase that could breakdown salicylic acid, a plant defense signaling molecule, and genes to make a variety of isoprenoids, including the plant hormone zeatin.

## **Conclusions:**

Rhizosphere Saccharibacteria likely depend on other bacteria for basic cellular building blocks. We propose that isotopically labeled CO<sub>2</sub> is incorporated into plant-derived carbon and then into the DNA of rhizosphere organisms capable of nucleotide synthesis, and the nucleotides are recycled into Saccharibacterial genomes.

**Keywords:** Stable isotope probing, complete genome, Saccharibacteria, metagenomics, rhizosphere

## Background

The Candidate Phyla Radiation (CPR) comprises a large fraction of the bacterial domain [1, 2]. An episympiotic lifestyle has been demonstrated experimentally for only one group within the CPR, a member of the phyla Saccharibacteria, formerly TM7 [3]. Bacteria from this phylum are detected across a diversity of environments, suggesting that they may vary in their metabolic capacities. Originally, discovered in a peat bog [4], these bacteria have been described from human mouths [3], aquifers [5], sediment [6], and soil [7]. 16S rRNA gene surveys indicate that Saccharibacteria also occur in the rhizosphere, but little is known about their metabolism and how they differ from related organisms growing in other environments [8, 9].

There are four previously reported, closed and circularized genomes for Saccharibacteria (from a sludge bioreactor [10], human mouth [3] and two from sediment [11]) and some partial genomes. Based on genomic analyses, it was previously predicted that Saccharibacteria are anaerobic fermenters, despite several genomes encoding a cytochrome-*o*-ubiquinol oxidase (Complex IV). Complex IV is the final step in the aerobic electron transport chain, however, it has been suggested that Saccharibacteria use it for O<sub>2</sub> scavenging, given the lack of other electron transport chain components. However, it has been documented that Saccharibacteria are able to grow in the presence of oxygen [12]. Here, we found that a Saccharibacteria population in rhizosphere soil incorporated <sup>13</sup>C into its genomic DNA, indicating that this population was active and dividing. We report the complete genome for this organism and show how analysis of its metabolism sheds light on the pathway by which the label was incorporated. We compare the genome to those of other complete Saccharibacteria to address the question of how this bacterium is adapted to live in soil and specifically in the rhizosphere.

## Results

DNA was extracted from soil removed from the roots of an *Avena fatua* plant grown for six weeks in a <sup>13</sup>CO<sub>2</sub> atmosphere and separated into three fractions based on density. The isotopically light, middle and heavy fractions were sequenced and subjected to genome-resolved

metagenomic analyses. From the rhizosphere middle fraction, accounting for 210 Mbp of contiguous sequence, a large scaffold was assembled *de novo* and could be circularized. Local assembly errors were identified and corrected and three scaffolding gaps were filled by manual curation. The complete, closed genome is 1.45 Mb in length with a GC content of 49.95%. The genome was most abundant in the middle fraction at 15x coverage, but was also present in the isotopically heavy and light fractions at ~3x coverage. The genome was undetectable in the background soil (**Figure 1**). DNA density is a function of both the extent of isotopic enrichment and the GC content, low GC is less dense than higher GC DNA [13]. This genome has a lower GC content than the rest of the assembly (average GC content of scaffolds larger than 1000 bp is 66%), which indicates  $^{13}\text{C}$  was incorporated, as nucleotides, into the DNA. The middle fraction where the genome was mainly detected had a density of 1.737-1.747 g/ml. Given that natural abundance DNA with 49.95% GC content would have a density of ~1.71 g/ml [14], we predict that the genome is at least 50% labeled.

The genome has 1531 protein coding sequences (**Table S1**) and a full complement of tRNAs (46 in total). The 5S rRNA, 23S rRNA and 16S rRNA genes are in a single locus that also includes Ala and Ile tRNA genes. Based on the sequence of the 16S rRNA gene, the genome was assigned to be a member of the Saccharibacteria phyla. The closest 16S rRNA gene sequence in NCBI is from an unpublished study entitled “Effects of pinewood nematode (*Bursaphelenchus xylophilus*) infected-*Pinus massoniana* on soil bacterial communities”. The cloned 16S rRNA gene shares 94% identity with the sequence reported here (**Figure 2**). The most closely related genomically described organism is *Saccharimonas aalborgensis* from activated sludge with 84% identity across the full-length 16S rRNA gene. We propose the name *Candidatus* “Teamsevenus rhizospherense” for the organism described here, given the derivation of the genome from the rhizosphere. As this is the first genome from a distinct Class, the representatives of which mostly are found in soil, we propose the complete taxonomic descriptor: Phylum: *Candidatus* Saccharibacteria, Class: *Candidatus* Soliteamseven, Order: *Candidatus* Teamsevenales, Family: *Candidatus* Teamsevenaceae, Genus: *Candidatus* Teamsevenus, Species: *Candidatus* rhizospherense.

