


METHODS PAPER

# MetaFunc: taxonomic and functional analyses of high throughput sequencing for microbiomes

Arielle Kae Sulit<sup>1,2,\*</sup> , Tyler Kolisnik<sup>2</sup>, Frank Antony Frizelle<sup>1</sup>, Rachel Purcell<sup>1</sup> and Sebastian Schmeier<sup>2</sup>

<sup>1</sup>Department of Surgery, University of Otago, Christchurch, New Zealand

<sup>2</sup>School of Natural Sciences, Massey University, Auckland, New Zealand

\*Corresponding author. Email: [iel\\_sulit@yahoo.com](mailto:iel_sulit@yahoo.com)

(Received 26 July 2022; revised 06 November 2022; accepted 13 December 2022)

## Abstract

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 metagenomes or metatranscriptomes, very few of these are able to link taxonomic identity to function or are limited by their input types or databases used. Here we present MetaFunc, a workflow which takes RNA sequences as input reads, and from these (1) identifies species present in the microbiome sample and (2) provides gene ontology annotations associated with the species identified. In addition, MetaFunc allows for host gene analysis, mapping the reads to a host genome, and separating these reads, prior to microbiome analyses. Differential abundance analysis for microbe taxonomies, and differential gene expression analysis and gene set enrichment analysis may then be carried out through the pipeline. 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 showed how MetaFunc can be applied to metatranscriptomic datasets of colorectal cancer.

**Keywords:** Metatranscriptomics; microbiome; functional annotation; host correlation

## Background

Metagenomic or metatranscriptomic studies of microbiome communities allow for characterisation 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 programmes (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

© The Author(s), 2023. Published by Cambridge University Press on behalf of The Nutrition Society. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

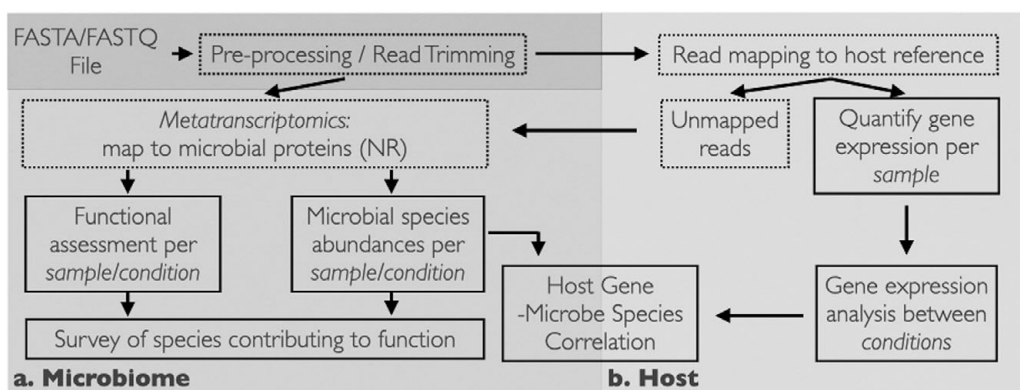
identification of bacterial genera (Bashiardes et al., 2016; Ternes et al., 2020) and are not present in other microbes aside from Bacteria and Archaea (Ye et al., 2019). Thus 16S based taxonomic identification, and subsequent functional predictions, may be unsuitable for species-level identification, and for recognising 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) highlight the limitations of using the MetaPhlan2 package, and therefore HUMAnN2, 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, using RNA sequences as input. 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 identifies microbes based on their gene expression, 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 (eg. 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.

## Protocol

### Workflow

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



**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.

trimmed for adapters and undergo 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**

MetaFunc parses through Kaiju results and gathers taxonomy IDs of species for taxonomic characterisation per sample and their corresponding protein accession numbers, which are subsequently annotated with GO terms (Figure 1a).

#### *Taxonomy*

Each classified 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 (then multiplying by 100). 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. The pipeline would also include the lineage of the taxonomies using TaxonKit (Shen and Ren, 2021).

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 (eg. 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 normalisation factors are calculated by *calcNormFactors* with default settings. *exactTest* is then applied to calculate differential abundance with *p*-values adjusted using Benjamini and Hochberg correction or false discovery rate (FDR).

#### *Proteins*

Kaiju outputs the accession number(s) of the protein match(es) with the highest BLOSUM62 alignment score of the read after translation into six 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 (then multiplying by 100).

#### *GO: 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, Fungi, 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 one 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 one *nr* entry are considered.

To facilitate GO 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 (Sulit et al., 2021a, 2021b) but users can make their own updated databases following instructions from <https://gitlab.com/schmeierlab/metafunc/metafunc-nrgo.git>.

#### *GO: protein annotation*

For each sample, the pipeline obtains only the proteins that are from taxonomy IDs that passed cutoffs in the section "Taxonomy" described above. Their scaled proportional read counts, as in the section "Proteins" 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 *GO: Database Construction*. Each GO term set annotating an accession number is then updated by accessing parent terms related to the GO terms by "*is\_a*" or "*part\_of*" using *GOATOOLS* (Klopfenstein et al., 2018). Note that this update takes the entire set of GOs annotating the accession number into consideration such that no GO terms or path/s to the top of the GO directed acyclic graph (DAG) is doubled. *GOATOOLS* also parses other information regarding the go term such as description, namespace, and depth through the *go-basic.obo* file (Ashburner et al., 2000). The proportional and scaled read counts are then added to all GO terms annotating a protein, including updated terms. Finally, the 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 groups as columns, with percentage within a namespace as values.

#### *Visualisation*

To facilitate the exploration of results from MetaFunc, MetaFunc automatically builds an R shiny application, such that users can view and interact with the taxonomy and GO 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*

Many microbiome communities are often associated with a host genome (Figure 1b). 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 programme 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, for example, the molecular signatures database (GSEA, n.d.; 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), expressed as transcript per million (TPM), and top most significant differentially abundant (DA) microbes, expressed as percent abundance. Results of these correlations are summarised 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.

#### *Tutorial/manual*

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

#### **Illustration of tool use**

##### ***Dataset PRJNA413956: matched colorectal cancer 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 tumours and corresponding adjacent non-tumour colorectal tissue samples. Raw sequencing data were 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 table of percent abundances of species that are identified in each sample and an average of these abundances across members of the same group if a grouping condition is applied. 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 (CRC) 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 CRC cohorts in meta-analyses of several datasets (Dai et al., 2018; Thomas et al., 2019). We also searched for *Bifidobacterium* species, *Bifidobacterium bifidum* and *Bifidobacterium longum*; Bifidobacteria are thought to confer protection from CRC (Wei et al., 2018).

The bars in Figure 2a show the average percent abundance of the species between samples from tumour and matched non-tumour tissue as identified through MetaFunc. As MetaFunc provides a per sample data, we are also able to plot individual values of CRC (red) and matched normal (blue) samples.

As seen in Figure 2a, MetaFunc identified *F. nucleatum*, *P. micra*, and *P. asaccharolytica* as being relatively more abundant (ie. have higher average percent abundance) in the CRC group while the *Bifidobacterium* species are relatively more abundant in the normal group.

