Investigating the microbial community of the human gut in colorectal cancer

Nicole Bowers, Brittany Hicks, Christina Kang-Yun, Katelyn Polemi, Kelly Sovacool

Contents

Guidelines: https://drive.google.com/file/d/1YNqYiMJYiVWXiSR9yWUbQTPAoUpVa9Id/view

Abstract

Colorectal cancer is the third leading cause of cancer related deaths globally (1). Colorectal cancer (CRC) initiates in the large intestine emerging from glandular epithelial cells. Due to a selective advantage, obtained from a series of genetic or epigenetic mutations, these unregulated cells grow and can potentially develop into colorectal adenocarcinoma (2). Many studies have used the cancer microbiome to investigate how these mutations occur, however, they have almost exclusively focused on bacteria. Human viruses have been implicated in the development of many cancers such as HPV, HTLV-1 and HIV (3). Through mutagenic and manipulative abilities, viruses cause unregulated growth by disrupting the normal function of cells. The association of colorectal cancer and the gut virome remains unknown, therefore, in this study we propose to investigate the differences in the bacterial and viral community composition and their effect on CRC development. We performed operational taxonomic unit (OTU) clustering of 16S sequences for bacterial taxonomy and operational viromic unit (OVU) clustering of metaviromes for viral taxonomy to investigate the differences in human colorectal bacterial and viral community composition. Using machine learning classification of OTUs and OVUs from healthy or cancerous samples, we aim to compare performance of the random forest model vs logistic regression. Overall, the goal of the current study is to classify whether samples are healthy or cancerous based on bacterial and viral taxa. These results provide…

Introduction

Colorectal cancer is a leading cause of cancer-related death in the United States and worldwide. Its risk and severity have been linked to colonic bacterial community composition The leading cancer-related death in the United States is colorectal cancer (4). Due to screening techniques and improvements in treatment the rate of colorectal cancer has fallen. The primary screening technique is colonoscopy. A new technique gaining popularity is Exact Sciences Cologuard, a combination test which looks for both the genetic material found in colon cancer and some more advanced colon polyps and it detects hemoglobin in the stool using fecal immunochemical testing (Cologuard). While colon cancer screening techniques are advancing, we aim to understand the environment of colorectal cancer.

Gut dysbiosis is any change to the normal microbiome that could change the relationship between the host and associated microbes (5). This is well documented gastrointestinal (GI) disorders including irritable bowel syndrome, peptic ulcers, and even gastric and colon cancers (5). Gut dysbiosis is documented in patients with colorectal cancers (6).

At present, the majority of cancer microbiome studies focus on the bacteria(citation needed). Colorectal cancer is previously linked to changes to the colonic bacterial composition yet the gut virome

is vastly unexplored (6).

Virome is a viral community associated with a particular ecosystem or holobiont. In mammals, the viruses infect host cells and the variety of organisms that inhabit us. Identifying and understanding the virome within a host requires genetic and transcriptional identification of the mammal. This virome may be able to shed light on the host's genetics in states of health and states of disease. The virome is composed of both nucleic acids (DNA and RNA).

Human viruses associated with many cancers, due to their manipulative and mutagenic abilities (4). Viral metagenomic taxa can distinguish colorectal cancer patients from control subjects (6). Furthermore, a subset of markers were identified as high-risk from a subset of patients with colorectal cancer.

Here we address the knowledge gap with respect to whether bacterial and viral community composition differ between healthy patients and those diagnosed with adenoma and carcinoma.

We used the data from **hannigan paper**_ and __sited in report_ to try to understand the viruses and bacteria in the gut microbiome in colorectal cancer.

Methods

Study design and data collection

study design / data collection based on hannigan and zackular papers

(4, 7)

The data are available in the NCBI Sequence Read Archive as PRJNA389927. All code used in preparation of this report is available in the following GitHub repository: https://github.com/kelly-sovacool/bioinf545-group3-project.

