Muribaculaceae genomes assembled from metagenomes suggest genetic drivers of differential response to acarbose treatment in mice

Byron J. Smith[1]

Richard A. Miller[2]

Thomas M. Schmidt[3*]

[1]Gladstone Institutes, San Francisco, CA, USA
[2]Department of Pathology and Geriatrics Center, University of Michigan, Ann Arbor, MI, USA
[3]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA

*Corresponding Author: schmidti@umich.edu

Abstract

The drug acarbose (ACA) is used to treat diabetes, and by inhibiting alpha-amylase in the small intestine increases the amount of starch entering the lower digestive tract. This results in changes to the composition of the microbiota and its fermentation products. Acarbose also increases longevity in mice, an effect that could be related to increased production of the short-chain fatty acids propionate and butyrate. In experiments replicated across three study sites, two distantly related species in the bacterial family *Muribaculaceae* were dramatically more abundant in ACA treated mice, distinguishing these responders from other members of the family. Members of the *Muribaculaceae* likely produce propionate and are abundant and diverse in the guts of mice, although few isolates are available. We reconstructed genomes from metagenomes (MAGs) for eight populations of *Muribaculaceae* to examine what distinguishes species that respond positively to acarbose. We found two closely related MAGs (B1-A and B1-B) from one responsive species that both contain a polysaccharide utilization locus with a predicted extracellular alpha-amylase. These also shared a periplasmic neopullulanase with another, distantly related MAG (B2) representative of the only other responsive species. This gene differentiated these three MAGs from MAGs representative of non-responding species. Differential gene content in B1-A and B1-B may be associated with the inconsistent response of this species to acarbose across study sites. This work

demonstrates the utility of culture-free genomics for inferring the ecological roles of gut bacteria including their response to pharmaceutical perturbations.

Importance

The drug acarbose is used to treat diabetes by preventing the breakdown of starch in the small intestine, resulting in dramatic changes in the abundance of some members of the gut microbiome and its fermentation products. In mice, several of the bacteria that respond most positively are classified in the family *Muribaculaceae*, members of which produce propionate as a primary fermentation product. Propionate has been associated with gut health and increased longevity in mice. We found that genomes of the most responsive *Muribaculaceae* showed signs of specialization for starch fermentation, presumably providing them a competitive advantage in the large intestine of animals consuming acarbose. Comparisons among genomes support existing models for the ecological niches occupied by members of this family. In addition, genes encoding one type of enzyme known to participate in starch breakdown were found in all three genomes from responding species, but none of the others.

Background

The mammalian gut microbiome is a complex ecological system that influences energy balance (1), pathogen resistance (2), and inflammation (3), among other processes with importance to host health. Understanding how the bacterial inhabitants of the gut respond to pharmaceutical and dietary perturbations is a major step in developing a predictive framework for microbiome-based therapies. Acarbose (ACA) is an alpha-glucosidase inhibitor prescribed for the treatment of type 2 diabetes mellitus because it reduces the absorption of glucose from starch in the small intestine (4). In rats, ACA has been shown to increase the amount of starch entering the lower digestive system after a meal (5). ACA treatment also changes the composition of the gut microbiota and its fermentation products in many rodents. (5–14). Interestingly, long-term treatment with ACA has been shown to substantially increase longevity in male mice and to a lesser extent in females (15–17).

Previously we have shown that the relative abundance of a number of bacterial species as well as the concentrations of propionate and butyrate respond to long term treatment with ACA (14). This

study was notable in being replicated across three sites: The University of Michigan (UM) in Ann Arbor, The University of Texas Health Science Center at San Antonio (UT), and The Jackson Laboratory (TJL) in Bar Harbor, Maine. At UM and TJL one highly abundant bacterial species was enriched nearly 4-fold in ACA treated mice. This species, defined at a 97% identity threshold of the 16*S* rRNA gene V4 region and designated as OTU-1, was classified as a member of the family *Muribaculaceae* in order *Bacteroidales*. OTU-1 was also present and abundant at UT but was not significantly more abundant in ACA treated mice relative to controls. Instead, a different *Muribaculaceae* species, designated OTU-4, was found to be highly abundant and 4-fold enriched in ACA-treated mice, but was nearly absent at UM and TJL. Other *Muribaculaceae* were also identified as among the most abundant members of the mouse gut microbiota across the three sites, although none of these were found to be enriched in ACA treatment.

The family *Muribaculaceae*—previously referred to as the S24-7 after an early clone (18, 19), or sometimes as *Candidatus Homeothermaceae* (20)—has only a handful of published cultivars (21–23) despite being a common and abundant inhabitant of the mammalian gut, especially in mice (20). Previous studies have suggested that the *Muribaculaceae* specialize on the fermentation of complex polysaccharides (20), much like members of the genus *Bacteroides*, which is also a member of the order *Bacteroidales*. Genomic analysis has also suggested that the capacity for propionate production is widespread in the family (20).

Recently, techniques have been developed to reconstruct genomes of uncultivated members of bacterial communities (24, 25). Based on 30 such metagenome assembled genomes (MAGs) they reconstructed using this approach, Ormerod and colleagues (20) proposed that the *Muribaculaceae* fall into three distinct carbohydrate utilization guilds, which they describe as specialists on alpha-glucans, plant glycans, and host glycans, respectively. While it is reasonable to expect that alpha-glucan specialists would benefit the most from the large influx of starch to the gut resulting from ACA treatment, this prediction has not been tested, and physiological inferences based on the genome content of members of this clade have been largely divorced from biological observations.

Experimental perturbations of complex microbial communities present an opportunity to observe ecological features of many bacterial taxa without cultivated members and generate hypotheses about their physiology. Given the observed, dramatically increased relative abundance of OTU-1 and OTU-4 (here referred to as “responders”) in mice treated with ACA, we hypothesize that these species are capable of robust growth on starch, while the other *Muribaculaceae* found in the study (“non-responders”), lack the genomic features necessary for the utilization of polysaccharides that

reach the colon in greater quantities following ACA treatment. Alternatively, responders may be resistant to the inhibitory effects of ACA, or benefit from elevated levels of intermediate starch degradation products. Since isolates of the *Muribaculaceae* strains in these mice are not available for characterization, a comparative genomic approach is taken to explore their functional potential.

Most of the research on the genomic components of polysaccharide degradation in gram negative bacteria has been carried out in the genus *Bacteroides*, and in particular *B. thetaiotaomicron* (26). Starch utilization in *B. thetaiotaomicron* is dependent on an ensemble of eight proteins, SusRABCDEFG that enable recognition, binding, hydrolysis, and import of starch and related polysaccharides (27). Homologs of SusC and SusD characterize all known polysaccharide utilization systems in this clade (28), are encoded in Sus-like genomic regions known as polysaccharide utilization loci (PULs), and are widespread in the phylum *Bacteroidetes* (29). The molecular range of these systems is determined by the carbohydrate-active enzymes and structural proteins they encode, based on the specificity of glycoside hydrolase (GH) and carbohydrate binding module (CBM) domains, which have been extensively cataloged in the dbCAN database (30, 31).

Here MAGs from the feces of mice at UT and UM are analyzed to explore two closely related questions about the niche of OTU-1 and OTU-4 in the lower digestive system. First, why do these species each increase in relative abundance with ACA treatment, while other species of *Muribaculaceae* do not? And second, why is the response of OTU-1 site specific? Despite similar patterns of abundance at their respective sites, the two responding species seem to be only distantly related, sharing just 90% of nucleotides in their 16S rRNA gene V4 hypervariable region (14). We nonetheless find genomic evidence that OTU-1 and OTU-4 occupy overlapping niches, specializing in the degradation of alpha-glucans, a role not held by the other *Muribaculaceae* described in this study. In addition, we identify two distinct genomic variants of OTU-1, referred to as B1-A and B1-B, which are differentially distributed between UM and UT and have functionally relevant differences in gene content.

Reconstructing genomes from metagenomes allows for the comparison of the functional potential of *Muribaculaceae* at UM and UT. This work demonstrates the utility of culture-free genomics to understand the potential ecological roles of these key members of the mouse gut microbial community and explore several hypotheses that may explain differences in the distribution and response of these bacteria to ACA treatment. Hypotheses derived from this analysis provide a foundation for future physiological studies in recently obtained cultivars. While a large fraction of

host-associated bacterial species are without isolated representatives (32), let alone characterized (33), combining experimental data from complex communities with the analysis of reconstructed genomes provides a powerful tool for expanding understanding to these understudied taxa.