We calculated the GC skew and cumulative GC skew across the closed *T. rhizospherense* genome and found the symmetrical pattern typical for bacteria, with a single peak and trough indicative of the terminus and origin of replication (**Figure 3**). This result both validates the

accuracy of the circularized genome and confirms that Saccharibacteria use the typical bacterial pattern of bi-directional replication from a single origin to the terminus (as do some Peregrinibacteria, another group of CPR bacteria[15]). The start of the genome was adjusted to correspond to the predicted origin, which lies between the DNA polymerase III subunit beta and the chromosomal replication initiator protein. It has a full set of ribosomal proteins, except for L30, which is uniformly absent in CPR bacteria [16].

### **Biosynthetic pathways**

The *T. rhizosphaerense* genome encodes a number of enzymes for the conversion of nucleotides to NMP, NDP and NTP and formation of RNA. In addition, we identified genes to phosphorylate G, C and U. However, the organism seems unable to synthesize 5-phospho-alpha-D-ribose-1-diphosphate (PRPP). Further, it lacks essentially all of the steps for synthesis of nucleotide bases and the pathways that would convert PRPP to inosine monophosphate or uridine monophosphate. *S. rhizophera* may have a novel nucleotide biosynthesis pathway, but this is unlikely as nucleotide biosynthesis remains highly conserved across domains [17]. Thus, *S. rhizophera* likely did not *de novo* synthesize its nucleotides but rather must have acquired nucleotides from an external source. The genome encodes several nucleases, an external micrococcal nuclease, and an oligoribonuclease for the breakdown of externally-derived DNA and RNA (**Figure 4**). The mechanism for DNA and RNA import is unknown, as we did not identify nucleotide transporters. However, there are a number of transporters with unidentified specificity that could be involved in DNA or nucleotide uptake or the type IV pili could be responsible for this function. A large portion of the genome is dedicated to DNA and RNA repair mechanisms. There are twenty 8-oxo-dGTP-diphosphatase genes that prevent the incorporation of oxidized nucleotides. This may be an adaptation for life in a mostly aerobic environment, a seemingly unusual condition for members of the Saccharibacteria phylum, as all other complete genomes were found in mostly anaerobic environments and encode few genes with this function.

The *T. rhizosphaerense* genome does not encode the ability to synthesize any amino acids *de novo* from central metabolites. However, it encodes genes to generate amino acids from precursors (e.g., valine and leucine from 2-oxoisovalerate, isoleucine from 2-methyl-2-oxopentanoate, and histidine from L-histidinol phosphate) and to interconvert some amino acids (serine and glycine). We identified genes for proteases that could breakdown externally-derived

proteins. No amino acid specific transporters were annotated, but several transporters of unknown function could import the amino acids. There is little evidence to suggest that externally derived amino acids are broken down for use in the TCA (the only TCA cycle gene identified is a fumarate reductase subunit) or other cycles.

*T. rhizosphaerense* appears unable to synthesize fatty acids, yet it encodes three copies of the 3-oxoacyl-(acyl-carrier protein) reductase and five copies of acyl-CoA thioesterase I, indicating the capacity for fatty acid conversion. We predict that *T. rhizosphaerense* takes up phosphatidylcholine (possibly derived from eukaryotes), 1,2 diacyl sn-glycerol-3P (from bacteria), or phosphatidylethanolamine (the main bacterial phospholipid) and may be able to interconvert the compounds using a gene annotated as phospholipase D. We identified a putative phosphatidate cytidyltransferase that could add a head group to 1,2 diacyl sn-glycerol-3P forming CDP-diacylglycerol. This may be able to be converted into three products: phosphatidylglycerophosphate (via CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase), or to cardiolipin (via cardiolipin synthase) or to phosphatidyl-1D-myo-inositol (via CDP-diacylglycerol-inositol 3-phosphatidyltransferase).

Interestingly, phosphatidyl-1D-myo-inositol is the precursor for generation of phosphatidylinositol mannosides, glycolipids that are decorated by a chain of mannose molecules and that are found in the cell walls of *Mycobacterium* [18]. There are several genes for the first step in phosphatidylinositol mannosides biosynthesis, which involves modification of phosphatidyl-1D-myo-inositol by addition of mannose. These include phosphatidylinositol alpha-mannosyltransferase (three copies) and a single copy of alpha-1,6-mannosyltransferase. Subsequently, a polyprenol-P-mannose  $\alpha$ -1,2-mannosyltransferase (CAZy glycosyltransferase family 87) adds another mannose group. Other mannose additions may involve the three copies of dolichol-phosphate mannosyltransferase, which transfer mannose from GDP-mannose to dolichol phosphate a mannose carrier involved in glycosylation. Thus, although *T. rhizosphaerense* appears to be unable to synthesize fatty acids, it appears that it may be able to generate membrane lipids, including phosphatidylinositol mannosides, if provided 1,2 diacyl sn-glycerol-3P.

The genome encodes a near-complete pathway for peptidoglycan biosynthesis. Given this, the one missing enzyme (murC, a ligase) may be too distant for identification and we predict the capacity for peptidoglycan synthesis.

This Saccharibacterial species appears to be reliant on external sources of cofactors. For instance, the genome encodes no genes in the nicotinate synthesis pathway, only one gene in the synthesis pathways for thiamine and riboflavin and two genes for the synthesis of biotin. However, there are a number of genes present for the biosynthesis of folate and the ability to activate, but not synthesize, B12 into its coenzyme form.

