MetaFunc: Taxonomic and Functional Analyses of High Throughput Sequencing for Microbiomes

Arielle Kae Sulit[1], Tyler Kolisnik[2], Frank A Frizelle[3], Rachel Purcell[3], Sebastian Schmeier[2]
[1]*School of Natural and Computational Sciences, Massey University, Auckland, New Zealand (a.sulit@massey.ac.nz)*
[2]*School of Natural and Computational Sciences, Massey University, Auckland, New Zealand*
[3]*Department of Surgery, University of Otago, Christchurch, New Zealand*

Abstract

Background: The identification of functional processes taking place in microbiome communities augment traditional microbiome taxonomic studies, giving a more complete picture of interactions taking place within the community. While there are applications that perform functional annotation on metagenome or metatranscriptomes, very few of these are able to link taxonomic identity to function and are limited by their input types or databases used.

Results: Here we present MetaFunc, a workflow which takes input reads, and from these 1) identifies species present in the microbiome sample and 2) provides gene ontology (GO) annotations associated with the species identified. MetaFunc can also provide a differential abundance analysis step comparing species between sample conditions. In addition, MetaFunc allows mapping of reads to a host genome, and separates these reads, before proceeding with the microbiome analyses. From the host reads, MetaFunc is able to identify host genes, perform differential gene expression analysis, and gene-set enrichment analysis. A final correlation analysis between microbial species and host genes can also be performed. Finally, MetaFunc builds an R shiny application that allows users to view and interact with the microbiome results. In this paper we show how MetaFunc can be applied to metatranscriptomic datasets of colorectal cancer.

Conclusion: MetaFunc is a one-stop shop microbiome analysis pipeline that can identify taxonomies and their respective functional contributions in a microbiome sample through GO annotations. It can also analyse host reads in a microbiome sample, providing information on host gene expression, and allowing for correlations between the microbiome and host genes. MetaFunc comes with a user-friendly R shiny application that allows for easier visualisation and exploration of its results. MetaFunc is freely available through https://gitlab.com/schmeierlab/workflows/metafunc.git.

Keywords: *metatranscriptomics, microbiome, function*

Introduction

Metagenomic or metatranscriptomic studies of microbiome communities allow for characterization of functional contributions as well as taxonomic load, by allowing the identification and quantification of genes possibly contributed by the microbial community. The ability to identify functional processes from the microbiome gives a more complete picture of microbe-microbe and/or microbe-host interactions that drive community dynamics (Langille, 2018).

There are existing bioinformatics programs (Nayfach et al., 2015; Sharma et al., 2015; Silva et al., 2016) that perform functional annotation on metagenomes and metatranscriptomes, but most of these are unable to link taxonomies (the microbes under study) to their respective functional processes. Existing packages with this capacity include PICRUSt and PICRUSt2 (Douglas et al., 2019; Langille et al., 2013), and HUMAnN2 (Franzosa et al., 2018). PICRUSt and PICRUSt2 predict metagenome function by inferring genes present in OTUs based on their phylogenetic similarities to other OTUs with known gene content (Douglas et al., 2019; Langille et al., 2013). However, they do not directly measure the genes involved,

but rather rely on 16S gene marker sequences, which, being highly conserved, are useful for the identification of bacterial genera (Bashiardes et al., 2016; Ternes et al., 2020) and are not present in other microbes aside from Bacteria and Archaea (Ye et al., 2019). Thus 16S based taxonomic identification, and subsequent functional predictions, may be unsuitable for species level identification, and for recognizing other microbes aside from Bacteria and Archaea. HUMAnN2's taxonomic profiling, meanwhile, is reliant on MetaPhlAn2 (Segata et al., 2012; Truong et al., 2015), which uses clade-specific marker genes from reference genomes. Benchmarking efforts by Ye et al., 2019 highlights the limitations of using the MetaPhlAn2 package that includes non-customizable databases and marker gene-based databases, which results in relatively lower precision and recall in its classification.

To augment such meta-omic studies, we present here a simple, straight-forward pipeline named MetaFunc, a snakemake workflow (Köster and Rahmann, 2012) that maps function to a microbiome (and optionally host) sample. MetaFunc uses Kaiju (Menzel et al., 2016) as its main taxonomic classifier. Kaiju uses protein translations of input reads to generate taxonomic profiles. By generating protein-based classifications using metatranscriptomic reads, MetaFunc is able to determine which microbes are more metabolically active, allowing more focus on the functional contributions of microbes. MetaFunc then uses protein accession numbers from Kaiju results to obtain the set of Gene Ontology (GO) terms associated with the microbiome community. Furthermore, Kaiju outputs provide a direct protein – taxonomy ID relationship that makes it possible for MetaFunc to establish which organisms are contributing to the functional GO terms. MetaFunc also has options for pre-processing of reads before running Kaiju; trimming of input reads with fastp (Chen et al., 2018) can be performed in addition to pre-mapping to a host genome (e.g. human) using STAR (Dobin et al., 2013). The unmapped reads following STAR processing are the input

used by MetaFunc for microbe identification, while host gene expression information can be obtained from STAR-mapped reads. Thus, MetaFunc allows simultaneous investigation of host and microbe community active functional processes as well as active host genes and microbes.

Implementation

Workflow

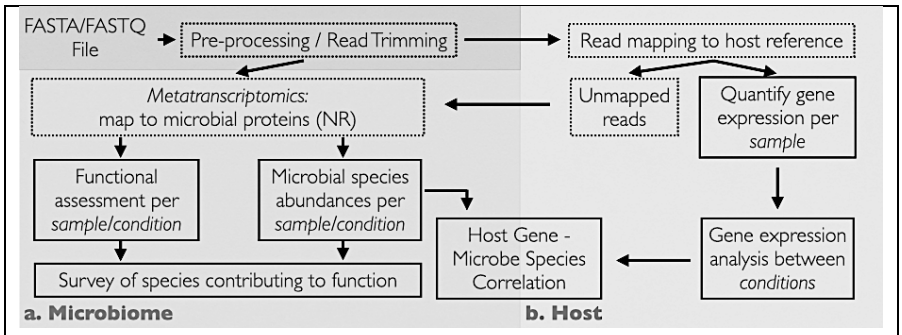


Figure 1. MetaFunc Workflow. The workflow uses FASTQ or FASTA as input and processes reads through the microbiome pipeline to give microbial abundance and function (**a**) and/or host gene analysis (**b**) which will first map reads to a host before sending unmapped reads to the microbiome pipeline. Applying host read analysis will give gene expression analysis results as well as host gene-microbial species correlation. Solid boxes indicate steps with an output while dotted boxes indicate intermediate steps in the pipeline.
NR: NCBI Blast nr database

Figure 1 shows the workflow that takes place within MetaFunc. Paired-end and/or single-end sequencing reads are used as input in fasta or fastq format. If trimming and mapping are not enabled, reads are used as input to Kaiju and subsequent microbiome analyses (Figure 1a). If trimming is enabled, reads are trimmed for adapters and quality controls using fastp. If mapping is enabled, either the trimmed reads or raw input reads are first mapped to a designated host genome using STAR. Unmapped reads after host mapping are then used as input to Kaiju. STAR results are then used to obtain host gene information (Figure 1b).

Microbiome Analysis (Figure 1a)

MetaFunc parses through Kaiju results and gathers taxonomy IDs of species for taxonomic characterization per sample and their corresponding protein accession numbers, which are subsequently annotated with gene ontology terms.

Taxonomy

Each read matches to a taxonomy ID in Kaiju. MetaFunc gathers the species level matches and adds up the raw reads matching to each species taxonomy ID. In cases of strain level identification, MetaFunc adds this count to its parent species. It also obtains scaled read counts in percentages by dividing the final read count of each taxonomy ID by the total reads that have mapped to species level taxonomies. For a dataset, the pipeline removes any taxonomy ID that is less than 0.001% in abundance in all samples of the dataset; this filter removes thousands of species that are likely to be false positives while retaining more confident classifications. Any remaining false classifications are thought not to affect downstream analyses, as the levels would be too low to impact true abundance (Ye et al., 2019), however, this value can be adjusted by the user. The taxonomy IDs that have passed the cutoff are then used in subsequent analyses. It should be noted that the pipeline still uses the original scaled percent abundances even after filtering.

For a dataset, the MetaFunc pipeline outputs two tables containing species as rows and samples as columns with values being raw read counts or percent abundance for each species in the samples. If the user wishes to compare groups or conditions (e.g. disease state vs control), the pipeline calculates the average percent abundance of species among samples belonging to a group and this table is also given as an output. Differential abundance of microbes between groups is also carried out in MetaFunc using edgeR

(McCarthy et al., 2012; Robinson et al., 2010). Raw read count tables are first filtered using the function *filterbyExpr* with threshold of 1 which is user-adjustable, and normalization factors are calculated by *calcNormFactors* with default settings. *exactTest* is then applied to calculate differential abundance with *p*-values adjusted using FDR.

Proteins

Kaiju outputs the accession number(s) of the protein matches with the highest BLOSUM62 alignment score of the read after translation into 6 open reading frames (ORF). It is possible to have more than one best protein match if two or more protein matches have equal scores in Kaiju. In order to account for this, we use proportional read counts per protein accession number where one read is divided by the number of best protein matches it has. Similar to that for taxonomy IDs, the pipeline adds up the proportional read counts per protein accession number of a species. Scaled reads as percent abundances are obtained by dividing the proportional count of each accession number by the total read counts that have mapped to a species.

