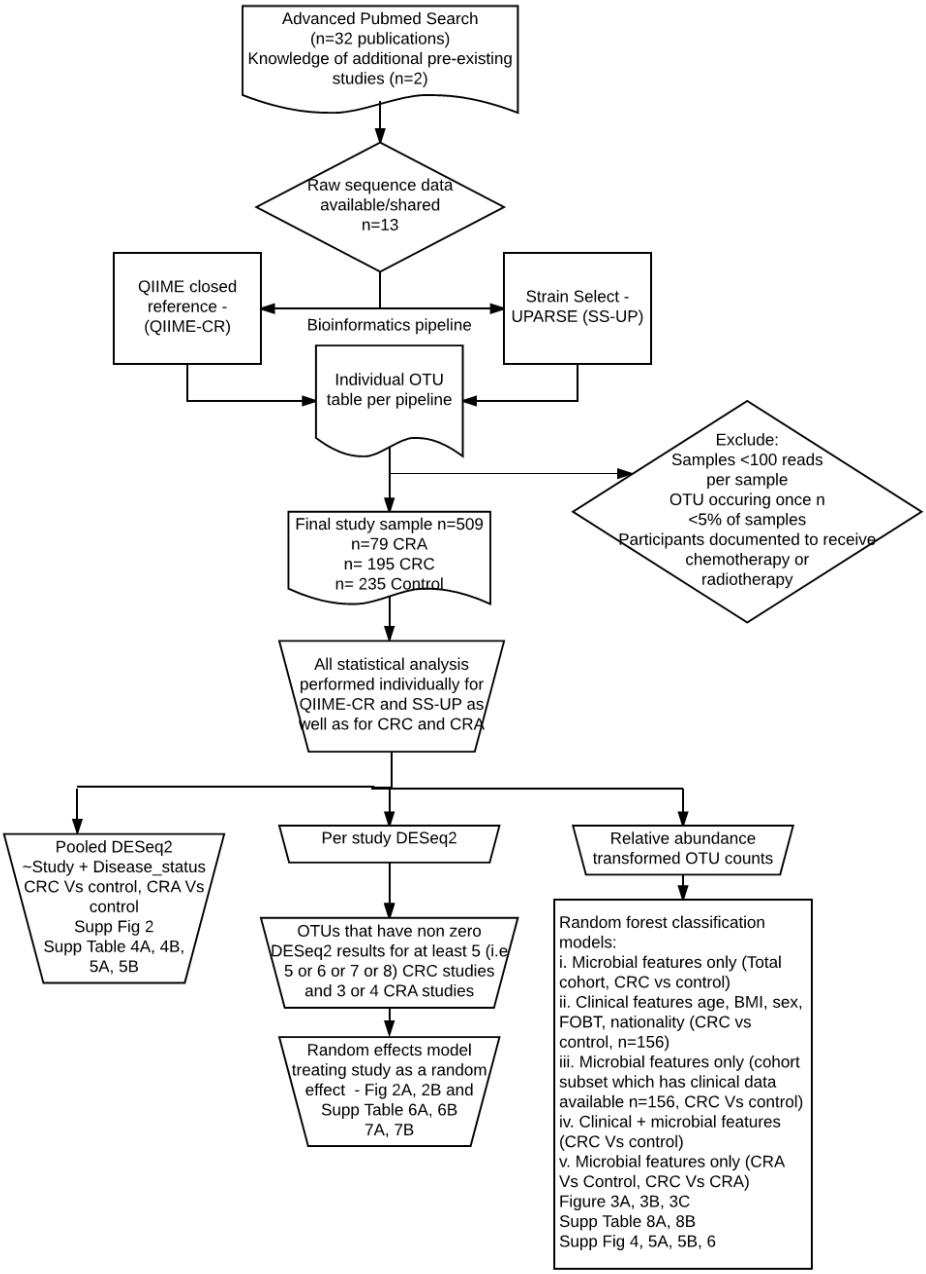
Supplemental Methods

Analysis flow:



The final search term using PubMed advanced search was ((((((((((bacterial microbiome OR gut microbiome OR microbiota OR microbial)) AND (fecal or feces)) AND (colorectal cancer[Title] OR colon cancer[Title] OR colorectal adenoma[Title] OR adenomatous polyp[Title] or colorectal carcinoma[Title])) AND ("2006/01/01"[PDAT] : "2016/04/01"[PDAT])) AND humans[MeSH Terms]) NOT review[Publication Type]) AND Humans[Mesh])). The manuscript required the terms bacterial microbiome, gut microbiome or microbiota in its main text, the terms colorectal cancer or colorectal adenoma or adenomatous polyp or colorectal carcinoma in the title, included human subjects only and published within the last 10 years.

