Supporting Information

Sela *et al.* 10.1073/pnas.0809584105

SI Text

Bacterial Strains, Growth Conditions, and Genomic DNA Extraction. *B. longum* subsp. *infantis* ATCC15697 was acquired from the American Type Culture Collection with strains JCM1272 and JCM11346 acquired from the Japan Collection of Microorganisms (Wako, Saitama, Japan). Cultures were propagated in de Man–Rogosa–Sharpe (MRS) broth (Becton Dickinson) at 37 °C under anaerobic conditions. All media were also supplemented with 1% (wt/vol) L-cysteine. Bacterial DNA was extracted by standard techniques for whole-genome shotgun sequencing or the MasterPure Gram-positive DNA Purification Kit (Epicentre) for Illumina sequencing.

HMO Growth Assay. Cell growth on semisynthetic MRS medium supplemented with 2% (wt/vol) HMO was monitored by optical density at 600 nm using a BioTek PowerWave 340 plate reader. The plate reader was run in discontinuous mode, with absorbance readings performed in 60 min intervals, and preceded by a 30 sec shaking at medium speed. Cultures were grown in biologically independent triplicates and the resulting growth data were expressed as the mean of these replicates.

Genome Sequencing and Assembly. Three whole genome shotgun libraries were constructed for end sequencing. A 3 kb insert library was constructed by randomly shearing 3 μg of genomic DNA by using a Hydroshear (GeneMachines) and end repairing by using T4 DNA polymerase and Klenow fragment (New England Biolabs). End repaired fragments were size selected from an agarose gel and purified by using the QIAquick Gel Extraction Kit (Qiagen). Approximately 200 ng of sheared DNA was ligated into 100 ng of linearized, dephosphyorylated pUC18 vector (Roche) at 24.5 °C for 90 min using the Fast Link™ DNA ligation Kit (Epicentre). A portion of the ligation product was electroporated into ElectroMAXDH10B cells (Invitrogen), and plated on agar plates containing ampicillin. White colonies were picked by using the robotic QPIX (Genetix)colony picker and grown in media containing glycerol to provide stocks for subsequent sequencing. Eight kb libraries were constructed similarly to the 3 kb, except 10 μg of DNA was sheared and agarose size selection was in the 6–8 kb range. Purified fragments were ligated into pMCL200 with kanamycin resistance. Fosmid libraries were constructed by using the CopyControl™ Fosmid Production Kit (Epicentre). DNA was sheared and fragments size selected on an agarose pulsed-field gel, excised and purified before ligation in the pCCqFos vector. The ligated vector was packaged by using MaxPlax™ Lambda Packaging Extract (Epicentre) and used to transduce *E. coli* (EP300).

DNA for sequencing was produced by using Templiphi™ (GE Healthcare) on aliquots of subclones grown in 384-well plates according to product specifications. Briefly, cells placed in denaturation buffer were heated to 95 °C for 5 min and cooled to 4 °C before the addition of Templiphi reagent. Subsequently, plates were incubated for 18 h. at 30 °C and aliquots were removed for sequencing. Standard cycle sequencing from both ends of the subclones by using universal primers was performed with BigDye Terminators (Applied Biosystems) and resolved on ABI Prism 3730xl capillary sequencers. Detailed library construction and sequencing protocols can be found online: http://www.jgi.doe.gov/sequencing/protocols/index.html.

A total of 48,830 end reads were attempted for the initial draft assembly. Poor quality ($\leq$100 bases) and vector reads were removed before assembly with PHRAP, PGA and Arachne (1) assemblers. Initial assemblies produced between 32 and 49 major contigs ($>$ 2kb, $\geq$ 10 reads each). Finishing was carried out by using CONSED (2). A total of 1,198 finishing reads were performed to solve mis-assemblies, close gaps and improve quality resulting in a circular genome of 2,832,748 bp.

Sequence Analysis and Annotation. Critica (v1.05) and Glimmer (v3.1) were run to model genes by using default settings that permit overlapping genes and using ATG, GTG, and TTG as potential start codons. The results were combined, and a BLASTP search of the translations vs. GenBank's nonredundant database (NR) was conducted. The alignment of the N terminus of each gene model vs. the best NR match was used to pick a preferred gene model. If no BLAST match was returned, the longest model was retained. Gene models that overlapped by greater than 10% of their length were flagged for revision or deletion, giving preference to genes with a BLAST match. The revised gene/protein set was searched against the Swiss-Prot/TrEMBL, PRIAM, protein family (Pfam), TIGRFam, Interpro, KEGG, and COGs databases, in addition to BLASTP vs. NR. From these results, product assignments were made. Initial criteria for automated functional assignment set priority based on PRIAM, TIGRFam, Pfam, Interpro profiles, pairwise BLAST vs. Swiss-Prot/TrEMBL, KEGG, and COG groups. Manual corrections to automated functional assignments were completed on an individual gene-by-gene basis as needed.

Bioinformatic Analyses. The *B. longum* subsp. *infantis* ATCC15697 genome sequence was analyzed in the Integrated Microbial Genomes data management, analysis, and annotation platform (3). IS element and prophage sequences were identified with IS finder (http://www-is.biotoul.fr/) and Prophage finder (4) respectively. Deviation from mean dinucleotide relative abundance (ρ^*XY) was examined in 10 kb windows (5). Potential genomic islands were identified by examining codon usage patterns with the SIGI-HMM program (6). Variances in GC content was profiled by the DNA segmentation algorithm hosted at http://tubic.tju.edu.cn/GC-Profile/ (7). Whole genome comparisons of *B. longum* subsp. *infantis* ATCC15697 were performed with BLASTN pairwise alignments (Expect threshold = 10) against *B. longum* subsp. *longum* NCC2705 and *B adolescentis* ATCC15703. The results of these comparisons were visualized by using the Artemis Comparison Tool (8).

The evolutionary history of the SBP proteins was inferred using the Neighbor-Joining method (9). The bootstrapping of 500 replicates was used to determine the statistical confidence of the phylogenetic relationships (10). Branches having bootstrap node support less than 50% were collapsed. The evolutionary distances were computed using the Poisson correction method with the units reflecting the number of amino acid substitutions per site. Amino acid positions containing gaps or missing data were eliminated from the dataset. There were a total of 192 positions examined in the final dataset. Phylogenetic analyses were conducted in the MEGA4 suite of bioinformatic tools (11).

The phylogeny of the five additional sequenced HMO$^+$ *B. longum* isolates was inferred from concatenation of *clpC*, *dnaB*, *dnaG*, *dnaJ1*, *purF*, *rpoC*, and *xfp* gene sequences as described in (12). Identical tree topology was obtained by using parsimony, tree-puzzle, and neighbor-joining methods as implemented in the PAUP phylogeny package. Nucleotide sequences were obtained from sequenced strains (i.e., UCD44, UCD49, JCM1272, JCM7010, and JCM11346) by consensus generation from

BLASTN alignments of Illumina reads to the *B. longum* subsp. *infantis* ATCC 15697 reference using the perl scripts blastView3.pl and bv3ToColumns.pl (J.C., unpublished work). GenBank nucleotide sequences used to generate the tree include *B. longum* subsp. *longum* DJO10A, *B. suis* LMG 21814, *B. breve* UCC2003, and *B. animalis* 25527. Branch lengths are in the same units (number of nucleotide substitutions per site) as those of the evolutionary distances used to construct the tree. Bootstrap values were computed by resampling 10,000 times.

Illumina Sequencing-by-Synthesis of Related Strains. Genomic DNA libraries were prepared for Illumina sequencing according to the manufacturer's specifications. Briefly, DNA was fragmented by nebulization, end repaired, ligated with Illumina specific adaptors, and size selected on an agarose gel. Purified fragments (approximaltey 130 bp) were first amplified for 18 cycles and then quantitated with a Bioanalyzer (Agilent Technologies), diluted, and loaded onto the Illumina cluster generation station to prepare the flow cell for sequencing. Forty-one cycles of single base addition/detection were performed resulting in 40 base read lengths. Reads passing noise filters were used for BLAST analysis (blastall -W 11 q -1 -F 'm D' -b 10 -v 10 -K 10 -e 1e-4 -p blastn).