Gene Ontology: Database Construction

Metafunc relies on Kaiju's nr_euk database for its taxonomic identification and corresponding protein matches. The nr_euk database is built on a subset from NCBI BLAST *nr* database containing Archaea, Bacteria , Viruses, and other Microbial Eukaryotes (see https://raw.githubusercontent.com/bioinformatics-centre/kaiju/master/util/kaiju-taxonlistEuk.tsv). Identical sequences in the *nr* database are compiled into 1 entry and Kaiju only outputs the first protein accession number of an entry that has multiple identical

sequences (Menzel et al., 2016). Thus we needed to construct the protein-to-GO database such that all functional terms of any protein compiled in 1 *nr* entry are considered.

To facilitate gene ontology annotations, we constructed an sqlite database in which GO annotations of a protein accession number from Kaiju can be looked up. We first gathered relevant NCBI *nr* database entries, converted all of the proteins of an *nr* entry into UniProt (Huang et al., 2011; The UniProt Consortium, 2017) entries, and then gathered corresponding GO annotations using the Gene Ontology Annotation (GOA) database for all those proteins (Camon et al., 2004). All GO annotations of one *nr* entry are then linked to the first protein of that entry in an sqlite database, which is used to annotate Kaiju protein accession matches with GO IDs. For more detailed information, please see the Notes section of the pipeline's documentation page (https://metafunc.readthedocs.io/en/latest/notes.html). For MetaFunc, we provide pre-made databases for download but users can make their own updated databases following instructions from https://gitlab.com/schmeierlab/metafunc/metafunc-nrgo.git.

Gene Ontology: Protein Annotation

For each sample, the pipeline obtains only the proteins that are from taxonomy IDs that passed cutoffs in the *Taxonomy* section described above. Their scaled proportional read counts, as in the *Proteins* section above, are still scaled against the total number of reads that mapped to a species. In order to compare groups or conditions, the pipeline first calculates the average of the corresponding proportional reads and scaled proportional reads of a protein accession number among samples of a group. It then searches for the GO terms annotating the (*nr*) protein using the created sqlite database described in *Gene Ontology: Database Construction.* Each GO term is then updated by accessing parent terms

related to the GO terms by *'is_a'* or *'part_of'*. The proportional and scaled read counts are then added to all GO terms annotating a protein, including updated terms. Finally, percentage of reads covering a GO term within a namespace (Biological Process, Molecular Function, and Cellular Component) is calculated by dividing the scaled read count of a GO term by the total scaled read counts covering a namespace and multiplying by 100. The final output table of the pipeline is a contingency table with GO IDs of all namespaces as rows and samples or group as columns, with percentage within a namespace as values.

Visualisation

To facilitate exploration of results from MetaFunc, MetaFunc automatically builds an R shiny application, such that users can view and interact with the taxonomy and gene ontology tables. The application allows users to select GO terms and identify the species whose proteins are annotated with the searched for term. Conversely, users may search for a species and obtain all GO terms associated with the searched for species. See the pipeline's documentation page for more information (https://metafunc.readthedocs.io/en/latest/rshiny.html).

Host Analyses (Figure 1b)

Many microbiome communities are often associated with a host genome. Reads belonging to the host genome have the capacity to misclassify as microbiome (Ye et al., 2019) and filtering of host reads has been a part of many microbiome studies, either prior to sequencing or *in silico* (Hugerth and Andersson, 2017; Macklaim and Gloor, 2018; Xia et al., 2018). The MetaFunc pipeline offers the option of mapping reads to a host genome using

the program STAR and using the unmapped reads from this step as input to Kaiju for the microbiome analysis.

MetaFunc also allows additional analyses of host reads after STAR mapping. Host genes are quantified using *featureCounts* (Liao et al., 2014) of the subread package. If comparisons between groups are indicated, edgeR is used to perform differential gene expression analysis (DGEA). Additionally, supplying a gene matrix transposed (*.gmt*) file from e.g. the molecular signatures database (MSigDB) (Liberzon et al., 2011; Subramanian et al., 2005) allows for gene set enrichment analysis (GSEA) of host genes using the *clusterProfiler* package (Yu et al., 2012).

Host Gene – Microbe Species Correlation

When a comparison between groups is specified, the pipeline also performs spearman correlation analysis between the top most significant differentially expressed genes (DEGs) and top most significant differentially abundant (DA) microbes. Results of these correlations are summarized in a matrix on which hierarchical clustering is performed and a heatmap is generated using Clustergrammer (Fernandez et al., 2017). Through this heatmap and table, a user can investigate the strength of correlation (*rho*) between a DA microbe and a DEG, and which microbes and genes have similar patterns of correlations.

For a more detailed description of the workflow, usage instructions, and results, documentation of the MetaFunc pipeline may be found in https://metafunc.readthedocs.io/en/latest/index.html.

Results (Usage Example)

Dataset PRJNA413956: Matched Colorectal Cancer (CRC) and Adjacent Non-Tumour Tissue

In order to demonstrate the utility of the MetaFunc pipeline, we obtained publicly available transcriptomics data from the study of Li et al., 2018 consisting of 10 tumour and corresponding adjacent non-tumour colorectal tissue samples. Raw sequencing data was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104836 and input to the pipeline and the full workflow carried out, generating data for host, microbiome, and host-microbiome correlation.

Microbiome Results

Taxonomy

The MetaFunc pipeline outputs a percent abundance average across members of the same group. We ran the pipeline with the intent of comparing microbiome species and function between colon cancer samples and non-tumour matched samples.

Previous studies have already established that certain microbes associate more with colorectal cancer samples compared to healthy controls. We searched for *Fusobacterium nucleatum*, *Parvimonas micra,* and *Porphyromonas asaccharolytica* in the averaged group results. These microbes have previously been found to be more abundant in colorectal cancer cohorts in meta-analyses of several datasets (Dai et al., 2018; Thomas et al., 2019). We also searched for *Bifidobacterium* species *B. bifidum* and *B. longum*; *Bifidobacteria* are thought to confer protection from colorectal cancer (Wei et al., 2018). Figure 2A shows the average percent abundance of the species between samples from tumour and matched non-tumour tissue as identified through MetaFunc.

As seen on Figure 2A, MetaFunc identified *F. nucleatum, P. micra,* and *P. asaccharolytica* as being relatively more abundant in the CRC group while the *Bifidobacterium* species are relatively more abundant in the normal group.

MetaFunc also has a step that utilizes edgeR to perform differential abundance on per sample species read counts, stratified according to CRC and non-tumour grouping. This resulted in a total of 117 species that are significantly different between the groups (*FDR < 0.05*). There are 59 species upregulated and 58 downregulated in colon cancer samples. Through the MetaFunc results, we identified *Tanerella forsythia* as the most prominent enriched species in the colon cancer cohort with a *logFC = 7.40*. *T. forsythia* is a known oral pathogen, thought to be part of the so-called Red complex of periodontal pathogens, along with *Porphyromonas gingivalis*, and *Treponema denticola (Malinowski et al., 2019)*. Members of this Red Complex have been found to be enriched in subtype CMS1 of colorectal cancers (Purcell et al., 2017), the subtype most associated with immune process activation in CRC (Dienstmann et al., 2017; Guinney et al., 2015; Inamura, 2018).

Microbiome Function

MetaFunc is intended to enable comparisons of the functional potential of the microbiome between groups. MetaFunc uses gene ontology annotations of protein matches from Kaiju. To demonstrate, we focused on polyamine biosynthetic processes GO terms. Polyamines (PAs) are polycations found to play important biological functions in cell growth. These molecules have been found to be associated with tumour progression and growth (Gerner and Meyskens, 2004; Soda, 2011; Tofalo et al., 2019). Although cells are able to biosynthesize polyamines and even export them, a large source of cellular polyamines comes from uptake from their surroundings and, importantly, the microbiota is thought to

