Supplemental figures and tables:

Figure S1. Correlation analysis between 16S rRNA amplicon and metagenomic taxonomic relative abundance estimates. Twelve example species across samples from Gopalakrishnan (n=24), Matson (n=39), and Frankel studies (n=44) were used for comparison. In the Gopalakrishnan and Matson data, paired whole genome sequencing (WGS) and 16S rRNA amplicon libraries were developed and sequenced separately by the authors. For the Frankel data, 100bp 16S rRNA amplicon V4 fragments were bioinformatically identified within the WGS datasets, and subsequently extracted and analyzed. Top panel displays results for all 12 species; bottom panel displays results by individual species. Dotted black line represents expected correlation of relative abundance.

Figure S2. Alpha diversity analysis across all cohorts using Chao1, observed species, Shannon, and Simpson Reciprocal measures. A) Forest plots are shown for each metric within each cohort and as fixed or random effects model. P values for random effects model and I² heterogeneity values indicated for each metric. B) Statistical analysis shown for each metric within cohorts with mean value for responder (R) and nonresponder (NR) patients. Results of statistical testing with Mann-Whitney and t-test shown for each alpha diversity metric. For the Chaput cohort, analysis by chao1 is limited by original sequencing depth in the sample and is thereby omitted.

Figure S3. Complete species list for each index identified through re-analysis of raw sequencing data or as reported by the authors. Species associated with responders (R) are indicated in blue, nonresponders (NR) in orange; and signals consistent across all three indexes in green. Of note, as *Faecalibacterium prausnitzii* is a single species genus, genus and species level taxonomic identification are used interchangeably.

Figure S4: Sensitivity and specificity analysis of the Integrated Microbiome Prediction Index for response prediction in responder (R) vs nonresponder (NR), further stratified by inclusion (+) or exclusion (-) of stable disease (SD) if available. Two cohorts (Frankel and Routy) had available SD metadata for individual patients, while the other three (Chaput, Gopalakrishnan, and Matson) defined patients only as R vs NR even if SD data was used to make the clinical assessment. A. The AUC for each cohort is shown for each R-associated, NR-associated, and Sum indices. Grey boxes indicate AUC including SD patients as indicated by original authors while blue boxes exclude all SD patients. B. The Integrated Microbiome Prediction Index is shown without SD patients for the Frankel and Routy cohorts. Y-axis displays sensitivity, and the x-axis displays specificity. Mean area under the curve (AUC) was calculated without SD patients; mean AUC and standard error of the mean (SEM) is displayed on each graph.

Figure S5. Limited gut microbiome compositional changes are observed during treatment with checkpoint inhibitors. Beta-diversity analysis and differential abundance analysis comparing pre-treatment and paired longitudinal samples from the Chaput (A, B) and Routy (C, D) cohorts revealed stable gut microbiome composition after initiation of therapy. (A, C) Beta-diversity distances (Bray-Curtis metric) between paired baseline and on-therapy samples per patient were significantly closer to each other (blue) than samples from other patients (red) across multiple months (\*p <0.005 Mann-Whitney test). (B, D) Species-specific differential abundance analysis for each cohort indicated no significant changes in relative abundance between and on-therapy samples. X-axis displays baseline percentage abundance of each species. Y-axis displays -1\*Log10 unadjusted paired Mann-Whitney test p-values comparing percent abundance of species at indicated timepoint (by color) to baseline.

Figure S6. Differential abundance analysis of antibiotic resistance genes and virulence factors between R and NR for WGS datasets. We first cataloged read-level matches using DIAMOND translated search against multiple databases annotated for antibiotic resistance (A-C) and virulence factors (D-F). A hit to a functional gene required a match of ≥80% amino acid identity and ≥90% query coverage. Differential abundance analysis was performed for each functional gene of interest between responders (R) and nonresponders (NR) for each study (Mann-Whitney test), normalizing for high-quality sequencing depth per sample. Y-axis values display -1\*Log10 (Mann-Whitney p-values) per functional gene of interest, with values ≥2 indicating P ≤0.01 (dashed horizontal lines). X-axis values display Log2 (Fold Difference R/NR), with values ≤-1 or ≥1 indicating depletion or enrichment in R, respectively (dashed vertical lines). Abbreviations: ARDB, Antibiotic Resistance Genes Database; CARD, The Comprehensive Antibiotic Resistance Database; NDARO, National Database of Antibiotic Resistant Organisms; PATRIC\_VF, Pathosystems Resource Integration Center Virulence Factor; VFDB, Virulence Factor Database; Victors, the Victors virulence factor database.

Figure S7. 16S rRNA amplicon data processing for predicted functional elements using Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) based on taxonomic composition. Available 16S rRNA amplicon sequences, or mined V4 fragments that were bioinfomatically identified from whole genome sequencing data in Routy and Frankel cohorts, were subsequently extracted and analyzed. Forest plots for each pathway are shown across all cohorts with both fixed effect and random effects models; p values are indicated for the random effects models. A positive shift indicates enrichment inresponders (R) while a negative shift indicates enrichment in nonresponders (NR). Heterogeneity analysis included: estimates of I² (percentage of variability in effect estimates due to study heterogeneity), τ²-statistic (between-study variance), and the p-value from Cochran’s test.

Table S1. Complete sample manifest and clinical characteristics available for each cohort. Abbreviations: R, responder; NR, nonresponder; WGS, whole genome sequencing; PFS, progression free survival. T0, pre-treatment or initial treatment sample; T1, 1 month; T2, 2months; Ztox, sample at time of irAE; Zposttox, sample after irAE; LTbenefit, long-term benefit.

Table S2. Microbiome features of ICI responding patients and statistics with correction for multiple hypothesis testing. Each study is indicated by the first author and year of publication. Major microbial signatures identified are listed along with each statistical test indicated and adjusted p values after correction for false discovery rate (FDR). Abbreviations: R, responder; NR, nonresponder; adj., adjusted for FDR.

Table S3: All microbiome features identified through re-analysis. For each feature, both fixed and random effects Hedge’s G Standardized Mean Difference (SMD) are shown along with p values. Heterogeneity analysis includes estimates of I2 (percentage of variability in effect estimates due to study heterogeneity) and τ2-statistic (betweenstudy variance). The percent positive rate across both responder (R) and nonresponder (NR) samples is listed by cohort. The color range displays relative % positive rates per cohort (green, high; red, low).

Table S4: Cross validation of each bacterial signal in all indexes. For each index, each bacterial signal was removed from the index to assess the individual contribution as part of a leave-on-out analysis. Hedge’s G for fixed and random effects model, along with p values and heterogeneity analysis, are shown for each signal. Heterogeneity analysis includes estimates of I2 (percentage of variability in effect estimates due to study heterogeneity) and τ2-statistic (between-study variance). Sensitivity and specificity was also calculated, and the AUC is shown for each cohort as well as an average across all studies. The color range displays relative AUC for individual cohorts and mean AUC (green, high; red, low).

Table S5. Effect of prior antibiotic exposure on alpha diversity metrics in responder (R) vs nonresponder (NR) patients. Using the Routy cohort and data analyzed with the bioinformatics platform herein, samples were stratified by clinical response and prior antibiotic exposure. Only pretreatment or initial sample was analyzed for each patient. Mean and standard error of the man (SEM) are shown. Statistical testing performed with both t-test and Mann Whitney.