MetaFunc also has a step that utilises 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 were 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  $\log_2$  FC = 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 CRCs (Purcell et al., 2017), the subtype most associated with immune process activation in CRC (Dienstmann et al., 2017; Guinney et al., 2015; Inamura, 2018).

**Function** MetaFunc is intended to enable comparisons of the functional potential of the microbiome between groups. MetaFunc uses GO 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).

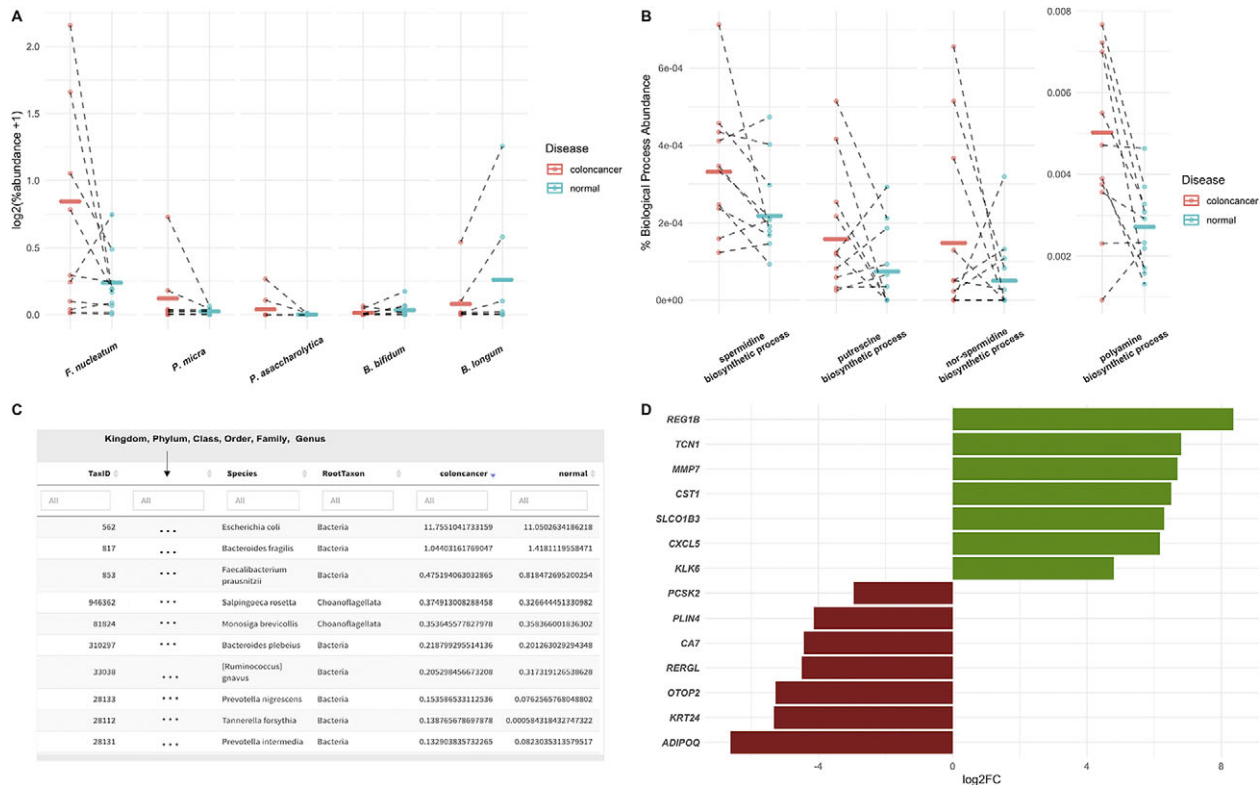
The bars in Figure 2b show the percent of reads among biological process GOs covering PA biosynthetic processes in the CRC and Normal conditions, superimposed with the individual values of samples from the CRC (red) and Normal (blue) groups. From Figure 2b, we saw that several of the polyamine biosynthetic processes were relatively more abundant (ie. higher percent of reads among biological process GOs) in the CRC cohort compared to the normal cohort, using protein annotations.

We used the built-in MetaFunc shiny application to facilitate an inquiry into the microbes species that may contribute to polyamine synthesis. 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 CRC and normal samples. Of these TaxIDs, we identified *Escherichia coli* and *B. fragilis* to be most abundant in both cohorts. However, differences in the 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 was seen in *T. forsythia*, which was previously found to be significantly more abundant in the CRC 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 results

The dataset we used for this study was from a total RNA transcriptomics run aiming to identify long non-coding RNAs (lncRNAs) and mRNAs in CRC 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 DEGs between CRC and matched normal samples through edgeR. MetaFunc results showed a total of 1,476 DEGs with an FDR < 0.05 and  $|\log_2$ 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), as well as all the genes they had randomly selected for expression confirmation via qPCR. Figure 2d shows their fold change as found through MetaFunc.

MetaFunc is also able to perform host gene set enrichment analysis using the DEGs. Significant gene sets ( $p$ .adjust < 0.05) with the highest normalised positive enrichment scores (NES) included such terms as ribosome biogenesis, DNA replication, mitotic nuclear division, and condensed chromosome (see



**Figure 2. MetaFunc Microbiome and Host Analyses of Dataset PRJNA413956. (a) Average percent abundance of selected bacterial species in CRC tissue compared to matched non-tumour (normal) samples.** From MetaFunc tabulated results, we plotted the percent abundances of selected bacteria in CRC and matched normal samples. Raw values were first  $\log_2$  transformed, with prior addition of 1 as a pseudocount to account for 0 values. Individual points represent per sample transformed values in red (CRC) and blue (Normal). Per group means are represented by the horizontal lines. Dotted lines connect matched CRC and normal samples. **(b) Percent abundance of specific polyamine biosynthetic process GO terms among all biological process GOs in a sample/group compared between CRC (red) and normal (blue) samples.** Values were calculated as described in section “GO: protein annotation” and output in MetaFunc tables or in the R Shiny application. These values were plotted, overlaying group means (horizontal lines) and individual values (data points). **(c) Screenshot from MetaFunc R shiny application.** This view shows the first 10 species with proteins contributing to the GO *polyamine biosynthetic process*. The R Shiny application columns include a URL (not shown in screenshot), which is linked to the NCBI’s Taxonomy Browser, the Species Taxonomy ID, Lineage (indicated as “...” in screenshot), Root Taxon, 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 representative upregulated and downregulated human genes** (Li et al., 2018) between CRC and matched normal samples in this study. Fold change values were obtained from the edgeR results of the pipeline. All these genes are significant (FDR < 0.05) in both this study and the source publication.

Supplementary Table S1), many of which appear to be related to cell division or replication, consistent with the findings of the source publication (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 DA 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 DA 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, included the gene Colorectal Neoplasia Differentially Expressed (*CRNDE*).

Conversely, we investigated which species correlated with *CRNDE*. The highest correlations were with microbes *Candida lusitanae*, *Cupriavidus necator*, and *Streptococcus pyogenes*. All correlations were determined to be significant. The same species were among the highest correlations of *TCN1*, and *WNT2*. *TCN1* was among the top DEGs in cancer identified in this study as well as in the source publication (Li et al., 2018). *WNT2* meanwhile is part of the Wnt/ $\beta$ -catenin pathway, which has roles in cell proliferation, cell migration, and cell differentiation. *WNT2* is responsible for the hyperactivation of  $\beta$ -catenin and is known to be upregulated in CRC (Jung et al., 2015).

