

Figure S1

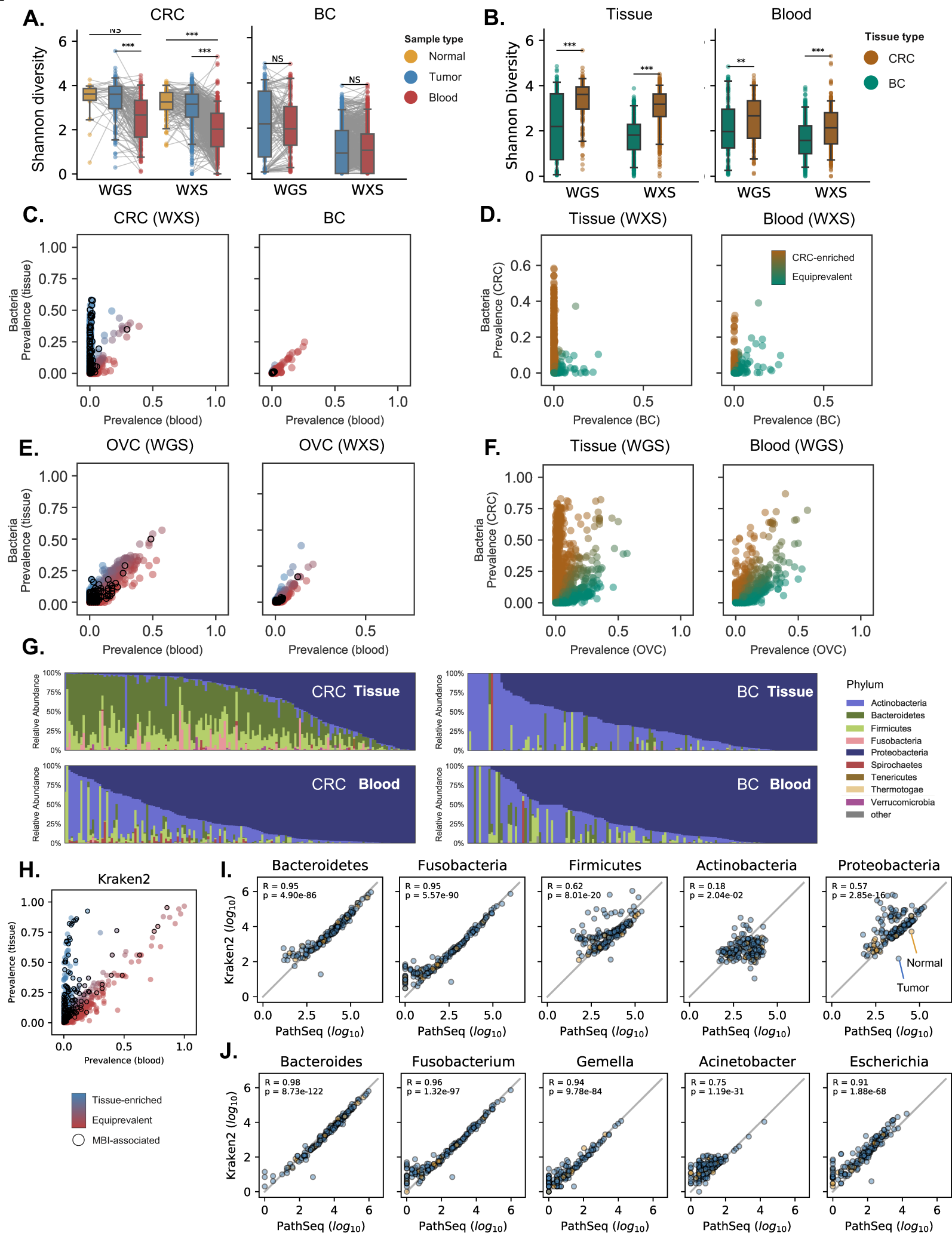


Figure S2