***in Vivo* HMO Metabolism.** Bifidobacteria were inoculated into hexane-washed RGM-EE medium containing 0.2% fructose and grown anaerobically at 37 °C (13). Aliquots (1.5 ml) were centrifuged with washed pellets resuspended in RGM-EE containing 0.2% of the ^{13}C-labeled carbon source (e.g., [U-^{13}C]fructose) to OD600 = 0.110, and incubated at 37 °C. Isotopically-labeled sugars and metabolites were obtained from Omicron Biochemicals, Inc. Aliquots (1.5 ml) were harvested at 0 min, 30 min, 1 h, 4 h, and 24 h time points, pelleted at 4 °C, washed (1X TE buffer), and resuspended in modified single-phase Bligh-Dyer solvent (chloroform:methanol:water, 1:2:0.8 by volume).

For the metabolic chase experiments the ^{13}C-labeled cultures were grown at 37 °C to the four hour time point (final OD600 = 0. 350). Cells were centrifuged anaerobically at room temperature and washed (1X RGM-EE), suspended in 20 ml RGM-EE (OD600 = 0.175) and divided into 3 × 6 ml aliquots. 1 ml from each tube was removed to represent the zero time point, pelleted, and washed with TE. 100 μl of a 0.2% solution of the unlabeled chase sugars were added to the respective aliquot tubes, and incubated at 37 °C. Samples were harvested (1.5 ml) at 30 min, 1 h, 4 h, and 24 h, pelleted, and washed (TE buffer) before Bligh-Dyer fractionation and analysis.

Analytes were derivatized by suspending cells in modified single-phase Bligh-Dyer solvent, converted to two-phase by addition of one volume each of chloroform and water, and centrifuged (2000 × *g*) to partition (14, 15). The aqueous phases was recovered and allowed to evaporate until dry. The carbohydrate-containing residues were hydrolyzed with 2M (TFA), and aldononitrile acetate derivatives were prepared as described previously (16). The Bligh-Dyer phase, containing lipophilic metabolites, was evaporated and fatty acid methyl esters and pyrrolidide derivatives prepared (17). The ^{13}C/^{12}C isotopic ratios were determined from the GC-EI-MS spectra of the derivatized fatty acid and monosaccharide components using selected isotopomer ions as previously described (16).

Ethyl acetate extracts containing sugar aldonitrile acetates, or fatty acid esters, or pyrrolidides were analyzed by GC-MS on an Agilent 6890N GC system coupled with an Agilent 5973 mass selective detector. An automated injector (Agilent 7683 series) was used to introduce the samples, and ionization was by electron impact (e.i.) mode with positive ion detection. Ion extraction full scale scans were collected over the mass range 50–500 mass units.

Proteomics Sample Preparation and Protein Digestion. *B. longum* subsp. *infantis* ATCC15697 cells were grown in modified MRS + 2% HMO as previously described (18). Cells were harvested during late exponential phase and normalized to 10 ml of OD600 of 1.0 by dilution. After centrifugation, the cell pellet was washed three times with PBS and resuspended in 1 ml of lysis buffer (pH 9.0) containing 100 mM of Tris and 8M of urea. The cells were disrupted with a bead beater (FastPrep; QBiogen) with 300 μg of silica beads (Sigma–Aldrich) for 6 × 30 sec pulses. Between each cycle, samples were cooled on ice. Beads and cell debris were removed by centrifugation and the soluble fraction was stored at −80 °C. The protein concentration was measured by the Bio-Rad protein assay kit (BioRad) according to the manufacturer's procedure.

For reduction of whole cell protein, 4 μl of 450 mM DTT (Sigma–Aldrich) was added to 200 μg of the cell-free extract and incubated for 45 min at 55 °C. Without alkylation, the reduced proteins were digested by 2.5 μg of mass spectrometry grade trypsin (Promega) overnight at 37 °C. The tryptic peptides were purified by C18 Ziptip (Millipore) according to the manufacture's protocol. The Ziptip was prepared by washing with 50% acetonitrile (ACN)/H_2O followed by 0.1% (vol/vol) TFA in H_2O. The tryptic peptide solution was subsequently loaded onto the Ziptip and washed with 0.1% (vol/vol) TFA in H_2O. The peptides were eluted with 50% ACN in H_2O. The purified sample was dried before mass spectrometry analysis.

Multidimensional Protein Identification Technology (MudPIT). Peptides were directly loaded on a Michrom PolySulfoethyl Aspartamide SCX-enrichment microtrap using a CTC PAL autosampler. Peptides were sequentially eluted off the SCX trap onto an Agilent ZORBAX 300SB C_18, reverse phase trap cartridge using 8 salt injections of ammonium formate (0 mM, 10 mM, 20 mM, 50 mM, 100 mM, 200 mM, 500 mM, 800 mM). After each salt injection the Agilent C_18 trap was switched in-line with a Michrom Magic C_18 AQ 100 μm × 150 mm C_18 column connected to a Thermo-Finnigan LTQ iontrap mass spectrometer through a Michrom Advance Plug and Play nano-spray source. Peptides were separated by using a gradient of 0–40% B (A = 0.1% formic acid, B = 100% acetonitrile) in 20 min (30 min total run time). MS and MS/MS spectra were acquired by using a top 10 method, where the top 10 ions in the MS scan were subjected to automated low energy CID. A total of eight separate LC-MS/MS runs were acquired for each sample for a total run time of four hours.

Peptide Database Analysis. Tandem mass spectra were extracted and charge state was deconvoluted by BioWorks version 3.3. Deisotoping was not performed. All MS/MS samples were analyzed by using X! Tandem (www.thegpm.org; version 2008.01.01.1). X! Tandem was set up to query a database constructed from the *B. longum* subsp. *infantis* ATCC15697 predicted proteome and assuming complete tryptic digestion. X! Tandem was searched with a fragment ion mass tolerance of 0.40 Da and a parent ion tolerance of 1.8 Da. No fixed modification was specified in X! Tandem. Deamidation of asparagine and glutamine, oxidation of methionine and tryptophan, sulfone of methionine, tryptophan oxidation to formylkynurenin of tryptophan and acetylation of the N terminus were specified.

Protein Identification. Scaffold (version 2.00.03, Proteome Software Inc.) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they exceeded a -Log(E) score greater than 2.0 by the X! Tandem. Protein identifications were accepted if they could be established at greater than 99.0% probability and contained at least 2 identified peptides. Protein probabilities were assigned

by the Protein Prophet algorithm (19). Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principle of parsimony.

1. Jaffe DB, *et al.* (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 13:91–96.
2. Gordon D, Desmarais C, Green P (2001) Automated finishing with autofinish. *Genome Res* 11:614–625.
3. Markowitz VM, *et al.* (2008) The integrated microbial genomes (IMG) system in 2007: Data content and analysis tool extensions. *Nucleic Acids Res* 36:D528–D533.
4. Bose M, Barber RD (2006) Prophage Finder: A prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol* 6:223–227.
5. van Passel MW, Luyf AC, van Kampen AH, Bart A, van der Ende A (2005) Deltarhoweb, an online tool to assess composition similarity of individual nucleic acid sequences. *Bioinformatics* 21:3053–3055.
6. Waack S, *et al.* (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics* 7:142–153.
7. Gao F & Zhang CT (2006) GC-Profile: A web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucleic Acids Res* 34:W686–W691.
8. Carver TJ, *et al.* (2005) ACT: The Artemis Comparison Tool. *Bioinformatics* 21:3422–3423.
9. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425.
10. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783–791.
11. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24:1596–1599 .
12. Ventura M, *et al.* (2006) Analysis of bifidobacterial evolution using a multilocus approach. *Int J Syst Evol Microbiol* 56:2783–2792.
13. Hespell RB, Wolf R, Bothast RJ (1987) Fermentation of xylans by Butyrivibrio fibrisolvens and other ruminal bacteria. *Appl Environ Microbiol* 53:2849–2853.
14. Bligh EG, Dyer WJ (1959) A rapid method of total lipid extraction and purification. *Can J Biochem Physiol* 37:911–917.
15. Ray BL, Painter G, Raetz CR (1984) The biosynthesis of gram-negative endotoxin. Formation of lipid A disaccharides from monosaccharide precursors in extracts of Escherichia coli. *J Biol Chem* 259:4852–4859.
16. Price NP (2004) Acylic sugar derivatives for GC/MS analysis of 13C-enrichment during carbohydrate metabolism. *Anal Chem* 76:6566–6574.
17. Christie WW (2007) The chromatographic analysis of lipids. *The Lipid Handbook* eds Gunstone FD, Harwood JL, Dijkstra A (Chapman and Hall, London), 3rd Ed, p 426455.
18. Locascio RG, *et al.* (2007) Glycoprofiling of bifidobacterial consumption of human milk oligosaccharides demonstrates strain specific, preferential consumption of small chain glycans secreted in early human lactation. *J Agric Food Chem* 55:8914–8919.
19. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 75:4646–4658.