QIIME-CR processing:

For the QIIME-CR pipeline, quality filtering and demultiplexing for the 454 datasets was done using the split_libraries.py command in QIIME 1.8 [1]. Minimum and maximum read lengths were chosen based on the target amplicon length to filter out truncated or erroneously long reads for both QIIME-CR and SS-UP. The filtering lengths used for each are summarized in Supplementary Table 1. Additionally, we used the default parameters for quality filtering (i.e., exclusion of sequences with >6 ambiguous bases, homopolymer runs >6 nucleotides, mismatches to the primer or barcode sequence). For Illumina data, we used the multiple_join_paired_ends.py and multiple_split_libraries_fastq.py scripts from QIIME 1.9, as they could process multiple files simultaneously. The quality filtering parameters were set to default (i.e reads were truncated at the first instance of a low-quality base call (q <20) and reads were excluded if <75% of the length of the original read). QIIME 1.9.0 was used only for initial fastq processing for the large MiSeq-based studies. OTU clustering and taxonomy assignment for all studies was performed using QIIME 1.8.0.

Quality-filtered and demultiplexed datasets from both the 454 and Illumina studies were assigned to reference based OTUs using pick_closed_reference_otus.py, which employed uclust 1.2.22q [2] with reverse strand matching enabled. In this strategy, input sequences were aligned to a pre-defined cluster centroid in the reference database (Greengenes_13_8).[3] A sequence was retained only if it matched the reference dataset at a threshold of 97% identity. A disadvantage of this approach is the disregard

of reads that are dissimilar to a reference. For one study [4], fasta-formatted sequence files were shared on the MG-RAST repository, but qual files were omitted. Hence quality filtering was not possible or this study and only length trimming was done prior to clustering for both the QIIME-CR and SS-UP pipelines. In two studies, [5 6] 454 was used to collect both F and R reads but since they were not paired, reads were assessed as the sum of two libraries of single ended reads.

SS-UP Processing:

Strain Select – UPARSE (SS-UP) (Second Genome, Inc) pipeline utilized the StrainSelect database, a collection of high-quality sequence and annotation data derived from bacterial and archaeal strains that can be obtained from an extant culture collection (http://secondgenome.com/StrainSelect) (publication in preparation), and conducts *de novo* clustering of all sequences without strain hits using the UPARSE algorithm (SS-UP). For SS-UP, Illumina paired-end sequenced reads were merged using USEARCH fastq_mergepairs with default settings except for dataset-specific cutoffs for fastq_minmergelen and fastq_maxmergelen (Supplementary Table 1). All resulting merged sequences were compared against StrainSelect v2014-02-20 using USEARCH's usearch_global. 454 single-end reads were first quality trimmed from the N-terminal end using PrinSeq-lite [7] and parameters '-trim_ns_left 1 -trim_ns_right 1 -min_len $MIN_LEN -trim_qual_right 20' (minimal length values per dataset are summarized in Supplementary Table 1) before comparison to StrainSelect using USEARCH's usearch_global. Distinct strain matches were defined as those with ≥ 99% identity to a 16S sequence from the closest matching strain and a lesser identity (even by one base) to the second closest matching strain. Those distinct hits were summed per strain and a strain-level OTU abundance table was created. The remaining sequences were filtered by overall read quality using USEARCH's fastq_maxee and a MAX_EE value of 1, length-trimmed to the lower boundary of the 95% interval of the read length distribution (for datasets with an uneven read length distribution length-trimming to the shortest read length is strongly affected by very short reads; the 95% interval is used to compensate for this outlier effect), de-replicated, sorted descending by size and clustered at 97% identity with USEARCH (fastq_filter, derep_fulllength, sortbysize, cluster_otus). USEARCH cluster_otus discards likely chimeras. A

representative consensus sequence per *de novo* OTU was. For each study, *de novo* OTUs with abundance of less than 3 in a study were discarded as spurious. All sequences that went into the comparison against StrainSelect but did not end up in a strain OTU were then mapped to the set of representative consensus sequences (≥97% identity) to generate a *de novo* OTU abundance table. Representative strain-level OTU sequences and representative *de novo* OTU sequences were assigned a Greengenes [3] taxonomic classification via mothur's bayesian classifier [8] at 80% confidence; the classifier was trained against the Greengenes reference database (version 13_5) of 16S rRNA gene sequences. Both Greengenes version 13_5 used for SS-UP and version 13_8 used for QIIME-CR contain the same set of reference sequences. In the 13_8 version, additional taxonomic terms were manually curated, but the reference OTUs and phylogenetic trees remained unchanged. Where standard taxonomic names have not been established, a hierarchical taxon identifier was used (for example "97otu15279"). Strain-level OTU abundances and taxonomy-mapped *de novo* OTU abundances from all studies were merged and used for further analysis. The SS-UP approach allowed all high-quality sequences to be counted, and the taxonomic classification of the *de novo* OTUs permitted *de novo* OTUs with conserved taxonomy to be compared across studies.

Samples with < 100 sequences after quality filtering and OTU assignment for either bioinformatics pipeline were excluded from both all further analysis. In all cases, any sample that had <100 sequences in one pipeline had <100 sequences in the other.

