This is the accepted manuscript made available via CHORUS. The article has been published as:

Extreme value analysis of gut microbial alterations in colorectal cancer

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

Phys. Rev. E **99**, 032413 — Published 13 March 2019

DOI: 10.1103/PhysRevE.99.032413

Extreme value analysis of gut microbial alterations in colorectal cancer

S.D. Song,[1] P. Jeraldo,[2,3] J. Chen,[4] and N. Chia[2,3,*]

[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: February 13, 2019)

Gut microbes play a key role in colorectal carcinogenesis, yet reaching a consensus on microbial signatures remains a challenge. This is in part due to a 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, reliable 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 [4–6].

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]. The microbial profiles of healthy subjects can then be compared to the profiles of subjects with CRC. Specifically, one common approach is to identify taxa that are enriched in CRC subjects, then narrow down the list of microbes that potentially drive CRC progression. This may be sound in principle, but results are difficult to interpret in practice when inconsistencies arise 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 [10–16]. Meanwhile, 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. Recent meta-analyses systematically analyze previously published 16S gene sequence data in an effort to identify consistent signatures. However, these results only confirm a limited number of species, and even highlight that the majority of results from individual studies do not agree [9, 16].

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 [22] or zero-inflated Poisson distributions [23, 24], 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 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, 25]. 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.

*Electronic address: Chia.Nicholas@mayo.edu

A power-law-like distribution of microbial relative abundances should not, in fact, be surprising. Observations of power laws are prevalent in nearly every field, and reflect the universal importance of extreme events across fields ranging from economics to ecology, describing events from wealth distribution in the US to wildfire sizes [26]. Even just within the field of physics, power law behaviors have received special attention; complex social networks [27], avalanches [28], scientific citations [29], and football goals [30] have all been characterized to follow power laws. While we cannot claim a power law distribution in our own work, understanding the characteristics of power laws can nonetheless guide the analysis of power-law-like data and reveal important insights. For example, one insight gained is that the extreme values comprising the heavy tail of power laws are often the most influential and informative to study [31, 32]. A Gaussian approximation of a power law tail would underestimate the extreme values, thereby underestimating the mean. This insight deems approaches based on mean values less appropriate and calls for extreme value analysis (EVA).

We examine the fecal microbiota of subjects who are healthy (n = 486), have adenomas (n = 233), or have CRC (n = 17) by using 16S rRNA gene sequences from Hale et al. [6]. The relative abundance of each taxon commonly cited in the field was determined for all samples at the genus level, the smallest taxonomic category able to be identified accurately. Average sequence read count was 6219, and samples with fewer than 2000 reads were removed as done in the original analysis by Hale et al. [6]. Removing these samples protects against the sensitivity to sequencing depth variations that relative abundances at the lower limit face.

While the relative abundances of high-abundance microbes show a distribution more similar to a symmetrical Gaussian distribution, low-abundance microbes display a highly right-skewed distribution with a heavy tail. Typical examples of genera displaying these two types of distributions are shown in Fig. 1.

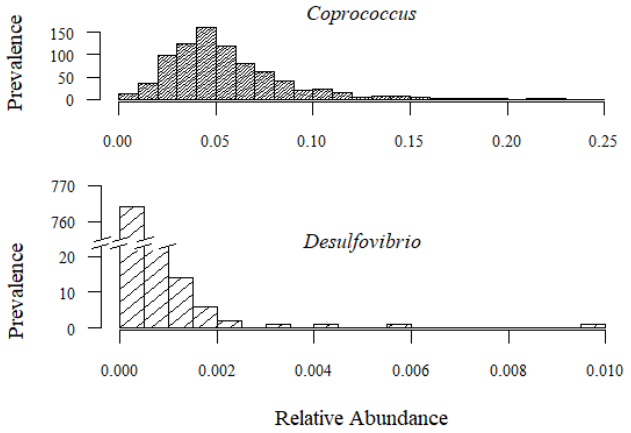

FIG. 1: Relative abundance distributions of typical high-abundance (*Coprococcus*) and low-abundance (*Desulfovibrio*) microbes. The former is more Gaussian, while the latter is heavily right-skewed.