16S rRNA gene sequence processing

The 16S rRNA gene sequences associated with this study were previously reported by (7). The fastq sequence and metadata files were downloaded from the NCBI sequence read archive with the sra toolkit. The 16S rRNA gene sequences were analyzed with the mothur software package (v1.43.0) (**???**) according to the MiSeq standard operating procedure (8). Mothur allows contigs from forward and reverse reads to be assembled, pre-processing steps for primer removal and quality screening, and the identification of unique sequences. After these steps were taken to reduce sequencing and PCR errors, sequences were aligned to the SILVA database (9), screened for chimeras using VSEARCH, and clustered into operational taxonomic units (OTUs) using the default 97% similarity threshold. Alpha and beta diversity metrics were calculated with the vegan R package (10).

Metagenome and virome quality control

Trimmomatic (v.0.39) (11) was used to remove adapter sequences and low-quality reads in metagenome and virome sample sets. The read quality was assessed using FastQC (v.0.11.9) (12) before and after adapter trimming to confirm removal of adapters and low quality reads. Unpaired reads were dropped using the repair function in BBTools (v.37.62) (13). Then, reads mapping to the human GRCh38 reference genome were removed using the BWA-MEM algorithm (BWA v.0.7.12) (14). The paired unmapped reads were used for taxonomic profiling and gene annotation.

Virome assembly

Closely related viral contig sequences were categorized into operational viral units (OVUs) in the absence of taxonomic identity. OVUs were defined with the CONCOCT algorithm (v0.4.0). This algorithm applies a variation-based Bayesian approach that bins related contigs by similar tetramer and coabundance profiles within samples (15).

Classification modeling

The caret R package (16) was used to train random forest models to classify samples as healthy or cancerous based on taxonomic abundance data. Only binary classifier models were trained, rather than three-class models, to reduce the time and computing resources required. Two models were trained; one with OTU (bacterial) abundances as features and the other with OVU (viral) abundances as features. For both models, the data were split into training and test sets with 65% of the data for training. The models were validated with five-fold cross validation and tuned for mtry values to maximize the area under the receiver operating curve (AUC).

Metagenome taxonomic profiling and gene annotation

The human genome-free reads were used to profile the metagenomic taxonomy at the species, genus, family, and phylum levels using MetaPhlAn2 (17). The results were visualized using the R package phyloseq (18).

The paired reads that did not map to the human genome were aligned to the Integrated Gene Catalog (IGC) (19) of 1,267 gut microbiome samples consisting of approximately 10 million genes with the BWA MEM algorithm for metagenome annotation (14). Annotated genes were extracted from the alignment results using functions geneList and countKegg modified from the MGS-Fast pipeline (20). The differences in gene abundance between healthy, adenoma, and cancer groups were assessed using the R package edgeR (21). Up to 480 top KEGG numbers of genes that were significantly different between healthy and other groups were selected and the relevant KEGG pathway was determined using the KEGG mapper tool (22).

Results

Bacterial diversity in colorectal cancer

The influence of colorectal cancer on bacterial diversity was evaluated amongst the three disease classes. A total of 3,904 OTUs were identified. Alpha and beta diversity metrics are used to characterize and distinguish ecological communities in broad strokes. As a metric of alpha diversity, the Shannon diversity index describes diversity within communities in terms of richness and evenness, taking into account the abundance of the species present (23). As seen in Fig. 1A, individuals with healthy colons do not appear to have a distinctly different gut microbiome from a species richness and evenness perspective. Although the mean Shannon diversity decreased along the progression from healthy, to adenoma, to cancer (Fig. 1A), these perceived differences were not statistically significant among the disease groups (analysis of variance [ANOVA] p-value = 0.3). This result suggests that individuals with cancerous and healthy colons may have a select group of OTUs that dominate the environment since their index values are not relatively high.

In contrast to Shannon diversity, which quantifies diversity within communities, Bray-Curtis dissimilarity is a beta diversity metric that quantifies diversity *between* communities (23). The pairwise Bray-Curtis distances of sample-wise OTU abundances were calculated and plotted as a multi-dimensional scaling plot (Fig. 1B). There does not appear to be distinct clustering of the disease classes. Instead, each disease class is fairly dispersed. There was no statistically significant clustering of the disease groups ([ANOVA] p-value > 0.23).