### **Central metabolism and energy generation**

Interestingly, the *T. rhizosphaerense* genome encodes a simple, two-subunit cellulosome that may be used to attach to and degrade plant or microbially-derived cellulose to cellobiose. Cellobiose is likely converted to D-glucose via one of 14 different glycosyl hydrolases. The genome also encodes genes for production of D-glucose via breakdown of starch/glycogen and trehalose. It also encodes several genes for hydrolysis of 1,3- $\beta$ -glucan, one of the most common fungal cell wall polysaccharides [19], to D-glucose. Sugars are a significant portion of plant root exudation, these sugars could be fed directly into the *T. rhizosphaerense* metabolism [20]. D-glucose can be converted to D-glucose-6P and D-fructose-6P and fed into the glycolysis pathway. The genome lacks a gene to phosphorylate  $\beta$ -D-fructose-6P, but this reaction is accomplished via pentose phosphate pathway. The product of these reactions is pyruvate. Pyruvate is not converted to acetyl-CoA. Interestingly, no reaction appears to generate or use acetyl-CoA, as reported for Saccharibacteria RAAC3. Acetyl-CoA metabolism also appears to be lacking in the Saccharibacteria represented by the four other publically available complete genomes.

The *T. rhizosphaerense* genome encodes a phosphoketolase that converts D-xylose-5P derived from the pentose phosphate pathway to acetyl-P that then can be converted to acetate, with production of ATP. Thus, we predict growth via fermentation under anaerobic conditions. ATP also can be produced by a F-type  $H^+$  transporting ATPase.

In glycolysis,  $NAD^+$  is consumed to form NADH that can be regenerated via what appears to be a linear electron transport chain that includes a single subunit NADH dehydrogenase (NDH-II). We identified several residues expected for function. Modeling revealed a close secondary structure match between the *T. rhizosphaerense* protein and NDH-II characterized from *E. coli*. The large evolutionary distance between *T. rhizosphaerense* and *E. coli* likely accounts for differences in some active site residues.

The genome encodes a few genes in an incomplete pathway for production of quinone-based molecules, two of which are in multicopy (five copies of genes annotated as ubiG and two copies as ubiE). We suspect that quinone is scavenged from an external source. We identified genes for a cytochrome-o-ubiquinol oxidase. Subunit I contains nearly all functional residues as well as the residues that distinguish it from cytochrome-c oxidases [21]. Five FNR family transcriptional regulators may serve to detect O<sub>2</sub>. If O<sub>2</sub> is available, electrons could be passed from the NDH-II to a complete cytochrome-o-ubiquinol oxidase, which pumps four protons with the reduction of oxygen.

Several genes such as a blue-copper protein and NADH-quinone oxidoreductase subunit L (NuoL)-related protein also may be involved in electron transfer and contribute to membrane potential in NDH-II systems [22]. Three cytosolic NAD(P)H:quinone oxidoreductase genes were identified. These may reduce semiquinone (SQ) formed under conditions of high O<sub>2</sub> availability to prevent reaction of SQ with O<sub>2</sub> to form oxygen radicals [23]. We identified a novel protein predicted to be associated with the cytoplasmic membrane with two of the domains found in subunits of the Na<sup>+</sup>-translocating NADH-quinone reductase and an FeS cluster domain. We suspect that this may function in the electron transport chain, converting NADH to NAD<sup>+</sup>. The subunit F-like domain may oxidize NADH and transfer electrons to the iron-sulfur domain and the subunit B-like domain could form the Na<sup>+</sup> translocating channel [24].

### **Other predicted capacities indicative of lifestyle**

The genome lacks CRISPR-Cas defense. However, we identified a restriction modification system that could serve this function.

We predict the capacity for twitching motility, guided by one of four chemotaxis response regulators. In addition to DNA uptake, type IV pili could be used for attaching to other cells, root surfaces or solids. There are also several genes for pseudo-pili, and an autotransporter adhesin that may also be involved in cellular adhesion. Also, annotated were two CAZy carbohydrate-binding module family 44 genes for binding to cellulose and the capacity for biosynthesis of cellulose that could be used to attach to plant surfaces [25].

We identified genes encoding for laccase and pyranose 2-oxidase, which may be used for lignin breakdown or detoxification of phenolics, and genes for the detoxification of lignin



byproducts, including 3-oxoadipate enol-lactonase and 4-carboxymuconolactone decarboxylase [26].

Interestingly, we identified genes whose roles may be to modulate plant physiology. For example, we identified genes for the production of cis-zeatin (a plant hormone) from isoprenoid precursors (but a pathway for formation of the precursor isopentenyl-PP was not present). The genome encodes a protein annotated as “Lonely Guy,” a cytokinin (a plant growth hormone) activating enzyme. We also found a gene encoding salicylate hydroxylase, which breaks down salicylic acid, a plant defense signaling molecule.