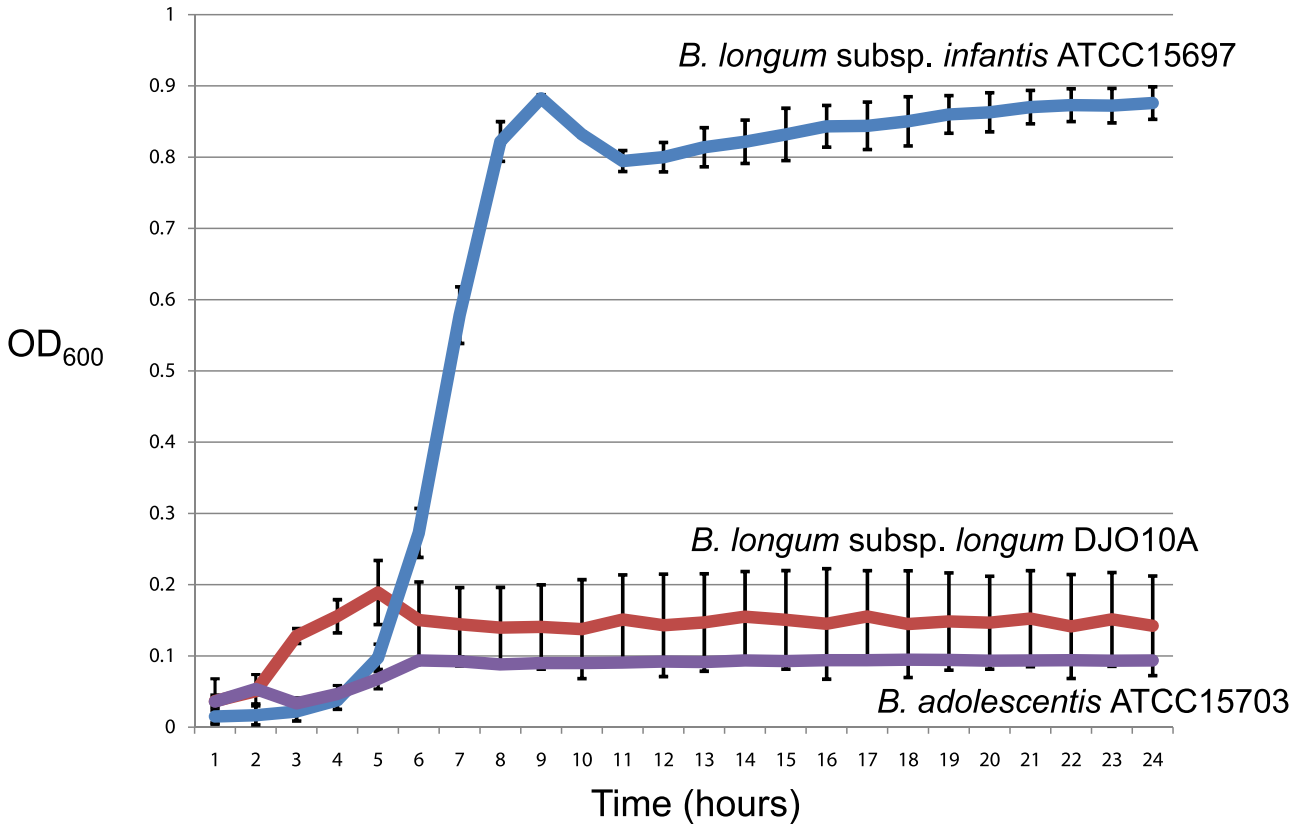

Fig. S1. Growth curves of sequenced bifidobacteria using HMO as a sole carbon source. *B. longum* subsp. *infantis* ATCC15697, *B. longum* subsp. *longum* DJO10A, and *B. adolescentis* ATCC15703 were grown on a semisynthetic MRS medium supplemented with 2% (wt/vol) HMO. Growth was measured as optical density of the medium at 600 nm. The growth methods used were previously described in ref. 18. Fermentations were performed in triplicate with error bars representing the standard deviations of the mean.

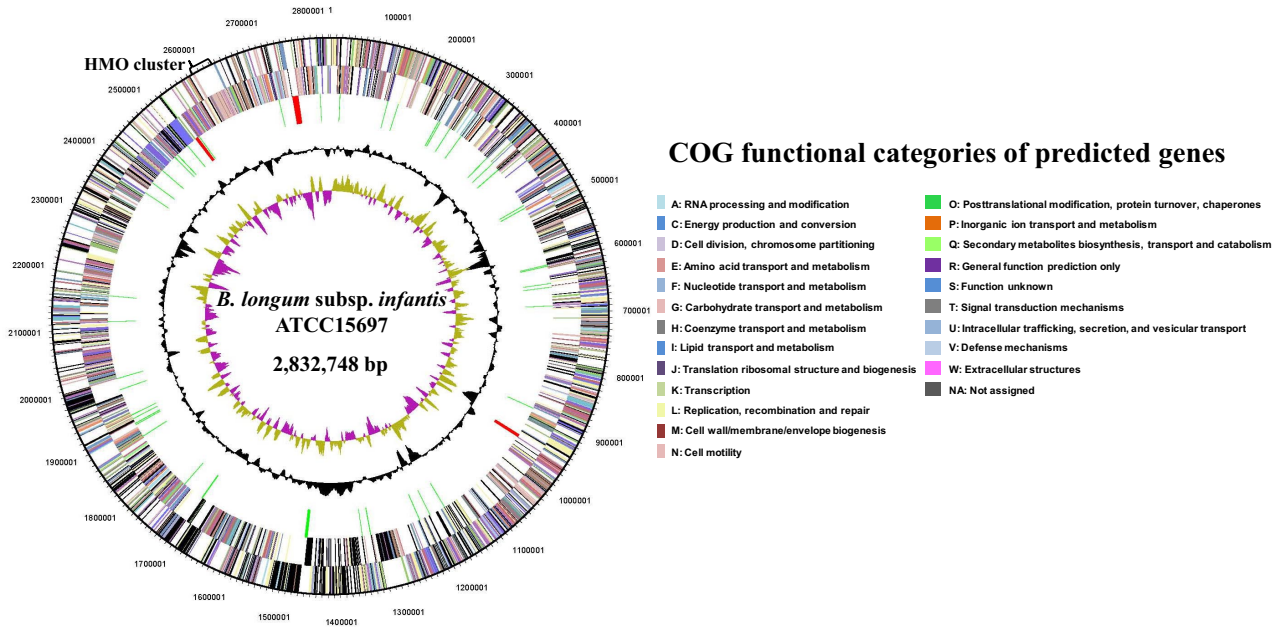

Fig. S2. Circular genome map of *B. longum* subsp. *infants* ATCC15697. Proceeding from the outer circle inward, circle (a) denotes genome coordinates in base pairs. Circle (b) indicates coding regions by strand with color corresponding to the COG functional assignment. Circle (c) indicates RNA genes with rRNAs represented in red and tRNA in green. Circle (d) depicts G+C % deviation from the mean (59.86%). Circle (e) illustrates GC skew (G–C/G+C), values >0 are in gold and <0 in purple.

