Extreme value analysis of gut microbial alterations in colorectal cancer

S.D. Song, P. Jeraldo, J. Chen, and N. Chia

1 Neuroscience Program, Wellesley College, 106 Central St., Wellesley, MA 02481
2 Microbiome Program, Center for Individualized Medicine, Mayo Clinic, 200 First St. SW, Rochester, MN 55905
3 Department of Surgery, Mayo Clinic, 200 First St. SW, Rochester, MN 55905
4 Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, 200 First St. SW, Rochester, MN 55905

(Dated: July 26, 2018)

Gut microbes play a key role in colorectal carcinogenesis, yet reaching a consensus on which microbes remains challenging in part due to reliance on mean value estimates. We present an extreme value analysis for overcoming these limitations. By characterizing a power law fit to the relative abundances of microbes, we capture the same microbial signatures as more complex meta-analyses. Importantly, we show that our method is robust to the variations inherent in microbial community profiling and point to future directions for developing sensitive, robust analytical methods.

PACS numbers: 87.19.xj, 87.23.Cc, 89.75.Da

Colorectal cancer (CRC) is the third most commonly diagnosed cancer in the United States [1], resulting in an estimated 50,000 deaths annually [2]. Development of sporadic, i.e., non-hereditary, CRC is a complex process typically defined by the adenoma-carcinoma sequence, where there is first a transition from a normal colon epithelium to an adenomatous growth followed by a transition to a cancerous tumor [3]. Recently, evidence has been mounting that alterations in the gut microbiome—the approximately 100 trillion microbes residing in the gut—play a crucial role in this transition from normal epithelium to cancerous tumor [4,5].

Profiling the taxonomic composition of the gut microbiome has been made possible due to recent advances in 16S rRNA sequencing, a technique that quantitatively sequences hypervariable regions of the microbial rRNA present in a sample and assigns taxonomy accordingly [7,8]. This allows for comparison between the microbial profiles of healthy subjects versus subjects with CRC. Identifying taxa that are enriched in CRC subjects serves to narrow down the list of microbes that potentially drive CRC progression. While this approach is sound in principle, results can be difficult to interpret, particularly when there are inconsistencies between different studies [5,9]. For instance, oral pathogen Peptostreptococcus stomatis is often considered a known associate of CRC, with several studies having corroborated these results [10,11]. However, other studies show a very weak [4] or even no association at all [5,17,20]. At the same time, these studies often find a number of other possible signatures; thus, instead of clarifying the microbial drivers, these studies create additional confusion.

One commonly proposed reason for the inconsistencies between studies is the existence of multiple mechanisms by which microbes can promote CRC [5,9,18,19]. The result is that CRC drivers in one case may be uninvolved in another, with even the few confirmed CRC signatures displaying enrichment in, at most, a subset of cancerous stool or tumor biopsy samples [9,16,21]. Typical analyses used to find these signatures rely on either the mean or a rank ordering of taxa abundances and therefore cannot reliably detect trends that only occur in a subset of samples. In other words, these measures only detect general and unidirectional shifts when, in fact, microbial effects are neither general nor linear. Further exacerbating these issues, mean value estimates often assume symmetrical Gaussian distributions, which do not reflect real microbial distributions. The field is in crucial need of reliable analysis methods to parse out the signal from the noise.

In this study, we make use of extremal distributions as a more sensitive and consistent means of identifying the major microbial signatures associated with CRC progression. In doing so, we show that the relative abundance distributions of putatively causative taxa follow a bounded power law whose tail, i.e., extreme values, differs between normal, adenoma, and CRC samples. We use a permutation test with extreme value test statistics to quantify the differences in these asymmetrical distributions. We show our extremal analysis to be robust by replicating our findings in a separate, European cohort [11], also corroborating the results of recent meta-analyses [9,16,22]. Employing a power law distribution to understand the role of potential microbial culprits in CRC progression will motivate, guide, and simplify future development of analytical methods for microbial data.