Also identified was N-acyl homoserine lactone hydrolase, a gene for quorum quenching of other soil microbes and a gene to form 3',5'-cyclic-AMP, which may be involved in intracellular signaling. We predict the presence of genes that confer resistance to bacterially produced antibiotics (beta-lactams, streptomycin, oleandomycin, methylenomycin A, vancomycin, and general macrolides). A gene annotated as phosphatidylglycerol lysyltransferase may produce lysylphosphatidylglycerol, a membrane lipid involved in cationic antimicrobial peptide resistance [27]. We also found a gene that may confer resistance to fusaric acid, an antibiotic made by a common fungal pathogen of grass [28]. This fungus was found to be growing in the rhizosphere (data not shown). Interestingly, there is a gene that encodes a 2,487 amino acid protein that contains domains found in insecticidal proteins, Yen-Tc toxin, rearrangement hotspot repeats, and a protective antigen domain. The protein is predicted to be secreted and may be a toxin. The genome contains at least 16 genes related to toxin/antitoxin systems which have been implicated in stress response [29].

### **Comparative genomics**

The *T. rhizosphaerense* genome is 28% larger than the largest reported complete Saccharibacteria genome, which is from an anaerobic bioreactor. It is 38% larger than the average size of all complete Saccharibacteria genomes (**Table 1**). By re-annotating and analyzing these previously reported complete genomes we found that the *T. rhizosphaerense* genome encodes nearly twice as many unannotated genes as the other Saccharibacteria.

**Table 1.** Genome statistics for the *T. rhizosphaerense* and other complete Saccharibacteria genomes

	Genome Size (bp)	Gene number	Average gene size (bp)	Total coding sequence (bp)	Coding density	Cazy genes	Unduplicated cazy genes
<i>S. rhizosphaerense</i>	1450269	1520	843.3	1280911	88	58	27
scnpilot	900471	918	896.0	822552	91	30	17
GW2011 GWC2.44_17	1038683	1093	860.3	939465	90	40	20
RAAC3_TM7	845464	921	835.5	769452	91	28	17
TM7x	705138	711	919.2	653574	93	30	17
<i>Saccharimonas aalborgensis</i>	1013781	1056	876.5	925533	91	30	17
<i>S. rhizosphaerense</i> difference from average	<b>37.9%</b>	<b>38.2%</b>	<b>-4.1%</b>	<b>35.8%</b>	<b>-3.4%</b>	<b>45.5%</b>	<b>34.8%</b>

There are 130 *T. rhizosphaerense* functional annotations that were not found in any other Saccharibacteria, with 152 genes falling into these annotations. A few appear to be involved in amino acid metabolism, others in transcriptional regulation, DNA repair, sugar metabolism, transport, non-homologous end-joining, three genes annotated as NAD(P)H:quinone reductase and a dihydropteroate synthase for use in folate synthesis. The genome appears to encode a nickel superoxide dismutase for oxidative stress response. None of the other genomes encode the plant hormone related genes: “Lonely Guy” or salicylate hydroxylase. The genome also encodes transaldolase, an addition to the pentose phosphate pathway, pyruvate, water dikinase and the ability to take methylglyoxal to L-Lactate. It is the only Saccharibacteria with a twin-arginine protein translocation system (TatBC), although the function of the proteins with predicted TAT motifs are unknown. Of the *T. rhizosphaerense* genes involved in phosphatidylinositol mannoside production, CDP-diacylglycerol-inositol 3-phosphatidyltransferase is unique to *T. rhizosphaerense* and alpha-1,6-mannosyltransferase is also found in *Saccharimonas aalborgensis*.

*T. rhizosphaerense* has a notably larger repertoire of carbohydrate active enzymes than occurs in the other Saccharibacteria, including 58 genes with 27 unique annotations. Included in the set and not found in other Saccharibacteria genomes are genes predicted to confer the ability to hydrolyze hemicellulose (AG-oligosaccharides and mannoooligosaccharides), amino sugars (galactosaminide), and pectin (oligogalacturonides).

The *T. rhizosphaerense* genome lacks several annotated genes found in all other Saccharibacteria. *T. rhizosphaerense* is unable to make (p)ppGpp, an alarmone that down regulates gene expression. Also lacking is dihydrolipoyl dehydrogenase, a subunit of pyruvate

dehydrogenase (the function of which is unclear in these bacteria). Additionally not found are a *recJ* gene that encodes for a single stranded DNA exonuclease, a glutamine amidotransferase involved in cobalamin synthesis, the *murC* gene involved in peptidoglycan synthesis, and the *PilW* gene involved in pilus stability.

## Discussion

The *T. rhizospherense* genome does not encode the capacity to generate the ribose backbone of DNA or bases and thus it must acquire nucleotides from external sources. Importantly, its DNA is labeled with  $^{13}\text{C}$  that likely originated from  $^{13}\text{CO}_2$  fixed by the plant over the six-week study period. There are several sources of labeled DNA that might have been available to *T. rhizospherense*. We suspect that it obtained DNA from bacteria that lived off plant-derived carbon. However, some of the *T. rhizospherense* DNA might have been from phage that killed bacteria that grew on the  $^{13}\text{C}$ -labeled plant exudates. This is plausible because we identified labeled phage DNA (data not shown). A comprehensive analysis of the stable isotope probing experiment and overall community composition will be described in a separate publication. Alternatively, *T. rhizospherense* might acquire DNA from fungi or the plant, but we consider this less likely given the predicted obligate association between Saccharibacteria and other bacteria.

