SUPPLEMENTAL MATERIAL

Manuscript: GENOME/2011/131482

Single cell sequencing provides clues about the host interactions of segmented filamentous bacteria (SFB)

Sünje J. Pamp[a], Eoghan D. Harrington[a], Stephen R. Quake[b,c,d], David A. Relman[a,e,f,1], Paul C. Blainey[d]

[a]Department of Microbiology and Immunology, [b]The Howard Hughes Medical Institute, Departments of [c]Applied Physics, [d]Bioengineering, and [e]Medicine, Stanford University, Stanford, CA 94305, USA; [f]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA 94304, USA

[1]To whom correspondence should be addressed.
relman@stanford.edu

Index

Fig. S8. C-terminal sequence conservation among SFB.Cluster.1 proteins.

Fig. S9. Polymorphisms in SFB-specific Cluster.1 and Cluster.3 proteins among different SFB isolates (SFB-mouse-SU, SFB-mouse-Yit, SFB-mouse-NYU, SFB-mouse-Japan, and SFB-rat-Yit).

Fig. S10. Relative abundance of KEGG metabolic pathways in SFB-co, SFB-mouse-SU, and other clostridia.

Fig. S11. Predicted three-dimensional structure of the SFB bacterial dynamin-like protein (BDLP).

Fig. S12. SFB ADP-ribosyltransferase (ADPRT) sequence types in SFB genomes.

Fig. S13. Read depth for sequences from individual SFB filaments, and random & observed SNP distribution.

Fig. S14. Examples of Inter-filament variability.

Fig. S15. Multiple genome sequence comparison of SFB-mouse-SU, SFB-mouse-Yit, SFB-mouse-NYU, SFB-mouse-Japan, SFB-rat-Yit.

3) SUPPLEMENTAL TABLES

Table S1: Additional genome information for the individual SFB assemblies (SFB-1 to SFB-5) and the co-assemblies (SFB-co, SFB-mouse-SU).

Table S2: Clusters of SFB-specific proteins in SFB-co and SFB-mouse-SU.

Table S3: Clusters of SFB-specific proteins in mouse and rat SFB genomes.

Table S4: Number of SNPs in protein-coding regions from pairwise comparisons.

4) SUPPLEMENTAL REFERENCES

1) SUPPLEMENTAL METHODS AND RESULTS

Microfluidic device and single SFB filament isolation.

Aliquots of fecal material from an SFB-monocolonized mouse, obtained from Dr. Yoshinori Umesaki, Yakult Central Institute for Microbiological Research, Tokyo, Japan (Umesaki et al. 1995), were resuspended in PBS, and larger debris removed through sequential washing steps. The 48-channel microfluidic devices (Figure S2) were produced by the Stanford Microfluidics Foundry. These devices were similar to those previously described (Blainey et al. 2011). The device was pre-treated for 10 minutes with pluronic F127 at 0.2% in 1x PBS before filling with 1x PBS containing 0.01% Tween-20 and 0.01% pluronic F127 to reduce cell adhesion. Bovine serum albumin (BSA) was added to the treated cells at a final concentration of 0.1 mg/mL. Individual cells were separated from the bulk sample based on morphologic features using a laser trap, passed through two valves in an "air lock" configuration, opening one valve at a time to allow the trapped cell, but not fluid to pass through (Fig. S2A). Each trapped cell was moved about 1 mm from the bulk sample to the reaction chamber using the laser trap. No bacterial morphotypes or 16S rRNA sequence types, other than those characteristic of SFB were found in the fecal material as examined by fluorescent *in situ* hybridization (FISH) and broad range 16S rDNA sequencing (see Methods below, Fig. S2, SI Movies 1-5).

On-chip single SFB filament amplification.

The Repli-G midi MDA reagents (Qiagen) were used to amplify DNA from individual cells in 60 nL volumes on the device. First, cells contained within 0.75 nL PBS with 0.02% Tween-20 were flushed into the first lysis chamber with 3.5 nL lysozyme (10 mg/ml) and incubated for one hour at room temperature. Second, cells were flushed into the second lysis chamber with 3.5 nL buffer DLB (supplemented with 0.1 M dithiothreitol) to complete the cell lysis and denature the genomic DNA. Then, ~50 nL of reaction mix (45 µL were prepared from 29 µL Repli-G reaction buffer, 10 µL 20 mM H2O with 0.6% Tween, 2 µL Repli-G enzyme, and 2 µL Repli-G stop solution) was added to each of the 48 reactions. The device was then transferred to a hot plate set to 32°C and incubated overnight. The reaction volume was recovered by fitting the recovery ports on the chip with plastic pipet tips (P10 size), and by flushing the products into the pipet tips with the TRIS solution pumped into the reagent

port at 8 psi. Reaction products were examined for the presence and identity of 16S rDNA sequences through bacterial broad range and SFB-specific PCR and sequencing of PCR products (see below).

DNA extraction, 16S rDNA amplification, and sequencing.

DNA from an aliquot of fecal material from an SFB-monocolonized mouse (Umesaki et al. 1995) was extracted using the QIAamp DNA Stool Mini Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's recommendations. The 16S rRNA gene was amplified using broad range bacterial-specific primers Bact8FM (5-AGAGTTTGATCMTGGCTCAG-3) and Bact1391R (5-GACGGGCGGTGTGTRCA-3) (Palmer et al. 2007). PCRs were performed in triplicate 25-cycle reactions with 5 min at 95°C, 25 cycles of 30 sec at 94°C, 30 sec at 55°C and 90 sec at 72°C, followed by 8 min at 72°C. PCR products were gel-purified (Qiagen), pooled, cloned with the TOPO TA cloning kit (Invitrogen, Carlsbad, CA, USA), and inserts from 95 plasmids from each reaction sequenced on both strands. To examine products subsequent to cell sorting and multiple displacement amplification (MDA) (see above) for the presence of SFB-specific 16S rDNA sequences, 16S rRNA genes were amplified as follows: For SFB-specific PCR, primers SFB747F (5-TAACTGACGCTGAGGCATGAG-3) and SFB1266R (5-TAAGTTTTGCTCACTATCRC-3) were designed based on the high-quality SFB 16S rDNA sequences of the SILVA SSU reference database (http://www.arb-silva.de/), and examined with RDP ProbeMatch (http://rdp.cme.msu.edu/probematch/search.jsp) and probeCheck (http://131.130.66.200/cgi-bin/probecheck/content.pl?id=home). For bacterial broad-range PCR with the individual MDA-amplified single filament DNA, primers Bact8FM and Bact1391R were used. PCRs were performed in 35 cycle reactions with 5 min at 95°C, 35 cycles of 30 sec at 94°C, 30 sec at 55°C and 90 sec at 72°C, followed by 8 min at 72°C. PCR products were gel-purified (Qiagen), and sequenced directly on both strands. Five samples, from which only SFB-specific 16S rDNA sequences were obtained from both SFB-specific as well as bacterial broad-range PCR, were selected for pyrosequencing.

Fluorescent *in situ* hybridization (FISH).

Bacterial cells were fixed according to a protocol designed for fixation of Gram-positive bacteria (Roller et al. 1994). Briefly, an aliquot of ~50 mg fecal material was washed twice in PBS, resuspended in 50% ethanol/PBS (1:1, vol/vol) and fixed for 2 h at 4°C. Fecal material was obtained from an SFB-monocolonized mouse (Umesaki et al. 1995), and as a control from a laboratory mouse (strain FVB/N)

with a complex gut microbiota that tested positive for SFB in a screen of mouse fecal samples using a PCR assay with SFB-specific primers. The samples were washed twice in PBS and hybridized with a Cy3-labeled SFB-specific probe SFB1266R (5-TAAGTTTTGCTCACTATCRC-3) (see above) and a Cy5-labeled EUB338 probe (5-GCTGCCTCCCGTAGGAGT-3) (Amann et al. 1990) in 35% formamide, similar to procedures previously described (Fuchs et al. 2007). Subsequently, the samples were washed in 48°C pre-heated washing buffer with a stringency adjusted to 0.08M NaCl, then resuspended in distilled H_2O, and distributed into the wells of silane-coated microscope slides (Tekdon, Florida, USA). After air-drying, the samples were mounted using a 4:1 mix of Citifluor (Citifluor Ltd, London, U.K) and Vecta Shield (Vector Laboratories, Inc., Burlingame, CA) (Fuchs et al. 2007). Image acquisition was performed with a Zeiss LSM 510 confocal laser scanning microscope (Carl Zeiss, Germany) equipped with a NeHe laser and detector and filter sets for simultaneous monitoring of Cy3 and Cy5 fluorescence. All wells were thoroughly inspected. In the sample from the SFB-monocolonized mouse all cells displayed Cy3 and Cy5 fluorescence; in contrast, the sample derived from the SFB-positive FVB/N mouse (housed at the Stanford Research Animal Facility) with a complex gut microbiota many bacterial cells displayed only Cy5 fluorescence. All bacteria with segmented filamentous morphology visible with transmitted light, in both samples, exhibited Cy3- and Cy5-positive signals. No bacteria of any morphology other than filamentous-segmented exhibited a Cy3-positive signal. Images were obtained using a 63x/1.4 NA Plan-Apochromat oil objective and analyzed using Imaris software package (Bitplane AG, Switzerland).

Creation of sequencing library.

Four microliters of DNA from each first-round reaction that was positive for SFB 16S rDNA and negative for other 16S rDNA sequence types, was re-amplified using the Repli-G midi kit (Qiagen). This template solution was denatured by the addition of 3.5 µl buffer DLB for 5 minutes at room temperature and neutralized by addition of 3.5 µl stop solution. A reaction mix consisting of 29 µl reaction buffer, 10 µl water, and 1 µl enzyme was prepared on ice, and then added to the denatured template. Reactions were incubated at 30°C for 12 hours and then diluted 10-fold in 10 mM TRIS with 0.02% Tween-20 for and storage at -60°C. A shotgun library was prepared from approximately 5 µg of the second round MDA product according to the Roche/454 protocol for "Titanium" shotgun libraries with the following modifications. Custom barcoded adaptor oligos (IDT, Coralville, Iowa) were used to enable the pooling

of multiple libraries in a single emulsion PCR reaction and picotiter plate region during sequencing. To obtain dsDNA sequencing libraries and shorten the library preparation process, the library immobilization, fill-in, and single-stranded library isolation steps were omitted.

Sequencing library DNA quantification, and shotgun sequencing.

Sequencing library DNA were quantified using digital PCR as previously reported (White et al. 2009), with the exceptions that 48.770 digital arrays (Fluidigm Corp, San Francisco, CA) were used for the microfluidic dPCR step, and that amplification primers complimentary to the Titanium adaptor sequences were used. Briefly, serial dilutions of the sequencing libraries were made in 10 mM TRIS buffer with 0.02% Tween-20. 48 sample preparations were then combined according to the Fluidigm dPCR protocol with a reaction buffer containing thermostable DNA polymerase, dNTPs, GE sample loading reagent (Fluidigm), and the primers and probe necessary to carry out the universal Taqman amplification/detection scheme. The samples were loaded in the array and run on the Biomark thermocycler for 45 cycles. Sample analysis was carried out using the default parameters for dPCR analysis using the Fluidigm analysis software. The quantified library was diluted to 2 x 10^6 molecules per microliter in 10 mM TRIS with 0.02% Tween-20 and aliquotted for storage at -60°C. DNA pyrosequencing of the shotgun library was carried out on the Roche 454 Genome Sequencer FLX instrument using "Titanium" chemistry. The SFB libraries were sequenced in three runs of the instrument. A total of 762,520 reads were obtained for the five SFB: 67,890,293 bases for SFB-1; 54,287,496 bases for SFB-2; 38,102,428 bases for SFB3; 44,753,181 bases for SFB-4; and 50,333,241 bases for SFB-5. The G+C contents of the individual genome read sets each formed a major peak centered near 28% GC. The dispersion of the read G+C content was typical of a single microbial genome sampled at the same average read length (Fig. S3A).

Assembly, Gene Prediction, and Annotation.

Reads from the shotgun pyrosequencing runs were binned by individual SFB filament and trimmed using the sfffile tool (Roche, 454 Life Sciences, Branford, CT) permitting one mismatch in each 10 bp barcode. Reads from each SFB filament were individually assembled, as well as co-assembled *de novo* using Genome Sequencer FLX System Software (Newbler) version 2.5.3 (Roche) at default parameters,

except for specifying an increased expected read depth in excess of the actual value (SFB-1 to SFB-5, and SFB-co). In addition, the reads from all SFB genomes were co-assembled by mapping the reads against the complete SFB genome sequence SFB-mouse-Yit (AP012209, (Prakash et al. 2011)) using Genome Sequencer FLX System Software (Newbler) version 2.5.3, resulting in SFB-mouse-SU. For each assembly chimeric reads, i.e. reads that mapped to more than one contig, were excluded. Identification and removal of contaminant reads were performed with the use of SmashCell (Harrington et al. 2010) based on tetranucleotide frequencies, GC content, and taxonomic affiliation from hits against NCBI GenBank that emerged as distinct clusters in principle coordinate analysis (PCA) plots, and self-organizing maps (SOM) (Harrington et al. 2010), http://asiago.stanford.edu/SmashCellReleases/dev/. Assembled contigs were inspected using the Hawkeye program from the AMOS package (Schatz et al. 2007) and Tablet viewer (Milne et al. 2010). Ribosomal RNA genes were predicted using Meta_RNA (Huang et al. 2009) and tRNA genes using tRNAscan-SE (Lowe and Eddy 1997) (Fig. S3B+C). These regions were then masked out before training the Prodigal algorithm and predicting protein-coding genes (Hyatt et al. 2010). Protein annotation was partly performed in SmashCell (Harrington et al. 2010) by comparing protein sequences against NCBI Complete Microbial Genomes (downloaded 2010-02-01), STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database version 8.2 (Jensen et al. 2009), KEGG (downloaded 2010-11-27) and Uniprot (release 2010-09). In addition, functional protein domains were assigned by comparing protein sequences against Pfam 25.0 (Finn et al. 2010), secretory proteins were predicted by SignalP (Bendtsen et al. 2004) and SecretomeP (Bendtsen et al. 2005), and transmembrane domains were predicted by TMHMM (Krogh et al. 2001). Contigs, genes and protein sequences of particular interest were also compared against the NCBI GenBank database, and InterPro.

Rarefaction analysis to determine approximate genome size.

Subsets of the co-assembled reads were re-assembled using Newbler version 2.5.3 in a rarefaction analysis to determine the approximate genome size. The total assembly size was predicted from an asymptote that formed just below 1.63 Mb, indicating that the SFB co-assembly represents an essentially complete genome at 1.625 Mb (Figure S3A).

Shared sequence analysis to estimate genome size.

Among closely-related single-cell genomes covered randomly by sequence reads, a prediction of the consensus genome size can be made based on the quantity of corresponding sequence found among pairs of single-cell assemblies of known size: e.g. if 50% were shared between two assemblies of 1.0 Mb each, the size would be estimated at 2.0 Mb. We used BLASTn comparisons of the 5 individual SFB filament genomes to determine the amount of shared sequence, and estimated the genome size at 1.612 +/- 0.011 Mb (SEM).

Phylogenetic tree construction of individual genes and proteins.

16S rDNA sequences were aligned using SINA (SILVA INcremental Aligner, http://www.arb-silva.de/aligner/), and protein sequences were aligned using MUSCLE (Edgar 2004). Both 16S rDNA sequences and protein sequences of the DNA polymerase III alpha subunit, translation elongation factor Tu, RecA protein, phenylalanyl-tRNA synthetase beta chain, and the preprotein translocase subunit SecA were identical among individual SFB, and one representative sequence was selected. Maximum likelihood (ML) phylogenetic inferences were derived using the PHYML (Guindon and Gascuel 2003) implementation in Geneious (Drummond et al. 2011), with the HKY85 model (Hasegawa et al. 1985) for nucleotide substitutions and the JTT matrix (Jones et al. 1992) for amino acid substitutions. Support for the resulting inferred relationships was assessed in 100 bootstrap replicates.

