# ABSTRACT

Previously, we reported the induction of T helper 17 cells by a specific member of the mouse gut microbiota, segmented filamentous bacteria. We here present a draft genome sequence of SFB, isolated from mono-colonized mice. At 1.57 Mb, the SFB genome is less than half the size of related *Clostridium* species, and represents one of the smallest *Clostridia* genomes sequenced to date. By comparative genomic analysis, we determined SFB to be functionally related to several pathogenic, gut-residing, and minimal genera, such as *Finegoldia*, *Mycoplasma*, *Borrelia*, and *Clostridium*. Further analysis suggests that SFB is highly dependent on its host and fellow members of the gut microbiota for amino acids and essential nutrients, and that it may utilize host and dietary glycans for carbon, nitrogen, and energy. Production of acetate as a metabolite suggests that SFB may offer the host protection from bacterial infection in a manner similar to *Bifidobacteria*, a species also implicated in glycan foraging.

# INTRODUCTION

Segmented filamentous bacteria…

# RESULTS AND DISCUSSION

## General genome features

The SFB genome was assembled into 5 contigs totaling 1,569,870 bp with an average depth of sequencing coverage of 341x. A single contig, SFBMM\_CONTIG04, contains two complete rRNA operons and has approximately 3x the average genome-wide read coverage, suggesting that an additional 3-4 rRNA operons occupy the 3 inter-contig gaps in the assembly (**Table 8**). 38 tRNAs, while a minimal complement well below the *Clostridium* average (**Table 7**), provide specificity for all 20 canonical amino acids, plus selenocysteine. A single putative replication origin could be predicted by a distinct inflection point in GC skew around the dnaA gene, accompanied by 5 dnaA boxes, between 42,988 and 45,400 bp on SFBMM\_CONTIG01 [**19**, **20**] (**Figure 14**). Another infection point, located opposite the replication origin, marked the putative replication termination point, confirming bidirectional chromosome replication (**Figure 6**). As both the *ori* and *ter* regions were contained within a single contig, we were able to orient the remaining contigs by GC skew and strand bias. The contig order was predicted by analyzing the similarity of *k-mers* present on the ends of every contig pair (see Methods and **Figure 11**).

A total of 1,533 coding sequences (CDS) were predicted from the genomic sequence. The distribution of ORF lengths was similar to that of *Clostridium* species (**Figure 8**). Strand bias was very pronounced, with 82.5% of ORFs encoded on the leading strand of DNA replication. Annotation was performed using multiple pipelines (see Methods) and manual curation. The annotation assigned functions to 792 CDS, while another 345 CDS could be assigned to orthologous groups in the eggNOG, KEGG, or MBGD databases. Another 212 CDS showed a high degree of similarity to CDS in other genomes, and finally 184 (12% of total) CDS were unique to SFB. We then examined the homology of these CDS to other genomes. Of the total 1218 (79.4%) CDS that showed similarity to other genomes, all were highly similar to CDS found in members of the *Clostridium* genus. However, 15% of all coding sequences were more homologous to other genera, such as *Bacillus* and *Thermoanaerobacter* (a member of class V *Clostridia*), than to *Clostridium* (**Figure 6**).

To determine the extent of gene duplication and functional redundancy in the genome, we clustered all translated SFB ORFs using CD-HIT [**67**] at a 60% identity, 70% coverage cutoff. A total of 28 proteins in 14 paralogous clusters were located (**Table 16**), however, the annotation of the genome revealed more: 7 copies of N-acetylmuramoyl-L-alanine amidase, 6 copies of beta-glucosidase, and 5 putative polysaccharide deacetylases are among the most functionally duplicated genes. The remaining paralogous groups are found mostly in two prophage clusters, indicating that the SFB genome encodes few redundant functions.

At an average G+C content of 27.9%, the SFB genome is consistent with SFB’s phylogenetic placement among low-G+C Firmicutes species. Deviant G+C content was confined to two prophage clusters, rRNA genes, and a cluster of hypothetical proteins and transporters between 500,000 and 600,000 bp on SFBMM\_CONTIG01. Interestingly, codon usage bias correlated with the representation of tRNAs in the genome, with a third-position preference for A or T. This was consistent with *Clostridium* genomes but not *E. coli*, which showed different codon usage bias (**Figure 12**). We hypothesize that the overall bias toward A or T in SFB has driven the loss of tRNAs specific for higher-GC codons.

### SFB from mono-colonized mice are a single species

It is currently unknown if SFB represent a collection of closely related species in vivo. To determine the degree of heterogeneity in the in vivo SFB population inhabiting the intestines of mono-colonized mice, we used an established pipeline (see Methods) to check our reads for indels and SNPs. We found almost no evidence of polymorphisms, suggesting that the intestinal SFB population sequenced is genetically homogenous. This may reflect the highly adapted nature of SFB to their host and environment. Because SFB have been shown to be host-specific [**70**], comparison of SFB sequences from different host species will reveal important host-specific adaptations.

### Phylogenomic analysis places SFB in the family *Clostridiaceae*

A previous study of 16S rRNA sequences from rat, mouse, and chicken isolates placed SFB in a unique cluster amongst group I and II *Clostridium* [**25**]. A phylogenetic tree constructed from 16S rDNA genes on SFBMM\_CONTIG04 confirms this assignment (**Figure 10**), consistent with the ribosomal database project’s placement of SFB in the family *Clostridiaceae* [**29**]. A broader context AMPHORA [**21**] tree, comparing SFB to 1,053 microbial genomes based on 31 conserved “housekeeping” genes, also placed SFB in its own cluster, close, but separate from group I *Clostridium* (**Figure 9)**. Interestingly, SFB branches from a trunk that includes two other minimal, anaerobic opportunistic pathogens: *Finegoldia magna* [**33**] and *Peptoniphilus lacrimalis* [**34**].

Finally, a genome-wide pairwise MUMi [**71**] comparison of the SFB genome with a set of 1,700 draft and complete genomes confirmed that SFB is quite distinct from all organisms sequenced to date. For example, the MUMi value of the closest species, *C. botulinum*, on a scale from 0 to 1, is 0.96 (where 0 indicates both genomes are identical, and 1 indicates no overlap). These results suggest that SFB is a unique member of a novel cluster of *Clostridiaceae*.

### The SFB genome encodes two prophages, but no insertion elements

To detect evidence of recent horizontal gene transfer (HGT), we used three different methods [**36**, **37**, **38**] to find genomic islands, ISsaga [**39**] to search for insertion sequence (IS) elements, and Prophinder [**40**] to locate prophages. We were able to predict several small genomic islands (**Figure 13**, **Table 15**), but were unable to detect evidence of any IS elements. The lack of genomic islands containing IS elements suggest that there has been no recent HGT in the SFB genome.

We were, however, able to detect a 45kb prophage between 22,041 and 67,597 bp on SFBMM\_CONTIG01 and a 32kb prophage between 33,982 and 66,452 bp on SFBMM\_CONTIG02. By manual inspection of the SFB annotation, we found two small prophage clusters spanning 6kb and 45kb. Similar phage clusters have been describedin related species, such as *Clostridium acetyobutylicum* and *Clostridium perfringens*. Seven prophage CRISPR-associated proteins (SFBMM\_008790 to SFBMM\_008850) were identified in the gene annotation, and 3 arrays of CRISPR sequences were located as previously described [**41**].

As we were able to detect only X putative pseudogenes, and no extrachromosomal sequences (plasmids or active phages), we conclude that the genome of SFB is relatively stable at present.

### Regulation of transcription and gene expression

Transcriptional regulation in bacteria occurs in response to a number of environmental stimuli. In the gut, SFB is most likely exposed to various such signals derived from both the host and the activities of the surrounding microbiota. We therefore examined the representation of global regulatory proteins, such as sigma factors, DNA-binding proteins, and two-component systems.

The SFB genome contains 8 sigma factors, 3 of which are likely to be sporulation-specific, appearing adjacent to clusters of sporulation proteins. Another 4 comprise the general RNA polymerase sigma factor system, and one appears specific to the flagellar operons (Table).

Two-component systems consist of a membrane-bound histidine kinase sensor protein and a cytoplasmic DNA-binding response regulator. SFB has 9 such pairs of sensor/regulator proteins, all occurring adjacent to one another, and a single orphan histidine kinase. While the function of phosphate-, sugar-, and chemotaxis- related pairs could be predicted by homology to other species and location within an operon, the remaining 6 have only general functional prediction (**Table 9**).

SFB have as many as 23 other annotated response regulators with only general functional prediction, including 3 with homology to *fur* family proteins implicated in regulation of iron uptake (**Table 10**). To locate transcriptional regulators with only distant homology to other genomes, we searched for nucleic acid binding helix-turn-helix motifs in SFB ORFs. Using this approach, we were able to locate an additional 16 putative transcriptional regulators, including an unclassified PTS EIIA/PRD/HTH fusion protein (SFBMM\_008590), which led us to investigate SFB’s PTS regulatory apparatus (**Table 11**). We located a PRD/RBD domain-containing protein (SFBMM\_008610) and a deoR-family transcriptional regulator (SFBMM\_005640) adjacent to two ascorbate and glucoside transport and utilization clusters, hinting that SFB possess a complex mechanism of regulating usage of these sugars, integrated with the PTS system. Interestingly, 7 helix-turn-helix proteins were located within the phage cluster on SFBMM\_CONTIG02.

## Comparative functional genomics of Segmented Filamentous Bacteria

In order to assess the SFB genome's functional potential in comparison with other microbes, we assigned its open reading frames to orthologous groups in two different ways. First, we examined KEGG Orthologs [**1**], which confidently annotated 792 of SFB's 1,533 ORFs with a total of 717 KOs; these were subsequently compared to the 1,209 high-quality KEGG microbial genomes and to five eukaryotic outgroups (human, mouse, fly, worm, and Arabidopsis). This catalog provides moderate coverage of SFB at the per-gene level (52%), comparable to that of the 30 *Clostridia* included natively in KEGG (45%), and it allowed convenient characterization of KEGG functional modules and pathways as detailed below. To improve per-gene coverage, we additionally analyzed a set of 236,073 orthologous groups derived from MBGD [**2**], annotating 1,003 of SFB's ORFs (66%) for whole-genome comparison with 1,153 microbial genomes (see Methods and **Figure 1**). These KO and MBGD gene family catalogs were employed to compare SFB with over 1,100 finished microbial genomes, allowing us to identify the most closely functionally related organisms to complement the phylogenetic analysis above.

### Organisms in class *Clostridia* possess similar functional repertoires to SFB

We used the Tversky index [**3**,**4**] with α=0.75 to determine organisms carrying a similar complement of genes to SFB (see Methods), highlighting primarily species from the genera *Clostridia* and *Thermoanaerobacter* (**Table 1** and **Figure 2**). The former are phylogenetically quite related to SFB, as detailed above by our 16S-rRNA analysis, and it is thus reassuring that they are found to carry a similar functional complement, particularly using the more specific MBGD catalog. However, while the organisms with comparable functional repertoires to SFB are enriched for endosymbionts, anaerobes, and modest thermophiles, it is notable that even the most similar organisms (top 20 genomes) share at best a fraction of SFB's gene families (ave. 571 sd. 19 of 717 different KOs, 80%; ave. 628 sd. 27 of 1,003 in MBGD, 63%). Similarly, these organisms tend to carry much larger gene complements than SFB (ave. 533 sd. 181 additional distinct genes in KO; ave. 1,199 sd. 181 in MBGD), and we turned to a lower Tversky α value in order to assess organisms more strictly functionally similar to SFB.

### "Minimal" organisms including the Mycoplasmas are not functionally dissimilar from SFB

Applying the Tversky index with α=0.25 to emphasize organisms with few gene families in addition to those carried by SFB indicated similarity with several "minimal" *Mycoplasma* and *Ureaplasma* species [**5**], in addition to the *Borrelia* genus (**Table 1** and **Figure 3**). These are again all endosymbionts that are at best obligate and at worst pathogenic, most with reduced genomes of size comparable to that of SFB [**6**,**7**]. This may be interpreted as the functional consequence of SFB's evolution as an obligate commensal (as observed for other bacteria [**8**] like some *Streptococcus* species [**9**]), but again, even the most similar organisms show only partial overlap with SFB's gene complement (ave. 332 sd. 69 of 717 different KO, 46%; ave. 288 sd. 86 of 1,003 in MBGD, 29%) and tend to carry a number of additional genes (ave. 160 sd. 84 in KO; ave. 238 sd. 175 in MBGD). Using either catalog's definition of orthologous gene families, SFB carries a strikingly distinct gene complement as compared to any of the ~1,200 currently sequenced archaea and bacteria.

For example, even in the more conservative KO catalog, SFB carries 94 genes (13%) not found in any of the 20 most similar organisms (see **Table 6**). These include several sporulation proteins (stages II to V for a total of 14 KOs), six regulators in the OmpR family, several amino acid biosynthesis proteins, and a variety of enzyme-coding genes. Nine uncharacterized conserved gene families are included in the genes appearing in at most one additional organism, and conversely, a core of 132 gene families (18.4%) appears in both SFB and all of the most similar organisms. Finally, 11 additional genes occur in at least 90% of the 20 most similar organisms but not SFB, comprising mainly proteins from the large ribosomal subunit. These are in addition to multiple energy-processing pathways common among other organisms (≥50%) yet absent in SFB, e.g. many sugar transporters and metabolic enzymes, the entire F-type ATPase {\*\*\*}, several spermidine/putrescine transporters, and a range of membrane proteins. These differences suggest that SFB has evolved a reduced genome comparable in scope to that of other "minimal" organisms [**6**,**8**], but with a unique set of retained pathways relying on targeted interactions with the host in order to sustain basic metabolic processes.

### SFB metabolic potential shares features of both generalist gut commensal Clostridia and of minimal Mycoplasmas

In order to more concisely compare SFB with other bacteria in terms of individual functional and metabolic pathways, we collapsed the total 13,118 KO families in these data to 371 functional modules (small ~5-20 gene pathways as defined by KEGG). These modules substantially improved the consistency of the above comparisons, highlighting almost exclusively the *Clostridia* and related *Thermoanaerobacteria* for α=0.75 and the *Mycoplasma* for α=0.25 (**Table 1**). Three strains of *Lactococcus lactis* and six of *Streptococcus pyogenes* were also included using the former criteria, which links SFB to this collection of organisms based on the presence of basic essential pathways (DNA maintenance and replication, transcription and translation, glycolysis, sugar and phosphate transport, fatty acid processing, etc.) More specifically, the collection of Firmicutes at α=0.75 also near-uniformly shares a rich collection of transport modules (including iron, nickel, cobalt, zinc, and several uncharacterized ABC systems), multiple phosphotransferase systems, and portions of the Sec (M00335) secretion system. SFB possesses few modules not present in these organisms, a notable exception being two modules related to metal metabolism, 2-oxoglutarate: ferredoxin oxidoreductase (M00311) and phosphonate transport (M00223). Conversely, as detailed in **Figure 4** and **Table 4**, SFB lacks a wide range of otherwise common functionality, primarily amino acid and nucleotide biosynthetic pathways (proline, cysteine, tryptophan, M00015/M00021/M00023, and adenine, guanine, and pyrimidine, M00049/M00050/M00052) and several phosphotransferase systems (specifically 10 modules including glucose, maltose, arbutin, sucrose, rehalose, fructose, mannose, and galactose components, mostly membrane-bound sugar specific permeases). Coupled with SFB's abundance of amino acid and phosphonate transport transporters, this suggests that the organism relies heavily on the host particularly for production of specific amino acids.

Although this set of metabolic modules is, as expected, representative of what might be expected for mildly thermophilic, anaerobic commensals in the gut community, a more surprising set of commonalities emerges when comparing SFB pathways with the *Mycoplasma* at α=0.25. Many of the same pathways are shared, e.g. the 20 most similar organisms at this parameter setting possess basic essential pathways, sugar, metal, and phosphate transport. More evident is the preference for F-type in place of V-type ATPases in SFB compared to these "minimal" organisms, among which the only three organisms lacking V-type ATPases (in addition to SFB) actually lack the F-type as well (whereas SFB possesses 7 of the 8 genes of the F-type ATPase module). In contrast to the Clostridia, few other pathways are lacking in SFB relative to the *Mycoplasma*, pyruvate oxidation(M00307), the M00248 multi-drug antibiotic transporter, ATP synthase (M00164), and again spermidine/putrescine transport being the only common exceptions (**Figure 5** and **Table 5**) The metabolism specific to SFB and lacking in these otherwise similar organisms is telling, consisting mainly of mannose/cellobiose/beta-glucoside components (mono/disaccharides particularly available in the gut {Stoll, 2010?}), iron/zinc transport M00240/2 and processing M00311 {\*\*\*}, phosphate/phosphonate transport M00222/3, and portions of the Shikimate pathway (M00022). SFB thus appears to occupy a functional "midpoint" between the pathogenic/parasitic *Mycoplasma* and the range of generalist Clostridia representative of the mammalian gut microbiota.

### Variable metabolic modules in SFB are enriched for small molecule phosphotransferases

Each of the contrasts above provides a perspective on the SFB genome's functional repertoire using one of three gene family catalogs: KEGG Orthology KO gene families, more specific MBGD orthologous gene families, and KEGG functional modules. We further refined each of these three catalogs into "core" members present in at least 75% of the 1,200 microbial reference genomes and "variable" members present in at least 5% but at most 25% (**Table 2**). SFB carries an expected complement of core genes and somewhat fewer than average variable gene families, unsurprising due to its small genome size. Its 14 characterized core modules are as described above (basic essential pathways, metal transport, etc.), but strikingly, five of its eight variable modules are small molecule phosphotransferases (M00271, M00274-6, and M00283), in addition to the V-type ATPase mentioned above. This core/variable separation additionally supports the functional contrasts among SFB and related genera, as its variable gene families and modules are particularly associated with the *Clostridia* and *Thermoanaerobacter* (**Table 2**). Its core genes and pathways are neither particularly associated with the *Clostridia* nor with the *Mycoplasma*, however, and instead are loosely similar with several generally thermophilic, anaerobic, and pathogenic bacteria. Much of SFB's genome falls outside of both these core and variable gene and pathway sets and represents relatively unique biology.

### SFB demonstrates moderate, but only moderate, functional similarity with "model" pathobionts

Finally, we compared the SFB genome's functional capacity with a set of bacteria of specific interest due to their phylogenetic relatedness, functional similarity, or related genome size and/or niche (**Table 3**). Particularly in terms of the mainly well-characterized biology represented in KEGG orthologous families and modules, SFB remains similar to the *Clostridia* and minimal pathobionts, including *F. magna* (sharing e.g. the filamentous growth module M00256 and several transport systems), *T. denticola* (sharing iron/zinc/cobalt transporters M00240/2 and M00245/6), *G. vaginalis*, (sharing sugar/ribose transporters), and *B. burgdorferi* (sharing the cellobiose PTS). This functional similarity is not observed in e.g. the *Helicobacteria* or *C. jejuni*, despite their similar genome sizes and host environments, even using the comprehensive MDGB gene family catalog. Still, with MDGB's emphasis on largely uncharacterized orthologous families, SFB remains overall genomically dissimilar even with the most phylogenetically related *Clostridia*, suggesting that a wealth of minimal gut commensal biology remains to be characterized from this unique organism.