As Saccharibacteria depend on extracellular nucleosides, they may have difficulty regulating the correct nucleoside concentrations. This is known to be a mutagenic condition and a possible reason for the large number of DNA repair mechanisms [30]. The lifestyle of *T. rhizospherense* is novel compared to those of other CPR in that it is predicted to have the capacity to grow aerobically. Consistent with this, it has a large number of oxidative stress response genes. A close relationship with plant roots also may explain the large number of these genes, as plants release reactive oxygen species in response to pathogens and stress. It is the first member of the Saccharibacteria to be described from the rhizosphere and it appears to be adapted to life there via an expanded genetic repertoire for carbohydrate degradation relative to other Saccharibacteria.

It is possible that *T. rhizospherense* is a symbiont of a soil dwelling bacterium, with a lifestyle analogous to that of the Saccharibacteria TM7x that was cultured as an epibiont of a mouth dwelling Actinobacteria [3, 31]. A similar association is indicated by the antibiotic

resistance mechanisms that are encoded by *T. rhizospherense* that protect against compounds produced by Actinobacteria, methylenomycin A, oleandomycin, vancomycin and streptomycin. Additionally, the ability for *T. rhizospherense* to produce phosphatidylinositol mannosides provides an intriguing connection between this organism and the clades of Actinobacteria that produce phosphatidylinositol mannosides, *Mycobacterium* being the best studied. In *Mycobacterium*, this cell wall component may protect the cell from antibiotics and act as a virulence factor during infection [18].

We do not know the extent of the labeling of cellular components of *T. rhizospherense*, other than the DNA, so we cannot rule out the possibility that this organism makes direct use of plant exudates for purposes other than DNA synthesis. However, the presence of machinery to degrade a variety of complex carbohydrates including cellulose and fungal cell walls suggests that it may not rely completely on plant exudates. *T. rhizospherense* has many genes for attachment that could be used to connect to a microbial host or plant surfaces. It may directly interact with the plant, given genes for the modulation of signaling molecules zeatin and salicylic acid.

## Conclusions

To our knowledge, this is the first study to generate a complete, circularized genome from a soil dwelling bacteria *de novo* from any soil metagenome. It is the first stable isotope-informed genome-resolved metagenomic study in the rhizosphere and thus the first soil microbiome study to make use of stable isotope probing to track carbon from atmospheric CO<sub>2</sub> into the genome of a novel rhizosphere organism. The evidence indicates that <sup>13</sup>C was incorporated into the *T. rhizospherense* DNA through its use of nucleic acids that were synthesized by other organisms that took up plant exudates. *T. rhizospherense* is the first genomically described rhizosphere-associated member of the Saccharibacteria phylum and many of its predicted metabolic capacities distinguish this organism from related Saccharibacteria that live in anaerobic environments. The rhizosphere alternates between aerobic and anaerobic conditions, and it appears that *T. rhizospherense* may have adapted to these cycles by having the capacity to perform fermentation and to grow aerobically using an unusual electron transport chain.

## Materials and methods

### *Labelling*

Soil (0-10cm) was collected from the University of California Hopland Research and Extension Center (Hopland, CA, USA), from an area where *Avena* spp. are a common grass. Microcosms were constructed and plant growth conditions were regulated as described previously [32]. The *A. fatua* plants were grown in labelling chambers with a 400 ppm atmosphere, the native CO<sub>2</sub> was replenished with 99 atm% <sup>13</sup>CO<sub>2</sub>. Rhizosphere and bulk soils were sampled at vegetative (6 weeks) and flowering (9 weeks) stages; microcosms were destructively harvested for rhizosphere soil and from bulk soil bags which exclude root ingrowth. Rhizosphere soil was washed off the roots and DNA was extracted from 0.5g of soil using a phenol:chloroform extraction detailed in [32].

### *Stable isotope probing*

To separate isotopically enriched DNA from unenriched DNA, each sample was separated based on density in a cesium chloride density gradient formed in an ultracentrifuge. Gradients were generated according to the method previously described [33]. Briefly 5.5 µg of DNA was added to the gradient buffer to create a solution with a density of 1.735 g/mL. Then 5.2 mL of the solution was transferred to an ultracentrifuge tube (Beckman Coulter Quick-Seal, 13 X 51 mm). Tubes were spun in an Optima L-90K ultracentrifuge (Beckman Coulter, Brea, California, USA) using a VTi65.2 rotor at 44000 rpm (176284 RCF<sub>avg</sub>) at 20°C for 109 h with maximum acceleration and braking. Then the content of the ultracentrifuge tube was separated into ~32 fractions using a syringe pump to deliver light mineral oil at 0.25mL/min to displace the gradient solution from the pierced bottom of the tube. Each fraction was approximately 12 drops (~144µL). The density of each fraction was measured using an AR200 digital refractometer (Reichert Inc., Depew, New York, USA). The DNA for each fraction was precipitated and quantified as previously described [33]. Fractions were then binned based on density and by comparison between the rhizosphere samples and the associated bulk sample (light= 1.692-1.737 g/mL; middle= 1.738-1.746 g/mL; heavy= 1.747-1.765 g/mL)

### *Sequencing*

Each fraction, light, middle, and heavy was then sent to the UC Davis Genome Center DNA Technologies Core for sequencing. Each sample was sequenced using an Illumina HiSeq 3000 (Illumina Inc., Hayward, California, USA) with paired-end libraries prepared with the Kapa Hyper protocol and a read length of 150 bp.

