Supporting Information

Jones et al. 10.1073/pnas.1519288112

SI Materials and Methods

DNA Extraction. A 3-mL volume of lysis buffer [20 mM Tris·HCl (pH 8.0), 2 mM sodium EDTA, 1.2% Triton X-100] was added to 0.5 g of stool sample and was vortexed until homogenous. A 1.2-mL volume of homogenized sample and 15 μL of Proteinase K (P2308; Sigma Aldrich) enzyme was aliquoted to a 1.5-mL tube with garnet beads (12830-50-BT; Mo Bio). Bead tubes were incubated at 65 °C for 10 min and then 95 °C for 10 min. Tubes were placed in a Vortex Genie 2 (Scientific Industries) to perform bead beating for 13 min, and the sample subsequently was spun in an Eppendorf Centrifuge 5424. Supernatant (800 μL) was transferred to a deep well block. DNA extraction and purification were performed using a Chemagic MSM I (Perkin-Elmer) following the manufacturer's protocol. The Zymo OneStep Inhibitor Removal kit (D6035; Zymo Research) was used following the manufacturer's instructions. DNA samples were quantified using Quant-iT on an Eppendorf AF2200 plate reader.

Taxonomic Assignment, Assembly, and Functional Analysis. Microbiome sequences were processed and analyzed with Human Longevity Inc.'s (HLI) microbiome annotation pipeline. Raw BCL data were de-multiplexed and converted to paired end reads of 2 × 125 bp in FASTQ format, trimming the adapter sequence. Reads then were filtered using Trimmomatic (1). After removal of low-quality bases and reads shorter than 90 nt, duplicated read pairs were identified with the program cd-hit-dup (2) by matching the first 50 bases from both R1 and R2 reads (cd-hit-dup parameter -u 50). Reads were aligned to human genome hg38 using the Burrows–Wheeler Aligner (3), and all reads that mapped were excluded from downstream analysis. All nonhuman reads were mapped to HLI's reference genome database, a collection of ~11,900 genomes of bacteria, archaea, viruses, and eukaryotes downloaded from the National Center for Biotechnology Information (NCBI) including both complete and draft genomes. For mock-community samples, reads were aligned to the 20 reference genomes in the mock sample. After read-mapping, an in-house implementation of an expectation maximization (EM) algorithm, similar to the GRAMMy algorithm (4) was used to process the reads that were ambiguously mapped to multiple genomes to estimate RGA. The genome coverage, which is the total length of mapped reads divided by the reference genome length, was calculated for each reference genome based on the EM program's assignment of reads to genomes. The relative abundance of a reference genome is the genome coverage divided by the sum of all of the genome coverages. The relative abundances were aggregated at each taxonomic rank: species, genus, family, order, class, and phylum. Nonhuman reads were assembled using IDBA-UD (5) to generate scaffolds, and reads were mapped to scaffolds using BWA. The EM algorithm was applied to this mapping result to assign reads to scaffolds and to calculate the coverage for all scaffolds. Scaffolds were assigned to the best-matched genomes based on the EM algorithm's assignment of probabilities of mapping reads to reference genomes and reads to scaffolds. Scaffolds that were assigned to the same species were put in a species bin if the total lengths of scaffolds were larger than 33% of average genome length of the reference genomes in this species. Scaffolds longer than 100 kb without a genome assignment were put in an "unknown" species bin category. ORFs were predicted from scaffolds using MetaGene (6) and were compared with several reference protein or domain families, including COG, KOG, Pfam, TIGRFAM, and a comprehensive protein sequence database with RPS-BLAST (COG and KOG), Hmmer3 (Pfam and TIGRFAM) (7), and NCBI BLASTP+ (protein db). Only the nonoverlapping top score alignments were used to calculate the protein family abundance. Per-genome read-depth bed files were generated using a combination of SAMtools (3, 8) and genomeCoverageBed from BEDTools (9, 10). Replicates were compared and shown to be highly concordant. The replicate pairs were combined into a single file to improve coverage depth representation. Per genome and per preparation method, these files were used to determine the fraction of each genome covered and the mean read-depth coverage per 10-kb bin. These files also were used to estimate the probability of coverage per base normalized by total genome read depth.

1. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
2. Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658–1659.
3. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
4. Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F (2011) Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PLoS One* 6(12):e27992.
5. Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28(11):1420–1428.
6. Noguchi H, Park J, Takagi T (2006) MetaGene: Prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* 34(19):5623–5630.
7. Finn RD, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38(Database issue):D211–D222.
8. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
9. Dale RK, Pedersen BS, Quinlan AR (2011) Pybedtools: A flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* 27(24):3423–3424.
10. Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.

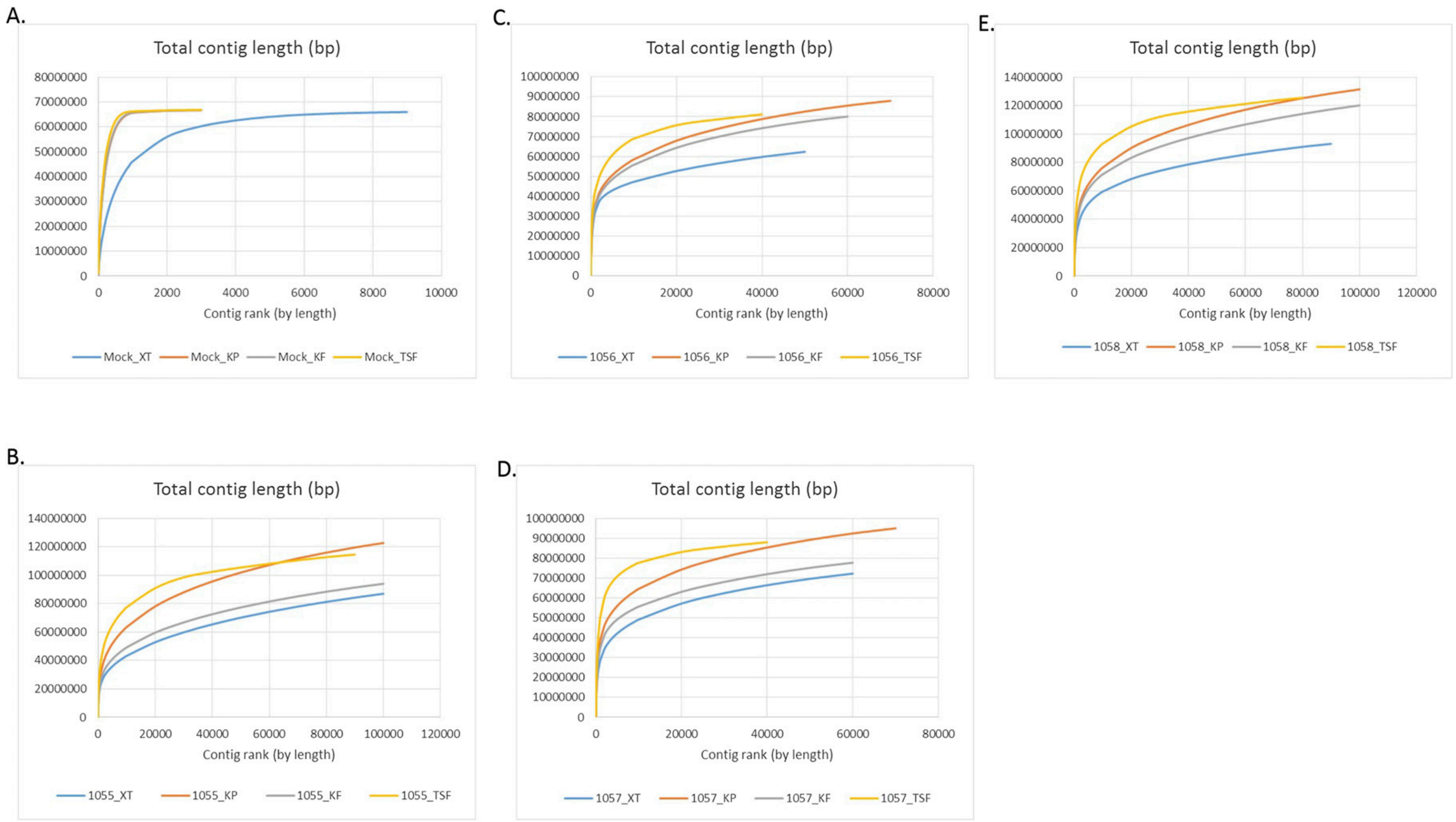

Fig. S1. Accumulated total length of contigs. The *x* axis represents contig rank in order of increasing length, and the *y* axis represents the accumulated total length of contigs in base pairs for mock community (*A*), 1055 stool samples (day 0) (*B*), 1056 stool samples (day 3) (*C*), 1057 stool samples (day 7) (*D*), and 1058 stool samples (week 8) (*E*). Blue, XT; orange, KP; gray, KF; yellow, TSF.

