Supplementary Information: A genomic toolkit for the mechanistic dissection of intractable human gut bacteria

Jordan E Bisanz[1], Paola Soto-Perez[1], Cecilia Noecker[1], Alexander A Aksenov[2], Kathy N Lam[1], Grace E Kenney[3], Elizabeth N Bess[1], Henry J Haiser[4], Than S Kyaw[1], Feiqiao B Yu[5], Vayu M Rekdal[3], Connie WY Ha[6], Suzanne Devkota[6], Emily P Balskus[3], Pieter C Dorrestein[2], Emma Allen-Vercoe[7], and Peter J Turnbaugh[1,5,8]*

[1]Department of Microbiology and Immunology, University of California San Francisco, San Francisco, California 94143, USA
[2]Collaborative Mass Spectrometry Innovation Center, Department of Pediatrics, Center for Microbiome Innovation, Department of Pharmacology, Skaggs School of Pharmacy and Pharmaceutical Sciences, San Diego CA 92093, USA.
[3]Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge MA 02138, USA
[4]Faculty of Arts and Sciences Center for Systems Biology, Harvard University, Cambridge MA 02138, USA
[5]Chan Zuckerberg Biohub, 499 Illinois Street, San Francisco California 94158, USA
[6]Department of Medicine, Division of Gastroenterology, Cedars-Sinai Medical Center, Los Angeles California, 90048, USA
[7]Molecular and Cellular Biology, University of Guelph, Guelph ON N1G 2W1, Canada
[8]Lead Contact