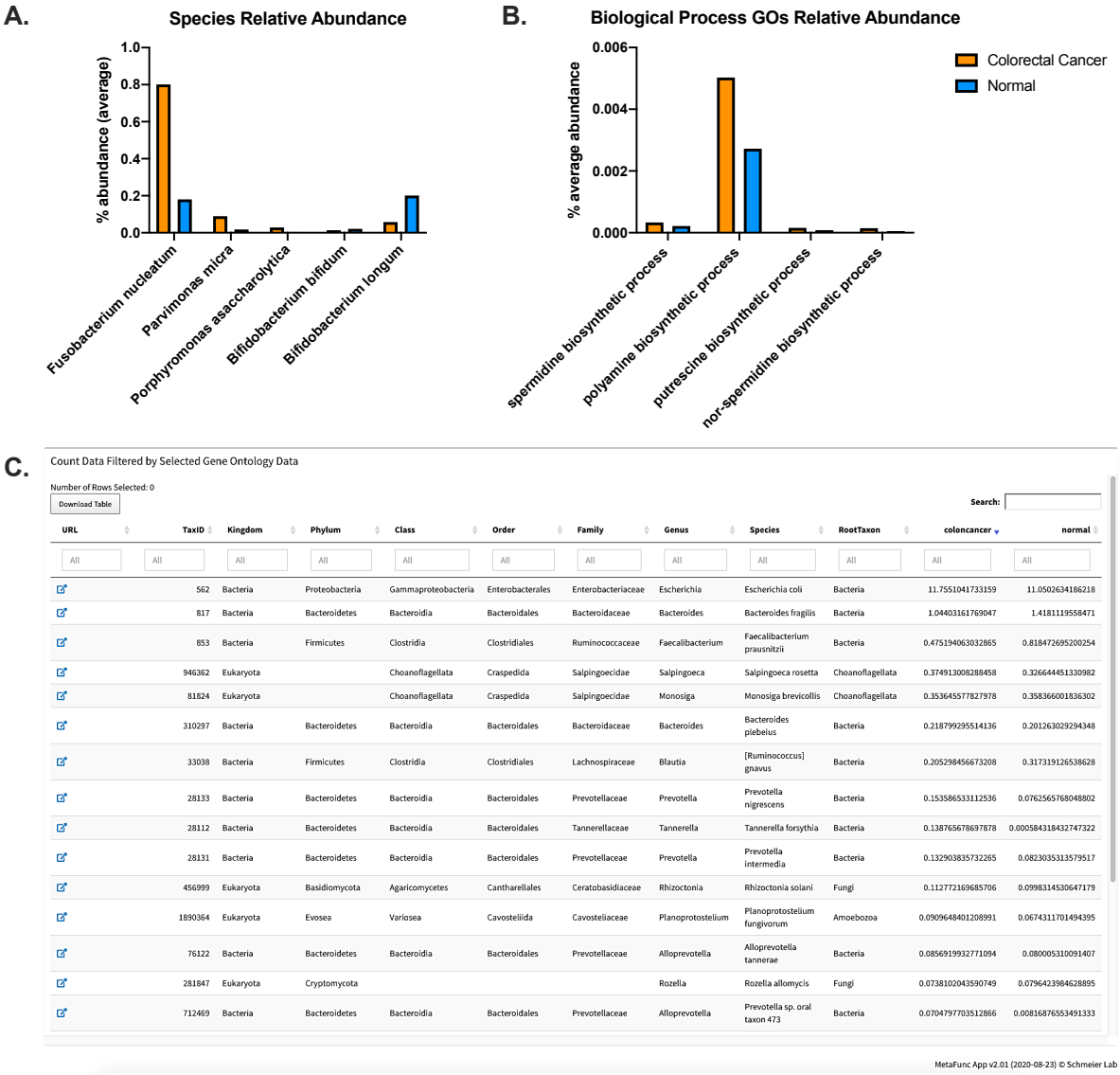
be an essential source (Soda, 2011; Thomas et al., 2019; Tofalo et al., 2019) with spermidine and putrescine being the most common of bacterial PAs (Tofalo et al., 2019). Figure 2B shows the percent reads among biological process GOs covering PA biosynthetic processes in the Colorectal Cancer and Normal conditions. From Figure 2B, we see that several of the polyamine biosynthetic processes are relatively more abundant in the colorectal cancer cohort compared to the normal cohort, using protein annotations.

We use the built-in MetaFunc shiny application to facilitate an inquiry into the microbes species that may contribute to polyamine syntheses. To illustrate, we searched for *'polyamine biosynthetic process'* in the 'GO to TaxIDs' tab of the application, and obtained a total of 126 TaxIDs contributing to the GO term in both colorectal cancer and normal samples. Of these TaxIDs, we identify *E.coli* and *B. fragilis* to be most abundant in both cohorts. However,differences in relative abundance of some microbial species can be identified between cancer and normal cohorts, notably several of which are oral pathogens from the genus *Prevotella*. A striking difference in abundance can be seen in *Tannerella forsythia,* which was previously found to be significantly more abundant in the colorectal cancer cohort via edgeR (Figure 2C). These data suggest that *T. forsythia* represents one of the bacterial species that most contributes to increased polyamine synthesis in CRC samples in this cohort.

Host

The dataset we used for this study was from a total RNA trancriptomics run aiming to identify long non-coding RNAs (lncRNAs) and mRNAs in colorectal cancer samples (Li et al., 2018). Therefore, we first mapped the reads to the human genome using the STAR mapping utility of the pipeline, subsequently using only the unmapped reads for the microbiome

analyses. From the reads mapped to the human genome, MetaFunc was able to obtain counts of reads covering human genes and using these, obtained differentially expressed genes between CRC and matched normal samples through edgeR. MetaFunc results showed a total of 1476 differentialy expressed genes with an $FDR < 0.05$ and $|\log_2 \text{fold change}|>2$. From these, we found all the top 5 upregulated and top 5 downregulated genes as reported in the source publication (Li et al., 2018). Figure 2D shows their fold change as found through MetaFunc.



URL	TaxID	Kingdom	Phylum	Class	Order	Family	Genus	Species	RootTaxon	coloncancer	normal
	562	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Escherichia	Escherichia coli	Bacteria	11.7551041733159	11.0502634186218
	817	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides fragilis	Bacteria	1.044031617690 47	1.4181119558471
	853	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium	Faecalibacterium prausnitzii	Bacteria	0.475194063032865	0.818472695200254
	946362	Eukaryota		Choanoflagellata	Craspedida	Salpingoecidae	Salpingoeca	Salpingoeca rosetta	Choanoflagellata	0.374913008288458	0.326644451330982
	81824	Eukaryota		Choanoflagellata	Craspedida	Salpingoecidae	Monosiga	Monosiga brevicollis	Choanoflagellata	0.353645577827978	0.358366001836302
	310297	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides plebeius	Bacteria	0.218799295514136	0.201263029294348
	33038	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia	[Ruminococcus] gnavus	Bacteria	0.205298456673208	0.317319126538628
	28133	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	Prevotella nigrescens	Bacteria	0.153586533112536	0.0762565768048802
	28112	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Tannerellaceae	Tannerella	Tannerella forsythia	Bacteria	0.138765678697878	0.000584318432747322
	28131	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	Prevotella intermedia	Bacteria	0.132903835732265	0.0823035313579517
	456999	Eukaryota	Basidiomycota	Agaricomycetes	Cantharellales	Ceratobasidiaceae	Rhizoctonia	Rhizoctonia solani	Fungi	0.112772169685706	0.0998314530647179
	1890364	Eukaryota	Evosea	Variosea	Cavosteliida	Cavosteliaceae	Planoprotostelium	Planoprotostelium fungivorum	Amoebozoa	0.0909648401208991	0.0674311701494395
	76122	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Alloprevotella	Alloprevotella tannerae	Bacteria	0.0856919932771094	0.080005310091407
	281847	Eukaryota	Cryptomycota				Rozella	Rozella allomycis	Fungi	0.0738102043590749	0.0796423984628895
	712469	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Alloprevotella	Prevotella sp. oral taxon 473	Bacteria	0.0704797703512866	0.00816876553491333

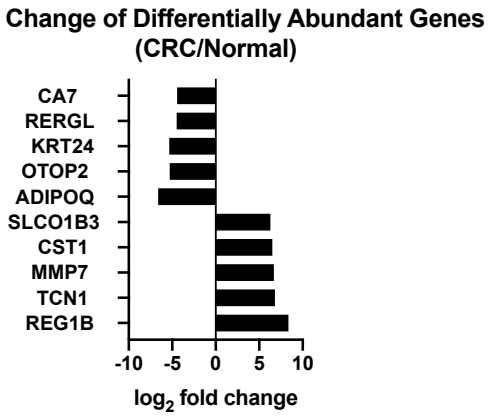


Figure 2. A. Average percent abundance of selected bacterial species in CRC and matched normal samples. **B.** Percent abundance in biological process GOs of polyamine biosynthetic processes in CRC and normal groups. **C.** Screenshot from MetaFunc R Shiny Application. This view shows the first 15 species with proteins contributing to the GO *Polyamine Biosynthetic Process*. Columns show URL which is linked to the NCBI's Taxonomy Browser, the Species Taxonomy ID, Lineage, Root Taxonomy, and percent abundances of the species in the two groups being compared: CRC and normal samples. Note that percent abundances refer to the total abundance of the species in question, not just the proteins contributing to the GO term. Results shown are sorted from highest to lowest percent abundance in the colon cancer cohort. **D.** Fold change of top upregulated and downregulated human genes (Li et al., 2018) between CRC and matched normal samples as found in this study through MetaFunc.

MetaFunc is also able to perform host gene set enrichment analysis using the differentially expressed genes. Significant gene sets (*p.adjust < 0.05*) with the highest normalized enrichment scores (NES) included such terms as '*GO_RIBOSOME_BIOGENESIS', 'GO_DNA_REPLICATION', 'GO_MITOTIC_NUCLEAR_DIVISION', 'GO_CONDENSED_CHROMOSOME' (see Table S1),* many of which appear to be related to cell division or replication, consistent with the findings of Li et al., 2018 that the upregulated lncRNAs they found were involved in *'mitosis', 'cell cycle process',* and *'mitotic cell cycle'*.

Host – Microbiome Correlations

We set MetaFunc's default abundance cutoff for microbial identification to 0.001% to remove most probable contaminants and so as not to lose any other meaningful taxonomies. It has been shown in a prior study (Ye et al., 2019), however, that most classifiers call false positives at below 0.01% abundance. We therefore applied this 0.01% cutoff in looking at the host-microbiome correlations in this dataset to narrow our focus on microbes that are more likely to be involved in our test case.