## Lifestyle and metabolism

SFB have a complete EMP glycolysis pathway, and the non-oxidative portion of the pentose phosphate pathway (PPP), but lack 6-phosphogluconolactonase, a key enzyme in the oxidative PPP, and nearly all components of the TCA cycle. Similar to *Clostridium* species, SFB possess the ability to synthesize fatty acids but lack the ability to degrade them. Import of fructose, mannose, mannitol, L-ascorbate, and oligosaccharides is accomplished by SFB’s PTS system, which consists of at least 20 genes (**Table 19**), representing enrichment versus SFB’s functional relatives. SFB also possess a putative sialic acid:sodium symporter that may catalyze the uptake of N-acetylneuramininate, a common terminal sugar found on mammalian glycans. After import into the cell, these substrates enter glycolysis and converge at phosphoenolpyruvate (PEP), which feeds back into the PTS system or proceeds to acetyl-CoA via reduction of a ferredoxin-like iron-sulfur cluster protein by a multi-subunit ferredoxin oxidoreductase. Reduced ferredoxin can be recycled back to its oxidized form with concomitant release of hydrogen, though it has been suggested that some *Clostridium* couple this oxidation to the reduction of NAD+ to NADH [**45**] which then serves as a cofactor in NADH-dependent reactions. Without the TCA cycle, acetyl-CoA then proceeds to fatty acid biosynthesis, or is converted to acetate (**Table 12** and **Table 14**).

### SFB form spores, but other minimal organisms don’t

In papers originally describing SFB, it was noted that spore-like intrasegmental bodies (up to two per cell) could be found in segments not attached to epithelial cells, in addition to true spores [**32**]. The SFB genome contains all the sporulation and germination gene sets found in spore-forming *Clostridium* species [**30**, **31**], in agreement with reports that SFB form spores, and their phylogenetic placement among spore-formers. Interestingly, while other species have multiple *ger* operons, presumably regulating the response to germinants, SFB have only one such operon. Consisting of three proteins, only gerAA is above the protein alignment twilight zone threshold at 58% homology to gerAA in related species, while gerAB has a best hit to NR of 28% identity, and gerAC, 23%. This may indicate that SFB’s germination cue is unusual. Further, our comparative genomic analysis showed that at least 14 sporulation-related genes (see **"Minimal" organisms including the Mycoplasmas are not functionally dissimilar from SFB**) are not found in SFB’s “minimal” relatives, suggesting that the ability to sporulate is unique to SFB among functionally similar organisms.

### SFB lack key components of major cofactor and amino acid biosynthesis pathways

As we demonstrated above, genes involved in the metabolism of essential nutrients are missing in SFB, or are under-represented as compared to its functional relatives (**Table 4** and **Figure 6**). This reflects SFB’s adaptation to life as an obligate symbiont.

Specifically, amino acid metabolism is almost entirely absent from SFB, except for lysine biosynthesis, an aspartate/glutamate/asparagine/glutamine (DENQ) pathway, and glycine hydroxymethyltransferase (SFBMM\_007260) catalyzing the interconversion of glycine and serine. SFB appear to import all amino acids via a polar amino acid ABC transporter, two putative amino acid antiporters, or alternately, they may import oligopeptides via one of several ABC transporters for cytosolic degradation (**Table 19**). DENQ amino acids are likely maintained as a nitrogen reservoir, participating in several trans- and de- amination reactions, notably the conversion of fructose-6P to glucosamine-6P for synthesis of peptidoglycan, or the reverse for utilization of N-acetylglucosamine in glycolysis.

Nearly all enzymes required for synthesis of biotin, thiamine, folate, and riboflavin, but notably not pyroxidal phosphate (PLP) or coenzyme-A (CoA), are missing. SFB appear to have the same glutamine-dependent PLP synthesis machinery (pdxS/pdxT) as *B. subtilis* [**35**], the presence of which is consistent with SFB’s DENQ pathway and the requirement of PLP as a cofactor for transamination.

Interestingly, SFB lack all enzymes to synthesize coenzyme-A but a putative pantothenate kinase (SFBMM\_011800), pantetheine-P adenyltransferase (SFBMM\_009620), dephospho-CoA kinase (SFBMM\_014840), and holo-[ACP] synthase (SFBMM\_014530) which together may constitute a complete pathway for conversion of pantetheine to coenzyme-A (CoA) and acyl-carrier-protein (ACP) (**Table 18**). For comparison, *Clostridium*, *Finegoldia*, and *Borrelia* have complete pantothenate (vitamin B5) to CoA pathways, while *Mycoplasma*, *Treponema*, and *Bifidobacterium* have the same subset as SFB. This suggests that minimal, obligate symbionts such as SFB rely on metabolic intermediaries from the conversion of vitamin B5 to CoA by the host or by other members of the microbiota. Similarly, SFB must rely on host, microbiota, and dietary sources for other essential cofactors.

### Transporters, especially of metal ions, are retained despite or due to genome reduction

SFB are similar to the United States. The United States began importing goods from overseas, then lost the ability to manufacture those same goods at home, and now seek only to secure their continued ability to import without the option of reverting back to their former way of life.

As might be expected from an organism that appears to have lost its ability to synthesize its own cofactors, a large portion of SFB’s genome (8%) is dedicated to transport machinery.

In particular, metal ion uptake and efflux machinery constitute a large fraction of SFB’s transporters. Several putative transporters for zinc, iron, cobalt, nickel, magnesium, and chromate are present. In addition, SFB have P-type ATPase genes implicated in the efflux of toxic heavy metals, such as cadmium, from the cell. Iron uptake systems are the most represented metal transporters in the genome. At least 31 genes were annotated as part of multiple mechanisms for acquiring different forms of iron from the environment. These include a hydroxamate siderophore ABC transporter, two feoA/feoB ferrous iron transporters, and three uncharacterized iron complex ABC transporters with homology to vibrioferrin (a catecholate siderophore) and heme importers (**Table 19**). While SFB does not appear able to synthesize its own siderophores, it seems likely that it has evolved to utilize those synthesized by neighboring species in the gut, and perhaps host iron in the form of heme.

The two feoB genes in the SFB genome share more than 50% identity with other Clostridium feoB genes. Prokaryotic FeoB proteins are involved in free Fe2+ (as opposed to Fe3+) transport from the environment and FeoB homologs are common in intestinal bacteria, such as H. pylori, where FeoB is thought to provide the major pathway for Fe2+ uptake and is essential for colonization of the murine gastric mucosa (Velayudhan et al.).

Finally, SFB possesses three proteins that encode the Ferric uptake regulator (FUR) family of metal ion uptake regulatory proteins, which are responsible for controlling the intracellular concentration of iron in many bacteria. Interestingly, the three SFB Fur genes are highly identical to Fur genes from different groups of microorganisms – fur1 is 60% identical to Acetovibrio and Fusibacterium, fur2 is 50% identical to Thermoanaerobacter and fur3 is 40% identical to cluster I clostridia (botulinum, beijernickii, perfingens, butyricum).

As we showed above (see **Variable metabolic modules in SFB are enriched for small molecule phosphotransferases**), the set of PTS transporters encoded for in the SFB genome are relatively unique as compared to SFB’s functional neighbors. Specific for saccharides present in mucin, laminin, and dietary copmonents of the host ileum, we believe this indicates a unique retention of genes necessary to degrade and utilize glycans (**Figure 7**).

## Factors in host-microbe interaction

### Extracellular and secreted proteins

PSort predicted 34 extracellular and 20 cell wall proteins (**Table 17**) in SFB. LPxTG-type of surface proteins are a major group of surface proteins found in Gram-positive bacteria [**42**]. This group of proteins including many adhesins [**43**] typically contain a conserved sortase reorganization and cleavage motif (the LPxTG motif) and many other sequence characteristics [**42**]. We were not able to find LPxTG-type of surface protein in SFB when using the criteria described by Boekhorst et al [**42**]. No sortase was found in SFB, suggesting that SFB may not have LPxTG-type surface protein as other Gram-positive bacteria including many Clostridium species, or that the features of SFB LPxTG-type surface proteins are more divergent. In fact, many Clostridium species seem to lack LPxTG-type of surface protein [**44**, **45**]. 214 ORFs were found to have signal peptides, presumably are secreted by the Sec-dependent pathway [**44**]. Sec-dependent secretion system proteins (SecA, secD, secE, secF, secG, secY and yajC) that are similar to other Clostridium found in SFB. No Clostridium type of S-layer protein, or cell wall-binding repeat 1 (PF01473) or cell wall-binding repeat 2 (PF04122) domain protein, or bacterial Ig-like group 2 domain (PF02368) or Bacterial Ig-like group4 domain (PF07532) was found in SFB.

### Fibronectin may be important in epithelial attachment

Fibronectin binding proteins in *Clostridium* species have been shown to bind soluble and immobilized fibronectin [**46**, **47**] and play an important role in intestinal colonization [**48**]. The presence of a fibronectin-binding protein (fbpA, SFBMM\_005870) in SFB may facilitate their colonization of the host GI tract. Although a previous study suggests that *Clostridium* fbpA may be localized to the cell surface [31], SFB and *Clostridium* fbpA seem to lack the N-terminal signal peptides or sortase recognition (LPxTG) motif, and so may be translocated through the cell membrane using another secretion pathway [**44**].

The largest coding sequence in the SFB genome is a 1901 amino acid CDS containing 1-3 fibronectin type III domains (SFBMM\_015470). Aligning this CDS back to the genome revealed two homologous ORFs (30% identity) of similar size (SFBMM\_010700, SFBMM\_004940), none of which had significant similarity to any sequences deposited in NR across more than 6% of its length. There was little evidence for misassembly in the vicinity of these ORFs, all other reading frames were riddled with stop codons, and it seems unlikely for reading frames this large to be preserved by chance. All are putatively extracellularly localized. Given the relevance of bacterial attachment to host fibronectin in gut colonization, these paralogous genes may serve as adhesins. Future experiments will determine their true function.

### Polysaccharide capsule genes

The ability to synthesize a polysaccharide capsule is a virulence-determining factor in many pathogenic bacteria, where it may facilitate attachment to a host mucosal surface or evasion of engulfment. A commensal gut microbe, *B. fragilis*, synthesizes 9 different polysaccharide capsules, one of which (PSA) is implicated in the restraint and induction of Th17 and Tregs cells, respectively.

All *Clostridium* genomes sequenced to date have contained the genes necessary to produce a capsule [**45**]. While SFB possess two key genes required for capsule biosynthesis, O-antigen polymerase and UDP-GluNAc 2-epimerase, and a putative exocellular polysaccharide synthesis operon (**Table 13**), they appear to lack a gene coding for capsule chain length determinant protein (PF02706).

### SFB lack toxins in pathogenic *Clostridium*

*Clostridium* is a genus defined by the diverse toxins and virulence factors produced by its member species. *C. botulinum* strains produce 7 different neurotoxins, while *C. difficile* produce at least 2 enterotoxins, and *C. perfringens* strains may produce an ulcerative beta-toxin [reference].

Two lipases - plcA, pathogenicity-related phosphatidylinositol phospholipase C (PI-PLC) and Phospholipase

### SFB have redundant sets of flagellar and chemotaxis genes, yet their mobility is undocumented

SFB genome has a complete set of genes that encode flagella proteins and chemotaxis-associated proteins, which are found in all Clostrium species except C. perfringens, suggesting that SFB is mobile. SFB appears to loss the type IV pilus encoding gene clusters during its genome reduction, although pilT (SFBMM\_007500, which is required for sporulation) and gspE (SFBMM\_005880) were retained. Type IV pili were found in Clostridium perfringens, Clostridium botulinum, Clostridium difficile and Clostridium tetani, which enable Clostridium of gliding motility [**49**].

### SFB may utilize host and dietary glycans

The specific set of amino acid metabolism, PTS transporters, and associated sugar utilizing enzymes present seem to suggest that SFB utilize N-glycans for carbon and energy. Annotated beta-glucoside and cellobiose transporters may import cleaved disaccharides, while N-acetylglucosamine/mannose and N-acetylneuraminate monosaccharides may be imported by mannose (manXYZ) and putative sialic acid (nanP) transporters, respectively. SFB possess all the enzymes necessary to prepare these substrates for glycolysis, several in multiple copies (**Table 12**). Glycoproteins may be cleaved by one of nine putative extracellular peptidases, imported by SFB’s oligopeptide ABC transporter, and degraded as a source of nitrogen and amino acids. Several putatively extracellular glycoside hydrolases are consistent with this model. Indeed, although SFB does not possess the glycan foraging activity of major gut symbionts, such as *B. theta*, genes involved in glycan metabolism were the only BRITE gene category that was overrepresented in SFB in comparison to the rest of the clostridia (**Figure 6**).

Previously, we reported that B3GNT7 and FUT2 were upregulated 6-fold in Taconic vs. Jackson mice, among the genes most induced by SFB *in vivo* [**21**]. Implicated in transferring N-acetylglucosamine and fucose, respectively, to glycoproteins in mice, and consistent with a report that glycan forager *Bifidobacterium bifidum* induces the upregulation of B3GNT5 when exposed to Caco-2 human intestinal cells [**23**], this may indicate that SFB and bifidobacteria induce the production of a food source in the form of host glycans. Further, it was shown that SFB induce production of asialo GM1 glycolipids on the epithelial surface of the ileum in mice [**26**], and that the interaction of flagella with asialo GM1 can induce the production of mucin in lung epithelial cells [**27**]. It seems worth future investigation to determine if SFB induce mucin production in an asialo GM1- and flagellin- dependent manner, perhaps while avoiding TLR5 stimulation polarized to the basolateral surface [**28**] of the gastric epithelium.

### Peptidoglycan deacetylation may be important in SFB avoidance of the innate immune system

Modifications to the peptidoglycan layer are implicated in the successful evasion of host antimicrobial enzymes by pathogenic bacteria. Lysozyme, in particular, can degrade bacterial cell walls and cause cell lysis. Bacteria have been reported to modify their cell walls by N-deacetylation of N-acetylglucosamine, and O-acetylation and N-glycosylation of N-acetylmuramic acid [**50**]. SFB possess 6 polysaccharide deacetylases predicted to be extracellularly localized (**Table 13**) that may function in the N-deacetylation of N-acetylglucosamine sugars in the cell wall. This may account, in part, for SFB’s unique ability to thrive embedded in the mucosal layer, rich in host antimicrobial factors.

### Acetate production may account, in part, for SFB’s protective effect against infection

It has been shown that SFB may exert a protective effect against *Listeria monocitogenes*, pathogenic *E. coli*, *Salmonella enteritidis*, and *Citrobacter rodentium* infection [reference]. In a previous study, *Bifidobacteria* were shown to exert a similarly protective effect against enteropathogenic *E. coli* through production of acetate [**24**]. As SFB possess a single pathway for fermentation of pyruvate, resulting in the production of acetate, it is tempting to speculate that this accounts, in part, for the observed protection against bacterial infection.

## SFB cannot be detected in human stool

To determine if an SFB-like species is present in the human gut, we aligned sequenced reads from 122 individuals from the MetaHIT project [**51**], and X individuals from the Human Microbiome Project [reference] to the assembled SFB genome. We defined “present” as having 0.5% or greater genome coverage at an identity threshold of X%. At this cutoff, low-abundance microbes such as *Enterococcus faecium*, *E. coli,* and *Methanobrevibacter smithii* could be detected, while SFB could not. This suggests that SFB are absent from human stool, persist in an amount below our level of detection, or are scarce in those individuals sequenced. We hypothesize that SFB are most likely to be found in populations that have lived alongside animals known to harbor SFB for many generations, perhaps enough time in close proximity for a zoonotic event to occur. It will be interesting to see if metagenomic sequencing of more individuals at greater depths in the future will allow the detection of SFB.

# METHODS

## DNA isolation

DNA was isolated from the feces of mice mono-colonized with SFB by mechanical extraction.

## Genome sequencing and assembly

The genome was assembled from one run of 454 XLR mate-paired sequencing, which generated 1,287,974 reads with 427,851,871 base sequences. Pyrosequencing reads were first filtered by aligning to *Mus musculus* genome build 37.1 from NCBI using blastn with an e-value cutoff of 1e-20, then assembled with Newbler (2010-04 pre-release) using default parameters. Intra-scaffold gaps were filled using 454 mate-paired reads not used in the initial assembly. Insertion, deletions, and SNPs in the 454 assembled contigs were corrected by SOLiD reads using the BWA aligner [**52**] and SAMtools [**53**]. The remaining contigs <1kb were aligned to NT using BLASTN and removed if found likely to be from a contaminant source. The final assembly consists of 5 contigs with a total size of 1,569,870 bp and an N50 of 1,317,732 bp.

## Orienting and ordering of contigs

The putative *ori* and *ter* regions of the SFB chromosome were contained within a single 1.3Mb contig. This precluded the remaining contigs from varying in their orientation, as they all must fall within the GC skew minus portion of the chromosome. The reverse complement of SFBMM\_CONTIG04 was taken relative to the orientation it was assembled in, for consistency with the rest of the genome.

To order the contigs, segments from each end of every contig were extracted, ranging from 100 bp to 4 kb, skip 100. For each segment size, the occurrence of *k-mers* of length 2 through 9 bp was counted. Therefore, at each end of every contig, a certain length sequence was extracted and the frequency that a certain length *k-mer* occurred was calculated. For each such segment and *k-mer* length, the Euclidean distance of *k-mer* occurrence frequencies between all possible combinations of contig ends was calculated, and those pairs minimizing the distance were counted as “votes” for that contig order. As performed on the SFB genome assembly, containing 5 contigs, 1600 such votes were cast. The results were visualized in cytoscape with edge thickness corresponded to votes cast for that contig order (**Figure 11**). Based on this analysis, we assigned an order of SFBMM\_CONTIG01, SFBMM\_CONTIG03, SFBMM\_CONTIG02, SFBMM\_CONTIG04, and SFBMM\_CONTIG05. PCR experiments to confirm this ordering are underway.

## ORF prediction and annotation

The prediction of protein coding genes on SFB contigs was accomplished by Glimmer 3 [**54**] and GeneMark [**55**]. tRNAScan [**56**] was used for tRNA prediction, RNAmmer [**57**] for rRNA prediction, and RFAM/infernal for other non-coding RNA genes [**58**, **59**] Gene annotation was accomplished by submission to RAST [**60**], KAAS [**10**], IMG/ER [**61**], and a prokaryotic annotation pipeline created by the Human Genome Sequencing Center at the Baylor College of Medicine. Domain families for each called ORF were found using Pfamscan v1.3 [**62**] with HMMer v3.0b3 and database Pfam-A v24.0. Differing annotations for the same ORF were resolved manually by choosing the annotation most consistent with PSI-BLAST hits to NR and Pfam domains present in the ORF. In instances where there was a convincing alignment (>50% identity across >70% of the ORF) to multiple proteins annotated differently, the ORF was labeled a “hypothetical protein” or by the conserved domains it contained. All annotations were then adjusted to be maximally compliant with NCBI’s bacterial genome submission guide.