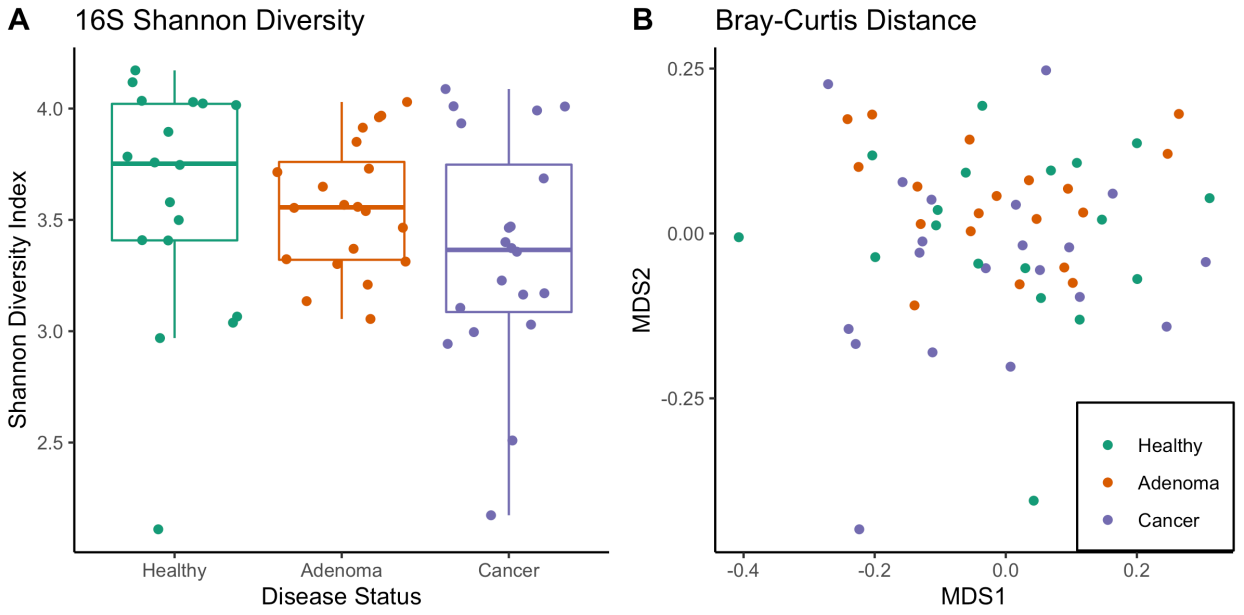

Figure 1: *(A) Shannon diversity index across disease states and (B) Multi-dimensional scaling plot of Bray-Curtis Distances.*

Classification modeling

Random forest models were trained on OTU abundances from 16S bacterial sequence data and on OVU abundances from viral metagenomes to classify samples as healthy or cancerous. Both models had modest performance with an AUROC of 0.765 for the bacterial model and an AUROC of 0.73 for the viral model (Fig. 2). The bacterial abundance data consisted of 3,704 OTUs, while the viral abundance data contained only 70 OVUs. This discrepancy in the number of features as inputs to the random forest models may have contributed to the bacterial model outperforming the viral model for classifying samples as healthy or cancerous. An OTU containing members of the genus *Fusobacterium* was by far the most important feature of the bacterial model. The viral metagenomes were not aligned to a reference database, so OVUs were not able to be identified for feature importance determination. Previous work by Hannigan *et al.* additionally included models trained on bacterial whole metagenomes and models combining bacterial and viral abundance data.