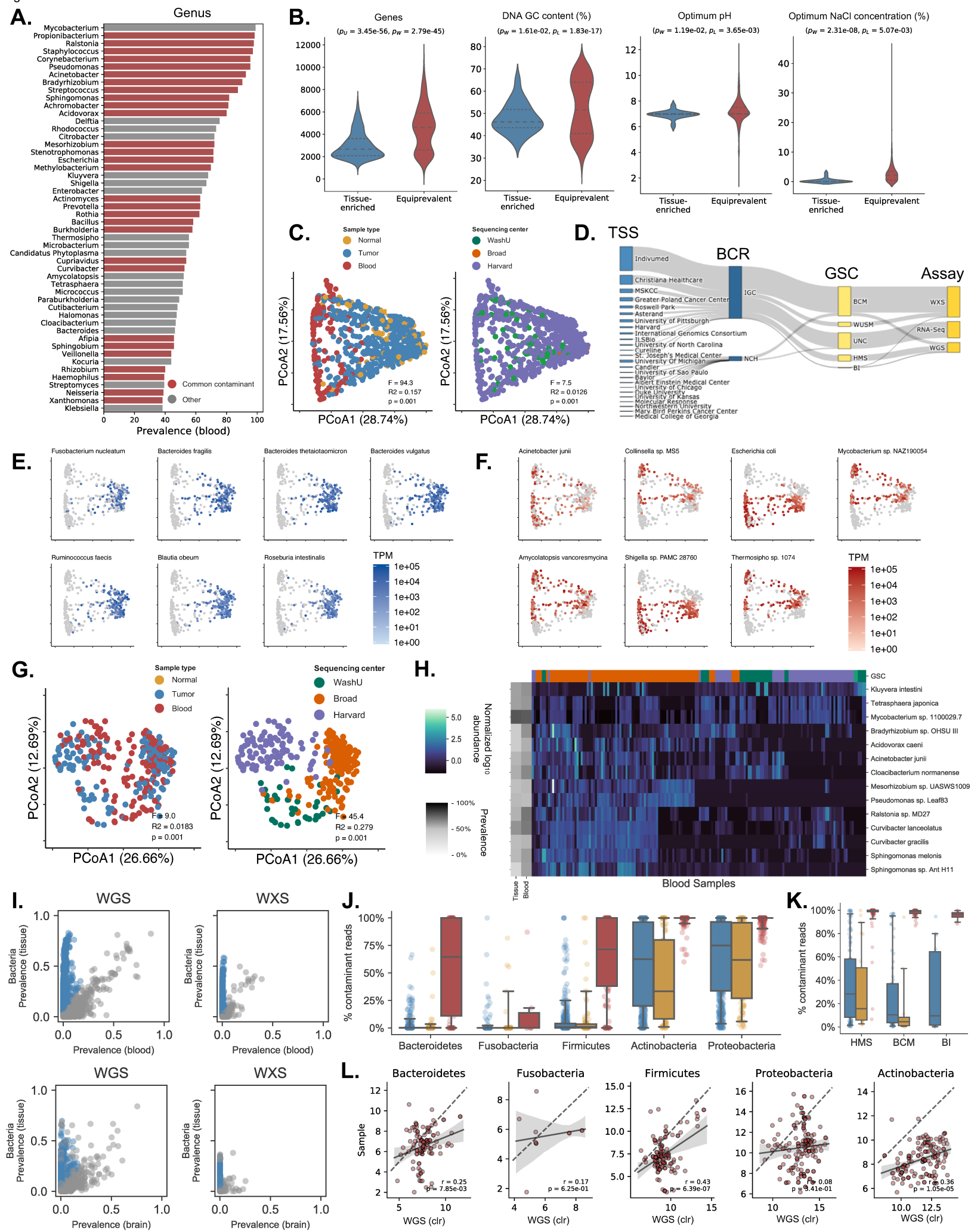
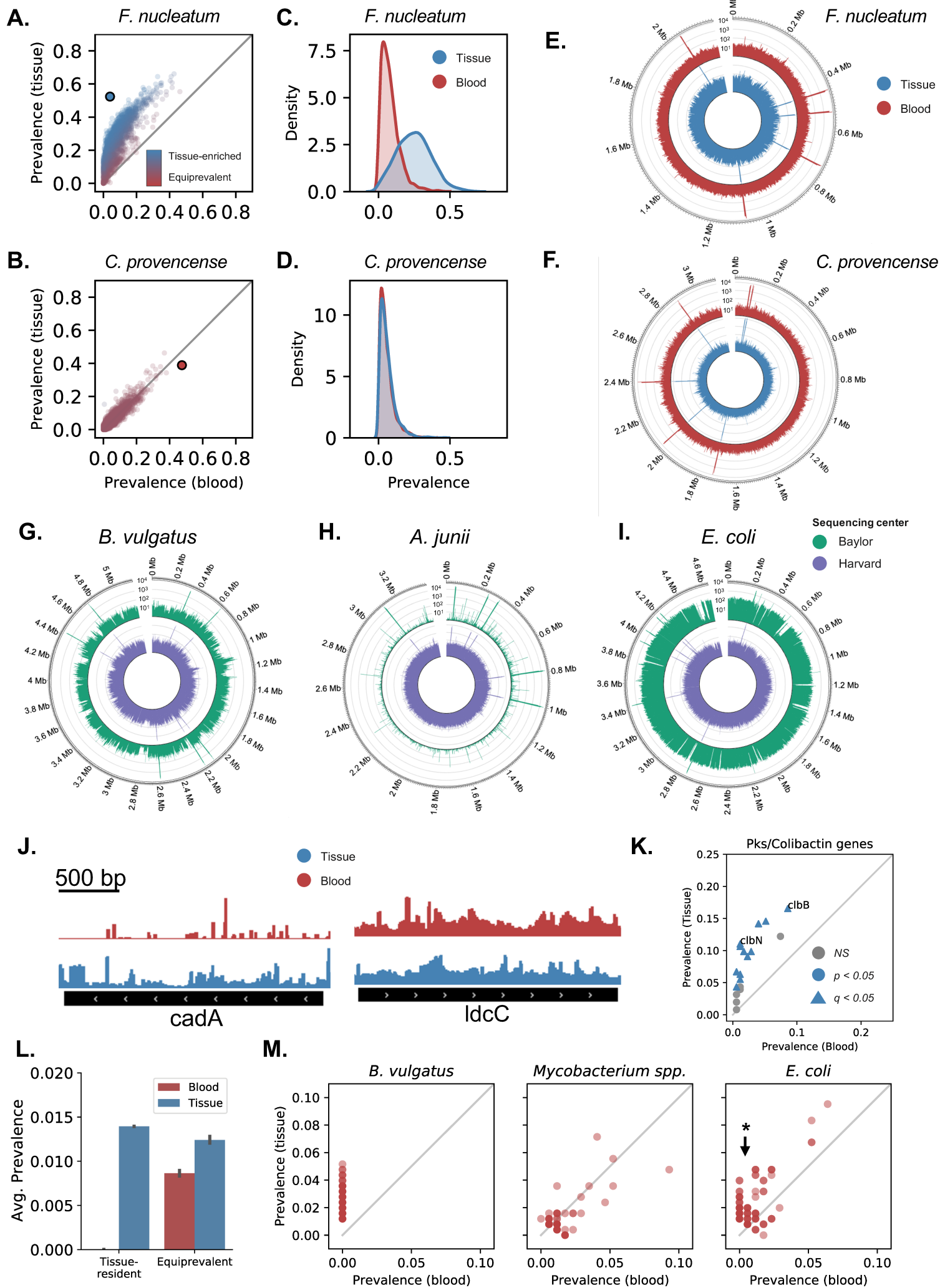