Comparative genomics and cluster analysis.

Comparative genomics analysis of gene and protein sequences derived from the SFB assemblies and those of other bacteria were performed with SmashCell (Harrington et al. 2010). Comparative analyses of genome and assembly sizes and number of genes were based on a local copy of NCBI's Complete Microbial Genomes from 2010-02-01. Comparative analyses of predicted proteomes were performed by comparing the proteins from SFB-co and selected members of the *Clostridiaceae* 1 against each other using BLASTp with the default parameters. The resulting hits were filtered (bitscore ≥ 60) and used to create an adjacency list representation with the proteins as nodes and the percent bitscore (bitscore/self-hit bitscore*100) as the edge weight. The network described by this adjacency list was analysed using the Markov cluster (MCL) algorithm to produce a list of protein clusters approximating

protein families (http://micans.org/mcl/) (Enright et al. 2002; Van Dongen 2008). A matrix was then constructed whereby each cell contained the number of proteins from a species belonging to a given cluster. This matrix was then analyzed using principal component analysis (PCA) to explore the differences between species. The patterns of variation observed were stable across a range of values for MCL's inflation parameter (1.2-5.0), and the results shown here (Figs. 2+S7) were produced with a value of 3.0.

Sequence analysis, including CRISPR loci, and protein structure prediction.

Proteins were aligned using MUSCLE (Edgar 2004) and sequence logos, as well as hydrophobicity and isoelectric point (PI) profiles generated with Geneious (Drummond et al. 2011). Protein structures were predicted using the Protein Homology/Analogy Recognition Engine V 2.0 PHYRE^2 (http://www.sbg.bio.ic.ac.uk/~phyre2) (Kelley and Sternberg 2009). Amino acid conservation was examined through PRofile ALIgNEment (PRALINE) (http://www.ibi.vu.nl/programs/pralinewww/) using default settings (Simossis and Heringa 2005). To identify potential clustered regularly interspaced short palindromic repeats (CRISPRs), assemblies were examined with CRISPRFinder (http://crispr.u-psud.fr/Server/) (Grissa et al. 2007). Three CRISPR arrays (CRISPR-1, CRISPR-2, CRISPR-3) were found. CRISPR-1 was identified in SFB-5, SFB-co and in fragmented form in SFB-4, and is composed of 10 direct repeats (DR) and 9 spacers, and located downstream of seven CRISPR-associated genes (Cas). It is situated in the neighborhood of genes for a ribose ABC transporter and PTS system compounds. CRISPR-2 and -3 were recovered from SFB-1, SFB-2, SFB4, SFB-co and in fragmented forms from SFB-5 and are composed of 8 DR (7 spacers) and 5 DR (4 spacers), respectively. The two arrays are separated from another by 1.7 kb. CRISPR-2 and -3 are located in the neighborhood of phage-related genes, a characteristic that has been observed previously in *Clostridium difficile* 630 in which CRISPRs have been identified in two prophages (Sebaihia et al. 2006). All spacers have a typical length of 34-36 bp. The sequence of the DRs of the CRISPR arrays are (nearly) identical and have similarity to NC_009253_3 of *Desulfotomaculum reducens* MI-1 (*Clostridia*), NC_015172_2 of *Syntrophobotulus glycolicus* DSM 8271 (*Clostridia*), and NC_002570_2 of *Bacillus halodurans* C-125 (*Bacilli*). The average GC content of the phage1 and phage2 elements is 31.7%. Four predicted restriction modification systems with proteins

similar to type I-, type II- and type III-restriction modification enzymes were identified in the SFB genomes. SFB also encode for proteins known to be involved in homologous recombination.

Whole genome sequence comparison

Whole and nearly-complete mouse SFB genomes (SFB-mouse-SU (this study), SFB-mouse-Yit (Prakash et al. 2011), SFB-mouse-NYU (Sczesnak et al. 2011), SFB-mouse-Japan (Kuwahara et al. 2011)) were analyzed using Mauve multiple genome alignment software (Darling et al. 2004). Whole and nearly complete mouse and rat SFB genomes (SFB-mouse-SU (this study), SFB-mouse-Yit (Prakash et al. 2011), SFB-mouse-NYU (Sczesnak et al. 2011), SFB-mouse-Japan (Kuwahara et al. 2011), SFB-rat-Yit (Prakash et al. 2011)) were compared using Artemis Comparison Tool (ACT) (Carver et al. 2008). To facilitate comparisons with SFB-mouse-NYU, the five contigs of SFB-mouse-NYU were reordered, the reverse-complement sequence was used where necessary, and the sequence was split at the origin of replication. To estimate phylogenetic relationships between the complete and nearly-complete SFB genomes, all five whole and nearly-complete genome sequences were aligned using Kalign2 (Lassmann et al. 2009). The alignment was manually inspected and regions missing from some genomes, such as rRNA operons and phage elements, were removed. The final alignment consisted of 1,455,482 sequence positions, and a maximum likelihood inference was derived using PHYML (Guindon and Gascuel 2003) with the HKY85 model (Hasegawa et al. 1985) for nucleotide substitutions. Support for the resulting, inferred relationships was assessed with 100 bootstrap replicates. Accession numbers for SFB genomes published elsewhere: AP012209 (SFB-mouse-Yit), AGAG01000000 (SFB-mouse-NYU), NC_015913 (SFB-mouse-Japan), AP012210 (SFB-rat-Yit) (Kuwahara et al. 2011; Prakash et al. 2011; Sczesnak et al. 2011).

Single nucleotide polymorphism (SNP) detection and error detection analysis.

To identify single nucleotide polymorphisms (SNPs) three approaches were applied. First, BLAST pairwise comparisons of nucleotide and amino acid sequences from all protein coding genes between the individual SFB assemblies were performed including self-alignments. Best reciprocal hit (BRH) pairs were identified and those that aligned over 98% of the length of both individual proteins (to limit the effect of paralogy, gene fragmentation and sequencing-derived frameshift errors) were selected for

subsequent analysis. Each pair of proteins was aligned using PyCogent's (Knight et al. 2007) Needleman-Wunsch implementation and the resulting alignment was used to create a codon alignment of the nucleotide sequences from which putative SNPs were identified (Table S3). Nucleotide polymorphisms were inspected, and those excluded from further investigation that occurred at homopolymeric regions. Genes with potential SNPs were examined further through nucleotide and protein alignments with MUSCLE (Edgar 2004) and PRALINE (Simossis and Heringa 2005). In addition, we searched for the genes reported here as carrying nucleotide polymorphisms in all assemblies with BLASTn, as some were not initially reported because they exhibited a target length below 98% due to their location at the end of contigs. Contigs that harbored genes with SNPs were inspected using the Hawkeye program from the AMOS package (Schatz et al. 2007) and Tablet viewer (Milne et al. 2010) to check for potential sequence variation among reads.

In a second approach, SNPs were identified as discrepancies between aligned single-filament assemblies and the co-assembly. We attempted to validate these by mapping reads (filtered using the MOTHUR package (Schloss et al. 2009) (http://www.mothur.org): no homopolymers greater than 10, no ambiguous bases, average quality score greater than 26) from each single cell against the co-assembly. 'High-confidence' variants were identified where a sequence variant was found in reads from both strands where the sequencing depth was at least five reads, and where the total variant fraction exceeded 70%. High-confidence variants were also called from regions with a depth of three or more reads if both strands were covered and the variant fraction was 100%. About 34% of the SNPs were supported by high-confidence variants. Among the 'unvalidated' SNPs, we were able to identify more than 70% as arising from probable homopolymeric errors, with the remainder split roughly evenly between substitutions and indels that arose from other sequencing errors, and were true variants insufficiently supported by the raw data to be deemed 'validated'. The data are summarized in Table 3.

In a third approach, SNPs were identified as discrepancies between aligned single-filament assemblies and the reference, SFB-mouse-Yit. We attempted to validate these by mapping reads (filtered using the MOTHUR package (Schloss et al. 2009) (http://www.mothur.org): no homopolymers greater than 10,

no ambiguous bases, average quality score greater than 26, trimmed where a sliding window of 50 bases drops in average quality score to less than 19) from each single filament against the reference, excluding chimeric reads (Fig. S13A). A number of measures were taken to preclude the possibility that mis-mapping resulted in artifactual variant calls. First, we selected SFB-mouse-Yit as the reference for two reasons: 1) it is the most closely related genome to that of our cells, allowing definitive alignments to be made; and 2) it has the highest quality assembly and the lowest chance of misassemblies or compression of repeats or paralogous genes. Next, we set the mapping criteria to be highly stringent, requiring 90–95% nucleotide identity over 40 base pairs. To test for miscalls resulting from mismapping of paralogous genes, we generated simulated reads from the reference with average depth, length distribution, and error density & distribution comparable to the filtered experimental reads, and mapped these back to the genome, finding no called variants. Finally, the loci with variants discussed in the text were compared via BLAST against the reference to ensure that no other regions exist in the reference with homology sufficient to create mapping ambiguities at the locus in question. In fact, paralogous gene pairs checked in this manner exhibited sufficient nucleotide sequence divergence to conclude that the mapping criteria applied here were sufficiently stringent to eliminate the possibility of mismapping.

Two classes of variants were identified, 'high-penetrance' variants, where essentially all the reads from a single filament exhibited a different sequence at a given locus, and 'partial-penetrance' variants, where a subset of the reads from a single filament revealed a coherent variant present in the dataset. These variants are inconsistent with sequencing error, exhibiting deep and consistent coverage of only the variant sequence and reference-matching sequence at such loci. In total, we identified 846 high-penetrance variants and 442 partial-penetrance variants in the single filaments. In some cases, all five single filaments shared the same variant, and in other cases, there were differences between the individual filaments. In most cases where inter-filament heterogeneity existed, a single variant appeared in a subset of the cells, and the remainder matched the reference.

A number of lines of reasoning and evidence support the assertion that the majority of these variants represent differences in the genomes of the SFB filaments, and are not MDA errors. First, the enzyme

we overproduced and purified for MDA, wild-type phi29 DNA polymerase (DNAP), is a proofreading enzyme with a per-insertion error rate approaching 10^{-6} (Esteban et al. 1993). Second, MDA allows multiple coverage of each strand of the original template DNA, allowing multiple independent sampling of the original genomic sequences. Third, the statistics of variant occurrence in the genome are inconsistent with random MDA errors, with variants both strongly clustered in some regions while larger regions remain variant-free. These findings are contrary to what would be expected for a random distribution of mutations (Fig. S13B).

Because we sequenced MDA products amplified independently from single filaments, we can provide a different class of evidence: correlations between the variants observed in independently handled SFB filaments. For example, our mapping exercise identified a total of 1287 variants in the five single-filament datasets across a total of 556 sites in the genome. At 288 of these 556 sites, the same variant was identified in two or more cells. This distribution of variants is extremely unlikely to occur by chance in a random mutational process. For example, when 1287 variants are randomly distributed over the five single filament genomes of 1,585,112 bp each, one would expect the same variant in different filaments at the same genomic locus to be extremely rare. Under this random model where variants are equally likely to arise at any position in any filament, the number of variants per genomic locus is binomially-distributed with $p = 0.000162$ for five trials. Specifically, one would expect to find 1,585,109 sites with no variants in any of the cells, 1287 sites with a variant observed in only one filament, and less than one site with variants observed in two or more filaments. The chance of observing two or more variants at a given site is 2.64×10^{-7}. Hence, the probability of observing 288 sites with variants in two or more cells is vanishingly small, estimated to be less than 10^{-109}. This argument is further supported by the fact that in nearly all the sites where variation was seen in more than one filament, the same change occurs in all the cells with variation.

Finally, we can use a different type of correlation to argue that the partially-penetrant variants we observe in the data from single filaments are not exclusively the result of MDA errors or other random sources of error, but represent intra-filament variability. If the partially-penetrant variants were the result of random MDA errors, we would not expect a relationship between the fractional penetrance of

variants and their genomic proximity. Conversely, if there is intra-filament heterogeneity and we sequence two different 'alleles' originating from two or more cells, we would expect a strong local correlation in the variant penetrance. This is because MDA amplification bias with a small number of molecules results in widely- and (mostly) randomly-varying sequencing depth along the genome. We observed this in the five SFB filaments we amplified (Fig. S13A). This variation in coverage depth is, however, strongly correlated along the genome; e.g., if a given position for a filament has a sequencing depth of 100 reads, it is very likely that the depth at the adjacent nucleotide is close to 100 reads. In MDA, this correlation extends beyond the sequencing read length to approximately 10 kb Fig. S13C), and is likely attributable to the finite processivity of phi29 DNAP in strand displacement synthesis under MDA conditions. A consequence of this correlation in sequencing depth is that if a small number of molecules representing two 'alleles' of a genomic locus are amplified by MDA, the ratio of reads representing variants in the products is expected to be a random value that is strongly correlated within the correlation length of MDA. We observed a striking correlation between genomic proximity and the similarity of variant penetrance values, and this correlation is lost beyond the MDA correlation distance. This is visible in Figure S13D where we plot the difference in penetrance fraction as a function of genomic separation for all pairs of partially-penetrant variants in each of the five filaments. A marked depletion in variant pairs with disparate degrees of penetrance is evident within the MDA correlation length for all five datasets. Thus, many of the partially-penetrant variants likely represent true heterogeneity between cells making up each of the SFB filaments we sequenced. The interpretation of these variants as intra-filament heterogeneity on this basis is reinforced by the observation that many of the same partially-penetrant variants are observed independently in two or more of the five filaments.