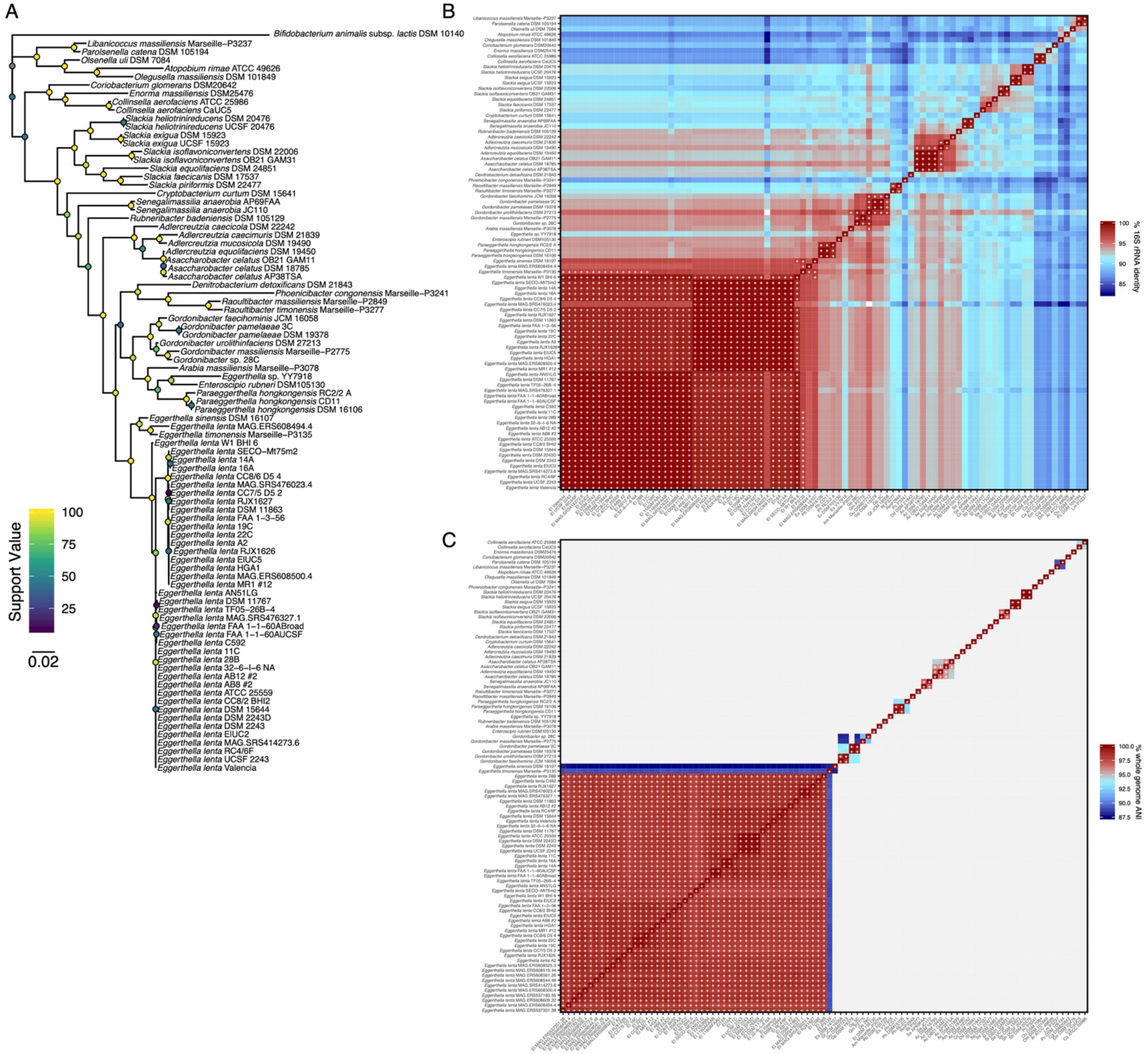

Figure S1. Phylogenetic and taxonomic analysis of Coriobacteriia related to Figure 1. (A) Phylogenetic tree of Coriobacteriia strains based on 16S rRNA alignment**.** 16S rRNA genes were extracted from genomes, aligned using DECIPHER and a tree generated with FastTree. *B. lactis* was included as an outgroup. **(B)** 16S rRNA sequence identity between genomes. **(C)** Whole genome average nucleotide identity between genomes. In both panels B and C, a white + denotes a comparison with >97% or >95% 16S rRNA or ANI respectively corresponding to the species boundary.