Statistical Analysis

The R package phyloseq was used for determining global community properties such as alpha diversity, beta diversity metrics such as the Bray-Curtis and Jaccard index, principle coordinate scaling of Bray-Curtis dissimilarities, Firmicutes/Bacteroidetes (F/B) ratio and differential abundance analysis. Two-sample permutation t-tests using Monte-Carlo resampling were used to compare the alpha diversity estimates and F/B ratio across CRC and controls and CRA and controls. Permutational analysis of variance (PERMANOVA) was used to test whether within group distances were significantly different from between group distances using the *adonis* function in the vegan package. Multivariate homogeneity of group dispersions was tested with vegan using the

betadisper function. Differential abundance of QIIME OTUs and SS-UP OTUs across CRC cases and controls was evaluated adjusting for *Study* as a confounding factor in the DESeq2 design (~ Study + disease status). OTUs were considered significantly different if their False Discovery Rate (FDR) adjusted Benjamin Hochberg (BH) p value was <0.1 and estimated $\log_2$-fold change was > 1.5 or < -1.5.

The Random Effects model (REM) considered the eight studies with CRC-control samples as a sample of a larger number of studies and inferred the likely outcome if a new study were performed. The CRC-fecal microbiome studies were dissimilar in terms of their methods as well as patient demographics. These differences may introduce heterogeneity among true effects. The RE model treats this heterogeneity as random. Specifically, in addition to the pooled analysis mentioned above we estimated study by study DESeq2 $\log_2$ fold changes as effect size estimates and the standard error associated with them as corresponding sampling variances as an input for the REM. OTUs that occurred as differentially abundant by DESeq2 in at least 5 studies (i.e 5 or 6 or 7 or 8 studies) for the CRC vs control comparison and either 3 or 4 studies for the CRA vs control comparison were retained for the analysis. The resulting RE model p-values were FDR corrected for multiple comparisons across taxa OTUs and forest plots were plotted for significant OTUs. We also plotted relative abundances of these OTUs across several studies to estimate how the log fold changes in cases as compared to controls reflected in the prevalence of the actual OTUs.

To determine the predictive power of microbial taxa For the random forest classifier, the number of predictor features randomly sampled for splitting at each node in the decision tree commonly known as *mtry* was tuned as (0.5, 1, 1.5, 1.75, 2, 2.5, 3.0)*(square root of total number of microbial predictors). Models were internally cross-validated ten-fold times with five repeats to avoid over-fitting. Tuning area under receiver operating characteristic (AUROC) curve with the largest value was used to select the optimal model. RF models to predict disease outcome were built for clinical markers only (for studies where clinical metadata was available (n= 3 studies, 156 samples)), microbial markers only (for all samples and studies (n= 8 studies, 344 samples) as well as the subset of samples for which complete clinical metadata was available n=3 studies, 156

samples)), and a combination of both clinical and microbial markers (n= 3 studies, 156 samples). Continuous variables among the clinical metadata such as age and BMI were centered and scaled prior to building the RF models. To estimate if any particular study disproportionately affected the optimal AUROC value of the classifier, we conducted a *leave one study out* analysis and estimated the classifier accuracy after each study was omitted. We also determined classifiers for individual studies to compare how the composite classifier fared with homogenously processed features from individual studies. Recursive feature elimination using 10 fold cross-validation with five repeats was used to identify the most informative microbial taxa for classification using the rfe function. To determine the generalizability of the composite microbial biomarker, the leave one study out cohort (test set) classifier was used to predict the disease outcome in the study that was left out (validation set) using the predict.train function. ROC's were plotted for the above models using the pROC package. [9] Differences in the AUROC were tested statistically with DeLong's test within the package.

References

1. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nature methods 2010;**7**(5):335-6 doi: 10.1038/nmeth.f.303[published Online First: Epub Date]|.
2. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England) 2010;**26**(19):2460-1 doi: 10.1093/bioinformatics/btq461[published Online First: Epub Date]|.
3. DeSantis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology 2006;**72**(7):5069-72 doi: 10.1128/aem.03006-05[published Online First: Epub Date]|.
4. Mira-Pascual L, Cabrera-Rubio R, Ocon S, et al. Microbial mucosal colonic shifts associated with the development of colorectal cancer reveal the presence of different bacterial and archaeal biomarkers. J Gastroenterol 2015;**50**(2):167-79 doi: 10.1007/s00535-014-0963-x[published Online First: Epub Date]|.
5. Wang T, Cai G, Qiu Y, et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. The ISME journal 2012;**6**(2):320-9 doi: 10.1038/ismej.2011.109[published Online First: Epub Date]|.
6. Wu N, Yang X, Zhang R, et al. Dysbiosis signature of fecal microbiota in colorectal cancer patients. Microbial ecology 2013;**66**(2):462-70 doi: 10.1007/s00248-013-0245-9[published Online First: Epub Date]|.
7. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics (Oxford, England) 2011;**27**(6):863-4 doi: 10.1093/bioinformatics/btr026[published Online First: Epub Date]|.
8. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology 2009;**75**(23):7537-41 doi: 10.1128/aem.01541-09[published Online First: Epub Date]|.
9. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics 2011;**12**:77 doi: 10.1186/1471-2105-12-77[published Online First: Epub Date]|.