2) SI FIGURES

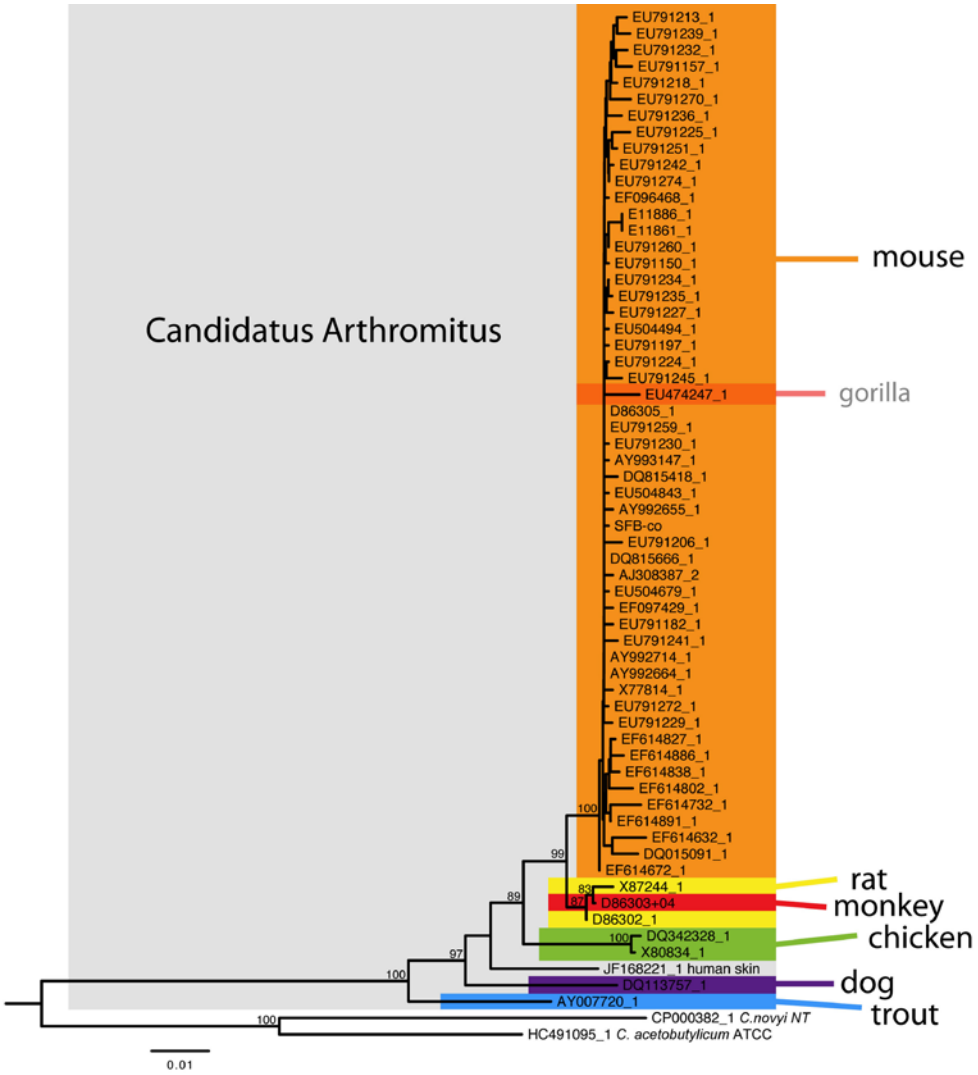
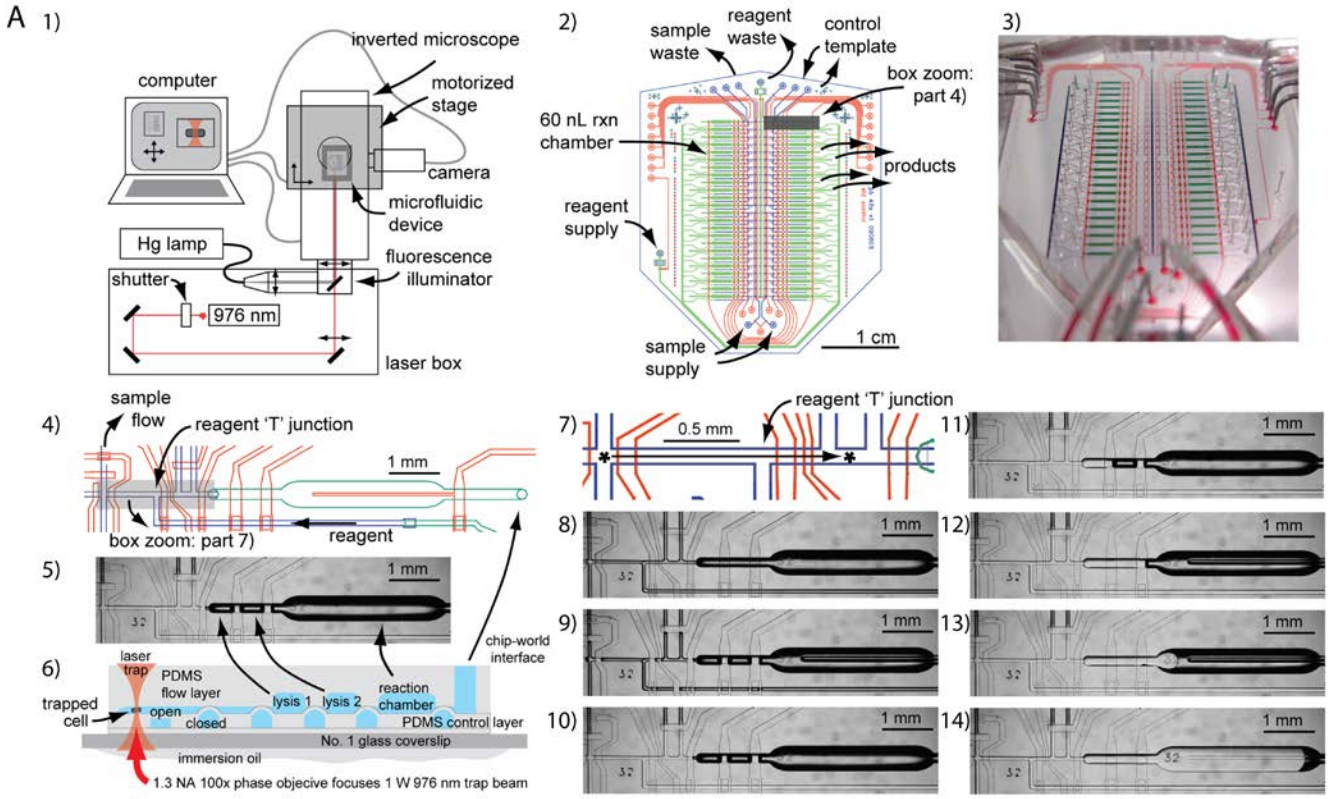
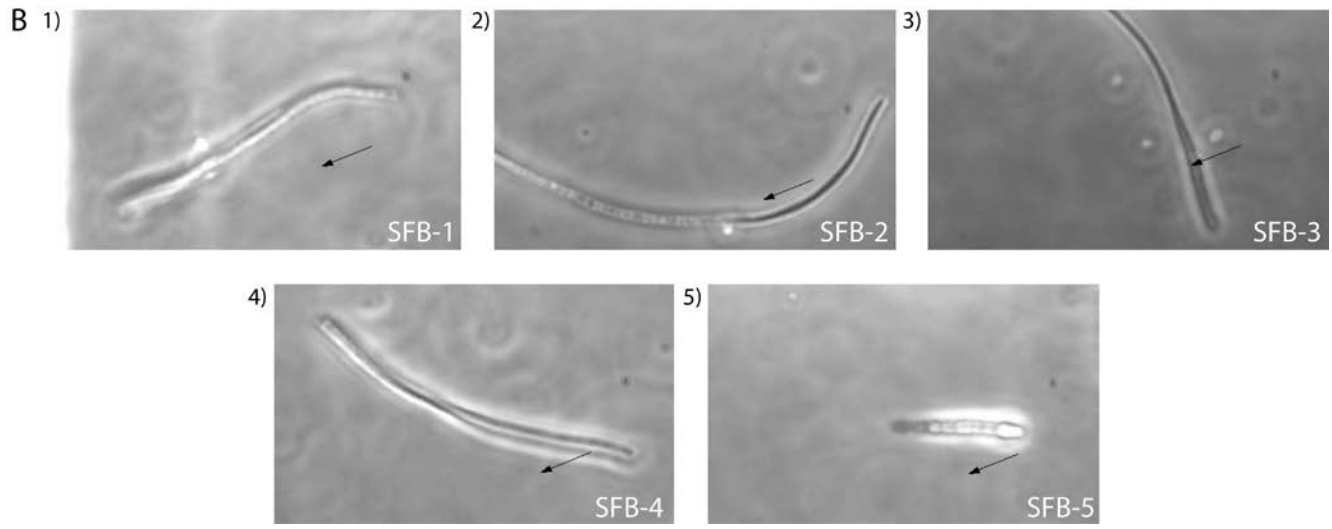


Fig. S1. Phylogenetic tree based on 16S rDNA sequences of members of the clade Candidatus *Arthromitus*, i.e. segmented filamentous bacteria (SFB) from a variety of different hosts, within the family *Clostridiaceae* 1. Sequences were aligned using SINA, and phylogeny inferred using the maximum likelihood method. Bootstrap statistical support values for branchings in the Candidatus *Arthromidus* clade ≥ 75 are shown.


A
1)
inverted microscope
computer
motorized stage
camera
microfluidic device
Hg lamp
fluorescence illuminator
shutter
976 nm
laser box
2)
sample waste
reagent waste
control template
box zoom: part 4)
60 nL rxn chamber
products
reagent supply
sample supply
1 cm
3)
4)
sample flow
reagent 'T' junction
1 mm
box zoom: part 7)
reagent
5)
1 mm
6)
laser trap
PDMS flow layer
chip-world interface
trapped cell
open
closed
lysis 1
lysis 2
reaction chamber
PDMS control layer
No. 1 glass coverslip
immersion oil
1.3 NA 100x phase objecive focuses 1 W 976 nm trap beam
7)
reagent 'T' junction
0.5 mm
8)
9)
10)
11)
12)
13)
14)
1 mm


B
1)
SFB-1
2)
SFB-2
3)
SFB-3
4)
SFB-4
5)
SFB-5

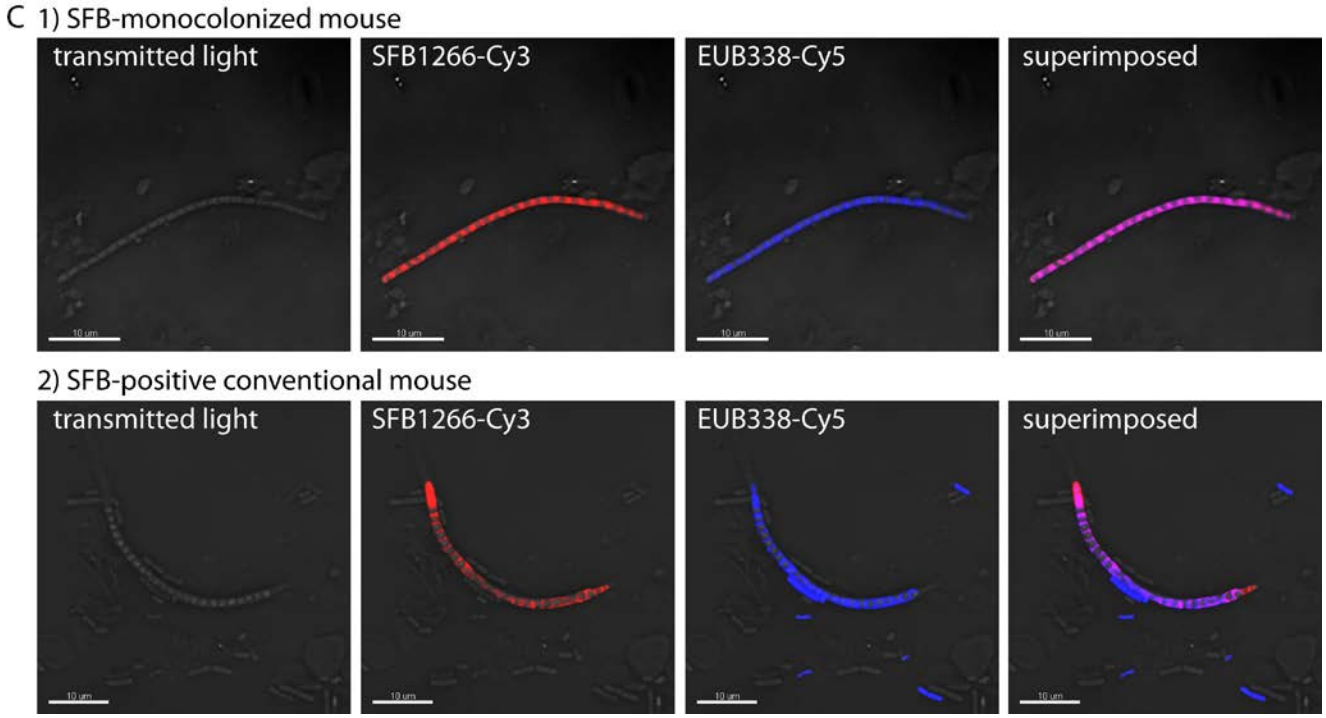
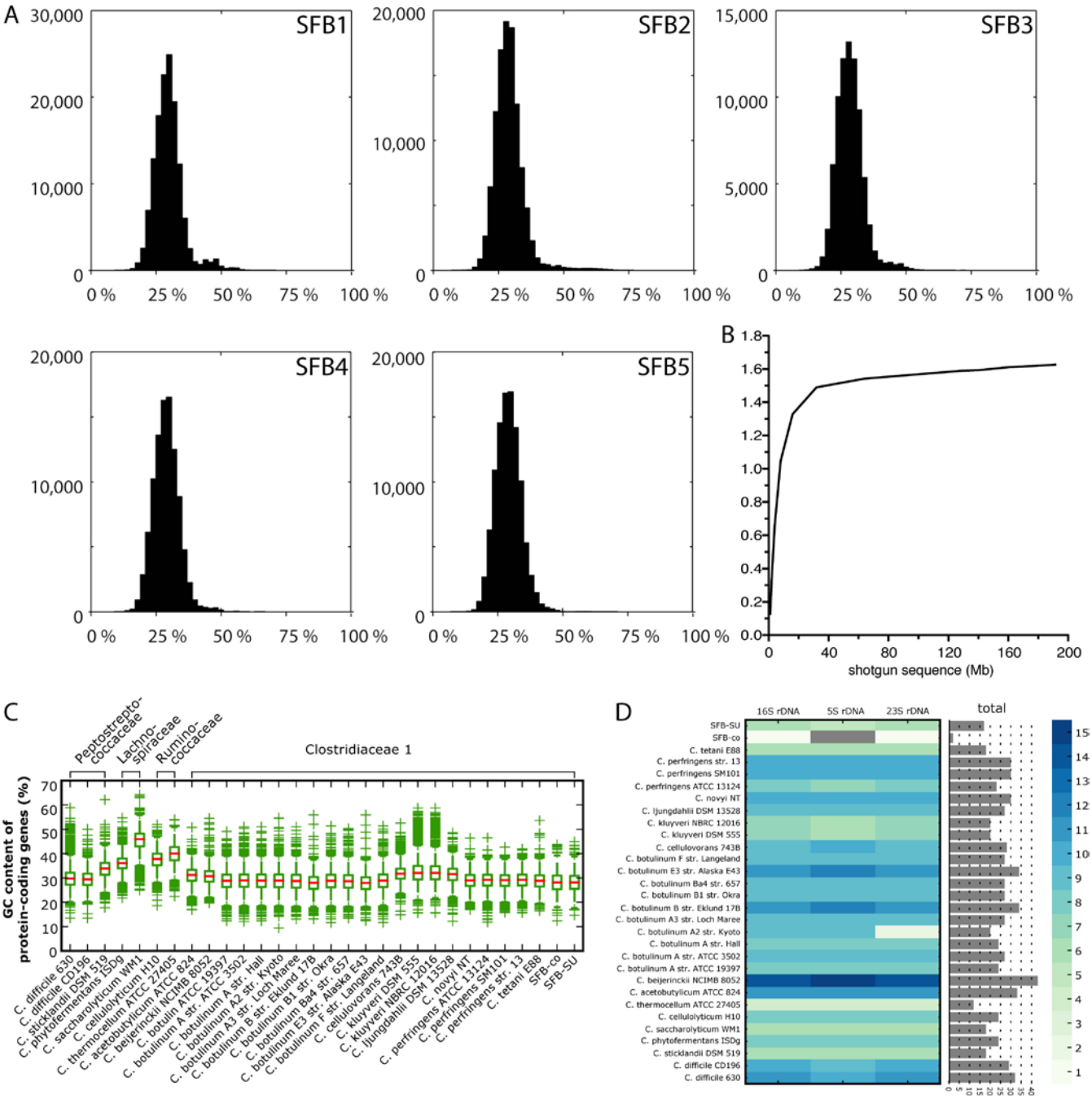


Fig. S2. Optofluidic device for the isolation of individual SFB filaments and amplification of their genomes (A), sorted SFB filaments (B), and identification of SFB in a murine fecal sample using fluorescent *in situ* hybridization (FISH) (C). (A) 1) Computer-controlled microscope fitted for fluorescence imaging and laser trapping. 2) Plan view of the two-layer 48-channel microfluidic device used in this study. Control lines (25 micron depth, square profile, bottom layer) are shown in red, flow lines (10 micron depth, rounded profile, top layer) are shown in blue, and large channels/chambers (60 micron depth, rounded profile, top layer), are shown in green. 3) Photograph of a similar 48-channel microfluidic device with tubing to power control lines attached. 4) Plan view zoom of box in part 2) showing single sorting/amplification channel. Cells suspension flows vertically in blue channel at left, reagents are supplied to the indicated 'T' junction from a supply line dedicated to one of the 48 reaction channels. Each reagent solution is flushed to the left from the T-junction to back-wash the blue channel before being applied for the single-cell reaction by redirection to the right of the T junction. 5) Plan view micrograph of the device region shown in part 4). 6) Elevation view (cross section) schematic indicating components visible in part 5), and layup of the microfluidic device, including laser-trapped cell. 7) Plan view zoom of box in part 4) showing path by which cells are sorted using the optical trap. Cells traverse about 1.5 mm of channel containing clean buffer across two valves, which are opened sequentially to allow cell to pass. 8-13) Micrographs depicting device and MDA reaction setup. 8) Bare device with air-filled channels. 9) Device with control lines filled with water (low-contrast channels) and pressurized (valves closed, visible where control channels cross air-filled flow lines). 10) Device with reagent and sample lines pre-filled with buffer (high-contrast channels: air; low-contrast channels: buffer). Sorting takes place with this device configuration. 11) Lysis chamber 1 (3.5 nL capacity) after reagent flush and dead-end fill. 12) Lysis chamber 2 (3.5 nL capacity) after reagent flush and dead-end fill. 13) Reaction chamber (60 nL capacity) initial filling by dead-end method after reagent flush. 14) Reaction chamber with nearly-complete dead-end fill. (B)

Snapshots from movies of sorted SFB-1 (1), SFB-2 (2), SFB-3 (3), SFB-4 (4), and SFB-5 (5) filaments. Each movie shows the last stage of the sorting for each individual SFB, right after a filament was transported through the second gate valve (see Fig. S2B-7, the area marked with the asterix to the right of the reagent 'T'-junction and second gate valve). The laser trap is fixed near the end of the fiducial arrow. The arrowhead appears just under 1 micron long. SI Movies 1-5 are available at http://genome.cshlp.org/ and http://asiago.stanford.edu/. (C) Bacterial cells in fecal samples from the 1) SFB-monocolonized mouse and a 2) conventional SFB-positive mouse were fixed according to a protocol established for Gram-positive bacteria. Red label indicates SFB-specific 16S rRNA (probe SFB1266-Cy3) and blue indicates bacterial 16S rRNA (probe EUB338-Cy5). Some unlabeled cells in 2) may be Gram-negative bacteria, as fixation was optimized for Gram-positive bacteria. Confocal transmission and superimposed fluorescent images are shown.