Results

Recovered population genomes are of high quality and resemble other *Muribaculaceae* genomes

MAGs were constructed for 8 populations classified as members of the family *Muribaculaceae*, including for two species, OTU-1 and OTU-4, previously shown to respond positively to ACA. For OTU-1, two closely related genomic variants were recovered, here designated B1-A and B1-B, possessing 0.63 and 0.36 Mbp of unshared sequence, respectively (Table 2). We designate the MAG constructed for OTU-4 as B2. MAGs obtained from non-responding species are designated B3 through B7. All 8 novel MAGs are estimated to be mostly complete and all had less than 1% estimated contamination based on the recovery of ubiquitous, single-copy genes (Table 1). The median N50 statistic was approximately 71 kbp, suggesting that assembly was suitable for inferring the genomic context of functional genes. Estimated genome sizes, GC%, and number of predicted genes are all similar to previously published MAGs belonging to the family *Muribaculaceae*, as well as the finished genome for *Muribaculum intestinale* strain YL27.

To confirm the assertion that each of the reconstructed genomes is representative of a previously described *Muribaculaceae* species identified in these mice (14), the per-library mapping rates of each genome were compared to the relative abundance of the associated 16S rRNA gene in amplicon libraries. Pearson correlation coefficients between the fraction of reads mapped and species relative abundance were above 0.86 for all MAGs,

***Table 1**: Summary of novel MAGs compared to the genome of Muribaculum intestinale YL27*

Genome / MAG	Completeness[1]	Scaffolds	Length[2]	N50	GC	in Smith et al., 2019	Top nr BLAST Hit (Identity)
YL-27[3]	99%	1	3.3	33,070	50.1%	*n/a*	*n/a*
B1-A	97%	228	3.2	41,412	46.6%	OTU-1	WP_123406077.1 (99.92%)
B1-B	97%	152	3.0	59,916	46.9%	OTU-1	WP_123406077.1 (100%)
B2	98%	65	2.6	79,454	50.5%	OTU-4	WP_128713622.1 (92.52%)
B3	86%	98	2.2	63,818	54.0%	OTU-6	OKY86749.1 (90.84%)
B4	98%	31	2.7	148,039	55.2%	OTU-5	WP_123486179.1 (91.80%)
B5	86%	50	2.5	78,179	55.7%	OTU-8	WP_123486179.1 (92.43%)
B6	99%	110	3.2	87,115	48.3%	OTU-30	WP_123613567.1 (100%)
B7	98%	97	2.5	59,037	53.9%	OTU-39	WP_123541885.1 (100%)

[1] Estimated by CheckM (34)

[2] Total length in Mbp

[3] *Muribaculum intestinale* YL-27 reference genome

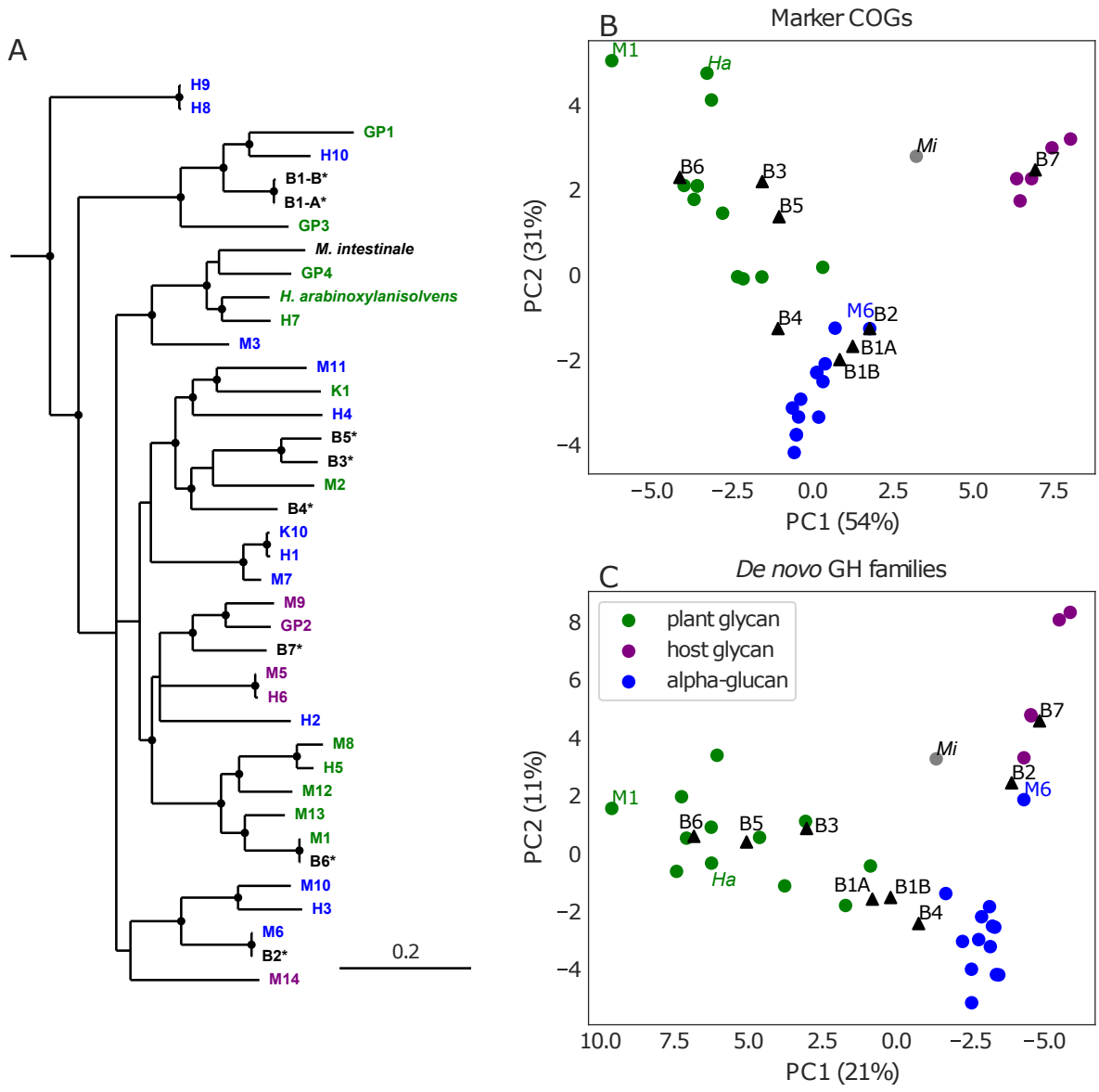

Figure 1*: Comparison of novel and previously described Muribaculaceae genomes. Novel MAGs ("B1-A", "B1-B", and "B2" through "B7") are combined with the finished genome for M. intestinale strain YL27, as well as 30 previously described MAGs that are hypothesized to reflect three polysaccharide utilization guilds: specializing on alpha-glucans (points and labels colored blue), host glycans (violet), and plant glycans (green) (20). (**A**) MAGs described in this study were placed in a phylogenetic context using a maximum-likelihood concatenated gene tree based on an amino acid alignment of 9 shared, single-copy genes, and four other Bacteroidales species as an outgroup (not shown). Nodes with less than 70% confidence are collapsed into polytomies and topological support greater than 95% is indicated (black dots). Branch lengths indicate an estimate of expected substitutions per site. (**B**, **C**)*

*Functional comparisons were visualized by plotting the first two principal components of an ordination on annotation counts of either (**B**) eight COGs previously identified as maximally discriminatory between hypothesized guilds (20), or (**C**) de novo clusters based on sequence similarity of GH domain containing proteins. PCA was performed with the 30 previously constructed MAGs from family Muribaculaceae and the percent of variation described by the first two components is included in the axis labels. All genomes were then projected onto that space. Novel MAGs (black triangles) are labeled, as are the previously described MAGs M1, M6, and the proposed H. arabinoxylanisolvens (Ha), and the finished genome of M. intestinale (Mi, gray circle).*

Phylogenetics

To better understand the evolutionary relationships between these organisms, a concatenated gene tree was constructed to assess relationships among the 8 new MAGs and those previously described (20), and *M. intestinale* YL27. The tree was rooted by four other *Bacteroidales* species: *Bacteroides ovatus* (ATCC-8483), *Bacteroides thetaiotaomicron* VPI-5482, *Porphyromonas gingivalis* (ATCC-33277), and *Barnesiella viscericola* (DSM-18177). Most internal nodes were supported with high topological confidence (>95% bootstrap support), and the placement of the MAGs derived by Ormerod and colleagues was highly consistent with their published tree. To further check the robustness of our phylogeny, a second maximum likelihood tree was constructed based on the *rpoB* gene, which is generally not thought to be transmitted horizontally (despite exceptions (35)), This approach also recapitulated the published topology (results not shown). The estimated phylogeny shows that the newly generated MAGs encompass most of the documented diversity of *Muribaculaceae*. Two of the new MAGs, B2 and B6, appear to be closely related to previous MAGs M6, and M1, respectively (20). Nonetheless, this phylogenetic analysis suggests that the majority of the MAGs derived here had not been described at the time of this study. To further link our MAGs with recently generated MAGs and isolate genomes, we searched translated *rpoB* sequences against the GenBank non-redundant protein database (updated 2020-06-22) using BLASTp. Accession numbers for the closest hits as well as the amino acid identity are reported in Table 1. Despite a growing number of *Muribaculaceae* genomes deposited in public repositories, four of the eight MAGs reconstructed here have *rpoB* genes with less than 93% amino acid identity to their highest scoring match.