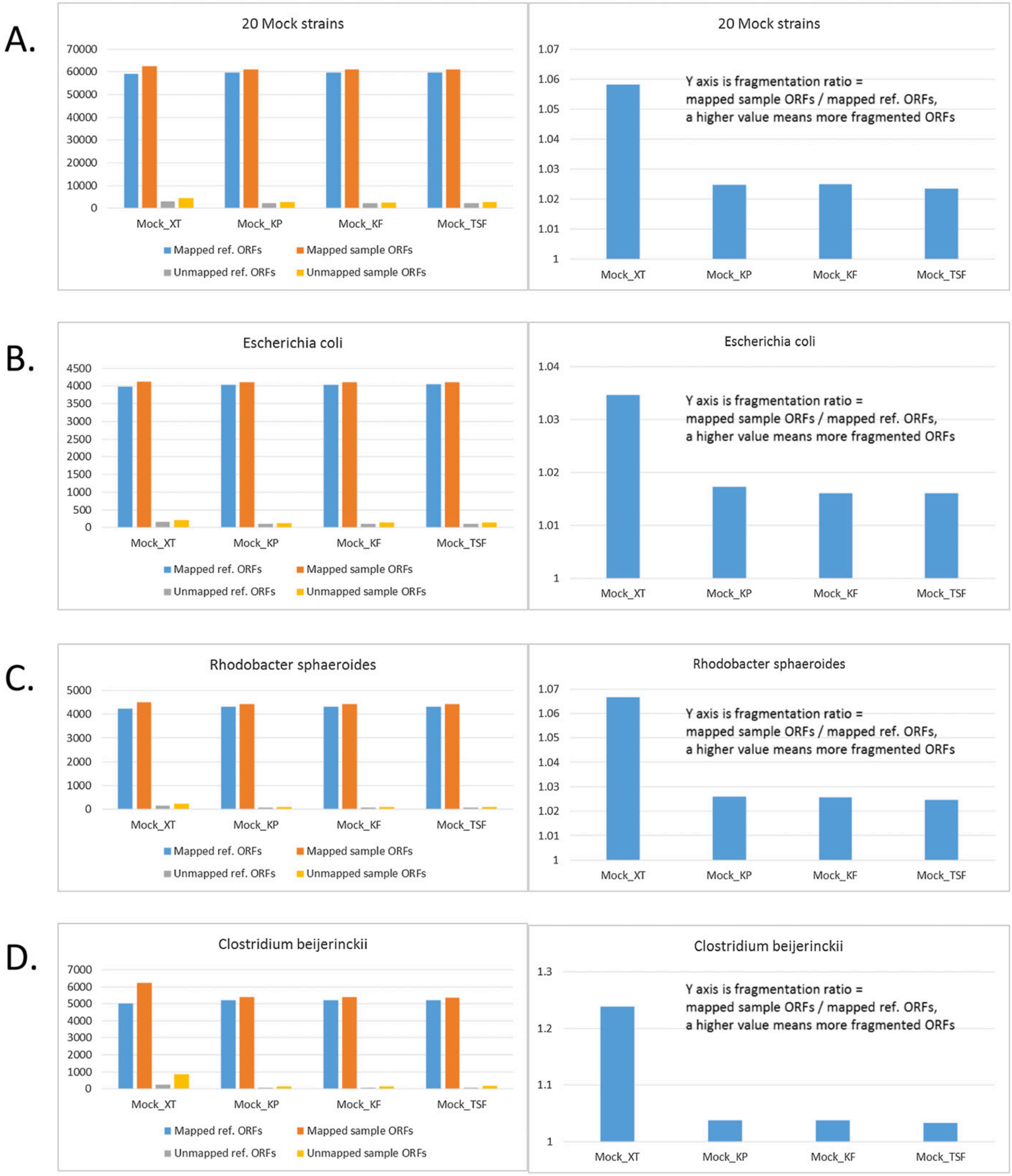

Fig. S2. Prediction of the function and pathway landscape of the mock community across library preparations. ORFs predicted from the assembled scaffolds and true ORFs called from the complete mock reference genomes are compared using cd-hit-2d at ≥98% sequence ID over 95% of the length of predicted ORFs to find matched ORFs in these two sets. (*A*) ORF analysis across all 20 organisms of the mock community. (*B*) ORF analysis for *E. coli*. (*C*) ORF analysis for *R. sphaeroides*. (*D*) ORF analysis for *C. beijerinckii*.