#### Dataset PRJNA404030: consensus molecular subtypes of CRC samples

To illustrate MetaFunc’s capacity to compare more than two sample groups, we used MetaFunc to analyse transcriptome reads from the study of Purcell and colleagues (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

**Table 1.** Spearman correlation between DA microbes and DGEs in CRC.

Gene name	Gene ID	TaxID	Species	rho	p-value
<i>CRNDE</i>	ENSG00000245694.10	28112	<i>Tannerella forsythia</i>	0.29	0.22
		36911	<i>Clavispora lusitanae</i>	0.70	0.00063
		106590	<i>Cupriavidus necator</i>	0.65	0.0019
		1314	<i>Streptococcus pyogenes</i>	0.63	0.0027
<i>TCN1</i>	ENSG00000134827.8	106590	<i>Cupriavidus necator</i>	0.71	0.00042
		36911	<i>Clavispora lusitanae</i>	0.61	0.0045
		1314	<i>Streptococcus pyogenes</i>	0.60	0.0048
<i>WNT2</i>	ENSG00000105989.10	1314	<i>Streptococcus pyogenes</i>	0.84	4.07E-06
		106590	<i>Cupriavidus necator</i>	0.75	0.00015
		36911	<i>Clavispora lusitanae</i>	0.75	0.00016



TruSeq Stranded Total RNA Library preparation kit. For these samples, fastq-mcf from ea-utils (Aronesty, 2011, 2013) and SolexaQA++ (Cox et al., 2010) were used to trim reads, which were then run 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. Metafunc was used with default parameters, except for the following options: trimming was set to false, and featureCounts with reverse stranded option was used.

### Microbiome results

**Taxonomy** MetaFunc performed pairwise differential abundance analysis on the three groups using edgeR. From MetaFunc's results, we considered a species to be significantly abundant in a subtype if it is significantly abundant compared to both of the other subtypes. For instance, a significantly abundant species in CMS1 must be significantly abundant in the CMS1 versus CMS2 and CMS1 versus CMS3 comparisons. Using this definition, only CMS1 had species that were significantly abundant (FDR < 0.05) compared to both CMS2 and CMS3. Figure 3a shows the false discovery rate (FDR; diamonds) and log<sub>2</sub> fold change (bars) of the species in CMS1 compared to CMS2 (blue) and CMS3 (brown).

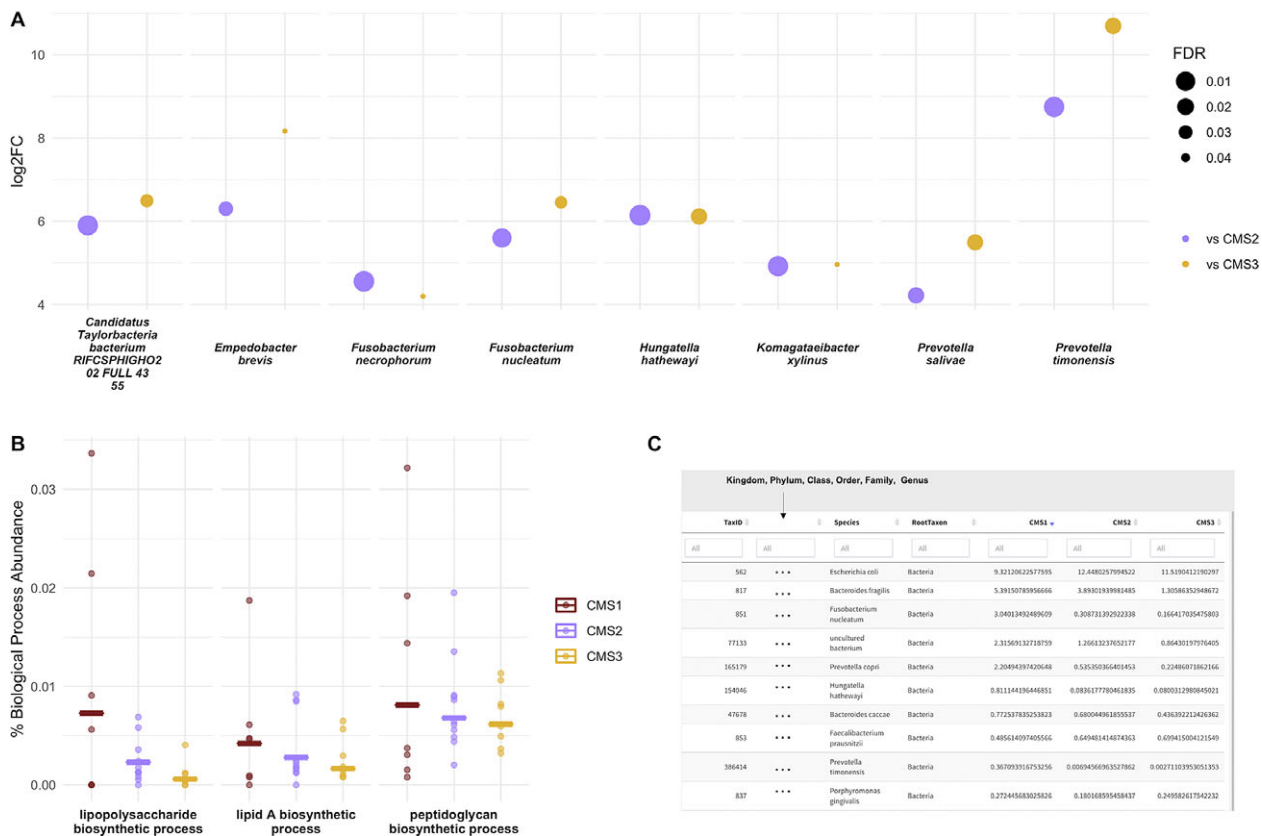
We take note of species in the genera *Prevotella* and *Fusobacterium*, which have previously been associated with CRC. *Fusobacterium nucleatum* in particular has strong evidence of an association with CRC (Dai et al., 2018; Gao et al., 2015; Ye et al., 2017). Most of these are also members of the oral 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 characterised by immune responses, which are usually triggered when the human immune system recognises such molecules. We used the MetaFunc R shiny application to search for terms “lipopolysaccharide biosynthetic process,” “lipid A biosynthetic process” and “peptidoglycan biosynthetic process,” and their relative abundances. Unsurprisingly, all PAMPs were relatively more abundant in CMS1 (Figure 3b).

Using the MetaFunc R shiny application, we also searched for which species might be contributing to the above terms. Figure 3c is a screenshot 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 saw microbes that were among those identified to be significantly abundant in CMS1 such as *F. nucleatum*, *Hungatella hathewayi*, and *Prevotella* species. These microbes have previously been associated with CRC (Dai et al., 2018; Gao et al., 2015; Wirbel et al., 2019; Ye et al., 2017).

