Title: Complete genomes of clade G6 Saccharibacteria suggest a divergent ecological niche and lifestyle

Author: Jonathon L. Baker[1,*]
[1] Genomic Medicine Group
J. Craig Venter Institute
4120 Capricorn Lane
La Jolla, CA 92037

*Corresponding Author: JLB: jobaker@jcvi.org

ORCID: JLB: 0000-0001-5378-322X

Running title: Highly divergent Clade G6 Saccharibacteria

Keywords: Saccharibacteria, oral microbiome, TM7, nanopore sequencing

ABSTRACT

Saccharibacteria (formerly TM7) have reduced genomes, a small size, and appear to have a parasitic lifestyle dependent on a bacterial host. Although there are at least 6 major clades of Saccharibacteria inhabiting the human oral cavity, cultured isolates or complete genomes of oral Saccharibacteria have been previously limited to the G1 clade. In this study, nanopore sequencing was used to obtain three complete genome sequences from clade G6. Phylogenetic analysis suggested the presence of at least 3-5 distinct species within G6, with two discrete taxa represented by the 3 complete genomes. G6 Saccharibacteria were highly divergent from the more well-studied clade G1, and had the smallest genomes and lowest GC-content of all Saccharibacteria. Pangenome analysis showed that although 97% of shared pan-Saccharibacteria core genes and 89% of G1-specific Core Genes had putative functions, only 50% of the 244 G6-specific Core Genes had putative functions, highlighting the novelty of this group. Compared to G1, G6 encoded divergent metabolic pathways. G6 genomes lacked an F1F0 ATPase, the pentose phosphate pathway, and several genes involved in nucleotide metabolism, which were all core genes for G1. G6 genomes were also unique compared to G1 in that they encoded lactate dehydrogenase, adenylate cyclase, limited glycerolipid metabolism, a homolog to a lipoarabinomannan biosynthesis enzyme, and the means to degrade starch. These differences at key metabolic steps suggest a distinct lifestyle and ecological niche for clade G6, possibly with alternative hosts and/or host-dependencies, which would have significant ecological, evolutionary, and likely pathogenic, implications.

37 **IMPORTANCE**

38 Saccharibacteria are ultrasmall, parasitic bacteria that are common members of the oral
39 microbiota and have been increasingly linked to disease and inflammation. However, the lifestyle
40 and impact on human health of Saccharibacteria remains poorly understood, especially for the 5
41 clades (G2-G6) with no complete genomes or cultured isolates. Obtaining complete genomes is
42 of particular importance for Saccharibacteria, because they lack many of the "essential" core
43 genes used for determining draft genome completeness and few references exist outside of clade
44 G1. In this study, complete genomes of 3 G6 strains, representing two candidate species, were
45 obtained and analyzed. The G6 genomes were highly divergent from G1, and enigmatic, with
46 50% of the G6 core genes having no putative functions. The significant difference in encoded
47 functional pathways is suggestive of a distinct lifestyle and ecological niche, probably with
48 alternative hosts and/or host-dependencies, which would have major implications in ecology,
49 evolution, and pathogenesis.

50 **OBSERVATION**

51 Saccharibacteria (formerly TM7) have an ultrasmall cell size, reduced genomes, and are thought
52 to be obligate epibionts, dependent on physically-associated host species (1-3). Common
53 constituents of the oral microbiota, Saccharibacteria have been increasingly linked to
54 inflammation and disease (4-6). Saccharibacteria contains at least 6 distinct clades (G1-G6)(7,
55 8), however all currently available human-associated complete genomes and cultured isolates
56 belong to clade G1, leaving clades G2-G6 quite poorly understood. Several recent publications
57 have provided the first draft genomes from clades G3, G5, and G6 (4, 8-11). Obtaining complete
58 genomes is of particular importance for Saccharibacteria, because they lack many of the
59 "essential" single-copy core genes that are typically used to estimate genome completion, as well
60 as complete reference genomes outside of the G1 clade.