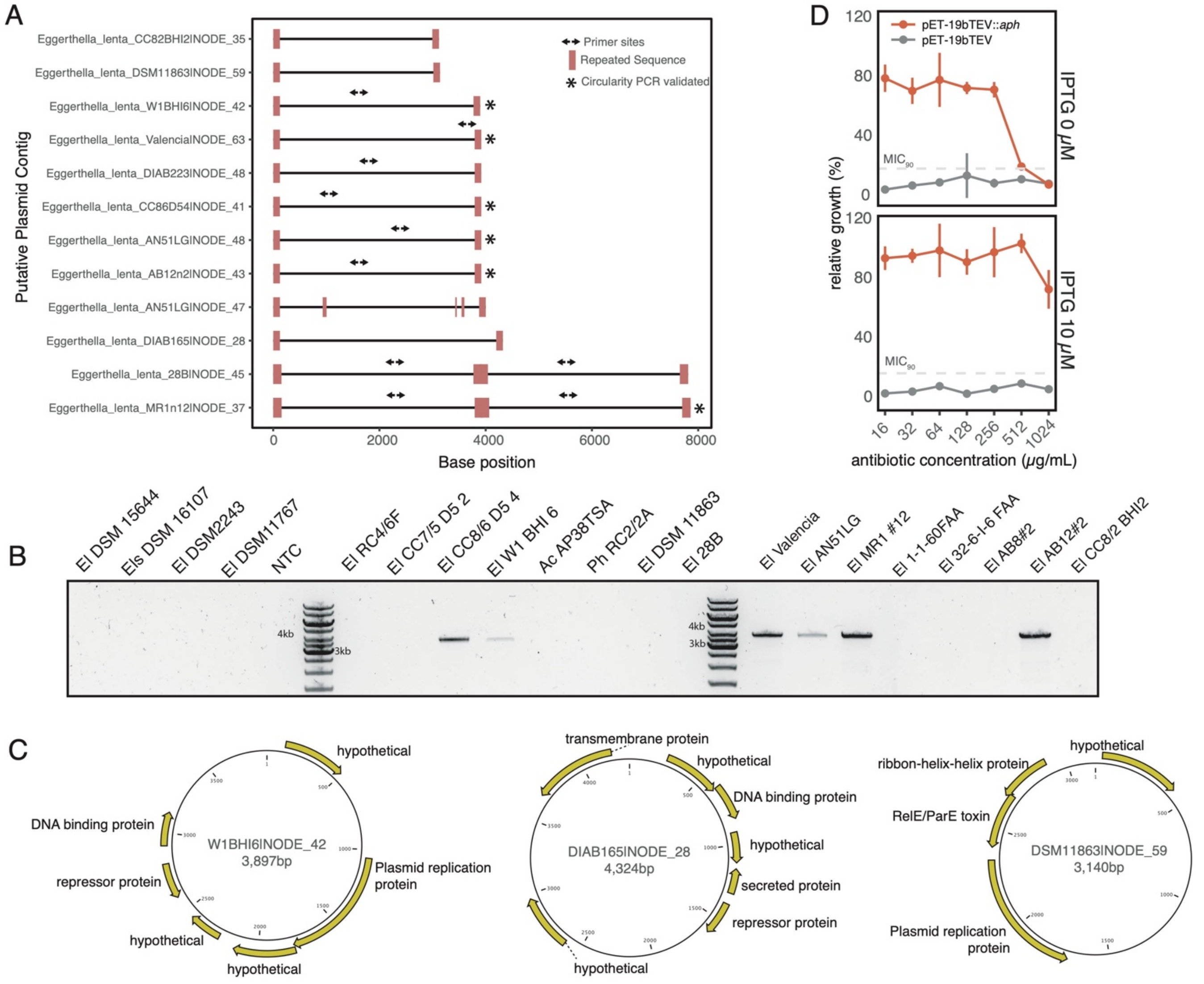

Figure S2. Mapping putative plasmids related to Figure 2. (A) Visualization of putative plasmids (>3kbp, >5-fold increase in coverage versus median of genome) reveals 12 putative plasmids with repeated termini. **(B)** PCR verification of circularity for W1BHI6-type plasmids. **(C)** Manual inspection and BLAST suggests 3 classes of plasmids in *E. lenta* strains. **(D)** Heterologous expression of *aph* confirms ability to impart resistance to kanamycin.