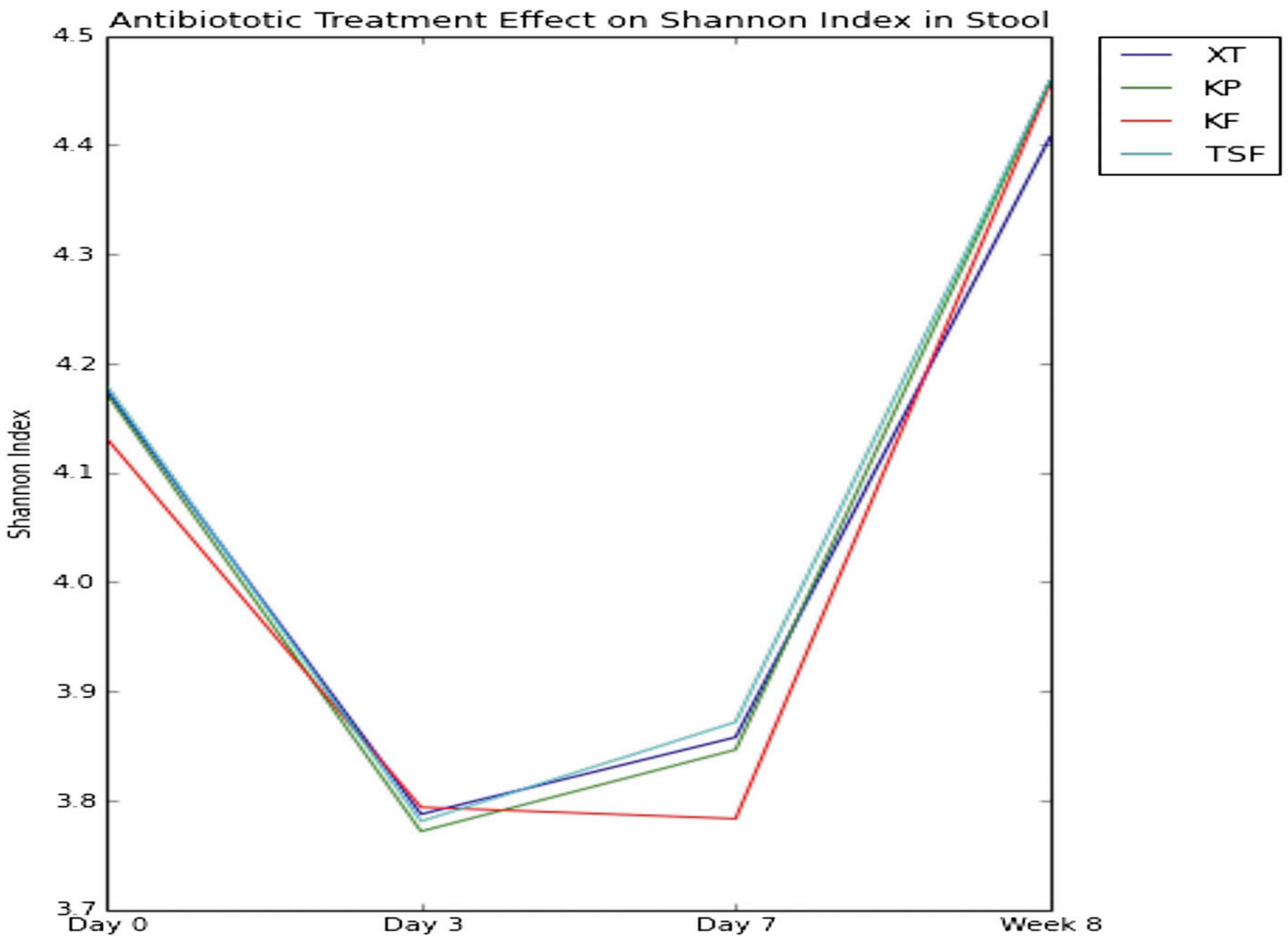

Fig. S3. Shannon index plot analysis of clinical stool specimens following antibiotic treatment across time points and library protocols. The *y* axis represents the diversity index, and the *x* axis represents the time point. Red, KF; dark blue, XT; green, KP; light blue, TSF.