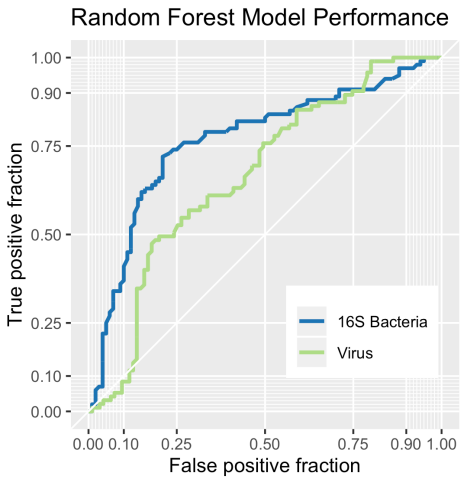

Figure 2: *Model performance as measured by the receiver operating characteristic (ROC) curve for models trained on bacterial abundance or viral abundance data.*

Taxonomonic abundance

moar werdz go here

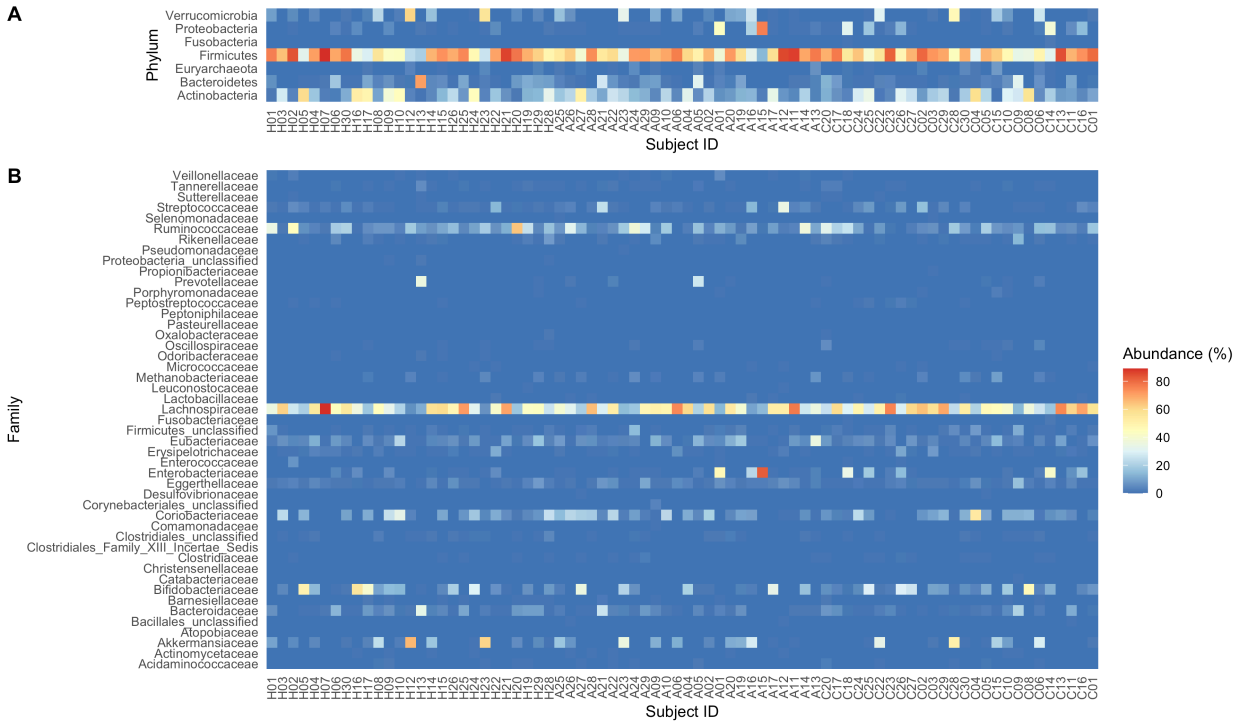

Figure 3

Metagenome annotation and quantification

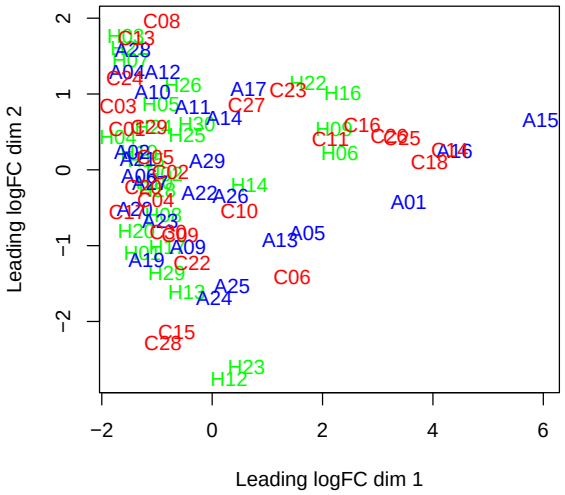

Figure 4: *Multidimensional scaling (MDS) abundance plot based on KEGG annotation counts derived from gut microbiome of patients that are healthy (H) or diagnosed with adenoma (A) or cancer (C).*

The gut metagenomes from patients that are healthy or diagnosed with adenoma or cancer was aligned to the IGC database to annotate the genes with KEGG numbers. KEGG number counts were calculated for all samples, then were used to determine whether specific genes were more or less abundant in patients with adenoma or cancer as compared to those who are healthy using edgeR. Seven samples, including the negative control, with extremely low read counts were excluded from subsequent analyses, as they strongly skewed the results. Multidimensional scaling (MDS) analysis of the KEGG number counts between different groups did not show clear clustering within groups (Fig. 4).

In the subsequent differential gene abundance analysis, abundance of 6,104 distinct genes

were compared between healthy vs adenoma and healthy vs cancer patients. In both cases, there were approximately 130 genes that were less abundant in adenoma and cancer patients, whereas around 600 to 1,000 genes were more abundant. Differentially abundant genes that were found in both adenoma and cancer patient groups were selected and used to map to metabolic pathways using the KEGG mapper (see Tables 1 and 2).