#### *Genome reconstruction, annotation and analysis*

Reads were trimmed using Sickle (<https://github.com/najoshi/sickle>); BBtools (<https://sourceforge.net/projects/bbmap/>) was used to remove Illumina adapters and trace contaminants; finally reads were assembled using IDBA-UD (-step 20, -maxk 140, -mink 40) [34]. A single 1.45 Mb scaffold was recovered and was able to be circularized, then scaffolding errors and completeness were assessed as described in [16] and manually by mapping reads to the scaffold using Bowtie2 on default settings [35]. The scaffold was visualized in Geneious [36]. Genes were predicted using Prodigal [37]. Predicted ORFs were functionally described using a multidatabase search pipeline. Sequence similarity searches were performed with USEARCH [38] against UniRef100 [39], Uniprot [40], and the KEGG database [41]. Additional gene annotations were assigned using HMMs that were constructed based on KEGG Orthologies [41]. All the proteins assigned to a KO were clustered using MCL [42] with inflation parameter (-I) of 1.1, based on global percent identity. Clusters were aligned using MAFFT v7 [43], and HMMs were constructed using the HMMER suite [44]. Carbohydrate-active enzymes were identified using dbCAN [45]. Domain level functional annotations were done using InterProScan [46]. tRNAs were predicted using tRNAScan-SE [47] and cellular localization was predicted using PSORTb with the gram-positive setting, which was demonstrated though previous imaging [48, 49]. The GC skew of the genome was calculated based on previously published tools [50]. Protein modelling was done with Swiss Model and NDH-2 modelling incorporated *Caldalkalibacillus thermarum* NDH-2 protein structure [51, 52]. Representative Saccaribacteria 16S rRNA sequences were obtained from NCBI and aligned using ssu-align [53] then a maximum-likelihood tree was constructed with RAxML by using the GTRCAT model with 100 bootstraps.

## Declarations

**Ethics approval and consent to participate:** Not applicable.

**Consent for publication:** Not applicable.

**Availability of data and material:** The genome will be made available on NCBI upon acceptance of this manuscript and via <http://ggkbase.berkeley.edu/tm7cure/organisms/342358>. The mapped read file can be also be downloaded from this site.

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

**Funding:** This research was supported by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research Genomic Science program under Awards DE-SC0010570 and DE-SC0016247 to MF and DE-SC10010566 to JB. AJP was supported by the German Science Foundation under DFG1603-1/1. ES was supported by a grant from the National Science Foundation Graduate Research Fellowships Program.

**Authors' contributions:** SS and MF designed the labeling experiment. SS and DH carried out the labeling. ES and SB conducted the density centrifugation. ES and JB analyzed the data, with input from AP. ES and JB wrote the manuscript and all authors read and approved the final manuscript.

**Acknowledgements:** This work was made possible by samples and expertise provided by personnel at the Hopland Research and Extension Center (Hopland, CA, USA). Sequencing was done at the UC Davis Genome Center and technical advice was provided by Lutz Froenicke. The HMM pipeline was developed by Dr. David Burstein and protein modeling assistance was provided by Dr. Cindy Castelle.

## References:

1. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*. 2015;523:208–11. doi:10.1038/nature14486.
2. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. *Nat Microbiol*. 2016;1:16048. doi:10.1038/nmicrobiol.2016.48.
3. He X, McLean JS, Edlund A, Yooseph S, Hall AP, Liu S-Y, et al. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc Natl Acad Sci U S A*. 2015;112:244–9. doi:10.1073/pnas.1419038112.
4. Rheims H, Rainey FA, Stackebrandt E. A molecular approach to search for diversity among bacteria in the environment. *J Ind Microbiol*. 1996;17:159–69. doi:10.1007/BF01574689.
5. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science*. 2012;337:1661–5. doi:10.1126/science.1224041.
6. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun*. 2016;7:13219. doi:10.1038/ncomms13219.
7. Podar M, Abulencia CB, Walcher M, Hutchison D, Zengler K, Garcia JA, et al. Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl Environ Microbiol*. 2007;73:3205–14. doi:10.1128/AEM.02985-06.
8. Correa-Galeote D, Bedmar EJ, Fernández-González AJ, Fernández-López M, Arone GJ. Bacterial Communities in the Rhizosphere of Amilaceous Maize (*Zea mays* L.) as Assessed by Pyrosequencing. *Front Plant Sci*. 2016;7:1016. doi:10.3389/fpls.2016.01016.
9. Beckers B, Op De Beeck M, Weyens N, Boerjan W, Vangronsveld J. Structural variability and niche differentiation in the rhizosphere and endosphere bacterial microbiome of field-grown poplar trees. *Microbiome*. 2017;5:25. doi:10.1186/s40168-017-0241-2.
10. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen re L, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol*. 2013;31. doi:10.1038/nbt.2579.
11. Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, et al. Small Genomes and Sparse Metabolisms of Sediment-Associated Bacteria from Four Candidate Phyla.



doi:10.1128/mBio.00708-13.

