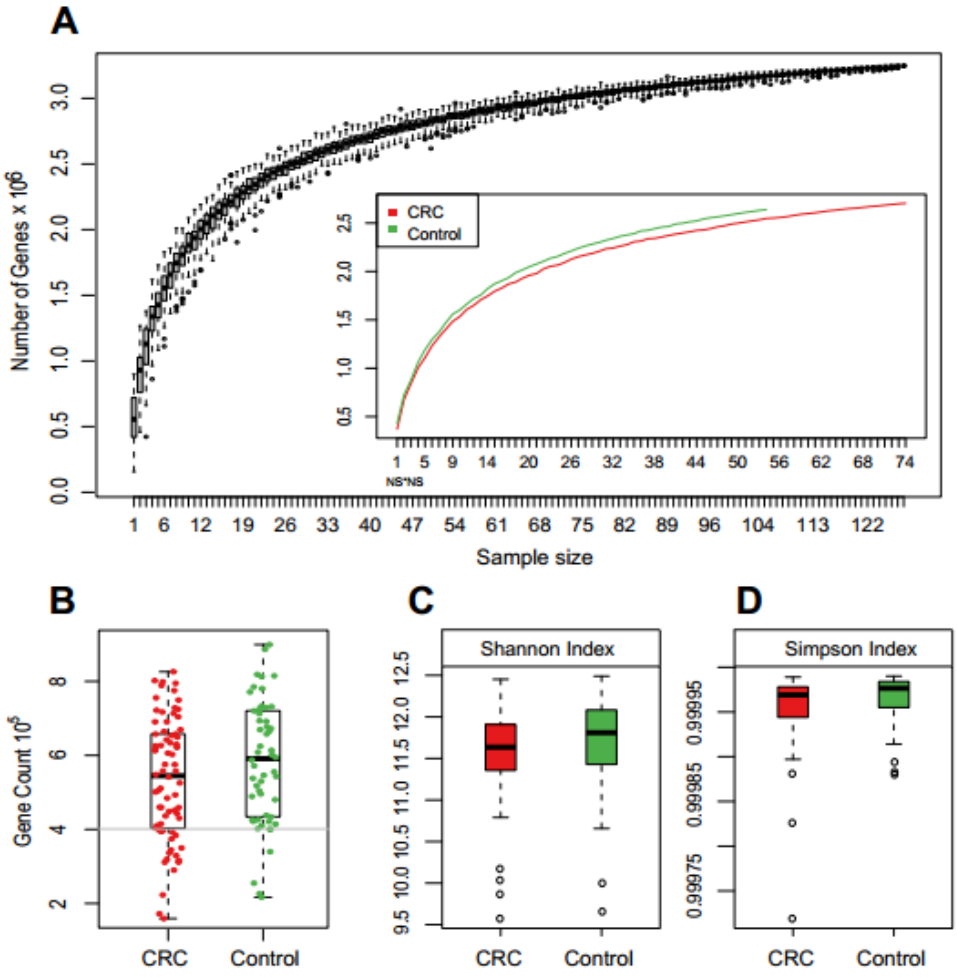
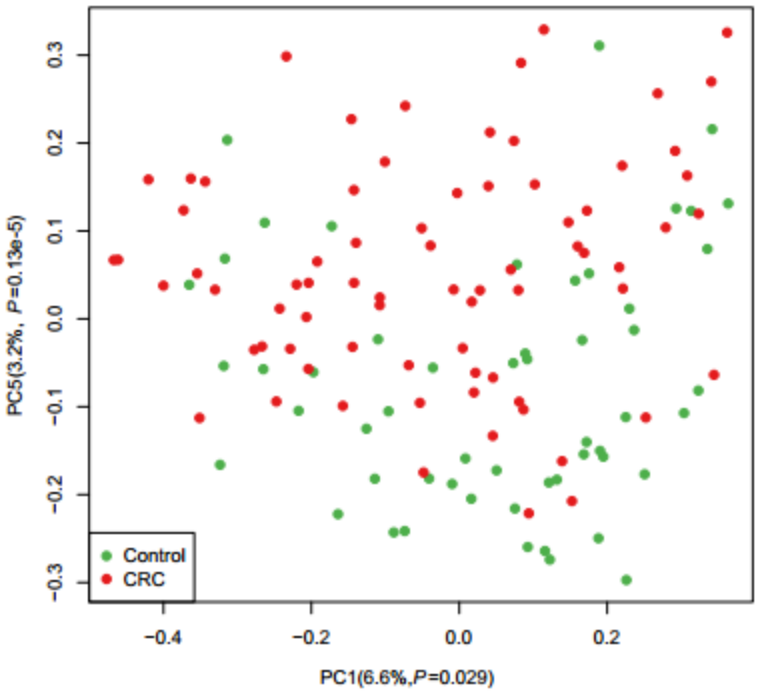
1. Supplementary Figures