Figure S3



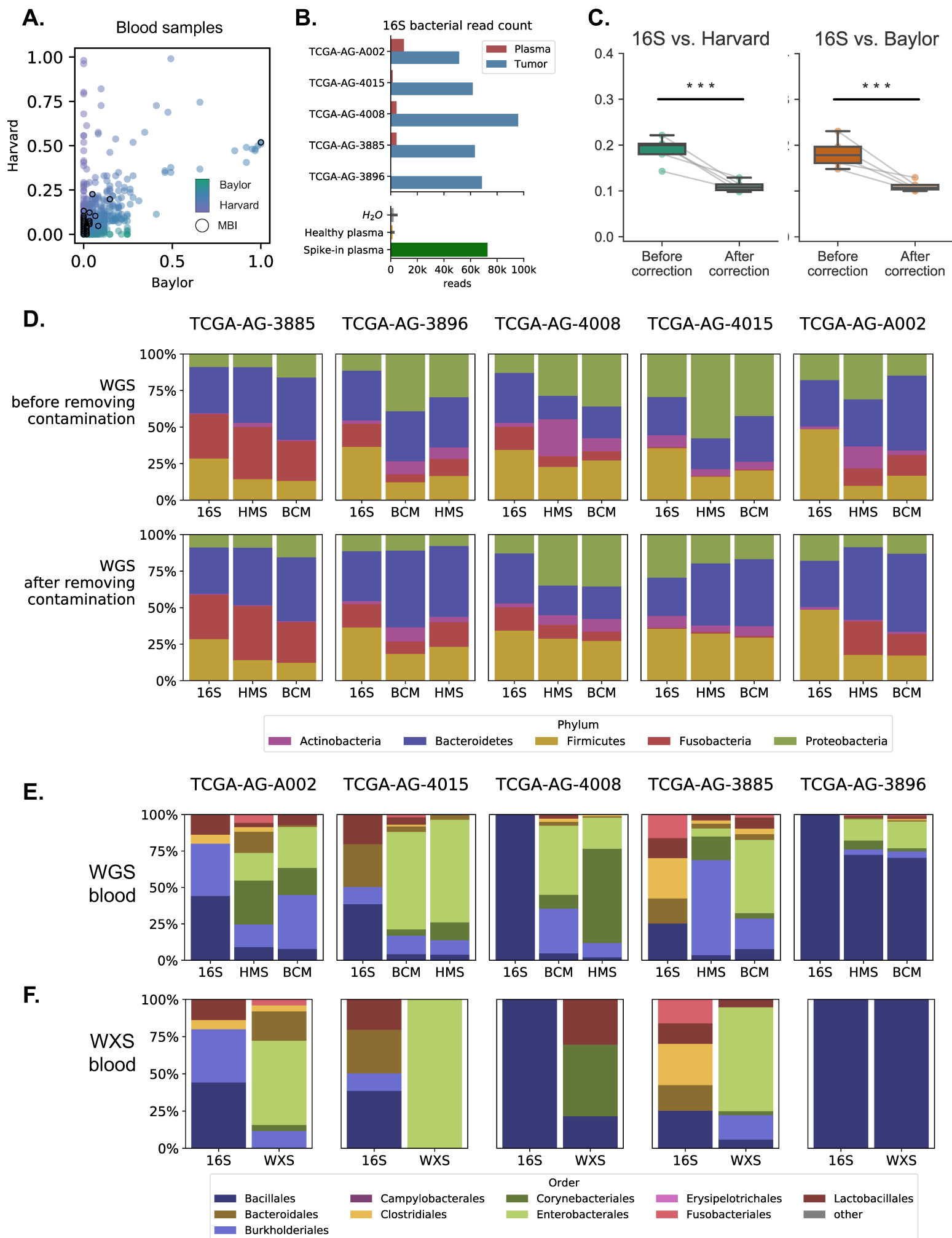
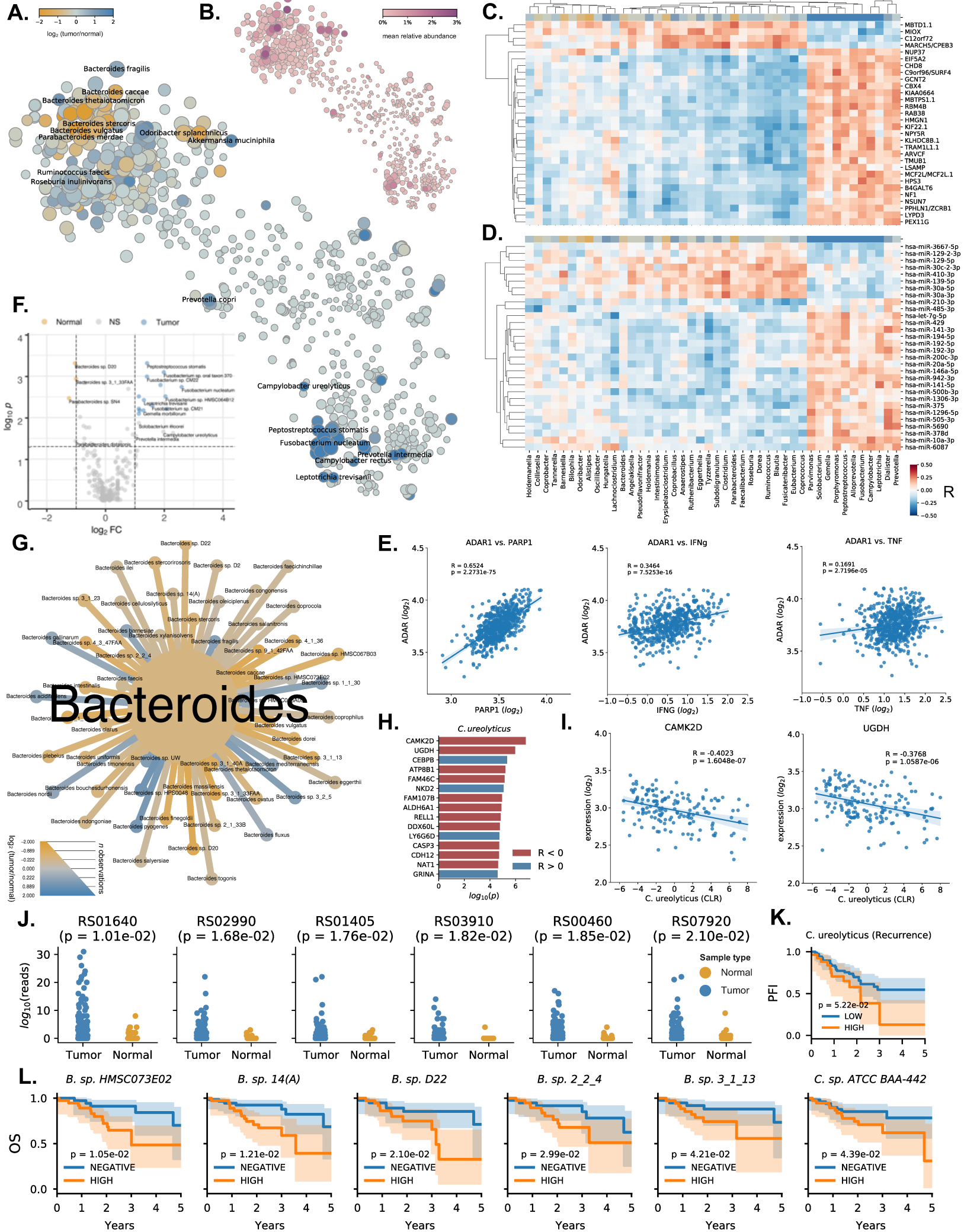
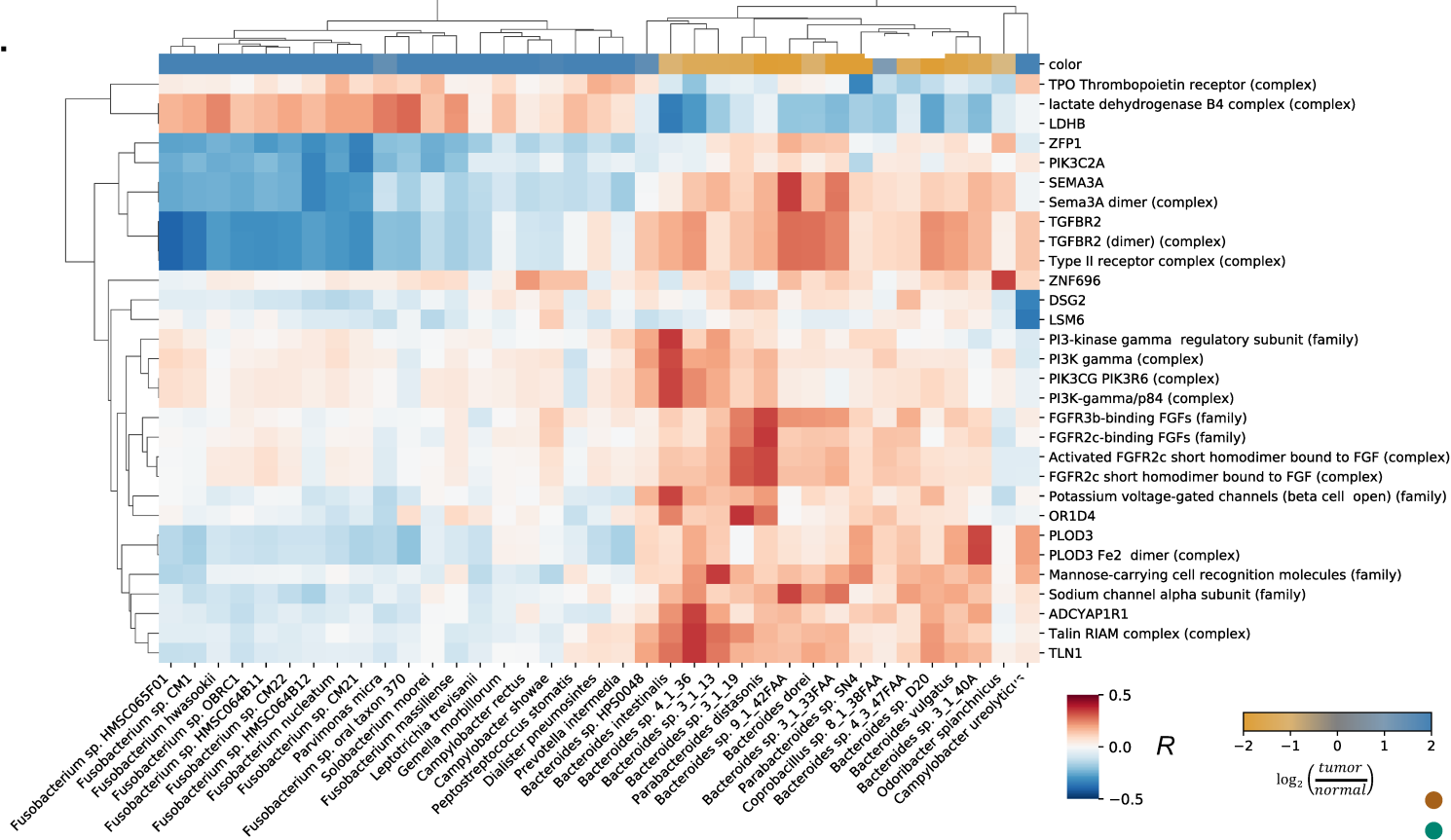