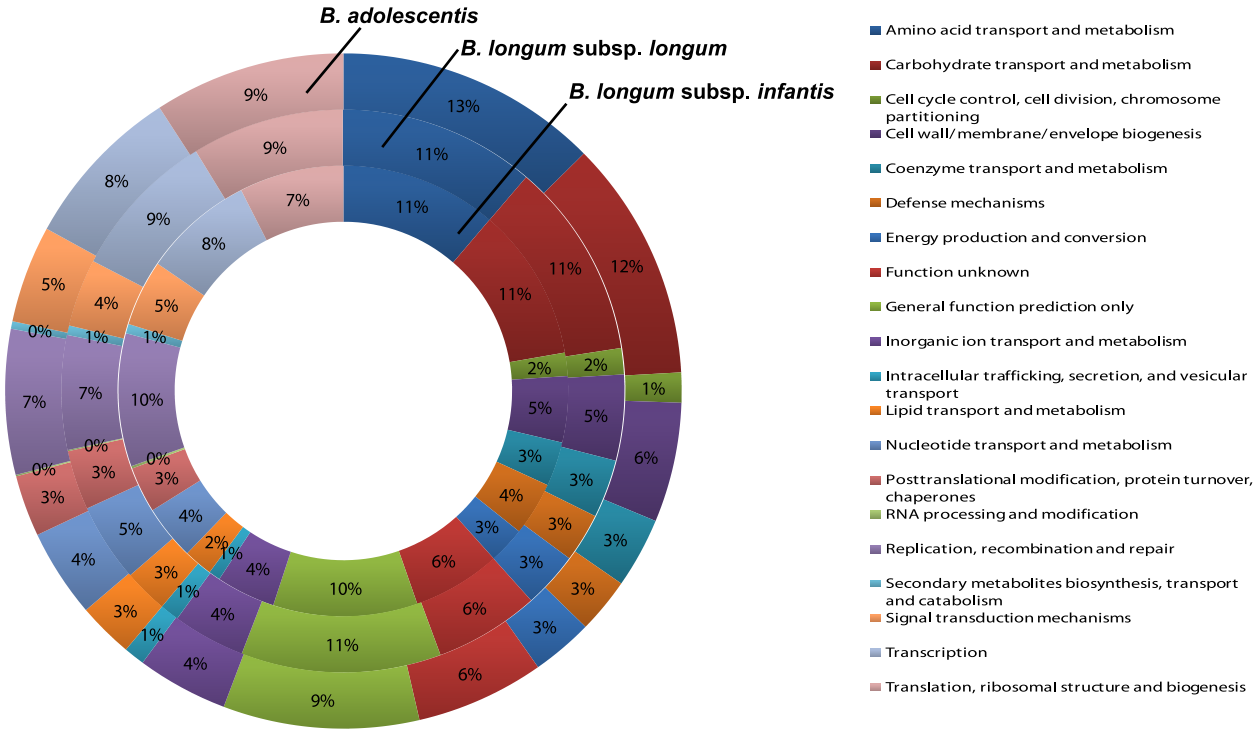

Fig. S3. Comparison of COG functional categories between fully sequenced bifidobacterial genomes. Each colored segment represents the relative contribution of a functional category as a percentage of total COGs. The inner ring depicts *B. longum* subsp. *infantis* ATCC15697; the middle ring *B. longum* subsp. *longum* NCC2705; and the outer ring *B. adolescentis* ATCC1703.

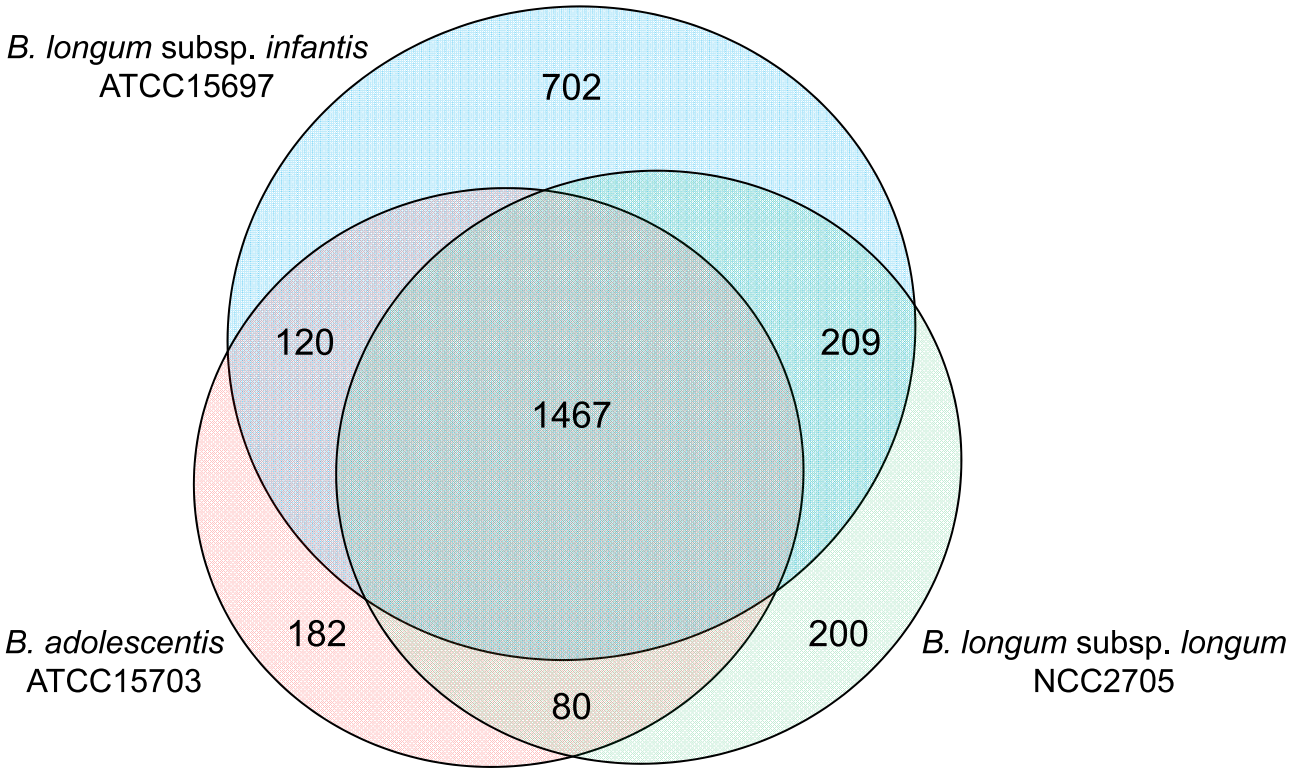

Fig. S4. Venn diagram of homologs shared between sequenced bifidobacterial genomes. Surface areas are proportional to members contained in each set. Shared homologs were identified by BLASTP pairwise alignments (E <1e-2; min identity 20%). Differences in sets due to paralogies were not considered and set values are expressed as the number of ATCC15697 homologs. The exception being the set of shared homologs between ATCC15703 and NCC2705, to the exclusion of ATCC15697, is expressed as homlogs contained in the NCC2705 genome.