As shown in Fig. 2, the right-skewed distributions do, in fact, approximate power laws. However, it should be noted that power laws are characterized by extreme behavior that extends to infinity, while relative abundances are bound by 1. In this study we only consider low-abundance genera with a mean relative abundance of less than 0.005. This left us with 11 out of the 18 total taxa characterized [45] (*Akkermansia, Bifidobacterium, Bilophila, Desulfovibrio, Epulopiscium, Fusobacterium, Lachnobacterium, Parvimonas, Peptostreptococcus, Pseudomonas*, and *Streptococcus*).

Upon closer inspection, the length of the power law tail appears to differ according to disease status (data not shown). We therefore propose that the power law tails, i.e. extreme values, of CRC drivers are the values that differ between the normal, adenoma, and cancer stages of CRC progression. In order to assess this, we employ extreme value analysis to quantify these differences. Examining data from subjects at each stage of the adenoma-carcinoma sequence can then distinguish a microbe's role in the various stages of carcinogenesis [33]. 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 suggest a later effect.

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, hence we do not make faulty assumptions about distributional symmetry. We can also choose any extreme value measure to test. One such example is a simple maximum [34]; 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. This statistic represents the maximum while eliminating noise, and it defines the "leading edge" of the abundance values.

While we utilize the previously mentioned insight about power law tails to focus our analysis on extreme values, importantly, our non-parametric approach does not rely on estimates of power law parameters, or even a rigorous claim that the relative abundances follow a strictly power law distribution.

We test this approach for different values of x (max_3, max_5, and max_7) and find nearly identical results despite different sample sizes in each group (see Appendix A). These results indicate that these defined sizes of the leading edge are sufficient to produce reliable results, independent of sample size. We carry on by performing the permutation test using only the max_5 test statistic.