Localization of putative protein products was determined using PSort-B v3.0.2 [**63**] in gram positive mode, and LipoP 1.0 [**64**] with default parameters. Proteins were classified as putative lipoproteins if a signal peptidase II sequence was located by LipoP. Sortase recognition motifs (LPxTG) for gram-positive surface proteins were identified by regular expression match of conserved sequences ([LYFPSIV][PGSA]X[TSA][GANS] and N[PSA][QK]T[NA]) [**42**, **43**]. Transmembrane helices were predicted with TMHMM Server v2.0 [**65**]. Signal peptide sequences were predicted by SignalP 3.0 [**66**].

## Phylogenetic analysis

16S rRNA sequences were predicted from finished and draft genomes downloaded from NCBI using RNAmmer [**57]**. The predicted 16S rRNAs were filtered by length (>=1300bp and <=1700bp) to remove short or possibly misassembled genes. One copy of the 16S rRNAs from Clostridiales or Clostridiaceae species were extracted for phylogenetic analysis. Infernal [**59]** was used for multiple sequence alignment and phylogenetic distances were inferred using the maximum-likelihood method [**68**]. The tree topology was obtained using the NEIGHBOR program of the PHYLIP package [**68]** based on 1000 resamplings. The resulting tree was visualized using iTOL [**69**].

## MetaHit WGS read mapping analysis

Illumina raw WGS reads of 124 individuals from the MetaHIT project [**51**] were aligned to SFB contigs and 6 other reference genomes (Clostridium perfringens ATCC13124, Enterococcus faecalis V583, Enterococcus faecium TX1330, Escherichia coli MG1655, Lactobacillus johnsonii NCC533, Methanobrevibacter smithii ATCC35061; accession numbers: CP000246.1, NC\_004668.1, NZ\_ACHL00000000, U00096.2, AE017198.1, NC\_009515.1, respectively) using BWA aligner [**52**]. Reads aligned with an identity of 95% or higher were considered good matches and used in relative abundance and genome coverage analysis. Relative abundance was defined as the percentage of mapped reads divided by total reads, and genome coverage as the percentage of genome bases aligned to reads.

## Polymorphism detection

454 reads were extracted from native SFF format into FASTQ, and those used by Newbler in the final assembly were kept. SOLiD reads were converted from native csfasta to FASTQ. Both were aligned with BWA v0.5.9-r16 to the assembled SFB genome. Samtools [**53**] “mpileup” was used as described in the documentation to locate convincing SNPs by fraction of coverage at a base position corresponding to different nucleotides.

## Annotating SFB ORFs with KEGG and MBGD orthologous families

The 1,533 ORFS detected for the SFB genome were mapped into the KEGG Orthology (KO) database [**1**] (as of March, 2011) as determined by the KEGG Automatic Annotation Server (KAAS [**10**]) and into the MBGD database [**2**, **11**] (version 2010-02) of orthologous families using blastn [**12**, **13**]. KO and MBGD include 1,191 and 1,153 microbial genomes, respectively. We found a total of 792 SFB ORFs (52%) with homologs in one of the 13,118 KOs (corresponding to 717 distinct KO families), whereas 1,003 ORFs (66%) had a confident match in at least one of the ~4M genes in MBGD belonging to one of the 236,073 gene families with at least 5 orthologs. This information was analyzed as two matrixes reporting the abundance of each gene family in each genome (all 1,153 organisms in MBDG, 1,186 high-quality KEGG microbial genomes, and a set of five eukaryotic outgroups including human, mouse, fly, worm, and Arabidopsis) and in the SFB genome. Cytoscape [**14**] was used to visualize these data before further processing (**Figure 1**).

## Assessing functional similarity of SFB with microbial reference genomes

To assess the functional similarity between SFB and other microbes, we compared the corresponding functional profiles in these two matrixes. Specifically, in order to consider not only the number of shared gene families between SFB and another genome X, but also the number of genes lacking in one of the two compared organisms, we used the Tversky index [**3**, **4**], which is a generalization of the Jaccard index [**15**] defined as:

where is the number of gene families shared between SFB and X, the number of gene families in SFB but not X, the number of gene families in X but not SFB, and the parameter α is a positive value weighting the relative importance of "extra" genes versus "missing" genes. α=0.75 encourages greater similarity to genomes containing a high fraction of SFB's genes (i.e. X tends to be a superset of SFB), and α=0.25 upweights genomes with few genes not in the SFB genome (i.e. X tends to be a subset of SFB). Ranking all genomes in MBGD and KEGG according to their Tversky similarity with respect to SFB, we identified the 20 closest organisms both at α=0.75 and at α=0.25.

## Supervised selection of model organisms for targeted comparison with SFB

In addition to this unsupervised comparison of SFB with all organisms included in MBGD and KEGG, we further investigated the relationship of SFB to a manually curated set of organisms selected based on several biological criteria. Specifically, we selected a set of 13 strains representative of 13 species and 9 genera: [[need to be completed with the criteria used to select the model organisms]].

## Detection of pathways differentiating SFB from functionally similar organisms

We assessed the metabolic potential of SFB in comparison with the ~1,200 organisms in MBGD and KEGG by identifying pathways and small metabolic modules present in their genomes. Functional units of approximately 5 to 20 genes describing small pathways and structural complexes were defined using KEGG modules [**17**], which were organized into larger pathways and functional classes by the BRITE hierarchy [**1**]. Coverage (presence/absence) of each of the 371 KEGG modules was determined based on the fraction of its KOs present in each genome, and the metabolic profiles obtained in this manner for all genomes was processed as above to identify the genomes closest to SFB. The over- or under-enrichment of single pathways in SFB compared to its 20 closest organisms (using both Tversky α=0.75 and α=0.25) and model organisms was calculated as the *z*-score of SFB module abundance with respect to their average abundance in the set of similar genomes and visualized on the BRITE hierarchy using our tool for circular dendrogram visualization [**18**] (**Figure 4** and **Figure 5**).

## Detection and analysis of SFB core and accessory genes and metabolic modules

We refined this analysis of gene families by separating out orthologs present in the majority of organisms as *core* families and those less ubiquitous (but still commonly conserved) as *variable* families. Core and variable genes in the MBGD and KO catalogs were calculated as those present in at least 75% and in 5-25% of organisms, respectively. We defined a catalog of core and variable pathways identically using the KEGG module matrix detailed above. The comparisons described above for assessing the functional similarity of SFB to other organisms were then applied to the six subsets of core and variable MBGD gene families, KO genes, and KEGG modules (**Table 2**).

# FIGURES

## Figure: All reference genomes networked by KO similarity

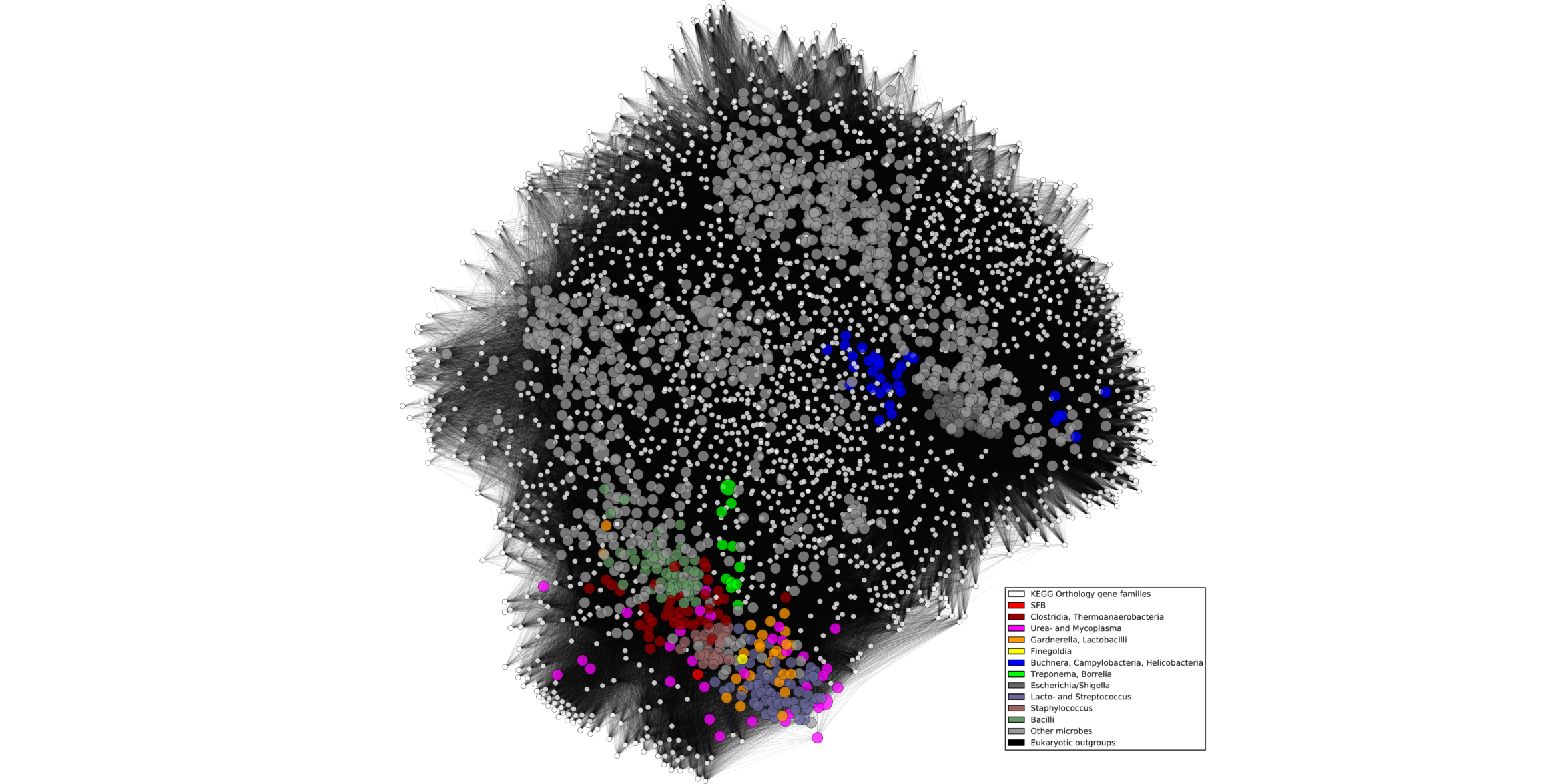


Figure 1. Overview of microbial functional similarities based on shared orthologous gene families. Small white nodes represent 13,118 KEGG Orthology gene families, and each of 1,191 larger nodes represents an organism (connected to each gene family in its genome and sized proportionally to its degree). SFB is highlighted to show overall similarity to the *Clostridia*, *Mycoplasma*, and other small genome pathobionts as detailed in the text.

## Figure: SFB similarity by KEGG modules, α=0.75

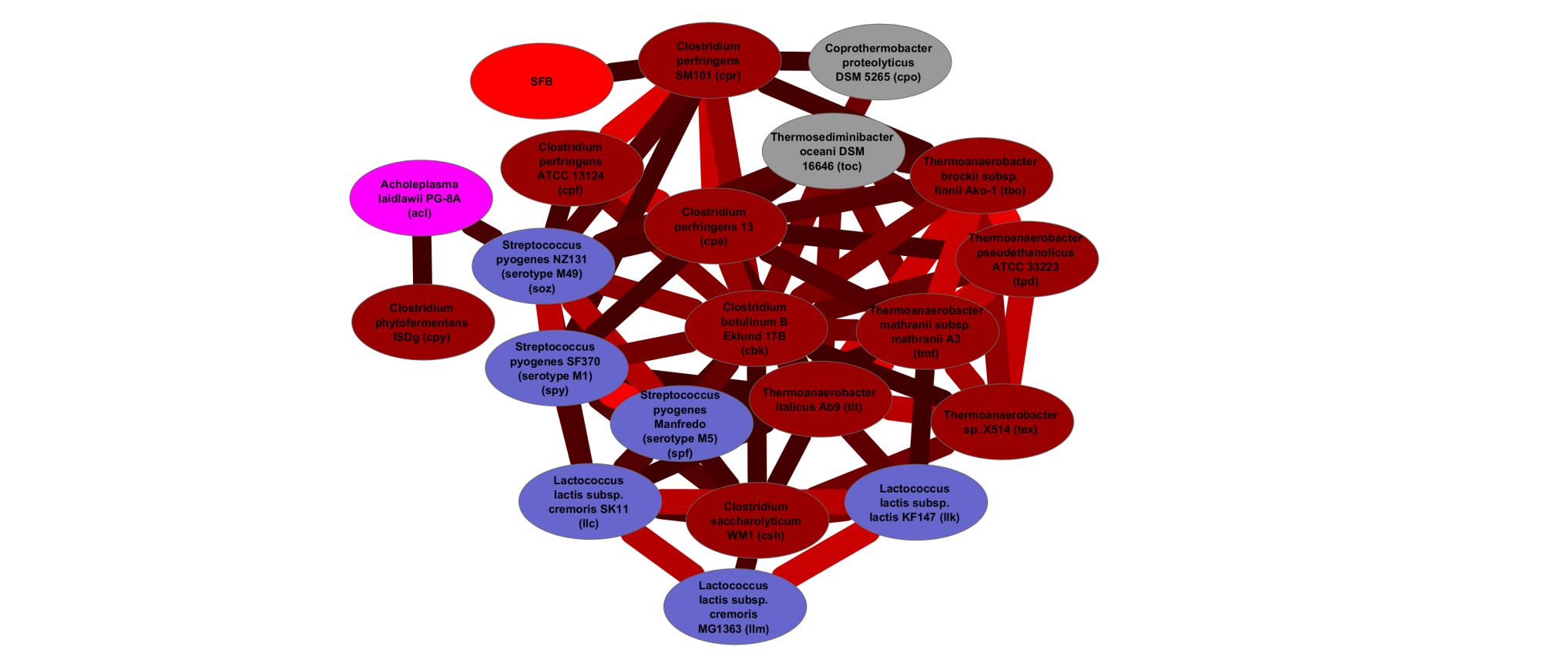


Figure 2. SFB's genomic similarity to other organisms based on shared metabolic modules. The fractional presence/absence of each of 371 KEGG metabolic modules was calculated within 1,191 microbial genomes, and the 20 most similar to SFB are shown here using two measures. The Tversky index [3, 4] is a generalization of the Jaccard index [15] used to calculate similarities between sets of continuous values. With higher α parameter values, it will associate genomes with few missing pathways relative to SFB; A) at α=0.75, the *Clostridia* and several *Strepto*- and *Lactococci* are found to carry a similar metabolic repertoire to SFB. B) {\*\*\*SEE FIGURE 2 FOR PANEL B. TO BE JOINED FOR THE PAPER\*\*\*} At lower α values, the Tversky index highlights genomes with few extra pathways not found in SFB, and with α=0.25, the *Mycoplasmas* in particular are shown to share a metabolic core.

## Figure: SFB similarity by KEGG modules, α=0.25

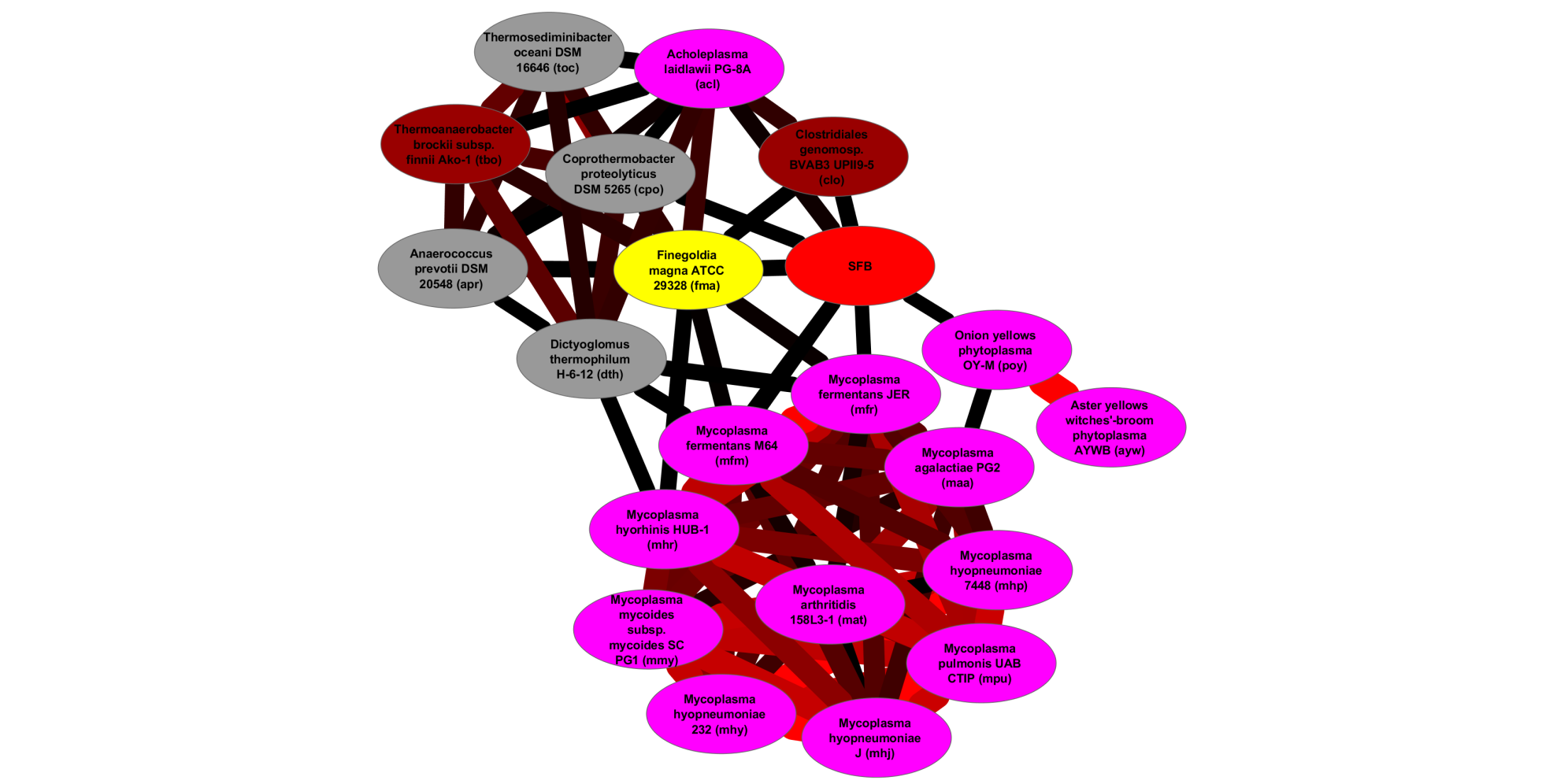


Figure 3. THIS IS THE PANEL B OF FIGURE 1. WE NEED TO JOIN THE TWO PANEL FOR THE PAPER

## Figure: KEGG modules over- and under-represented in SFB, α=0.75

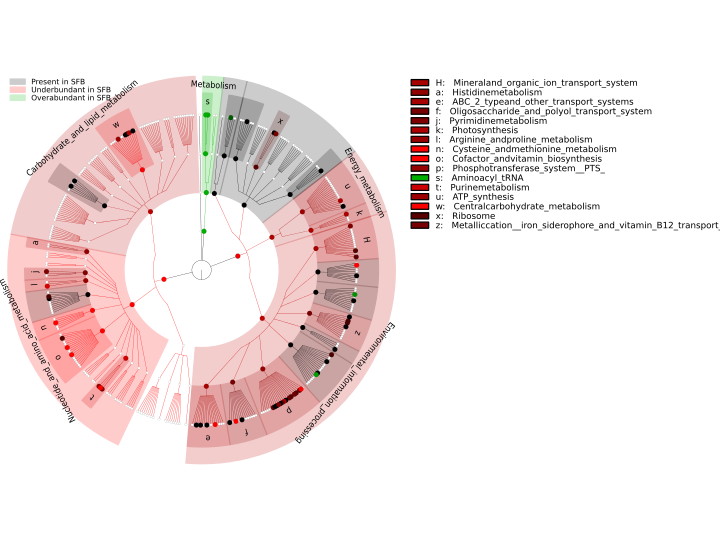


Figure 4. KEGG modules over- and under-represented in the SFB genome compared to its 20 most similar organisms. The KEGG functional hierarchy [1] is shown here, with leaves highlighting metabolic modules over- or under-enriched in SFB relative to functionally similar organisms. White circles represent modules with abundance consistent between SFB and other organisms (absolute z-score <1); green and red indicate z-scores >1.0 and <-1.0, respectively. Similar organisms were computed using A) Tversky α = 0.75 and B) α = 0.25. See Table 4 for a detailed list of KEGG modules differentiating SFB from functionally similar organisms such as the Clostridia.