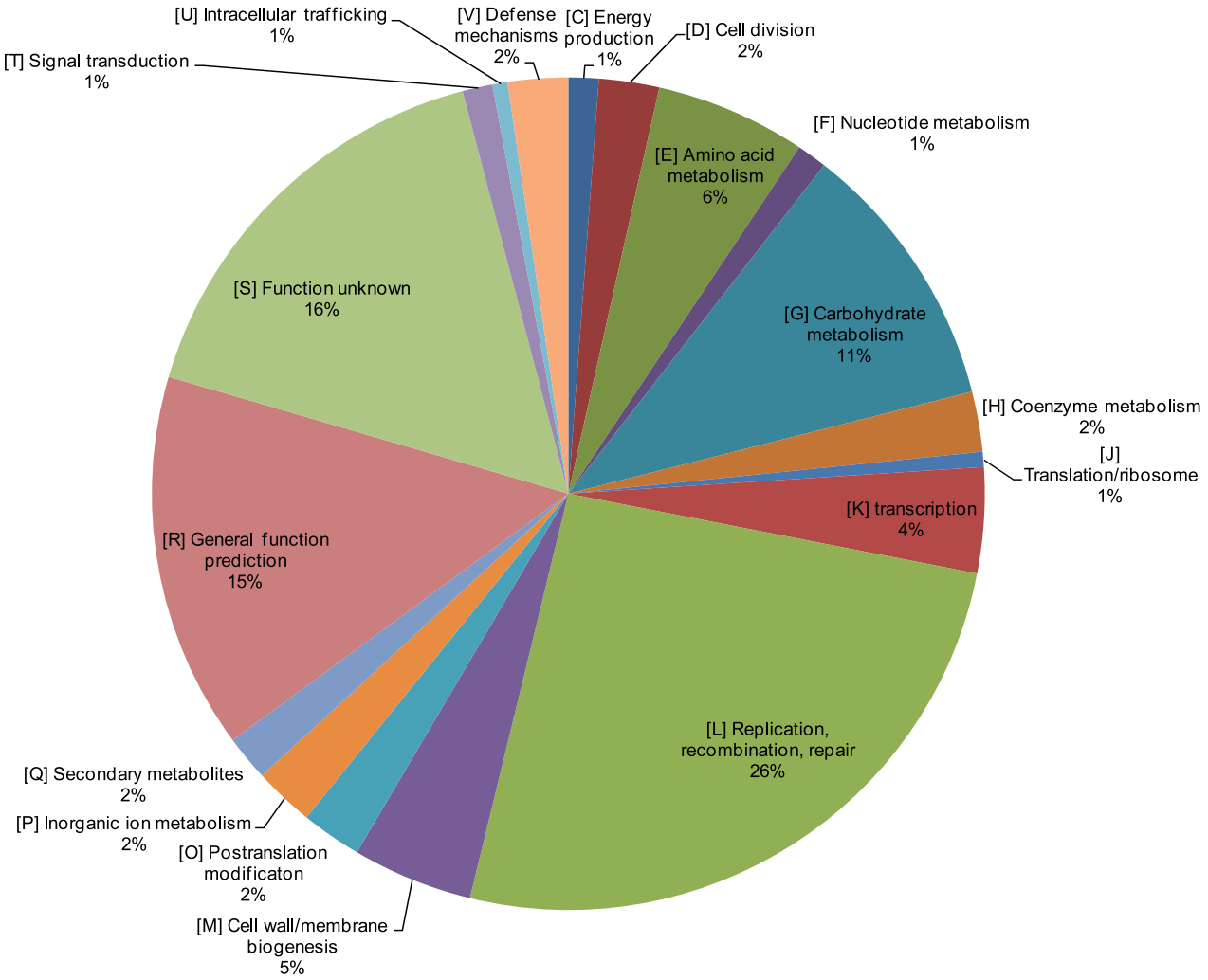

Fig. S5. Functional distribution of genes unique to ATCC15697. 171 genes (of 702) identified as unique to ATCC15697 when compared to NCC2705 and ATCC15703 can be assigned to at lease one COG.

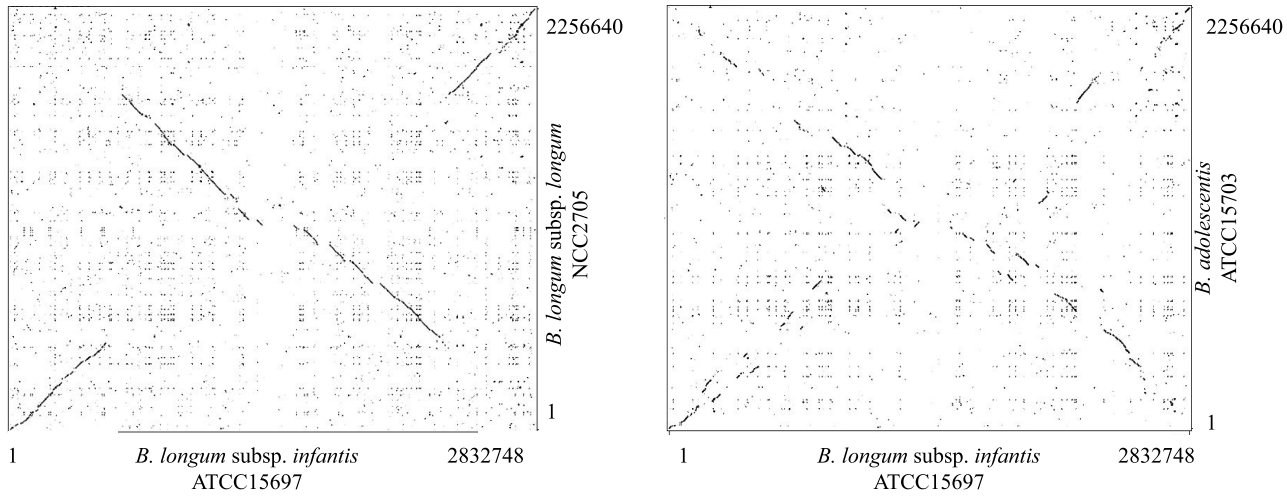

Fig. S6. Genome colinearity of *B. longum* subsp. *infantis* ATCC15697 with *B. longum* subsp. *longum* NCC2705 and *B. adolescentis* ATCC15703. The dot plots were generated by using PipMaker percent identity plotting software (http://pipmaker.bx.psu.edu/pipmaker/).

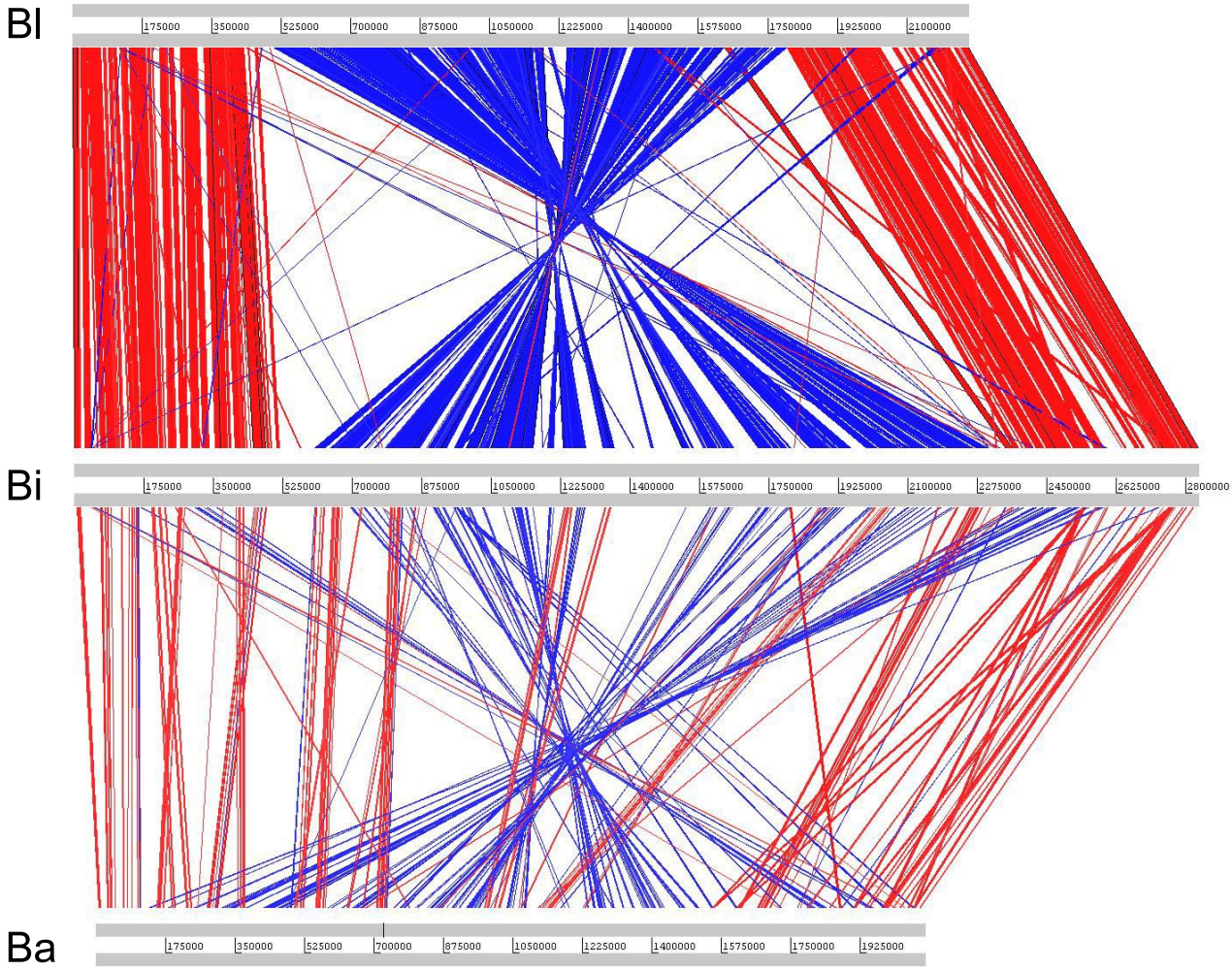

Fig. S7. Chromosome alignment of ATCC15697 (Bi) with NCC2705 (Bl), and ATCC15703 (Ba). The diagram depicts forward and reverse DNA strands (gray bars) with accompanying base coordinates. Similar regions that are more than 700 bases in length are indicated by red (colinear) and blue (inverted) lines.