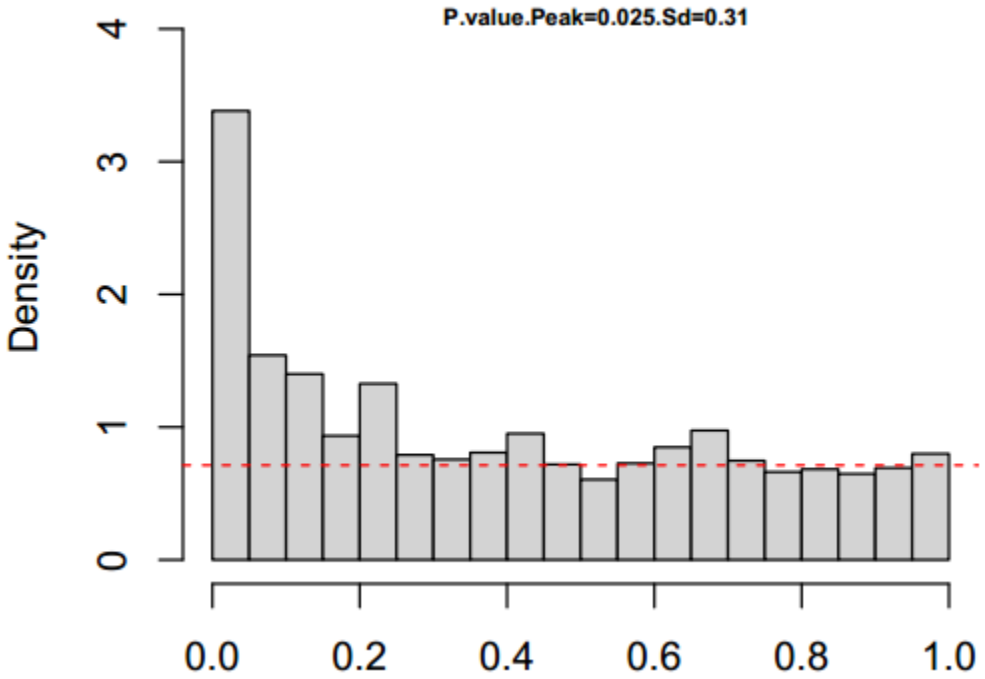
Supplementary Figure 1. Microbial dysbiosis in colorectal cancer.

(A) Rarefaction curves showing cumulative number of genes sequenced with increasing sample size. The numbers are close to saturation given the current sample sizes for all 128 samples (grey). Inset: CRC samples (red) have significantly lower gene counts compared to healthy controls (green). Only the first three points corresponding to sizes of up to 3 samples each are not significant (NS). **(B)** Gene count distribution of 128 CRC case and healthy control individuals. The grey line shown corresponds to 400,000 genes, below which 80% individuals had cancer. **(C)**The Shannon index of the CRC case and healthy control microbiomes from gene abundances. **(D)** Simpson index of the CRC case and healthy control microbiomes from gene abundances.