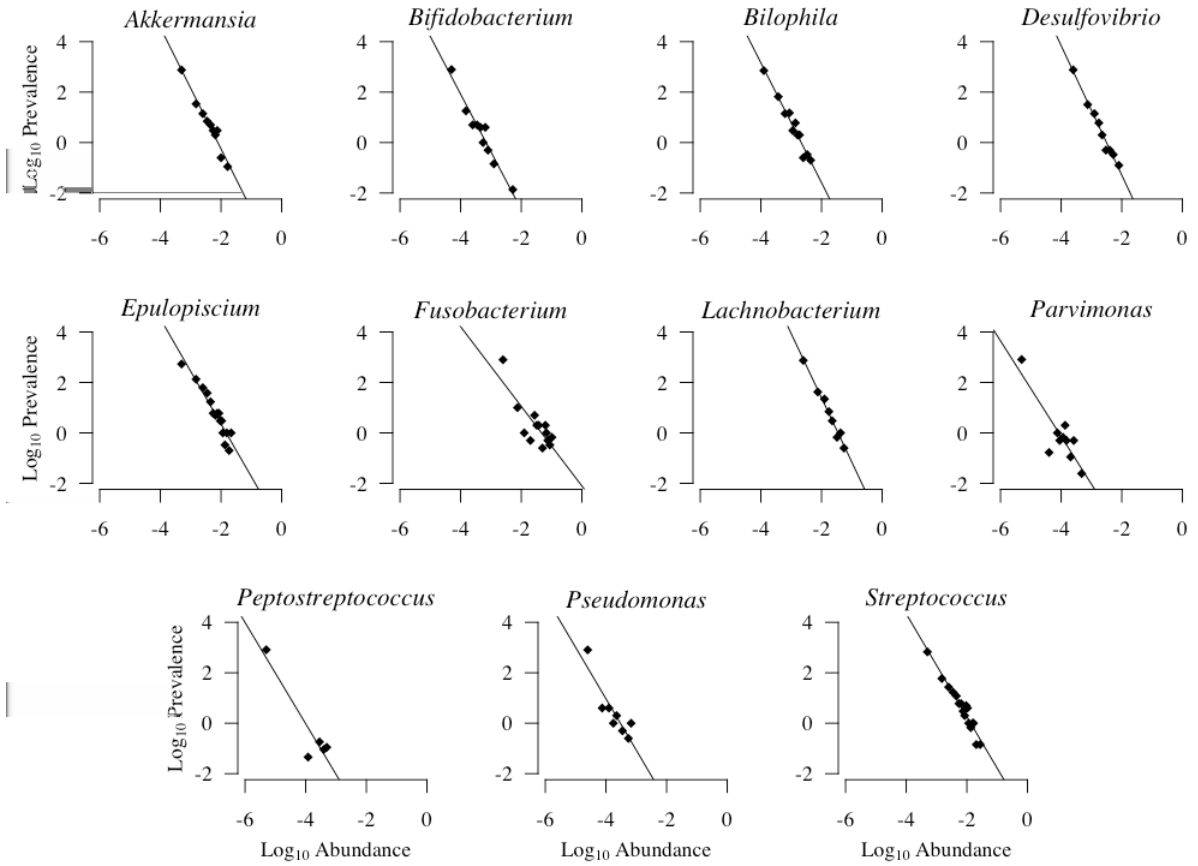

FIG. 2: Log-log transformations reveal that the 11 genera examined approximate power law relative abundance distributions.

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 are able to conclude whether the observed extreme values in each group are higher or lower than expected. Note that the calculated expectations for extreme values depend on sample size, as larger samples are more likely to include more extreme values than smaller samples. Correspondingly, cohorts with different sample sizes between groups lead to different expectations for extreme values in each group as well.

When applied to *Desulfovibrio*, we find that the max_5 statistic 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 are also found for *Fusobacterium* and *Peptostreptococcus*, both enriched in cancer (Fig. 4).

As additional confirmation, we implement an exploratory parametric EVA method that explicitly applies the power law model to estimate an expected number of extreme values (see Appendix B for detailed methods). Briefly, we normalize a power law fit to the microbial abundances and 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* ($p = 0$ in the normal group, $p = .015$ in the adenoma group, $p = .009$ in the cancer group) and are non-significant but trending in the same direction for *Desulfovibrio* ($p = .065$ in the normal group) and *Fusobacterium* ($p = .141$ in the normal group). The uncertainty in the power law parameters lowers the detected significance, but overall these results strengthen our findings based on the permutation test. Improving the power law model and its corresponding PDF would further enhance the applicability of this approach.

Finally, we examine if our results can be generalized to other studies. Inconsistencies between different studies arise due to many different factors including batch effects [35, 36] and sequencing biases from collection [37–39], primer design [40], PCR conditions [41], and lab-to-lab variation [37, 42]. However, an advantage of our approach is that the 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 replicate our results in a separate European cohort (61 normal, 42 adenoma, and 53 cancer) obtained from Zeller et al. [11] that used different collection methods, primer design, and sequencing