61 A recent, short-read-based oral microbiome study provided 21 Saccharibacteria draft
62 genomes from clades G1, G3, and G6 (4), with several being high quality (high N50, relatively
63 contiguous, low predicted contamination). Therefore, nanopore sequencing of the same saliva
64 samples that had produced the draft genomes, followed by long-read and/or hybrid assembly,
65 was used to improve these genomes, resulting in 3 complete, circular G6 genomes: JB001
66 (662,051 bp), JB002 (639,751 bp), and JB003 (663,165 bp). Table 1 is a summary of the
67 genomes improved during this study and the Supplemental Methods contain a full description of
68 the DNA extraction, sequencing, assembly, and analysis methods. These methods are a modified
69 version of a previously reported protocol (Baker 2021, in-press). Although the G1 and G3 "near
70 complete" improved genomes that were obtained are useful in their own right, they are still
71 incomplete, and/or may contain contamination, therefore the 3 complete G6 genomes are the
72 focus of this report, and the near complete genomes are briefly discussed in the Supplemental
73 Methods.

74 Phylogenetic analysis using concatenated protein sequences was performed using Anvi'o
75 (12), and included the 8 improved/completed genomes from this study, all 26 complete

76 Saccharibacteria genomes available on NCBI (as of 1 April 2021), and 90 Saccharibacteria draft
77 genomes from 5 recent studies (Table S1). JB001, JB002, and JB003 were indeed members of
78 Saccharibacteria clade G6 (Figure 1A, Figure S1), and represent the only human-associated,
79 complete Saccharibacteria genomes outside of clade G1. Notably, G6 had the smallest genomes
80 and the lowest GC-content of all Saccharibacteria (Figure 1A). Percent average nucleotide
81 identity (ANI) between the G6 genomes was calculated using Anvi'o and suggested that there are
82 at least 3-5 distinct species within the clade (Figure 1B; a cutoff of 95% ANI is frequently used to
83 estimate the species level (13, 14)). JB001, JB003, JCVI_1_bin.12, and G6_32_bin_33_unicycler
84 appear to be the same species, with an ANI of ≥95%, despite their source from different human
85 subjects and independent genome assembly (Figure 1B). JB002 and T-C-M-Bin-00022 were
86 over 98% ANI, likely representing the same distinct species, while CMJM-G6-HOT-870 and T-C-
87 M-Bin-00011 were ~98% ANI and formed what is likely an additional G6 species (Figure 1B).
88 CLC Genomics Workbench was used to perform whole genome alignment for JB001, JB002,
89 JB003, and the G1 reference strain, TM7x (Figure 1C). While JB001 and JB003 were completely
90 syntenic, and there were moderate differences between JB001/JB003 and JB002, TM7x and the
91 G6 Saccharibacteria have undergone many genomic re-arrangements and instances of gene
92 gain/loss since their last common ancestor (Figure 1C).

93 To examine functional and metabolic differences between the G6 clade and the more well-
94 understood G1 clade, pangenome analysis was performed using Anvi'o (15) on the 3 complete
95 G6 genomes and 4 diverse G1 complete genomes (Figure 2, Table S3). This identified 223 "pan-
96 Saccharibacteria Core Genes" appearing in all genomes, as well as all 94 "G1 Core Genes", and
97 244 "G6 Core Genes" (Figure 2A). While 97% of the pan-Saccharibacteria Core Genes and 89%
98 of the G1 Core Genes had known COG functions and pathways, only 50% of the G6 Core Genes
99 had known COG functions and pathways (Figure 2A), highlighting the enigmatic nature of this
100 clade. The likely reason for the lower number of G1 core genes is the larger amount of known
101 diversity within the G1 clade and the genomes analyzed here (8, 9), leading to less conservation