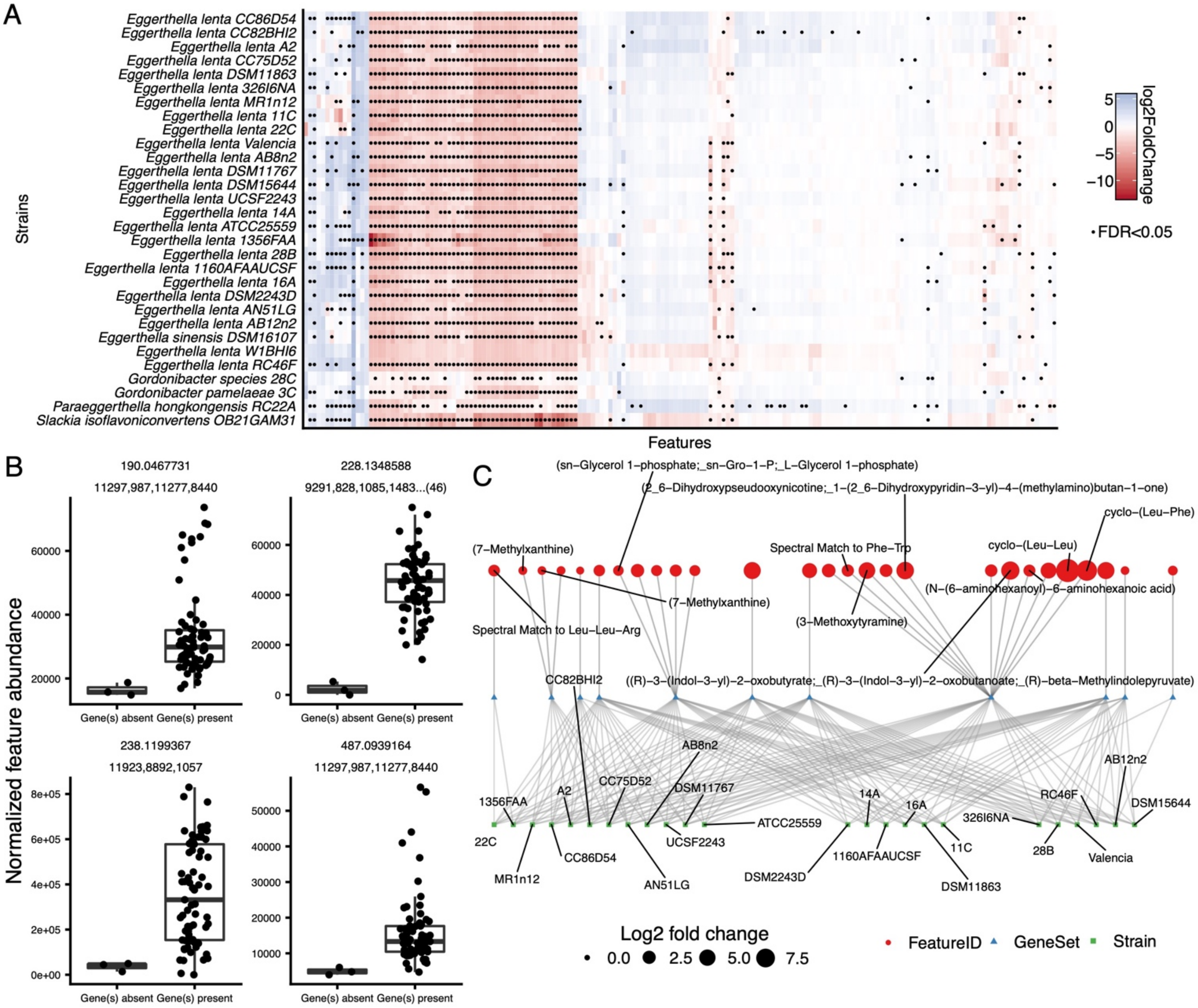

Figure S3. Untargeted metabolomic analysis of *E. lenta* strains related to Figure 2. (A) Normalized $\log_2$ fold change of the abundances of metabolite features compared with sterile media controls. All features that passed quality checks are shown. A black dot indicates a significant difference from media controls based on a Student's *t* test with a false discovery rate cutoff of 0.05. Features and strains are ordered based on hierarchical clustering with complete linkage. **(B)** Examples of metabolite features associated with variable gene presence/absence patterns across *E. lenta* strains. Each boxplot shows the peak abundances of a metabolite feature in cultures of strains with and without one or more genes. Each plot is labeled with the metabolite feature *m/z* as well as the IDs of the relevant gene ortholog clusters. **(C)** Visualization of putative links between gene families and metabolite features. Red nodes indicate metabolite features, blue nodes indicate gene families, and green nodes indicate strains. Strains are linked to gene families in their genome, and gene families are linked to metabolite features with which they are strongly associated (see Methods and **Table S4**). Metabolite labels are based on GNPS library identifications, except those in parentheses, which are putative MAGI identifications. Unlabeled metabolite nodes represent unidentified features.