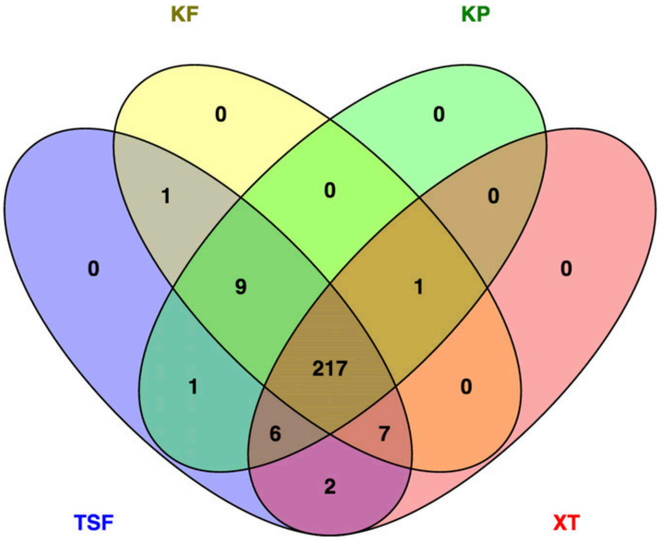

Fig. S4. Venn diagram analysis was used to identify microbial species/strains significantly modulated in common following administration of antibiotic. Purple (TSF), yellow (KF), green (KP), and pink (XT) shadings represent organisms that were significantly modulated in stool specimens following antibiotic treatment from sequencing data from the TSF, KF, KP, and XT libraries.

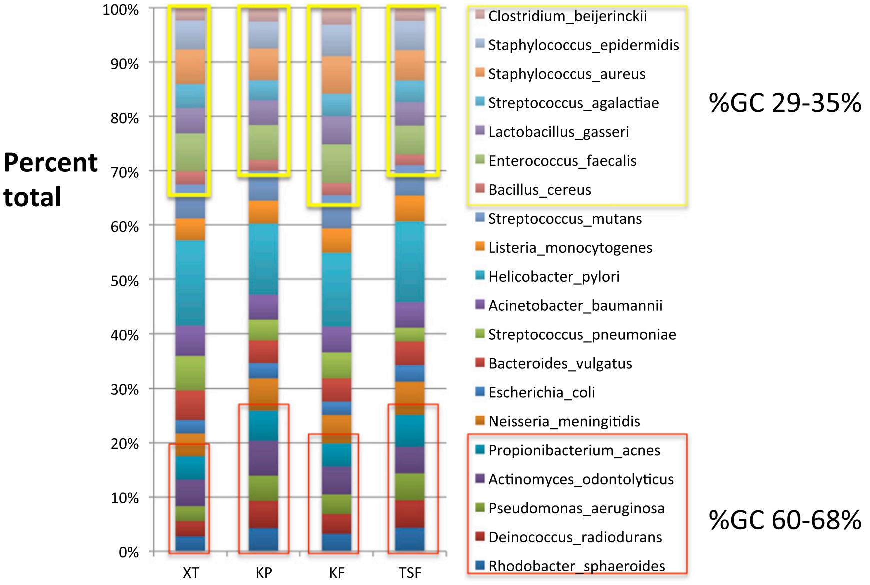

Fig. S5. Relative genome abundance measurements of the mock community across library methods. High %G+C content (60–68%) species are highlighted by a red box, and species with a low %G+C content (29–35%) are highlighted with a yellow box.