## Figure: KEGG modules over- and under-represented in SFB, α=0.25

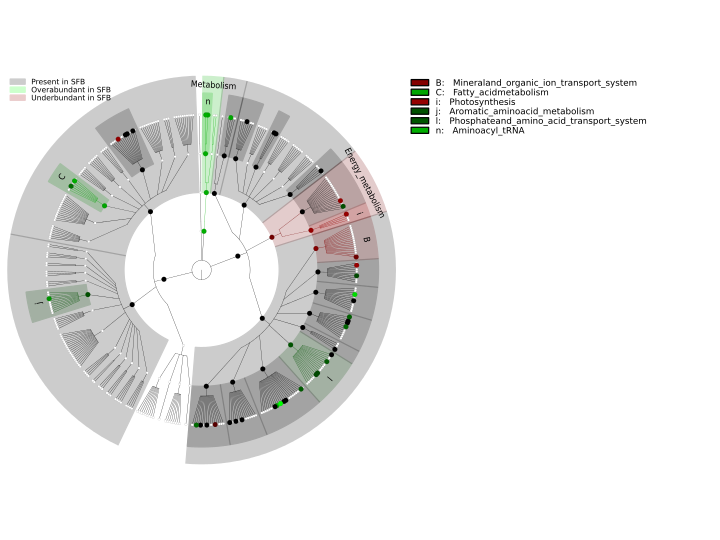


Figure 5. KEGG modules over- and under-represented in the SFB genome compared to its 20 most similar organisms (Tversky α = 0.25). The list of KEGG modules with absolute z-score higher than 1.0 are reported in Figure 5. {\*\*\*THIS SHOULD BE PANEL B OF FIGURE 4\*\*\*}

## Figure: SFB wheel diagram

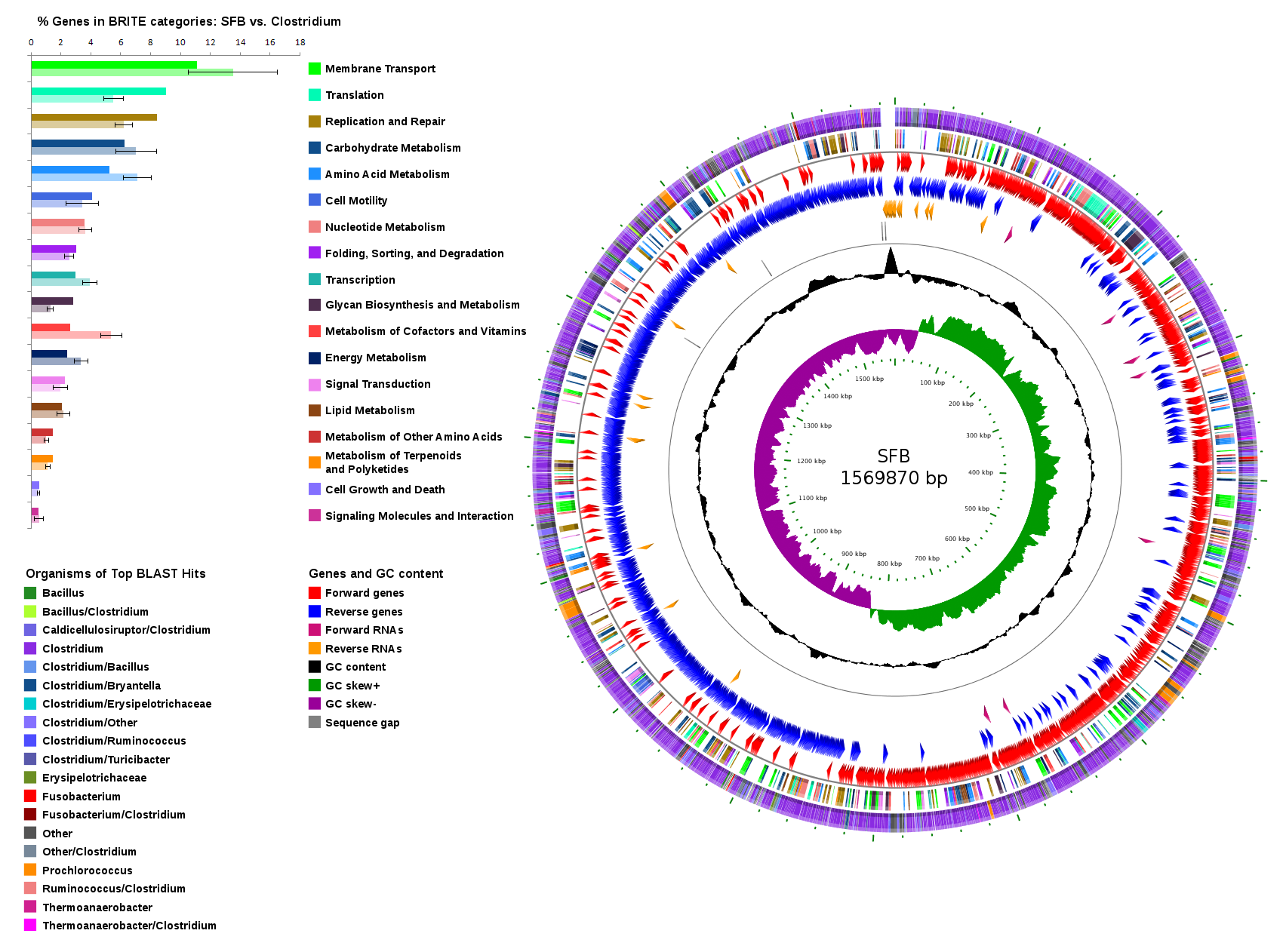


Figure 6. SFB wheel diagram

## Figure: SFB pathway overview graphic

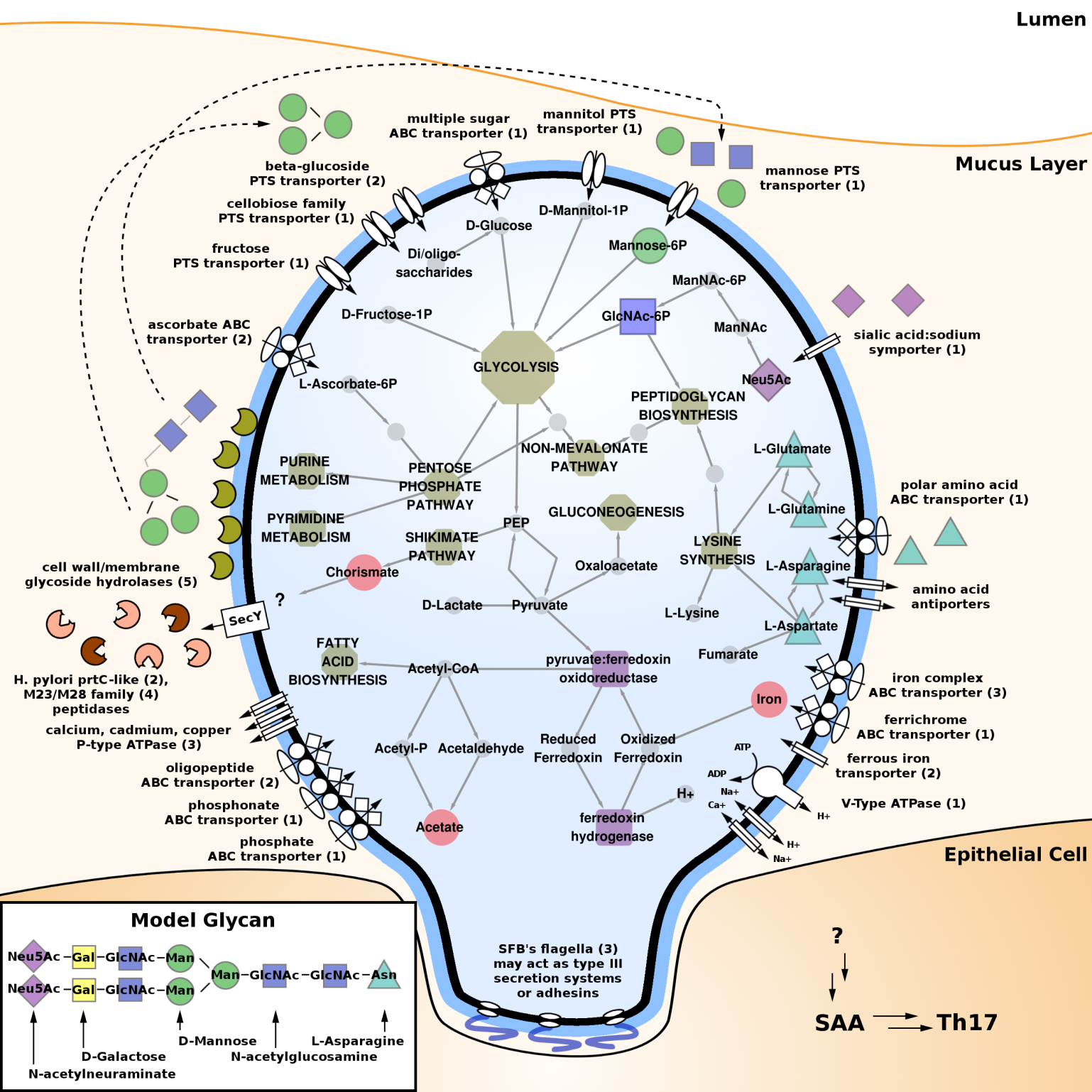


Figure 7. SFB pathway overview cartoon.

## Figure: ORF size distribution vs. Clostridium

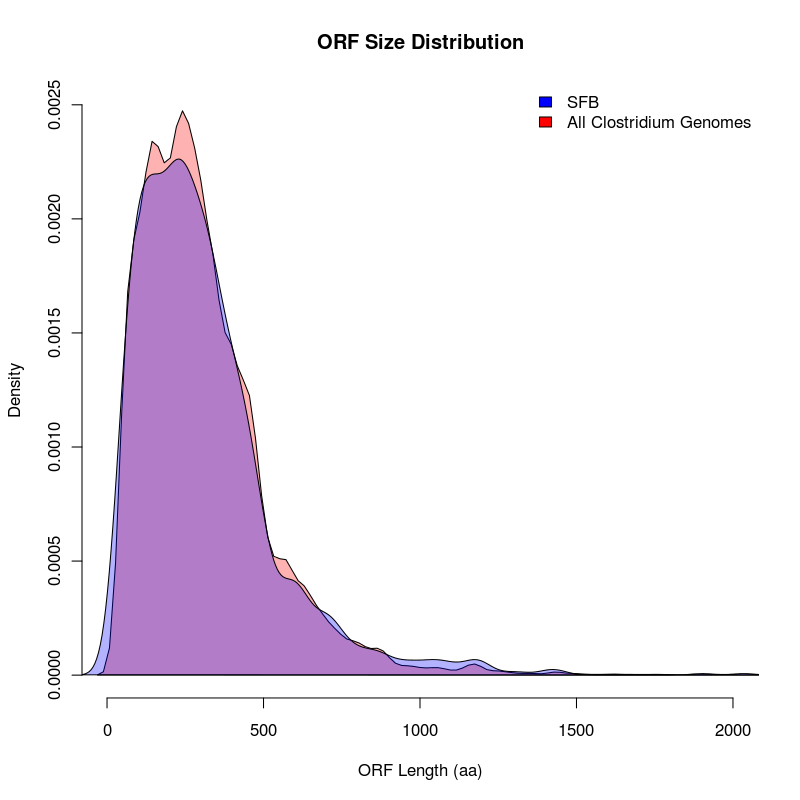


Figure 8. SFB ORF length distribution vs. Clostridium.

## Figure: Phylogenomic tree

Figure 9. Phylogenomic tree

## Figure: Phylogenetic tree

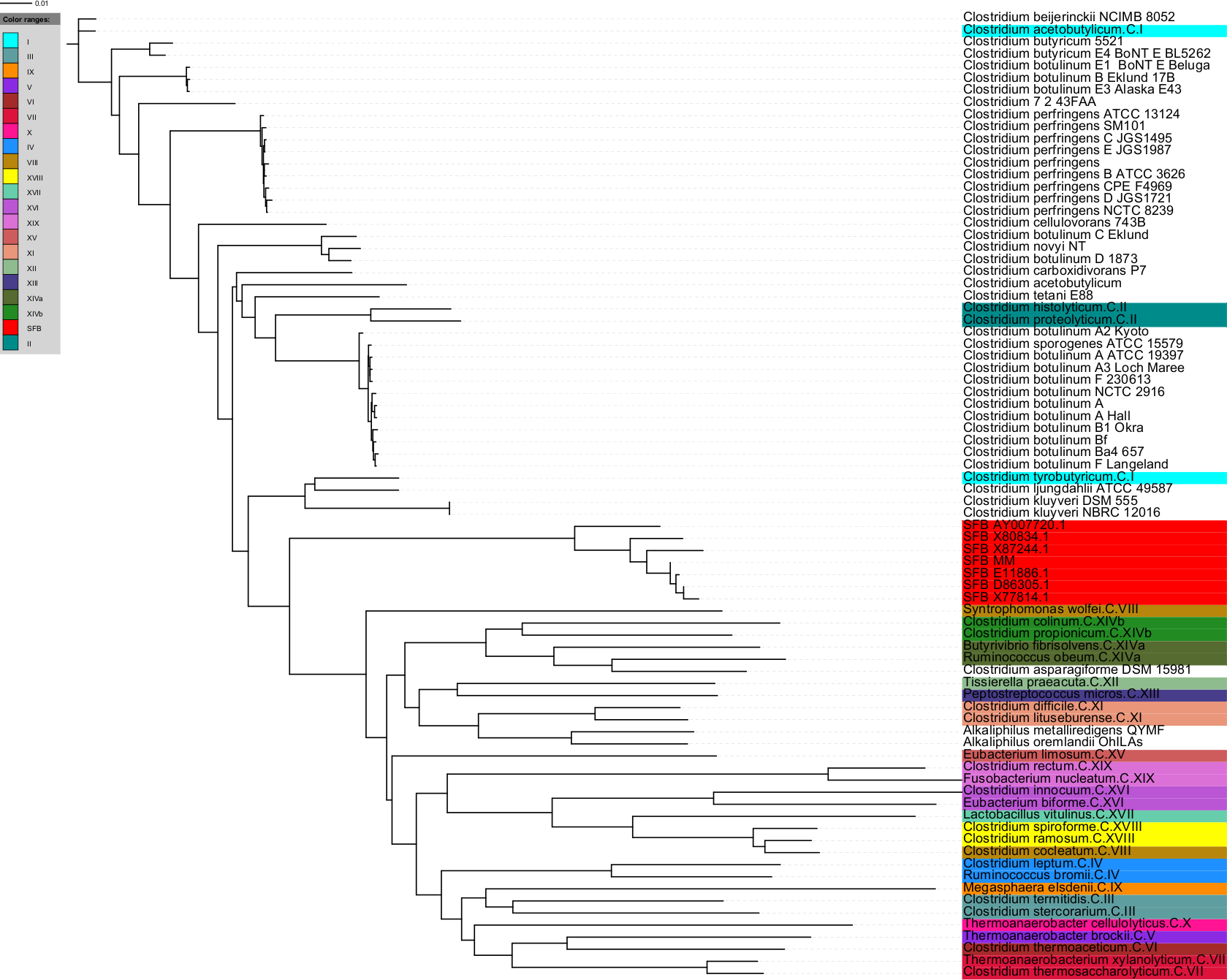
****

Figure 10. Phylogenetic tree

## Figure: Network of contig edge *k-mer* distance

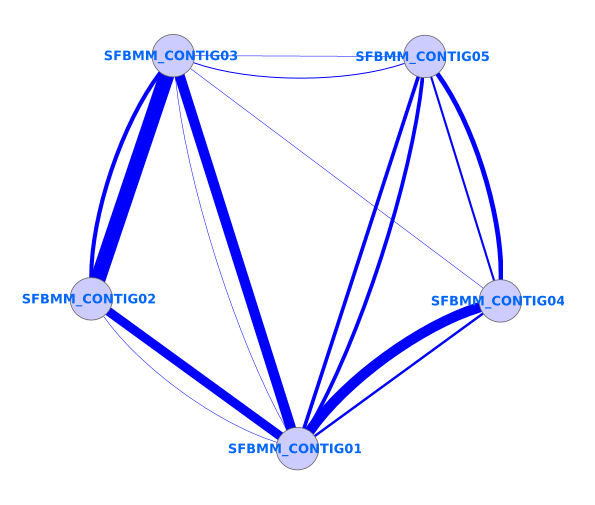


Figure 11. The *k-mer* profile of both ends of every contig was determined, and the Euclidean distance between the profiles of each contig end pair iteratively calculated as described in Methods. Edge thickness correlates with overall lower distance between *k-mer* profiles. Two edges connecting node pairs represent the L-R or R-L orientation. No edge is visualized in cases where a particular connection never minimized the Euclidean distance.