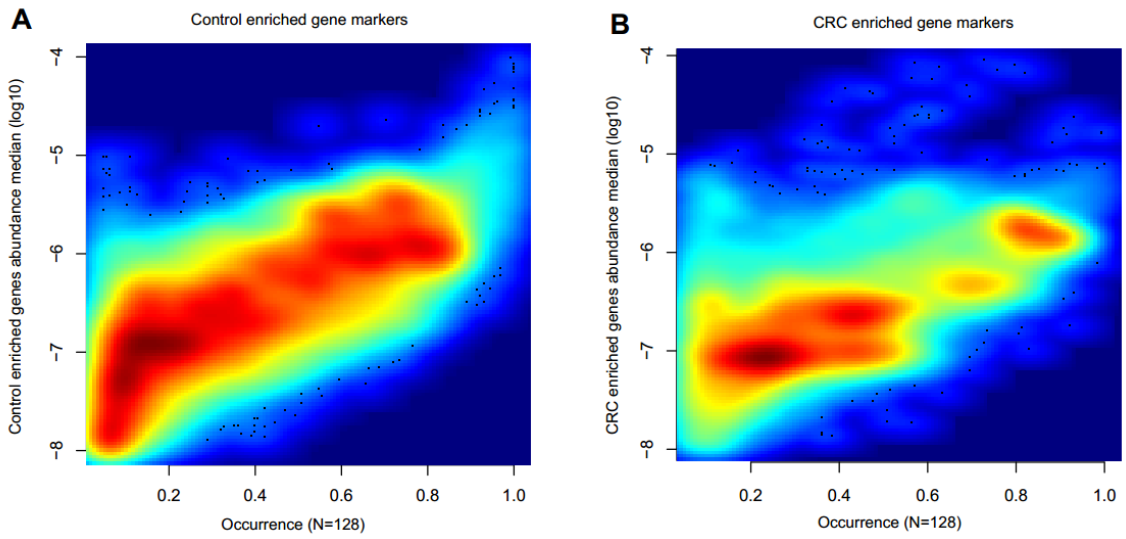
Supplementary Figure 2. Principal component analysis using 2,110,489 genes identified in cohort C1.

PC1 and PC5 are associated with CRC status (see Supplementary Table S5), but do not show separation between gut microbiomes of CRC patients and control individuals based on gene profiles. See **Fig. 2A** where a strong separation is observed.



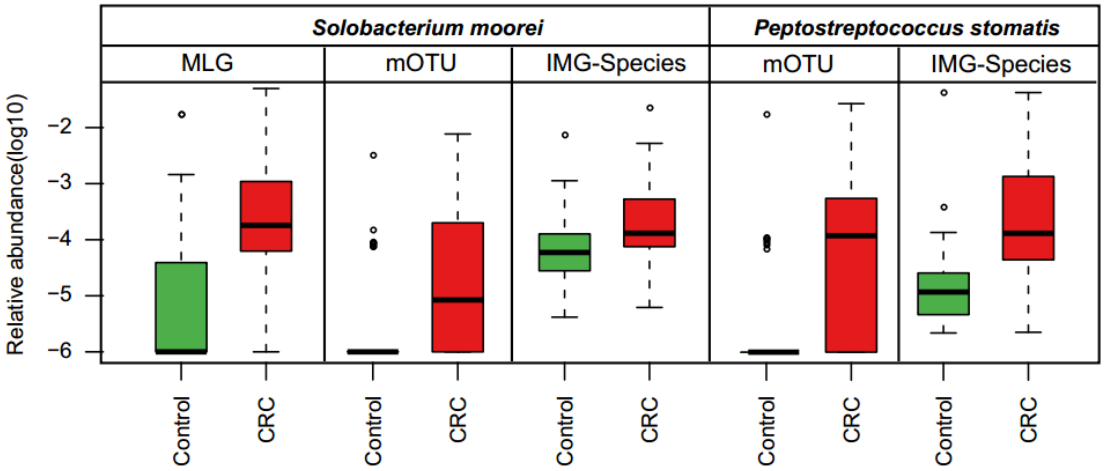
Supplementary Figure 3. Distribution of *P*-value association statistics of all microbial genes in cohort C1.

The association analysis of CRC p-value distribution identified a disproportionate over-representation of strongly associated markers at lower P-values, with the majority of genes following the expected P-value distribution under the null hypothesis. This suggests that the significant markers likely represent true rather than spurious associations.