The less and more abundant genes found in adenoma and cancer patients mapped to approximately 20 and 550 genes in the KEGG pathways, respectively. Less abundant genes in adenoma and cancer patients were slightly enriched in pathways in metabolism, genetic information processing, signal transduction, transport and catabolism, immune system, and bacterial infectious disease (Table 1). Contrastingly, more abundant genes in adenoma and cancer patients were most enriched in pathways in metabolism, environmental information processing, and cellular proccesses amongst other pathways. Though human genome was filtered out through pre-processing, pathways pertaining to human diseases and organismal systems were also enriched in the case of genes more abundant in adenoma and cancer patients.

Discussion

taxonomic abundance, gene abundance

The alpha and beta diversity of bacteria of the gut microbiome did not significantly shift in response to the presence of colorectal cancer. Standard alpha and beta diversity metrics did not adequately capture bacterial community differences of the healthy and cancerous gut. This illuminates that examining diversity metrics alone may not be an adequate tool for identifying a healthy versus a cancerous colon.

random forest discriminates disease classes better than diversity metrics

Acknowledgements

Author order was determined by alphabetizing by last name. Christina Kang-Yun constructed the workflow for preprocessing metagenome data, taxonomic profiling, and gene annotation and prepared the methods and results for these analyses.

insert contribution statement here

References

1. **Rawla P**, **Sunkara T**, **Barsouk A**. 2019. Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors. Gastroenterology Rev **14**:89–103. doi:10.5114/pg.2018.81072.

2. **Ewing I**, **Hurley JJ**, **Josephides E**, **Millar A**. 2014. The molecular genetics of colorectal cancer.

Table 1: *Genes that are less abundant in adenoma or cancer patients as compared to healthy patients.*