102 across the G1 pangenome. A larger pangenome analysis, examining all 11 G6 genomes and 14
103 diverse G1 genomes is available in Figure S2 and Table S4. This generated similar results, but
104 note that this analysis contains incomplete draft genomes which are incomplete and/or may
105 contain contamination. A complete metabolic network illustrating the known KEGG pathways
106 identified in the three sets of core genes identified in Figure 2A is shown in Figure 2B. Both G1
107 and G6 genomes encode partial cell wall metabolism, glycolysis (missing phosphofructokinase),
108 and arginine biosynthesis pathways, and do not encode fatty acid metabolism, a TCA cycle, or
109 amino acid metabolism (other than arginine) (Figure 2B). Notable pathways present in G6
110 genomes but absent in G1 include: maltase glucoamylase (to metabolize starch), fructose
111 bisphosphate aldolase (a glycolytic step), adenylate cyclase, lactate dehydrogenase, partial
112 lipoarabinomannan (LAM) biosynthesis, and partial glycerolipid metabolism. Conversely, G1
113 genomes encode the non-oxidative phase of the pentose phosphate pathway, an F1F0 ATPase,
114 alpha galactosidase, and several steps in nucleotide metabolism, which were not present in the
115 G6 genomes (Figure 2B). Between JB001 and JB002, most differences were genes with
116 unknown functions, therefore the differences in the KEGG pathways encoded were minor (Figure
117 S3). The G6 genomes examined did not contain predicted elements of a CRISPR system.
118 Although it is not known how Saccharibacteria obtain needed metabolites from the host, a type
119 IV pilus-like system is generally well-conserved across the group, has been proposed as a
120 candidate mechanism (8, 9), and was present in the G6 genomes here. The species-level clade
121 that included JB001 and JB003 encoded a ~10,000bp putative prophage element, which was
122 flanked by homologs to the PinE invertase and contained a T4SS VirD4 homolog and 4
123 hypothetical proteins, all with ~95% homology to a similar region in *Streptococcus salivarius.*
124 Taken together, these analyses indicate that Saccharibacteria clade G6 is highly divergent
125 from clade G1, and may have a different lifestyle, host, and host-dependencies. This is in line
126 with the recent hypothesis that G6 reside on the tongue (G6 are referred to as 'T2' in reference
127 9) and have a long history of association with animal hosts, while G1 reside in dental plaque and

128 were a much more recent acquisition from the environment (8, 9). Interestingly, the species-level
129 clade containing JB002 (the most reduced Saccharibacteria genome, with only 615 genes) was
130 the only Saccharibacteria group that resided both on the tongue and in dental plaque (9).
131 Although all cultured isolates of Saccharibacteria were epibionts of *Actinomyces* spp., they were
132 all G1 strains. Residing in a different environment, G6 may have distinct host species, possibly
133 *Streptococcus,* given the acquired homologous sequence. It is likely that G6 fallen into the
134 'unknown' taxonomic bucket in the majority of past microbiome studies, thus the role of G6 in
135 human health remains to be elucidated. The high percentage of genes with unknown functions
136 further adds to the obscurity of this clade. Overall, this article highlights an urgent need for study
137 of Saccharibacteria, since almost nothing is known about the lifestyle, host, or ecological impact
138 of Saccharibacteria clade G6, and even less still is understood about clades G2, G3, G4, and G5.

ACKNOWLEDGEMENTS

I thank Karrie Goglin-Almeida, Jelena Jablanovic, and Kara Riggsbee for performing the library preparation and sequencing, and Jeffrey S. McLean for helpful discussions. This research was supported by NIH/NIDCR K99-DE029228.

DATA AVAILABILITY

The complete genome sequences of JB001, JB002, and JB003 have been deposited in GenBank under the accession numbers: CP072208, SUB9758535, and SUB9758549. The BioProject accession for this project is PRJNA624185. The short reads used to generate the assemblies are available in the SRA database with the accession numbers SRX4318838, SRX4318837, and SRX4318835. The long reads used to generate the assemblies are available in the SRA dataset with the accession numbers SRX10387815 and SUB9758617.