We examined the fecal microbiota of subjects who are healthy (n = 549), have adenomas (n = 253), or have CRC (n = 20) by using 16S rRNA gene sequences from Hale et al. [6]. The relative abundance of each taxon commonly cited in the field was determined at the genus level for all samples. While the relative abundance of high-abundance microbes show a distribution...
FIG. 1: Relative abundance distributions of typical high-abundance (Coprococcus) and low-abundance (Desulfovibrio) microbes. The former is more Gaussian, while the latter is heavily right-skewed.

FIG. 2: Log-log transformation reveals that Desulfovibrio follows a power law relative abundance distribution.

Characterization of low-abundance taxa as following a power law distribution of relative abundances should not, in fact, be surprising. Observations of power laws are prevalent in nearly every field, ranging from economics to ecology, describing events from wealth distribution in the US to wildfire sizes [23]. Despite even a couple mentions of power laws related to the human microbiome [23–26], the potential application of power laws in the analysis of microbial data has largely gone unnoticed. Yet much can be gleaned from understanding the limiting distribution that events follow [23–27, 28]. One insight from other fields is that the extreme values comprising the heavy tail of power laws are often the most influential and informative to study [29, 30]. Additionally, power laws characteristically lack a finite mean [23], deeming approaches based on mean values less appropriate. These realizations call for extreme value analysis (EVA).

Specific to this study, we propose that the power law tails of CRC drivers are the values that differ between the normal, adenoma, and cancer stages of CRC progression. Examining data from subjects at each stage of the adenoma-carcinoma sequence can distinguish a microbe’s role in the etiology of CRC [31]. For example, a microbe that exhibits depletion of extreme values in normal samples but enrichment in adenoma samples would indicate a role in the early transition to cancer, whereas a microbe that exhibits only an enrichment in cancer samples would not.

In order to focus our analysis on extreme rather than mean values, we perform a permutation test on an extreme value measure to determine the role of microbial taxa in CRC etiology. An advantage of using a permutation test is that it is non-parametric, i.e., we do not make faulty assumptions about distributional symmetry, and we can choose any extreme value measure to test. One such example is a simple maximum [32]; however, this statistic is overly dependent on a single data point, making it sensitive to random fluctuations and technical artifacts. Instead, we found averaging over the x greatest values ($\max_x$) to be a more reliable measure of extreme behavior. We tested this approach for different values of $x$ ($\max_3$, $\max_5$, and $\max_7$) and found nearly identical results (see Appendix). We carried on by performing the permutation test using only the $\max_5$ test statistic.

By comparing the group $\max_5$ values generated from the 10,000 random permutations of normal, adenoma, and cancer to the observed $\max_5$ values, we were able to conclude whether the observed extreme values in each group are higher or lower than expected. When applied to Desulfovibrio, we find that the $\max_5$ is lower than expected in the normal group and higher than expected in the adenoma group, indicating a role for this genus in the transition from normal to adenoma (Fig. 3). Applying these methods to the 10 other taxa studied, results were also found for Fusobacterium and Peptostreptococcus, both enriched in cancer (Table I).

As additional confirmation, we implement an ex-
3

FIG. 3: Permutation distributions of the average of maximum 5 relative abundances test statistic (max5) for Desulfovibrio. 10,000 random permutations were generated to produce reference distributions of the expected max5 values in each group: normal, adenoma, and cancer. Lines indicate the corresponding observed values. Note that the observed max5 appears significantly lower than expected in normal and higher than expected in adenoma.