A
SFB1
SFB2
SFB3
SFB4
SFB5
0 %
25 %
50 %
75 %
100 %
B
shotgun sequence (Mb)
C
Peptostrepto-coccaceae
Lachno-spiraceae
Rumino-coccaceae
Clostridiaceae 1
GC content of protein-coding genes (%)
C. difficile 630
C. difficile CD196
C. sticklandii DSM 519
C. phytofermentans ISDg
C. saccharolyticum WM1
C. cellulolyticum H10
C. thermocellum ATCC 27405
C. acetobutylicum ATCC 824
C. beijerinckii NCIMB 8052
C. botulinum ATCC 19397
C. botulinum A str. ATCC 3502
C. botulinum A str. Hall
C. botulinum A2 str. Kyoto
C. botulinum A3 str. Loch Maree
C. botulinum B str. Eklund 17B
C. botulinum B1 str. Okra
C. botulinum Ba4 str. 657
C. botulinum E3 str. Alaska E43
C. botulinum F str. Langeland
C. cellulovorans 743B
C. kluyveri DSM 555
C. kluyveri NBRC 12016
C. novyi NT
C. ljungdahlii DSM 13528
C. perfringens ATCC 13124
C. perfringens SM101
C. perfringens str. 13
C. tetani E88
SFB-co
SFB-SU
D
16S rDNA
5S rDNA
23S rDNA
total

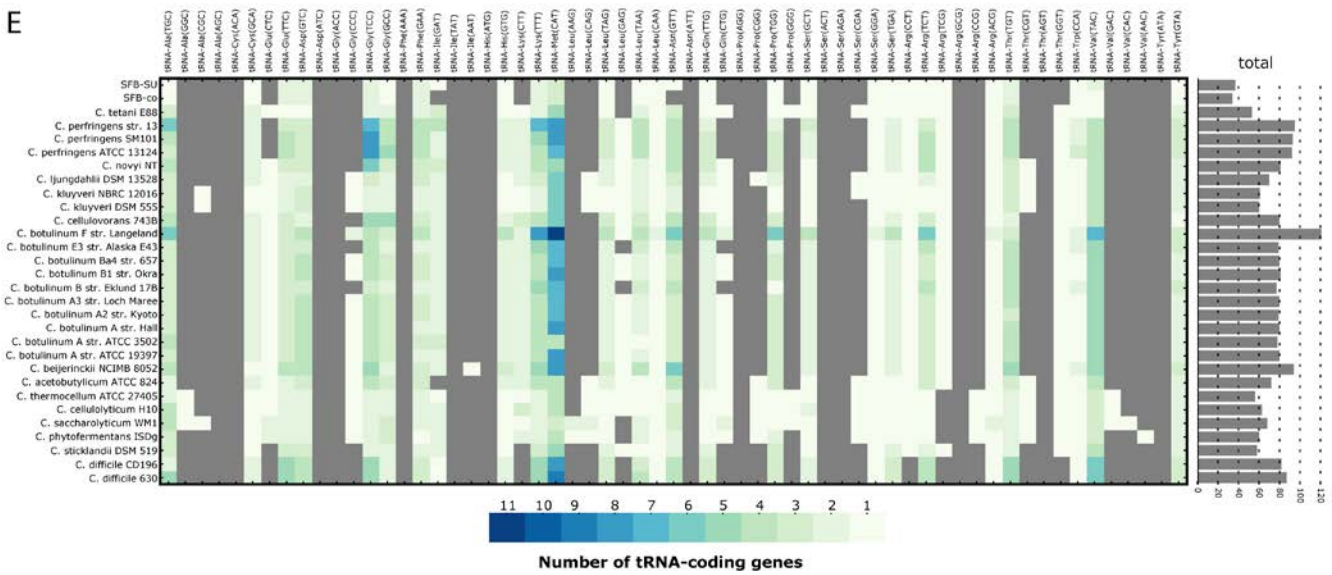


Fig. S3. Read G+C content distribution, and Genome statistics. (A) Distribution of read G+C content for read sets obtained for each individual SFB filament (SFB-1 to SFB-5). (B) Rarefaction analysis to determine the approximate genome size of SFB. The total assembly size is predicted from an asymptote around 1.62 Mb. (C) Distribution of GC content of protein-coding genes in SFB-co, SFB-mouse-SU, and other clostridia. The box designates the lower and upper quartile values with a line (red) at the median. (D) Heatmap displaying the number of 16S, 5S, and 23S rDNA-coding genes observed in SFB-co, SFB-mouse-SU, and other clostridia. The number of total rRNAs are summarized to the right. (E) Heatmap displaying the number of tRNA-coding genes observed in SFB-co and other clostridia. tRNA species are listed as columns, including anticodons, and clostridial strains are listed in rows. The number of total tRNAs are summarized to the right.

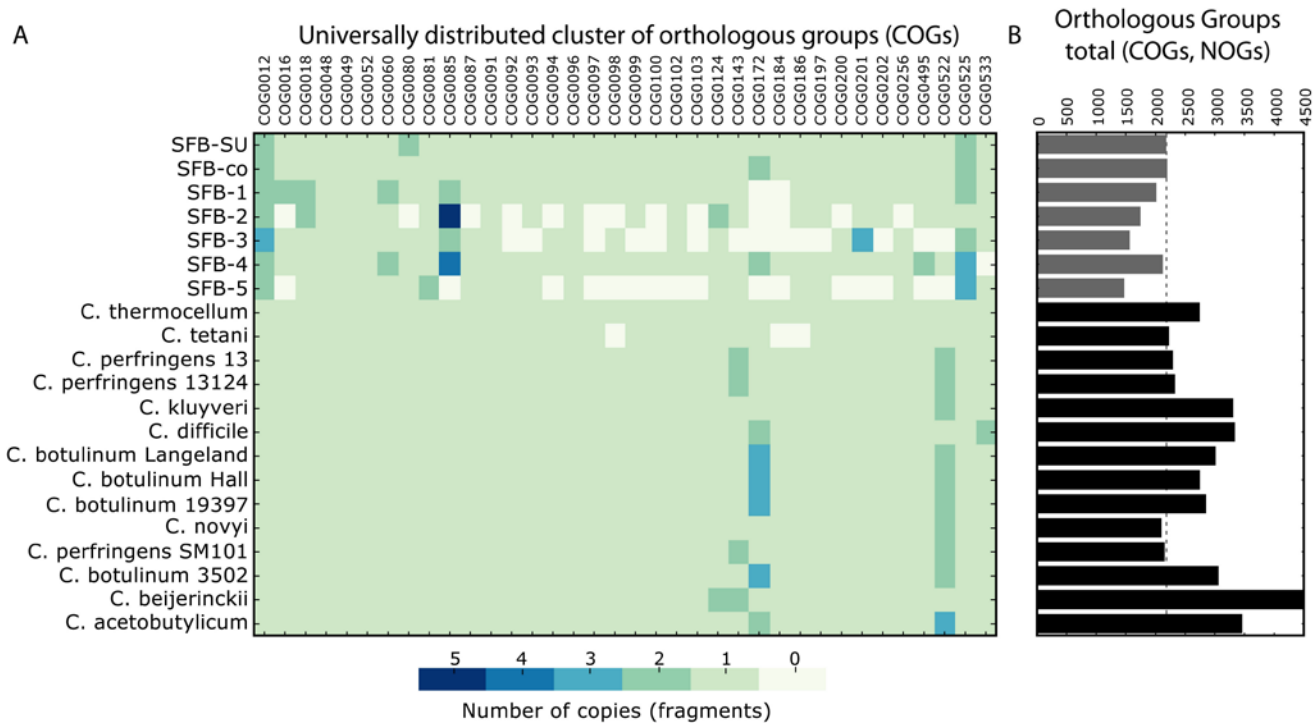


Fig. S4. Clusters of orthologous groups (COGs) in SFB genomes. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins)-based analysis of the predicted proteomes of the SFB assemblies in comparison to other clostridia. (A) Distribution of the 36 universal distributed COGs (Ciccarelli et al. 2006) in the individual SFB assemblies and complete clostridial genomes. Note, in the case of the individual SFB assemblies occasional COG sequences were distributed over a number of contigs such as for COG0085 in SFB-2 and SFB-4 due to gene fragmentation. (B) Total numbers of orthologous groups (OGs) observed for the SFB assemblies and complete clostridial genomes. Strain-designation for each bar as in (A).

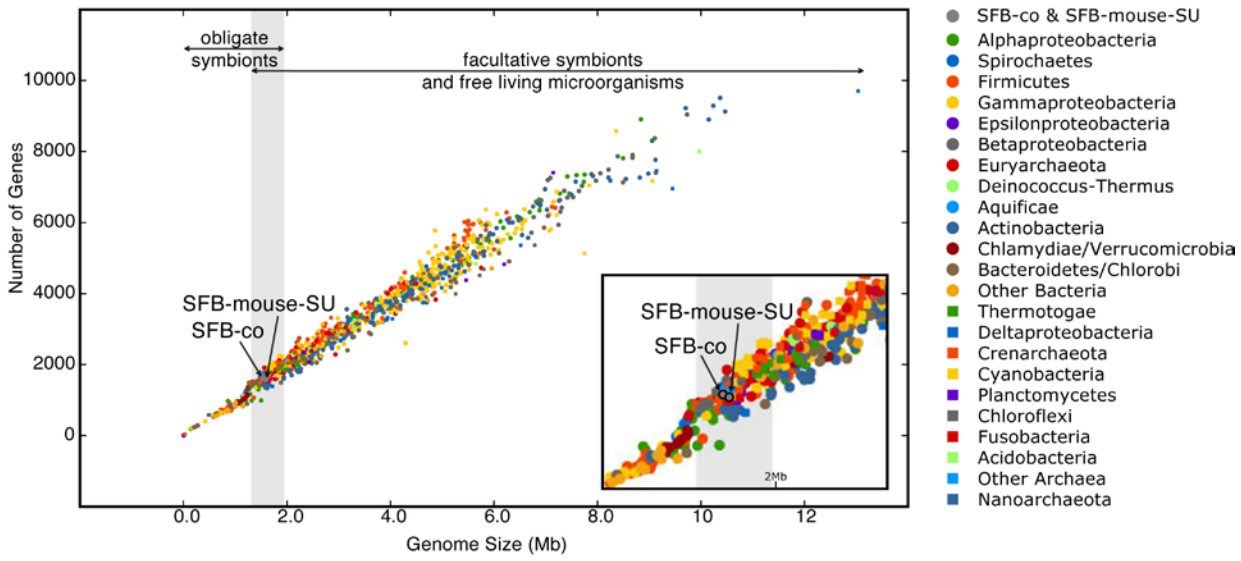


Fig. S5. Relationship between the genome size and number of genes for SFB-co, SFB-mouse-SU, and 1247 complete microbial genomes. Microbial strains are color-coded by taxonomic affiliation at the phylum or class level. Insert image is a close-up of the region around SFB-co and SFB-mouse-SU. Information on microorganism habitat was obtained from NCBI and (Podar et al. 2008).

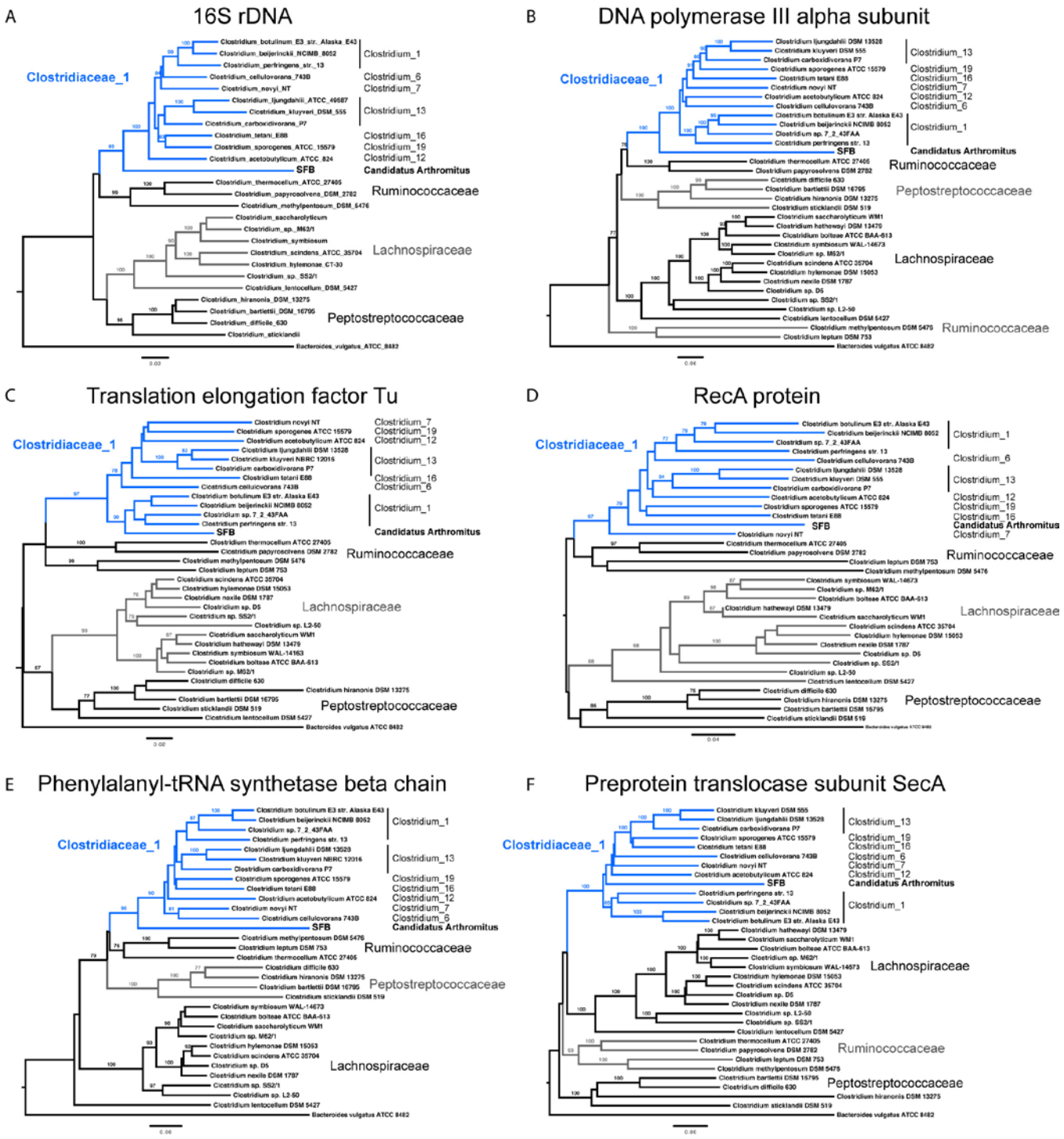


Fig. S6. Phylogenetic analysis of 16S rDNA genes and conserved protein sequences from SFB and other clostridia. (A) 16S rDNA sequences were aligned using SINA, and phylogeny inferred using the maximum likelihood method. (B-F) Protein sequences (PolC, DNA polymerase III alpha subunit; EFTu, Translation elongation factor Tu; RecA, Recombinase A; PheT, Phenylalanyl-tRNA synthetase beta chain; SecA, Protein translocase subunit SecA) were aligned using MUSCLE, and phylogeny inferred using the maximum likelihood method. Bootstrap statistical support values ≥ 75 are shown.