## Figure: Codon usage bias in *E. coli*, *Clostridium*, and SFB

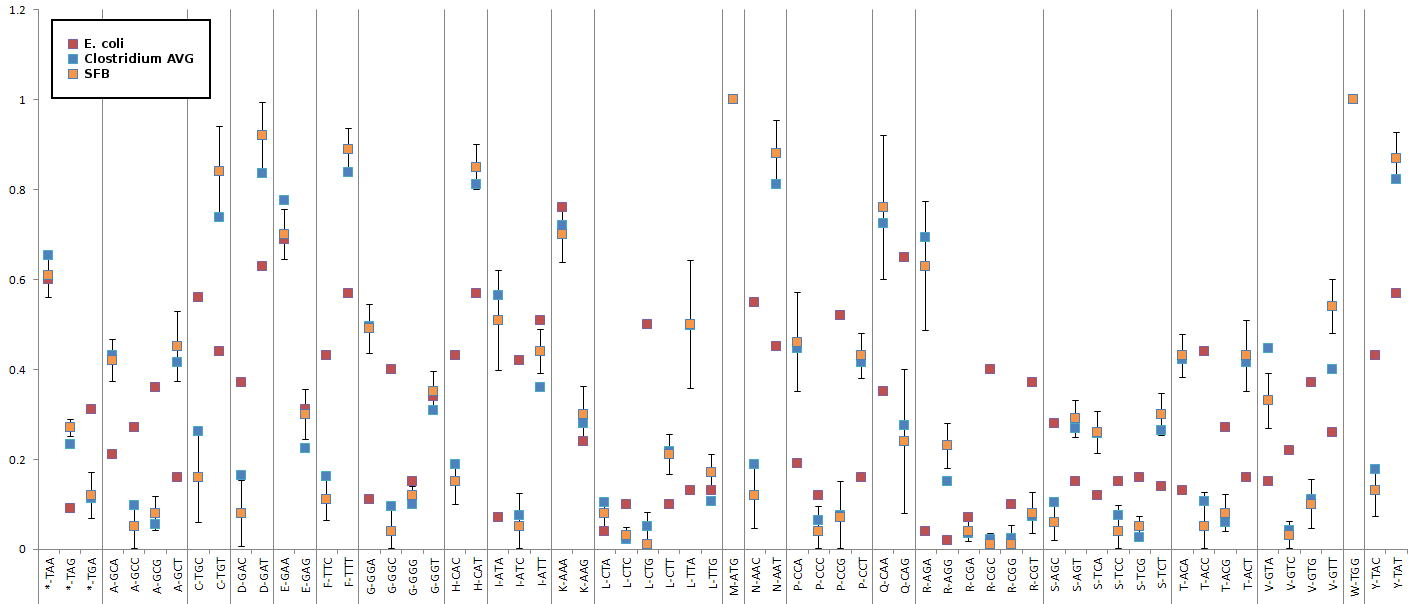


Figure 12. Codon usage bias in SFB is shifted toward codons ending in T or A, consistent with *Clostridium* species, while codon usage bias is different in a distantly related species, *E. coli*.

## Figure: Predicted genomic islands

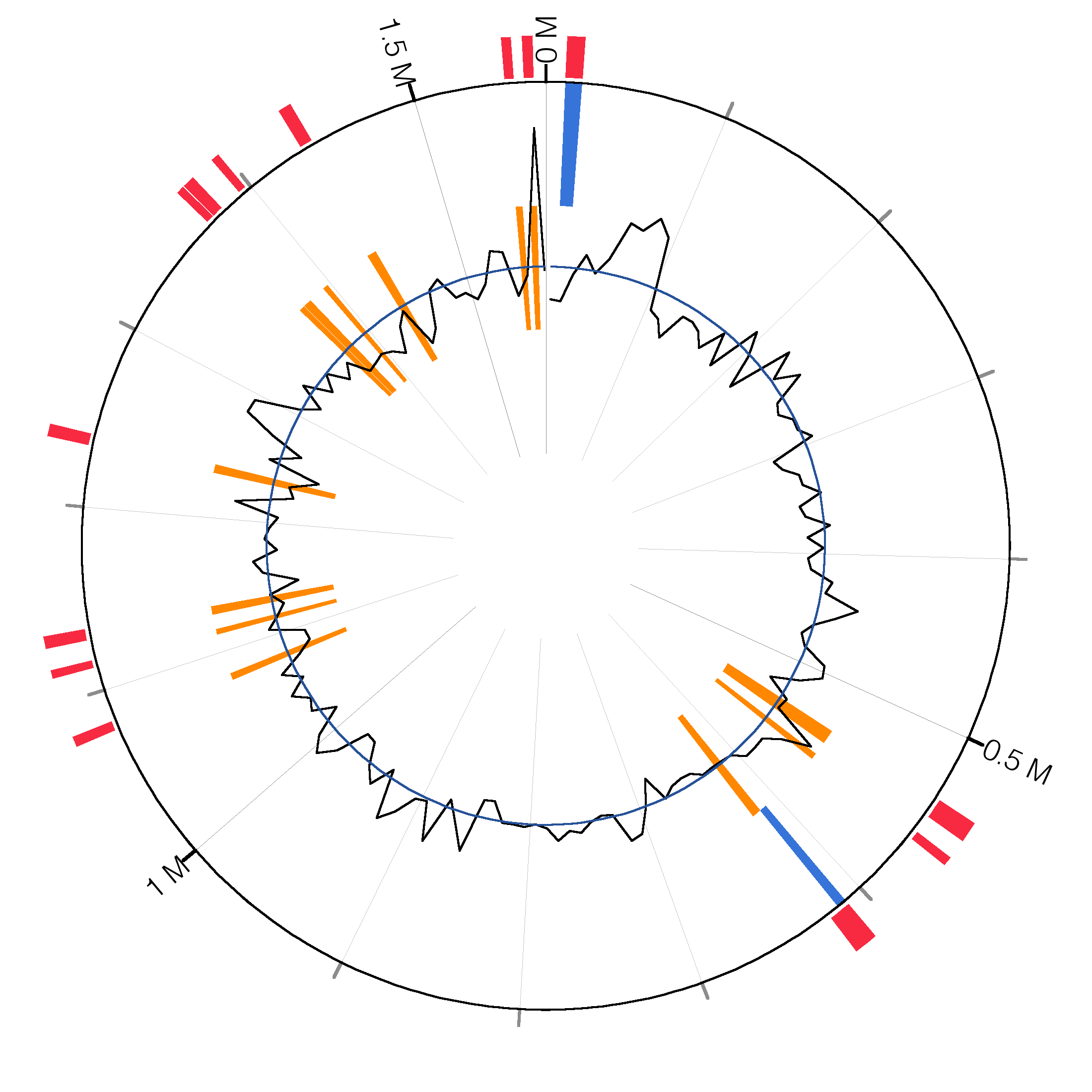


Figure 13. Genomic islands predicted in SFB.

## Figure: Ori-Finder

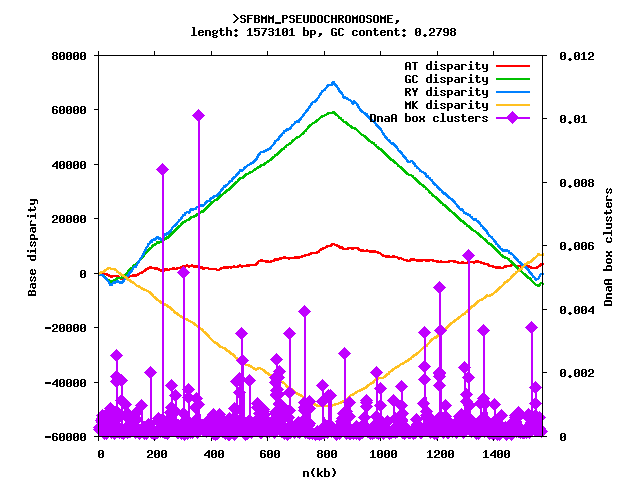


Figure . Ori-finder

# TABLES

## Table: Genera functionally similar to SFB by MDGB/KO/KEGG

Table 1. Organisms functionally similar to SFB based on shared gene families and metabolic modules. 1,191 microbial reference genomes were sorted by the specific orthologous gene families (using MDGB [2]), general gene families (using the KEGG Orthology [1]), or metabolic modules (small ~5-20 gene pathways) shared with SFB. Genera appearing at least twice among the 20 most similar organisms are shown here, using the Tversky similarity index with α=0.25 (emphasizing organisms with few pathways not carried by SFB) and with α=0.75 (emphasizing organisms missing few pathways carried by SFB). Percentages in parentheses refer to the fraction of the total number of available genomes in the correspondent genus present within the 20 genomes most similar to SFB.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Total genomes in genus | # of organisms within the 20 genomes sharing the most gene families | | | | | |
| Genus | MDGB | | KO | | KEGG | |
| α = 0.25 | α = 0.75 | α = 0.25 | α = 0.75 | α = 0.25 | α = 0.75 |
| Borrelia | **8** | 6 (75%) |  | 6 (75%) |  |  |  |
| Clostridium | **30** | 2 (7%) | 16 (53%) | 2 (7%) | 12 (40%) |  | 3 (10%) |
| Lactococcus | **4** |  |  |  |  |  | 3 (75%) |
| Mycoplasma | **26** | 10 (38%) |  | 4 (15%) |  | 10 (38%) |  |
| Streptococcus | **50** |  |  |  |  |  | 6 (12%) |
| Thermoanaerobacter | **7** |  | 4 (57%) |  | 6 (86%) |  | 6 (86%) |
| Ureaplasma | **3** |  |  | 3 (100%) |  |  |  |

## Table: Genera functionally similar to SFB by CORE/VARIABLE gene families

Table 2. Organisms similar to SFB in carriage of "core" and "variable" gene families and metabolic modules. The MDGB, KO, and KEGG module catalogs were split into core (present in at least 75% of available genomes) and variable (present in 5% to 25%) subsets, and the reference genomes most similar to SFB in these subsets are shown here. SFB carries a core gene set of average size and a slightly reduced variable gene and pathway set attributable to its reduced genome size.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Genus | Total genomes in genus | # of organisms within the 20 genomes sharing the most **CORE** gene families | | | | | |
| MDGB  α = 0.25 α = 0.75 | | KO  α = 0.25 α = 0.75 | | KEGG  α = 0.25 α = 0.75 | |
| Bifidobacterium | 14 |  |  |  |  |  | 2 (14%) |
| Borrelia | 8 | 7 (88%) |  | 6 (75%) |  | 8 (100%) |  |
| Clostridium | 30 |  | 11 (37%) |  |  |  |  |
| Haemophilus | 10 |  |  | 2 (20%) |  |  |  |
| Lactobacillus | 25 | 5 (20%) |  |  |  |  |  |
| Mycoplasma | 26 |  |  | 5 (19%) |  | 2 (8%) |  |
| Propionibacterium | 3 |  |  |  | 2 (67%) |  |  |
| Streptococcus | 50 | 6 (12%) | 8 (16%) |  |  |  | 9 (18%) |
| Thermoanaerobacter | 7 |  |  |  | 2 (29%) |  |  |
| Thermus | 3 |  |  |  |  |  | 2 (67%) |
| Treponema | 3 | 2 (67%) |  |  |  | 2 (67%) |  |
| Tropheryma | 2 |  |  |  |  |  | 2 (100%) |
| Ureaplasma | 3 |  |  |  |  | 3 (100%) |  |
| Genus | Total  genomes  in genus | # of organisms within the 20 genomes sharing the most **VARIABLE** gene families | | | | | |
| MDGB  α = 0.25 α = 0.75 | | KO  α = 0.25 α = 0.75 | | KEGG  α = 0.25 α = 0.75 | |
| Blattabacterium | 2 |  |  |  |  | 2 (100%) |  |
| Clostridium | 30 | 15 (50%) | 18 (60%) | 6 (20%) | 12 (40%) |  | 6 (20%) |
| Lactococcus | 4 |  |  |  |  | 2 (50%) |  |
| Mycoplasma | 26 |  |  |  |  | 3 (12%) |  |
| Porphyromonas | 2 |  |  |  |  | 2 (100%) |  |
| Streptococcus | 50 |  |  | 5 (10%) |  | 4 (8%) | 9 (18%) |
| Thermoanaerobacter | 7 | 4 (57%) | 2 (29%) | 6 (86%) | 6 (86%) | 2 (29%) | 4 (57%) |

## Table: Model organisms similarity to SFB

Table 3. SFB's similarity in gene and pathway carriage to selected model microbes. Based on phylogenetic, functional, genomic, or niche similarity, we selected 13 organisms with which to individually compare SFB's MDGB orthologous family carriage, KOs, or KEGG metabolic modules. Comparisons were performed using the Dice index, i.e. Tversky α=0.5 for simplicity [3, 4, 16]. As detailed in the text, SFB demonstrates considerable phylogenetic and functional similarity with the Clostridia and moderate functional similarity with several reduced genome pathobionts. It is functionally distinct even from these, however, and particularly so with respect to gut organisms with potentially similar environmental niches (e.g. C. jejuni or H. pylori).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Organism | MDGB | | KO | | KEGG modules | |
| Total | Dice | Total | Dice | Total | Dice |
| SFB | 1003 | 1 | 792 | 1 | 31 | 1 |
| Borrelia burgdorferi | 669 | 0.39 | 526 | 0.51 | 25 | 0.43 |
| Buchnera aphidicola | 483 | 0.35 | 498 | 0.42 | 27 | 0.34 |
| Campylobacter jejuni | 1281 | 0.34 | 1027 | 0.45 | 52 | 0.39 |
| Clostridium botulinum | 1991 | 0.43 | 1573 | 0.54 | 61 | 0.52 |
| Clostridium kluyveri | 1982 | 0.41 | 1529 | 0.52 | 59 | 0.46 |
| Clostridium novyi | 1515 | 0.48 | 1176 | 0.59 | 51 | 0.50 |
| Clostridium perfringens | 1628 | 0.47 | 1307 | 0.58 | 58 | 0.58 |
| Finegoldia magna | 1429 | 0.39 | 875 | 0.57 | 30 | 0.63 |
| Gardnerella vaginalis | 1053 | 0.32 | 698 | 0.47 | 37 | 0.55 |
| Helicobacter hepaticus | 1154 | 0.34 | 984 | 0.46 | 47 | 0.31 |
| Helicobacter pylori | 1017 | 0.35 | 893 | 0.47 | 42 | 0.34 |
| Mycoplasma pneumoniae | 340 | 0.32 | 360 | 0.45 | 23 | 0.47 |
| Treponema denticola | 1276 | 0.36 | 1024 | 0.50 | 29 | 0.53 |

## Table: KEGG modules over- and under-represented in SFB, α=0.75

Table 4. KEGG modules over- (z-score > 1.0) and under-enriched (z-score < -1.0) in SFB with respect to its 20 closest organisms (Tversky α=0.75).

| **KEGG module code** | **KEGG module name** | **z-score** |
| --- | --- | --- |
| M00140 | C1-unit interconversion, prokaryotes | < -3.5 |
| M00021 | Cysteine biosynthesis, serine => cysteine | < -3.5 |
| M00276 | PTS system, mannose-specific II component | < -3.5 |
| M00120 | Coenzyme A biosynthesis, pantothenate => CoA | < -3.5 |
| M00221 | Simple sugar transport system | < -3.5 |
| M00248 | Antibiotic transport system | < -3.5 |
| M00048 | Inosine monophosphate biosynthesis, PRPP + glutamine => IMP | < -3.5 |
| M00194 | Maltose/maltodextrin transport system | < -3.5 |
| M00015 | Proline biosynthesis, glutamate => proline | -2.933 |
| M00273 | PTS system, fructose-specific II component | -2.4 |
| M00007 | Pentose phosphate pathway, non-oxidative phase, fructose 6P => | -2.4 |
| M00164 | ATP synthase | -2.4 |
| M00157 | F-type ATPase, bacteria | -2.4 |
| M00307 | Pyruvate oxidation, pyruvate => acetyl-CoA | -2.1 |
| M00050 | Guanine nucleotide biosynthesis, IMP => GDP/dGDP,GTP/dGTP | -2.0 |
| M00049 | Adenine nucleotide biosynthesis, IMP => ADP/dADP,ATP/dATP | -2.0 |
| M00209 | Osmoprotectant transport system | -1.7 |
| M00052 | Pyrimidine ribonucleotide biosynthesis, UMP => UDP/UTP,CDP/CTP | -1.7 |
| M00270 | PTS system, trehalose-specific II component | -1.7 |
| M00178 | Ribosome, bacteria | -1.7 |
| M00266 | PTS system, maltose and glucose-specific II component | -1.6 |
| M00303 | PTS system, N-acetylmuramic acid-specific II component | -1.5 |
| M00272 | PTS system, arbutin-, cellobiose-, and salicin-specific II component | -1.5 |
| M00279 | PTS system, galactitol-specific II component | -1.5 |
| M00269 | PTS system, sucrose-specific II component | -1.3 |
| M00299 | Spermidine/putrescine transport system | -1.2 |
| M00268 | PTS system, arbutin-like II component | -1.2 |
| M00265 | PTS system, glucose-specific II component | -1.2 |
| M00188 | Sulfonate/nitrate/taurine transport system | -1.1 |
| M00125 | Riboflavin biosynthesis, GTP => riboflavin/FMN/FAD | -1.1 |
| M00238 | D-Methionine transport system | -1.1 |
| M00246 | Nickel transport system | -1 |
| M00245 | Cobalt transport system | -1 |
| M00023 | Tryptophan biosynthesis, chorismate => tryptophan | -1 |
| M00311 | 2-oxoglutarate:ferredoxin oxidoreductase | 1.2 |
| M00331 | Type II general secretion system | 2 |
| M00223 | Phosphonate transport system | 2.2 |
| M00360 | Aminoacyl-tRNA biosynthesis, prokaryotes | 2.4 |

## Table: KEGG modules over- and under-represented in SFB, α=0.25

Table 5. KEGG modules over- (z-score > 1.0) and under-enriched (z-score < -1.0) in SFB with respect to its 20 closest organisms (Tversky α=0.25).