TABLE I: Taxa with significantly different than expected extreme values at each stage of CRC progression. Left- (<) or right- (>) sided p-values are shown from the permutation test using the max5 test statistic. * p < .05; ** p < .01

| Genus            | Normal | Adenoma | Cancer |
|------------------|--------|---------|--------|
| Desulfovibrio    | .012*  | .009**  | .138   |
| Direction        | <      | >       | <      |
| Fusobacterium    | .167   | .465    | .038*  |
| Direction        | <      | <       | >      |
| Peptostreptococcus| .013*  | .292    | 0**    |
| Direction        | <      | <       | >      |

The exploratory parametric EVA method that explicitly applies the power law model to estimate an expected number of extreme values. For this, we normalize a power law fit to the microbial abundances to generate a probability density function (PDF). The PDF served to estimate the number of expected extreme values above a defined critical relative abundance value, which we then compare to the observed number of extreme values. Our results match the permutation test results for Peptostreptococcus and are non-significant but trending in the same direction for Desulfovibrio and Fusobacterium. The uncertainty in the power law parameters lowers the detected significance, but overall these results serve to confirm our findings based on the permutation test.

Finally, we examine the robustness of our approach. Inconsistencies between different studies that arise due to different microbial pathways of promoting CRC affect average enrichment or depletion. However, overall distributional characteristics should remain relatively stable. We therefore expect our extreme value approach, which makes assessments based off the observed distribution, to be more robust to these variations. In order to test this, we replicated our results in a separate European cohort (61 normal, 42 adenoma, and 53 cancer) obtained from Zeller et al. [11]. The same 11 genera were tested as previously described using the permutation test. Indeed, results for all of the signatures found in Hale et al. (Desulfovibrio, Fusobacterium, and Peptostreptococcus) were confirmed (Table II).

Overall, our results are very consistent, demonstrating the robustness of EVA. It is worth noting that the results of our relatively simple permutation model performed on a single dataset captures many of the same signatures that otherwise have only been confirmed by more complex meta-analyses, most of which show agreement on the importance of Fusobacterium and Peptostreptococcus [9, 10, 22]. The match in results is a testament to the equal, if not greater, sensitivity and increased robustness of EVA compared to other methods.

It is worth addressing the two results that differ between the American and European cohort. Specifically, the European cohort results show Porphyromonas and Akkermansia to have significant positive associations with cancer (Table II). However, these appear to be the result of differences in study design. Porphyromonas was completely undetected by Hale et al., making it methodologically impossible to assess by any method. For Akkermansia, cohort size for the cancer group was larger in the European cohort, making significance easier to assess in
the adenoma to cancer transition. In addition to study design-based limitations, it should also be noted that the p-values reported here were calculated without multiple hypothesis correction. However, the number of genera considered is relatively small, and our several replications of the results provide an additional level of stringency.

EVA also has some intrinsic limitations. By definition, EVA requires capturing rare events and thus performs best with a large number of samples. Fortunately, large-scale microbiome studies are becoming increasingly common as high-throughput sequencing becomes more accessible. In this context, EVA takes advantage of this progress while being less influenced by the variations inherent in microbial community profiling than mean value analyses. For the parametric testing, estimating the power law parameters was challenging because the power law fit represents a crude estimate of microbial extreme value behavior. Improving the power law model and its corresponding PDF would further enhance the applicability of this approach.

Future work may shift the paradigm of analytical methods for 16S rRNA sequencing data. Already, we have demonstrated the power of EVA to generate consistent results despite the multiple different mechanisms by which the microbiome drives CRC. Understanding that relative abundances of potential CRC drivers follow the ubiquitous power law distribution provides a guiding framework for developing future analytical tools. The characteristic lack of a finite mean exhibited by power laws explains the challenges of mean value analyses, motivating a shift towards EVA. Advancements in EVA, guided by this realization, will result in a clearer, simpler, and more accurate way to understand the role that key gut microbes play in the development and progression of CRC.

We would like to extend our gratitude to Prof. Vanessa Hale for providing the data, Profs. Marc Tetel and Cassandra Pattanayak for continuing support and encouragement, and members of the Chia group and Prof. Marina Walther-Antonio for useful discussions. This work was supported by a NIH grant (R01 CA 179243) as well as the Arnold and Mabel Beckman Foundation and Patterson Funds from the Neuroscience Program at Wellesley.