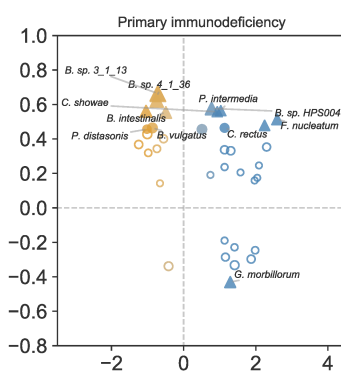
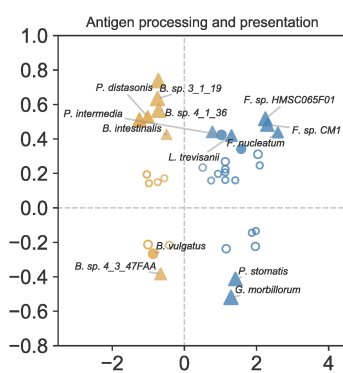
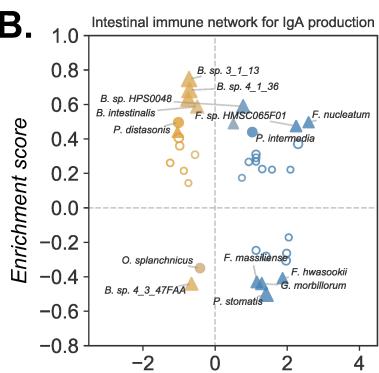