Novel protein families

Annotations based on alignment to a database of previously characterized sequences may provide only limited insight into the function of gene products, in particular for genes from largely unstudied families of bacteria. In order to identify previously uncharacterized orthologous groups of genes, *de novo* clustering (36) was carried out based on amino acid similarity of all putative genes found in the 8 novel MAGs, 30 previously generated MAGs, *M. intestinale*, four publicly available draft genomes from the family, and four reference *Bacteroidales* (see Fig. 1). The resulting clusters are referred to as operational protein families (OPFs). While a fraction of the 12,648 resulting OPFs may be due to spurious sequence similarity and without biological relevance, 5,767 had representatives in at least three genomes, increasing the likelihood that these reflect evolutionarily conserved protein sequences. Of these, only 2,404 had members annotated with any COG, KO, or putative function. The remaining 3,363 unannotated OPFs encompass 17,831 predicted protein sequences across the 47 genomes. >

Ordination of gene content

To compare the novel MAGs to other available MAGs and reference genomes, a previous published analysis was recreated, harnessing a set of 8 COGs found by Ormerod and colleagues to maximally differentiate the three hypothesized guilds (20). We replicated the original ordination analysis using our annotation of the publicly available MAGs, and then projected all other genomes onto this same space (see Fig. 1). Newly available genomes were compared to the three clusters hypothesized to represent specialization on alpha-glucans, plant glycans, and host glycans. While the 8 novel MAGs inhabit approximately the same volume as those in the original analysis, and some could be plausibly classified based on these criteria, the ambiguous placement of B4 and *M. intestinale* suggests that new genomes will present additional exceptions to the three-guild model.

It is notable that MAGs from responding species cluster with the proposed alpha-glucan guild, consistent with a functional potential for starch utilization absent in the non-responders. To expand on this descriptive analysis and to leverage the more comprehensive view provided by *de novo* clustering to explore differences and similarities in carbohydrate utilization potential, a second ordination of genomes was performed, this time based on OPF labels of predicted genes found to contain GH domains (Fig. 1). Similar to the previous ordination based on COGs, three groups of genomes approximately reflecting those proposed by Ormerod and colleagues are apparent.

However, the placement of B2 (as well as the closely related M6) relative to the proposed guilds are substantially different.

Analysis of MAGs from species responsive to ACA treatment suggests genes involved in starch utilization.

Based on the characterization of genes and genomic regions with a role in starch utilization in the closely related genus *Bacteroides*, it is plausible that alpha-amylase localized to the outer membrane may be common to starch utilizing bacteria in the order *Bacteroidales* (37). Indeed, B1-A and B1-B both have three OM-localized genes predicted to code for GH13 containing lipoproteins (B1A280, B1A301, B1A333 in B1-A and matching genes in B1-B), each in a separate PUL (see Fig. 2). While it also includes members without this activity, GH13 is the main family of alpha-amylases (38). These genomic regions also possess additional genes with carbohydrate-active domains that are expected to interact with alpha-glucans.

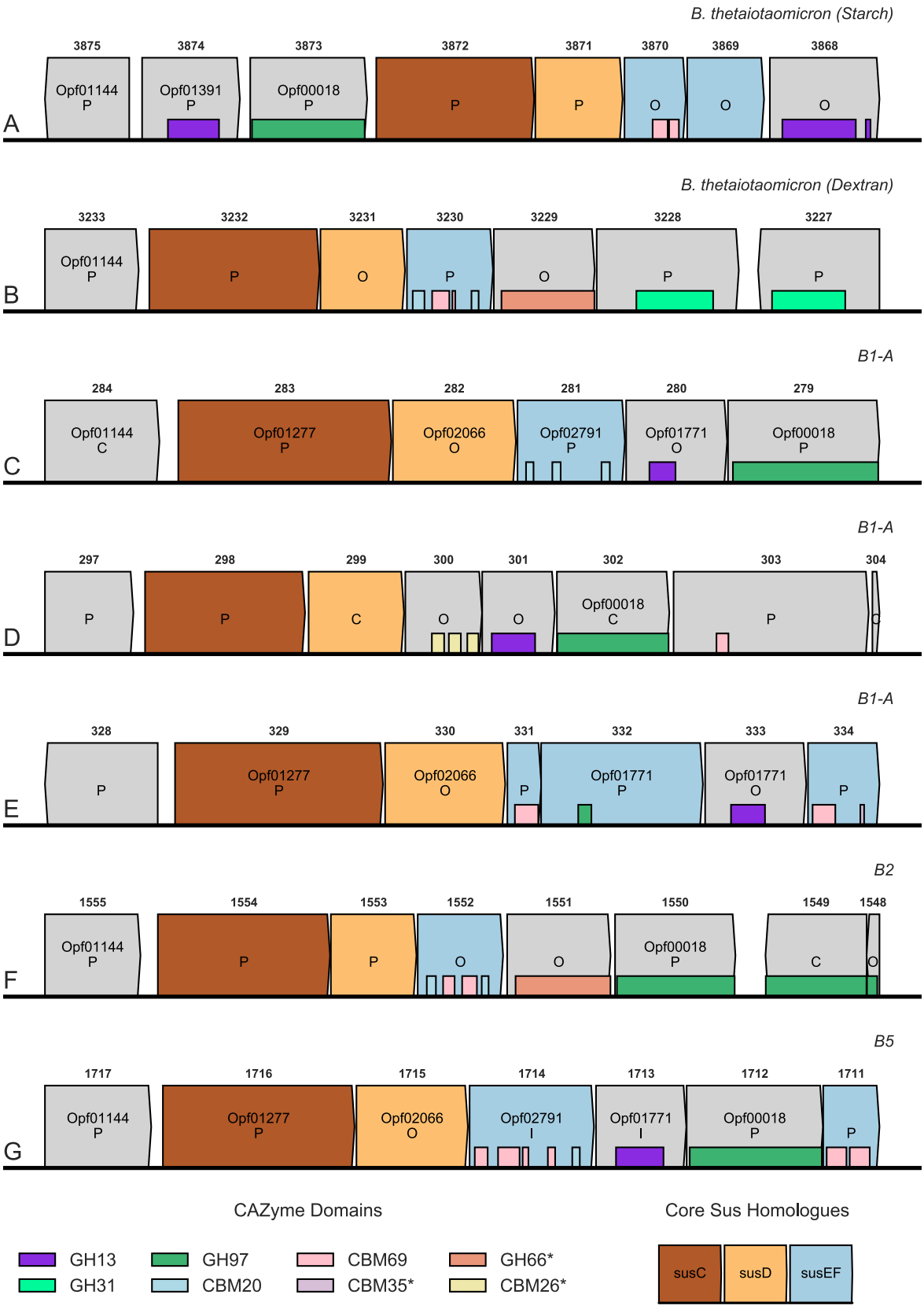
B. thetaiotaomicron (Starch)
3875
3874
3873
3872
3871
3870
3869
3868
Opf01144
P
Opf01391
P
Opf00018
P
P
P
O
O
O
A
B. thetaiotaomicron (Dextran)
3233
3232
3231
3230
3229
3228
3227
Opf01144
P
P
O
P
O
P
P
B
B1-A
284
283
282
281
280
279
Opf01144
C
Opf01277
P
Opf02066
O
Opf02791
P
Opf01771
O
Opf00018
P
C
B1-A
297
298
299
300
301
302
303
304
P
P
C
O
O
Opf00018
C
P
C
D
B1-A
328
329
330
331
332
333
334
P
Opf01277
P
Opf02066
O
P
Opf01771
P
Opf01771
O
P
E
B2
1555
1554
1553
1552
1551
1550
1549
1548
Opf01144
P
P
P
O
O
Opf00018
P
C
O
F
B5
1717
1716
1715
1714
1713
1712
1711
Opf01144
P
Opf01277
P
Opf02066
O
Opf02791
I
Opf01771
I
Opf00018
P
P
G
CAZyme Domains
GH13
GH31
GH97
CBM20
CBM69
CBM35*
GH66*
CBM26*
Core Sus Homologues
susC
susD
susEF