REFERENCES

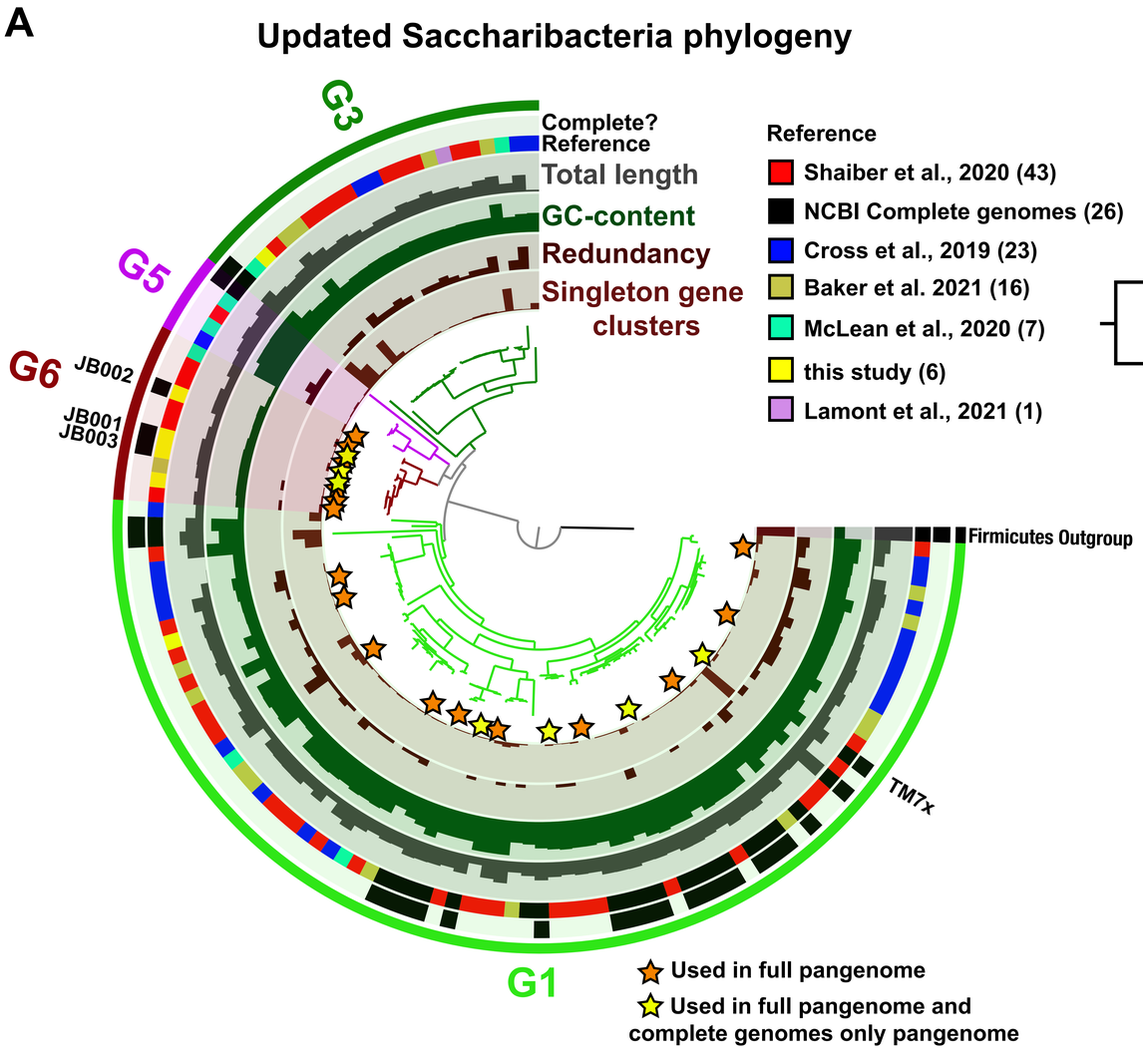
1. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:16048.
2. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523:208-11.
3. He X, McLean JS, Edlund A, Yooseph S, Hall AP, Liu SY, Dorrestein PC, Esquenazi E, Hunter RC, Cheng G, Nelson KE, Lux R, Shi W. 2015. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. Proc Natl Acad Sci U S A 112:244-9.
4. Baker JL, Morton JT, Dinis M, Alvarez R, Tran NC, Knight R, Edlund A. 2021. Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. Genome Res 31:64-74.
5. Abu Fanas S, Brigi C, Varma SR, Desai V, Senok A, D'Souza J. 2021. The prevalence of novel periodontal pathogens and bacterial complexes in Stage II generalized periodontitis based on 16S rRNA next generation sequencing. J Appl Oral Sci 29:e20200787.
6. Bor B, Bedree JK, Shi W, McLean JS, He X. 2019. Saccharibacteria (TM7) in the Human Oral Microbiome. J Dent Res 98:500-509.
7. Camanocha A, Dewhirst FE. 2014. Host-associated bacterial taxa from Chlorobi, Chloroflexi, GN02, Synergistetes, SR1, TM7, and WPS-2 Phyla/candidate divisions. J Oral Microbiol 6.
8. McLean JS, Bor B, Kerns KA, Liu Q, To TT, Solden L, Hendrickson EL, Wrighton K, Shi W, He X. 2020. Acquisition and Adaptation of Ultra-small Parasitic Reduced Genome Bacteria to Mammalian Hosts. Cell Rep 32:107939.
9. Shaiber A, Willis AD, Delmont TO, Roux S, Chen LX, Schmid AC, Yousef M, Watson AR, Lolans K, Esen OC, Lee STM, Downey N, Morrison HG, Dewhirst FE, Mark Welch JL, Eren AM. 2020. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. Genome Biol 21:292.
10. Cross KL, Campbell JH, Balachandran M, Campbell AG, Cooper SJ, Griffen A, Heaton M, Joshi S, Klingeman D, Leys E, Yang Z, Parks JM, Podar M. 2019. Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. Nat Biotechnol 37:1314-1321.
11. Lamont EI, Gadkari A, Kerns KA, To TT, Daubert D, Kotsakis G, Bor B, He X, McLean JS. 2021. Modified SHI medium supports growth of a disease-state subgingival polymicrobial community in vitro. Mol Oral Microbiol 36:37-49.
12. Eren AM, Esen OC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319.
13. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C, Huttenhower C, Segata N. 2019. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell 176:649-662 e20.
14. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114.
15. Delmont TO, Eren AM. 2018. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. PeerJ 6:e4320.