Figure S5

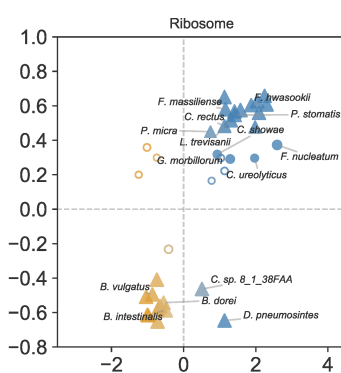
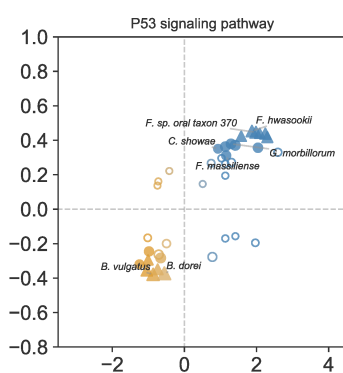
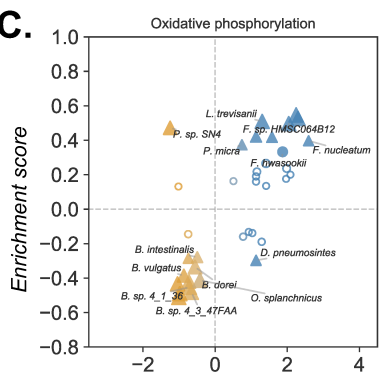
A.