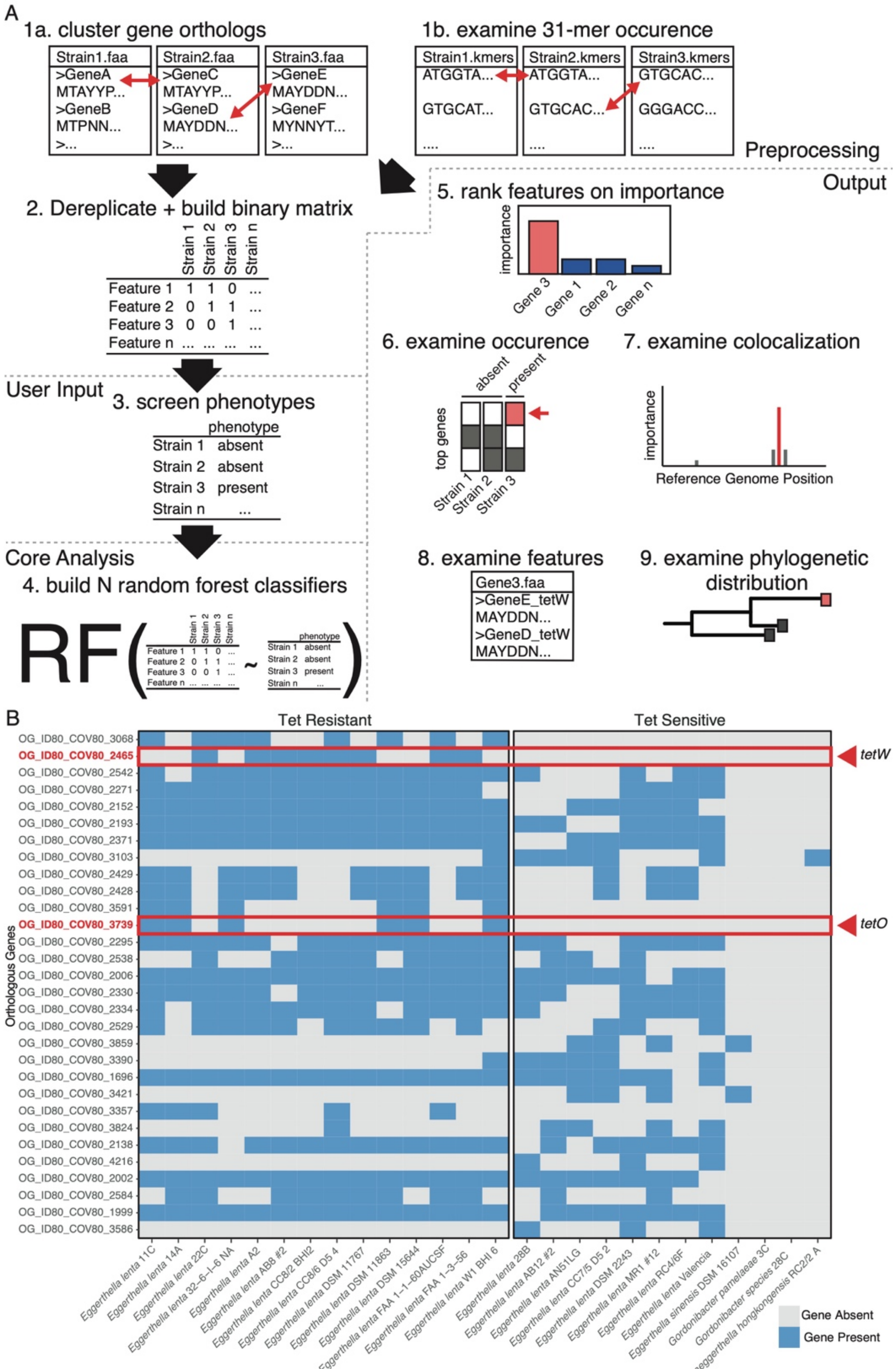

Figure S4. ElenMatchR descriptions related to Figure 4. (A) Description of ElenMatchR approach: Preclustered gene orthologs or 31-mers are dereplicated into co-occurring bins and used to construct a binary matrix. Next the user supplies a list of phenotypes which are used to build a variable number of random forest classifiers whose results are aggregated and summary statistics are provided. Features are then ranked on feature importance (mean decrease in GINI) and presented to the user with a heatmap of gene occurrence, manhattan plot to observe co-localized features, phylogenetic tree, tabular list of features with their metadata, and fasta files containing the top hits. **(B)** Representative ElenMatchR results for tetracycline resistance at stringent clustering thresholds. The top features plotted using default parameters (80% identity and coverage) include two tetracycline resistance homologs (*tetW* and *tetO*) which occur in a mutually-exclusive pattern demonstrating the utility of random forests in uncovering more complex gene correlations.

