

Investigating the microbial community of the human gut in colorectal cancer

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

Contents

Abstract	2
Introduction	2
Methods	2
Study design and data collection	2
16S rRNA gene sequence processing	2
Metagenome and virome quality control	2
Virome assembly	2
Classification modeling	2
Metagenome taxonomic profiling and gene annotation	2
Results	3
Metagenome annotation and quantification	3
Discussion	3
Acknowledgements	3
References	4

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

Abstract

Introduction

Methods

Study design and data collection

study design / data collection based on hannigan and zackular papers

(1, 2)

The data are available in the NCBI Sequence Read Archive as [PRJNA389927](https://www.ncbi.nlm.nih.gov/sra/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

Metagenome and virome quality control

Trimmomatic (v.0.39) (3) 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) (4) 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) (5). Then, reads mapping to the human GRCh38 reference genome were removed using the BWA-MEM algorithm (BWA v.0.7.12) (6). The paired unmapped reads were used for taxonomic profiling and gene annotation.

Virome assembly

Classification modeling

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 (7). The results were visualized using the R package phyloseq (8).

The paired reads that did not map to the human genome were aligned to the Integrated Gene Catalog (IGC) (9) of 1,267 gut microbiome samples consisting of approximately 10 million genes with the BWA MEM algorithm for metagenome annotation (6). Annotated genes were extracted from the alignment results using functions geneList and countKegg modified from the MGS-Fast pipeline (10). The differences in gene abundance between healthy, adenoma, and cancer groups were assessed using the R package edgeR (11). Up to 500 top KEGG numbers of genes that were

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	3	K01184
			K03280
			K11694
Carbohydrate metabolism	Pentose and glucuronate interconversions	1	K01184
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
Cellular Processes			
Transport and catabolism	Autophagy - yeast	1	K01336

References

1. **Hannigan GD, Duhaime MB, Ruffin MT, Koumpouras CC, Schloss PD.** 2017. The Diagnostic Potential & Interactive Dynamics of the Colorectal Cancer Virome. doi:[10.1101/152868](https://doi.org/10.1101/152868).
2. **Zackular JP, Rogers MAM, Ruffin MT, Schloss PD.** 2014. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prevention Research* **7**:1112–1121. doi:[10.1158/1940-6207.CAPR-14-0129](https://doi.org/10.1158/1940-6207.CAPR-14-0129).
3. **Bolger AM, Lohse M, Usadel B.** 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120. doi:[10.1093/bioinformatics/btu170](https://doi.org/10.1093/bioinformatics/btu170).
4. **Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S.** 2010. FastQC. Babraham Institute, Babraham, UK.
5. **Bushnell B.** 2018. BBTools: A suite of fast, multithreaded bioinformatics tools designed for analysis of dna and rna sequence data.
6. **Li H.** 2013. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem.
7. **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. *Nature Methods* **12**:902–903. doi:[10.1038/nmeth.3589](https://doi.org/10.1038/nmeth.3589).
8. **McMurdie S Paul J. AND Holmes.** 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](https://doi.org/10.1371/journal.pone.0061217).
9. **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**

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	104
Carbohydrate metabolism	Glycolysis / Gluconeogenesis	4
Energy metabolism	Oxidative phosphorylation	5
Lipid metabolism	Fatty acid biosynthesis	1
Nucleotide metabolism	Purine metabolism	5
Amino acid metabolism	Alanine, aspartate and glutamate metabolism	2
Metabolism of other amino acids	beta-Alanine metabolism	3
Glycan biosynthesis and metabolism	Lipopolysaccharide biosynthesis	8
Metabolism of cofactors and vitamins	Riboflavin metabolism	4
Metabolism of terpenoids and polyketides	Limonene and pinene degradation	2
Biosynthesis of other secondary metabolites	Isoquinoline alkaloid biosynthesis	2
Xenobiotics biodegradation and metabolism	Benzoate degradation	4
Genetic Information Processing		
Folding, sorting and degradation	Protein export	1
Replication and repair	DNA replication	1
Environmental Information Processing		
Membrane transport	ABC transporters	33
Signal transduction	Two-component system	36
Cellular Processes		
Cell growth and death	Cell cycle - yeast	1
Cellular community - prokaryotes	Quorum sensing	6
Cell motility	Bacterial chemotaxis	7
Organismal Systems		
Immune system	Complement and coagulation cascades	1
Aging	Longevity regulating pathway - worm	1
Human Diseases		
Cancer: overview	Pathways in cancer	1
Cancer: specific types	Hepatocellular carcinoma	1
Cardiovascular disease	Fluid shear stress and atherosclerosis	1
Infectious disease: bacterial	Epithelial cell signaling in Helicobacter pylori infection	1
Drug resistance: antimicrobial	beta-Lactam resistance	1
Drug resistance: antineoplastic	Platinum drug resistance	1

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](https://doi.org/10.1038/nbt.2942).

10. **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](https://doi.org/10.1093/gigascience/giz020).

11. **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](https://doi.org/10.1093/bioinformatics/btp616).

12. **Kanehisa M, Sato Y.** 2020. KEGG Mapper for inferring cellular functions from protein sequences. *Protein Science* **29**:28–35. doi:[10.1002/pro.3711](https://doi.org/10.1002/pro.3711).