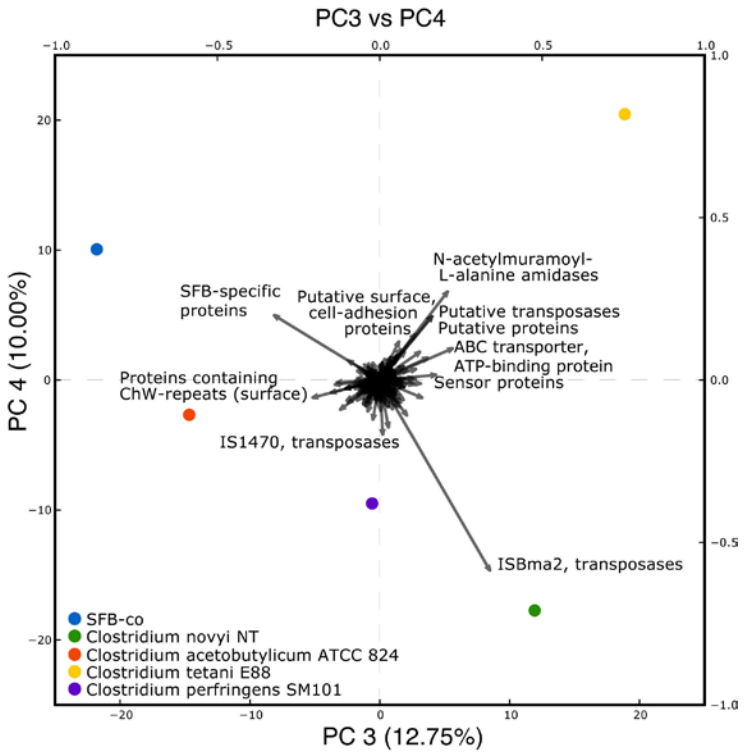


Fig. S7. PC3 vs. PC4 of the principal component analysis (PCA) of protein clusters from the predicted proteomes of SFB-co and four other members of the *Clostridiaceae* 1. A vector in close proximity of a clostridial strain indicates that the protein cluster dominates in that particular strain. The length of a vector indicates the influence of that particular protein cluster in relation to all clusters, whereby a long vector indicates a high influence.

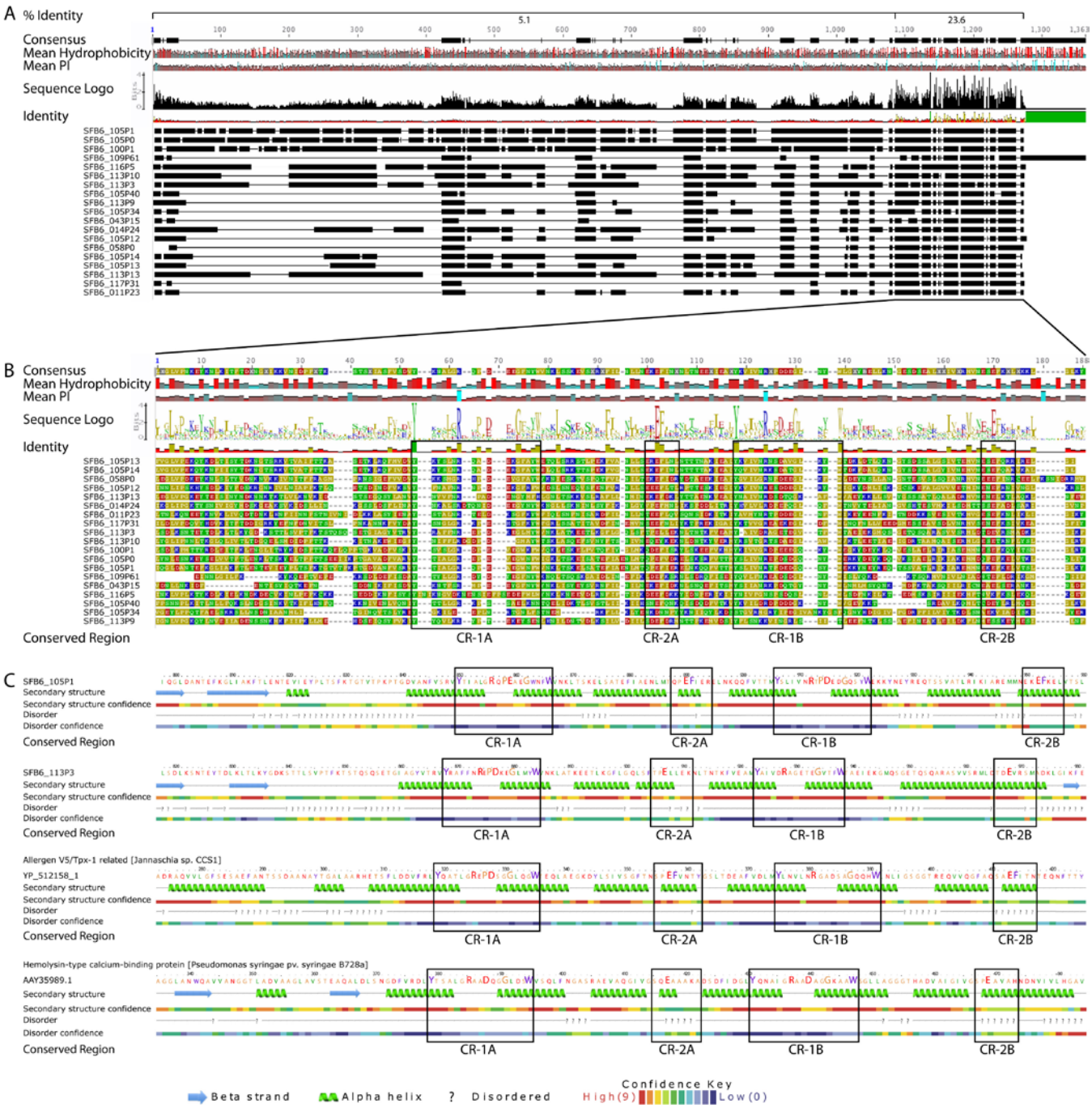


Fig. S8. Sequence conservation among SFB.Cluster.1 proteins. (A+B) Sequences were aligned using MUSCLE. Close-up of the C-terminal domain from (A) in (B). The four regions that exhibit high conservation are designated as CR-1A, CR-2A, CR-1B, and CR-2B, whereby CR-1A and CR-1B, and CR-2A and CR-2B exhibit similarity respectively. (C) Secondary structure prediction of the C-terminal domain from two SFB PF13946-domain-proteins (SFB6_105P1, and SFB6_113P3) and two proteins from other bacteria (YP_512158.1, and AAY35989.1) using the Protein Homology/Analogy Recognition Engine V 2.0 PHYRE[2] (Kelley and Sternberg 2009). Proteins from SFB-co are shown here, for the identical homologs in SFB-mouse-SU, see locus_tag IDs listed in Table S2.

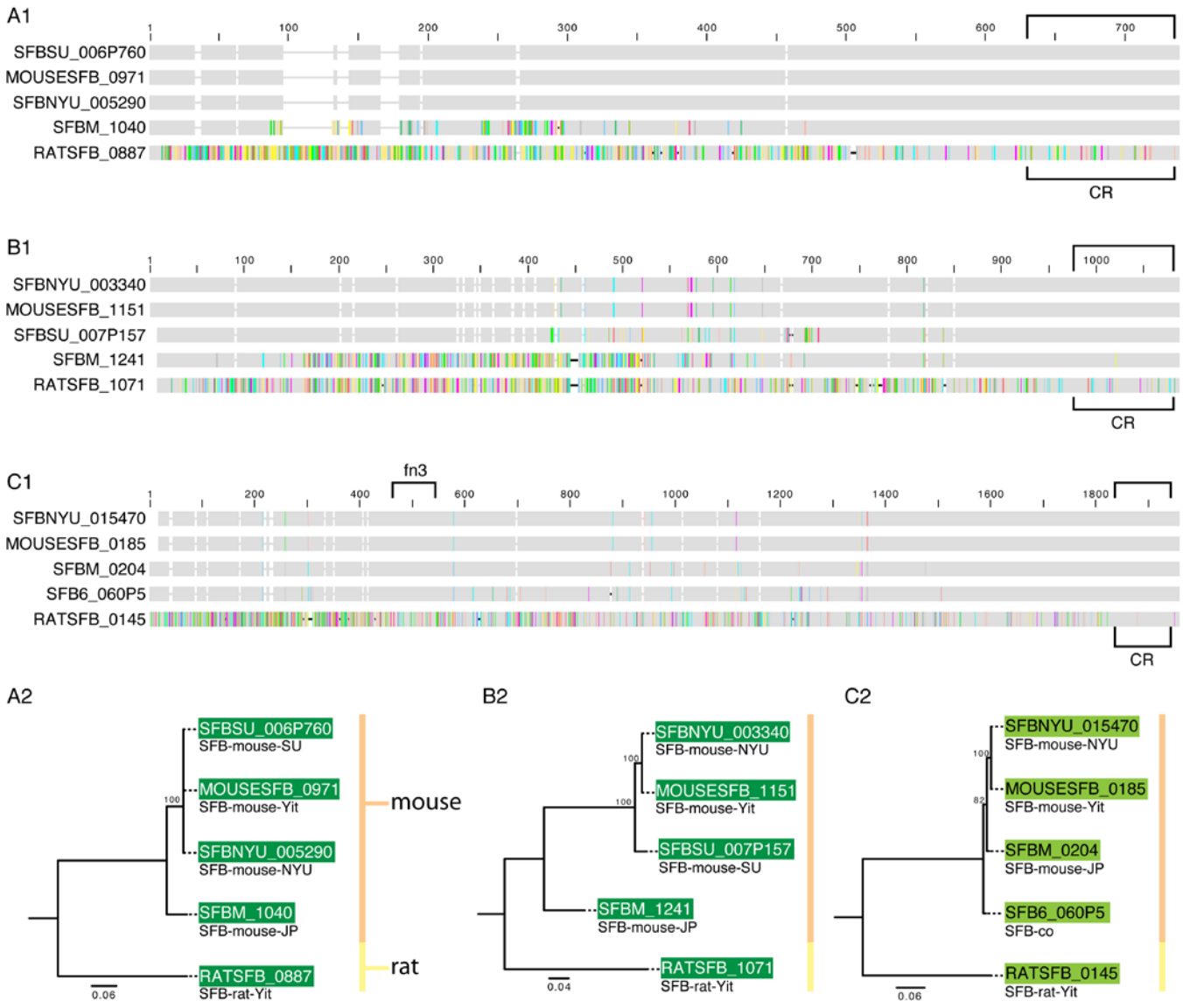


Fig. S9. Polymorphisms in SFB-specific Cluster.1 and Cluster.3 proteins among different SFB lineages. (A1+B1+C1) Protein sequence alignment. Identical amino acid residues are displayed as grey, and differing amino acid residues as colored vertical lines. The conserved C-terminal region (see Fig. S8) is indicated, respectively. The fibronectin type III domain (PF00041-fn3) in the SFB Cluster.3 proteins are indicated. (A2+B2+C2) Maximum likelihood trees based on protein alignments in (A1+B1+C1). Proteins are color-coded as in Fig. 3. Bootstrap statistical support values ≥ 75 are shown. A) Cluster.1 protein SFBSU_006P760 (SFB6_105P14, SFB-co) and homologs from other SFB. B) Cluster.1 protein SFBSU_007P157 (SFB6_113P3, SFB-co) and homologs from other SFB. C) Cluster.3 protein SFB6_060P5 (SFBSU_003P27, SFB-mouse-SU) and homologs from other SFB.

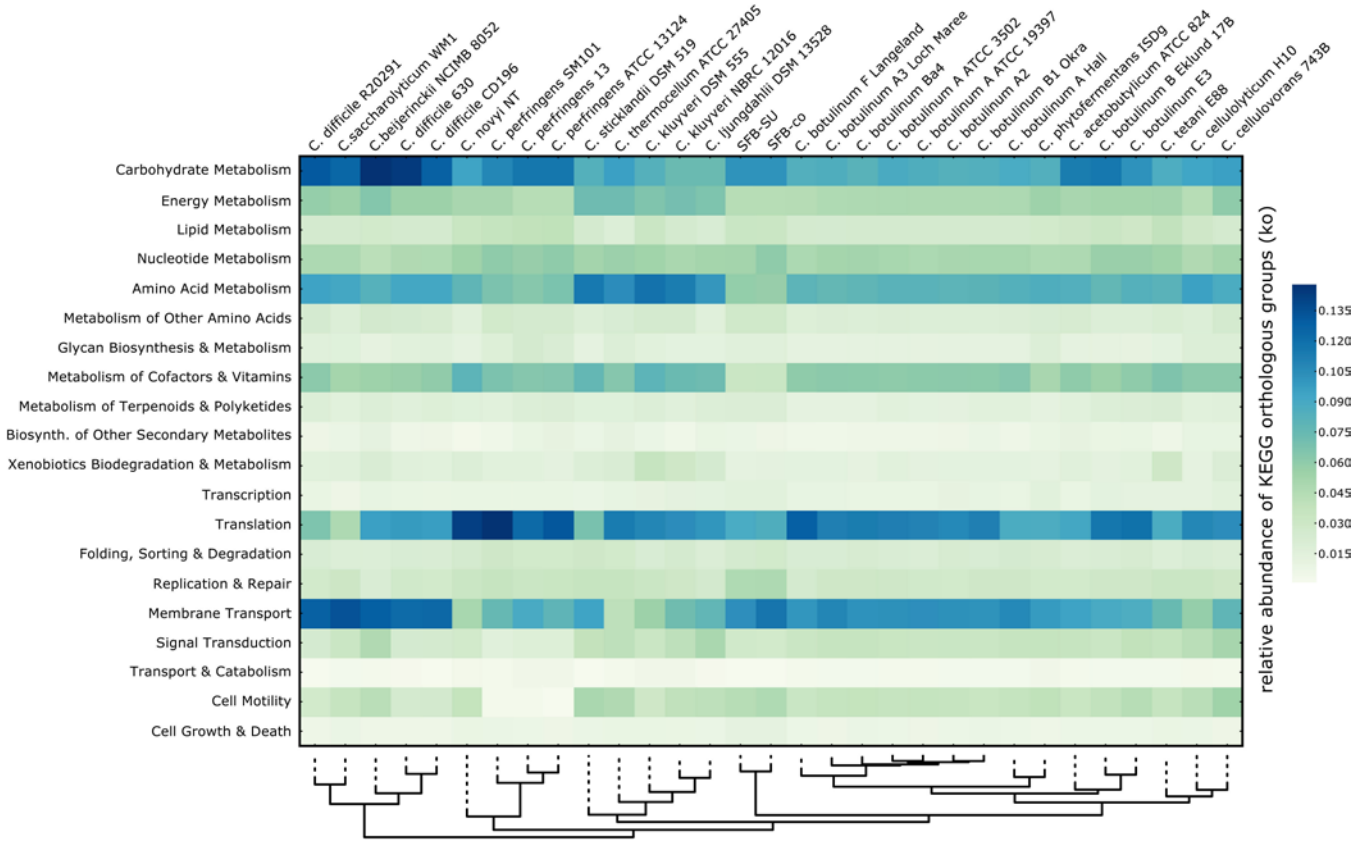


Fig. S10. Relative abundance of KEGG metabolic pathways in SFB-co, SFB-mouse-SU, and other clostridia clustered using a Euclidean distance metric and displayed as heatmap. The relative abundances based on the number of proteins assigned to a particular pathway per strain are displayed from low (light green) to high (dark blue).