In using the 0.01% cutoff, MetaFunc was able to only identify 19 differentially abundant microbes. Their correlations with the top 100 significantly abundant genes can be seen at the URL:

http://amp.pharm.mssm.edu/clustergrammer/viz/5f02a49e8ec9bb33170b865c/cor.deg-tax.matrix.tsv. Table 1 highlights some notable correlations between differentially abundant microbes and differentially expressed human genes. *T. forsythia*, although significantly abundant in CRC samples, do not correlate significantly with any DEGs in CRC. Among its highest correlations however includes the gene Colorectal Neoplasia Differentially Expressed (CRNDE).

Conversely, we investigated which species correlate with CRNDE. The highest correlations are with with microbes *C. lusitaniae, C. necator,* and *S. pyogenes.* All correlations were determined to be significant. The same species were among the highest correlations of TCN1, and WNT2. TCN1 is among the top DEGs in cancer identified in this study as well as Li et al., 2018. WNT2 meanwhile is part of the *Wnt/β-catenin* pathway, which has roles in cell proliferation, cell migration, and cell differentiation. WNT2 is responsible for hyperactivation of β-catenin and is known to be upregulagted in CRC (Jung et al., 2015).

Table 1. Spearman Correlation Between DA Microbes and DGEs in CRC

Gene Name	Gene ID	TaxID	Species	rho	pvalue
CRNDE	ENSG00000245694.10	28112	Tannerella forsythia	0.28604	0.2214856
		36911	Clavispora lusitaniae	0.697744	0.000625388
		106590	Cupriavidus necator	0.649624	0.001935333
		1314	Streptococcus pyogenes	0.633083	0.002733787
TCN1	ENSG00000134827.8	106590	Cupriavidus necator	0.712782	0.000420175
		36911	Clavispora lusitaniae	0.607519	0.004495101
		1314	Streptococcus pyogenes	0.604511	0.004753101
WNT2	ENSG00000105989.10	1314	Streptococcus pyogenes	0.837594	4.07E-06
		106590	Cupriavidus necator	0.748872	0.00014522
		36911	Clavispora lusitaniae	0.745865	0.000159702

Dataset PRJNA4040030: Consensus Molecular Subtypes (CMS) of CRC Samples

To illustrate MetaFunc's capacity to compare more than two sample groups, we used MetaFunc to analyse transcriptomic reads from the study of Purcell et al., 2017 (raw reads may be accessed at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA404030), which are grouped into four CRC consensus molecular subtypes (CMS). A total of 33 samples were collected during surgical resection of tumours, and sample preparation for RNA sequencing was carried out using the Illumina TruSeq Stranded Total RNA Library preparation kit. For this sample SolexaQA++ (Cox et al., 2010) was used to trim reads, which were then ran through Salmon (Patro et al., 2017) to quantify transcript expression. The publicly available

CRC CMS classifier (Guinney et al., 2015) was used to categorize samples into one of four CMSs. Of the 33 samples, only 27 were classified into a CMS and of these, only one sample was classified into CMS4. This sample was also removed from the dataset for lack of replicates leaving a total of 26 samples – 7 samples in CMS1, 11 in CMS2, and 8 in CMS3.

Microbiome Results

Taxonomy

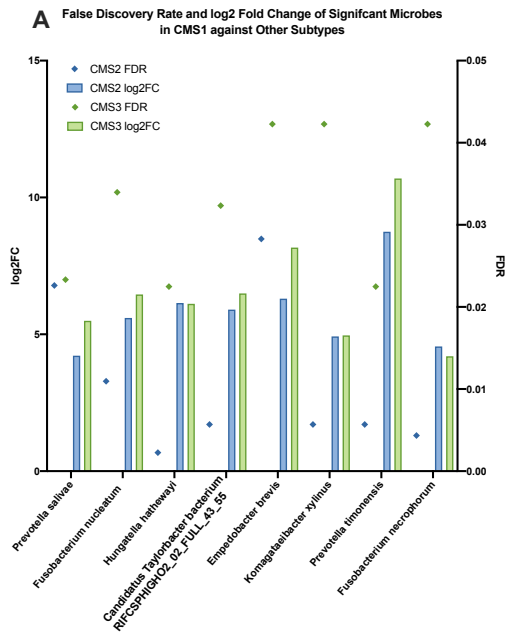
MetaFunc performed pairwise differential abundance analysis on the 3 groups using edgeR. From MetaFunc's results, we consider a species to be significantly abundant in a subtype if it is significantly abundant against both of the other subtypes. For instance, a significantly abundant species in CMS1 must be significantly abundant in the CMS1 vs CMS2 and CMS1 vs CMS3 comparisons. Using this definition, only CMS1 has species that are significantly abundant (*FDR <0.05*) compared to both CMS2 and CMS3. Figure 3A shows the False Discovery Rate (FDR; diamonds) and $\log_2$ fold change (bars) of the species in CMS1 compared to CMS2 (blue) and CMS3 (green).

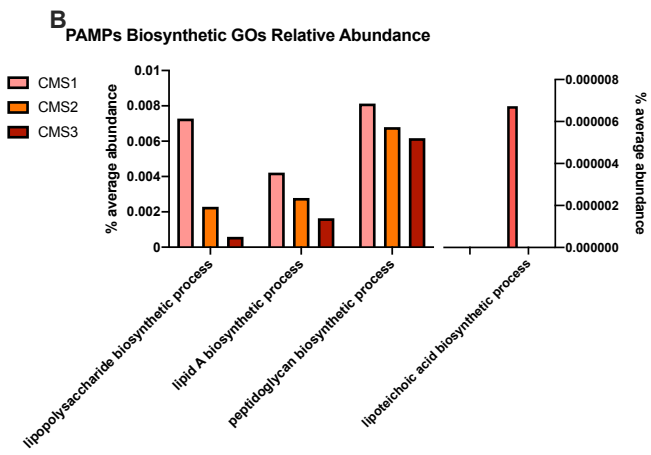
We take note of species in the genera *Prevotella* and *Fusobacterium,* which have previously been associated with colorectal cancer. *Fusobacterium nucleatum* in particular has strong evidence of an association with CRC (Dai et al., 2018; Gao et al., 2015; Ye et al., 2017). Most of these are also members of the oral microbiota, which have also previously been associated with cancer development particularly through inflammatory processes (Whitmore and Lamont, 2014). We found no species that were significantly abundant in CMS2 or CMS3 using the given criteria.

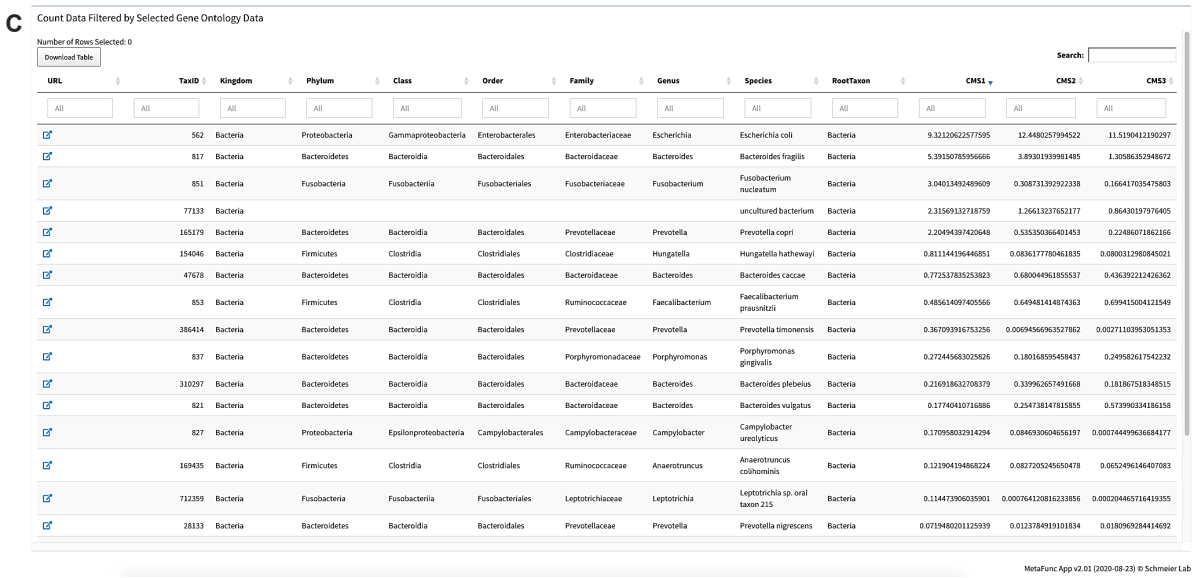
Function

Through the microbiome functional results of MetaFunc, we then investigated if processes relating to pathogen-associated molecular patterns (PAMPs) were contributed by the microbial communities, considering that CMS1 is characterized by immune responses, which are usually triggered when the human immune system recognizes such molecules. We used the MetaFunc R shiny application to search for terms *'lipopolysaccharide biosynthetic process', 'lipid A biosynthetic process', 'peptidoglycan biosynthetic process',* and *'lipoteichoic acid biosynthetic process'* and their relative abundances. Unsurprisingly, all PAMPs were relatively more abundant in CMS1 (Figure3B).

