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

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. 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 ggplot2 (v.3.3.0) (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 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 (12).

Results

Taxnomonic profile based on metagenome

The taxonomic profile was constructed based on the gut microbiome of healthy, adenoma, and cancer patients using MetaPhlAn2 (Fig. 1). There are subtle differences of composition between healthy patients and those who were diagnosed with adenoma or cancer. At the phylum level, some healthy patients have an more members of the phyla Verrucomicrobia, Bacteroidetes, and Actinobacteria as compared to other patient groups. In addition, some patients with adenoma or cancer have a greater abundance of species belonging to the Proteobacteria phylum. The gut microbiome of healthy patients are relatively more abundant with members of the families *Ruminococcaceae*, *Lachnospiraceae*, *Bifidobacteriaceae*, and *Akkermansiaceae*. Conversely, the gut microbiome of patients with adenoma or cancer seem to have greater abundance of members of the families *Enterobacteriaceae* and *Coriobacteriaceae*.

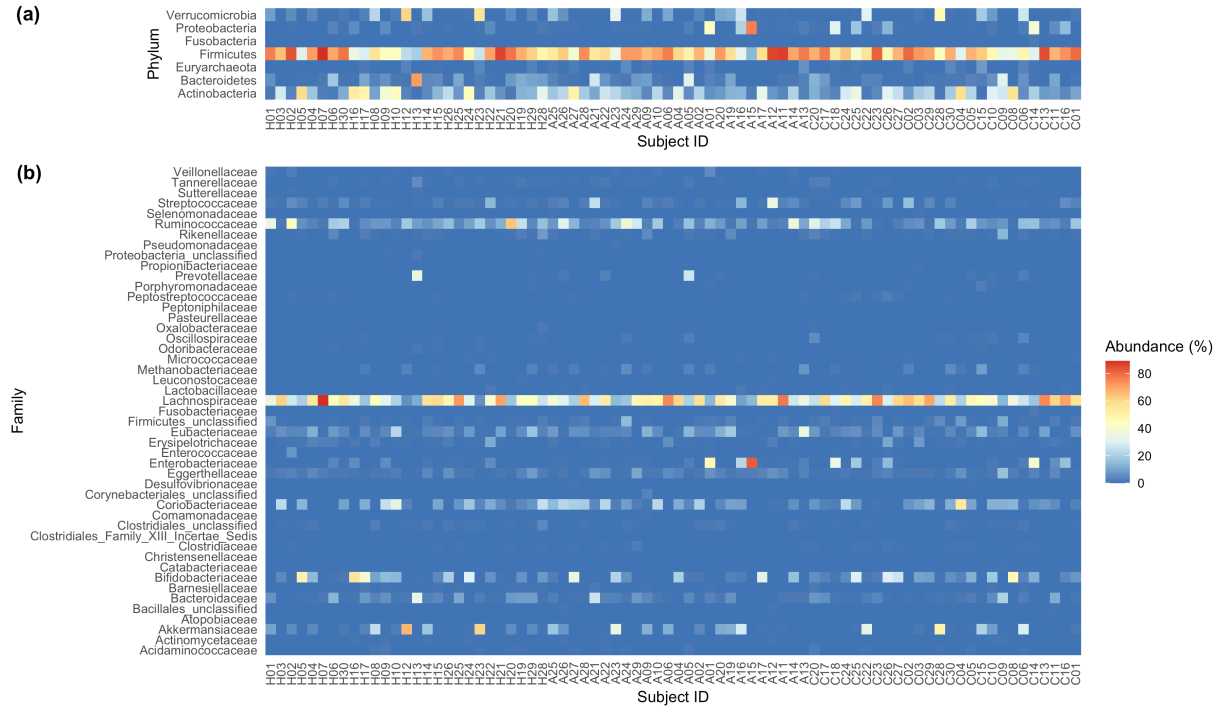

Figure 1: *Taxonomic profile constructed using MetaPhlAn2 at the (a) phylum and (b) family level based on metagenome of gut microbiome of patients that are healthy (H) or diagnosed with adenoma (A) or cancer (C).*

Metagenome annotation and quantification

The gut metagenome 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. 2).

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 Table 1).

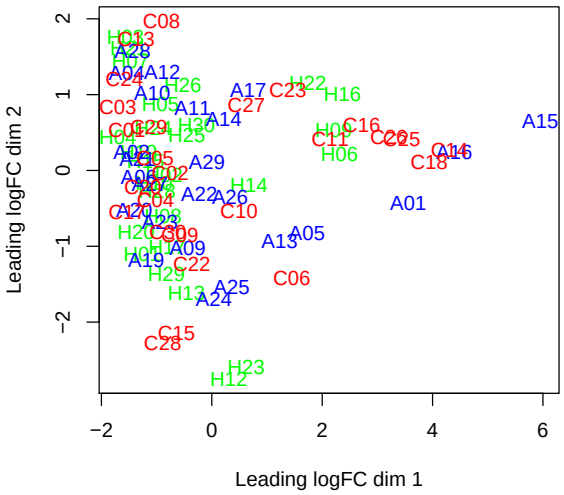

Figure 2: *Multidimensional scaling (MDS) abundance plot based on log2 fold-change (logFC) of KEGG annotation counts derived from gut microbiome of patients that are healthy (H) or diagnosed with adenoma (A) or cancer (C).*

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. 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

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, results, and discussion for these analyses.

insert contribution statement here

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.

2. **Zackular JP**, **Rogers MA**, **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.

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.

Table 1: *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

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.

8. **Wickham H**. 2016. Ggplot2: Elegant graphics for data analysisSecond edition. Springer, Cham.

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 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.

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.

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.

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.