KEGG Pathway		Number of Genes	KEGG No.
Metabolism			
Global and overview maps	Metabolic pathways	5	K00750
			K00895
			K01184
			K03280
			K11694
	Biosynthesis of secondary metabolites	2	K00750
			K00895
	Microbial metabolism in diverse environments	1	K00895
Carbohydrate metabolism	Glycolysis / Gluconeogenesis	1	K00895
	Pentose phosphate pathway	1	K00895
	Pentose and glucuronate interconversions	1	K01184
	Fructose and mannose metabolism	1	K00895
	Starch and sucrose metabolism	1	K00750
Glycan biosynthesis and metabolism	Lipopolysaccharide biosynthesis	1	K03280
	Peptidoglycan biosynthesis	1	K11694
Genetic Information Processing			
Folding, sorting and degradation	Ubiquitin mediated proteolysis	1	K10595
Replication and repair	Homologous recombination	1	K10875
Environmental Information Processing			
Signal transduction	Two-component system	1	K07770
Cellular Processes			
Transport and catabolism	Autophagy - yeast	1	K01336
Organismal Systems			
Immune system	Toll and Imd signaling pathway	1	K01446
Human Diseases			
Infectious disease: bacterial	Vibrio cholerae infection	1	K10954

Table 2: *Genes that are more abundant in adenoma or cancer patients as compared to healthy patients.*

KEGG Pathway		Number of Genes
Metabolism		
Global and overview maps	Metabolic pathways, etc.	193
Carbohydrate metabolism	Glycolysis / Gluconeogenesis, etc.	40
Energy metabolism	Oxidative phosphorylation, etc.	15
Lipid metabolism	Fatty acid biosynthesis, etc.	17
Nucleotide metabolism	Purine metabolism, etc.	4
Amino acid metabolism	Alanine, aspartate and glutamate metabolism, etc.	50
Metabolism of other amino acids	beta-Alanine metabolism, etc.	14
Glycan biosynthesis and metabolism	Lipopolysaccharide biosynthesis, etc.	8
Metabolism of cofactors and vitamins	Riboflavin metabolism, etc.	17
Metabolism of terpenoids and polyketides	Limonene and pinene degradation, etc.	7
Biosynthesis of other secondary metabolites	Isoquinoline alkaloid biosynthesis, etc.	6
Xenobiotics biodegradation and metabolism	Benzoate degradation, etc.	20
Genetic Information Processing		
Transcription	RNA polymerase	1
Folding, sorting and degradation	Protein export, etc.	2
Replication and repair	DNA replication, etc.	8
Environmental Information Processing		
Membrane transport	ABC transporters, etc.	47
Signal transduction	Two-component system, etc.	35
Cellular Processes		
Cell growth and death	Cell cycle - yeast	1
Cellular community - prokaryotes	Quorum sensing, etc.	27
Cell motility	Bacterial chemotaxis, etc.	12
Organismal Systems		
Immune system	Complement and coagulation cascades	1
Aging	Longevity regulating pathway - worm	1
Human Diseases		
Cancer: overview	Pathways in cancer, etc.	3
Cancer: specific types	Hepatocellular carcinoma, etc.	2
Cardiovascular disease	Fluid shear stress and atherosclerosis	1
Infectious disease: bacterial	Epithelial cell signaling in Helicobacter pylori infection, etc.	10
Drug resistance: antimicrobial	beta-Lactam resistance, etc.	3
Drug resistance: antineoplastic	Platinum drug resistance	1

Frontline Gastroenterology **5**:26–30. doi:10.1136/flgastro-2013-100329.

3. **Liao JB**. 2006. Viruses and human cancer. Yale J Biol Med **79**:115–122.

4. **Hannigan GD**, **Duhaime MB**, **Ruffin MT**, **Koumpouras CC**, **Schloss PD**. 2018. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. mBio **9**. doi:10.1128/mBio.02248-18.

5. **Neto AG**, **Hickman RA**, **Khan A**, **Nossa C**, **Pei Z**. 2017. The Upper Gastrointestinal Tract—Esophagus and Stomach, pp. 1–11. *In* The Microbiota in Gastrointestinal Pathophysiology. Elsevier.

6. **Nakatsu G**, **Zhou H**, **Wu WKK**, **Wong SH**, **Coker OO**, **Dai Z**, **Li X**, **Szeto C-H**, **Sugimura N**, **Lam TY-T**, **Yu AC-S**, **Wang X**, **Chen Z**, **Wong MC-S**, **Ng SC**, **Chan MTV**, **Chan PKS**, **Chan FKL**, **Sung JJ-Y**, **Yu J**. 2018. Alterations in Enteric Virome Are Associated With Colorectal Cancer and Survival Outcomes. Gastroenterology **155**:529–541.e5. doi:10.1053/j.gastro.2018.04.018.