Figure 2*: Polysaccharide utilization loci in Bacteroidales. Diagrams of the Sus operon (**A**) and the dextran associated PUL (**B**) of B. thetaiotaomicron VPI-5482 along with five putative starch-associated PULs identified in Muribaculaceae MAGs B1-A and B4 (**C**-**G**). B1-A PULs shown here are syntenic in B1-B (not shown). Predicted protein coding sequences are shown as boxes pointed in the direction of transcription. Homology to SusC, SusD, and SusEF is indicated. Protein regions with homology to starch-associated GHs, as well as GH66, and CBMs are shown as shallow rectangles, and are colored as indicated in the legend. Several OPFs are noted with members in multiple genomes, including clusters that contain SusR (Opf01144), SusA (Opf01391), and SusB (Opf00018). The inferred localization of each protein product is also indicated: cytoplasmic (genes labeled C), periplasmic (**P**), outer membrane (**O**), or inner membrane (**I**).*

Besides B1-A and B1-B, B5 is the only other MAG to possess a putative PUL coding for a full complement of predicted starch-active proteins. Several of the OPF annotations found in these presumptive starch PULs are shared by *B. thetaiotaomicron*, suggesting shared function. This set including SusC-homologs Opf01277, Opf02066, which includes relatives of SusD, and Opf02791 whose members possess CBM20 starch-binding domains. However, while B5 also has a GH13 containing lipoprotein (B51713), its predicted localization is on the inner membrane. It is unclear whether this explains B5's non-response in ACA-treated mice. Plausible OM-localized, GH13 containing proteins are not found in any non-responders. While this characteristic does not seem to perfectly discriminate responder from non-responders—B2 also lacks such a gene—it nonetheless demonstrates concordance between inferred genomic features and observed population dynamics of the corresponding species.

Despite the absence of a GH13 domain on the outer-membrane, it is plausible that B2 is capable of degrading starch using other enzymatic machinery. We speculate about one putative locus (see Fig. 2F), which has a similar gene content to characterized (39–41) dextran PULs in *B. thetaiotaomicron* and *B. ovatus*.

To expand the search for relevant genetic features, *de novo* protein clusters were filtered to those with members in B1-A, B1-B, and B2. Of these OPFs, several stood out as particularly relevant. Opf01144 includes SusR, the regulator of transcription of the starch utilization system in *B. thetaiotaomicron*, as well as its homolog in *B. ovatus*. It is an apparent subcluster of the larger family defined by K21557, and in many cases is encoded directly upstream of *susC* in putative PULs that are considered likely to have affinity for alpha-glucans. In B1-A and B1-B, two of the three putative starch PULs encode a member of Opf01144, and it is similarly located in PULs with starch-active

CBM and GH domains in B2 and B5. In addition, of the seven MAGs constructed by Ormerod *et al.* that encode a member of this cluster, five of them are classified to the alpha-glucan guild. It is plausible that members of Opf01144 share a functional role regulating transcriptional responses to alpha-glucans.

Opf01391, which recapitulates K21575, includes SusA: the periplasmic neopullulanase of *B. thetaiotaomicron* and an important component of starch utilization in that organism (42). This family is found in the MAGs associated with responders, B1-A, B1-B, and B2, and none of the other MAGs generated in this study. What's more, it's found in twelve of the thirteen alpha-glucan and a minority of the plant glycan guild members. Interestingly, although it is encoded by the Sus operon in *B. thetaiotaomicron* and its homologous locus in *B. ovatus*, in the *Muribaculaceae* members of Opf01391 do not in general appear to be encoded in PULs.

Unshared gene content in B1-A and B1-B

Two distinct genomic variants were associated with OTU-1 with one found in a majority of the UT mouse metagenomes, and the other ubiquitous at UM. Using the nucmer tool for genome alignment (43), 19.6% of the B1-A MAG sequence and 12.2% of B1-B were found to not align to the other. While these hundreds of kbp may in part reflect errors in genome recovery, much of the unaligned length suggests differences in gene content between these populations of OTU-1. This observation was confirmed by assessing the mapping of metagenomic reads against predicted protein coding genes in each variant. For each pairing of metagenomic read library to genomic variant, gene coverage was normalized by the median gene coverage in order to identify genes with conspicuously fewer reads in particular subsets of the mice. Metagenomic libraries manually chosen as unambiguous representatives of either B1-A or B1-B were used to systematically identify genes differentiating the two. The median normalized mapping depths in each set of libraries against predicted genes in each MAG were compared, providing a measure of the relative enrichment or depletion of genomic sequences between the two populations of OTU-1. Libraries have low coverage of large portions of either the B1-A or B1-B MAG (see Fig. 3), suggesting that mice are primarily inhabited by one of the two variants, and that a portion of genes are variant specific.

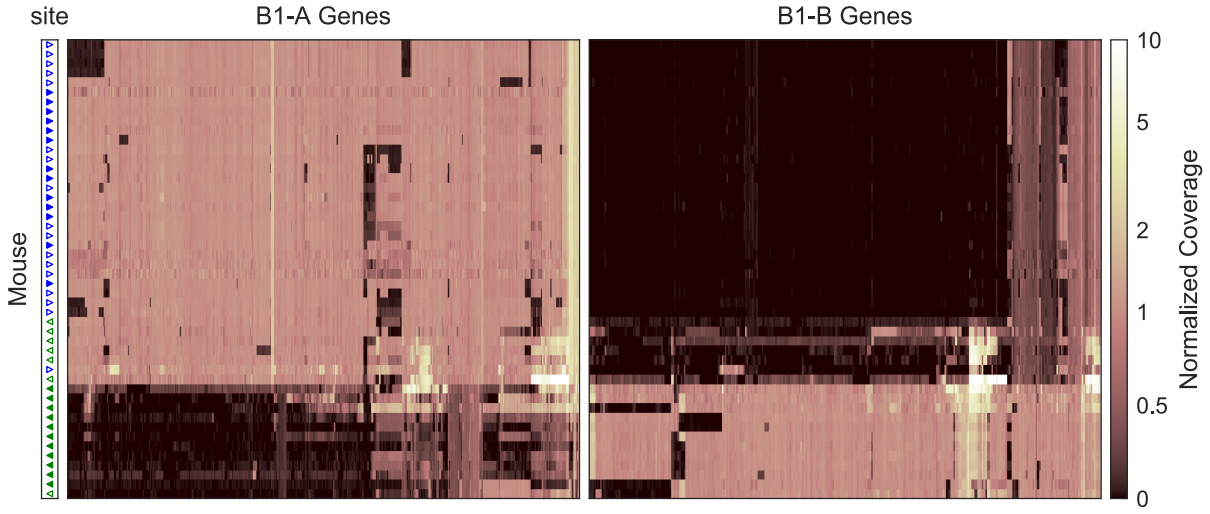

Figure 3*: Visualization of differential gene content in two OTU-1 populations. Heatmaps depict mapping coverage of metagenomes against putative protein coding genes in MAGs B1-A or B1-B normalized to the median coverage. Rows represent one or more pooled libraries for each mouse included in the study and columns represent individual genes. The site at which each mouse was housed is indicated by triangles in the far left column: UT (green, left pointing) or UM (blue, right). Filled triangles correspond to those mice representative of just B1-A or just B1-B (not a mixture) flagged for downstream analysis. Genes are shown only where the median normalized coverage ratio between these B1-A and B1-B specific metagenomes is greater than 1.5. Rows and columns are arbitrarily ordered to maximize visual distinction between variants.*

This analysis found 12.8% of predicted genes in B1-A were depleted at least 5-fold in B1-B populations, and 12.4% the reverse. While this observed depletion could indicate variation in copy number, differential gene content between variants is a more parsimonious explanation for most loci. These predicted genes reflect 2.7% of unique KOs in B1-A and 1.9% in B1-B. Interestingly, the fraction of variant specific OPFs is greater, 7.5% and 7.1% respectively, suggesting that *de novo* clustering could be more sensitive to potential differences in physiology.

***Table 2**: Summary of variant specific features in two highly similar MAGs*

	B1-A		B1-B	
	Total	Specific	Total	Specific
Nucleotide length[1]	3.23	0.63	2.96	0.36
Genes	2,710	348	2,496	309
OPFs[2]	2,308	173	2,202	157
KOs[2]	1,056	29	1,033	20
COGs[2]	716	8	709	3

[1] in Mbp

[2] unique

Given the observation that the relative abundance of OTU-1 was dramatically increased with ACA treatment at UM, while not being significantly affected at UT, and that B1-B was not found in metagenomes at UM, we hypothesized that differences in the genomic potential of B1-A and B1-B could explain the different response to ACA at the two sites.

Genomic regions apparently specific to B1-A—defined as an at least 5-fold enrichment in B1-A specific libraries relative to B1-B specific libraries—include just one PUL (SusC-homolog encoded by B1A00048). This locus includes a predicted outer membrane localized GH30-containing protein. Proteins that contain a GH30 domain have beta-glucosylceramidase, beta-1,6-glucanase, or beta-xylosidase activity (44). Given that this PUL also encodes a periplasmic, GH3 containing protein, it appears to be unlikely that it has specificity for starch. The B1-A also possesses numerous phage insertions not seen in B1-B. Conversely, a CRISPR operon including 25 repeat units (Cas9 encoded by B1B01367) appears to be specific to B1-B.