200 **FIGURE LEGENDS**

201 **Figure 1: JB001, JB002, and JB003 are clade G6 Saccharibacteria representing two**
202 **distinct species. (A) Phylogenetic tree of Saccharibacteria annotated with genome data.**
203 Phylogenetic analysis of the 123 Saccharibacteria genomes listed in Table S1. Firmicutes was
204 used as an outgroup. The bars in the innermost layer represent the number of singleton gene
205 clusters (i.e. genes appearing in only that one genome) in each genome. The bars in the second
206 layer represent the redundancy (likely contamination) within each genome. The bars in the third
207 layer represent the %GC content of each genome. The bars in the fourth layer represent the total
208 length in bp of each genome. The fifth layer displays the source/reference for each genome. The
209 sixth layer displays the genomes that are complete. The outermost layer, and the color of the
210 branches of the tree, illustrate which Saccharibacteria clade each genome is part of. Orange
211 stars indicate genomes that were used in the full pangenome analysis (Figure S2, Table S4).
212 Yellow stars indicate genomes that were used in the pangenome analysis of compete genomes
213 only (Figure 2, Table S3) as well as the full pangenome analysis (Figure S2, Table S4). A larger
214 version of this figure, with the name of each genome labeled, is available in Figure S1. Note that
215 CP025011_1_Candidatus_Saccharibacteria_bacterium_YM_S32_TM7_50_20_chromosome_c
216 omplete_genome and c_000000000001 (GCA_003516025.1_ASM351602v1_genomic.fa), the
217 only two complete genomes in clades G3 and G5, are from environmental, not oral, samples. The
218 raw data in the annotations of the tree is available in Table S1. **(B) Average nucleotide identity**
219 **(%ANI) of G6 genomes.** Heatmap of all-vs-all comparison of %ANI of all 11 G6 genomes. The
220 tree on the right is a scaled up version of the G6 portion of the phylogenetic tree in panel A. Full
221 percentage identity, which takes alignment length into account, is available in Table S2. **(C)**
222 **Whole genome alignment of TM7x vs complete G6 genomes.** Whole genome alignment
223 diagram produced by CLC Genomics Workbench. The tree on the right is based on the whole
224 genome alignment itself.