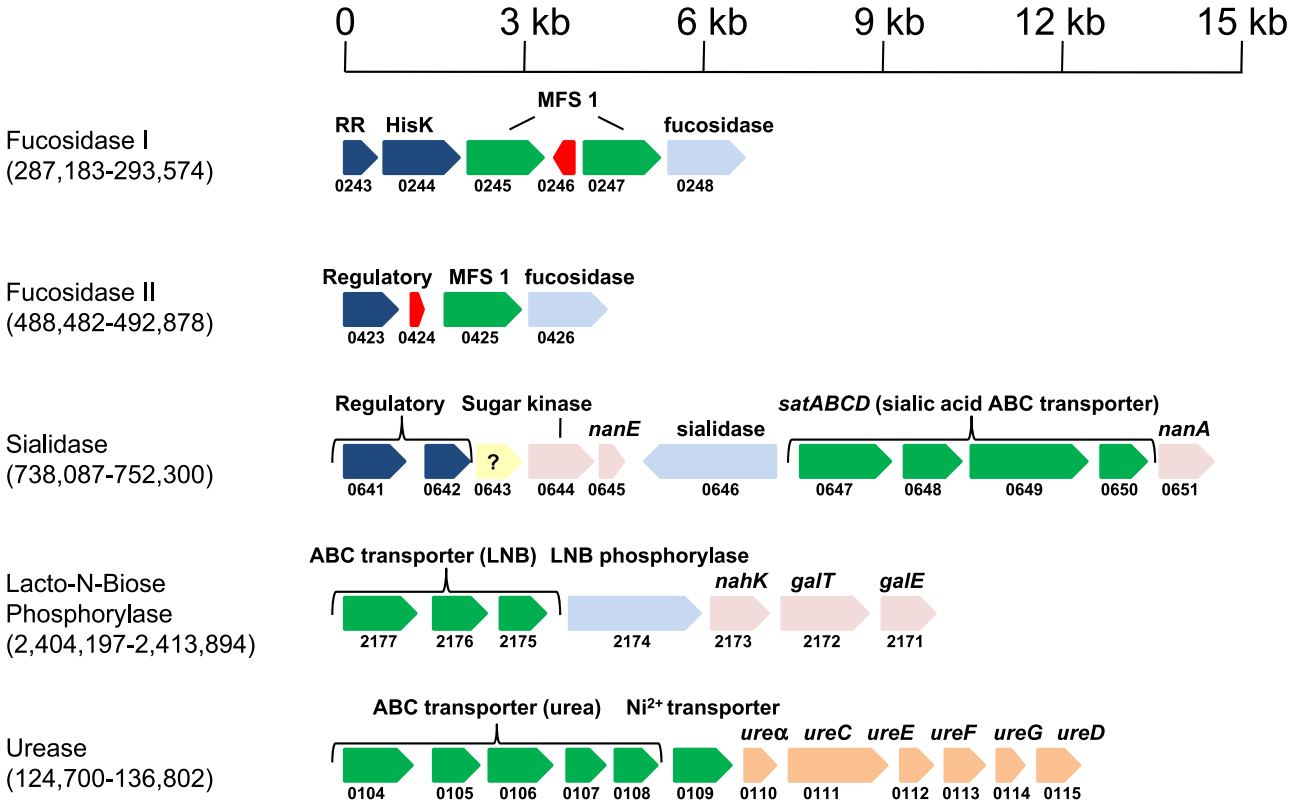

Fig. S8. Additional milk-related gene clusters. Genomic coordinates are given in bp in parentheses. Locus tags are listed underneath genes and are preceded by 'Blon_'. Genes are represented by arrows and are colored according to predicted function. Dark blue is related to transcriptional regulation; light blue represents glycosyl hydrolases; green are transport-related; red arrows denote gene fragments; rose represents carbohydrate related genes; whereas orange genes are urease-related ORFs. RR, response regulator of a two-component system; HisK, histidine (sensor kinase); MFS 1, major facilitator superfamily 1 (sugar) permease; ABC, ATP binding cassette; LNB, Lacto-N-Biose.

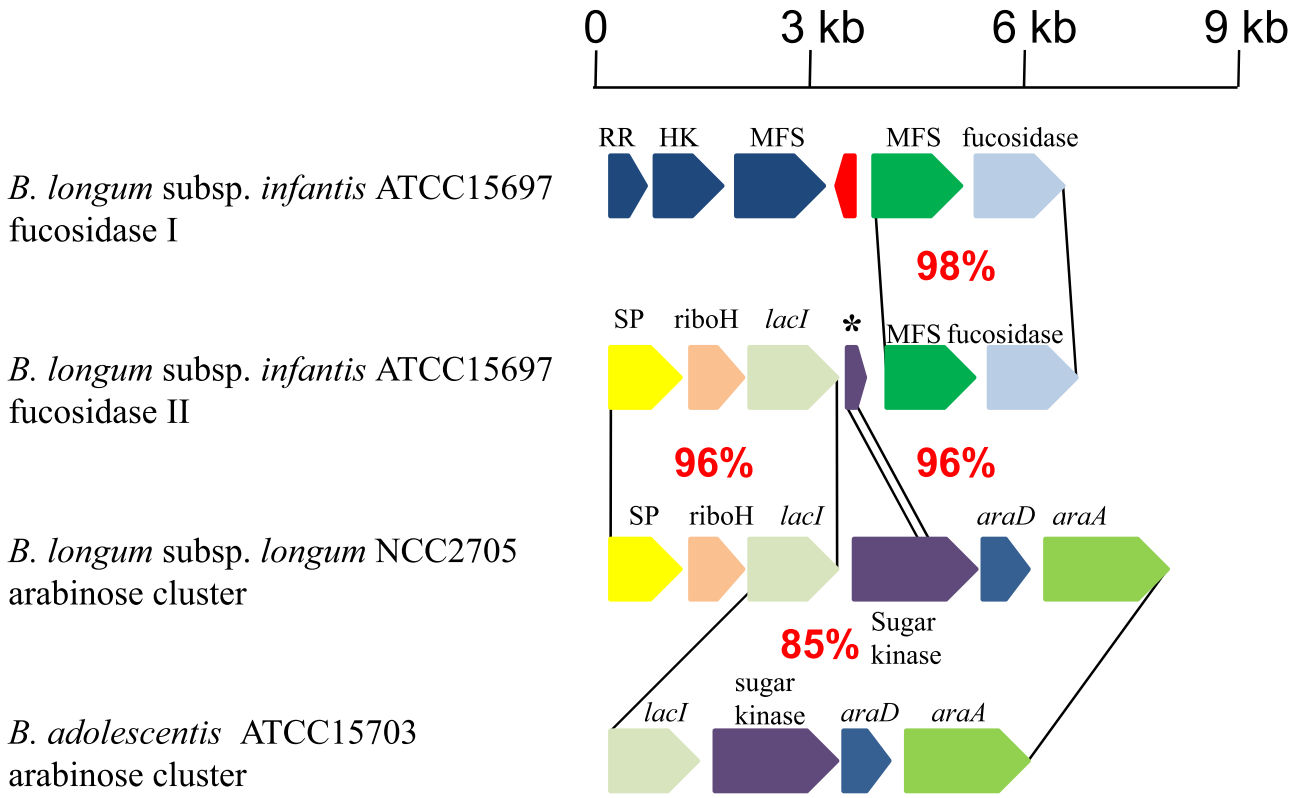

Fig. S9. Nonorthologous displacement of arabinose genes by fucosidase in *B. longum* subsp. *infantis* ATCC15697. Genes are represented by arrows and are colored according to predicted function and homology to other bifidobacterial sequences. High sequence identity is depicted by solid black lines between gene clusters. RR, response regulator of a two-component system; HisK, histidine (sensor kinase); MFS, major facilitator superfamily (sugar) permease; SP, signal peptidase; RiboH, ribonuclease H; *lacI*, lacI-like regulator; *araD*, L-ribulose-5-phosphate 4- epimerase; *araA*, arabinose isomerase. The sugar kinase in *B. longum* subsp. *longum* and *B. adolescentis* may function as a L-ribulokinase (*araB*). * denotes sugar kinase fragment. Nucleotide identities are indicated between homologous regions.