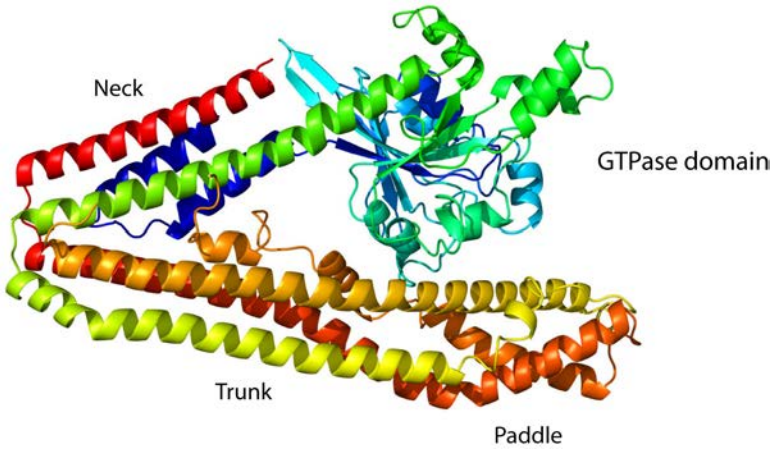
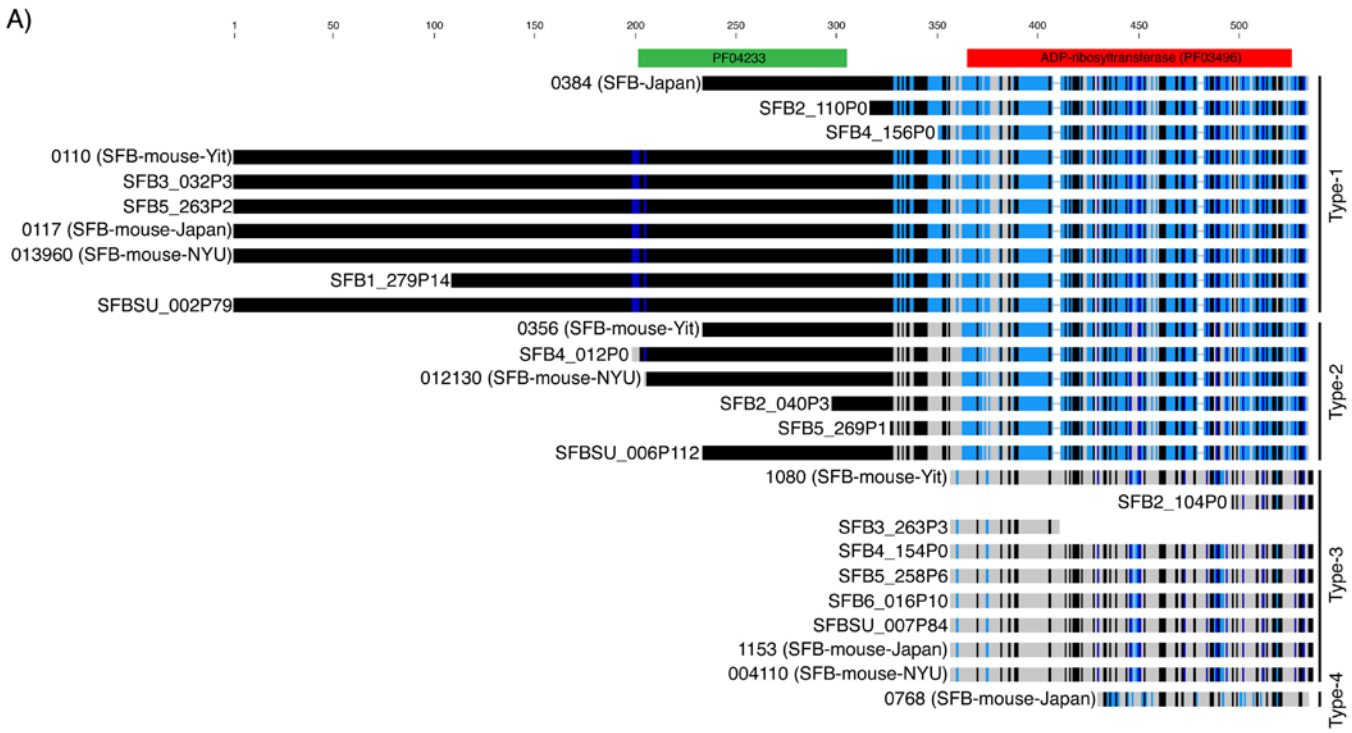


Fig. S11. Predicted three-dimensional structure of the SFB bacterial dynamin-like protein (BDLP) (SFB7_C6P653). 630 residues (85% of the sequence) have been modeled with 100% confidence by the highest scoring template, the bacterial dynamin-like protein from *Nostoc punctiforme* (PDB ID 2J69). The protein comprises the GTPase head, a four-helix neck and trunk bundle, and the paddle, which is described for *N. punctiforme* to mediate membrane-binding. Compare SFB BDLP 3D structure to Fig. 2a, and the nucleotide-free state in Fig. 2c in reference (Low and Lowe 2006) and Fig. 1A in reference (Low et al. 2009). Structure colored by rainbow from N (blue) to C (red) terminus.



B)

Length/domain type	Location in genome	SFB-mouse-SU	SFB-mouse-Yit	SFB-mouse-NYU	SFB-mouse-Japan	SFB-rat-Yit
		ADPRT Sequence Types				
I	phage 1	Type-1	Type-1	Type-1	Type-1	-
II	phage 2	Type-2	Type-2	Type-2	Type-1	-
III	chromos.	Type-3	Type-3	Type-3	Type-3	some nt seq similarity intergenetic
IV	phage 3	-	-	-	Type-4	-

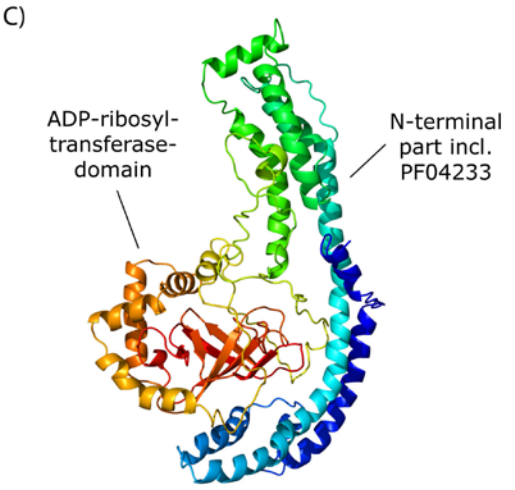


Fig. S12. ADP-ribosyltransferase (ADPRT) sequence types in SFB genomes. A) Magnified view of protein alignment of all SFB ADPRT sequence types found in the SFB genome sequences reported here, and other SFB genomes. Sequence Identity: Black, 100%; dark blue, 80-99%; light blue, 79-60%; grey, less than 60%. Classification of the sequences into 4 types, Type-1 to Type-4. The ADPRT-domain is indicated in red, and the PF04233-domain in green. B) SFB ADPRT sequence types found in SFB-mouse-SU, SFB-mouse-Yit, SFB-mouse-NYU, SFB-mouse-Japan, and their location within the genome. C) Predicted three-dimensional (3D) structure of the SFB ADP-ribosyltransferase Type-1 (SFBSU_002P79). The C-terminal ADPRT-domain has been modeled with 100% confidence by the highest scoring template, the Vip2 protein from *Bacillus cereus* (PDB ID 1QS2). The N-terminal protein part was modeled with low confidence. Compare predicted 3D structure of SFB-ADPRT-domain-containing protein with predicted 3D structure of EFV-toxin in *E. faecalis* V583 (Fig. 5D in (Fieldhouse et al. 2010)).

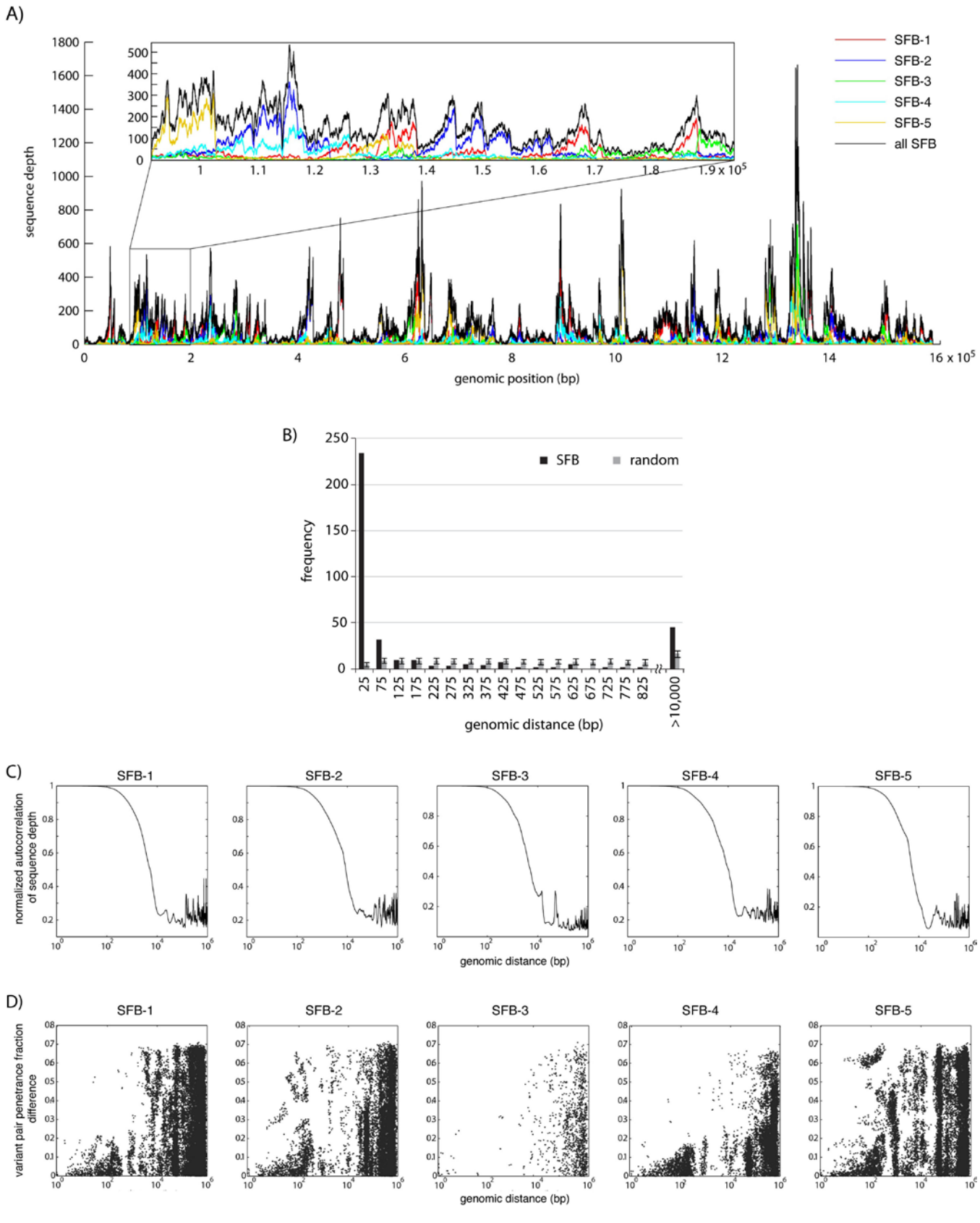

A)
sequence depth
genomic position (bp)
SFB-1
SFB-2
SFB-3
SFB-4
SFB-5
all SFB
B)
frequency
SFB
random
genomic distance (bp)
C)
SFB-1
SFB-2
SFB-3
SFB-4
SFB-5
normalized autocorrelation of sequence depth
genomic distance (bp)
D)
SFB-1
SFB-2
SFB-3
SFB-4
SFB-5
variant pair penetrance fraction difference
genomic distance (bp)

Fig. S13. Read depth for individual SFB filaments, and random & observed SNP distribution. (A) Read depth for sequences from individual SFB filaments. Filtered reads from each individual filament were mapped against the genome sequence SFB-mouse-Yit (AP012209). The observed uneven read distribution is a result of amplification biases during MDA (for example, see (Dean et al. 2001; Rodrigue et al. 2009). (B) Histogram of genomic distance between loci exhibiting genomic variation among individually amplified SFB filaments versus a random distribution. The distance distribution in the SFB data shows strong enhancement at separations below 100 bp and above 10,000 bp, as well as depletion at intermediate separations versus a random model where variants (sampled at the same genomic density) are equally likely to arise at all positions. These results show both strong clustering of groups of variants in the SFB genome as well as large regions of the genome where variants are less likely to be observed. (C) Autocorrelation of coverage depth in single-SFB filament datasets. (D) Pairwise penetrance fraction difference is correlated with genomic separation. The difference in penetrance fraction as a function of genomic separation for all pairs of partially-penetrant variants in each of the five filaments is plotted. A marked depletion in variant pairs with disparate degrees of penetrance is evident within the MDA correlation length for all five datasets compared with the relatively uniform distribution of penetrance disparity among pairs of variants separated by more than 20,000 bp.