225

Figure 1



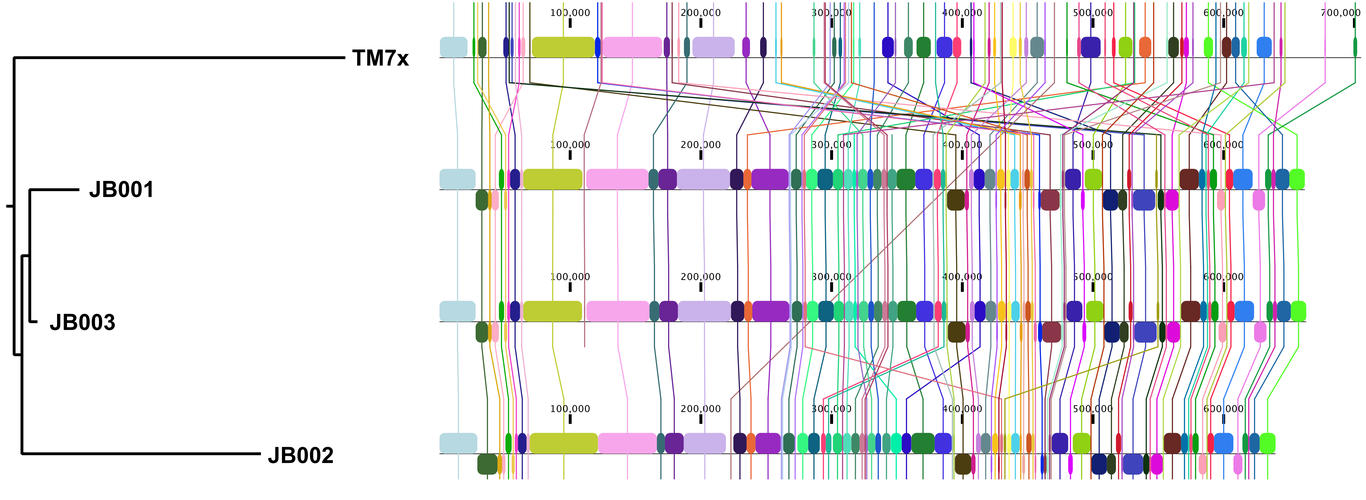
B

Average Nucleotide Identity (%ANI) of G6 genomes

ORAL_T-A-F_MAG_00016	100.0%	85.4%	86.1%	86.1%	86.1%	93.8%	84.9%	85.8%	82.2%	72.9%	72.8%
G6_32_bin_33_unicycler	85.8%	100.0%	94.7%	94.4%	94.4%	85.1%	80.4%	80.3%	80.2%	72.4%	72.3%
FGB1_strain_JCVI_1_bin_12	86.4%	94.8%	100.0%	96.2%	96.3%	85.6%	80.2%	80.6%	80.3%	73.2%	72.9%
JB003_complete_genome	86.4%	94.6%	96.3%	100.0%	99.8%	85.8%	80.3%	81.2%	80.2%	73.0%	73.5%
GCA_015257795_JB001	86.6%	94.5%	96.4%	99.7%	100.0%	86.1%	80.4%	81.0%	80.2%	72.8%	72.9%
ORAL_T-B-F_Bin_00014	93.6%	85.3%	85.8%	85.9%	85.9%	100.0%	85.2%	85.9%	82.2%	72.4%	72.3%
ORAL_T-C-M_Bin_00022	85.8%	80.7%	80.8%	80.9%	80.9%	85.7%	100.0%	96.7%	81.7%	72.9%	73.0%
JB002_complete_genome	85.7%	80.3%	80.7%	81.2%	81.0%	85.8%	96.6%	100.0%	81.7%	72.8%	73.6%
ORAL_T-B-F_Bin_00010	82.6%	80.0%	80.6%	80.5%	80.5%	82.3%	81.4%	81.8%	100.0%	72.6%	72.5%
ORAL_T-C-M_Bin_00011	73.0%	72.1%	72.9%	73.0%	73.0%	72.2%	72.1%	72.8%	72.7%	100.0%	98.1%
TM7_CMJM_G6_1_HOT_870	72.4%	72.0%	72.6%	72.8%	72.6%	72.0%	72.2%	73.1%	72.6%	98.0%	100.0%