### Host results

**Gene set expression analysis** MetaFunc calculated DEGs between subtypes in a pairwise manner (ie. CMS1 versus CMS2, CMS1 versus CMS3, CMS2 versus CMS3). From the DEGs of the results, MetaFunc was also able to calculate enriched gene sets for each comparison. Similar to identifying DA microbes, we obtained a final set of enriched gene sets for a subtype if it showed enrichment compared to both other subtypes ( $p.adjust < 0.05$ ). Unsurprisingly, we saw several host GO terms involved in immune response enriched in CMS1, including regulation of innate immune response, response to interferon gamma, and positive regulation of cytokine production among others. Enriched host GOs in CMS2 are involved in the cell cycle and ribosome biogenesis, with terms such as tRNA metabolic process, ribosomal large subunit biogenesis, and DNA replication initiation, while host GOs enriched in CMS3 involve metabolic processes, for example, primary xenobiotic metabolic process, flavonoid



**Figure 3. MetaFunc Microbiome Analysis of Dataset PRNJA4040030. (a) Microbes that are significantly more abundant (FDR < 0.05) in CMS1 compared to CMS2 (purple) and CMS3 (yellow).** Microbes are considered DA in CMS1 if it is identified through edgeR as DA in both CMS1 versus CMS2 and CMS1 versus CMS3 comparisons. Log<sub>2</sub>FC (y-axis) is the log<sub>2</sub> of the fold-change between CMS1 and the other subtypes (eg. CMS1/CMS2); FDR (point sizes) is the false discovery rate adjusted *p*-values. **(b) Percent abundance of specific PAMPs biosynthetic process GO terms among all biological process GOs in a sample/group compared between CRC subtypes, CMS1 (red), CMS2 (purple), and CMS3 (yellow).** Values were calculated as described in section “GO: protein annotation” and output in MetaFunc tables or in the R Shiny application. These values were plotted, overlaying group means (horizontal lines) and individual values (data points). **(c) Screenshot of R shiny application showing the relative abundances of species associated with PAMPs biosynthetic processes compared among CMS1, CMS2, and CMS3.** This view shows the first 10 species, with the highest abundances in CMS1, with proteins contributing to any of the PAMPs biosynthetic processes described above. The application columns show a URL (not shown in screenshot), which is linked to the NCBI’s Taxonomy Browser, the Species Taxonomy ID, Lineage (shown as “...” in screenshot), Root Taxon, and percent abundances of the species in the three groups being compared: CMS1, CMS2, and CMS3. 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 CMS1 group.

metabolic process, and 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 Supplementary Tables S2–S7.

**Table 2.** Spearman correlation between DA microbes in CMS1 and DGEs in CMS1.

Gene name	Gene ID	TaxID	Species	rho	p-value
WARS1	ENSG00000140105.18	1802307	<i>Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55</i>	0.59	0.0015
WARS1	ENSG00000140105.18	851	<i>Fusobacterium nucleatum</i>	0.55	0.0035
RNF213	ENSG00000173821.19	1802307	<i>Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55</i>	0.54	0.0048
ICAM1	ENSG00000090339.9	1802307	<i>Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55</i>	0.50	0.01
RNF213	ENSG00000173821.19	851	<i>Fusobacterium nucleatum</i>	0.50	0.01
PARP14	ENSG00000173193.15	851	<i>Fusobacterium nucleatum</i>	0.47	0.02
PARP14	ENSG00000173193.15	1802307	<i>Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55</i>	0.47	0.02
PARP9	ENSG00000138496.16	851	<i>Fusobacterium nucleatum</i>	0.46	0.02
ICAM1	ENSG00000090339.9	851	<i>Fusobacterium nucleatum</i>	0.46	0.02
SLC15A3	ENSG00000110446.11	1802307	<i>Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55</i>	0.46	0.02
STAT1	ENSG00000115415.19	386414	<i>Prevotella timonensis</i>	0.44	0.02
CD163	ENSG00000177575.12	1802307	<i>Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55</i>	0.42	0.03
PARP14	ENSG00000173193.15	386414	<i>Prevotella timonensis</i>	0.42	0.03
CD163	ENSG00000177575.12	386414	<i>Prevotella timonensis</i>	0.42	0.03
ICAM1	ENSG00000090339.9	386414	<i>Prevotella timonensis</i>	0.41	0.04
PARP9	ENSG00000138496.16	1802307	<i>Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55</i>	0.41	0.04
CD163	ENSG00000177575.12	851	<i>Fusobacterium nucleatum</i>	0.41	0.04
SLC15A3	ENSG00000110446.11	851	<i>Fusobacterium nucleatum</i>	0.40	0.04
STAT1	ENSG00000115415.19	851	<i>Fusobacterium nucleatum</i>	0.40	0.04
PML	ENSG00000140464.20	1802307	<i>Candidatus Taylorbacteria bacterium RIFCSPHIGHO2_02_FULL_43_55</i>	0.39	0.05
GBP1	ENSG00000117228.10	28448	<i>Komagataeibacter xylinus</i>	-0.39	0.05
CEBPA	ENSG00000245848.3	154046	<i>Hungatella hathewayi</i>	-0.41	0.04
GNLY	ENSG00000115523.16	28448	<i>Komagataeibacter xylinus</i>	-0.43	0.03

### Host-microbiome results

Next, using correlation results from MetaFunc, we investigated which of the top significantly DEGs correlated with the significantly abundant microbes in CMS1. We obtained the following statistically significant correlations between host and microbiome abundances shown in Table 2.

Some of these correlations may be found in <http://maayanlab.cloud/clustergrammer/viz/610d8b3c97f268000ea37f41/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 Spearman correlations ( $\rho$ ) between DA microbes and DEGs were quite small in value (the highest value being  $\sim |0.59|$  between *WARS1* and *Candidatus Taylorbacteria bacterium RIFCSPHIGHO2\_02\_FULL\_43\_55*). Nevertheless, several of the genes appeared to have a 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 two microorganisms have previously been associated with CRC.

### Comparison of MetaFunc results to HUMAnN2

HUMAnN2 (Franzosa et al., 2018) is one of the packages most frequently used to assess functional pathways of the microbiome, and to determine which organisms are contributing to the functional pathways. HUMAnN2 works by pre-screening which taxonomies are present in a sample using MetaPhlan2, afterwards aligning the reads to pangenomes of the classified taxonomies for gene hits. Unclassified reads then undergo an organism-agnostic translated search (Franzosa et al., 2018). MetaPhlan2 has a rather limited database for the pre-screening of organisms (Ye et al., 2019), resulting in a high level of unmapped reads and a 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. Part of HUMAnN2's tiered search uses diamond (Franzosa et al., 2018), which requires higher memory and run time compared to Kaiju, used by MetaFunc (Ye et al., 2019). From taxonomy identification, using Kaiju, to the generation of GO tables, took MetaFunc 11.39 hours to complete, while a comparable analysis using HUMAnN2 took 65.9 hours to complete, almost six times slower than MetaFunc on the same machine (CentOS Linux release 7.9.2009). Notably, HUMAnN2 has an additional pathway abundance and pathway coverage analysis absent from MetaFunc. Runs for HUMAnN2 may be accessed at [https://github.com/asulit08/Humannn2\\_PRJNA413956](https://github.com/asulit08/Humannn2_PRJNA413956).