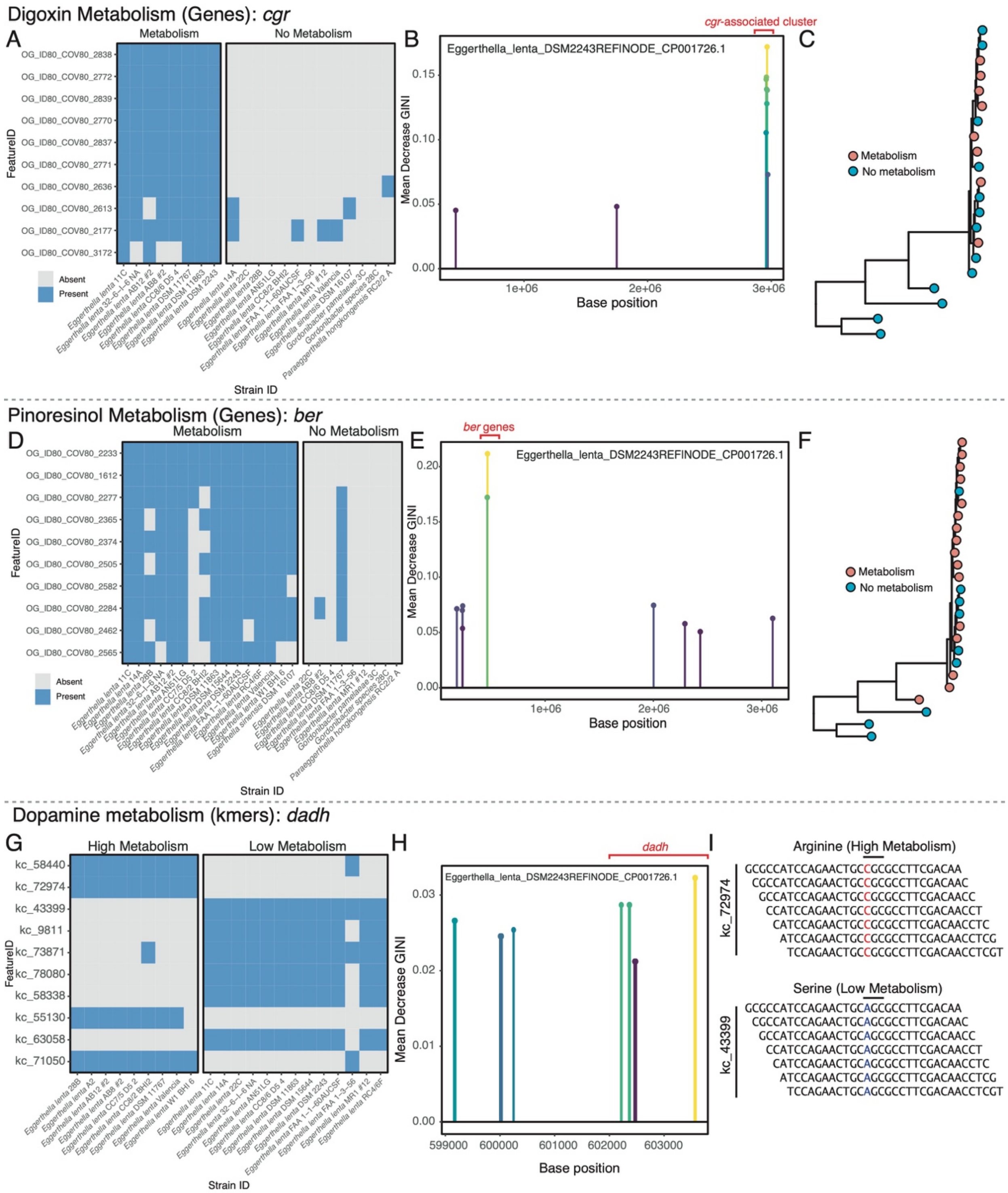

Figure S5. ElenMatchR Demonstration Cases related to Figure 4. (A) Heatmap of gene occurrence in digoxin metabolizing and non-metabolizing strains reveals a set of genes conserved to metabolizers. **(B)** These genes are co-localized into an island at ~3Mbp in the reference genome termed the Cgr-associated gene cluster. **(C)** Digoxin metabolism does not display a strong signal of phylogenetic correlation. **(D)** Pinoresinol metabolism is correlated with two genes present in all metabolizers. **(E)** These two genes, a reductase (*ber*) and its putative regulator (*berR*), are located at a single genomic locus. **(F)** Pinoresinol metabolism is not correlated with phylogeny. **(G)** Dopamine metabolism by *E. lenta* perfectly associated with two clusters of k-mers (kc_72974 and kc_43399). **(H)** The top k-mers map to a single locus in the reference genome (~600 Kbp). **(I)** Tiling these k-mers reveals they cover a SNP which determines metabolism in isolates and mixed communities.