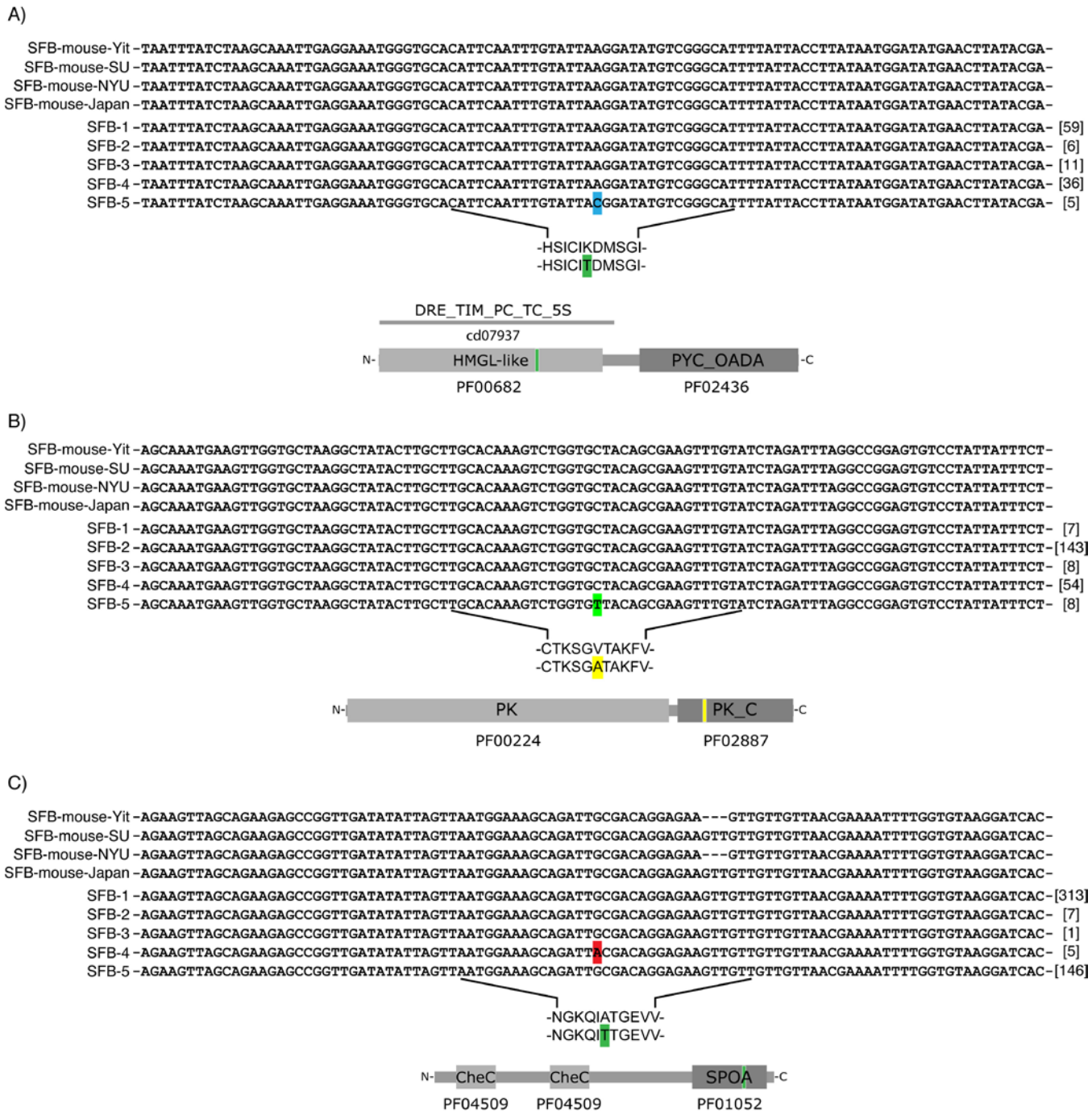
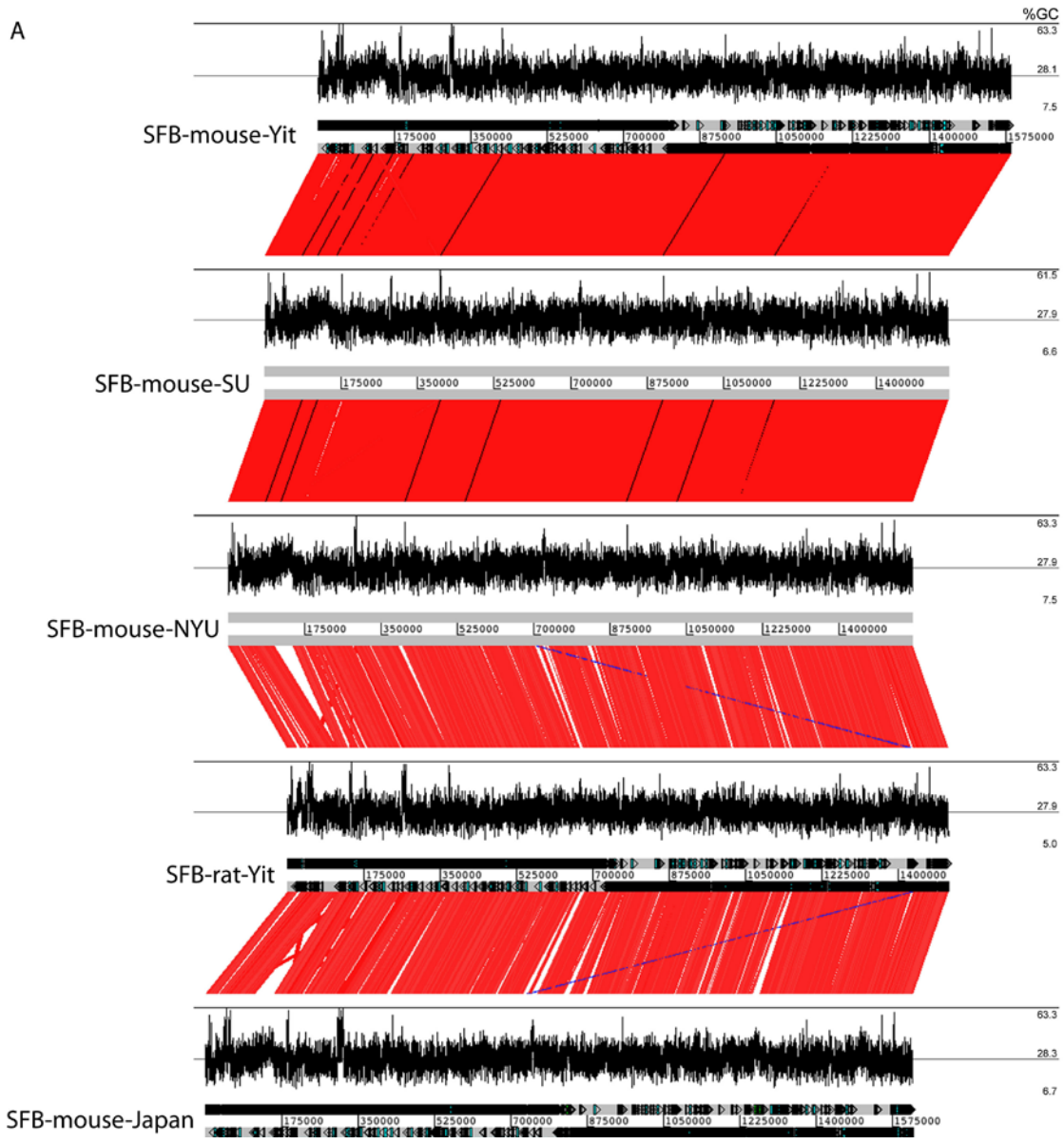


Fig. S14. Examples of inter-filament variability. A) Nucleotide substitution in the SFB-5 gene encoding for oxaloacetate decarboxylase subunit alpha, OadA (SFB5_131P2). The OadA-encoding gene is situated in the genetic neighborhood of ORFs predicted to encode for oxaloacetate decarboxylase gamma chain (OadG), a biotin/lipoyl attachment domain-containing protein, and Na+-transporting methylmalonyl-CoA/oxaloacetate decarboxylase beta subunit (OadB). Together they may form a oxaloacetate decarboxylase Na+ pump, that generates pyruvate and CO_2 from oxaloacetate (Studer et al. 2007). A to C transversion in SFB-5 (Fig. S2B5), leading to a predicted threonine (T) residue compared to lysine (K) in the other individual SFB. The lysine (K) residue is situated in the

DRE_TIM_PC_TC_5S (cd07937) domain, is conserved in many bacteria, and required for catalytic activity and potentially ion (Zn^{2+}) binding (Hall et al. 2004; Studer et al. 2007). B) Cytosin to thymine nucleotide conversion leading to a valine (V) to alanine (A) change in the pyruvate kinase (SFB5_060P2)-encoding enzyme in SFB-5. The pyruvate kinase (PK) catalyzes the conversion of phosphoenolpyruvate to pyruvate with concomitant phosphorylation of ADP to ATP. C) Nucleotide polymorphism in the flagellar motor switch protein (FliN)-encoding gene. A threonin residue in the SPOA (surface presentation of antigens)-domain of the SFB-4 protein (SFB4_057P10) appears to be present compared to an alanine (A) residue in the other SFB filaments. Read depths at the particular loci are specified for each individual SFB (SFB-1 to SFB-5), respectively.


A
%GC
63.3
28.1
7.5
SFB-mouse-Yit
175000
350000
525000
700000
875000
1050000
1225000
1400000
1575000
61.5
27.9
6.6
SFB-mouse-SU
63.3
27.9
7.5
SFB-mouse-NYU
63.3
27.9
5.0
SFB-rat-Yit
63.3
28.3
6.7
SFB-mouse-Japan

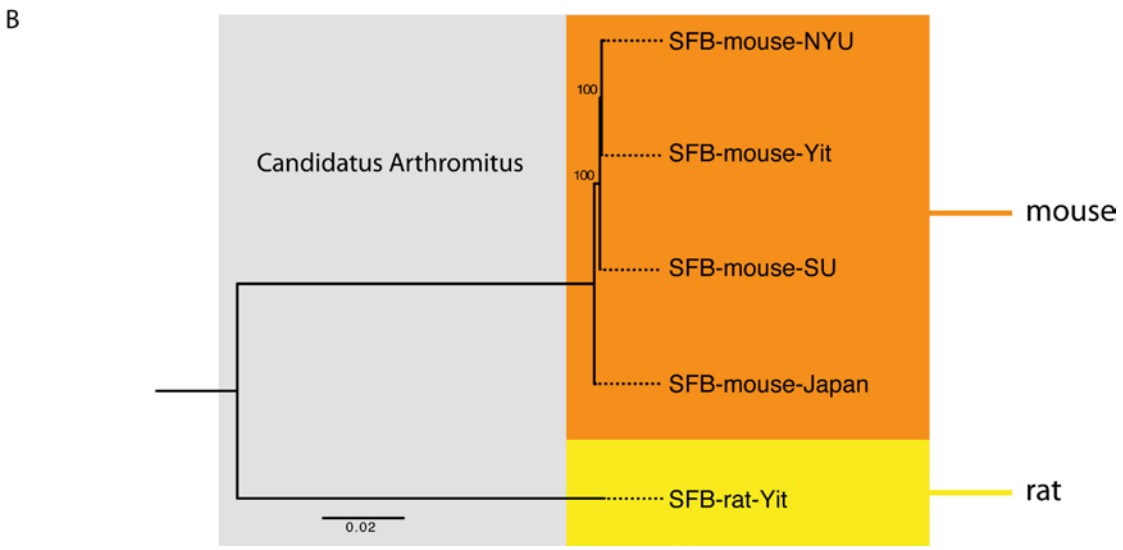


Fig. S15. Genome sequence comparison of five SFB genomes (SFB-mouse-SU, SFB-mouse-Yit, SFB-mouse-NYU, SFB-mouse-Japan, and SFB-rat-Yit). A) Multiple genome sequence alignment generated using Artemis Comparison Tool (ACT). GC content (%) for each genome are displayed above the genome sequence, respectively, using default window size (120 bases). Regions missing from SFB-rat-Yit compared to the SFB-mouse genomes include the phage elements. B) Phylogenetic analysis based on a Kalign2 multiple genome sequence alignment (consisting of 1,455,482 columns). Phylogeny was inferred using the maximum likelihood method. Bootstrap statistical support values ≥ 75 are shown.

3) SI TABLES

Table S1: Additional genome information for the individual SFB assemblies (SFB-1 to SFB-5) and the co-assemblies (SFB-co, SFB-mouse-SU).

	SFB-1	SFB-2	SFB-3	SFB-4	SFB-5	SFB-co	SFB-mouse-SU
Number of contigs	368	330	441	356	339	132	10
Number of contigs bases	1,270,456	1,135,433	951,018	1,342,936	928,891	1,529,892	1,566,160
N50	7,465	8,836	4,578	8,707	7,476	31,015	816,772
Number of reads	148,261	129,310	91,029	122,450	123,279	593,202	747080
Number of reads bases	51,277,095	44,498,803	28,768,162	38,684,608	38,952,490	194,190,653	226,643,329
Min contig length (bp)	190	177	247	144	167	112	1026
Max contig length (bp)	41,349	48,350	23,990	31,800	26,953	92,740	816,772
Mean contig length (bp)	3,452.33	3,440.71	2,156.50	3,772.29	2,740.09	11,590.09	156,616
Number of genes: all	1,573	1,395	1,330	1,658	1,219	1,613	1,557
Number of genes: coding	1,550	1,364	1,307	1,626	1,203	1,577	1,503
Number of genes: non-coding	23	31	23	32	16	36	54
%GC of genes: all (mean)	28.69	28.94	28.91	28.63	28.70	28.77	29.0
%GC of genes: coding (mean)	28.32	28.35	28.46	28.14	28.37	28.18	28.2
%GC of genes: non-coding (mean)	54.08	54.81	54.28	53.95	53.82	54.88	52.5
Contigs: multi-gene	278	251	304	269	227	112	10
Contigs: single-gene	90	79	137	86	112	20	0
Contigs: no-gene	0	0	0	1	0	0	0
Proteins bitscore 40+ -<60*	193	184	204	226	170	190	146
Total number of COGs	1,528	1,306	1,186	1,566	1,127	1,621	1,578

Number of unique COGs	945	859	771	973	741	1,041	1,049
Total number of NOGs	482	441	380	560	347	578	604
Number of unique NOGs	398	368	308	476	292	470	493
Number of proteins with ko (KEGG) assignments	809	716	649	816	596	803	770
Total number of Pfam-A domains	1459	1264	1125	1524	1066	1664	1705
Assembly size (bp) when reads are mapped to SFB-mouse-SU	1,486,989	1,491,070	1,399,153	1,502,331	1,279,456	-	-

* in BLASTP search against NCBI Complete Microbial Genomes 2010-12-14

Table S2: Clusters of SFB-specific proteins in SFB-co and SFB-mouse-SU.

Clusters with 4+ members are listed.

Protein SFB-co (SFB6_) SFB-mouse-SU (SFBSU_)	**Length (aa)**	**Secreted* / TM domain****	**Pfam-A** (release 25.0)
SFB.Cluster.1	Putative secreted proteins, similar to PF13946 (DUF4214)***		
SFB6_100P1 **SFBSU_006P743**	1149	yes / nd	-
SFB6_105P0 **SFBSU_006P746**	1095	yes / nd	-
SFB6_105P1 **SFBSU_006P747**	1084	yes / nd	-
SFB6_105P12 **SFBSU_006P758**	449	yes / nd	-
SFB6_105P13 **SFBSU_006P759**	566	yes / nd	-
SFB6_105P14 **SFBSU_006P760**	672	yes / nd	-
SFB6_105P34 **SFBSU_006P781**	481	yes / nd	-
SFB6_105P40 **SFBSU_006P787**	360	yes / yes	PF00427 (PBS_linker_poly)
SFB6_109P61 **SFBSU_006P647**	375	yes / yes	-
SFB6_011P23 **SFBSU_006P284**	428	yes / nd	-
SFB6_113P10 **SFBSU_007P160**	719	yes / yes	-
SFB6_113P13 **SFBSU_007P157**	1050	yes / yes	-
SFB6_113P3 **SFBSU_007P167**	981	yes / nd	PF03099 (*BPL_LplA_LipB*)
SFB6_113P9 **SFBSU_007P161**	374	yes / yes	-
SFB6_116P5 **SFBSU_007P24**	960	yes / yes	-
SFB6_117P31 **SFBSU_006P211**	236	yes / yes	-
SFB6_014P24 **SFBSU_006P70**	766	yes / yes	-
SFB6_043P15 **SFBSU_002P10**	309	yes / nd	PF00963 (Cohesin)
SFB6_058P0 **SFBSU_006P586**	278	nd / nd	-
SFB.Cluster.2	N-acetylmuramoyl-L-alanine amidases (autolysins)		
SFB6_106P19 **SFBSU_006P534**	222	yes / nd	PF01510 (Amidase_2)
SFB6_116P6 **SFBSU_007P25**	225	nd / nd	PF01510 (Amidase_2)
SFB6_024P13 **SFBSU_009P95**	233	yes / nd	PF01510 (Amidase_2)
SFB6_063P5	143	yes / yes	PF01510 (Amidase_2)

SFBSU_002P106 YP_878230.1	431	yes / nd	PF01510 (Amidase_2), PF01471 (*PG_binding_1*), PF01832 (*Glucosaminidase*)
SFB.Cluster.3	Putative secreted proteins (C-term similar to SFB.Cluster.1 C-term)		
SFB6_060P5 **SFBSU_003P27******	1916	yes / yes	PF00041 (fn3)
SFB6_105P46 **SFBSU_007P2**	1432	yes / yes	-
SFB6_117P0 (N-term) **SFB6_087P2** (C-term) **SFBSU_006P242**	2034	yes / yes	-
SFB6_117P1 **SFBSU_006P241**	1461	yes / yes	-
SFB.Cluster.4			
SFB6_018P10 **SFBSU_002P128**	291	yes / nd	-
SFB6_018P7 **SFBSU_002P131**	304	nd / nd	-
SFB6_018P6 **SFBSU_002P132**	335	nd / nd	-
SFB6_018P8 **SFBSU_002P130**	298	yes / nd	-

* Secretory proteins predicted by SignalP-NN gram+, and/or SignalP-HMM gram+, or Secretome2.0 gram+, nd = not detected.
** Transmembrane domains (TM) predicted by TMHMM-2.0, nd = not detected.
*** https://pfamsvn.sanger.ac.uk/svn/pfam/trunk/Data/Families/PF13946/
**** SFBSU_003P27 is truncated as a result of a stop codon generated in the process of co-assembly.