Results showed that for the 20 samples analysed, 8.4–22.9% of reads were 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. We also detected only 87 species across the 20 samples using HUMAnN2, compared with a total of 4,267 species using Kaiju in the MetaFunc pipeline. Further, HUMAnN2 was only able to detect Bacteria and Viruses in the samples, while MetaFunc analysis was able to detect Fungi and Archaea as well. We also investigated the concordance of the microbial GO terms that had been classified to a taxonomy from the MetaFunc run with that of HUMAnN2. We focused on only the Bacteria and Viruses – related GO terms as found in the HUMAnN2 run. We found that the majority (69–100%) of the GOs found in HUMAnN2 was also found in the MetaFunc run. There were more unique GO terms found in the MetaFunc run, which may be due to the higher number of species detected with MetaFunc (Supplementary Figure S1).

We investigated the same species and polyamine (PA) biosynthetic process GO terms in our HUMAnN2 results as we had in the MetaFunc run of dataset PRJNA413956 (Supplementary Figures S2 and S3). We see in Supplementary Figure S2 that abundances of the species in CRC and normal groups have the same trends in HUMAnN2 results as in that of MetaFunc (Kaiju) results. In HUMAnN2 runs, however, we were not able to find *B. bifidum* among the identified species in the

**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
<i>ICAM1</i>	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
<i>SLC15A3</i>	Solute carrier (SLC) 15A3	Membrane transporter; highly expressed in macrophage populations	Upregulated by LPS via NF- $\kappa$ B pathway; influences pro-inflammatory cytokine production triggered by TLR-4	Song et al., 2018; Wang et al., 2014
<i>CD163</i>	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; are correlated with CCL4	Argyle and Kitamura, 2018; Bayoumi et al., 2016; De la Fuente López et al., 2018; Pinto et al., 2019
<i>STAT1</i>	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
<i>PARP 9</i>	Poly(ADP-ribose) polymerase family member 9	Involved in cell migration	Possible role in metastasis	Vyas and Chang, 2014
<i>PARP 14</i>	Poly(ADP-ribose) polymerase family member 14	Involved in IL-4 signalling and cell migration	Involved in anti-apoptotic effects	Vyas and Chang, 2014
<i>RNF213</i>	Ring finger protein 213	Involved in PI3K-AKT pathway for cell growth	Involved in endothelial angiogenesis	Ohkubo et al., 2015
<i>WARS1</i>	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

PRJNA413956 cohort. Meanwhile, we also see the same trends in the abundances of PA biosynthetic process GO terms in CRC samples compared to matched normal samples in HUMAnN2 runs, as in our MetaFunc run (Supplementary Figure S3), except for nor-spermidine biosynthetic process, which was not seen using HUMAnN2. Differences in abundance values were noted when comparing individual samples, however, direct comparison between HUMAnN2 and MetaFunc is difficult as raw read-counts scaled to species-classified reads are used in MetaFunc, while HUMAnN2 uses reads-per-kilobase (RPK)-based relative abundances.

## 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 show that the average abundance of microbes known to contribute to CRC progression are higher in cancer samples while those protective against CRC have higher average abundance in normal samples. Through MetaFunc, we also identified that *Tannerella forsythia*, a known oral pathogen and part of the Red Complex that causes 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, contributed 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 identifying the DA microbes in CMS1 compared to both CMS2 and CMS3 subtypes, as identified by MetaFunc's edgeR step. *Fusobacteria* have long been associated with CRC 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 may reflect 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 generated by HUMAnN2 to be limited, because of its reliance on marker genes. For our purposes, we found MetaFunc invaluable for investigating novel microbes that did not have marker gene representation, in addition to being faster for larger amounts of data, and compatible with downstream analysis programs. We showed in this paper that results from the pipeline are biologically meaningful and corroborate previous literature. It was meant to be an alternative or a complement to HUMAnN2 in this regard. Although similar trends were seen in taxa and gene ontologies of interest between CRC and matched normal samples, fewer test reads were designated as taxa using HUMAnN2 compared to MetaFunc in our comparative analysis. Unfortunately, direct comparison was not possible because HUMAnN2 and MetaFunc use different abundance outputs.

We 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 the advice of including negative control samples in sequencing (Salter et al., 2014). 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), *C. 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 analyses host and microbiome reads, providing a user-friendly, interactive R-shiny application to investigate results, most useful for those with candidate microbes and function in mind, or for exploratory analyses of the characteristics of a user's dataset. It should be noted that these values are based on raw counts and percent abundances. Microbiome datasets are considered compositional (Gloor et al., 2017; Gloor and Reid, 2016), and this should be taken into consideration during further analysis. We reiterate that values shown in the shiny application (eg. average of microbial relative abundances within a group), are to be used as initial comparisons and description of the data, and care should be taken in its interpretation, especially in the light of compositional data analysis. Further downstream analysis, such as differential abundance of microbes, can also facilitate parsing of tables in the shiny application. A gold standard for differential abundance analysis in microbiome datasets is currently non-existent and different tools reach different results (Calgaro et al., 2020; Nearing et al., 2022). We offer edgeR in MetaFunc as we believe it is a good initial tool to explore DA microbes, though this is offset by being prone to false positives (Thorsen et al., 2016). MetaFunc results provide potential starting points for more in-depth analyses or hypothesis generation for experimental procedures. In this regard, we provide results in ".tsv" formats for use in other downstream bioinformatics applications, so users might apply their own analyses of choosing.

Correlation analysis on compositional data has the same contentious issue as differential abundance. Although there is published literature supporting the use of Spearman rank correlation coefficient in this analysis (Cremonesi et al., 2018; Dai et al., 2018; Geng et al., 2014), there are dissenting voices stating that there are spurious correlations, especially in compositional data (Aitchison, 1982; Faust et al., 2012; Friedman and Alm, 2012; Lovell et al., 2015; Pearson, 1897), and as such, conclusions from such correlations are meaningless (Lovell et al., 2015). Nevertheless, Spearman correlation serves a useful purpose, especially for an initial exploration of the data. Should users choose other analyses methods, intermediate results are provided with the pipeline.

This method was developed specifically for an RNA-seq (transcriptomic/metatranscriptomic) dataset, allowing for the common analysis applied to such studies. It is intended for an initial complete analysis of the data, with only a single configuration file and sample sheet necessary once installation of the tool has been done. Users can augment this analysis by accessing host gene and microbiome count files supplied by the pipeline and use this as input in other applications. Users can also potentially use the microbiome aspect of the pipeline on a metagenomic dataset, and can adjust this in the configuration file. 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 also be carried out.

While the methodology of this paper focuses on RNA sequences, metagenomic content could affect variation seen in microbial community gene expression (Franzosa et al., 2014). It should be noted that gene copy number, for instance, could affect transcript counts. Counts seen with

metatranscriptomic data would also reflect a species' gene expression contribution as opposed to abundance. It would be prudent to take this into consideration when interpreting biological implications of the results.