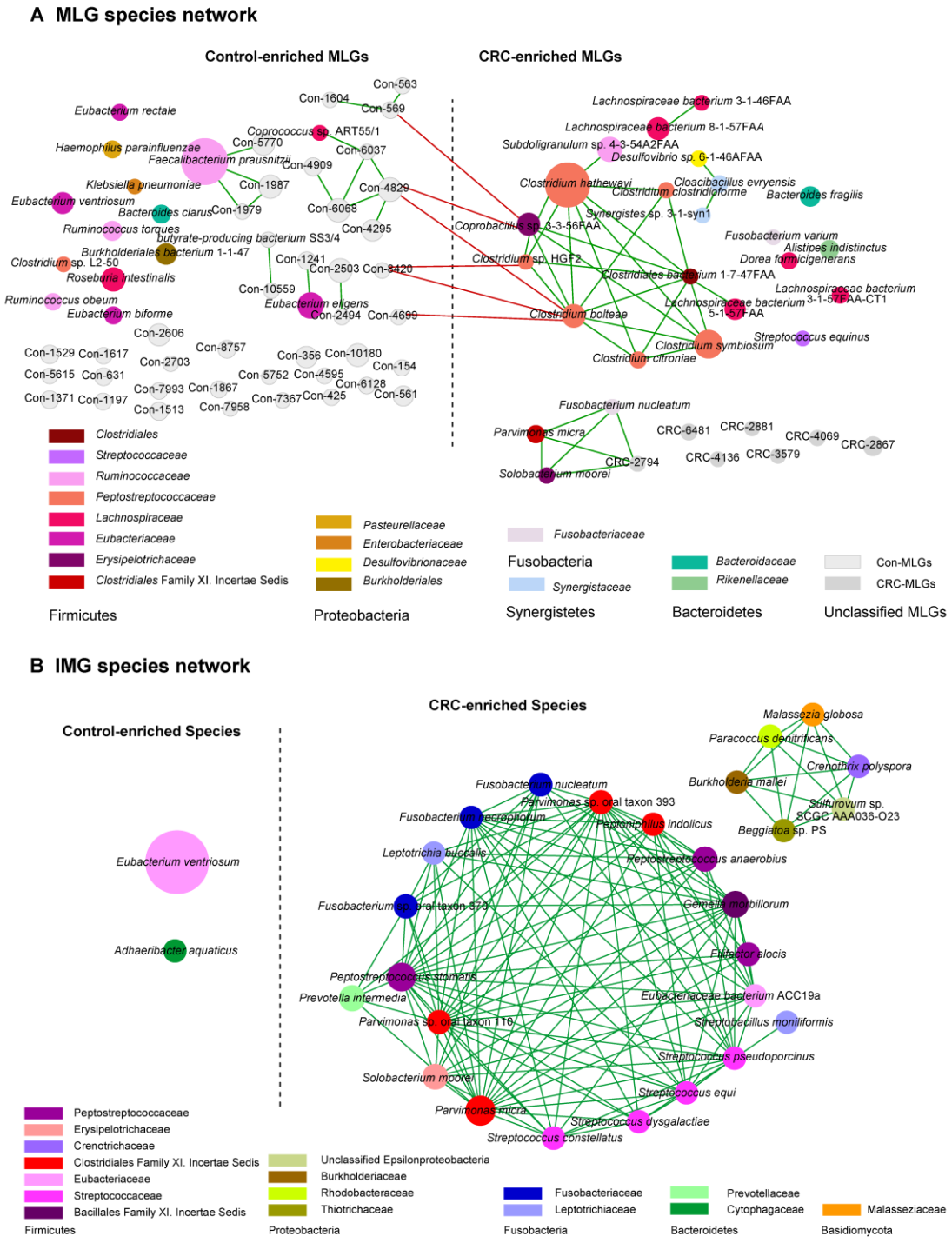


Supplementary Figure 4. 2-dimensional histogram of abundance-vs-occurrence rate of CRC-associated gene markers.

The CRC-associated gene markers selection were based on the significant enrichment in CRC case or healthy control. We computed the occurrence rate and median relative abundance for the CRC-enriched gene markers and control-enriched gene markers in all 128 samples from C1, and generated a 2-dimensional histogram following previously described methods[25] to show the distribution of all marker genes. **(A)** Control-enriched gene markers exhibit a wider range of occurrence rate and relative abundance. **(B)** CRC-enriched gene markers are mostly present in low occurrence rate and low relative abundance.