|  |  |  |
| --- | --- | --- |
| **KEGG module code** | **KEGG module name** | **z-score** |
| M00221 | Simple sugar transport system | -2.1 |
| M00157 | F-type ATPase, bacteria | -2.0 |
| M00164 | ATP synthase | -2.0 |
| M00307 | Pyruvate oxidation, pyruvate => acetyl-CoA | -2.0 |
| M00299 | Spermidine/putrescine transport system | -1.5 |
| M00248 | Antibiotic transport system | -1.2 |
| M00223 | Phosphonate transport system | 1.0 |
| M00212 | Ribose transport system | 1.0 |
| M00222 | Phosphate transport system | 1.0 |
| M00159 | V-type ATPase, prokaryotes | 1.1 |
| M00236 | Polar amino acid transport system | 1.2 |
| M00276 | PTS system, mannose-specific II component | 1.2 |
| M00242 | Zinc transport system | 1.3 |
| M00240 | Iron complex transport system | 1.4 |
| M00256 | Cell division transport system | 1.5 |
| M00083 | Fatty acid biosynthesis, elongation | 1.5 |
| M00022 | Shikimate pathway, phosphoenolpyruvate + erythrose-4P => chorismate | 2.0 |
| M00311 | 2-oxoglutarate:ferredoxin oxidoreductase | 2.3 |
| M00086 | beta-Oxidation, acyl-CoA synthesis | 2.4 |
| M00360 | Aminoacyl-tRNA biosynthesis, prokaryotes | 2.4 |
| M00275 | PTS system, cellobiose-specific II component | 2.4 |
| M00359 | Aminoacyl-tRNA biosynthesis, eukaryotes | 2.4 |
| M00331 | Type II general secretion system | 3.0 |
| M00271 | PTS system, beta-glucosides-specific II component | > 3.5 |

## Table: KOs in SFB but not similar organisms, α=0.75

Table 6. KOs present in SFB but not in the 20 most similar organisms (Tversky alpha = 0.25).

|  |
| --- |
| K00971: mannose-1-phosphate guanylyltransferase [EC:2.7.7.22] |
| K07474: phage terminase small subunit |
| K00058: D-3-phosphoglycerate dehydrogenase [EC:1.1.1.95] |
| K00208: enoyl-[acyl-carrier protein] reductase I [EC:1.3.1.9] |
| K00680 |
| K00970: poly(A) polymerase [EC:2.7.7.19] |
| K00996: undecaprenyl-phosphate galactose phosphotransferase [EC:2.7.8.6] |
| K01191: alpha-mannosidase [EC:3.2.1.24] |
| K01205: alpha-N-acetylglucosaminidase [EC:3.2.1.50] |
| K01297: muramoyltetrapeptide carboxypeptidase [EC:3.4.17.13] |
| K01305: beta-aspartyl-dipeptidase (metallo-type) [EC:3.4.19.-] |
| K01412: mitochondrial processing peptidase [EC:3.4.24.64] |
| K01476: arginase [EC:3.5.3.1] |
| K01485: cytosine deaminase [EC:3.5.4.1] |
| K01567 |
| K01605: methylmalonyl-CoA decarboxylase beta chain [EC:4.1.1.41] |
| K01771: 1-phosphatidylinositol phosphodiesterase [EC:4.6.1.13] |
| K01953: asparagine synthase (glutamine-hydrolysing) [EC:6.3.5.4] |
| K01960: pyruvate carboxylase subunit B [EC:6.4.1.1] |
| K02405: RNA polymerase sigma factor for flagellar operon FliA |
| K02454: general secretion pathway protein E |
| K02483: two-component system, OmpR family, response regulator |
| K03080: L-ribulose-5-phosphate 4-epimerase [EC:5.1.3.4] |
| K03294: basic amino acid/polyamine antiporter, APA family |
| K03327: multidrug resistance protein, MATE family |
| K03483: mannitol operon transcriptional antiterminator |
| K03610: septum site-determining protein MinC |
| K03718: Lrp/AsnC family transcriptional regulator, regulator for asnA, asnC and gidA |
| K03771: peptidyl-prolyl cis-trans isomerase SurA [EC:5.2.1.8] |
| K03781: catalase [EC:1.11.1.6] |
| K03802: cyanophycin synthetase [EC:6.-.-.-] |
| K04074: cell division initiation protein |
| K04076: Lon-like ATP-dependent protease [EC:3.4.21.-] |
| K04478: monofunctional glycosyltransferase [EC:2.4.1.-] |
| K04750: PhnB protein |
| K04769: AbrB family transcriptional regulator, stage V sporulation protein T |
| K05540: tRNA-dihydrouridine synthase B [EC:1.-.-.-] |
| K05546: alpha 1,3-glucosidase [EC:3.2.1.84] |
| K05770: benzodiazapine receptor |
| K05916: nitric oxide dioxygenase [EC:1.14.12.17] |
| K06012: spore protease [EC:3.4.24.78] |
| K06283: putative DeoR family transcriptional regulator, stage III sporulation protein D |
| K06284: transcriptional pleiotropic regulator of transition state genes |
| K06310: spore germination protein |
| K06331: spore coat protein I |
| K06378: stage II sporulation protein AA (anti-sigma F factor antagonist) |
| K06379: stage II sporulation protein AB (anti-sigma F factor) [EC:2.7.11.1] |
| K06382: stage II sporulation protein E [EC:3.1.3.16] |
| K06385: stage II sporulation protein P |
| K06387: stage II sporulation protein R |
| K06390: stage III sporulation protein AA |
| K06393: stage III sporulation protein AD |
| K06394: stage III sporulation protein AE |
| K06397: stage III sporulation protein AH |
| K06398: stage IV sporulation protein A |
| K06399: stage IV sporulation protein B [EC:3.4.21.116] |
| K06405: stage V sporulation protein AC |
| K06406: stage V sporulation protein AD |
| K06407: stage V sporulation protein AE |
| K06409: stage V sporulation protein B |
| K06410: dipicolinate synthase subunit A |
| K06411: dipicolinate synthase subunit B |
| K06436: spore coat assemly protein |
| K06438: similar to stage IV sporulation protein |
| K07012 |
| K07149 |
| K07217: Mn-containing catalase |
| K07322: regulator of cell morphogenesis and NO signaling |
| K07642: two-component system, OmpR family, sensor histidine kinase BaeS [EC:2.7.13.3] |
| K07659: two-component system, OmpR family, phosphate regulon response regulator OmpR |
| K07699: two-component system, response regulator, stage 0 sporulation protein A |
| K07718: two-component system, sensor histidine kinase YesM [EC:2.7.13.3] |
| K07772: two-component system, OmpR family, torCAD operon response regulator TorR |
| K07774: two-component system, OmpR family, response regulator TctD |
| K07775: two-component system, OmpR family, response regulator ResD |
| K07816: putative GTP pyrophosphokinase [EC:2.7.6.5] |
| K07979: GntR family transcriptional regulator |
| K08316: ribosomal RNA small subunit methyltransferase D [EC:2.1.1.171] |
| K08641: D-alanyl-D-alanine dipeptidase [EC:3.4.13.-] |
| K08978: putative membrane protein |
| K09704: hypothetical protein |
| K10974: cytosine permease |
| K11166: dehydrogenase/reductase SDR family member 7B [EC:1.1.-.-] |
| K11991: tRNA-specific adenosine deaminase [EC:3.5.4.-] |
| K12264: anaerobic nitric oxide reductase flavorubredoxin |
| K12942: aminobenzoyl-glutamate transport protein |
| K13051: beta-aspartyl-peptidase (threonine type) [EC:3.4.19.5] |
| K13282: cyanophycinase [EC:3.4.15.6] |
| K13643: Rrf2 family transcriptional regulator, iron-sulfur cluster assembly transcription factor |
| K13685: UDP-N-acetylglucosamine:undecaprenyl-P N-acetylglucosaminyl 1-P transferase [EC:2.7.8.-] |
| K13688: cyclic beta-1,2-glucan synthetase [EC:2.4.1.-] |
| K13695: probable lipoprotein NlpC |
| K13788: phosphate acetyltransferase [EC:2.3.1.8] |
| K14058: tRNA 2-thiocytidine biosynthesis protein TtcA |

## Table: General genome features

Table 7. SFB features compared to the Clostridium average.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **SFB** | **Clostridium AVG** | **Clostridium SD** |
| Genome Size | 1569870 | 3974341 | 739004 |
| G+C % | 27.9 | 30.57 | 4.04 |
| Features | 1589 | 3678 | 639 |
| ORFs | 1533 | 3525 | 612 |
| ORF Density | 1024 | 1127 | 57 |
| tRNAs | 38 | 77 | 15 |
| rRNA Operons | 2 | 9 | 2 |
|  |  |  |  |
|  |  |  |  |
| Signal Peptidase I | 80 | 192 | 50 |
| Signal Peptidase II | 46 | 84 | 26 |
| Cell Wall | 20 | 29 | 9 |
| Membrane | 326 | 862 | 153 |
| Extracellular | 34 | 67 | 26 |
|  |  |  |  |
|  |  |  |  |
| Annotated (All) | 1137 | 2424 | 539 |
| Annotated (KAAS) | 792 | 1481 | 178 |
| Hypothetical | 396 | 1101 | 454 |

## Table: Depth of coverage, length, and features by contig

Table 8. Coverage, length, and features of SFB contigs.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Contig** | **Size (bp)** | **Features** | **Coverage** | **Coverage (normalized)** |
| SFBMM\_CONTIG01 | 1317732 | 1329 | 332.965281282 | 0.9760267 |
| SFBMM\_CONTIG02 | 128113 | 144 | 373.371973399 | 1.094471512 |
| SFBMM\_CONTIG03 | 110903 | 98 | 340.238638823 | 0.997347214 |
| SFBMM\_CONTIG04 | 9855 | 10 | 1029.14762581 | 3.016757654 |
| SFBMM\_CONTIG05 | 3267 | 5 | 331.156364749 | 0.970724192 |
| Average | - | - | 341.143619715 | 1 |

## Table: Two-component systems

Table 9. SFB’s two-component systems

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Histidine kinase** | **Response regulator** | **Predicted function** | **Adjacent to** | **Reference** |
| SFBMM\_001100 | SFBMM\_001090 | Unknown | Sporulation, housekeeping genes |  |
| SFBMM\_004310 | SFBMM\_004300 | Unknown | Iron ABC transporters |  |
| SFBMM\_004680 | SFBMM\_004670 | Unknown | Flagellar operon |  |
| SFBMM\_007600 | SFBMM\_007590 | Phosphate regulation | Phosphate ABC transporter | [1] |
| SFBMM\_007860 | SFBMM\_007870 | Sugar sensing, araC family | Multiple sugar ABC transporter | [2] |
| SFBMM\_010110 | SFBMM\_010090 | Chemotaxis | Flagellar operon | [3] |
| SFBMM\_010770 | - | Unknown | Hypothetical genes |  |
| SFBMM\_011190 | SFBMM\_011200 | Unknown | Hypothetical genes |  |
| SFBMM\_012290 | SFBMM\_012300 | Unknown | Hypothetical genes |  |
| SFBMM\_013300 | SFBMM\_013310 | Unknown | Sialic acid utilization operon |  |

[1] <http://www.ncbi.nlm.nih.gov/pubmed/2824439>

[2] <http://www.ncbi.nlm.nih.gov/pubmed/10524254>

[3] <http://www.ncbi.nlm.nih.gov/pubmed/21283116>

## Table: Transcriptional regulators

Table 10. SFB’s annotated transcriptional regulators

|  |  |
| --- | --- |
| **Locus tag** | **Annotation** |
| SFBMM\_000230 | PadR family transcriptional regulator |
| SFBMM\_002380 | LacI family transcriptional regulator |
| SFBMM\_003670 | AsnC family transcriptional regulator |
| SFBMM\_004040 | fur family transcriptional regulator |
| SFBMM\_005030 | recombination regulator |
| SFBMM\_005640 | DeoR family transcriptional regulator, fructose operon |
| SFBMM\_007360 | iron-sulfur cluster assembly transcription regulator |
| SFBMM\_007420 | fur family transcriptional regulator |
| SFBMM\_007650 | phosphate transport system regulatory protein |
| SFBMM\_008070 | GntR family transcriptional regulator |
| SFBMM\_008670 | TetR/AcrR family transcriptional regulator |
| SFBMM\_009220 | arginine repressor protein |
| SFBMM\_010010 | carbon storage regulator |
| SFBMM\_012480 | TetR/AcrR family transcriptional regulator |
| SFBMM\_012610 | ArsR family transcriptional regulator |
| SFBMM\_012680 | redox-sensing transcriptional repressor |
| SFBMM\_013160 | fur family transcriptional regulator |
| SFBMM\_014490 | PemK family protein |
| SFBMM\_014650 | cell envelope-related transcriptional attenuator domain-containing protein |
| SFBMM\_014710 | cell envelope-related transcriptional attenuator domain-containing protein |
| SFBMM\_014740 | AbrB family transcriptional regulator |
| SFBMM\_015030 | LacI family transcriptional regulator |
| SFBMM\_015170 | TetR/AcrR family transcriptional regulator |
| SFBMM\_015370 | ArsR family transcriptional regulator |

## Table: HTH-containing ORFs

Table 11. Putative transcriptional regulators containing a helix-turn-helix motif.

|  |  |  |  |
| --- | --- | --- | --- |
| **Locus tag** | **Domains/motifs** | | **Predicted to regulate** |
| SFBMM\_000720 | Cupin\_2 | PF0788 | Unknown |
| HTH\_3 | PF01381 |
| SFBMM\_007670 | HTH\_3 | PF01381 | Unknown |
| SFBMM\_008590 | PRD | PF00874 | Ascorbate, beta-glucoside uptake and utilization operon |
| PTS\_EIIA\_2 | PF0035 |
| HTH\_11 | PF08279 |
| SFBMM\_010620 | HTH\_6 | PF01418 | Unknown |
| SFBMM\_012200 | HTH\_3 | PF01381 | Unknown |
| SFBMM\_012610 | HTH\_5 | PF01022 | Unknown |
| SFBMM\_013660 | HTH\_3 | PF01381 | Unknown |
| SFBMM\_013990 | DUF2442 | PF10387 | Unknown |
| HTH\_3 | PF01381 |
| SFBMM\_014040 | HTH\_3 | PF01381 | Unknown |
| SFBMM\_014160 | HTH\_3 | PF01381 | Unknown |
| SFBMM\_014170 | HTH\_3 | PF01381 | Unknown |
| SFBMM\_014180 | HTH\_3 | PF01381 | Unknown |
| SFBMM\_014200 | Peptidase\_S24 | P00717 | Unknown |
| HTH\_3 | PF01381 |
| SFBMM\_014330 | HTH\_WhiA | PF02650 | Unknown |
| SFBMM\_015370 | HTH\_5 | PF01022 | Unknown |
| SFBMM\_015410 | PRD | PF00874 | Mannitol uptake and utilization operon |
| HTH\_11 | PF08279 |

## Table: Glycan metabolism

Table 12. SFB proteins putatively involved in glycan metabolism.

|  |  |  |
| --- | --- | --- |
| **Locus tag** | **Predicted function** | **Predicted localization** |
| **Transporters** | | |
| SFBMM\_005570 | PTS system, cellobiose-specific IIC component | Membrane |
| SFBMM\_005730 | PTS system, cellobiose-specific IIB component | Cytoplasmic |
| SFBMM\_005740 | PTS system, cellobiose-specific IIC component | Membrane |
| SFBMM\_006070 | PTS system, cellobiose-specific IIA component | Cytoplasmic |
| SFBMM\_008640 | PTS system, beta-glucoside-specific IIABC component | Membrane |
| SFBMM\_008740 | PTS system, beta-glucoside-specific IIABC component | Membrane |
| SFBMM\_012520 | PTS system, mannose-specific IID component | Membrane |
| SFBMM\_012530 | PTS system, mannose-specific IIC component | Membrane |
| SFBMM\_012540 | PTS system, mannose-specific IIAB component | Cytoplasmic |
| SFBMM\_007880 | multiple sugar ABC transporter, permease protein | Membrane |
| SFBMM\_007890 | multiple sugar ABC transporter, permease protein | Membrane |
| SFBMM\_007910 | multiple sugar ABC transporter, substrate-binding protein | Cell Wall / Extracellular |
| SFBMM\_007940 | multiple sugar ABC transporter, ATP-binding protein | Membrane |
| SFBMM\_004930 | oligopeptide ABC transporter, substrate-binding protein | Cell Wall |
| SFBMM\_005470 | oligopeptide ABC transporter, substrate-binding protein | Cell Wall / Extracellular |
| SFBMM\_011340 | oligopeptide ABC transporter, substrate-binding protein | Cell Wall / Extracellular |
| SFBMM\_011360 | oligopeptide ABC transporter, substrate-binding protein | Cell Wall / Extracellular |
| SFBMM\_011370 | oligopeptide ABC transporter, ATP-binding protein | Membrane |
| SFBMM\_011380 | oligopeptide ABC transporter, ATP-binding protein | Membrane |
| SFBMM\_011390 | oligopeptide ABC transporter, permease protein | Membrane |
| SFBMM\_011400 | oligopeptide ABC transporter, permease protein | Membrane |
| SFBMM\_013190 | putative sialic acid:sodium symporter | Membrane |
|  | | |
| **Glycoside hydrolases** | | |
| SFBMM\_003570 | GH94 family glycoside hydrolase | Membrane |
| SFBMM\_005720 | beta-glucosidase | Cytoplasmic |
| SFBMM\_005760 | beta-glucosidase | Cytoplasmic |
| SFBMM\_007830 | alpha-glucosidase | Cytoplasmic |
| SFBMM\_007900 | alpha-mannosidase | Cytoplasmic |
| SFBMM\_007960 | endo-beta-N-acetylglucosaminidase | Cell Wall / Extracellular |
| SFBMM\_008620 | beta-glucosidase | Cytoplasmic |
| SFBMM\_008630 | beta-glucosidase | Cytoplasmic |
| SFBMM\_008730 | beta-glucosidase | Cytoplasmic |
| SFBMM\_010210 | beta-N-acetylhexosaminidase | Cell Wall / Extracellular |
| SFBMM\_014470 | alpha-N-acetylglucosaminidase | Cell Wall / Extracellular |
|  | | |
| **Peptidases** | | |
| SFBMM\_000610 | M23/M37 family peptidase | Extracellular |
| SFBMM\_002400 | oligoendopeptidase F | Cytoplasmic |
| SFBMM\_002420 | oligoendopeptidase F | Cytoplasmic |
| SFBMM\_002960 | M23/M37 family peptidase | Extracellular |
| SFBMM\_005480 | M28 family peptidase | Extracellular |
| SFBMM\_006890 | putative trypsin-like protease | Extracellular |
| SFBMM\_007470 | U32 family peptidase | Cytoplasmic |
| SFBMM\_007680 | M28 family peptidase | Cell Wall / Extracellular |
| SFBMM\_009670 | putative trypsin-like protease | Cell Wall / Extracellular |
| SFBMM\_010600 | U32 family peptidase | Cytoplasmic |
| SFBMM\_011130 | M23/M37 family peptidase | Extracellular |
| SFBMM\_011490 | M28 family peptidase | Extracellular |
| SFBMM\_011830 | N(4)-(beta-N-acetylglucosaminyl)-L-asparaginase | Cytoplasmic |
| SFBMM\_013050 | M23/M37 family peptidase | Extracellular |
| SFBMM\_013260 | O-sialoglycoprotein endopeptidase | Cytoplasmic |
|  | | |
| **N-acetylglucosamine, N-acetylneuraminate, and mannose utilization** | | |
| SFBMM\_001070 | UDP-N-acetylglucosamine diphosphorylase | Cytoplasmic |
| SFBMM\_002180 | phosphoglucosamine mutase | Cytoplasmic |
| SFBMM\_003030 | N-acetylglucosamine-6-phosphate deacetylase | Cytoplasmic |
| SFBMM\_005330 | glutamine-fructose-6-phosphate transaminase | Cytoplasmic |
| SFBMM\_007920 | N-acetylmannosamine kinase | Cytoplasmic |
| SFBMM\_013180 | N-acetylmannosamine kinase | Cytoplasmic |
| SFBMM\_013210 | N-acetylneuraminate lyase | Cytoplasmic |
| SFBMM\_013220 | N-acylglucosamine-6-phosphate 2-epimerase | Cytoplasmic |
| SFBMM\_013330 | N-acylglucosamine-6-phosphate 2-epimerase | Cytoplasmic |
| SFBMM\_014810 | mannose-6-phosphate isomerase | Cytoplasmic |