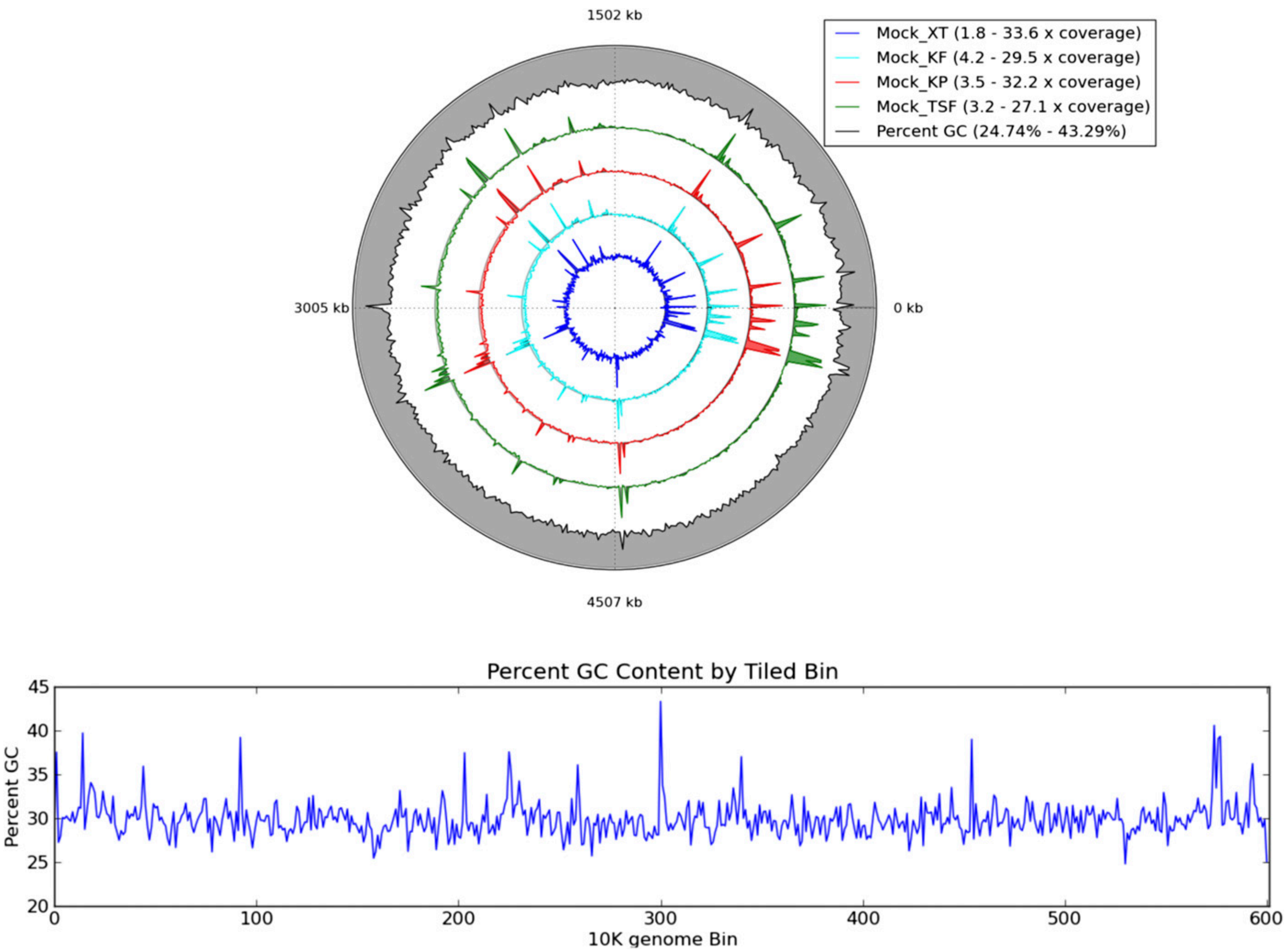

Fig. S6. (*Upper*) Map of mean GC content and mean relative sequencing depth by library preparation method across the genome of *C. beijerinckii*. The complete genome for the organisms is used, including any known plasmids, and (*Lower*) is subdivided into 10-kb bins for mean analysis. The outer gray ring depicts the delta from 50% GC content for a sequence bin. The four colored inner rings depict the delta of the average sequencing depth for the bin from the average sequencing depth of the whole genome. Maximum and minimum values per ring are given in the legend.