Most strikingly, a 16 kbp region (from B1A01498 to B1A01514) specific to B1-A was found to contain several genes with homology to cell capsule and exopolysaccharide synthesizing enzymes. Based on annotations with KEGG orthologous groups, these include homologs of *tuaG* (K16698), *tagE* (K00712), *gmhB* (K03273), *gmhA*/*lpcA* (K03271), *hddA* (K07031), *exoO* (K16555), *waaH* (K19354), and *tagF* (K09809). Interestingly, the B1-B MAG contains a different such region of about 6.5 kbp (B1B00851 to B1B00856) with *wfeD* (K21364), *pglJ* (K17248), and *epsH* (K19425). For

each, several of the OPFs in the respective regions were not found anywhere in the opposing genome, suggesting that the makeup of each variant's exterior surface might be distinctly different.

Discussion

Mice are a key model system for study of the mammalian gut microbiome, with an outsized importance in testing mechanistic hypotheses for the roles of this community in host health (45). The generalizability of observations made in mice is a constant concern (45), in part due to extensive difference in taxonomic composition compared to humans (21). Bacteria classified in the family *Muribaculaceae* are abundant in the murine gut microbiome (20). While these bacteria are also found in humans (although at lower abundance), only a few members of this clade have been cultivated and described (21–23). As a result, the ecological roles of these bacteria have not yet been characterized, limiting the value of the mouse as a model system. Better understanding the ecology of *Muribaculaceae* in the murine gut will increase the transferability of microbiome findings from mice to humans. Attempts to study these organisms make use of genomes reconstructed from metagenomic reads, and have suggested—in the absence of experimental data—that members of the family consume a diversity of polysaccharides in the lower gut.

Here we have extended that approach to eight new genomes, and associated those with species for which changes in relative abundance in response to ACA treatment have been experimentally assessed. This enabled us to explore why two responding species, represented by MAGs B1-A, B1-B, and B2, increase with ACA treatment, while the other species of *Muribaculaceae* do not. Annotations of reconstructed genomes suggest that the responders may possess starch degradation capabilities absent in the non-responders.

We examine the three-guild model proposed by Ormerod and colleagues (20) by reproducing their dimensional reduction approach with the addition of these new genomes. In this analysis, annotations of B1-A, B1-B, and B2 are consistent with a hypothesized guild of alpha-glucan degrading species, supporting their interpretation. A more nuanced approach to annotation was also applied by constructing *de novo* clusters of proteins based on homology. Interestingly, this analysis indicates that B2, and the closely related M6, share physiological potential with MAGs in the host-glycan guild, suggesting that a more detailed examination can identify specific functions that discriminate responders from non-responders. This approach is bolstered by the phylogenetic

and genomic distinction between B2 and both B1-A and B1-B, reducing the confounding effects of shared evolutionary history.

By including otherwise unannotated genes, genomic comparisons based on OPFs may reflect shared functional potential better than applying previously defined orthologies. Besides the identification of potentially novel gene families, *de novo* homology clustering (36) also enables differentiation of sub-groups not captured by standard annotations. For instance, hypothetical genes annotated as homologs of SusC, SusD, and SusEF, were clustered into 119, 162, and 33 different OPFs respectively. It is plausible that this sub-clustering captures differences in protein structure with importance in oligo- and polysaccharide recognition, import, and binding. Combined with annotation of characterized functional domains, these clusters more narrowly predict the polysaccharide utilization ranges of uncultured organisms. Testing these predictions will require characterization of the metabolic potential of these genes after obtaining cultivars or through heterologous expression in appropriate hosts.

A detailed analysis of PULs identified multiple loci shared in both B1-A and B1-B that appear to be adapted to the degradation of starch or related carbohydrates, due to the presence of an OM localized GH13 containing protein (46). Counterintuitively, B2 had no such PUL, suggesting that its response to ACA may result from other enzymatic capabilities. Of particular interest is a PUL encoding proteins with GH97, CBM20, and CBM69 domains, all of which have documented activity on starch (47, 48). While the only outer-membrane localized hydrolase in this PUL is a GH66, and members of this family have characterized activity on the alpha-1,6 linkages between glucose monomers in dextran (49), it is plausible that this PUL can be repurposed and confers some ability to grow on starch.

In addition, a gene encoding a SusA homolog was identified in B1-A, B1-B, and B2 but in none of the non-responders. While it is unclear how expression of this important component of starch utilization might be regulated, given that it is not located in a PUL in any of the responding populations, SusA is important for growth on amylopectin in *B. thetaiotaomicron* (42). Since inhibition by ACA is variable across enzymes (50), it is possible that ACA treatment results in elevated production of dextrin and maltooligosaccharides in the lower guts of mice due to residual alpha-amylase activity, even at levels sufficient to prohibit host digestion. Periplasmic hydrolysis of these starch breakdown products may be sufficient for increased abundance of these species in ACA treated mice.

It is notable that of the closely related variants, B1-A and B1-B associated with OTU-1, B1-B is found at UT and not UM. We previously observed site-specificity of the ACA response of this species, in which OTU-1 did not have a significantly increased abundance in treated mice at UT, while it was the most dramatic change at UM. Differences in the functional potential due to differences in gene content of populations found at each of the sites is one possible explanation for this pattern. Intriguingly, while we do not conjecture a mechanistic link, an ACA-by-site interaction effect on longevity has been previously observed in the mouse colonies sampled here, with male mice at UT showing a larger increase in median longevity with ACA treatment than those at UM (15, 17).

Despite evidence that large differences in gene content can be found between even closely related populations (51, 52), studies reconstructing genomes from metagenomes have just started to consider these pangenome dynamics (53–56). The discovery of two populations of OTU-1 therefore demonstrates the value of considering pangenome dynamics, and presents a potential explanation for the observed site-specific response of that taxon. The finding that both variants have the same complement of three PULs apparently specializing in starch utilization and the same SusA homolog does not support the hypothesis that differences in starch utilization potential account for these abundance patterns. We did, however, identify numerous differences in the gene content of B1-A and B1-B, including variant specific loci that may influence the structure and function of the outer surface of the cell. Capsule variation is known to greatly affect both ecological and host interactions (57).

While these results do not establish a mechanistic explanation for differences in the response of B1-A, B1-B at UM and UT, nor conclusively identify the pathways that enable starch utilization in B2, they do suggest a number of genomic features that likely contribute to previously observed patterns in taxon abundance. Future studies utilizing metatranscriptomic analysis might demonstrate active expression of these genes, or differential expression in mice treated with acarbose compared to controls. Likewise, even in the absence of a B2 cultivar, the potential role of its dextran PUL in enrichment under acarbose treatment could be tested using available cultivars, like *B. thetaiotaomicron*, that posses a homologous gene cluster.

Conclusions

In this study we have reconstructed and described genomes representing 7 species in the family *Muribaculaceae* from the mouse fecal microbiome, and have found features that differentiate those

bacterial species that respond positively to ACA treatment from those that do not. This analysis suggests that utilization of starch and related polysaccharides enables increased population size in mice treated with ACA—an alpha-amylase inhibitor. In addition, two distinct genomic variants of one species were identified that differ in functional gene content, potentially explaining site-specific differences in response. By combining observed changes in relative abundance during experimental manipulation with inferred functional gene content, we are able to study mammalian symbionts in the absence of cultured representatives. This sequence-based approach is broadly applicable in microbial ecology and enables improved understanding of *in situ* dynamics within complex microbial communities.

Methods

Mouse treatment, sample collection, extraction and sequencing

Mice were bred, housed, and treated as described in (15). Briefly, genetically heterogeneous UM-HET3 mice at each study site were produced by the four-way cross between (BALB/cByJ x C57BL/6J) F1 mothers and (C3H/HeJ x DBA.2J) F1 fathers, as detailed in (58). Mice were fed LabDiet 5LG6 (TestDiet Inc.) from weaning onwards. Starting at 8 months of age, mice randomly assigned to treatment were fed chow with 1,000 ppm ACA (Spectrum Chemical Manufacturing Corporation). Mice were housed 3 males or 4 females to a cage. Colonies were assessed for infectious agents every 3 months, and all tests were negative.

Individual fecal pellets were collected from a single mouse per cage. 16S rRNA gene libraries and metabolite analyses of these samples are as described previously (14). From this collection, a subset of samples were non-randomly selected for metagenomic sequencing in order to test several unrelated hypotheses about SCFA production. Samples were from 54 mice, with at least six treated and control representatives of both males and females at each site.