Using the MetaFunc R shiny application, we also searched for which species might be contributing to the above terms. Figure 3C is a screencap of the application showing the species contributing to any of the terms in Figure 3B. Figure 3C is arranged from highest to lowest relative abundance in CMS1 and we see microbes that were among those identified to be significantly abundant in CMS1 such as *Fusobacterium nucleatum, Prevotella timonensis* , and *Hungatella hathewayi*.







URL	TaxID	Kingdom	Phylum	Class	Order	Family	Genus	Species	RootTaxon	CMS1	CMS2	CMS3
	562	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	Escherichia	Escherichia coli	Bacteria	9.32120622577595	12.4480257994522	11.5190412190297
	817	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides fragilis	Bacteria	5.39150785956666	3.89301939981485	1.30586352948672
	851	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Fusobacteriaceae	Fusobacterium	Fusobacterium nucleatum	Bacteria	3.04013492489609	0.308731392922338	0.166417035475803
	77133	Bacteria						uncultured bacterium	Bacteria	2.31569132718759	1.26613237652177	0.86430197976405
	165179	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	Prevotella copri	Bacteria	2.20494397420648	0.535350366401453	0.224860718621 66
	154046	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Hungatella	Hungatella hathewayi	Bacteria	0.811144196446851	0.0836177780461835	0.0800312980845021
	47678	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides caccae	Bacteria	0.772537835253823	0.680044961855537	0.436392212426362
	853	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium	Faecalibacterium prausnitzii	Bacteria	0.485614097405566	0.649481414874363	0.699415004121549
	386414	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	Prevotella timonensis	Bacteria	0.367093916753256	0.00694566963527862	0.00271103953051353
	837	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Porphyromonas	Porphyromonas gingivalis	Bacteria	0.272445683025826	0.180168595458437	0.249582617542232
	310297	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides plebeius	Bacteria	0.216918632708379	0.339962657491668	0.181867518348515
	821	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides vulgatus	Bacteria	0.177404107168 86	0.254738147815855	0.573990334186158
	827	Bacteria	Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	Campylobacter	Campylobacter ureolyticus	Bacteria	0.170958032914294	0.0846930604656197	0.00074449963668 4177
	169435	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Anaerotruncus	Anaerotruncus colihominis	Bacteria	0.121904194868224	0.0827205245650478	0.0652496146407083
	712359	Bacteria	Fusobacteria	Fusobacteriia	Fusobacteriales	Leptotrichiaceae	Leptotrichia	Leptotrichia sp. oral taxon 215	Bacteria	0.114473906035901	0.000764120816233856	0.000204465716419355
	28133	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	Prevotella nigrescens	Bacteria	0.0719480201125939	0.0123784919101834	0.0180969284414692

Figure 3. A. Significantly abundant (*FDR < 0.05*) microbes in CMS1. **B.** Relative abundances of PAMPs biosynthetic processes in different CRC subtypes. **C.** Screenshot of R shiny application showing the relative abundances of species associated with PAMPs biosynthetic processes compared among CMS1, CMS2, and CMS3.

Host Results

Gene Set Expression Analysis

MetaFunc calculated differentially expressed genes between subtypes in a pairwise manner (i.e. CMS1 vs CMS2, CMS1 vs CMS3, CMS2 vs CMS3). From the DEGs of the results, MetaFunc was also able to calculate for enriched gene sets for each comparison. Similar to identifying DA microbes, we obtain a final set of enriched gene sets for a subtype if it shows enrichment compared to both other subtypes (*p.adjust < 0.05*). Unsurprisingly, we saw several GO terms involved in immune response enriched in CMS1, including *'GO_REGULATION_OF_INNATE_IMMUNE_RESPONSE',* *'GO_RESPONSE_TO_INTERFERON_GAMMA',* and *'GO_POSITIVE_REGULATION_OF_CYTOKINE_PRODUCTION* among others. Enriched GOs in CMS2 are involved in the cell cycle and and ribosome biogenesis, with terms such as *'GO_TRNA_METABOLIC_PROCESS'*, '*GO_RIBOSOMAL_LARGE_SUBUNIT_BIOGENESIS'*, and '*GO_DNA_REPLICATION_INITIATION*' while GOs enriched in CMS3 involve metabolic processes, e.g. *'GO_PRIMARY_ALCOHOL_METABOLIC_PROCESS ',* *'GO_FLAVONOID_METABOLIC_PROCESS', and 'GO_LIPID_CATABOLIC_PROCESS*. These results are consistent with the description of these three CRC subtypes in the original CMS study (Guinney et al 2015). The top enriched gene sets for each subtype can be found in Tables S2-S7.

Host – Microbiome Results

Next, using MetaFunc's correlation results, we investigated which of the top significant differentially expressed genes correlated with the significantly abundant microbes in CMS1.

We obtained the following statistically significant correlations between host and microbiome abundances shown in Table 2.

Table 2. Spearman Correlation Between DA Microbes in CMS1 and DGEs in CMS1

Gene Name	Gene ID	TaxID	Species	rho	pvalue
WARS1	ENSG00000140105.18	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.586767	1.63E-03
ICAM1	ENSG00000090339.9	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.555309	3.23E-03
WARS1	ENSG00000140105.18	**851**	*Fusobacterium nucleatum*	0.540513	4.36E-03
CCL4	ENSG00000275302.2	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.534792	4.88E-03
RNF213	ENSG00000173821.19	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.528979	5.46E-03
CCL4	ENSG00000275302.2	**851**	*Fusobacterium nucleatum*	0.508376	8.01E-03
ICAM1	ENSG00000090339.9	**851**	*Fusobacterium nucleatum*	0.502906	8.83E-03
RNF213	ENSG00000173821.19	**851**	*Fusobacterium nucleatum*	0.487863	1.15E-02
PARP14	ENSG00000173193.15	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.476321	1.39E-02
STAT1	ENSG00000115415.19	**386414**	*Prevotella timonensis*	0.474651	1.43E-02
PARP14	ENSG00000173193.15	**851**	*Fusobacterium nucleatum*	0.471453	1.50E-02
PARP9	ENSG00000138496.16	**851**	*Fusobacterium nucleatum*	0.452991	2.01E-02
ICAM1	ENSG00000090339.9	**386414**	*Prevotella timonensis*	0.429085	2.87E-02
CD163	ENSG00000177575.12	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.425372	3.03E-02
CD163	ENSG00000177575.12	**386414**	*Prevotella timonensis*	0.418729	3.33E-02
PARP14	ENSG00000173193.15	**386414**	*Prevotella timonensis*	0.418038	3.36E-02
CD163	ENSG00000177575.12	**851**	*Fusobacterium nucleatum*	0.414017	3.55E-02
STAT1	ENSG00000115415.19	**851**	*Fusobacterium nucleatum*	0.41265	3.62E-02
GBP1P1	ENSG00000225492.6	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.40568	3.98E-02
PARP9	ENSG00000138496.16	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.404514	4.04E-02
STAT1	ENSG00000115415.19	**1802307**	*Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*	0.400752	4.25E-02
GBP1	ENSG00000117228.10	**28448**	*Komagataeibacter xylinus*	-0.404514	4.04E-02
EPCAM-DT	ENSG00000234690.7	**154046**	*Hungatella hathewayi*	-0.414707	3.52E-02
GNLY	ENSG00000115523.16	**28448**	*Komagataeibacter xylinus*	-0.424004	3.09E-02

Some of these correlations may be found in http://amp.pharm.mssm.edu/clustergrammer/viz/5ee05cec8ec9bb6c154446a5/cor.deg-tax.matrix.tsv. This is the hierarchical cluster obtained when correlating top DA microbes and top DGEs in CMS1 compared to CMS2. It is to be noted that there may be correlations in

this clustering that are not found in CMS1 compared to CMS3 and are therefore not reported in Table 2.

The correlations (*rho*) between DA microbes and DGEs were quite small in value (the highest value being ~ |0.59| between WARS1 and *Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55*). Nevertheless, several of the genes appear to have relevant function with regards to CRC and immune responses. Table 3 shows information for genes that correlated with *Fusobacteria* and *Prevotella* species in our analyses. These 2 microorganisms have previously been associated with CRC.

Table 3. Gene Information of DEGs correlated with DA Microbes in CMS1.