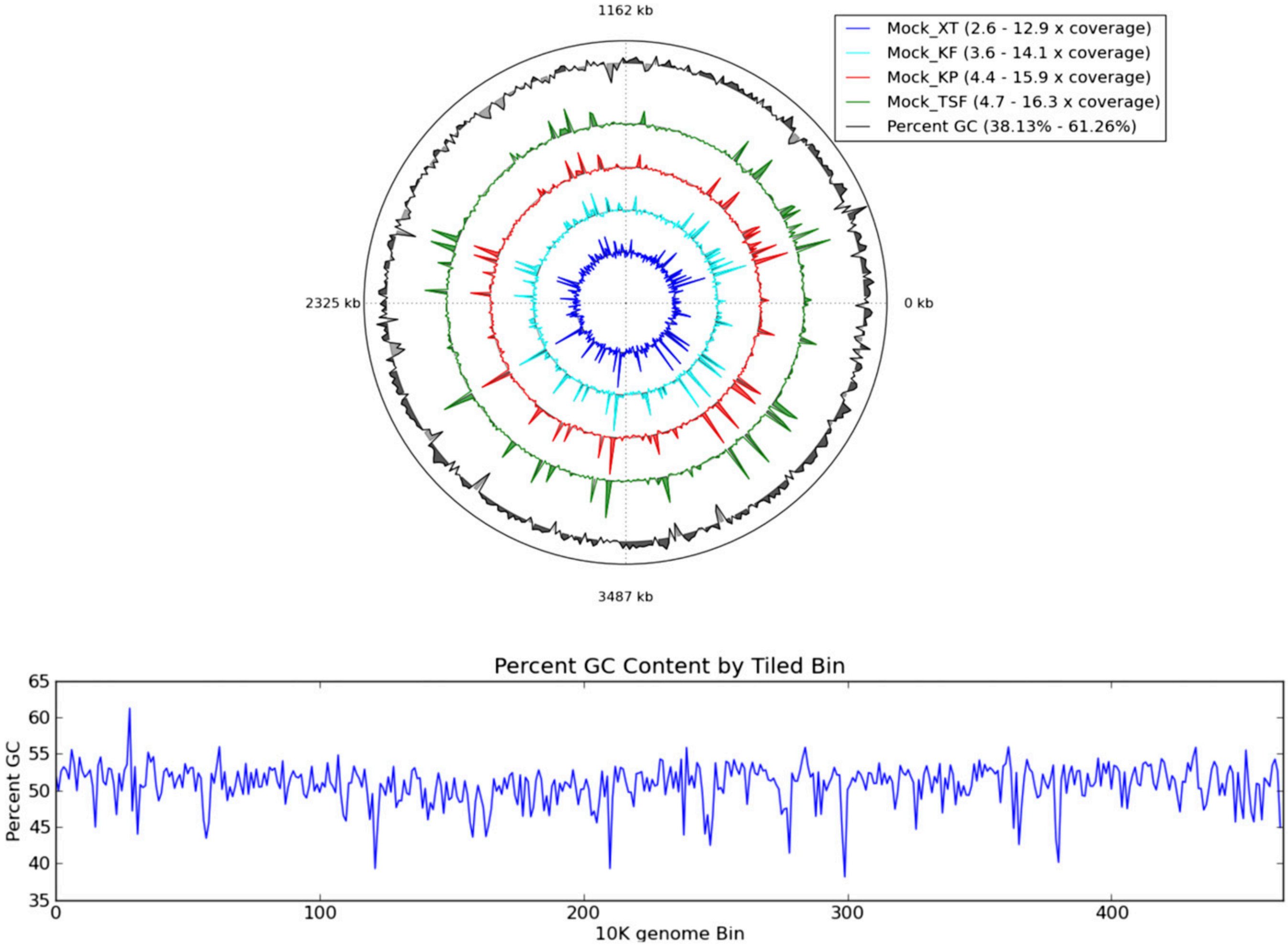

Fig. S7. (*Upper*) Map of mean GC content and mean relative sequencing depth by library preparation method across the genome of *E. coli*. The complete genome for the organism, including any known plasmids, is used and (*Lower*) is subdivided into 10-kb bins for mean analysis. The gray outer ring depicts the delta from 50% GC content for a sequence bin. The four colored inner rings depict the delta of the average sequencing depth for the bin from the average sequencing depth of the whole genome. Maximum and minimum values per ring are given in the legend.

Other Supporting Information Files

Dataset S1 (XLSX)