## Conclusion

Here we presented MetaFunc, a single pipeline for analysing host and microbiome sequencing reads and their relationships. We found that we identified more microbes in our test datasets using MetaFunc compared to HUMAnN2, while microbes and functions of interest were comparable between the two. 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 highlight that these microorganisms could contribute to CRC progression through polyamine production.

For a dataset with more than two 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 was able to identify candidate microorganisms that differentiate sample groups and provide insight on the functional capacities of these candidates.

**Acknowledgement.** The authors would like to thank Dr. Olin Silander for valuable technical and academic advice for this manuscript.

**Disclosure statement.** The authors have no competing interests.

**Supplementary materials.** To view supplementary material for this article, please visit <http://doi.org/10.1017/gmb.2022.12>.

**Data availability statement.** MetaFunc is freely available through <https://gitlab.com/schmeierlab/workflows/metafunc.git>, and full documentation can be found in <https://metafunc.readthedocs.io/en/latest/>.

**Author contributions.** Conceptualisation: A.K.S, R.P., S.S.; Formal analysis: A.K.S, F.A.F., R.P., S.S.; Funding Acquisition: R.P.; Investigation: A.K.S., R.P., S.S.; Methodology: A.K.S, T.K., R.P., S.S.; Software: A.K.S, T.K., S.S.; Supervision: R.P., S.S.; Validation: A.K.S, S.S.; Visualisation: A.K.S; Writing – original draft: A.K.S; Writing – review and editing: A.K.S, T.K., F.A.F, R.P., S.S. A.K.S. and S.S. developed and co-wrote the pipeline which ultimately led to MetaFunc, and were involved with the majority of the design. T.K. developed the shiny application that is integrated in the pipeline. A.K.S. wrote the manuscript with editorial input from T.K., S.S. and R.P. R.P. further contributed to the design of the pipeline. F.A.F. provided guidance about all clinical aspects of the manuscript.

**Funding.** This work was supported in part by the Maurice and Phyllis Paykel Trust, Gut Cancer Foundation (NZ), with support from the Hugh Green Foundation, Colorectal Surgical Society of Australia and New Zealand (CSSANZ) and The Health Research Council of New Zealand.

**Notes on Contributors.** A.K.S recently finished her PhD in genetics from Massey University and is currently continuing her research on the microbiome in colorectal cancer at the University of Otago, Christchurch as a postdoctoral fellow. T.K. is a PhD student in computer science from Massey University. He is working on using computational modeling to study biomarkers in colorectal cancer. F.A.F is a professor of colorectal surgery at University of Otago, Christchurch and a colorectal surgeon at Christchurch Hospital with the Canterbury District Health Board. His research interests lie in the management and outcomes of patients with colorectal disease. RP is a senior research fellow at the University of Otago, Christchurch. Her research focuses on the microbiome of colorectal cancer. SS is a data scientist specializing in biological high-throughput data. He used to be an independent research group leader at Massey University. Currently, he is a senior scientist at Evotec, SE.

## References

- Aitchison J (1982) The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44**(2), 139–160. <https://doi.org/10.1111/j.2517-6161.1982.tb01195.x>
- Argyle D and Kitamura T (2018) Targeting macrophage-recruiting chemokines as a novel therapeutic strategy to prevent the progression of solid tumors. *Frontiers in Immunology* **9**, 2629. <https://doi.org/10.3389/fimmu.2018.02629>



- Aronesty E (2011) ea-utils: Command-line tools for processing biological sequencing data [C<sup>++</sup>]. Available at <https://github.com/ExpressionAnalysis/ea-utils> (original work published 2015) (accessed 16 August 2018).
- Aronesty E (2013) Comparison of sequencing utility programs. *The Open Bioinformatics Journal* 7(1), 1–8. <https://doi.org/10.2174/1875036201307010001>
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM and Sherlock G (2000) Gene ontology: Tool for the unification of biology. *Nature Genetics* 25(1), 25–29. <https://doi.org/10.1038/75556>
- Aslani N, Janbabaie G, Abastabar M, Meis JF, Babaieian M, Khodavaisy S, Boekhout T and Badali H (2018) Identification of uncommon oral yeasts from cancer patients by MALDI-TOF mass spectrometry. *BMC Infectious Diseases* 18(1), 24. <https://doi.org/10.1186/s12879-017-2916-5>
- Bashiardes S, Zilberman-Schapira G and Elinav E (2016) Use of metatranscriptomics in microbiome research. *Bioinformatics and Biology Insights* 10, 19–25. <https://doi.org/10.4137/BBIS34610>
- Bayoumi A, Sayed A, Broskova Z, Teoh J-P, Wilson J, Su H, Tang Y-L and Kim I (2016) Crosstalk between long noncoding RNAs and microRNAs in health and disease. *International Journal of Molecular Sciences* 17(3), 356. <https://doi.org/10.3390/ijms17030356>
- Calgaro M, Romualdi C, Waldron L, Risso D and Vitulo N (2020) Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biology* 21(1), 191. <https://doi.org/10.1186/s13059-020-02104-1>
- Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R and Apweiler R (2004). The gene ontology annotation (GOA) database: Sharing knowledge in Uniprot with gene ontology. *Nucleic Acids Research* 32 (database issue), D262–D266. <https://doi.org/10.1093/nar/gkh021>
- Chen S, Zhou Y, Chen Y and Gu J (2018) Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34(17), i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
- Cox MP, Peterson DA and Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11(1), 485. <https://doi.org/10.1186/1471-2105-11-485>
- Cremonesi E, Governa V, Garzon JFG, Mele V, Amicarella F, Muraro MG, Trella E, Galati-Fournier V, Oertli D, Däster SR, Drosier RA, Weixler B, Bolli M, Rosso R, Nitsche U, Khanna N, Egli A, Keck S, Slotta-Huspenina J, Terracciano LM, Zajac P, Spagnoli GC, Eppenberger-Castori S, Janssen KP, Borsig L and Izzi G (2018) Gut microbiota modulate T cell trafficking into human colorectal cancer. *Gut* 67, 1984–1994. <https://doi.org/10.1136/gutjnl-2016-313498>
- Dai Z, Coker OO, Nakatsu G, Wu WKK, Zhao L, Chen Z, Chan FKL, Kristiansen K, Sung JYJ, Wong SH and Yu J (2018) Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* 6(1), 70. <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 JC, Chahuán I, Gutiérrez R, Lopez-KF, Alvarez K, Kronberg U, López S, Sanguinetti A, Moreno N, Abedrapo M, González M-J, Quera R and Hermoso-R MA (2018) The relationship between chemokines CCL2, CCL3, and CCL4 with the tumor microenvironment and tumor-associated macrophage markers in colorectal cancer. *Tumor Biology* 40(11), 1010428318810059. <https://doi.org/10.1177/1010428318810059>
- Desnos-Ollivier M, Moquet O, Chouaki T, Guérin A-M and Dromer F (2011) Development of echinocandin resistance in *Clavispora lusitanae* during saspofungin treatment. *Journal of Clinical Microbiology* 49(6), 2304–2306. <https://doi.org/10.1128/JCM.00325-11>
- Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S and Tabernero J (2017) Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer* 17(2), 79–92. <https://doi.org/10.1038/nrc.2016.126>
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M and Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1), 15–21. <https://doi.org/10.1093/bioinformatics/bts635>
- Douglas GM, Maffei VJ, Zaneveld J, Yurgel SN, Brown JR, Taylor CM, Huttenhower C and Langille MGI (2019). PICRUSt2: an improved and extensible approach for metagenome inference. *BioRxiv*, 672295. <https://doi.org/10.1101/672295>
- Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J and Huttenhower C (2012) Microbial co-occurrence relationships in the human microbiome. *PLoS Computational Biology* 8(7), e1002606. <https://doi.org/10.1371/journal.pcbi.1002606>
- Fernandez NF, Gundersen GW, Rahman A, Grimes ML, Rikova K, Hornbeck P and Ma'ayan A (2017) Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific Data* 4(1), 170151. <https://doi.org/10.1038/sdata.2017.151>
- Flemer B, Warren RD, Barrett MP, Cisek K, Das A, Jeffery IB, Hurley E, O'Riordain M, Shanahan F and O'Toole PW (2018) The oral microbiota in colorectal cancer is distinctive and predictive. *Gut* 67(8), 1454–1463. <https://doi.org/10.1136/gutjnl-2017-314814>
- Franzosa EA, McIver LJ, Rahnava G, Thompson LR, Schirmer M, Weingart G, Lipson KS, Knight R, Caporaso JG, Segata N and Huttenhower C (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods* 15(11), 962. <https://doi.org/10.1038/s41592-018-0176-y>