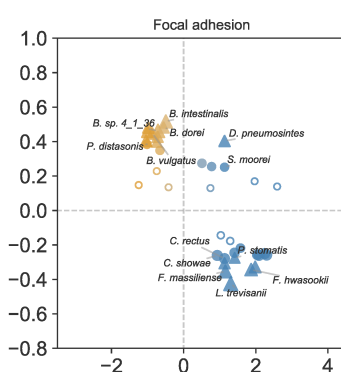
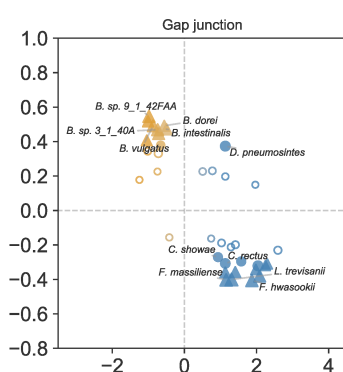
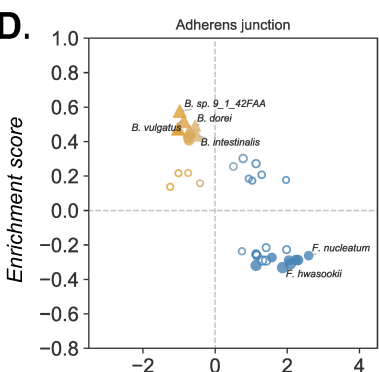
B.



C.



D.



E.

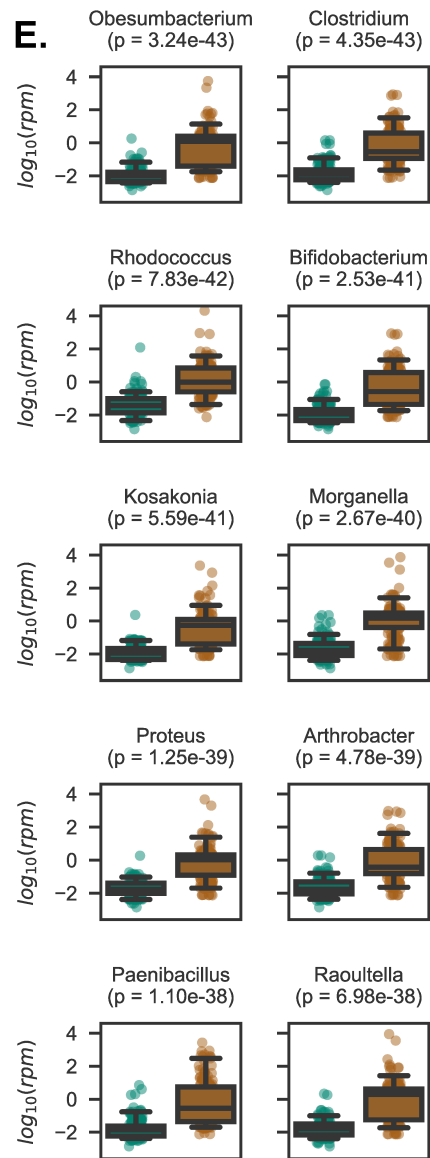
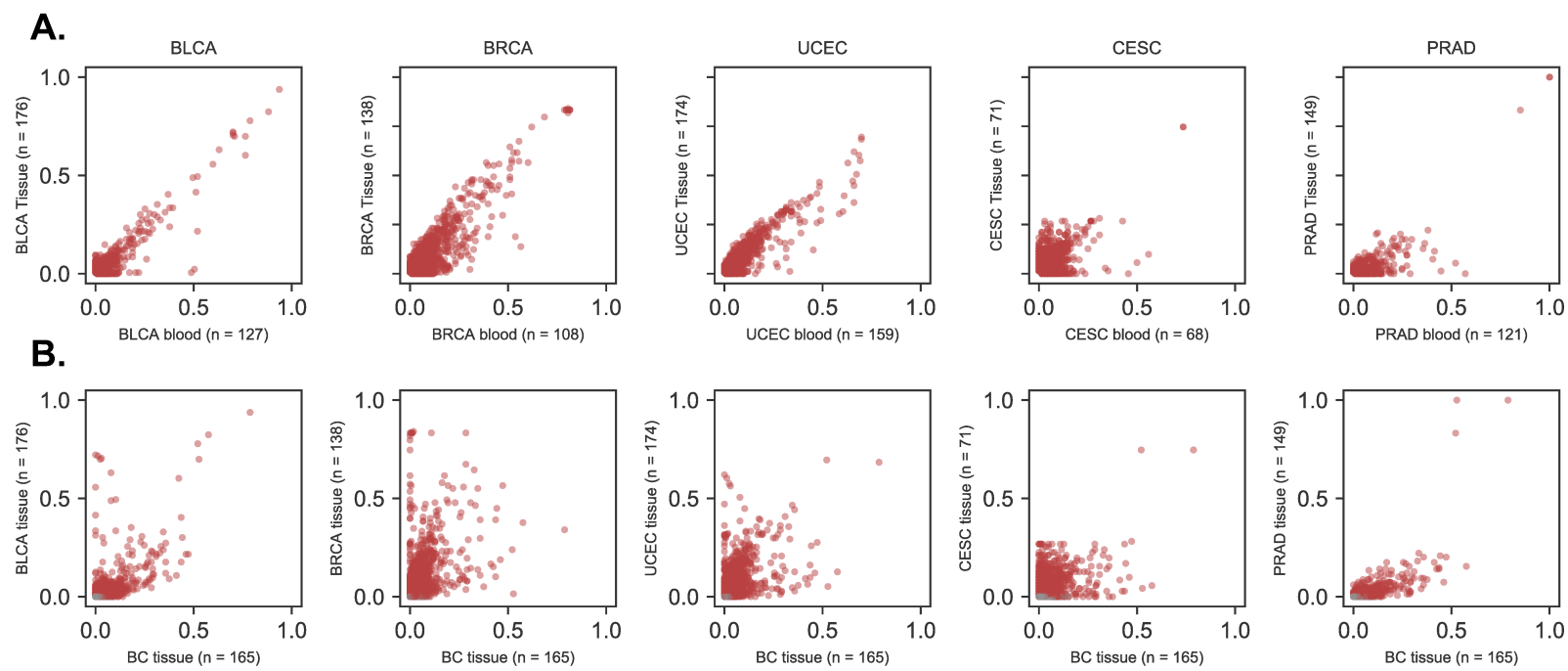