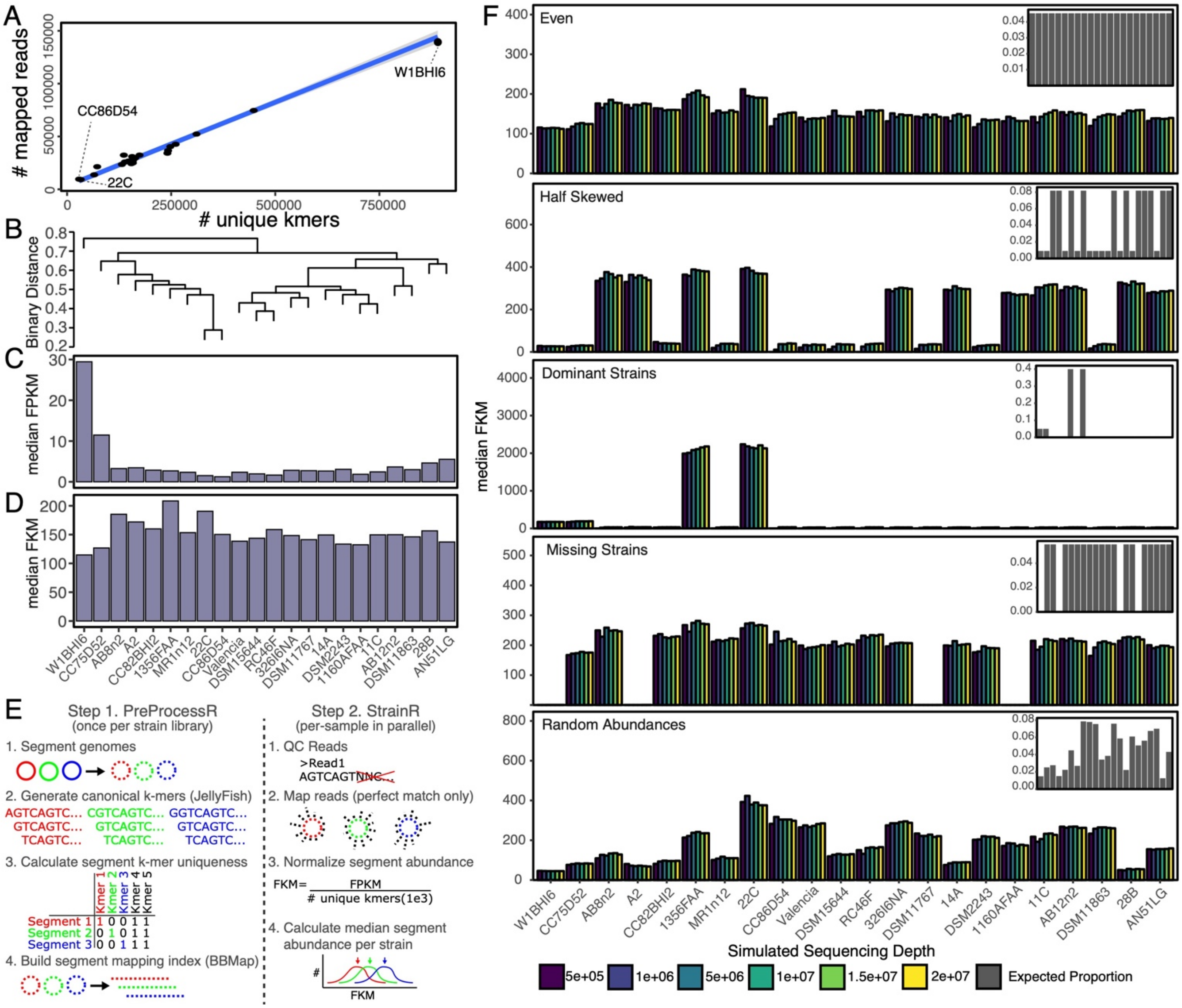

Figure S6. StrainR validation related to Figure 6. (A) Correlation of strain abundance with the number of unique k-mers in the genome highlights that normalization is required for accurately determining community composition. **(B)** UPGMA clustered tree of binary distances based on shared k-mers. **(C)** Uncorrected strain abundances in an even pool demonstrate an apparent 28-fold difference in abundance from the most to least abundant which is a function of shared k-mer profile (FPKM, fragments per thousand bases per million reads mapped). **(D)** Corrected strain abundances by using the median fragments per thousand unique k-mers per million reads mapped (FKM) reveal an approximately even community (1.8-fold between highest and lowest abundance). **(E)** Logical workflow of two-step normalization strategy: For any given library of strains, an index is first created using PreProcessR which is then used by StrainR for per-sample abundance normalizations. **(F)** Validation of community composition in *in silico* designed communities of even, skewed, and random abundances, and missing strains (inset: expected proportions).