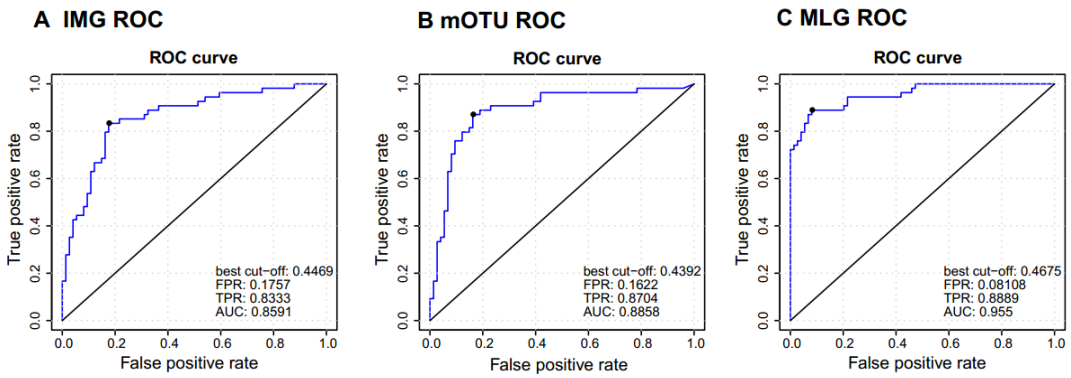


Supplementary Figure 5. Enrichment of *Solobacterium moorei* and *Peptostreptococcus stomatis* in CRC patient microbiomes in cohort C1.



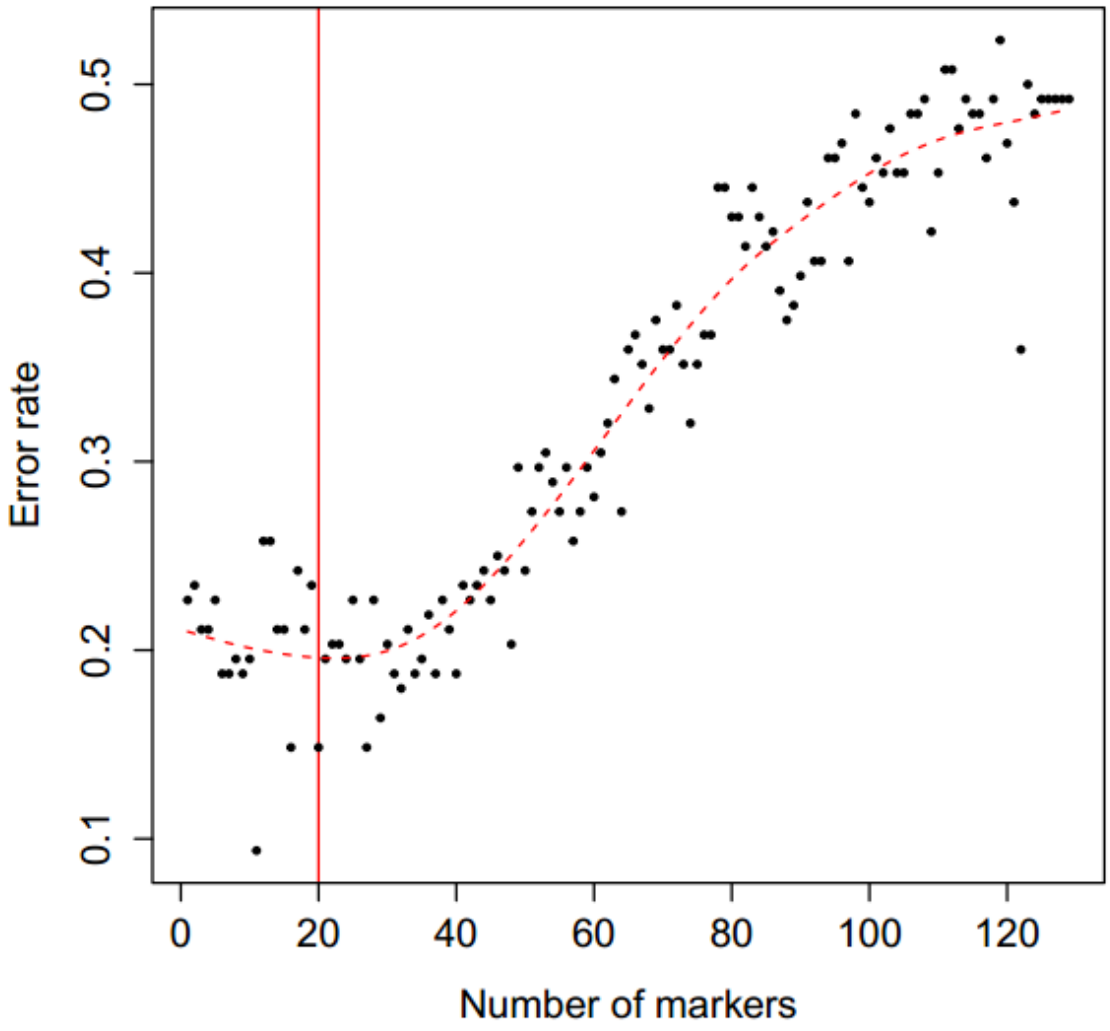
Supplementary Figure 6. Significant species network using IMG version 400 and MLG annotation with q-value<0.05 in cohort C1.