[1] B.K. Edwards, E. Ward, B.A. Kohler, C. Eheman, A.G. Zauber, R.N. Anderson, A. Jemal, M.J. Schymura, I. Lansdorp-Vogelaar, et al., Cancer 116, 544 (2010).
[2] R.L. Siegel, K.D. Miller, and A. Jemal, CA Cancer J. Clin. 66, 7 (2016).
[3] E.R. Fearon and B. Vogelstein, Cell 61, 759 (1990).
[4] J. Ahn, R. Sinha, Z. Pei, C. Dominianni, J. Wu, J. Shi, J.J. Goedert, R.B. Hayes, and L. Yang, J. Natl. Cancer Inst. 105, 1907 (2013).
[5] J.P. Zackular, N.T. Baxter, K.D. Iverson, W.D. Sadler, J.F. Petrovino, G.Y. Chen, and P.D. Schloss, mBio 4, e00692 (2013).
[6] V.L. Hale, J. Chen, S. Johnson, S.C. Harrington, T.C. Yab, T.C. Smyrk, H. Nelson, L.A. Boardman, B.R. Druliner, T.R. Levin, et al., Cancer Epidemiol. Biomarkers Prev. 26, 85 (2017).
[7] H. Hayashi, M. Sakamoto, and Y. Benno, Microbiol. Immunol. 46, 535 (2002).
[8] J.M. Janda and S.L. Abbott, J. Clin. Microbiol. 45, 2761 (2007).
[9] J.L. Drewes, J.R. White, C.M. Dejea, P. Fathi, T. Iyadurai, J. Vaidivelu, A.C. Roslandi, E.C. Wick, E.F. Monroe, M.F. Loke, et al., NPJ Biofilms Microbiomes 3, 34 (2017).
[10] T. Wang, G. Cai, Y. Qiu, N. Fei, M. Zhang, X. Pang, W. Jia, S. Cai, and L. Zhao, ISME J. 185, 320 (2014).
[11] G. Nakatsu, X. Li, H. Zhou, J. Sheng, S.H. Wong, W.K. Wu, S.C. Ng, H. Tsoi, Y. Dong, N. Zhang, et al., Nat. Commun. 6, 6528 (2015).
Appendix: Sensitivity of Extremal Averages

One method of measuring extreme values is to examine sample maxima; however, a single maximum value is often sensitive to random fluctuations and technical artifacts. Averaging a number of maximum values remedies this problem. Here, we considered using the average of 3 (\( \max_3 \)), 5 (\( \max_5 \)), and 7 (\( \max_7 \)) maximum values in order to derive the test statistic that is most suitable. All three measures performed nearly identically (Table III). The index of dispersion was used as a quantitative measure of the differences between the test statistics,

\[
D = \frac{\sigma^2}{\mu}
\]  

where \( \sigma^2 \) is the variance and \( \mu \) is the mean of the three p-values for each group. Indeed, the resulting indices of dispersion were very low, ranging from \( D = 0 \) to .016.

| Genus          | Test Statistic | Normal | Adenoma | Cancer |
|---------------|----------------|--------|---------|--------|
|               | \( \max_3 \)   | .011*  | .027*   | .141   |
| Desulfovibrio | \( \max_5 \)   | .012*  | .009**  | .138   |
|               | \( \max_7 \)   | .011*  | .007**  | .138   |
| Fusobacterium | \( \max_3 \)   | .255   | .563    | .038*  |
|               | \( \max_5 \)   | .167   | .465    | .038*  |
|               | \( \max_7 \)   | .163   | .392    | .038*  |
| Peptostreptococcus | \( \max_3 \)  | .013*  | .292    | 0**    |
|               | \( \max_5 \)   | .013*  | .292    | 0**    |
|               | \( \max_7 \)   | .013*  | .292    | 0**    |