- Leon-Cabrera S, Vázquez-Sandoval A, Molina-Guzman E, Delgado-Ramirez Y, Delgado-Buenrostro NL, Callejas BE, Chirino YI, Pérez-Plasencia C, Rodríguez-Sosa M, Olguín JE, Salinas C, Satoskar AR and Terrazas LI (2018) Deficiency in STAT1 signaling predisposes gut inflammation and prompts colorectal cancer development. *Cancers* **10**(9), 341. <https://doi.org/10.3390/cancers10090341>
- Li M, Zhao L, Li S, Li J, Gao B, Wang F, Wang S, Hu X, Cao J and Wang G (2018) Differentially expressed lncRNAs and mRNAs identified by NGS analysis in colorectal cancer patients. *Cancer Medicine* **7**(9), 4650–4664. <https://doi.org/10.1002/cam4.1696>
- Liao Y, Smyth GK and Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**(7), 923–930. <https://doi.org/10.1093/bioinformatics/btt656>
- Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P and Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**(12), 1739–1740. <https://doi.org/10.1093/bioinformatics/btr260>
- Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S and Bähler J (2015) Proportionality: A valid alternative to correlation for relative data. *PLoS Computational Biology* **11**(3), e1004075. <https://doi.org/10.1371/journal.pcbi.1004075>
- Macklaim JM and Gloor GB (2018) From RNA-seq to biological inference: Using compositional data analysis in meta-transcriptomics. In Beiko RG, Hsiao W and Parkinson J (eds), *Microbiome Analysis: Methods and Protocols*. New York, NY: Springer, pp. 193–213. [https://doi.org/10.1007/978-1-4939-8728-3\\_13](https://doi.org/10.1007/978-1-4939-8728-3_13)
- Malinowski B, Węsierska A, Zalewska K, Sokołowska MM, Bursiewicz W, Socha M, Ozorowski M, Pawlak-Osińska K and Wiciński M (2019) The role of *Tannerella forsythia* and *Porphyromonas gingivalis* in pathogenesis of esophageal cancer. *Infectious Agents and Cancer* **14**(1), 3. <https://doi.org/10.1186/s13027-019-0220-2>
- McCarthy DJ, Chen Y and Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**(10), 4288–4297. <https://doi.org/10.1093/nar/gks042>
- Menzel P, Ng KL and 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 PH, Wyman SK, Laurent TJ, Williams A, Eisen JA, Pollard KS and Sharpton TJ (2015) Automated and accurate estimation of gene family abundance from shotgun metagenomes. *PLoS Computational Biology* **11**(11), e1004573. <https://doi.org/10.1371/journal.pcbi.1004573>
- Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, Jones CMA, Wright RJ, Dhanani AS, Comeau AM and Langille MGI (2022) Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications* **13**(1), 342. <https://doi.org/10.1038/s41467-022-28034-z>
- Ohkubo K, Sakai Y, Inoue H, Akamine S, Ishizaki Y, Matsushita Y, Sanefuji M, Torisu H, Ihara K, Sardiello M and Hara T (2015) Moyamoya disease susceptibility gene RNF213 links inflammatory and angiogenic signals in endothelial cells. *Scientific Reports* **5**(1), 13191. <https://doi.org/10.1038/srep13191>
- Parks T, Barrett L and Jones N (2015) Invasive streptococcal disease: A review for clinicians. *British Medical Bulletin* **115**(1), 77–89. <https://doi.org/10.1093/bmb/ldv027>
- Patro R, Duggal G, Love MI, Irizarry RA and Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**(4), 417–419. <https://doi.org/10.1038/nmeth.4197>
- Pearson K (1897) Mathematical contributions to the theory of evolution—On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* **60**(359–367), 489–498. <https://doi.org/10.1098/rspl.1896.0076>
- Pinto ML, Rios E, Durães C, Ribeiro R, Machado JC, Mantovani A, Barbosa MA, Carneiro F and Oliveira MJ (2019) The two faces of tumor-associated macrophages and their clinical significance in colorectal cancer. *Frontiers in Immunology* **10**, 1875. <https://doi.org/10.3389/fimmu.2019.01875>
- Purcell RV, Visnovska M, Biggs PJ, Schmeier S and Frizelle FA (2017) Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Scientific Reports* **7**(1), 11590. <https://doi.org/10.1038/s41598-017-11237-6>
- Reinecke F and Steinbüchel A (2009) *Ralstonia eutropha* strain H16 as model organism for PHA metabolism and for biotechnological production of technically interesting biopolymers. *Journal of Molecular Microbiology and Biotechnology* **16**(1–2), 91–108. <https://doi.org/10.1159/000142897>
- Robinson MD, McCarthy DJ and Smyth GK (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ and Walker AW (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12**(1), 87. <https://doi.org/10.1186/s12915-014-0087-z>
- Sánchez-Rovira P, Jimenez E, Carracedo J, Barneto IC, Ramirez R and 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**(3), 394–398. [https://doi.org/10.1016/S0959-8049\(97\)10033-8](https://doi.org/10.1016/S0959-8049(97)10033-8)
- Schellerer VS, Langheinrich MC, Zver V, Grützmann R, Stürzl M, Gefeller O, Naschberger E and Merkel S (2019) Soluble intercellular adhesion molecule-1 is a prognostic marker in colorectal carcinoma. *International Journal of Colorectal Disease* **34**(2), 309–317. <https://doi.org/10.1007/s00384-018-3198-0>

- Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O and Huttenhower C (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9**(8), 811–814. <https://doi.org/10.1038/nmeth.2066>
- Sharma AK, Gupta A, Kumar S, Dhakan DB and Sharma VK (2015) Woods: A fast and accurate functional annotator and classifier of genomic and metagenomic sequences. *Genomics* **106**(1), 1–6. <https://doi.org/10.1016/j.ygeno.2015.04.001>
- Shen W and Ren H (2021) TaxonKit: A practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and Genomics* **48**(9), 844–850. <https://doi.org/10.1016/j.jgg.2021.03.006>
- Silva GGZ, Green KT, Dutilh BE and Edwards RA (2016) SUPER-FOCUS: A tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics (Oxford, England)* **32**(3), 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**(1), 95. <https://doi.org/10.1186/1756-9966-30-95>
- Song F, Yi Y, Li C, Hu Y, Wang J, Smith DE and Jiang H (2018) Regulation and biological role of the peptide/histidine transporter SLCl5A3 in toll-like receptor-mediated inflammatory responses in macrophage. *Cell Death & Disease* **9**(7), 1–15. <https://doi.org/10.1038/s41419-018-0809-1>
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**(43), 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
- Sulit AK, Kolisnik T, Frizelle FA, Purcell R and Schmeier S (2021a) *MetaFunc Databases: Kaiju Database* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.5602178>
- Sulit AK, Kolisnik T, Frizelle FA, Purcell R and Schmeier S (2021b) *MetaFunc Databases: Nr-go Database* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.5602157>
- Tachimori A, Yamada N, Sakate Y, Yashiro M, Maeda K, Ohira M, Nishino H and Hirakawa K (2005) Up regulation of ICAM-1 gene expression inhibits tumour growth and liver metastasis in colorectal carcinoma. *European Journal of Cancer* **41**(12), 1802–1810. <https://doi.org/10.1016/j.ejca.2005.04.036>
- Tanaka A, Zhou Y, Ogawa M, Shia J, Klimstra DS, Wang JY and Roehrl MH (2020) STAT1 as a potential prognosis marker for poor outcomes of early stage colorectal cancer with microsatellite instability. *PLoS One* **15**(4), e0229252. <https://doi.org/10.1371/journal.pone.0229252>
- Ternes D, Karta J, Tsenkova M, Wilmes P, Haan S and Letellier E (2020) Microbiome in colorectal cancer: how to get from meta-omics to mechanism? *Trends in Microbiology* **28**, 401–423. <https://doi.org/10.1016/j.tim.2020.01.001>
- The UniProt Consortium (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Research* **45**(D1), D158–D169. <https://doi.org/10.1093/nar/gkw1099>
- Thomas AM, 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 CM, Brenner H, Arumugam M, Bork P, Zeller G, Cordero F, Dias-Neto E, Setubal JC, Tett A, Pardini B, Rescigno M, Waldron L, Naccarati A and Segata N (2019) Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine* **25**(4), 667–678. <https://doi.org/10.1038/s41591-019-0405-7>
- Thorsen J, Breyndrod A, Mortensen M, Rasmussen MA, Stokholm J, Al-Soud WA, Sørensen S, Bisgaard H and Waage J (2016) Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* **4**(1), 62. <https://doi.org/10.1186/s40168-016-0208-8>
- Tofalo R, Cocchi S and Suzzi G (2019) Polyamines and gut microbiota. *Frontiers in Nutrition* **6**, 16. <https://doi.org/10.3389/fnut.2019.00016>
- Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C and Segata N (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**(10), 902–903. <https://doi.org/10.1038/nmeth.3589>
- Vyas S and Chang P (2014) New PARP targets for cancer therapy. *Nature Reviews. Cancer* **14**(7), 502–509. <https://doi.org/10.1038/nrc3748>
- Wang Y, Sun D, Song F, Hu Y, Smith DE and Jiang H (2014) Expression and regulation of the proton-coupled oligopeptide transporter PhT2 by LPS in macrophages and mouse spleen. *Molecular Pharmaceutics* **11**(6), 1880–1888. <https://doi.org/10.1021/mp500014r>
- Wei H, Chen L, Lian G, Yang J, Li F, Zou Y, Lu F and Yin Y (2018) Antitumor mechanisms of bifidobacteria. *Oncology Letters* **16**(1), 3–8. <https://doi.org/10.3892/ol.2018.8692>
- Whitmore SE and Lamont RJ (2014) Oral bacteria and cancer. *PLoS Pathogens* **10**(3), e1003933. <https://doi.org/10.1371/journal.ppat.1003933>
- Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, Fleck JS, Voigt AY, Palleja A, Ponnudurai R, Sunagawa S, Coelho LP, Schrotz-King P, Vogtmann E, Habermann N, Niméus E, Thomas AM, Manghi P, Gandini S, Serrano D, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Waldron L, Naccarati A, Segata N, Sinha R, Ulrich CM, Brenner H, Arumugam M, Bork P and Zeller G (2019) Meta-analysis of fecal metagenomes reveals global microbial

- signatures that are specific for colorectal cancer. *Nature Medicine* **25**(4), 679–689. <https://doi.org/10.1038/s41591-019-0406-6>
- Xia Y, Sun J and Chen D-G** (2018) Bioinformatic analysis of microbiome data. In Xia Y, Sun J and Chen D-G (eds), *Statistical Analysis of Microbiome Data with R*. Singapore: Springer, pp. 1–27. [https://doi.org/10.1007/978-981-13-1534-3\\_1](https://doi.org/10.1007/978-981-13-1534-3_1).
- Xia X, Wu WKK, Wong SH, Liu D, Kwong TNY, Nakatsu G, Yan PS, Chuang Y-M, Chan MW-Y, Coker OO, Chen Z, Yeoh YK, Zhao L, Wang X, Cheng WY, Chan MTV, Chan PKS, Sung JJY, Wang MH and Yu J** (2020) Bacteria pathogens drive host colonic epithelial cell promoter hypermethylation of tumor suppressor genes in colorectal cancer. *Microbiome* **8**(1), 108. <https://doi.org/10.1186/s40168-020-00847-4>
- Ye SH, Siddle KJ, Park DJ and Sabeti PC** (2019) Benchmarking metagenomics tools for taxonomic classification. *Cell* **178**(4), 779–794. <https://doi.org/10.1016/j.cell.2019.07.010>
- Ye X, Wang R, Bhattacharya R, Boulbes DR, Fan F, Xia L, Adoni H, Ajami NJ, Wong MC, Smith DP, Petrosino JF, Venable S, Qiao W, Baladandayuthapani V, Maru D and Ellis LM** (2017) *Fusobacterium Nucleatum* subspecies *Animalis* influences Proinflammatory cytokine expression and monocyte activation in human colorectal tumors. *Cancer Prevention Research* **10** (7), 398–409. <https://doi.org/10.1158/1940-6207.CAPR-16-0178>
- Yu G, Wang L-G, Han Y and He Q-Y** (2012) clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* **16**(5), 284–287. <https://doi.org/10.1089/omi.2011.0118>

---

**Cite this article:** Sulit A.K., Kolisnik T., Frizelle F.A., Purcell R., and Schmeier S. 2023. MetaFunc: taxonomic and functional analyses of high throughput sequencing for microbiomes. *Gut Microbiome*, **4**, e4, 1–21. <https://doi.org/10.1017/gmb.2022.12>