Table S3: Clusters of SFB-specific proteins in SFB-mouse and -rat genomes.

Sequence identities (%) are in comparison to the SFB-co/SFB-mouse-SU reference protein.

Protein SFB-co (SFB6_) SFB-mouse-SU (SFBSU_)	**SFB-mouse-Yit** (Identities in %)	**SFB-mouse-NYU** (Identities in %)	**SFB-mouse-Japan** (Identities in %)	**SFB-rat-Yit** (Identities in %)
SFB.Cluster.1	Putative secreted proteins, similar to PF13946 (DUF4214)*			
SFB6_100P1 **SFBSU_006P743**	MOUSESFB_0955 (99%)	SFBNYU_005460 (99%)	SFBM_1023 (99%)	RATSFB_0872 (72%)
SFB6_105P0 **SFBSU_006P746**	MOUSESFB_0957 (100%)	SFBNYU_005430 (100%)	SFBM_1026 (99%)	RATSFB_0874 (77%)
SFB6_105P1 **SFBSU_006P747**	MOUSESFB_0958 (100%)	SFBNYU_005420 (100%)	SFBM_1027 (99%)	RATSFB_0875 (78%)
SFB6_105P12 **SFBSU_006P758**	MOUSESFB_0969 (100%)	SFBNYU_005310 (100%)	SFBM_1038 (99%)	RATSFB_0885 (83%)
SFB6_105P13 **SFBSU_006P759**	MOUSESFB_0970 (100%)	SFBNYU_005300 (100%)	SFBM_1039 (99%)	RATSFB_0886 (56%)
SFB6_105P14 **SFBSU_006P760**	MOUSESFB_0971 (100%)	SFBNYU_005290 (100%)	SFBM_1040 (90%)	RATSFB_0887 (63%)
SFB6_105P34 **SFBSU_006P781**	MOUSESFB_0990 (100%)	SFBNYU_005070 (100%)	SFBM_1060 (100%)	RATSFB_0905 (84%)
SFB6_105P40 **SFBSU_006P787**	MOUSESFB_0996 (99%)	SFBNYU_005010 (100%)	SFBM_1066 (99%)	RATSFB_0911 (68%)
SFB6_109P61 **SFBSU_006P647**	MOUSESFB_0867 (100%)	SFBNYU_006470 (100%)	SFBM_0929 (100%)	RATSFB_0786 (85%)
SFB6_011P23 **SFBSU_006P284**	MOUSESFB_0517 (100%)	SFBNYU_010260 (100%)	SFBM_0553 (100%)	RATSFB_0456 (77%)
SFB6_113P10 **SFBSU_007P160**	MOUSESFB_1154 (100%)	SFBNYU_003310 (100%)	SFBM_1244 (100%)	RATSFB_1074 (85%)
SFB6_113P13 **SFBSU_007P157**	MOUSESFB_1151 (96%)	SFBNYU_003340 (96%)	SFBM_1241 (72%)	RATSFB_1071 (55%)
SFB6_113P3 **SFBSU_007P167**	MOUSESFB_1162 (99%)	SFBNYU_003240 (100%)	SFBM_1251 (100%)	RATSFB_1081 (75%)
SFB6_113P9 **SFBSU_007P161**	MOUSESFB_1155 (100%)	SFBNYU_003300 (100%)	SFBM_1245 (100%)	RATSFB_1075 (74%)
SFB6_116P5 **SFBSU_007P24**	MOUSESFB_1022 (100%)	SFBNYU_004730 (100%)	SFBM_1094 (100%)	RATSFB_0938 (80%)
SFB6_117P31 **SFBSU_006P211**	MOUSESFB_0450 (100%)	SFBNYU_011020 (100%)	SFBM_0481 (100%)	RATSFB_0391 (76%)
SFB6_014P24 **SFBSU_006P70**	MOUSESFB_0319 (100%)	SFBNYU_012560 (100%)	SFBM_0343 (99%)	RATSFB_0273 (81%)
SFB6_043P15 **SFBSU_002P10**	MOUSESFB_0048 (100%)	SFBNYU_014660 (100%)	SFBM_0048 (100%)	RATSFB_0049 (77%)
SFB6_058P0 **SFBSU_006P586**	MOUSESFB_0807 (100%)	SFBNYU_007150 (100%)	SFBM_0865 (99%)	RATSFB_0724 (46%)
SFB.Cluster.2	N-acetylmuramoyl-L-alanine amidases (autolysins)			
SFB6_106P19 **SFBSU_006P534**	MOUSESFB_0756 (100%)	SFBNYU_007660 (100%)	SFBM_0814 (100%)	RATSFB_0673 (79%)
SFB6_116P6 **SFBSU_007P25**	MOUSESFB_1023 (100%)	SFBNYU_004720 (100%)	SFBM_1095 (100%)	RATSFB_0939 (85%)
SFB6_024P13 **SFBSU_009P95**	MOUSESFB_0234 (100%)	SFBNYU_014930 (100%)	SFBM_0258 (100%)	RATSFB_0194 (90%)

SFB6_063P5 **SFBSU_002P106**	MOUSESFB_0132 (100%)	SFBNYU_013680 (100%)	SFBM_0143 (100%)	RATSFB_0673 (64%)
SFB.Cluster.3	Putative secreted proteins (C-term similar to SFB.Cluster.1 C-term)			
SFB6_060P5 **SFBSU_003P27**	MOUSESFB_0185 (98%)	SFBNYU_015470 (98%)	SFBM_0204 (98%)	RATSFB_0145 (64%)
SFB6_105P46 **SFBSU_007P2**	no ORF predicted	SFBNYU_004940 (100%)	SFBM_1073 (100%)	RATSFB_0918 (66%)
SFB6_117P0 (N-term) **SFB6_087P2** (C-term) **SFBSU_006P242**	MOUSESFB_0477 (100%)	SFBNYU_010700 (100%)	SFBM_0511 (99%)	RATSFB_0418 (71%)
SFB6_117P1 **SFBSU_006P241**	MOUSESFB_0476 (99%)	SFBNYU_010710 (99%)	SFBM_0510 (99%)	RATSFB_0417 (70%)
SFB.Cluster.4				
SFB6_018P10 **SFBSU_002P128**	MOUSESFB_0555 (100%)	SFBNYU_009850 (100%)	SFBM_0595 (100%)	RATSFB_0493 (98%)
SFB6_018P7 **SFBSU_002P131**	MOUSESFB_0154 (100%)	SFBNYU_013410 (100%)	SFBM_0168 (100%)	RATSFB_0107 (66%)
SFB6_018P6 **SFBSU_002P132**	MOUSESFB_0155 (100%)	SFBNYU_009810 (100%)	SFBM_0599 (100%)	RATSFB_0497 (93%)
SFB6_018P8 **SFBSU_002P130**	MOUSESFB_0553 (100%)	SFBNYU_009830 (100%)	SFBM_0597 (100%)	RATSFB_0495 (91%)

* https://pfamsvn.sanger.ac.uk/svn/pfam/trunk/Data/Families/PF13946/

Identity
(Compared to SFB-co/SFB-SU homolog)

100%
99%
98-96%
95-90%
89-80%
79-70%
69-60%
≤ 59%

Table S4: Number of SNPs in protein-coding regions from pairwise comparisons.

	SFB1 vs SFB2	SFB1 vs SFB3	SFB1 vs SFB4	SFB1 vs SFB5	SFB2 vs SFB3	SFB2 vs SFB4	SFB2 vs SFB5	SFB3 vs SFB4	SFB3 vs SFB5	SFB4 vs SFB5
Number of BRH pairs	483	284	607	371	290	486	306	328	229	444
Total codons aligned	124,732	61,592	151,554	91,414	72,329	126,228	77,093	77,177	51,630	116,984
Total codons that differ	63	3	37	23	14	10	71	4	10	43
sSNPs	21	1	10	17	6	3	45	1	7	24
nsSNPs	42	2	27	6	8	7	26	3	3	19
Nucleotide differences										
at codon position 1	32	2	17	4	6	3	16	1	2	9
at codon position 2	25	2	18	1	3	4	13	0	2	9
at codon position 3	32	1	23	19	7	7	53	3	7	32
total	89	5	58	24	16	14	82	4	11	50
Number of SNPs/Total codons aligned	7.14E-04	8.12E-05	3.83E-04	2.63E-04	2.21E-04	1.11E-04	1.06E-03	5.18E-05	2.13E-04	4.27E-04

4) SI REFERENCES

Amann RI, Binder BJ, Olson RJ, Chisholm SW, Devereux R, Stahl DA. 1990. Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl Environ Microbiol* **56**(6): 1919-1925.

Bendtsen JD, Kiemer L, Fausboll A, Brunak S. 2005. Non-classical protein secretion in bacteria. *BMC Microbiol* **5**: 58.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**(4): 783-795.

Blainey PC, Mosier AC, Potanina A, Francis CA, Quake SR. 2011. Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis. *PLoS One* **6**(2): e16626.

Carver T, Berriman M, Tivey A, Patel C, Bohme U, Barrell BG, Parkhill J, Rajandream MA. 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**(23): 2672-2676.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**(5765): 1283-1287.

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**(7): 1394-1403.

Dean FB, Nelson JR, Giesler TL, Lasken RS. 2001. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**(6): 1095-1099.

Drummond A, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S et al. 2011. Geneious v5.4.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5): 1792-1797.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**(7): 1575-1584.

Esteban JA, Salas M, Blanco L. 1993. Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J Biol Chem* **268**(4): 2719-2726.

Fieldhouse RJ, Turgeon Z, White D, Merrill AR. 2010. Cholera- and Anthrax-Like Toxins Are among Several New ADP-Ribosyltransferases. *PLoS Comput Biol* **6**(12): e1001029.

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K et al. 2010. The Pfam protein families database. *Nucleic Acids Res* **38**(Database issue): D211-222.

Fuchs BM, Pernthaler J, Amann R. 2007. Single cell identification by fluorescence in situ hybridization. In *Methods for General and Molecular Microbiology*, (ed. CA Reddy, TJ Beveridge, JA Breznak, G Marzluf, TM Schmidt, LR Snyder). ASM Press, Washington, D.C.

Grissa I, Vergnaud G, Pourcel C. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**(Web Server issue): W52-57.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**(5): 696-704.

Hall PR, Zheng R, Antony L, Pusztai-Carey M, Carey PR, Yee VC. 2004. Transcarboxylase 5S structures: assembly and catalytic mechanism of a multienzyme complex subunit. *Embo J* **23**(18): 3621-3631.

Harrington ED, Arumugam M, Raes J, Bork P, Relman DA. 2010. SmashCell: a software framework for the analysis of single-cell amplified genome sequences. *Bioinformatics* **26**(23): 2979-2980.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**(2): 160-174.

Huang Y, Gilna P, Li W. 2009. Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* **25**(10): 1338-1340.

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M et al. 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37**(Database issue): D412-416.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**(3): 275-282.

Kelley LA, Sternberg MJ. 2009. Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* **4**(3): 363-371.

Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC, Eaton M, Hamady M, Lindsay H, Liu Z et al. 2007. PyCogent: a toolkit for making sense from sequence. *Genome Biol* **8**(8): R171.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**(3): 567-580.

Kuwahara T, Ogura Y, Oshima K, Kurokawa K, Ooka T, Hirakawa H, Itoh T, Nakayama-Imaohji H, Ichimura M, Itoh K et al. 2011. The lifestyle of the segmented filamentous bacterium: a non-culturable gut-associated immunostimulating microbe inferred by whole-genome sequencing. *DNA Res* **18**(4): 291-303.

Lassmann T, Frings O, Sonnhammer EL. 2009. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic Acids Res* **37**(3): 858-865.

Low HH, Lowe J. 2006. A bacterial dynamin-like protein. *Nature* **444**(7120): 766-769.

Low HH, Sachse C, Amos LA, Lowe J. 2009. Structure of a bacterial dynamin-like protein lipid tube provides a mechanism for assembly and membrane curving. *Cell* **139**(7): 1342-1352.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**(5): 955-964.

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D. 2010. Tablet--next generation sequence assembly visualization. *Bioinformatics* **26**(3): 401-402.

Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO. 2007. Development of the human infant intestinal microbiota. *PLoS Biol* **5**(7): e177.

Podar M, Anderson I, Makarova KS, Elkins JG, Ivanova N, Wall MA, Lykidis A, Mavromatis K, Sun H, Hudson ME et al. 2008. A genomic analysis of the archaeal system Ignicoccus hospitalis-Nanoarchaeum equitans. *Genome Biol* **9**(11): R158.

Prakash T, Oshima K, Morita H, Fukuda S, Imaoka A, Kumar N, Sharma VK, Kim SW, Takahashi M, Saitou N et al. 2011. Complete genome sequences of rat and mouse segmented filamentous bacteria, a potent inducer of th17 cell differentiation. *Cell Host Microbe* **10**(3): 273-284.

Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW. 2009. Whole genome amplification and de novo assembly of single bacterial cells. *PLoS One* **4**(9): e6864.

Roller C, Wagner M, Amann R, Ludwig W, Schleifer KH. 1994. In situ probing of gram-positive bacteria with high DNA G + C content using 23S rRNA-targeted oligonucleotides. *Microbiology* **140 (Pt 10)**: 2849-2858.

Schatz MC, Phillippy AM, Shneiderman B, Salzberg SL. 2007. Hawkeye: an interactive visual analytics tool for genome assemblies. *Genome Biol* **8**(3): R34.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**(23): 7537-7541.

Sczesnak A, Segata N, Qin X, Gevers D, Petrosino JF, Huttenhower C, Littman DR, Ivanov, II. 2011. The genome of th17 cell-inducing segmented filamentous bacteria reveals extensive auxotrophy and adaptations to the intestinal environment. *Cell Host Microbe* **10**(3): 260-272.

Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, Thomson NR, Roberts AP, Cerdeno-Tarraga AM, Wang H et al. 2006. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. *Nat Genet* **38**(7): 779-786.

Simossis VA, Heringa J. 2005. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Res* **33**(Web Server issue): W289-294.

Studer R, Dahinden P, Wang WW, Auchli Y, Li XD, Dimroth P. 2007. Crystal structure of the carboxyltransferase domain of the oxaloacetate decarboxylase Na+ pump from Vibrio cholerae. *J Mol Biol* **367**(2): 547-557.

Umesaki Y, Okada Y, Matsumoto S, Imaoka A, Setoyama H. 1995. 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* **39**(8): 555-562.

Van Dongen S. 2008. Graph clustering via a discrete uncoupling process. *Siam Journal on Matrix Analysis and Applications* **30**(1): 121-141.

White RA, 3rd, Blainey PC, Fan HC, Quake SR. 2009. Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics* **10**: 116.