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.

make the repo public before submitting

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 1267 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). The top 500 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

Example of referring to a figure in text (Fig. 1) and including it inline with the text.

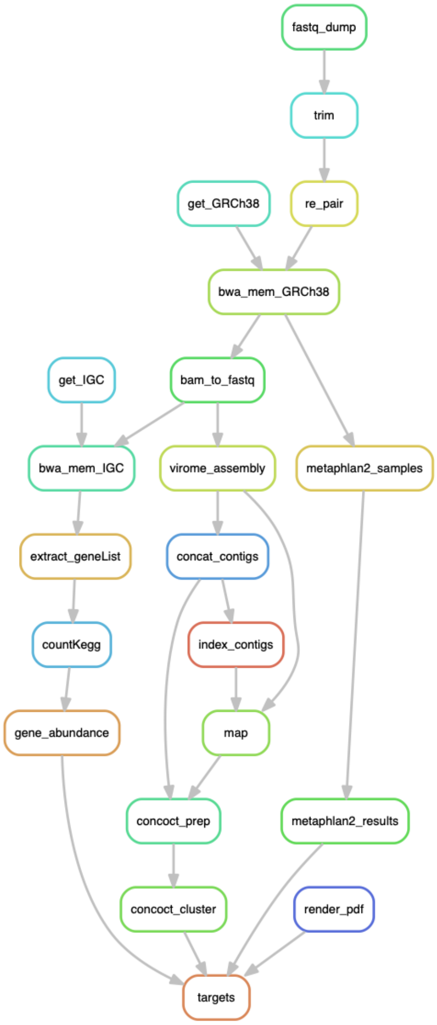

Figure 1: Example

remove example figure before submitting

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 and results 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 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.

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.

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. **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.

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.