Fecal samples were slurried with nuclease free water at a 1:10 (w/v) ratio, and most samples were spiked with *Sphingopyxis alaskensis* RB2256 prepared as described previously (14) before DNA extraction and sequencing. Based on alignment to the reference genome, sequenced reads from *S. alaskensis* can be distinguished from all endogenous bacteria in mouse feces. This spike was added as an internal standard to quantify total 16S rRNA gene abundance, and also provides a benchmark for the reconstruction of bacterial genomes. A small number of these were split for both spiked and

unspiked samples, which we used to validate this procedure. For each, 150 uL of this sample was transferred for extraction using the MoBio PowerMag Microbiome kit. Metagenomic libraries were prepared using standard procedures sequenced on the Illumina HiSeq 400 platform using the v4 paired-end 2x150 bp.

Assembly, binning, and MAG refinement

Raw metagenomic reads were deduplicated using FastUniq (59), adapters trimmed using Scythe (60), and quality trimmed using Sickle (61) to produce processed reads for all downstream analyses. The resulting paired-end reads were assembled into primary contigs using MEGAHIT (62). Reads were then mapped back to these contigs with Bowtie2 (63), and per-library coverage was estimated for each contig.

For all contigs >1000 bp in length, dimensional reductions built into CONCOCT (64) were applied to produce input data for a Gaussian mixture model (GMM) similar to the procedure used by that program for binning. However, unlike CONCOCT—due to computational limitations—the model was trained on just 10% of the input data, sampled randomly, before assigning bins to all contig. While this may have reduced the accuracy of the binning procedure, we believe that subsequent refinement steps mitigated the impact of this decision.

OTUs were classified taxonomically and relative abundance was calculated for matched libraries as described in (14). Bins were then recruited to one or more OTUs by calculating a Canonical partial least squares between OTU abundance and bin coverage as implemented in the scikit-learn machine learning library for Python (65). For bins recruited to OTUs classified as *Muribaculaceae*, contigs were re-clustered based on coverage across samples. First “trusted contigs” were manually selected which correlated closely with OTU abundance. The mean coverage of these was used to normalize the per-library coverage of all other contigs. Then, using a GMM, groups of contigs were clustered such that the normalized coverage across samples was consistent. These groups were used to inform the manual assignment of contigs to MAGs. Libraries in which MAGs had non-negligible coverage were identified and used in subsequent refinements. While clustering contigs associated with OTU-1 a number of groups containing on the order of 10^5 bp were found with a bimodal coverage distribution, low normalized coverage in a subset of libraries, and normal coverage in others. By this criterion, contigs in these “variable” groups were partitioned into two MAG variants, A and B, with the non-variable contig groups shared by both. To avoid challenges associated with admixture, only libraries that appeared on further inspection to have just one of the

two variants were considered in downstream refinement steps. The mice matching these libraries are highlighted in Fig. 3. Genomic variants were not found associated with any of the other *Muribaculaceae* OTUs described in this study.

For each MAG, several alternative refinement procedures were performed from which the best quality result was selected. Reads mapping to the curated contigs were digitally normalized (66–68) and reassembled with SPAdes (69). This reassembly as well as the original contigs were cleaned using a single pass of the Pilon assembly refinement tool (70). Finally, the per-library mapping depths of each position in these assemblies were compared to the average mapping depth of the "trusted contigs" selected earlier, and regions with low cosine similarity were excised from the final assemblies.

Genome completeness and contamination estimates were calculated based on ubiquitous single-copy genes using the program CheckM (34). Based on these results, the final assembly with the highest completeness and with contamination < 1% was selected from the various refinements.

Reference genomes

The *Muribaculum intestinale* genome sequence was obtained from GenBank (accession GCA002201515.1), as well as four additional draft genomes (GCA003024805.1, GCA003024815.1, GCA002633305.1, GCA002633115.1) enabling comparison of MAGs to cultured isolates. While other genomes labeled as *Muribaculaceae* had also been deposited at the time of this work, they were excluded from this analysis due to redundancy or apparent misidentification to the family. More recently, several other isolate genomes have become available but have also not been included here. Thirty previously constructed MAGs (20) were obtained from the SRA. For comparison, nucleotide sequences for *B. thetaiotaomicron* VPI-5482 (AE015928.1), *B. ovatus* (CP012938.1), *Barnesiella viscericola* (GCA000512915.1), and *Porphyromonas gingivalis* (GCA000010505.1), were also downloaded from GenBank.

Genome annotation

All genomes were initially annotated with Prokka (71) version 1.13, which uses Prodigal (72) for gene finding. Putative protein sequences were additionally annotated with domains from both the dbCAN database (30) release 6 of carbohydrate-active domains and Pfam (73) release 31.0, using HMMER3 (74, 75) version 3.1b2. Protein sequences were also annotated with KO numbers by

BLAST using the KEGG database as of March 2018 as the reference and taking the best hit with a maximum E-value of 1e-10.

Lipoproteins were predicted using LipoP (76) (version 1.0a) and a score cutoff of 5 and a margin cutoff of 2. Lipoproteins with an arginine at position +2 relative to the cleavage site were labeled as localized to the inner membrane. Periplasmic proteins were identified with SignalP (77) (version 4.1). Predicted protein sequences from all annotated genomes were locally all-by-all aligned using the DIAMOND implementation of the BLAST algorithm (78). Each pair was then assigned a similarity value as the bitscore of their best local alignment normalized by the greater of the two self-alignments. This results in a matrix of pairwise scores reflecting the proximity to perfect homology. Scores less than 0.2 were replaced with 0. Clusters were formed using the MCL algorithm (79) with an inflation parameter of 5.

SusCDEF homologs were identified based on relatively relaxed criteria, harnessing OPF assignments, domain predictions, and Prokka annotations to avoid false negatives while maintaining specificity. For each OPF, all KOs assigned to members were collected as plausible KOs for the cluster. Protein sequences in OPF clusters which included K21572 were flagged as putative SusC-homologs, as were sequences directly annotated as such by Prokka. Using a similar approach, proteins in clusters tagged with K21571 or with any of domains PF12771, PF14322, PF12741, PF07980 were identified as putative SusD. Proteins in clusters tagged with K21571, or with either PF14292 or PF16411, were considered SusEF homologs. PULs were identified by a SusC-homolog with its start codon within 5 kbp of a SusD-homolog's start on the same strand. Manual inspection supported the vast majority of these identifications.

Phylogenetics

Predicted amino acid sequences for ORFs from all MAGs and all reference genomes were search for homology to TIGRFAM protein clusters (80) using `hmmsearch` in the HMMER3 software package version 3.2.1 (74). Hits were filtered at the "trusted-cutoff" score threshold defined separately for each protein model. Sequences found in just one copy in every genome were used as taxonomic marker genes for phylogenetic analysis. Marker gene sequences were aligned to their respective models using `hmmalign` dropping unaligned residues. Aligned markers were then concatenated for each MAG and reference genome, and an approximate maximum likelihood phylogeny was estimated using the FastTree software version 2.1.10 (81) with the default parameters.

Acknowledgments

We would like to acknowledge the technical support from Randy Strong and others at The University of Texas Health Science Center at San Antonio for initial sample collection. We would also like to thank Nicole Koropatkin for feedback on the manuscript.

Funding

Funding for this work was provided by the Glenn Foundation for Medical Research.

Data Availability

Metagenomic libraries will be uploaded to the short read archive, and the SRA Accession will be added in an upcoming version of this manuscript. All code and additional data/metadata needed to reproduce this analysis are available at https://github.com/bsmith89/longev-mgen.

References

1. Turnbaugh PJ, Ley RE, Mahowald MA. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444 doi:10.1038/nature05414.

2. Britton RA, Young VB. 2012. Interaction between the intestinal microbiota and host in *Clostridium difficile* colonization resistance. Trends in Microbiology 20:313–9 doi:10.1016/j.tim.2012.04.001.

3. Syal G, Kashani A, Shih DQ. 2018. Fecal microbiota transplantation in inflammatory bowel disease—a primer for the internists. The American Journal of Medicine doi:10.1016/j.amjmed.2018.03.010.

4. Hiele M, Ghoos Y, Rutgeerts P, Vantrappen G. 1992. Effects of acarbose on starch hydrolysis. Digestive Diseases and Sciences 37:1057–1064 doi:10.1007/BF01300287.

5. Dehghan-Kooshkghazi M, Mathers JC. 2004. Starch digestion, large-bowel fermentation and intestinal mucosal cell proliferation in rats treated with the α-glucosidase inhibitor acarbose. British Journal of Nutrition 91:357 doi:10.1079/BJN20031063.

6. Zhao L, Zhang F, Ding X, Wu G, Lam YY, Wang X, Fu H, Xue X, Lu C, Ma J, Yu L, Xu C, Ren Z, Xu Y, Xu S, Shen H, Zhu X, Shi Y, Shen Q, Dong W, Liu R, Ling Y, Zeng Y, Wang X, Zhang Q, Wang J, Wang L, Wu Y, Zeng B, Wei H, Zhang M, Peng Y, Zhang C. 2018. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science 359:1151–1156 doi:10.1126/science.aao5774.