Gene Name	Protein Name	Relevant Protein/Gene Function	Association with CRC and/or Inflammation	Sources
ICAM1	Intercellular adhesion molecule 1	Mediates cell adhesion of cytotoxic T lymphocytes and natural killer cells	upregulation of ICAM1 inhibits tumour growth and metastasis; a soluble form (sICAM1) is increased in CRC tissues compared to normal, and is associated with an inflammatory tumour microenvironment	Sánchez-Rovira et al., 1998 Schellerer et al., 2019 Tachimori et al., 2005
CCL4	C-C motif chemokine ligand 4	Chemokine involved in macrophage migration	overexpressed in CRC; directly correlated with levels of VEGF and TNF-a; correlated with CD163 and related to poor clinical outcomes and lower cumulative survival	Argyle and Kitamura, 2018 Bayoumi et al., 2016 De la Fuente López et al., 2018
CD163	CD163 receptor	M2 Macrophage marker	M2 macrophages are anti-inflammatory macrophages and CD163+ tumour-associated macrophages are with mesenchymal transition and poor prognosis in CRC; is correlated with CCL4	Argyle and Kitamura, 2018 Bayoumi et al., 2016 De la Fuente López et al., 2018 Pinto et al., 2019
STAT1	Signal transducer and activator of transcription 1	Transcription factor for IFN signalling	upregulated in CRCs; correlated with PD-L1 and PD1 immune checkpoint inhibitors; pro-oncogenic in MSI CRCs	Leon-Cabrera et al., 2018 Tanaka et al., 2020
PARP 9	Poly(ADP-ribose) polemerase family member 9	Involved in cell migration	possible role in metastasis	Vyas and Chang, 2014
PARP 14	Poly(ADP-ribose) polemerase family member 14	Involved in IL-4 signalling and cell migration	involved in anti-apopotitc effects	Vyas and Chang, 2014
RNF213	Ring Finger Protein 213	Involved in PI3K-AKT pathway for cell growth	involved in endothelial angiogenesis	Ohkubo et al., 2015
WARS1	Tryptophanyl-TRNA Synthetase 1	Inhibitor of angiogenesis; Involved in IFN-g signalling	involved in immune responses; cleaved form potentially inhibits angiogenesis; increased levels indicate better CRC survival	Ghanipour et al., 2009 Jin, 2019

Microbiome Data Output from Kaiju Compared to Humann2 Output

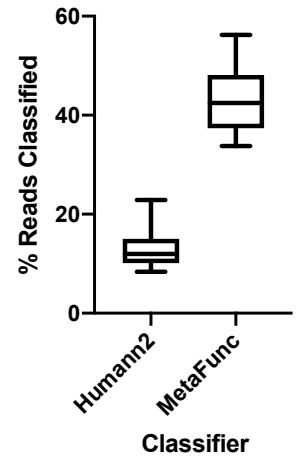


Figure 4. Percentage (%) of reads classified using Humann2 and MetaFunc pipeline using Kaiju as distributed across 20 samples.

Humann2 (Franzosa et al., 2018) is one of the packages most frequently used to assess functional pathways of the microbiome, and to ascertain which organisms are contributing to the functional pathways. Humann2 works by using MetaPhlAn2 to pre-screen for taxonomic classifications on which the database for gene hits will be based. MetaPhlAn2 has a rather limited database for pre-screening of organisms (Ye et al., 2019), resulting in a high level of unmapped reads and limited number of organisms identified.

We ran the same sequencing reads from the study PRJNA413956 (Li et al., 2018) through Humann2, first trimming with fastp and removing human-mapped reads using the same conditions as for the MetaFunc pipeline. To be more comparable, we changed the pre-screen threshold of Humann2 to 0.001% of mapped reads. Results show that for the 20 samples analyzed, 8.4% – 22.9% of reads mapped after nucleotide and protein alignment steps. In contrast, using Kaiju in the MetaFunc pipeline resulted in 33.8% – 56.2% reads mapped to microbial species through protein matches (Figure 4).

We also detected only 87 species across the 20 samples using Humann2, compared with a total of 4267 species using Kaiju in the MetaFunc pipeline.

As raw read counts that are scaled to species-classified reads as percent abundance are used in MetaFunc, while Humann2 uses reads-per-kilobase (RPK), it is difficult to compare results from Humann2 and MetaFunc directly.

Discussion

MetaFunc allowed us to investigate the relative abundances of known CRC – associated bacteria between CRC samples and matched normal tissues using the *PRJNA413956* dataset. MetaFunc results showed that microbes known to contribute to CRC progression are relatively more abundant in cancer samples while those protective in CRC are relatively more abundant in the normal samples. Through MetaFunc, we also identified that *Tannerella forsythia*, a known oral pathogen and part of the Red Complex that cause periodontal diseases (Malinowski et al., 2019), is significantly more abundant in CRC tissues than in normal tissues. Oral pathogens have previously been seen to associate with CRC samples (Flemer et al., 2018; Koliarakis et al., 2019; Thomas et al., 2019; Whitmore and Lamont, 2014). By investigating the R shiny application from MetaFunc, we also found that *T. forsythia, along* with bacteria in the *Prevotella* genera, contribute to polyamine biosynthetic processes indicating that some oral pathogens contribute to cancer progression by producing polyamines that could be taken up by the surrounding cells. Furthermore, we were able to find known bacteria in the MSI-Immune subset of CRCs by getting the common DA microbes in CMS1 against both CMS2 and CMS3 subtypes, as identified by MetaFunc's edgeR step. *Fusobacteria* have long been associated with colorectal cancer development (Dai et al., 2018; Gao et al., 2015; Thomas et al., 2019; Ye et al., 2017) while *Prevotella* includes species that inhabit the oral cavity; there have also been *Prevotella* species that were found to be abundant in CRC cohorts (Dai et al., 2018; Flemer et al., 2018; Gao et al., 2015). In line with this, PAMPs were also found to be relatively more abundant in the CMS1 cohort upon investigation through MetaFunc's R shiny application. The involvement of these bacteria in CMS1 as well as a relatively higher abundance of proteins contributing to biosynthesis of PAMPs in CMS1 indicate a role of microorganisms in

the immune responses that drive the development of CRC in these tumours. This is further supported by correlation with host genes involved in inflammation and/or CRC development as found using MetaFunc's spearman correlation step. The lack of significantly abundant microorganisms in CMS2 and CMS3 reflects that the CRC development in these subtypes are not as dependent on immune dysregulation.

We created MetaFunc with the aim of identifying microbes and their functional contribution in a microbiome environment. One of the most widely used packages for this is Humann2 (Franzosa et al., 2018) but we find the taxonomic identification given out by Humann2 too limited because of its reliance on marker genes. We do acknowledge that, especially at the 0.001% abundance cutoff, some of these species we are seeing could be false positives, or that these could be contaminants from sequencing and processing kits used (Goffau et al., 2018; Salter et al., 2014). We would caution users in interpreting data from microbes of very low abundances and would recommend following Salter and colleagues' (2014) advice of including negative control samples in sequencing. Indeed we could be seeing these effects upon looking at the microbes correlating with significantly abundant host genes in CRC samples from *PRJNA413956*. While *C. lusitaniae* is an opportunistic pathogen causing candidemia (Desnos-Ollivier et al., 2011; Krcmery et al., 1999) possibly exploiting the lowered immune responses in cancer patients (Aslani et al., 2018), and some *Streptococcus* species have previously been implicated in CRC (Kumar et al., 2017; Xia et al., 2020) with *S. pyogenes* having been known to cause invasive infections in humans (Parks et al., 2015), *Cupriavidus necator* (formerly known as *Ralstonia eutropha* (Reinecke and Steinbüchel, 2009)) is a soil bacterium that may be a sequencing contaminant in this dataset. *Cupriavidus* and *Ralstonia* species have been previously identified as common contaminants in meta-omics studies (Guo et al., 2019; Salter et al., 2014).

MetaFunc analyzes host and microbiome reads, in the end providing a user-friendly, interactive R-shiny application to investigate results, most useful for those with candidate microbes and function in mind. Downstream analysis such as differential abundance on microbes can also facilitate parsing of tables in the shiny application, while results in *'.tsv'* formats are also provided for use in other downstream bioinformatics applications. While this method was developed for a metatranscriptomic dataset, it is also suitable for metagenomic data input. As Kaiju (Menzel et al., 2016) identifies a single best protein match (or multiple matches with equal scores) of a read, we recommend its usage for short-read datasets. An exception could be made for long read sets in which the user is certain an input read will only span one protein.

We used the MetaFunc pipeline to compare genes and microbes between or among groups, but exploratory analyses of datasets from single groups can be carried out as well.

Conclusion

Here we present MetaFunc, a pipeline that is a one-stop shop of analyzing host and microbiome sequencing reads and their relationships. We find that we identify more microbes in our test datasets using MetaFunc compared to Humann2. We have used MetaFunc to determine that microbes previously known to have associations with CRC are indeed relatively more abundant in CRC samples compared to normal samples. Furthermore, we were able to use MetaFunc to pinpoint that these microorganisms could contribute to CRC progression through polyamine production.

For a dataset with more than 2 groups, we have also used MetaFunc to identify abundant bacteria in a CRC subtype associated with immune responses, while conversely, we have not been able to identify significant microbes in the other CRC subtypes. MetaFunc's spearman

correlation step showed that the significant bacteria correlate with human DEGs that function in immune responses and CRC progression. We showed that MetaFunc is able to identify candidate microorganisms that differentiate sample groups and provide insight on the functional capacities of these candidates.