7. **Zackular JP**, **Rogers MAM**, **Ruffin MT**, **Schloss PD**. 2014. The Human Gut Microbiome as a Screening Tool for Colorectal Cancer. Cancer Prev Res **7**:1112–1121. doi:10.1158/1940-6207.CAPR-14-0129.

8. **Kozich JJ**, **Westcott SL**, **Baxter NT**, **Highlander SK**, **Schloss PD**. 2013. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Appl Environ Microbiol **79**:5112. doi:10.1128/AEM.01043-13.

9. **Pruesse E**, **Quast C**, **Knittel K**, **Fuchs BM**, **Ludwig W**, **Peplies J**, **Glöckner FO**. 2007. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res **35**:7188–7196. doi:10.1093/nar/gkm864.

10. **Oksanen J**. 2018. Vegan: Community ecology package.

11. **Bolger AM**, **Lohse M**, **Usadel B**. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **30**:2114–2120. doi:10.1093/bioinformatics/btu170.

12. **Andrews S**, **Krueger F**, **Segonds-Pichon A**, **Biggins L**, **Krueger C**, **Wingett S**. 2010. FastQC.

13. **Bushnell B**. 2018. BBTools: A suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data.

14. **Li H**. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

15. **Alneberg J**, **Bjarnason BS**, **Bruijn I de**, **Schirmer M**, **Quick J**, **Ijaz UZ**, **Lahti L**, **Loman NJ**, **Andersson AF**, **Quince C**. 2014. Binning metagenomic contigs by coverage and composition. Nature Methods **11**:1144–1146. doi:10.1038/nmeth.3103.

16. **Kuhn M**. 2013. Caret: Classification and regression training **1**.

17. **Truong DT**, **Franzosa EA**, **Tickle TL**, **Scholz M**, **Weingart G**, **Pasolli E**, **Tett A**, **Huttenhower C**, **Segata N**. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods **12**:902–903. doi:10.1038/nmeth.3589.

18. **McMurdie PJ**, **Holmes S**. 2013. Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. PLOS ONE **8**:1–11. doi:10.1371/journal.pone.0061217.

19. **Li J**, **Jia H**, **Cai X**, **Zhong H**, **Feng Q**, **Sunagawa S**, **Arumugam M**, **Kultima JR**, **Prifti E**, **Nielsen T**, **Juncker AS**, **Manichanh C**, **Chen B**, **Zhang W**, **Levenez F**, **Wang J**, **Xu X**, **Xiao L**, **Liang S**, **Zhang D**, **Zhang Z**, **Chen W**, **Zhao H**, **Al-Aama JY**, **Edris S**, **Yang H**, **Wang J**, **Hansen T**, **Nielsen HB**, **Brunak S**, **Kristiansen K**, **Guarner F**, **Pedersen O**, **Doré J**, **Ehrlich SD**, **Bork P**, **Wang J**. 2014. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol **32**:834–841. doi:10.1038/nbt.2942.

20. **Brown SM**, **Chen H**, **Hao Y**, **Laungani BP**, **Ali TA**, **Dong C**, **Lijeron C**, **Kim B**, **Wultsch C**, **Pei Z**, **Krampis K**. 2019. MGS-Fast: Metagenomic shotgun data fast annotation using microbial gene catalogs. Gigascience **8**. doi:10.1093/gigascience/giz020.

21. **Robinson MD**, **McCarthy DJ**, **Smyth GK**. 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**:139–140. doi:10.1093/bioinformatics/btp616.

22. **Kanehisa M**, **Sato Y**. 2020. KEGG Mapper for inferring cellular functions from protein sequences. Protein Science **29**:28–35. doi:10.1002/pro.3711.

23. **Tuomisto H**. 2010. A diversity of beta diversities: Straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. Ecography **33**:2–22. doi:10.1111/j.1600-0587.2009.05880.x.