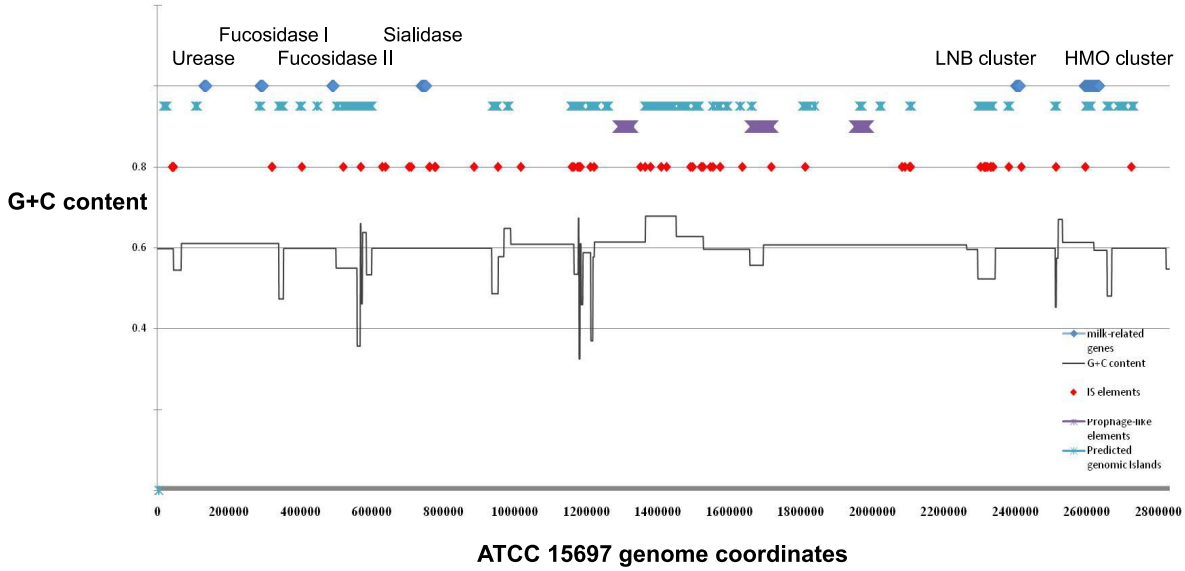

Fig. S10. Mobile genetic elements of the ATCC15697 genome. IS elements and predicted prophages are labeled in red and purple respectively. G+C content (mean 59.9%) is represented as a black line. Potential genomic islands based on irregular codon usage is denoted in turquoise.

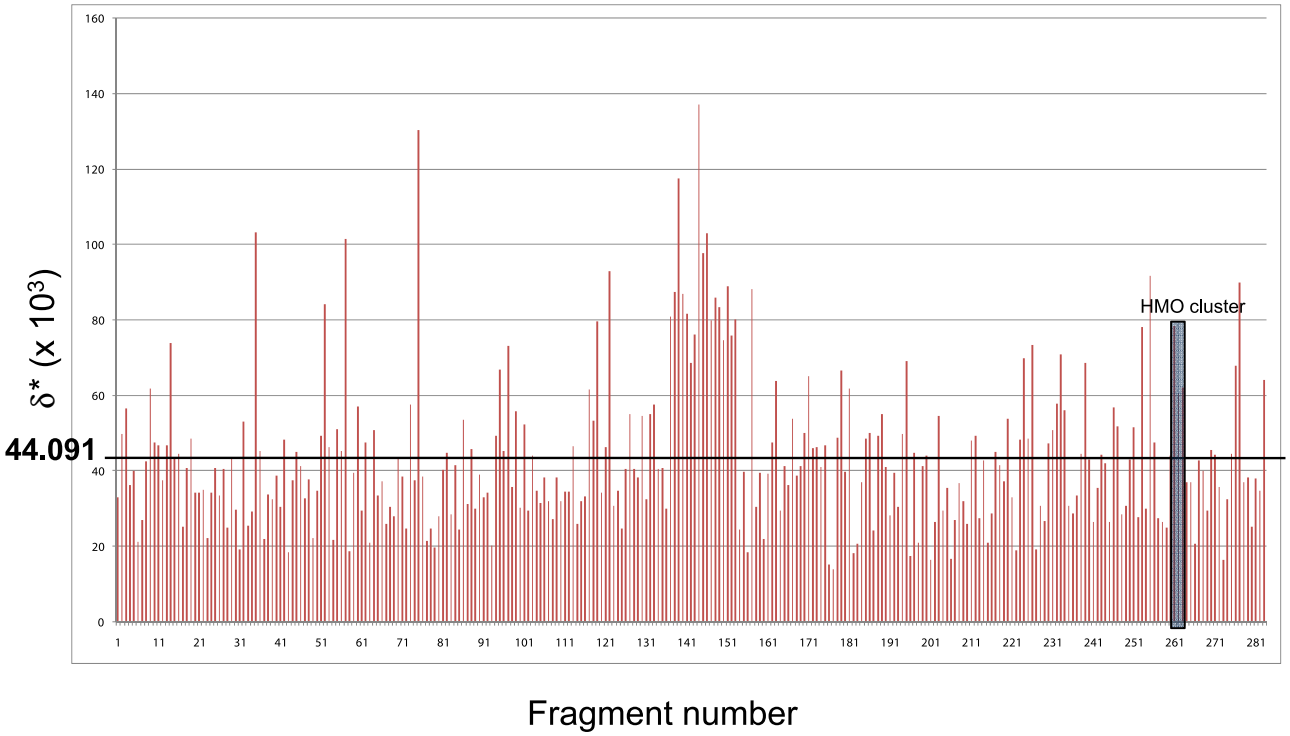

Fig. S11. The average dinucleotide relative abundance (genome signature) difference within the ATCC 15697 genome. The nonoverlapping window size was set at 10 kbp. The mean intragenomic δ* value was calculated at 44.091 for *B. longum* subsp. *infantis.*

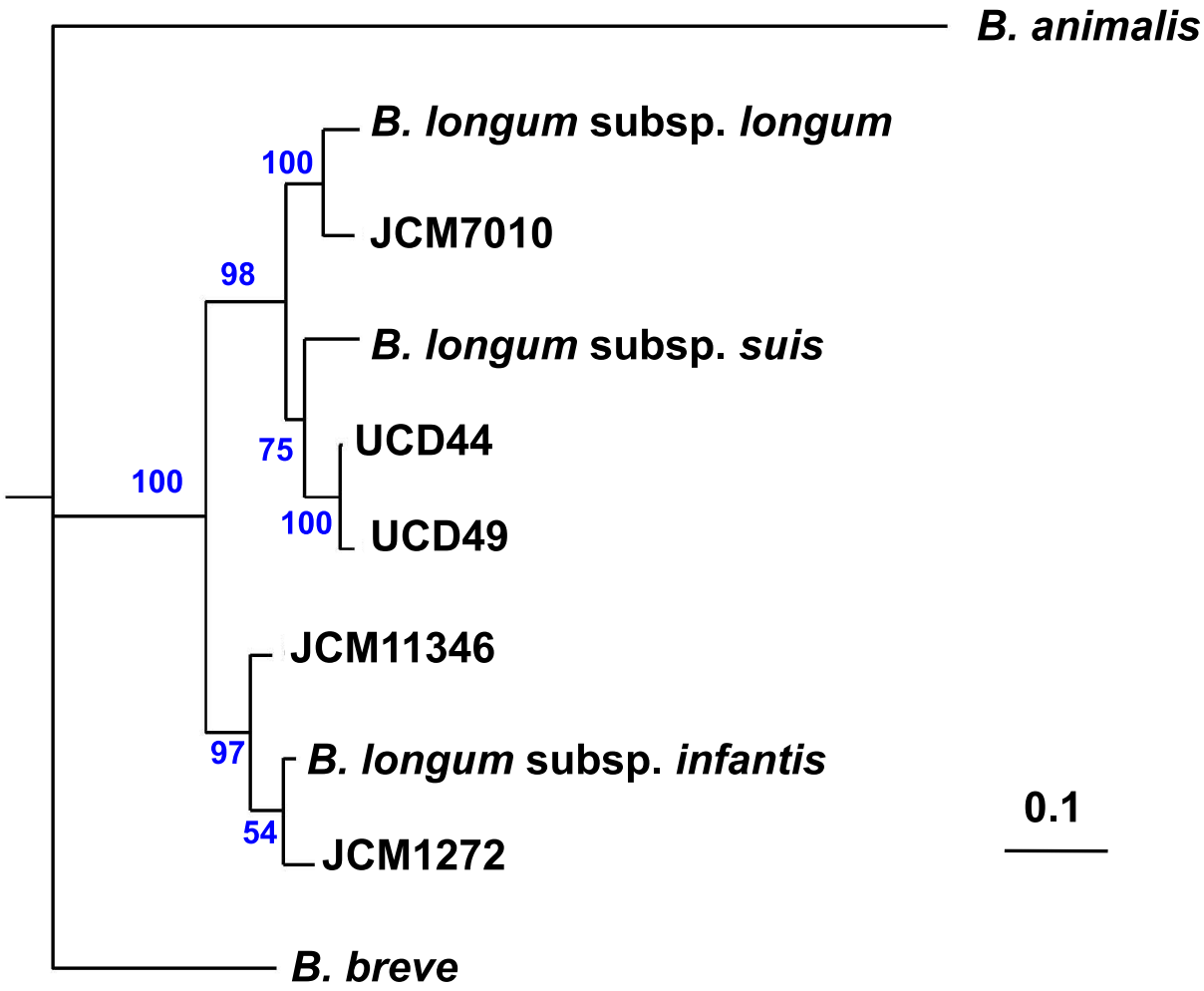

Fig. S12. Phylogenetic distribution of the five additional sequenced HMO$^+$ *B. longum* isolates.