(A) A co-occurrence network deduced from 85 MLGs significantly associated with CRC. Each MLG with more than 100 genes and at least 50% genes annotated to a single species was annotated with species name. The remaining MLGs were named Con or CRC MLGs according to their enrichment in control and CRC samples, respectively. Species are rearranged in two sides based on their enrichment in CRC or healthy microbiomes. Spearman correlation coefficient values lower than -0.5 (negative correlation) are indicated as red edges, and coefficient values higher than 0.5 (positive correlation) are indicated as green edges. Node size indicates the number of genes within the MLG, and node color shows their taxonomic annotation. **(B)** A co-occurrence network deduced from 28 IMG species significantly associated with CRC. Node size indicates the average of relative abundance for each species. See legend for panel **A** for other details.



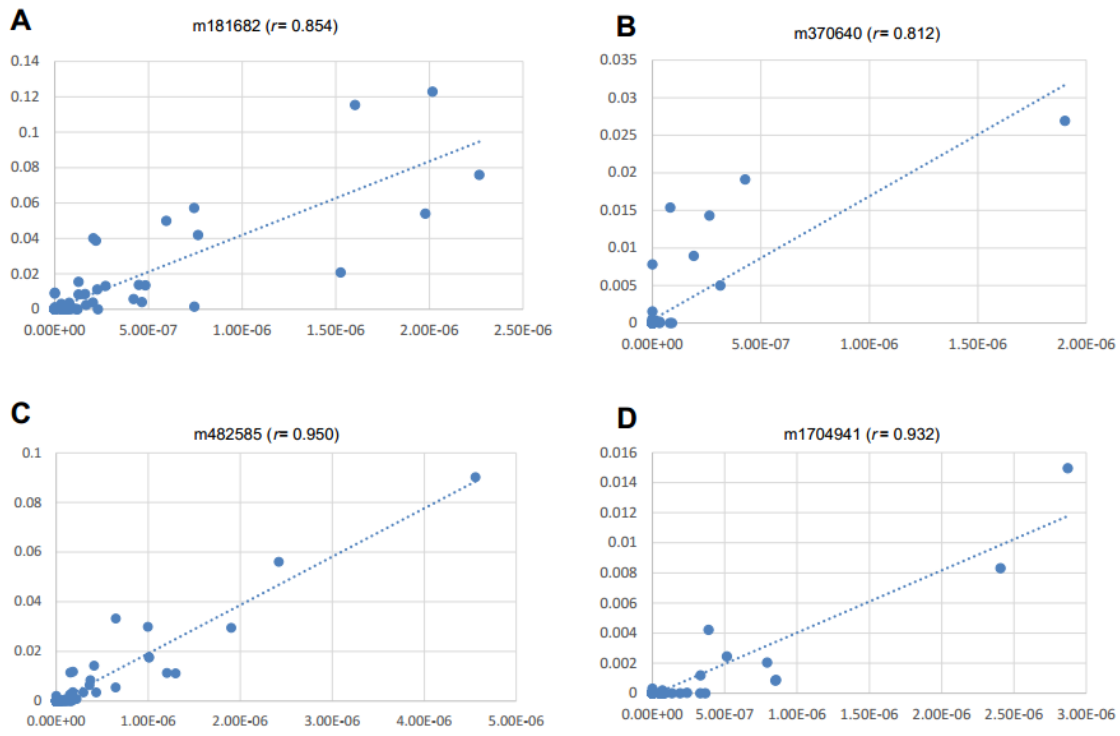
Supplementary Figure 7. The Receive-Operator-Curve of CRC specific species marker selection using random forest method and three different species annotation methods.