Figure S7



SUPPLEMENTAL FIGURE LEGENDS

Figure S1. WGS and WXS harbor colorectal bacterial reads distinct from blood and brain, Related to Figure 1

- (A) Matched analysis of Shannon diversity of bacteria in normal tissue (yellow), tumor tissue (blue), and blood (red) from CRC and BC patients in TCGA. By a paired, one-sided *t*-test, tumor and normal tissue from CRC patients have more bacterial diversity than CRC blood, but BC tumor tissue does not have more bacterial diversity than BC blood.
- (B) Shows diversity data from (A), but pools together tumor and normal tissue samples from BC (green) and CRC (brown) patients. By an unpaired, one-sided *t*-test CRC tissue has greater bacterial diversity than BC tissue, and CRC blood has greater bacterial diversity than BC blood.
- (C) Comparison of bacterial species prevalence in WXS data for blood and tissue samples from CRC and BC.
- (D) Comparison of bacterial species prevalence in WXS data for BC and CRC samples from tissue and blood.
- (E) Comparison of bacterial species prevalence in WGS and WXS data for blood and tissue samples from OVC.
- (F) Comparison of bacterial species prevalence in WGS data for OVC and CRC samples from tissue and blood.
- (G) Relative abundance of bacterial phyla in WXS data for tissue and blood samples from CRC and BC patients.
- (H) Comparison of bacterial species prevalence in tissue and blood using Kraken2 results.
- (C – D) Correlation between PathSeq and Kraken2 read counts estimated for selected phyla (I) and genera (J) in matched sequencing experiments from tumors (blue) and normal tissue (yellow).

Figure S2. Most equiprevalent taxa are common contaminants and are associated with particular sequencing centers, Related to Figure 2

- (A) Genera ranked by their prevalence in CRC blood samples. Common contaminants are marked in red.
- (B) Panel comparing gene count, G-C content, pH optima, and NaCl optima for tissue-enriched (blue) and equiprevalent (red) species. Dashed lines represent median and quartiles.
- (C) PCoA of WXS data for CRC samples, labeled by sample type (left) and sequencing center (right).
- (D) Schematic showing the path of TCGA sequencing samples from their tissue source site (TSS) to their genome characterization centers (GSC) for sequencing by various platforms.
- (E) Abundance of selected tissue-resident species overlaid on PCoA plot of WGS sequencing data (Figure 2D).
- (F) Abundance of putative contaminant species overlaid on PCoA plot of WGS sequencing data (Figure 2D).
- (G) PCoA of WGS data for BC samples, labeled by sample type (top) and sequencing center (bottom).
- (H) Heatmap clustering of bacterial species' normalized \log_{10} -abundance in blood samples from BC patients.
- (I) Species classified as tissue-resident by comparing WGS prevalence in tissue and blood samples (blue) are consistent with prevalence comparisons of BC tissue vs. CRC tissue and for analogous comparisons made with WXS data.
- (J) Fraction of contaminant bacterial reads from WXS data for normal (yellow), tumor (blue), and blood (red) samples from CRC patients, broken down by the five most prevalent phyla.
- (K) Fraction of contaminant bacterial reads from WGS data for each sample type at Harvard, Baylor, and Broad.
- (L) The CLR-normalized relative abundances of the five most prevalent phyla are uncorrelated between matched WGS and WXS blood samples. Dashed lines represent

Figure S3. Detecting tissue-resident and contaminant species with gene-level resolution, Related to Figure 3

- (A) Prevalence of genes belonging to *F. nucleatum* genes in blood vs. tissue; representative of tissue-resident species. The large blue dot indicates species-level prevalence.