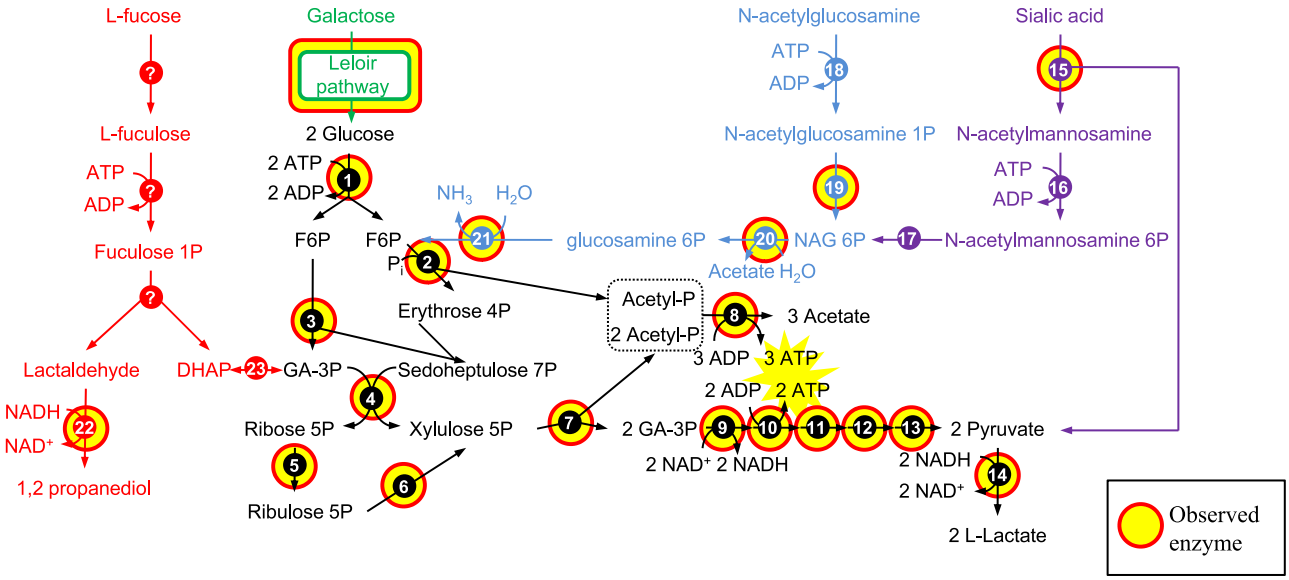

Fig. S13. Proteomic profile of HMO metabolic pathways in *B. longum* subsp. *infantis* ATCC15697. Genes involved in HMO catabolism are listed in Table S7 online. Proteins that were identified by using Multidimensional Protein Identification Technology (MudPIT) are highlighted in burgundy and gold. Note that GalT (Blon_2172) in the modified Leloir pathway did not exceed the threshold for detection.

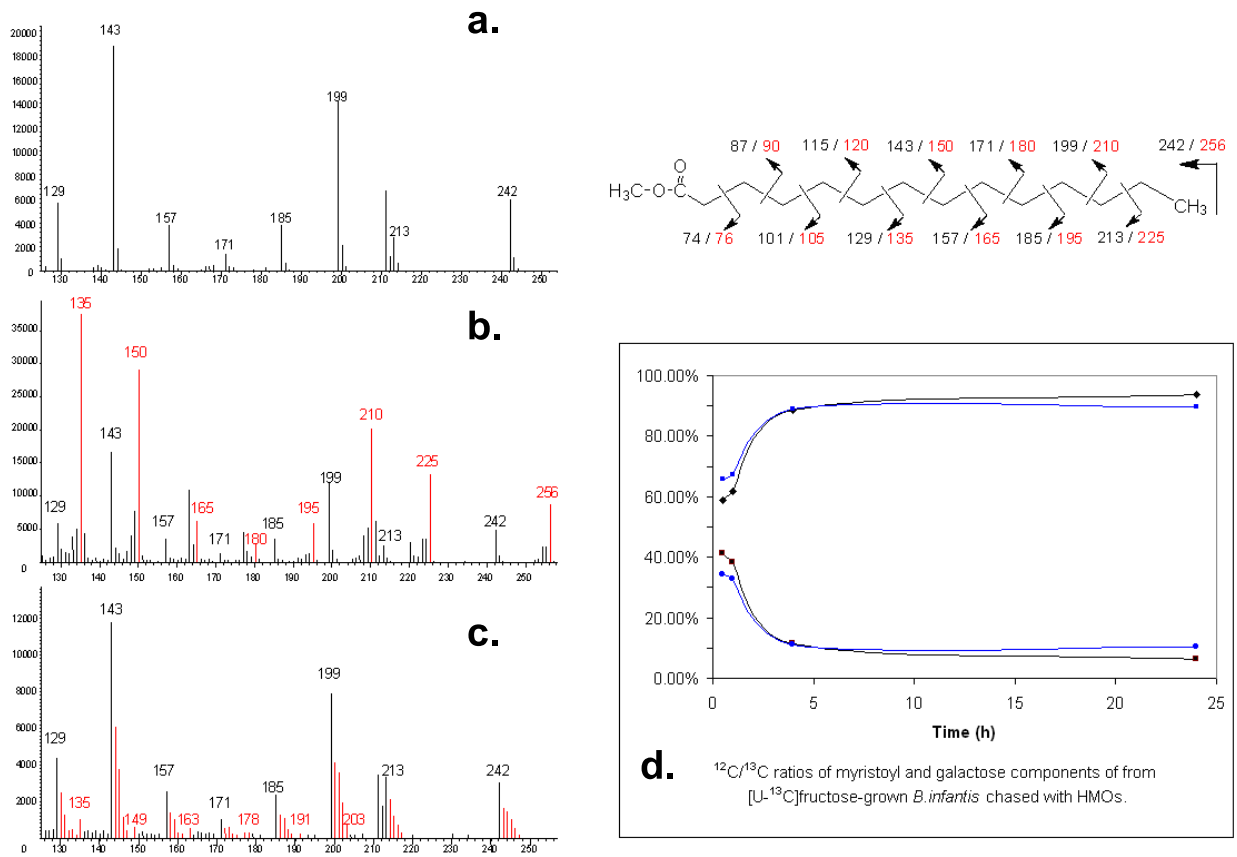

Fig. S14. Functional analysis of HMO and sialic acid metabolism in *B. longum* subsp. *infantis* ATCC15697. Mass spectrometric (GC-EI-MS) analysis of the incorporation of ^{13}C into myristate fatty acid methyl ester by *B. longum* subsp. *infantis* cultured for 24 h on unlabeled fructose (*A*); universally-labeled [U-^{13}C]fructose (*B*); or *N*-acetyl-[2-^{13}C]neuraminic acid (*C*) as sole sources of carbon. Peaks due to monoisotopic ions are black, and those due to carbon isotopomers are shown in red. The fragmentation ions arising from C-C bond cleavage are assigned on the chemical structure (top right). (*D*) The time course for a chase out of ^{13}C from [U-^{13}C]fructose-labeled cells by a pulse of unlabeled HMOs administered at time 0. The $^{12}C/^{13}C$ ratios for myristate (black lines) and galactose (blue lines) components are shown as representative fatty acids and sugars, respectively.

Other Supporting Information Files

Table S1 (PDF)
Table S2 (PDF)
Table S3 (PDF)
Table S4 (PDF)
Table S5 (PDF)
Table S6 (PDF)
Table S7 (PDF)
Table S8 (PDF)