## Table: SFB “virulence” factors

Table 13. SFB proteins putatively involved in epithelial cell attachment and host immune system avoidance/modulation

| **Locus tag** | **Predicted function** | **Predicted localization** |
| --- | --- | --- |
| **Polysaccharide deacetylases** | | |
| SFBMM\_003190 | putative polysaccharide deacetylase | Extracellular |
| SFBMM\_013320 | putative polysaccharide deacetylase | Extracellular |
| SFBMM\_006970 | putative polysaccharide deacetylase | Extracellular |
| SFBMM\_005410 | putative polysaccharide deacetylase | Extracellular |
| SFBMM\_005100 | putative polysaccharide deacetylase | Membrane |
| SFBMM\_006570 | putative polysaccharide deacetylase | Extracellular |
|  | | |
| **Exocellular polysaccharide biosynthesis** | | |
| SFBMM\_002920 | UDP-GluNAc:undecaprenyl-P GluNAc-1-P transferase | Membrane |
| SFBMM\_002930 | UDP-GluNAc 2-epimerase | Cytoplasmic |
| SFBMM\_006530 | N-acetylmannosaminyltransferase | Cytoplasmic |
| SFBMM\_006550 | GT1 family glycosyltransferase | Unknown |
| SFBMM\_006560 | O-antigen polymerase family protein | Membrane |
| SFBMM\_014580 | UTP-glucose-1-phosphate uridylyltransferase | Cytoplasmic |
| SFBMM\_014590 | GT2 family glycosyltransferase | Cytoplasmic |
| SFBMM\_014600 | mannose-1-phosphate guanylyltransferase | Unknown |
| SFBMM\_014610 | undecaprenyl-phosphate galactose phosphotransferase | Membrane |
| SFBMM\_014620 | polysaccharide biosynthesis protein | Membrane |
| SFBMM\_014630 | GT1 family glycosyltransferase | Cytoplasmic |
| SFBMM\_014650 | cell envelope-related transcriptional attenuator | Membrane |
| SFBMM\_014680 | GT1 family glycosyltransferase | Cytoplasmic |
| SFBMM\_014690 | GT1 family glycosyltransferase | Cytoplasmic |
| SFBMM\_014700 | acyltransferase domain-containing protein | Membrane |
| SFBMM\_014710 | cell envelope-related transcriptional attenuator | Membrane |
|  | | |
| **Adhesive structures** | | |
| SFBMM\_005870 | fibronectin/fibrinogen-binding protein | Unknown |
| SFBMM\_005890 | type IV pilus assembly protein | Membrane |
| SFBMM\_005900 | putative prepilin peptidase | Unknown |
|  |  |  |
| **Flagella** | | |
| SFBMM\_002500 | flagellar M-ring protein | Membrane |
| SFBMM\_002510 | flagellar hook-basal body complex protein | Unknown |
| SFBMM\_004530 | flagellar basal body-associated protein | Unknown |
| SFBMM\_004540 | flagellar biosynthesis protein | Membrane |
| SFBMM\_004550 | flagellar biosynthesis protein | Membrane |
| SFBMM\_004560 | flagellar biosynthesis protein | Membrane |
| SFBMM\_004570 | flagellar biosynthesis protein | Membrane |
| SFBMM\_004580 | flagellar biosynthesis protein | Membrane |
| SFBMM\_004590 | flagellar biosynthesis protein | Membrane |
| SFBMM\_004630 | flagellar basal-body rod protein | Extracellular |
| SFBMM\_004640 | flagellar basal-body rod protein | Extracellular |
| SFBMM\_006140 | flagellin | Cytoplasmic |
| SFBMM\_006150 | flagellar basal-body rod protein | Cytoplasmic |
| SFBMM\_006160 | flagellar basal-body rod protein | Unknown |
| SFBMM\_006170 | flagellar motor switch protein | Cytoplasmic |
| SFBMM\_006180 | flagellar assembly protein | Cytoplasmic |
| SFBMM\_006190 | flagellum-specific ATP synthase | Cytoplasmic |
| SFBMM\_006200 | flagellar export protein | Cytoplasmic |
| SFBMM\_006210 | fliK domain-containing protein | Cytoplasmic |
| SFBMM\_006220 | flagellar basal-body rod modification protein | Cytoplasmic |
| SFBMM\_006230 | putative flagellar protein | Cytoplasmic |
| SFBMM\_009380 | flagellin | Cytoplasmic |
| SFBMM\_009960 | flagellin | Cytoplasmic |
| SFBMM\_009970 | flagellar hook protein | Extracellular |
| SFBMM\_009990 | flagellar hook-associated protein 2 | Extracellular |
| SFBMM\_010000 | flagellar biosynthesis protein | Cytoplasmic |
| SFBMM\_010020 | flagellar assembly protein | Cytoplasmic |
| SFBMM\_010030 | flagellar hook-associated protein 3 | Cytoplasmic |
| SFBMM\_010040 | flagellar hook-associated protein 1 | Unknown |
| SFBMM\_010050 | flgN domain-containing protein | Unknown |
| SFBMM\_010060 | flagellar motor switch protein | Membrane |
| SFBMM\_010070 | flagellar motor switch protein | Membrane |
|  | | |
| **Hemolysins** | | |
| SFBMM\_000070 | putative hemolysin | Membrane |
| SFBMM\_009240 | putative hemolysin A | Cytoplasmic |
|  | | |
| **Phospholipases** | | |
| SFBMM\_008100 | phosphatidylinositol-specific phospholipase C | Extracellular |
| SFBMM\_008540 | cardiolipin synthase | Membrane |
| SFBMM\_012850 | phospholipase C | Unknown |
| SFBMM\_015700 | cardiolipin synthase | Membrane |

## Table: SFB metabolic enzymes

Table 14. Enzymes in the metabolic pathways of SFB.

| **Predicted function** | **Locus tag(s)** |
| --- | --- |
| **EMP glycolysis** | |
| glucokinase | SFBMM\_007950 |
| glucose-6-phosphate isomerase | SFBMM\_004850 |
| 6-phosphofructokinase | SFBMM\_003410, SFBMM\_014290 |
| 6-phospho-beta-glucosidase | SFBMM\_005720, SFBMM\_005760, SFBMM\_008620, SFBMM\_008630, SFBMM\_008730 |
| fructose-1,6-bisphosphate aldolase | SFBMM\_002540 |
| triose-phosphate isomerase | SFBMM\_015570 |
| glyceraldehyde 3-phosphate dehydrogenase | SFBMM\_015590 |
| phosphoglycerate kinase | SFBMM\_015580 |
| phosphoglycerate mutase | SFBMM\_015560 |
| enolase | SFBMM\_015550 |
| pyruvate kinase | SFBMM\_014280 |
| pyruvate phosphate dikinase | SFBMM\_010490 |
|  | |
| **Pentose phosphate pathway (oxidative phase; incomplete)** | |
| glucose-6-phosphate 1-dehydrogenase | SFBMM\_011910 |
| 6-phosphogluconate dehydrogenase | SFBMM\_011920 |
|  | |
| **Pentose phosphate pathway (non-oxidative phase; complete)** | |
| ribulose-phosphate 3-epimerase | SFBMM\_012710 |
| ribose 5-phosphate isomerase B | SFBMM\_009870 |
| transketolase | SFBMM\_000900, SFBMM\_000910 |
| ribose-phosphate pyrophosphokinase | SFBMM\_001080 |
|  | |
| **TCA cycle (incomplete)** | |
| 2-oxoglutarate ferredoxin oxidoreductase subunits alpha/beta/gamma/delta | SFBMM\_002140, SFBMM\_002150, SFBMM\_002160, SFBMM\_002190 |
| pyruvate carboxylase | SFBMM\_003880 |
|  |  |
| **Lysine biosynthesis (via acetylation)** | |
| aspartate kinase | SFBMM\_008300, SFBMM\_013570 |
| aspartate-semialdehyde dehydrogenase | SFBMM\_010940 |
| dihydrodipicolinate synthase | SFBMM\_010950 |
| dihydrodipicolinate reductase | SFBMM\_010960 |
| 2,3,4,5-tetrahydropyridine-2-carboxylate N-acetyltransferase | SFBMM\_005120 |
| N-acetyl-L,L-diaminopimelate aminotransferase | SFBMM\_005140 |
| N-acetyl-L,L-diaminopimelate deacetylase | SFBMM\_012900 |
| diaminopimelate epimerase | SFBMM\_013560 |
| diaminopimelate decarboxylase | SFBMM\_015310 |
| UDP-MurNAc-Ala-D-Glu-2,6-diaminopimelate ligase | SFBMM\_006780 |
| UDP-MurNAc-Ala-D-Glu-2,6-diaminopimelate-D-Ala-D-Ala ligase | SFBMM\_006790 |
|  |  |
| **Aspartate/asparagine** | |
| asparagine synthase | SFBMM\_014310 |
| aspartate aminotransferase | SFBMM\_008030 |
|  |  |
| **Glutamate/glutamine** | |
| glutamine synthetase | SFBMM\_000830 |
| glutamate dehydrogenase | SFBMM\_003430 |
|  |  |
| **Peptidoglycan biosynthesis** | |
| UDP-GluNAc 1-carboxyvinyltransferase | SFBMM\_000650, SFBMM\_002940 |
| UDP-MurNAc dehydrogenase | SFBMM\_014360 |
| UDP-MurNAc--L-Ala ligase | SFBMM\_001040 |
| UDP-MurNAc-L-Ala--D-Glu ligase | SFBMM\_001360 |
| UDP-MurNAc-Ala-D-Glu--2,6-diaminopimelate ligase | SFBMM\_006780 |
| alanine racemase | SFBMM\_014510, SFBMM\_015330 |
| D-Ala--D-Ala ligase | SFBMM\_002620 |
| UDP-MurNAc-Ala-D-Glu-2,6-diaminopimelate--D-Ala-D-Ala ligase | SFBMM\_006790 |
| phospho-MurNAc-pentapeptide-transferase | SFBMM\_006800 |
| UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase | SFBMM\_010760 |
| penicillin-binding protein 3 | SFBMM\_007480 , SFBMM\_014380 |
| penicillin-binding protein 1A | SFBMM\_007040 |
| monofunctional glycosyltransferase | SFBMM\_010650 |
| penicillin-binding protein 2 | SFBMM\_011690 |
| serine-type D-Ala-D-Ala carboxypeptidase | SFBMM\_003200, SFBMM\_004420, SFBMM\_009940, SFBMM\_011560 |
|  |  |
| **Peptidoglycan degradation** | |
| N-acetylmuramoyl-L-alanine amidase | SFBMM\_000500, SFBMM\_002050, SFBMM\_004720, SFBMM\_007660, SFBMM\_013680, SFBMM\_015140, SFBMM\_015390 |
|  |  |
| **Fatty acid biosynthesis** | |
| 3-oxoacyl-(acyl-carrier-protein) synthase III | SFBMM\_007270 |
| acetyl-coA carboxylase | SFBMM\_007310, SFBMM\_007330, SFBMM\_007340, SFBMM\_007350 |
| malonyl-CoA-[acyl-carrier-protein] transacylase | SFBMM\_007280 |
| 3-oxoacyl-(acyl-carrier-protein) synthase II | SFBMM\_007300 |
| 3-oxoacyl-[acyl-carrier-protein] reductase | SFBMM\_007290 |
| (3R)-hydroxymyristoyl-[acyl-carrier-protein] dehydratase | SFBMM\_007320 |
| enoyl-[acyl-carrier-protein] reductase | SFBMM\_002690 |
| Acyl-CoA thioesterase | SFBMM\_004500 |

## Table: Predicted genomic islands

Table 15. Genomic islands as predicted by SIGI-HMM and IslandPath-DIMOB.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Start** | **End** | | **Size** | **GI (genomic island) Prediction Program** |
| 10497 | 19388 | | 8891 | Predicted by multiple methods |
| 536558 | 547022 | | 10464 | Predicted by multiple methods |
| 557200 | 561840 | | 4640 | Predicted by multiple methods |
| 610435 | 615664 | | 5229 | Predicted by multiple methods |
| 615684 | 622338 | | 6654 | Predicted by multiple methods |
| 1077892 | 1082961 | | 5069 | Predicted by multiple methods |
| 1113102 | 1117273 | | 4171 | Predicted by multiple methods |
| 1128181 | 1134207 | | 6026 | Predicted by multiple methods |
| 1233225 | 1239382 | | 6157 | Predicted by multiple methods |
| 1369934 | 1374070 | | 4136 | Predicted by multiple methods |
| 1374641 | 1380671 | | 6030 | Predicted by multiple methods |
| 1393336 | 1397386 | | 4050 | Predicted by multiple methods |
| 1433963 | 1440639 | | 6676 | Predicted by multiple methods |
| 1549806 | 1554375 | | 4569 | Predicted by multiple methods |
| 1560124 | 1565201 | | 5077 | Predicted by multiple methods |
| 536558 | 547022 | | 10464 | SIGI-HMM |
| 557200 | 561840 | | 4640 | SIGI-HMM |
| 615684 | 622338 | | 6654 | SIGI-HMM |
| 1077892 | 1082961 | | 5069 | SIGI-HMM |
| 1113102 | 1117273 | | 4171 | SIGI-HMM |
| 1128181 | 1134207 | | 6026 | SIGI-HMM |
| 1233225 | 1239382 | | 6157 | SIGI-HMM |
| 1369934 | 1374070 | | 4136 | SIGI-HMM |
| 1374641 | 1380671 | | 6030 | SIGI-HMM |
| 1393336 | 1397386 | | 4050 | SIGI-HMM |
| 1433963 | 1440639 | | 6676 | SIGI-HMM |
| 1549806 | 1554375 | | 4569 | SIGI-HMM |
| 1560124 | 1565201 | | 5077 | SIGI-HMM |
| 10497 | 19388 | | 8891 | IslandPath-DIMOB |
| 610435 | 615664 | | 5229 | IslandPath-DIMOB |
| **Key:** | red | predicted by at least one method | | |
|  | blue | IslandPath-DIMOB | | |
|  | orange | SIGI-HMM | |  |

## Table: Paralogous gene families

Table 16. Gene families clustered at 60% identity, 70% coverage.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Cluster Seed** | **Cluster Hits** | **Annotation** |
| 1 | SFBMM\_012140 | SFBMM\_013970 | SPP1 family phage portal protein |
| 2 | SFBMM\_012150 | SFBMM\_014000 | PBSX family phage terminase large subunit |
| 3 | SFBMM\_001540 | SFBMM\_001670 | translation elongation factor Tu |
| 4 | SFBMM\_009380 | SFBMM\_009960 | flagellin |
| 5 | SFBMM\_010290 | SFBMM\_010310 | hypothetical protein |
| 6 | SFBMM\_007660 | SFBMM\_013680 | N-acetylmuramoyl-L-alanine amidase |
| 7 | SFBMM\_012100 | SFBMM\_013920 | phage minor structural protein GP20 |
| 8 | SFBMM\_004960 | SFBMM\_004970 | hypothetical protein |
| 9 | SFBMM\_012180 | SFBMM\_014010 | hypothetical protein |
| 10 | SFBMM\_012190 | SFBMM\_014020 | phage terminase small subunit |
| 11 | SFBMM\_012110 | SFBMM\_013940 | hypothetical protein |
| 12 | SFBMM\_008700 | SFBMM\_009000 | DUF898 domain-containing protein |
| 13 | SFBMM\_008210 | SFBMM\_009490 | stage 5 sporulation protein S |
| 14 | SFBMM\_012120 | SFBMM\_013950 | hypothetical protein |

## Table: Putatively extracellular and cell wall localized proteins

Table 17. PSort extracellular and cell wall localized proteins.

| **Locus tag** | **Predicted localization** | **PSort score** | **Signal peptide detected** |
| --- | --- | --- | --- |
| SFBMM\_002050 | Cellwall\* | 8.19 | No |
| SFBMM\_002450 | Cellwall | 9.2 | No |
| SFBMM\_002950 | Cellwall\* | 8.28 | Yes |
| SFBMM\_004930 | Cellwall | 8.97 | No |
| SFBMM\_004950 | Cellwall | 9.21 | Yes |
| SFBMM\_005180 | Cellwall | 9.17 | Yes |
| SFBMM\_005190 | Cellwall | 9.2 | No |
| SFBMM\_005200 | Cellwall | 9.94 | No |
| SFBMM\_005290 | Cellwall | 9.94 | Yes |
| SFBMM\_005340 | Cellwall | 8.97 | No |
| SFBMM\_005430 | Cellwall | 9.21 | Yes |
| SFBMM\_005440 | Cellwall | 9.21 | Yes |
| SFBMM\_005470 | Cellwall | 8.97 | No |
| SFBMM\_009330 | Cellwall | 8.97 | No |
| SFBMM\_011340 | Cellwall | 9.07 | Yes |
| SFBMM\_011360 | Cellwall | 9.21 | Yes |
| SFBMM\_012560 | Cellwall | 9.26 | Yes |
| SFBMM\_014950 | Cellwall | 9.21 | Yes |
| SFBMM\_015140 | Cellwall\* | 8.19 | No |
| SFBMM\_015390 | Cellwall\* | 8.19 | No |
| SFBMM\_000500 | Extracellular | 9.72 | No |
| SFBMM\_000610 | Extracellular | 9.6 | No |
| SFBMM\_000630 | Extracellular | 8.91 | No |
| SFBMM\_003160 | Extracellular | 8.91 | No |
| SFBMM\_003190 | Extracellular | 9.6 | No |
| SFBMM\_003460 | Extracellular | 9.13 | Yes |
| SFBMM\_003580 | Extracellular | 8.91 | No |
| SFBMM\_004630 | Extracellular | 9.72 | No |
| SFBMM\_004640 | Extracellular | 9.72 | No |
| SFBMM\_004960 | Extracellular | 9.13 | Yes |
| SFBMM\_004970 | Extracellular | 8.91 | No |
| SFBMM\_005150 | Extracellular | 8.91 | No |
| SFBMM\_005170 | Extracellular | 8.91 | No |
| SFBMM\_005410 | Extracellular | 9.6 | No |
| SFBMM\_005480 | Extracellular | 9.6 | No |
| SFBMM\_007070 | Extracellular | 8.91 | No |
| SFBMM\_007160 | Extracellular | 8.91 | No |
| SFBMM\_007660 | Extracellular | 9.6 | No |
| SFBMM\_008100 | Extracellular | 9.72 | No |
| SFBMM\_009160 | Extracellular | 8.91 | No |
| SFBMM\_009860 | Extracellular | 8.91 | No |
| SFBMM\_009970 | Extracellular | 9.6 | No |
| SFBMM\_009990 | Extracellular | 9.72 | No |
| SFBMM\_011130 | Extracellular | 9.6 | No |
| SFBMM\_011490 | Extracellular | 9.72 | No |
| SFBMM\_011740 | Extracellular | 9.73 | Yes |
| SFBMM\_013050 | Extracellular | 9.73 | Yes |
| SFBMM\_013460 | Extracellular | 8.91 | No |
| SFBMM\_013680 | Extracellular | 9.72 | No |
| SFBMM\_013690 | Extracellular | 8.91 | No |
| SFBMM\_015080 | Extracellular | 8.91 | No |
| SFBMM\_015090 | Extracellular | 10 | No |
| SFBMM\_015360 | Extracellular | 8.91 | No |
| SFBMM\_015470 | Extracellular | 8.91 | No |
| \* This protein may have multiple localization sites. | | | |