12. Kindaichi T, Yamaoka S, Uehara R, Ozaki N, Ohashi A, Albertsen M, et al. Phylogenetic diversity and ecophysiology of Candidate phylum Saccharibacteria in activated sludge. *FEMS Microbiol Ecol.* 2016;92:fiw078. doi:10.1093/femsec/fiw078.

13. Schildkraut CL, Marmur J, Doty P. Determination of the base composition of deoxyribonucleic acid from its buoyant density in CsCl. *J Mol Biol.* 1962;4:430–43. doi:10.1016/S0022-2836(62)80100-4.

14. Birnie GD, Rickwood D. Centrifugal separations in molecular and cell biology. Butterworths; 1978.

15. Anantharaman K, Brown CT, Burstein D, Castelle CJ, Probst AJ, Thomas BC, et al. Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum. *PeerJ.* 2016;4:e1607. doi:10.7717/peerj.1607.

16. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature.* 2015;523:208–11. <http://dx.doi.org/10.1038/nature14486>.

17. Peregrín-Alvarez JM, Sanford C, Parkinson J. The conservation and evolutionary modularity of metabolism. *Genome Biol.* 2009;10:R63. doi:10.1186/gb-2009-10-6-r63.

18. Kaur D, Guerin ME, Skovierová H, Brennan PJ, Jackson M. Chapter 2: Biogenesis of the cell wall and other glycoconjugates of *Mycobacterium tuberculosis*. *Adv Appl Microbiol.* 2009;69:23–78. doi:10.1016/S0065-2164(09)69002-X.

19. Douglas CM. Fungal B (1,3)-D-glucan synthesis. *Med Mycol.* 2001;39:55–66.

20. Sasse J, Martinoia E, Northen T. Feed Your Friends: Do Plant Exudates Shape the Root Microbiome? *Trends Plant Sci.* 2017. doi:10.1016/j.tplants.2017.09.003.

21. Iwata S, Wikström M, Abramson J, Riistama S, Larsson G, Jasaitis A, et al. The structure of the ubiquinol oxidase from *Escherichia coli* and its ubiquinone binding site. *Nat Struct Biol.* 2000;7:910–7. doi:10.1038/82824.

22. Mayer S, Steffen W, Steuber J, Götz F. The *Staphylococcus aureus* NuoL-like protein MpsA contributes to the generation of membrane potential. *J Bacteriol.* 2015;197:794–806. doi:10.1128/JB.02127-14.

23. Patridge E V, Ferry JG. WrbA from *Escherichia coli* and *Archaeoglobus fulgidus* Is an NAD(P)H:Quinone Oxidoreductase. *J Bacteriol.* 2006;188:3498–506.

doi:10.1128/JB.188.10.3498–3506.2006.

24. Steuber J, Vohl G, Casutt MS, Vorburger T, Diederichs K, Fritz G. Structure of the V. cholerae Na<sup>+</sup>-pumping NADH:quinone oxidoreductase. *Nature*. 2014;516:62–7.

doi:10.1038/nature14003.

25. Rodríguez-Navarro DN, Dardanelli MS, Ruíz-Saínz JE. Attachment of bacteria to the roots of higher plants. *FEMS Microbiol Lett*. 2007;272:127–36. doi:10.1111/j.1574-6968.2007.00761.x.

26. Billings AF, Fortney JL, Hazen TC, Simmons B, Davenport KW, Goodwin L, et al. Genome sequence and description of the anaerobic lignin-degrading bacterium *Tolomonas lignolytica* sp. nov. *Stand Genomic Sci*. 2015;10:106. doi:10.1186/s40793-015-0100-3.

27. Sohlenkamp C, Galindo-Lagunas KA, Guan Z, Vinuesa P, Robinson S, Thomas-Oates J, et al. The Lipid Lysyl-Phosphatidylglycerol Is Present in Membranes of *Rhizobium tropici* CIAT899 and Confers Increased Resistance to Polymyxin B Under Acidic Growth Conditions. *Mol Plant-Microbe Interact*. 2007;20:1421–30. doi:10.1094/MPMI-20-11-1421.

28. Gagkaeva TY, Gavrilova OP, Orina AS, Blinova EV, Loskutov IG. Response of wild *Avena* species to fungal infection of grain. *Crop J*. 2017. doi:10.1016/j.cj.2017.04.005.

29. Pandey DP, Gerdes K. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Res*. 2005;33:966–76. doi:10.1093/nar/gki201.

30. Nordman J, Wright A. The relationship between dNTP pool levels and mutagenesis in an *Escherichia coli* NDP kinase mutant. *Proc Natl Acad Sci U S A*. 2008;105:10197–202.

doi:10.1073/pnas.0802816105.

31. Bor B, Poweleit N, Bois JS, Cen L, Bedree JK, Zhou ZH, et al. Phenotypic and Physiological Characterization of the Epibiotic Interaction Between TM7x and Its Basibiont Actinomyces.