- (B) Prevalence of genes belonging to *C. provencense* genes in blood vs. tissue; representative of contaminant species. The large red dot indicates species-level prevalence.
- (C – D) Distribution of gene prevalence in blood (red) and tissue (blue) for *F. nucleatum* (C) and *C. provencense* (D).
- (E – F) Genome coverage of reads aligning to *F. nucleatum* (E) and *C. provencense* (F) in blood (red) and tissue (blue).
- (G – I) Genome coverage of *B. vulgatus* (G), *A. junii* (H), *E. coli* (I) in samples from Baylor and Harvard.
- (J) Distribution of sequencing reads aligned to *cadA* (left) and *ldcC* (right) genes in blood (red) and tissue (blue).
- (K) Prevalence of genes in the *pks* island encoding colibactin in tissue and blood samples. These genes are enriched in tissue from CRC patients, consistent with previous reports.
- (L) Barplot showing average prevalence of variants from species defined as tissue-resident and equiprevalent species in blood (red) and tissue (blue). Data are represented as mean \pm 95% CI.
- (M) Prevalence of nucleotide variants in selected tissue-resident (*B. vulgatus*), contaminant (*Mycobacterium spp.*), and mixed-evidence species (*E. coli*). *E. coli* appears to harbor both tissue-enriched and equiprevalent sequencing variants.

* Denotes tissue-resident variants.

Figure S4. Metagenomic analysis of original TCGA tissue and blood samples validate tissue-resident microbial compositions and equivalent species as contaminants, Related to Figure 4

- (A) Prevalence of species in blood samples sequenced at Baylor vs. Harvard.
- (B) Total bacterial reads from 16S results of tissue (blue), plasma (red) and controls (bottom panel). Data are represented as mean \pm 95% CI.
- (C) Weighted UniFrac distances between microbial compositions of tissue sequenced at Harvard and Baylor (WGS) compared with patient-matched tissue analyzed at Duke (16S), before and after correcting for contamination.

- (D) Relative abundances of bacterial phyla in 16S results for tissue compared with tissue samples sequenced using WGS at Harvard and Baylor.
- (E – F) Relative abundances of bacterial phyla in 16S results for blood compared with blood samples sequenced using WGS at Harvard and Baylor (E) and WXS (F).

Figure S5. Analysis of microbe-microbe associations in contamination-adjusted data yields clinically relevant bacterial coabundance groups, Related to Figure 5

- (A) T-distributed stochastic neighborhood embedding (T-SNE) visualization showing coabundance of tumor-associated (blue) and normal tissue-associated (yellow) species. Points are sized proportional to the species' prevalence in tissue.
- (B) T-SNE visualization showing coabundance of detected species and their relative abundances (purple). Points are sized according to the species prevalence in tissue.
- (C) Coabundance groups are predictive of gene methylation (Methylation 27K).
- (D) Coabundance groups are predictive of micro-RNA expression (miRNA-seq).
- (E) ADAR1, which distinguished of *Fusobacterium* and *Bacteroides* coabundance groups, is positively correlated with PARP1, IFN-gamma, and TNF-alpha.
- (F) Volcano plot showing species that are more abundant in tumor samples (blue) or matched normal tissue (yellow).
- (G) Heat-tree showing *Bacteroides spp.* and their associations with normal (yellow) and tumor (blue) tissue.
- (H) Barplot showing genes significantly correlated with *C. ureolyticus* expression in tissue samples from CRC patients.
- (I) Scatterplots showing correlation of *C. ureolyticus* abundance (CLR) and gene expression of CAMK2D and UGDH.
- (J) Several in the *C. ureolyticus* genome are significantly associated with tumor samples (blue) relative to adjacent normal tissue (yellow).
- (K) *C. ureolyticus* is predictive of progression-free interval (PFI) among among patients with CRC recurrence.

- (L) Several species in the *Bacteroides* cluster are negatively predictive of patient survival (OS).

Figure S6. Bacteria in TCGA sequencing data are prognostic of mucosal barrier injury, immune infiltration, and tumor staging, Related to Figure 6

- (A) Heatmap showing correlations between CLR-transformed abundances of tumor- and normal tissue-associated species with features in the PARADIGM pathway matrix. PARADIGM scores are inferred from gene expression data.
- (B – D) Comparison of differentially abundant species and their association with tissue type (x-axis) versus enrichment score (y-axis) for KEGG terms associated with immune pathways (B), cancer-related pathways (C), and cellular adhesion pathways (D).
- (E) Enrichment scores for tumor-associated (blue) and normal tissue-associated (yellow) species across pathologic stages for selected pathways.
- (F) Panel showing enrichment scores across pathologic stage for all tumor-associated (blue) and normal tissue-associated (yellow) species and selected pathways.
- (G) Abundance of species in BC blood (green) and CRC blood (brown) samples, ordered by statistical significance.

Figure S7. Contamination-adjusted tissue microbiome profiles for all gastrointestinal cancers in TCGA, Related to Figure 7

- (A) Prevalence of species in blood and tissue from non-gastrointestinal cancers.
- (B) Prevalence of species in BC tissue and disease-specific tissue from non-gastrointestinal cancers.