## Table: SFB cofactor synthesis enzymes

Table 18. SFB have incomplete pathways for synthesizing major cofactors, except PLP.

| **Predicted function** | **Locus tag(s)** |
| --- | --- |
| **Biotin (complete from dethiobiotin to biotin)** |  |
| biotin synthetase | SFBMM\_006590 |
| biotin-[acetyl-CoA-carboxylase] ligase | SFBMM\_003510 |
|  |  |
| **Riboflavin (incomplete)** | |
| riboflavin kinase / FMN adenylyltransferase | SFBMM\_008340 |
|  |  |
| **Folate (incomplete)** | |
| dihydrofolate reductase | SFBMM\_011320 |
|  | |
| **Pyridoxal Phosphate (complete)** | |
| pyridoxine biosynthesis protein | SFBMM\_007810 |
| glutamine amidotransferase | SFBMM\_007820 |
|  |  |
| **Coenzyme-A (complete from pantetheine to CoA)** | |
| putative pantothenate kinase | SFBMM\_011800 |
| pantetheine-P adenyltransferase | SFBMM\_009620 |
| dephospho-CoA kinase | SFBMM\_014840 |
| holo-[ACP] synthase | SFBMM\_014530 |
|  |  |
| **Thiamine (incomplete)** | |
| thiamine biosynthesis protein thiH | SFBMM\_006600 |
| thiamine pyrophosphokinase thiN | SFBMM\_009730 |
|  | |

## Table: Transporters

Table . Transporters

|  |  |  |
| --- | --- | --- |
| Locus tag | Predicted function | Reference |
| Other | | |
| SFBMM\_000150 | chromate ion efflux pump | [1] |
| SFBMM\_000160 | chromate ion efflux pump | [1] |
| SFBMM\_001380 | magnesium transporter | [2] |
| SFBMM\_002200 | p-aminobenzoyl-glutamate transporter | [3] |
| SFBMM\_002530 | proton-coupled thiamine transporter | [4] |
| SFBMM\_002580 | magnesium transporter | [5] |
| SFBMM\_003220 | sodium:proton antiporter | [6] |
| SFBMM\_004070 | MATE family efflux pump |  |
| SFBMM\_004160 | MFS family transporter |  |
| SFBMM\_004710 | MATE family efflux pump |  |
| SFBMM\_005400 | sodium/calcium exchanger protein |  |
| SFBMM\_005530 | sodium:dicarboxylate symporter |  |
| SFBMM\_006100 | sodium:dicarboxylate symporter |  |
| SFBMM\_007560 | P-type ATPase cadmium/zinc/cobalt transporter | [7] |
| SFBMM\_007990 | C4-dicarboxylate transporter/malic acid transport protein |  |
| SFBMM\_008020 | uracil:cation symporter | [8] |
| SFBMM\_009410 | uracil:cation symporter | [8] |
| SFBMM\_009700 | P-type ATPase calcium transporter | [9] |
| SFBMM\_009710 | putative formate/nitrite transporter |  |
| SFBMM\_010280 | amino acid antiporter |  |
| SFBMM\_010670 | CDF family efflux pump |  |
| SFBMM\_012250 | P-type ATPase calcium transporter | [9] |
| SFBMM\_012410 | MFS family transporter |  |
| SFBMM\_012580 | magnesium transporter | [2] |
| SFBMM\_013190 | putative sialic acid transporter |  |
| SFBMM\_013390 | RND family efflux pump |  |
| SFBMM\_013590 | CDF family efflux pump |  |
| SFBMM\_015100 | P-type ATPase copper transporter | [10] |
| SFBMM\_015380 | MATE family efflux pump |  |
| SFBMM\_015460 | amino acid antiporter |  |
|  |  |  |
| ABC transporters | | |
| SFBMM\_007880 | multiple sugar ABC transporter, permease protein |  |
| SFBMM\_007890 | multiple sugar ABC transporter, permease protein |  |
| SFBMM\_007910 | multiple sugar ABC transporter, substrate-binding protein |  |
| SFBMM\_007940 | multiple sugar ABC transporter, ATP-binding protein |  |
| SFBMM\_004930 | oligopeptide ABC transporter, substrate-binding protein |  |
| SFBMM\_005470 | oligopeptide ABC transporter, substrate-binding protein |  |
| SFBMM\_011340 | oligopeptide ABC transporter, substrate-binding protein |  |
| SFBMM\_011360 | oligopeptide ABC transporter, substrate-binding protein |  |
| SFBMM\_011370 | oligopeptide ABC transporter, ATP-binding protein |  |
| SFBMM\_011380 | oligopeptide ABC transporter, ATP-binding protein |  |
| SFBMM\_011390 | oligopeptide ABC transporter, permease protein |  |
| SFBMM\_011400 | oligopeptide ABC transporter, permease protein |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| PTS transporter components | | |
| SFBMM\_005570 | PTS system, cellobiose-specific IIC component |  |
| SFBMM\_005730 | PTS system, cellobiose-specific IIB component |  |
| SFBMM\_005740 | PTS system, cellobiose-specific IIC component |  |
| SFBMM\_006070 | PTS system, cellobiose-specific IIA component |  |
| SFBMM\_008640 | PTS system, beta-glucoside-specific IIABC component |  |
| SFBMM\_008740 | PTS system, beta-glucoside-specific IIABC component |  |
| SFBMM\_012520 | PTS system, mannose-specific IID component |  |
| SFBMM\_012530 | PTS system, mannose-specific IIC component |  |
| SFBMM\_012540 | PTS system, mannose-specific IIAB component |  |
| SFBMM\_005660 | PTS system, ascorbate-specific IIA component |  |
| SFBMM\_005670 | PTS system, ascorbate-specific IIC component |  |
| SFBMM\_005680 | PTS system, ascorbate-specific IIB component |  |
| SFBMM\_008560 | PTS system, ascorbate-specific IIB component |  |
| SFBMM\_008570 | PTS system, ascorbate-specific IIC component |  |
| SFBMM\_008580 | PTS system, ascorbate-specific IIA component |  |
| SFBMM\_009300 | PTS system, fructose-specific IIABC component |  |
| SFBMM\_015420 | PTS system, mannitol-specific IIABC component |  |
| SFBMM\_002070 | phosphoenolpyruvate-protein phosphotransferase |  |
| SFBMM\_008200 | phosphocarrier protein |  |
| SFBMM\_012590 | HPr kinase/phosphatase |  |

[1] <http://www.ncbi.nlm.nih.gov/pubmed/?term=2152903>

[2] <http://www.ncbi.nlm.nih.gov/pubmed/?term=15231793>

[3] <http://www.ncbi.nlm.nih.gov/pubmed/?term=9829935>

[4] <http://www.ncbi.nlm.nih.gov/pubmed/?term=16291685>

[5] <http://www.ncbi.nlm.nih.gov/pubmed/?term=17700703>

[6] <http://www.ncbi.nlm.nih.gov/pubmed/?term=1325937>

[7] <http://www.ncbi.nlm.nih.gov/pubmed/?term=12779235>

[8] <http://www.ncbi.nlm.nih.gov/pubmed/?term=7721693>

[9] <http://www.ncbi.nlm.nih.gov/pubmed/?term=2526682>

[10] <http://www.ncbi.nlm.nih.gov/pubmed/?term=16151212>

# REFERENCES

1. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355-360.
2. Uchiyama I, Higuchi T, Kawai M (2010) MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. Nucleic acids research 38: D361-365.
3. Tversky A (1977) Features of similarity. Psychological review 84: 327.
4. Bologa CG, Revankar CM, Young SM, Edwards BS, Arterburn JB, et al. (2006) Virtual and biomolecular screening converge on a selective agonist for GPR30. Nature chemical biology 2: 207-212.
5. Waites KB, Schelonka RL, Xiao L, Grigsby PL, Novy MJ. Congenital and opportunistic infections: Ureaplasma species and Mycoplasma hominis; 2009. Elsevier. pp. 190-199.
6. Yus E, Maier T, Michalodimitrakis K, Van Noort V, Yamada T, et al. (2009) Impact of genome reduction on bacterial metabolism and its regulation. Science 326: 1263.
7. Dandekar T, Huynen M, Regula JT, Ueberle B, Zimmermann CU, et al. (2000) Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames. Nucleic Acids Research 28: 3278.
8. Toft C, Andersson SGE (2010) Evolutionary microbial genomics: insights into bacterial host adaptation. Nature Reviews Genetics 11: 465-475.
9. Kilian M, Poulsen K, Blomqvist T, Havarstein L, Bek-Thomsen M, et al. (2008) Evolution of Streptococcus pneumoniae and its close commensal relatives. PLoS ONE 3: e2683.
10. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35: W182-185.
11. Uchiyama I (2007) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. Nucleic acids research 35: D343-346.
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.
13. Altschul SF, Madden TL, Schäffer aa, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25: 3389-3402.
14. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, et al. (2007) Integration of biological networks and gene expression data using Cytoscape. Nature protocols 2: 2366-2382.
15. Real R, Vargas JM (1996) The probabilistic basis of Jaccard's index of similarity. Systematic biology 45: 380.
16. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26: 297-302.
17. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, et al. (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36: D480-484.
18. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, et al. (2011) Metagenomic Biomarker Discovery and Explanation. in review.
19. Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, Dudek MR, Cebrat S. Where does bacterial replication start? Rules for predicting the oriC region. Nucleic Acids Res. 2004; 32: 3781-3791.
20. Feng Gao and Chun-Ting Zhang (2008) Ori-Finder: a web-based system for finding *oriC*s in unannotated bacterial genomes. *BMC Bioinformatics*. 9: 79.
21. Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9: R151.
22. Ivanov II, Atarashi K, Manel N, Brodie EL, Shima T, Karaoz U, et al. Induction of intestinal Th17 cells by segmented filamentous bacteria. Cell. 2009; 139: 485–498.
23. Turroni F, Bottacini F, Foroni E, Mulder I, Kim JH, et al. (2010) Genome analysis of Bifidobacterium bifidum PRL2010 reveals metabolic pathways for host-derived glycan foraging. Proc Natl Acad Sci U S A 107: 19514–19519.
24. Fukuda S, Toh H, Hase K, Oshima K, et al. 2011. Bifidobacteria can protect from enteropathogenic infection through production of acetate. Nature 469: 543–7.
25. Snel J, Heinen PP, Blok HJ, Carman RJ, Duncan AJ, Allen PC, et al. Comparison of 16S rRNA sequences of segmented filamentous bacteria isolated from mice, rats and chickens and proposal of “Candidatus arthromitus” Int J Syst Bacteriol. 1995;45:780–782.
26. Umesaki Y, Okada Y, Matsumoto S, Imaoka A, Setoyama H. Segmented filamentous bacteria are indigenous intestinal bacteria that activate intraepithelial lymphocytes and induce MHC class-II molecules and fucosyl asialo Gm1 glycolipids on the small-intestinal epithelial-cells in the ex-germ-free mouse. Microbiol Immunol. 1995; 39: 555–562.
27. McNamara N, Khong A, McKemy D, et al. ATP transduces signals from ASGM1 a glycolipid that functions as a bacterial receptor. Proc Natl Acad Sci USA. 2001;98:9086–91.
28. Rhee SH, Im E, Riegler M, Kokkotou E, O'Brien M, Pothoulakis C. Pathophysiological role of Toll-like receptor 5 engagement by bacterial flagellin in colonic inflammation. Proc Natl Acad Sci USA. 2005; 102: 13610-13615.
29. Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37 (Database issue): D141-D145; doi: 10.1093/nar/gkn879.
30. Paredes CJ, Alsaker KV, Papoutsakis ET. (2005) A comparative genomic view of clostridial sporulation and physiology. Nat Rev Microbiol. 3(12):969-78.
31. Seedorf H, Fricke WF, Veith B, Brüggemann H, Liesegang H, Strittmatter A, Miethke M, Buckel W, Hinderberger J, Li F, Hagemeier C, Thauer RK, Gottschalk G. (2008) The genome of Clostridium kluyveri, a strict anaerobe with unique metabolic features. Proc Natl Acad Sci U S A. 105(6):2128-33.
32. Davis CP, Savage DC. Effect of penicillin on the succession, attachment and morphology of segmented, filamentous microbes in the murine small bowel. Infect Immun. 1976;13:180–188.
33. Takatsugu G, Atsushi Y, Hideki H, Minenosuke M, Kozo T, Kenshiro O, Hidehiro T, Kazuaki M, Satoru K, Masahira H, et al.: Complete genome sequence of Finegoldia magna, an anaerobic opportunistic pathogen. DNA Research 2008, 15:39-47.
34. Marrazzo JM, Thomas KK, Fiedler TL, Ringwood K, Fredricks DN (2008) Relationship of specific vaginal bacteria and bacterial vaginosis treatment failure in women who have sex with women. Ann Intern Med 149: 20–28.
35. Raschle T, Amrhein N, Fitzpatrick TB (2005) On the two components of pyridoxal 5-phosphate synthase from Bacillus subtilis. J Biol Chem 280: 32291–32300.
36. Hsiao,W.W. et al. (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. PLoS Genet., 1, e62.
37. Waack,S. et al. (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics, 7, 142.
38. Langille,M.G.I. et al. (2008) Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinformatics, 9, 329.
39. Varani AM, Siguier P, Gourbeyre E, Charneau V, Chandler M. (2011) ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. Genome Biol. 12 (3): R30
40. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. (2008) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics. 24 (6): 863-5.
41. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics, 8 (1): 209
42. Boekhorst J, de Been MW, Kleerebezem M, Siezen RJ. (2005) Genome-wide detection and analysis of cell wall-bound proteins with LPxTG-like sorting motifs. J Bacteriol. 187 (14): 4928-34.
43. Sillanpää J, Nallapareddy SR, Qin X, Singh KV, Muzny DM, Kovar CL, Nazareth LV, Gibbs RA, Ferraro MJ, Steckelberg JM, Weinstock GM, Murray BE. (2009) A collagen-binding adhesin, Acb, and ten other putative MSCRAMM and pilus family proteins of Streptococcus gallolyticus subsp. gallolyticus (Streptococcus bovis Group, biotype I). J Bacteriol. 191 (21): 6643-53.
44. Bruggemann H, Baumer S, Fricke WF, Wiezer A, Liesegang H, Decker I, Herzberg C, Martinez-Arias R, Merkl R, Henne A, Gottschalk G. (2003) The genome sequence of Clostridium tetani, the causative agent of tetanus disease. Proc Natl Acad Sci U S A. 100 (3): 1316-21.
45. Brüggemann H, Gottschalk G. (2008) Comparative genomics of clostridia: link between the ecological niche and cell surface properties. Ann N Y Acad Sci. 1125: 73-81.
46. Hennequin C, Janoir C, Barc MC, Collignon A, Karjalainen T. (2003) Identification and characterization of a fibronectin-binding protein from Clostridium difficile. Microbiology, 149: 2779-87.
47. Katayama S, Nozu N, Okuda M, Hirota S, Yamasaki T, Hitsumoto Y. (2009) Characterization of two putative fibronectin-binding proteins of Clostridium perfringens. Anaerobe, 15 (4): 155-9.
48. Barketi-Klai A, Hoys S, Lambert-Bordes S, Collignon A, Kansau I. (2011) Role of fibronectin binding protein A in Clostridium difficile intestinal colonization. J Med Microbiol. 2011 Feb 24. [Epub ahead of print]
49. Varga JJ, Nguyen V, O'Brien DK, Rodgers K, Walker RA, Melville SB. (2005) Type IV pili-dependent gliding motility in the Gram-positive pathogen Clostridium perfringens and other Clostridia. Mol Microbiol. 62(3):680-94.
50. Davis KM, Weiser JN. (2011) Modifications to the peptidoglycan backbone help bacteria to establish infection. Infect Immun. 2011 Feb;79(2):562-70.
51. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J; MetaHIT Consortium, Bork P, Ehrlich SD, Wang J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 464 (7285): 59-65.
52. Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26: 589-95.
53. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25: 2078-9.
54. Delcher AL, Bratke KA, Powers EC, Salzberg SL. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics, 23:673-9.
55. Besemer J. and Borodovsky M. (1999) Heuristic approach to deriving models for gene finding. Nucleic Acids Research, 27:3911-3920.
56. Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research, 25:955-964.
57. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research, 35:3100-8.
58. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. (2005) Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Research, 33:D121-4.
59. E. P. Nawrocki, D. L. Kolbe, S. R. Eddy. (2009) Infernal 1.0: Inference of RNA Alignments. Bioinformatics, 25:1335-1337.
60. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, et al. (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9: 75.
61. Markowitz VM, Ivanova NN, Chen IMA, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics 2009; 25:2271-2278. doi: 10.1093/bioinformatics/btp393.
62. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–222.
63. N.Y. Yu, J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S.C. Sahinalp, M. Ester, L.J. Foster, F.S.L. Brinkman (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics 26(13):1608-1615
64. A. S. Juncker, H. Willenbrock, G. von Heijne, H. Nielsen, S. Brunak and A. Krogh. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. 12(8):1652-62, 2003
65. A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology, 305(3):567-580, January 2001.
66. Jannick Dyrløv Bendtsen, Henrik Nielsen, Gunnar von Heijne and Søren Brunak. (2004) Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol., 340:783-795.
67. Weizhong Li & Adam Godzik. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22:1658-1659.
68. Felsenstein, J. (1989) PHYLIP -- Phylogeny Inference Package (Version 3.2). Cladistics, 5:164-166.
69. Letunic I, Bork P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics, 23:127-8.
70. Tannock GW, Miller JR, Savage DC. Host specificity of filamentous, segmented microorganisms adherent to the small bowel epithelium in mice and rats. Appl Environ Microbiol. 1984 Feb; 47 (2): 441-2.
71. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 2008 Oct 13; 9 (10): R151.