*Microb Ecol*. 2016;71:243–55. doi:10.1007/s00248-015-0711-7.

32. Shi S, Nuccio E, Herman DJ, Rijkers R, Estera K, Li J, et al. Successional Trajectories of Rhizosphere Bacterial Communities over Consecutive Seasons. *mBio*. 2015;6:4.

doi:10.1128/mBio.00746-15.

33. Blazewicz SJ, Schwartz E, Firestone MK. Growth and death of bacteria and fungi underlie rainfall-induced carbon dioxide pulses from seasonally dried soil. *Ecology*. 2014;95:1162–72.

doi:10.1890/13-1031.1.

34. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell

- and metagenomic sequencing data with highly uneven depth. *Bioinformatics*. 2012;28:1420–8. doi:10.1093/bioinformatics/bts174.
35. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–9. doi:10.1038/nmeth.1923.
36. Kears M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012;28:1647–9. doi:10.1093/bioinformatics/bts199.
37. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11:119. doi:10.1186/1471-2105-11-119.
38. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26:2460–1. doi:10.1093/bioinformatics/btq461.
39. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*. 2007;23:1282–8. doi:10.1093/bioinformatics/btm098.
40. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. *Database*. 2011;2011:bar009–bar009. doi:10.1093/database/bar009.
41. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28:27–30. <http://www.ncbi.nlm.nih.gov/pubmed/10592173>.
42. Van Dongen S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J Matrix Anal Appl*. 2008;30:121–41. doi:10.1137/040608635.
43. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30:772–80. doi:10.1093/molbev/mst010.
44. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011;39 Web Server issue:W29–37. doi:10.1093/nar/gkr367.
45. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2012;40:W445–51. doi:10.1093/nar/gks479.
46. Zdobnov EM, Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. 2001;17:847–8.

doi:10.1093/bioinformatics/17.9.847.

47. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 1997;25:955–64.

<http://www.ncbi.nlm.nih.gov/pubmed/9023104>.

48. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, et al. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics.* 2010;26:1608–15.

doi:10.1093/bioinformatics/btq249.

49. Hugenholtz P, Tyson GW, Webb RI, Wagner AM, Blackall LL. Investigation of Candidate Division TM7, a Recently Recognized Major Lineage of the Domain Bacteria with No Known Pure-Culture Representatives. *Appl Environ Microbiol.* 2001;67:411–9.

doi:10.1128/AEM.67.1.411–419.2001.

50. Brown CT, Olm MR, Thomas BC, Banfield JF. Measurement of bacterial replication rates in microbial communities. *Nat Biotech.* 2016;34:1256–63. <http://dx.doi.org/10.1038/nbt.3704>.

51. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 2014;42:W252–8. doi:10.1093/nar/gku340.

52. Blaza JN, Bridges HR, Aragão D, Dunn EA, Heikal A, Cook GM, et al. The mechanism of catalysis by type-II NADH:quinone oxidoreductases. *Sci Rep.* 2017;7:40165.

doi:10.1038/srep40165.

53. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments.

*Bioinformatics.* 2009;25:1335–7. doi:10.1093/bioinformatics/btp157.

**Figure 1.** Stable isotope fraction determination. This plot shows the distribution of densities and concentrations of DNA extracted from rhizosphere and bulk soil following density centrifugation. The offset of the curves demonstrates clear enrichment of heavy DNA in the rhizosphere. The black dots on the curves represent individual fraction measurements. The three fractions are designated as light (blue shading), middle (yellow shading), and heavy (red shading). Numbers indicate the coverage of the *T. rhizosphaerense* genome in each fraction. The *T. rhizosphaerense* genome had < 1x coverage in each bulk soil fraction.

**Figure 2.** Phylogeny of Saccharibacteria based on 16S rRNA gene sequences. The maximum-likelihood tree shown was constructed from an alignment containing representative Saccharibacteria. Named branches indicate the complete genomes included in this study. Bootstrap values of >50% are displayed.

**Figure 3.** Plot of GC skew (black) and cumulative GC skew (green, window 1000 bp, slide of 10 bp) of the *T. rhizosphaerense* genome. The plot shows the predicted locations of the origin (red line, 1,201 bp) and terminus (blue line, 757,370 bp) of replication. The form of the plot is as expected for a correctly assembled, circularized genome that undergoes bi-directional replication from a single origin.

**Fig. 4.** Cell diagram of *T. rhizosphaerense*. A. fusaric acid resistance machinery, B. unidentified importer, C. glucan 1,3-beta-glucosidase, D. NADH dehydrogenase II, E. blue-copper protein, F. cytochrome-o-ubiquinol oxidase, G. F-type H<sup>+</sup>-transporting ATPase, H. peptidase, I. nuclease, J. root hair, K. cellulosome, L. type IV pilus, M. salicylate hydroxylase, N. zeatin production, O. restriction system for possible phage defense, P. various antibiotic resistance mechanisms, Q. intercellular attachment, R. scavenging of lipids, S. production of phosphatidyl myo-inositol mannosides, T. DNA repair machinery.

**Supplementary Table 1:** Genome summary, including HMM-based annotations.