7. Holt PR, Atillasoy E, Lindenbaum J, Ho SB, Lupton JR, McMahon D, Moss SF. 1996. Effects of acarbose on fecal nutrients, colonic pH, and short-chain fatty acids and rectal proliferative indices. Metabolism: Clinical and Experimental 45:1179–1187 doi:10.1016/S0026-0495(96)90020-7.

8. Wolever TMS, Chiasson JL. 2000. Acarbose raises serum butyrate in human subjects with impaired glucose tolerance. British Journal of Nutrition 84:57–61 doi:10.1017/S0007114500001239.

9. Weaver GA, Tangel CT, Krause JA, Parfitt MM, Jenkins PL, Rader JM, Lewis BA, Miller TL, Wolin MJ. 1997. Acarbose enhances human colonic butyrate production. The Journal of Nutrition 127:717–723 doi:10.1093/jn/127.5.717.

10. Weaver GA, Tangel CT, Krause JA, Parfitt MM, Stragand JJ, Jenkins PL, Erb TA, Davidson RH, Alpern HD, Guiney WB, Higgins PJ. 2000. Biomarkers of human colonic cell growth are influenced differently by a history of colonic neoplasia and the consumption of acarbose. The Journal of Nutrition 130:2718–25. doi:10.1093/jn/130.11.2718.

11. Wolin MJ, Miller TL, Yerry S, Bank S, Weaver GA, Zhang Y. 1999. Changes of fermentation pathways of fecal microbial communities associated with a drug treatment that increases dietary starch in the human colon changes of fermentation pathways of fecal microbial communities associated with a drug treatment that increases dietary starch in the human colon. Applied and Environmental Microbiology 65:2807–2812 doi:10.1128/AEM.65.7.2807-2812.1999.

12. Zhang X, Fang Z, Zhang C, Xia H, Jie Z, Han X, Chen Y, Ji L. 2017. Effects of acarbose on the gut microbiota of prediabetic patients: A randomized, double-blind, controlled crossover trial. Diabetes Therapy 8:293–307 doi:10.1007/s13300-017-0226-y.

13. Baxter NT, Lesniak NA, Sinani H, Schloss PD, Koropatkin NM. 2019. The glucoamylase inhibitor acarbose has a diet-dependent and reversible effect on the murine gut microbiome. mSphere 4:1–12 doi:10.1128/msphere.00528-18.

14. Smith BJ, Miller RA, Ericsson AC, Harrison DC, Strong R, Schmidt TM. 2019. Changes in the gut microbiome and fermentation products concurrent with enhanced longevity in acarbose-treated mice. BMC Microbiology 19:130 doi:10.1186/s12866-019-1494-7.

15. Harrison DE, Strong R, Allison DB, Ames BN, Astle CM, Atamna H, Fernandez E, Flurkey K, Javors MA, Nadon NL, Nelson JF, Pletcher S, Simpkins JW, Smith DL, Wilkinson JE, Miller RA. 2014. Acarbose, 17-α-estradiol, and nordihydroguaiaretic acid extend mouse lifespan preferentially in males. Aging Cell 13:273–282 doi:10.1111/acel.12170.

16. Strong R, Miller RA, Antebi A, Astle CM, Bogue M, Denzel MS, Fernandez E, Flurkey K, Hamilton KL, Lamming DW, Javors MA, de Magalhães JP, Martinez PA, McCord JM, Miller BF, Müller M, Nelson JF, Ndukum J, Rainger GE, Richardson A, Sabatini DM, Salmon AB, Simpkins JW, Steegenga WT, Nadon NL, Harrison DE. 2016. Longer lifespan in male mice treated with a weakly estrogenic agonist, an antioxidant, an α-glucosidase inhibitor or a Nrf2-inducer. Aging Cell 15:872–884 doi:10.1111/acel.12496.

17. Harrison DE, Strong R, Alavez S, Astle CM, DiGiovanni J, Fernandez E, Flurkey K, Garratt M, Gelfond JAL, Javors MA, Levi M, Lithgow GJ, Macchiarini F, Nelson JF, Sukoff Rizzo SJ, Slaga TJ, Stearns T, Wilkinson JE, Miller RA. 2019. Acarbose improves health and lifespan in aging HET3 mice. Aging Cell 18:e12898 doi:10.1111/acel.12898.

18. 2001. Uncultured bacterium partial 16S rRNA gene, clone S24-7 http://www.ncbi.nlm.nih.gov/nuccore/AJ400263.1.

19. Salzman NH, de Jong H, Paterson Y, Harmsen HJM, Welling GW, Bos NA. 2002. Analysis of 16S libraries of mouse gastrointestinal microflora reveals a large new group of mouse intestinal bacteria. Microbiology 148:3651–3660 doi:10.1099/00221287-148-11-3651.

20. Ormerod KL, Wood DLA, Lachner N, Gellatly SL, Daly JN, Parsons JD, Dal'Molin CGO, Palfreyman RW, Nielsen LK, Cooper MA, Morrison M, Hansbro PM, Hugenholtz P. 2016. Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. Microbiome 4:36 doi:10.1186/s40168-016-0181-2.

21. Lagkouvardos I, Pukall R, Abt B, Foesel BU, Meier-Kolthoff JP, Kumar N, Bresciani A, Martínez I, Just S, Ziegler C, Brugiroux S, Garzetti D, Wenning M, Bui TPN, Wang J, Hugenholtz F, Plugge CM, Peterson DA, Hornef MW, Baines JF, Smidt H, Walter J, Kristiansen K, Nielsen HB, Haller D, Overmann J, Stecher B, Clavel T. 2016. The Mouse Intestinal Bacterial Collection (miBC) provides

host-specific insight into cultured diversity and functional potential of the gut microbiota. Nature Microbiology 1:16131 doi:10.1038/nmicrobiol.2016.131.

22. Lagkouvardos I, Lesker TR, Hitch TCA, Gálvez EJC, Smit N, Neuhaus K, Wang J, Baines JF, Abt B, Stecher B, Overmann J, Strowig T, Clavel T. 2019. Sequence and cultivation study of *Muribaculaceae* reveals novel species, host preference, and functional potential of this yet undescribed family. Microbiome 7:28 doi:10.1186/s40168-019-0637-2.

23. Miyake S, Ding Y, Soh M, Low A, Seedorf H. 2020. Cultivation and description of *Duncaniella dubosii sp. nov.*, *Duncaniella freteri sp. nov.* and emended description of the species *Duncaniella muris*. International Journal of Systematic and Evolutionary Microbiology doi:10.1099/ijsem.0.004137.

24. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology 2:1533–1542 doi:10.1038/s41564-017-0012-7.

25. Lee STM, Kahn SA, Delmont TO, Shaiber A, Esen özcan C, Hubert NA, Morrison HG, Antonopoulos DA, Rubin DT, Eren AM. 2017. Tracking microbial colonization in fecal microbiota transplantation experiments via genome-resolved metagenomics. Microbiome 5:1–10 doi:10.1186/S40168-017-0270-X.

26. Martens EC, Koropatkin NM, Smith TJ, Gordon JI. 2009. Complex glycan catabolism by the human gut microbiota: The Bacteroidetes Sus-like paradigm. Journal of Biological Chemistry 284:24673–24677 doi:10.1074/jbc.R109.022848.

27. Foley MH, Cockburn DW, Koropatkin NM. 2017. The Sus operon: A model system for starch uptake by the human gut Bacteroidetes. Cellular and Molecular Life Sciences 73:2603–2617 doi:10.1007/s00018-016-2242-x.

28. Grondin JM, Tamura K, Déjean G, Abbott DW, Brumer H. 2017. Polysaccharide utilization loci: Fueling microbial communities. Journal of Bacteriology 199 doi:10.1128/JB.00860-16.

29. Fernández-Gómez B, Richter M, Schüler M, Pinhassi J, Acinas SG, González JM, Pedrós-Alió C. 2013. Ecology of marine Bacteroidetes: A comparative genomics approach. ISME Journal 7:1026–1037 doi:10.1038/ismej.2012.169.

30. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. dbCAN: A web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Research 40:445–451 doi:10.1093/nar/gks479.

31. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018. dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Research 46:W95–W101 doi:10.1093/nar/gky418.

32. Lagkouvardos I, Overmann J, Clavel T. 2017. Cultured microbes represent a substantial fraction of the human and mouse gut microbiota. Gut Microbes 8:493–503 doi:10.1080/19490976.2017.1320468.

33. Stewart EJ. 2012. Growing unculturable bacteria. Journal of Bacteriology 194:4151–4160 doi:10.1128/JB.00345-12.

34. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research 25:1043–1055 doi:10.1101/gr.186072.114.