References

Argyle, D., Kitamura, T., 2018. Targeting Macrophage-Recruiting Chemokines as a Novel Therapeutic Strategy to Prevent the Progression of Solid Tumors. Front Immunol 9. https://doi.org/10.3389/fimmu.2018.02629

Aslani, N., Janbabaei, G., Abastabar, M., Meis, J.F., Babaeian, M., Khodavaisy, S., Boekhout, T., Badali, H., 2018. Identification of uncommon oral yeasts from cancer patients by MALDI-TOF mass spectrometry. BMC Infectious Diseases 18. https://doi.org/10.1186/s12879-017-2916-5

Bashiardes, S., Zilberman-Schapira, G., Elinav, E., 2016. Use of Metatranscriptomics in Microbiome Research. Bioinform Biol Insights 10, 19–25. https://doi.org/10.4137/BBI.S34610

Bayoumi, A., Sayed, A., Broskova, Z., Teoh, J.-P., Wilson, J., Su, H., Tang, Y.-L., Kim, I., 2016. Crosstalk between Long Noncoding RNAs and MicroRNAs in Health and Disease. International Journal of Molecular Sciences 17, 356. https://doi.org/10.3390/ijms17030356

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., Apweiler, R., 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 32, D262–D266. https://doi.org/10.1093/nar/gkh021

Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cox, M.P., Peterson, D.A., Biggs, P.J., 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11, 485. https://doi.org/10.1186/1471-2105-11-485

Dai, Z., Coker, O.O., Nakatsu, G., Wu, W.K.K., Zhao, L., Chen, Z., Chan, F.K.L., Kristiansen, K., Sung, J.J.Y., Wong, S.H., Yu, J., 2018. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. Microbiome 6. https://doi.org/10.1186/s40168-018-0451-2

De la Fuente López, M., Landskron, G., Parada, D., Dubois-Camacho, K., Simian, D., Martinez, M., Romero, D., Roa, J.C., Chahuán, I., Gutiérrez, R., Lopez-K, F., Alvarez, K., Kronberg, U., López, S., Sanguinetti, A., Moreno, N., Abedrapo, M., González, M.-J., Quera, R., Hermoso-R, M.A., 2018. The relationship between chemokines CCL2, CCL3, and CCL4 with the tumor microenvironment and tumor-associated macrophage markers in colorectal cancer. Tumour Biol. 40, 1010428318810059. https://doi.org/10.1177/1010428318810059

Desnos-Ollivier, M., Moquet, O., Chouaki, T., Guérin, A.-M., Dromer, F., 2011. Development of Echinocandin Resistance in Clavispora lusitaniae during Caspofungin Treatment. Journal of Clinical Microbiology 49, 2304–2306. https://doi.org/10.1128/JCM.00325-11

Dienstmann, R., Vermeulen, L., Guinney, J., Kopetz, S., Tejpar, S., Tabernero, J., 2017. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. Nature Reviews Cancer 17, 79–92. https://doi.org/10.1038/nrc.2016.126

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Douglas, G.M., Maffei, V.J., Zaneveld, J., Yurgel, S.N., Brown, J.R., Taylor, C.M., Huttenhower, C., Langille, M.G.I., 2019. PICRUSt2: An improved and extensible approach for metagenome inference. Preprint at bioRxiv 672295. https://doi.org/10.1101/672295

Fernandez, N.F., Gundersen, G.W., Rahman, A., Grimes, M.L., Rikova, K., Hornbeck, P., Ma'ayan, A., 2017. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. Sci Data 4, 1–12. https://doi.org/10.1038/sdata.2017.151

Flemer, B., Warren, R.D., Barrett, M.P., Cisek, K., Das, A., Jeffery, I.B., Hurley, E., O'Riordain, M., Shanahan, F., O'Toole, P.W., 2018. The oral microbiota in colorectal cancer is distinctive and predictive. Gut 67, 1454–1463. https://doi.org/10.1136/gutjnl-2017-314814

Franzosa, E.A., McIver, L.J., Rahnavard, G., Thompson, L.R., Schirmer, M., Weingart, G., Lipson, K.S., Knight, R., Caporaso, J.G., Segata, N., Huttenhower, C., 2018. Species-level functional profiling of metagenomes and metatranscriptomes. Nature Methods 15, 962. https://doi.org/10.1038/s41592-018-0176-y

Gao, Z., Guo, B., Gao, R., Zhu, Q., Qin, H., 2015. Microbiota disbiosis is associated with colorectal cancer. Frontiers in Microbiology 6. https://doi.org/10.3389/fmicb.2015.00020

Gerner, E.W., Meyskens, F.L., 2004. Polyamines and cancer: old molecules, new understanding. Nature reviews. Cancer 4, 781–792. https://doi.org/10.1038/nrc1454

Ghanipour, A., Jirström, K., Pontén, F., Glimelius, B., Påhlman, L., Birgisson, H., 2009. The Prognostic Significance of Tryptophanyl-tRNA Synthetase in Colorectal Cancer. Cancer Epidemiol Biomarkers Prev 18, 2949–2956. https://doi.org/10.1158/1055-9965.EPI-09-0456

Goffau, M.C. de, Lager, S., Salter, S.J., Wagner, J., Kronbichler, A., Charnock-Jones, D.S., Peacock, S.J., Smith, G.C.S., Parkhill, J., 2018. Recognizing the reagent microbiome. Nat Microbiol 3, 851–853. https://doi.org/10.1038/s41564-018-0202-y

GSEA [WWW Document], n.d. URL https://www.gsea-msigdb.org/gsea/index.jsp (accessed 5.20.20).

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B.M., Morris, J.S., Simon, I.M., Gerster, S., Fessler, E., de Sousa e Melo, F., Missiaglia, E., Ramay, H., Barras, D., Homicsko, K., Maru, D., Manyam, G.C., Broom, B., Boige, V., Perez-Villamil, B., Laderas, T., Salazar, R., Gray, J.W., Hanahan, D., Tabernero, J., Bernards, R., Friend, S.H., Laurent-Puig, P., Medema, J.P., Sadanandam, A., Wessels, L., Delorenzi, M., Kopetz, S., Vermeulen, L., Tejpar, S., 2015. The Consensus Molecular Subtypes of Colorectal Cancer. Nat Med 21, 1350–1356. https://doi.org/10.1038/nm.3967

Guo, M., Xu, E., Ai, D., 2019. Inferring Bacterial Infiltration in Primary Colorectal Tumors From Host Whole Genome Sequencing Data. Front. Genet. 10. https://doi.org/10.3389/fgene.2019.00213

Huang, H., McGarvey, P.B., Suzek, B.E., Mazumder, R., Zhang, J., Chen, Y., Wu, C.H., 2011. A comprehensive protein-centric ID mapping service for molecular data integration. Bioinformatics 27, 1190–1191. https://doi.org/10.1093/bioinformatics/btr101

Hugerth, L.W., Andersson, A.F., 2017. Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. Front. Microbiol. 8. https://doi.org/10.3389/fmicb.2017.01561

Inamura, K., 2018. Colorectal Cancers: An Update on Their Molecular Pathology. Cancers 10, 26. https://doi.org/10.3390/cancers10010026

Jin, M., 2019. Unique roles of tryptophanyl-tRNA synthetase in immune control and its therapeutic implications. Experimental & Molecular Medicine 51. https://doi.org/10.1038/s12276-018-0196-9

Jung, Y.-S., Jun, S., Lee, S.H., Sharma, A., Park, J.-I., 2015. Wnt2 complements Wnt/β-catenin signaling in colorectal cancer. Oncotarget 6, 37257–37268.

Koliarakis, I., Messaritakis, I., Nikolouzakis, T.K., Hamilos, G., Souglakos, J., Tsiaoussis, J., 2019. Oral Bacteria and Intestinal Dysbiosis in Colorectal Cancer. Int J Mol Sci 20. https://doi.org/10.3390/ijms20174146

Köster, J., Rahmann, S., 2012. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28, 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Krcmery, V., Mateicka, F., Grausova, S., Kunova, A., Hanzen, J., 1999. Invasive infections due to *Clavispora lusitaniae*. FEMS Immunology & Medical Microbiology 23, 75–78. https://doi.org/10.1111/j.1574-695X.1999.tb01719.x

Kumar, R., Herold, J.L., Schady, D., Davis, J., Kopetz, S., Martinez-Moczygemba, M., Murray, B.E., Han, F., Li, Y., Callaway, E., Chapkin, R.S., Dashwood, W.-M., Dashwood, R.H., Berry, T., Mackenzie, C., Xu, Y., 2017. Streptococcus gallolyticus subsp. gallolyticus promotes colorectal tumor development. PLOS Pathogens 13, e1006440. https://doi.org/10.1371/journal.ppat.1006440

Langille, M.G.I., 2018. Exploring Linkages between Taxonomic and Functional Profiles of the Human Microbiome. mSystems 3, e00163-17. https://doi.org/10.1128/mSystems.00163-17

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G., Huttenhower, C., 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology 31, 814–821. https://doi.org/10.1038/nbt.2676

Leon-Cabrera, S., Vázquez-Sandoval, A., Molina-Guzman, E., Delgado-Ramirez, Y., Delgado-Buenrostro, N.L., Callejas, B.E., Chirino, Y.I., Pérez-Plasencia, C., Rodríguez-Sosa, M., Olguín, J.E., Salinas, C., Satoskar, A.R., Terrazas, L.I., 2018. Deficiency in STAT1 Signaling Predisposes Gut Inflammation and Prompts Colorectal Cancer Development. Cancers (Basel) 10. https://doi.org/10.3390/cancers10090341

Li, M., Zhao, L., Li, S., Li, J., Gao, B., Wang, F., Wang, S., Hu, X., Cao, J., Wang, G., 2018. Differentially expressed lncRNAs and mRNAs identified by NGS analysis in colorectal cancer patients. Cancer Medicine 7, 4650–4664. https://doi.org/10.1002/cam4.1696

Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., Mesirov, J.P., 2011. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740. https://doi.org/10.1093/bioinformatics/btr260