C

Whole genome alignment of TM7x vs complete G6 genomes

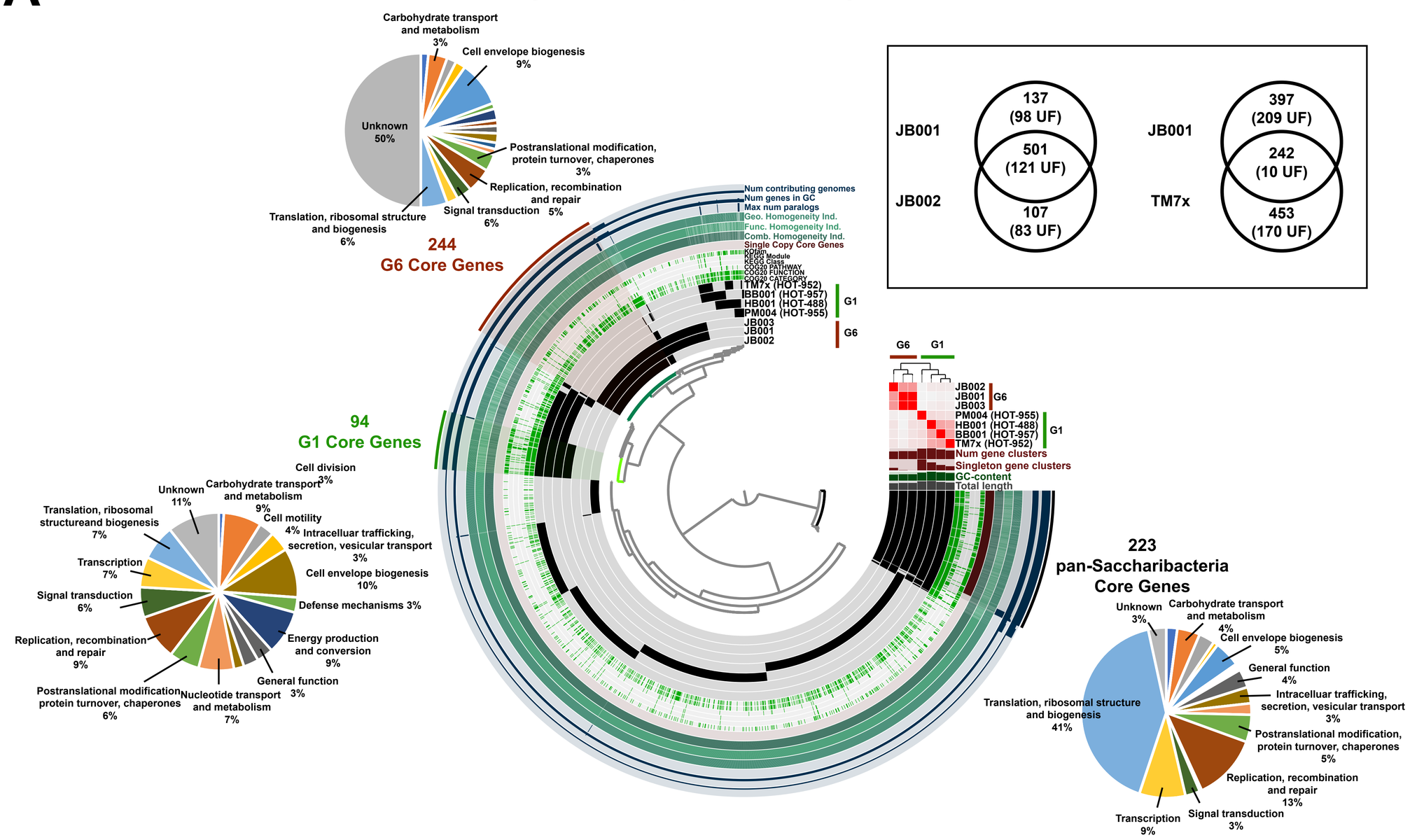


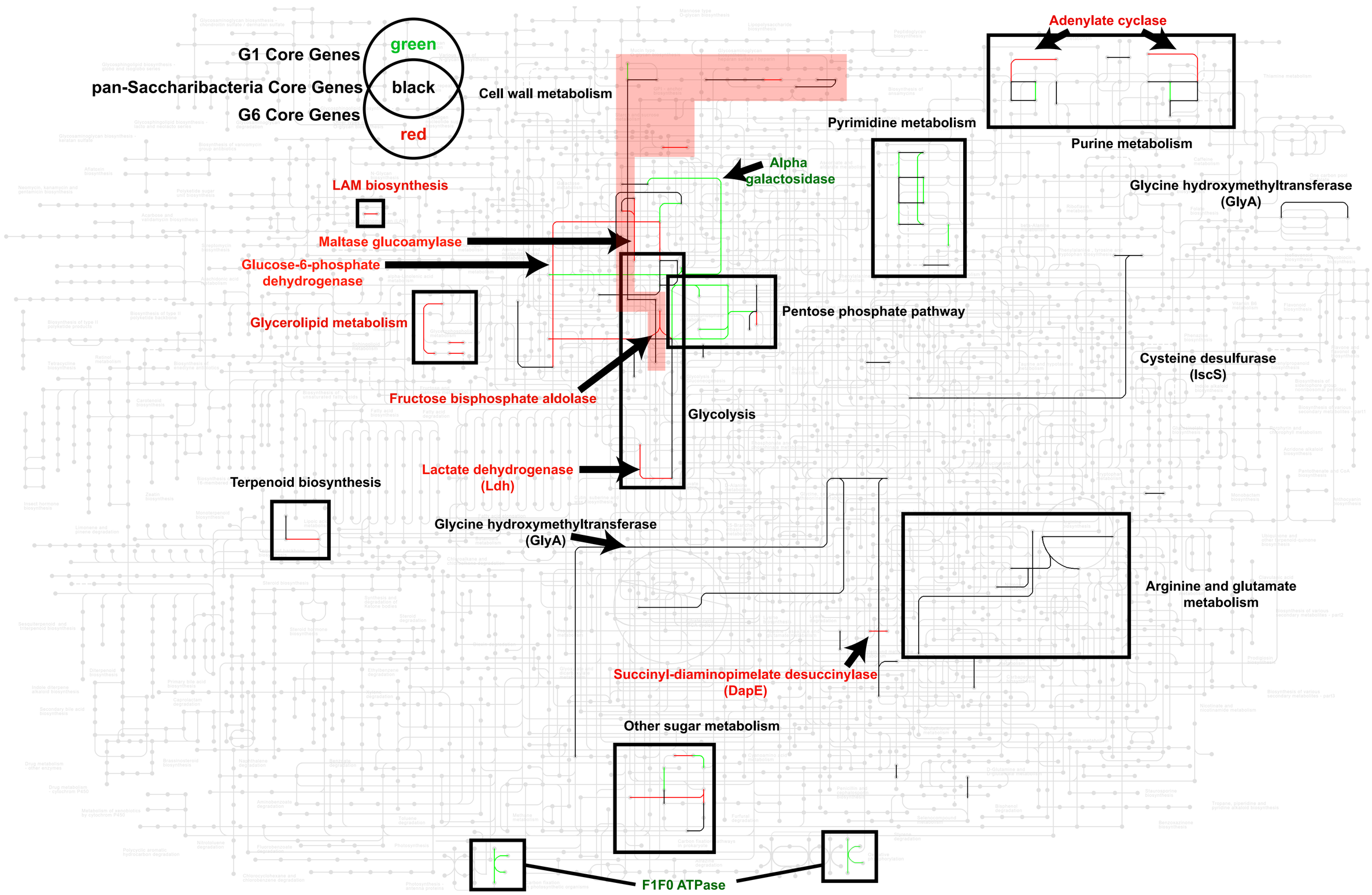
226 **Figure 2: Pangenome analysis of complete genomes in Saccharibacteria clade G1 vs.**
227 **clade G6 identifies core genes with encoding distinct functional pathways. (A) The**
228 **pangenome of complete G1 and G6 genomes.** The dendrogram in the center organizes the
229 2,279 gene clusters identified across in the genomes represented by the innermost 7 layers:
230 TM7x, BB001, HB001, PM004, JB003, JB001, and JB002. The data points within these 7 layers
231 indicate the presence of a gene cluster in a given genome. From inside to outside, the next 6
232 layers indicate known vs unknown COG category, COG function, COG pathway, KEGG class,
233 KEGG module, and KOfam. The next layer indicates single-copy pan-Saccharibacteria core
234 genes. The next 6 layers indicate the combined homogeneity index, functional homogeneity
235 index, geometric homogeneity index, max number of paralogs, number of genes in the gene
236 cluster, and the number of contributing genomes. The outermost layer highlights gene clusters
237 that correspond to the pan-Saccharibacteria Core Genes (found in all 7 genomes), the G1 Core
238 Genes (found in all G1 genomes and no G6 genomes), and the G6 Core Genes (found in all G6,
239 but no G1 genomes). The pie chart adjacent to each group of core genes indicates the breakdown
240 of COG categories of the gene clusters in the group. The 7 genome layers are ordered based on
241 the tree of the %ANI comparison, which is displayed with the red and white heatmap. The layers
242 underneath the %ANI heatmap, from top to bottom, indicate: the number of gene clusters, the
243 number of singleton gene clusters, the GC-content, and the total length of each genome. The
244 Venn diagrams in the inset show the number of overlapping and non-overlapping genes between
245 JB001 and JB002, and JB001 and TM7x. The number in parenthesis is the number of genes with
246 unknown functions (UF). **(B). KEGG pathways encoded by G1 and G6 core genes.** KEGG
247 **m**etabolic map overlaid with the pathways encoded by the pan-Saccharibacteria core genes
248 (black), G1 Core Genes (green), and G6 Core Genes (red), as indicated by the Venn diagram
249 key. Enzymes of interest are labeled with text and arrows. Pathways are indicated by labeled
250 boxes, the cell wall metabolism pathways is labeled with the red background to distinguish it due
251 to the odd shape and overlap with the glycolysis pathway space.

Figure 2

A

Pangenome of complete G1 and G6 genomes



B

KEGG pathways encoded by G1 and G6 core genes