

# Investigating the microbial community of the human gut in colorectal cancer

## Questions / Hypothesis

1. Are there differences in bacterial & viral community composition between human stool samples from patients diagnosed as normal, adenoma, and carcinoma?
2. Are there differences in model performance between random forest and logistic regression when classifying samples as healthy or cancerous based on bacterial & viral taxa?
  - Hannigan et al (1) used a random forest model with OTUs or OVUs as features. Recent work suggests that logistic regression may perform just as well as random forest for OTU classification but with fewer computing resources (2).
3. Are there differences in functional potential between metagenomes from normal, adenoma, and carcinoma samples?
  - There may be differences at the OTU level between conditions, but that may or may not translate to changes at the gene level in the bacterial community. Annotating genes with known functions is one step closer to characterizing the functional composition of the community (3).

## Datasets

1. Number of observations/samples in each group.
  - 90 total human stool samples. 30 each from normal, adenoma, & (colorectal) carcinoma.
2. Type of data
  - 16S rRNA gene sequence
  - metagenomes
  - metaviromes
3. Source of data
  - Paper: Hannigan et al 2017 (1)
  - NCBI BioProject ID: PRJNA389927

## Analysis plan

1. OTU clustering of 16S sequences (bacterial taxonomy) and OVU clustering of metaviromes (viral taxonomy).
2. Machine learning classification of samples as healthy or cancerous with OTUs & OVUs as model features. Compare performance of random forest vs other model(s) (e.g. logistic regression).
3. Assemble metagenomes, annotate genes with KEGG ids, and compare across samples.

## Tools

- CONCOCT (4)
- MGS-Fast (5)
- mothur (6)
- python (7)
- R (8)
  - caret (9)
  - R Markdown (10)
  - tidyverse (11)
  - vegan (12)
- snakemake (13)

We will use `conda` to manage our dependencies.

## Strengths of each person relative to project requirements

- Brittany - data analysis, scientific writing, mothur, R
- Christina - experimental design, data analysis, R, python.
- Katelyn - experimental design, scientific writing, carcinogenesis, R
- Kelly - 16S sequence analysis with mothur, pipelines with Snakemake, collaboration with git/GitHub, R/tidyverse, python.
- Nicole - scientific writing, colorectal carcinoma, problem solving with google, R, python, linux.

## 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. **Topçuoğlu BD, Lesniak NA, Ruffin M, Wiens J, Schloss PD.** 2019. Effective application of machine learning to microbiome-based classification problems. bioRxiv 816090. doi:10.1101/816090.
3. **Carr R, Borenstein E.** 2014. Comparative Analysis of Functional Metagenomic Annotation and the Mappability of Short Reads. PLOS ONE 9:e105776. doi:10.1371/journal.pone.0105776.
4. **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.
5. **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.
6. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Horn DJV, Weber CF.** 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microbiol 75:7537–7541. doi:10.1128/AEM.01541-09.
7. **Van Rossum G, Drake FL.** 2009. Python 3 reference manual. CreateSpace, Scotts Valley, CA.
8. **R Core Team.** 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
9. **Kuhn M.** 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28:1–26. doi:10.18637/jss.v028.i05.
10. 2020. Rmarkdown: Dynamic documents for r.
11. **Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Golemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H.** 2019. Welcome to the Tidyverse. Journal of Open Source Software 4:1686. doi:10.21105/joss.01686.
12. **Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H.** 2019. Vegan: Community ecology package.
13. **Köster J, Rahmann S.** 2012. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28:2520–2522. doi:10.1093/bioinformatics/bts480.