Macklaim, J.M., Gloor, G.B., 2018. From RNA-seq to Biological Inference: Using Compositional Data Analysis in Meta-Transcriptomics, in: Beiko, R.G., Hsiao, W., Parkinson, J. (Eds.), Microbiome Analysis: Methods and Protocols, Methods in Molecular Biology. Springer New York, New York, NY, pp. 193–213. https://doi.org/10.1007/978-1-4939-8728-3_13

Malinowski, B., Węsierska, A., Zalewska, K., Sokołowska, M.M., Bursiewicz, W., Socha, M., Ozorowski, M., Pawlak-Osińska, K., Wiciński, M., 2019. The role of Tannerella forsythia and Porphyromonas gingivalis in pathogenesis of esophageal cancer. Infectious Agents and Cancer 14, 3. https://doi.org/10.1186/s13027-019-0220-2

McCarthy, D.J., Chen, Y., Smyth, G.K., 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research 40, 4288–4297. https://doi.org/10.1093/nar/gks042

Menzel, P., Ng, K.L., Krogh, A., 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications 7, 11257. https://doi.org/10.1038/ncomms11257

Nayfach, S., Bradley, P.H., Wyman, S.K., Laurent, T.J., Williams, A., Eisen, J.A., Pollard, K.S., Sharpton, T.J., 2015. Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. PLOS Computational Biology 11, e1004573. https://doi.org/10.1371/journal.pcbi.1004573

Ohkubo, K., Sakai, Y., Inoue, H., Akamine, S., Ishizaki, Y., Matsushita, Y., Sanefuji, M., Torisu, H., Ihara, K., Sardiello, M., Hara, T., 2015. Moyamoya disease susceptibility gene RNF213 links inflammatory and angiogenic signals in endothelial cells. Scientific Reports 5, 13191. https://doi.org/10.1038/srep13191

Parks, T., Barrett, L., Jones, N., 2015. Invasive streptococcal disease: a review for clinicians. Br Med Bull 115, 77–89. https://doi.org/10.1093/bmb/ldv027

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C., 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods 14, 417–419. https://doi.org/10.1038/nmeth.4197

Pinto, M.L., Rios, E., Durães, C., Ribeiro, R., Machado, J.C., Mantovani, A., Barbosa, M.A., Carneiro, F., Oliveira, M.J., 2019. The Two Faces of Tumor-Associated Macrophages and Their Clinical Significance in Colorectal Cancer. Front. Immunol. 10. https://doi.org/10.3389/fimmu.2019.01875

Purcell, R.V., Visnovska, M., Biggs, P.J., Schmeier, S., Frizelle, F.A., 2017. Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. Scientific Reports 7. https://doi.org/10.1038/s41598-017-11237-6

Reinecke, F., Steinbüchel, A., 2009. Ralstonia eutropha strain H16 as model organism for PHA metabolism and for biotechnological production of technically interesting biopolymers. J. Mol. Microbiol. Biotechnol. 16, 91–108. https://doi.org/10.1159/000142897

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. https://doi.org/10.1093/bioinformatics/btp616

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., Walker, A.W., 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology 12, 87. https://doi.org/10.1186/s12915-014-0087-z

Sánchez-Rovira, P., Jimenez, E., Carracedo, J., Barneto, I.C., Ramirez, R., Aranda, E., 1998. Serum levels of intercellular adhesion molecule 1 (ICAM-1) in patients with colorectal cancer: inhibitory effect on cytotoxicity. European Journal of Cancer 34, 394–398. https://doi.org/10.1016/S0959-8049(97)10033-8

Schellerer, V.S., Langheinrich, M.C., Zver, V., Grützmann, R., Stürzl, M., Gefeller, O., Naschberger, E., Merkel, S., 2019. Soluble intercellular adhesion molecule-1 is a prognostic marker in colorectal carcinoma. Int J Colorectal Dis 34, 309–317. https://doi.org/10.1007/s00384-018-3198-0

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., Huttenhower, C., 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods 9, 811–814. https://doi.org/10.1038/nmeth.2066

Sharma, A.K., Gupta, A., Kumar, S., Dhakan, D.B., Sharma, V.K., 2015. Woods: A fast and accurate functional annotator and classifier of genomic and metagenomic sequences. Genomics 106, 1–6. https://doi.org/10.1016/j.ygeno.2015.04.001

Silva, G.G.Z., Green, K.T., Dutilh, B.E., Edwards, R.A., 2016. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. Bioinformatics 32, 354–361. https://doi.org/10.1093/bioinformatics/btv584

Soda, K., 2011. The mechanisms by which polyamines accelerate tumor spread. Journal of Experimental & Clinical Cancer Research 30, 95. https://doi.org/10.1186/1756-9966-30-95

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102

Tachimori, A., Yamada, N., Sakate, Y., Yashiro, M., Maeda, K., Ohira, M., Nishino, H., Hirakawa, K., 2005. Up regulation of ICAM-1 gene expression inhibits tumour growth and liver metastasis in colorectal carcinoma. Eur. J. Cancer 41, 1802–1810. https://doi.org/10.1016/j.ejca.2005.04.036

Tanaka, A., Zhou, Y., Ogawa, M., Shia, J., Klimstra, D.S., Wang, J.Y., Roehrl, M.H., 2020. STAT1 as a potential prognosis marker for poor outcomes of early stage colorectal cancer with microsatellite instability. PLOS ONE 15, e0229252. https://doi.org/10.1371/journal.pone.0229252

Ternes, D., Karta, J., Tsenkova, M., Wilmes, P., Haan, S., Letellier, E., 2020. Microbiome in Colorectal Cancer: How to Get from Meta-omics to Mechanism? Trends in Microbiology. https://doi.org/10.1016/j.tim.2020.01.001

Thomas, A.M., Manghi, P., Asnicar, F., Pasolli, E., Armanini, F., Zolfo, M., Beghini, F., Manara, S., Karcher, N., Pozzi, C., Gandini, S., Serrano, D., Tarallo, S., Francavilla, A., Gallo, G., Trompetto, M., Ferrero, G., Mizutani, S., Shiroma, H., Shiba, S., Shibata, T., Yachida, S., Yamada, T., Wirbel, J., Schrotz-King, P., Ulrich, C.M., Brenner, H., Arumugam, M.,

Bork, P., Zeller, G., Cordero, F., Dias-Neto, E., Setubal, J.C., Tett, A., Pardini, B., Rescigno, M., Waldron, L., Naccarati, A., Segata, N., 2019. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 25, 667–678. https://doi.org/10.1038/s41591-019-0405-7

Tofalo, R., Cocchi, S., Suzzi, G., 2019. Polyamines and Gut Microbiota. Front. Nutr. 6. https://doi.org/10.3389/fnut.2019.00016

Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., Segata, N., 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 12, 902–903. https://doi.org/10.1038/nmeth.3589

The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Research 45, D158–D169. https://doi.org/10.1093/nar/gkw1099

Vyas, S., Chang, P., 2014. New PARP targets for cancer therapy. Nat Rev Cancer 14, 502–509. https://doi.org/10.1038/nrc3748

Wei, H., Chen, L., Lian, G., Yang, J., Li, F., Zou, Y., Lu, F., Yin, Y., 2018. Antitumor mechanisms of bifidobacteria. Oncol Lett 16, 3–8. https://doi.org/10.3892/ol.2018.8692

Whitmore, S.E., Lamont, R.J., 2014. Oral Bacteria and Cancer. PLoS Pathogens 10, e1003933. https://doi.org/10.1371/journal.ppat.1003933

Xia, X., Wu, W.K.K., Wong, S.H., Liu, D., Kwong, T.N.Y., Nakatsu, G., Yan, P.S., Chuang, Y.-M., Chan, M.W.-Y., Coker, O.O., Chen, Z., Yeoh, Y.K., Zhao, L., Wang, X., Cheng, W.Y., Chan, M.T.V., Chan, P.K.S., Sung, J.J.Y., Wang, M.H., Yu, J., 2020. Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer. Microbiome 8, 108. https://doi.org/10.1186/s40168-020-00847-4

Xia, Y., Sun, J., Chen, D.-G., 2018. Bioinformatic Analysis of Microbiome Data, in: Xia, Y., Sun, J., Chen, D.-G. (Eds.), Statistical Analysis of Microbiome Data with R, ICSA Book Series in Statistics. Springer Singapore, Singapore, pp. 1–27. https://doi.org/10.1007/978-981-13-1534-3_1

Ye, S.H., Siddle, K.J., Park, D.J., Sabeti, P.C., 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 178, 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Ye, X., Wang, R., Bhattacharya, R., Boulbes, D.R., Fan, F., Xia, L., Adoni, H., Ajami, N.J., Wong, M.C., Smith, D.P., Petrosino, J.F., Venable, S., Qiao, W., Baladandayuthapani, V., Maru, D., Ellis, L.M., 2017. *Fusobacterium Nucleatum* Subspecies *Animalis* Influences Proinflammatory Cytokine Expression and Monocyte Activation in Human Colorectal Tumors. Cancer Prevention Research 10, 398–409. https://doi.org/10.1158/1940-6207.CAPR-16-0178

Yu, G., Wang, L.-G., Han, Y., He, Q.-Y., 2012. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS 16, 284–287. https://doi.org/10.1089/omi.2011.0118