35. Kim BJ, Hong SH, Kook YH, Kim BJ. 2013. Molecular evidence of lateral gene transfer in *rpoB* gene of *Mycobacterium yongonense* strains via multilocus sequence analysis. PLoS ONE 8:1–5 doi:10.1371/journal.pone.0051846.

36. Schloss PD, Handelsman J. 2008. A statistical toolbox for metagenomics: Assessing functional diversity in microbial communities. BMC Bioinformatics 9:1–15 doi:10.1186/1471-2105-9-34.

37. Shipman JA, Cho KH, Siegel HA, Salyers AA. 1999. Physiological characterization of SusG, an outer membrane protein essential for starch utilization by *Bacteroides thetaiotaomicron*. Journal of Bacteriology 181:7206–7211 doi:10.1128/JB.181.23.7206-7211.1999.

38. Janeček Š, Svensson B, MacGregor EA. 2014. A-Amylase: An enzyme specificity found in various families of glycoside hydrolases. Cellular and Molecular Life Sciences 71:1149–1170 doi:10.1007/s00018-013-1388-z.

39. Ravcheev DA, Godzik A, Osterman AL, Rodionov DA. 2013. Polysaccharides utilization in human gut bacterium *Bacteroides thetaiotaomicron*: Comparative genomics reconstruction of metabolic and regulatory networks. BMC Genomics 14:873 doi:10.1186/1471-2164-14-873.

40. Rogers TE, Pudlo NA, Koropatkin NM, Bell JSK, Moya Balasch M, Jasker K, Martens EC. 2013. Dynamic responses of *Bacteroides thetaiotaomicron* during growth on glycan mixtures. Molecular Microbiology 88:876–890 doi:10.1111/mmi.12228.

41. van Bueren AL, Saraf A, Martens EC, Dijkhuizena L. 2015. Differential metabolism of exopolysaccharides from probiotic lactobacilli by the human gut symbiont *Bacteroides thetaiotaomicron*. Applied and Environmental Microbiology 81:3973–3983 doi:10.1128/AEM.00149-15.

42. D'Elia JN, Salyers AA. 1996. Contribution of a neopullulanase, a pullulanase, and an α-glucosidase to growth of *Bacteroides thetaiotaomicron* on starch. Journal of Bacteriology 178:7173–7179 doi:10.1016/j.tim.2004.07.004.

43. Delcher AL. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Research 30:2478–2483 doi:10.1093/nar/30.11.2478.

44. St John FJ, González JM, Pozharski E. 2010. Consolidation of glycosyl hydrolase family 30: A dual domain 4/7 hydrolase family consisting of two structurally distinct groups. FEBS Letters 584:4435–4441 doi:10.1016/j.febslet.2010.09.051.

45. Nguyen TLA, Vieira-Silva S, Liston A, Raes J. 2015. How informative is the mouse for human gut microbiota research? Disease Models & Mechanisms 8:1–16 doi:10.1242/dmm.017400.

46. Koropatkin NM, Smith TJ. 2010. SusG: A unique cell-membrane-associated α-amylase from a prominent human gut symbiont targets complex starch molecules. Structure 18:200–215 doi:10.1016/j.str.2009.12.010.

47. Naumoff DG. 2005. GH97 is a new family of glycoside hydrolases, which is related to the α-galactosidase superfamily. BMC Genomics 6:1–12 doi:10.1186/1471-2164-6-112.

48. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. 2004. Carbohydrate-binding modules: Fine-tuning polysaccharide recognition. Biochemical Journal 382:769–781 doi:10.1042/BJ20040892.

49. Kim YM, Yamamoto E, Kang MS, Nakai H, Saburi W, Okuyama M, Mori H, Funane K, Momma M, Fujimoto Z, Kobayashi M, Kim D, Kimura A. 2012. *Bacteroides thetaiotaomicron* VPI-5482 glycoside hydrolase family 66 homolog catalyzes dextranolytic and cyclization reactions. FEBS Journal 279:3185–3191 doi:10.1111/j.1742-4658.2012.08698.x.

50. Kim MJ, Lee SB, Lee HS, Lee SY, Baek JS, Kim D, Moon TW, Robyt JF, Park KH. 1999. Comparative study of the inhibition of alpha-glucosidase, alpha-amylase, and cyclomaltodextrin glucanosyltransferase by acarbose, isoacarbose, and acarviosine-glucose. Archives of Biochemistry and Biophysics 371:277–283 doi:10.1006/abbi.1999.1423.

51. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J. 2008. The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. Journal of Bacteriology 190:6881–6893 doi:10.1128/JB.00619-08.

52. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. 2005. The microbial pan-genome. Current Opinion in Genetics and Development 15:589–594 doi:10.1016/j.gde.2005.09.006.

53. Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature Methods 13:435–438 doi:10.1038/nmeth.3802.

54. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. 2017. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Research 27:626–638 doi:10.1101/gr.216242.116.

55. Delmont TO, Eren AM. 2018. Linking pangenomes and metagenomes: The *Prochlorococcus* metapangenome. PeerJ 6:e4320 doi:10.7717/peerj.4320.

56. Segata N. 2018. On the road to strain-resolved comparative metagenomics. mSystems 3:e00190–17 doi:10.1128/mSystems.00190-17.

57. Merino S, Tomás JM. 2015. Bacterial capsules and evasion of immune responses. eLS 1–10 doi:10.1002/9780470015902.a0000957.pub4.

58. Miller RA, Harrison DE, Astle CM, Baur JA, Boyd AR, de Cabo R, Fernandez E, Flurkey K, a. Javors M, Nelson JF, Orihuela CJ, Pletcher S, Sharp ZD, Sinclair DA, Starnes JW, Wilkinson JE, Nadon NL, Strong R. 2011. Rapamycin, but not resveratrol or simvastatin, extends life span of genetically heterogeneous mice. Journals of Gerontology - Series A Biological Sciences and Medical Sciences 66 A:191–201 doi:10.1093/gerona/glq178.

59. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S. 2012. FastUniq: A fast *de novo* duplicates removal tool for paired short reads. PLoS ONE 7:1–6 doi:10.1371/journal.pone.0052249.

60. Buffalo V. 2018. Scythe - A Bayesian adapter trimmer https://github.com/vsbuffalo/scythe.

61. Joshi NA, Fass JN. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) https://github.com/najoshi/sickle.

62. Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2014. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676 doi:10.1093/bioinformatics/btv033.

63. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–359 doi:10.1038/nmeth.1923.

64. Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nature Methods 11:1144–1146 doi:10.1038/nmeth.3103.

65. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830 http://www.jmlr.org/papers/v12/pedregosa11a.html.

66. Wedemeyer A, Kliemann L, Srivastav A, Schielke C, Reusch TB, Rosenstiel P. 2017. An improved filtering algorithm for big read datasets and its application to single-cell assembly. BMC Bioinformatics 18:1–11 doi:10.1186/s12859-017-1724-7.

67. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. 2012. A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv 1–18 http://arxiv.org/abs/1203.4802.

68. Zhang Q, Pell J, Canino-Koning R, Howe AC, Brown CT. 2014. These are not the k-mers you are looking for: Efficient online k-mer counting using a probabilistic data structure. PLoS ONE 9 doi:10.1371/journal.pone.0101271.

69. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19:455–477 doi:10.1089/cmb.2012.0021.

70. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9 doi:10.1371/journal.pone.0112963.

71. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069 doi:10.1093/bioinformatics/btu153.

72. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119 doi:10.1186/1471-2105-11-119.

73. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD. 2012. The Pfam protein families database. Nucleic Acids Research 40:D290–301 doi:10.1093/nar/gkr1065.

74. Eddy SR. 2011. Accelerated profile HMM searches. PLoS Computational Biology 7:e1002195 doi:10.1371/journal.pcbi.1002195.

75. Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome informatics International Conference on Genome Informatics 23:205–11 doi:10.1142/9781848165632_0019.

76. Juncker AS, Willenbrock H, von Heijne G, Brunak S, Nielsen H, Krogh A. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Science 12:1652–1662 doi:10.1110/ps.0303703.

77. Petersen TN, Brunak S, Von Heijne G, Nielsen H. 2011. SignalP 4.0: Discriminating signal peptides from transmembrane regions. Nature Methods 8:785–786 doi:10.1038/nmeth.1701.

78. Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12:59–60 doi:10.1038/nmeth.3176.

79. Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30:1575–1584 doi:doi: 10.1093/nar/30.7.1575.

80. Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, Richter AR, White O. 2007. TIGRFAMs and Genome Properties: Tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Research 35:D260–4 doi:10.1093/nar/gkl1043.

81. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 - approximately maximum - likelihood trees for large alignments. PloS ONE 5:e9490 doi:10.1371/journal.pone.0009490.