A, IMG species annotation using clean reads to IMG version 400. **B,** mOTU species using published methods[26], **C**, All significant genes clustered using MLG methods[25] and the species annotation using IMG version 400.

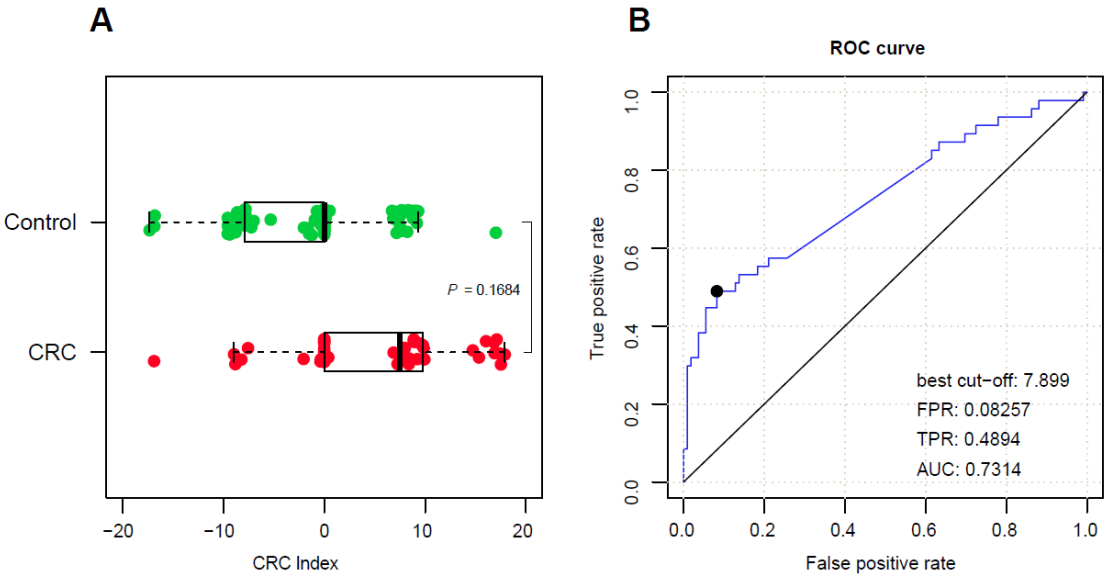


Supplementary Figure 8. Minimum redundancy maximum relevance (mRMR) method to identify 20 gene markers that differentiate CRC cases from controls in cohort C1.

Incremental search was performed using the mRMR method which generated a sequential number of subsets. For each subset, the error rate was estimated by a leave-one-out cross-validation (LOOCV) of a linear discrimination classifier. The optimum subset with the lowest error rate contained 20 gene markers.