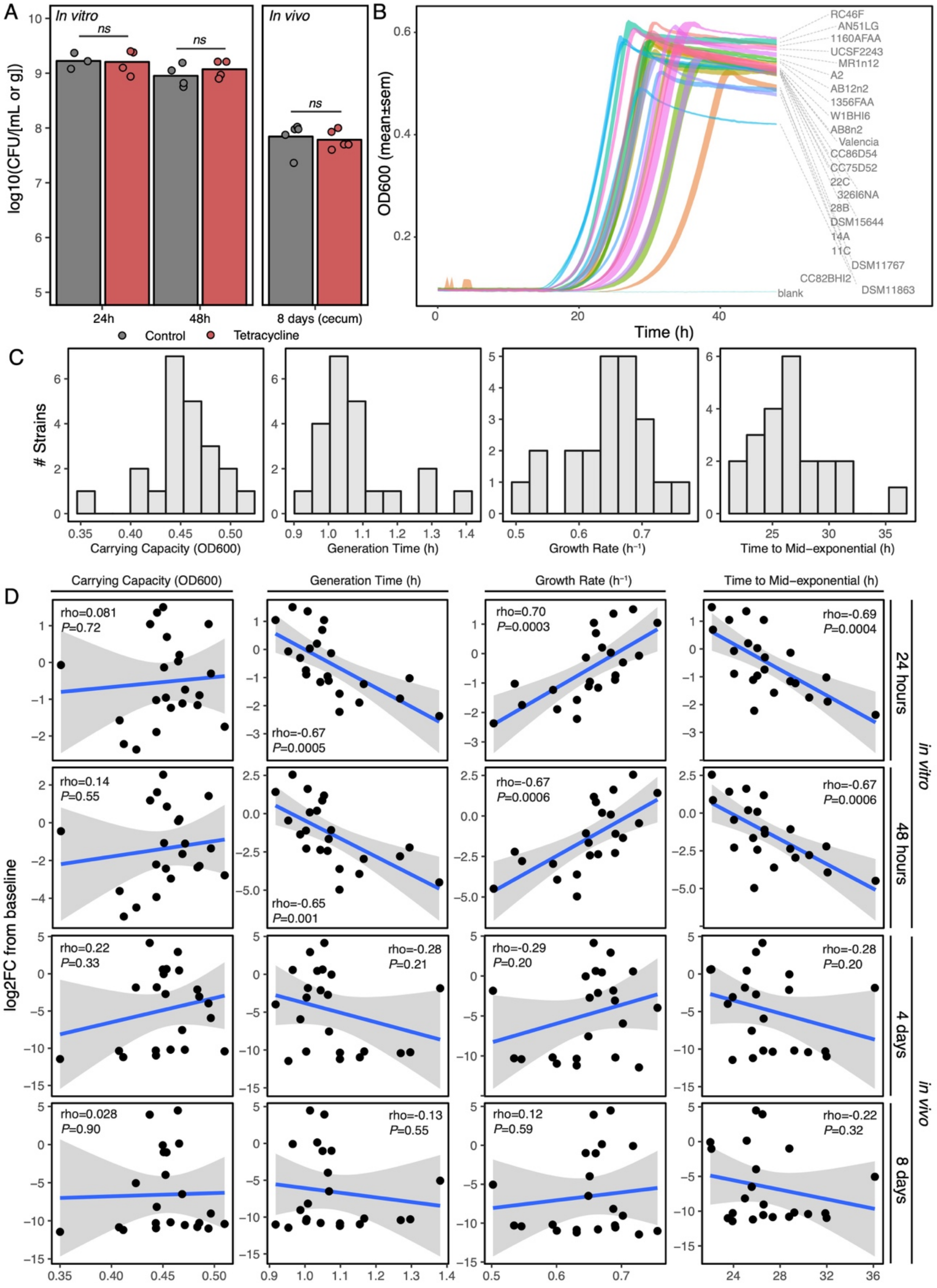

Figure S7. Growth in pooled communities and isolation related to Figure 6. (A) Total *E. lenta* CFU counts in *in vivo* and *in vitro* competitions. Total abundances are stable over time and unaffected by the addition of tetracycline selection. Statistical testing was conducted via Mann-Whitney U test. **(B) I**ndividual growth curves for strains present in pooled competition experiment demonstrate variable growth patterns. **(C)** Distribution of growth parameters among pooled *E. lenta* strains. **(D)** Spearman correlations of growth parameters with competitive outcomes demonstrates significant correlations between generation time, growth rate, and time to mid-exponential growth *in vitro* but not *in vivo*.