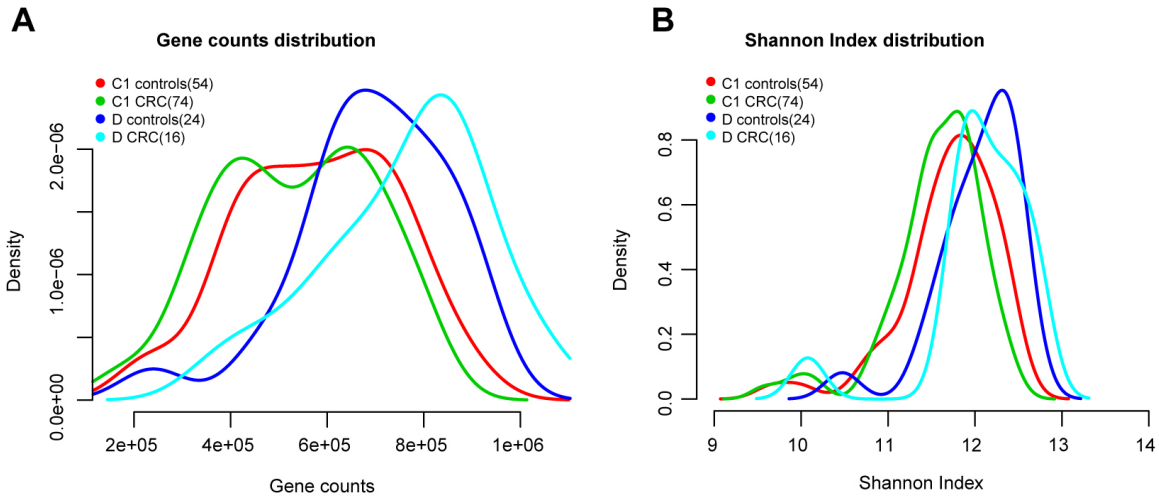
Supplementary Figure 9. Correlation between quantification by the metagenomic approach versus quantitative polymerase chain reaction (qPCR) for four gene markers.



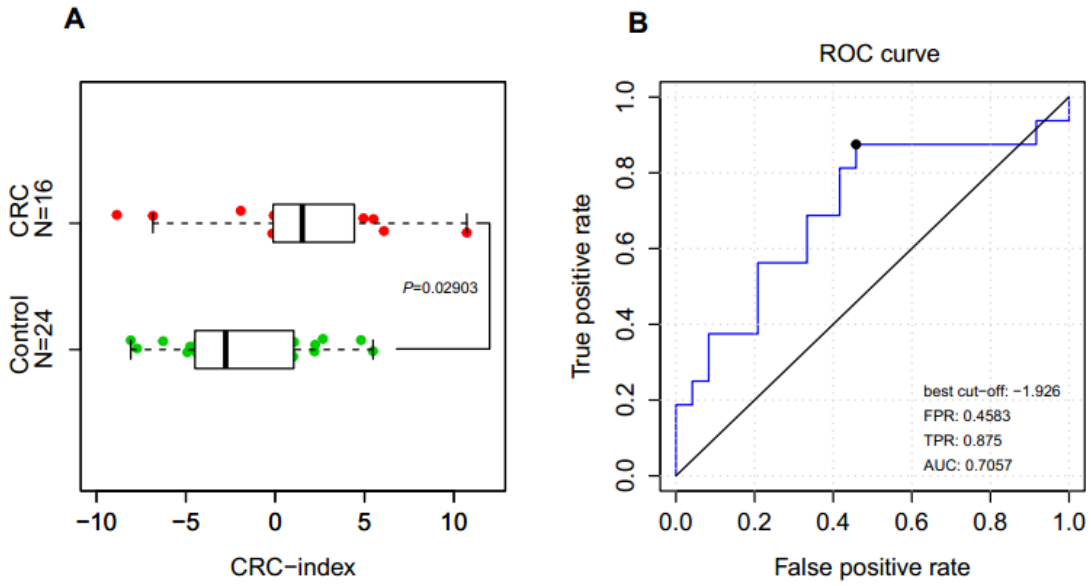
Supplementary Figure 10. Evaluating CRC index from four markers in Chinese

cohort C2 of 156 individuals.

(A) CRC index based on qPCR abundance of 4 gene markers shows marginal separation of CRC and control microbiomes. **(B)** ROC analysis reveals moderate potential for classification using CRC index, with an area under the curve of 0.73.



Supplementary Figure 11. Comparison of (A) gene richness (gene count) and (B) alpha-diversity (Shannon index) distribution in cohorts C1 and D.



Supplementary Figure 12. Evaluating CRC index in cohort D consisting of 40 individuals.

(A) CRC index based on 20 gene markers shows marginal separation of CRC and control microbiomes. **(B)** ROC analysis reveals moderate potential for classification using CRC index, with an area under the curve of 0.71.