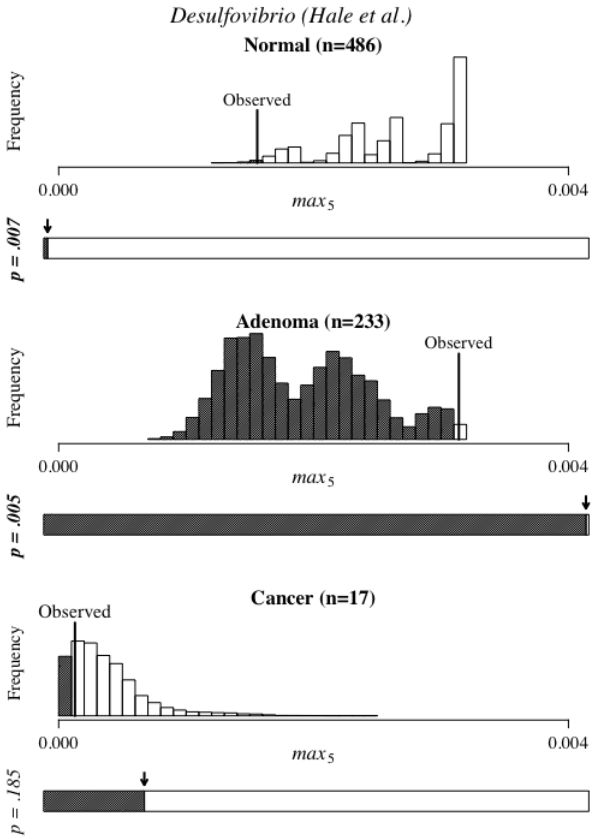

FIG. 3: Permutation distributions of the average of maximum 5 relative abundances test statistic (max_5) for *Desulfovibrio*. 10,000 random permutations were generated to produce expected distributions of the max_5 values in each group: normal, adenoma, and cancer. Lines indicate actually observed max_5 values. An observed value near the middle of the expected distribution would indicate an expected result, while one near the ends would indicate significance. Consequently, observed max_5 is significantly lower than expected in normal and higher than expected in adenoma. In order to visualize the significance and direction of results, barcharts are shown below each histogram, with x-axis representing left-sided p-values ranging from 0 to 1. One-sided p-values (whichever is smaller) are also labeled on the left of the barcharts (bold p-values $< .05$). Therefore, a significantly lower than expected result is indicated by a small shaded region, while a significantly higher than expected result is indicated by a large shaded region.

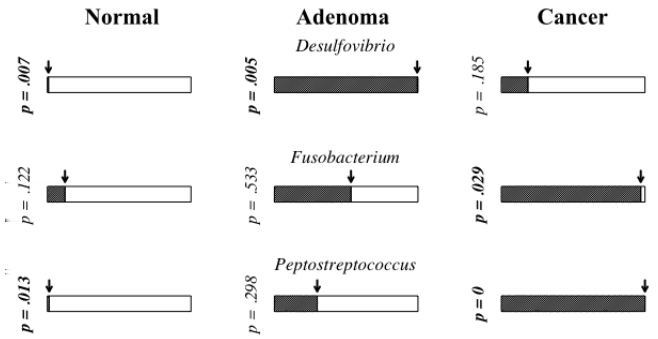

FIG. 4: Taxa from Hale et al. with significantly different than expected extreme values at each stage of CRC progression. One-sided p-values are shown from the permutation test using the max_5 test statistic. Note that results fBold p-values $< .05$

platforms processed in a different lab [46]. The same 11 genera are tested as previously described using the permutation test. Indeed, results for all of the signatures found in Hale et al. (*Desulfovibrio*, *Fusobacterium*, and *Peptostreptococcus*) are confirmed; *Desulfovibrio* is found to be depleted in normal, and *Fusobacterium* and *Peptostreptococcus* enriched in cancer but depleted in normal (Fig. 5).

It is worth addressing the two results that differ between the American and European cohort. Specifically, the European cohort results show *Akkermansia* and *Porphyromonas* to have significant positive associations with cancer (Fig. 5). However, these appear to be the result of differences in study design. For *Akkermansia*, cohort size of the cancer group was larger in the European cohort, making significance easier to assess in the adenoma to cancer transition. *Porphyromonas* was completely undetected by Hale et al., making it methodologically impossible to assess computationally. In addition to study design-based limitations, it should also be noted that the p-values reported here are 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.

Overall, our results are very consistent, demonstrating the robustness of EVA. It is important to note that 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, 16, 25]. The match in results is a testament to the equal, if not greater, sensitivity and reliability of EVA compared to other methods. Moreover, EVA identified *Desulfovibrio* and *Akkermansia* to be significantly associated with the transition to adenoma and cancer stage, respectively, a result that went undetected in the original analysis by Zeller et al. [11].

Nonetheless, EVA has some 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 [43, 44] 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.

Future work may shift the paradigm of analytical methods for 16S rRNA sequencing data. Already, we have demonstrated the power of EVA to generate con-

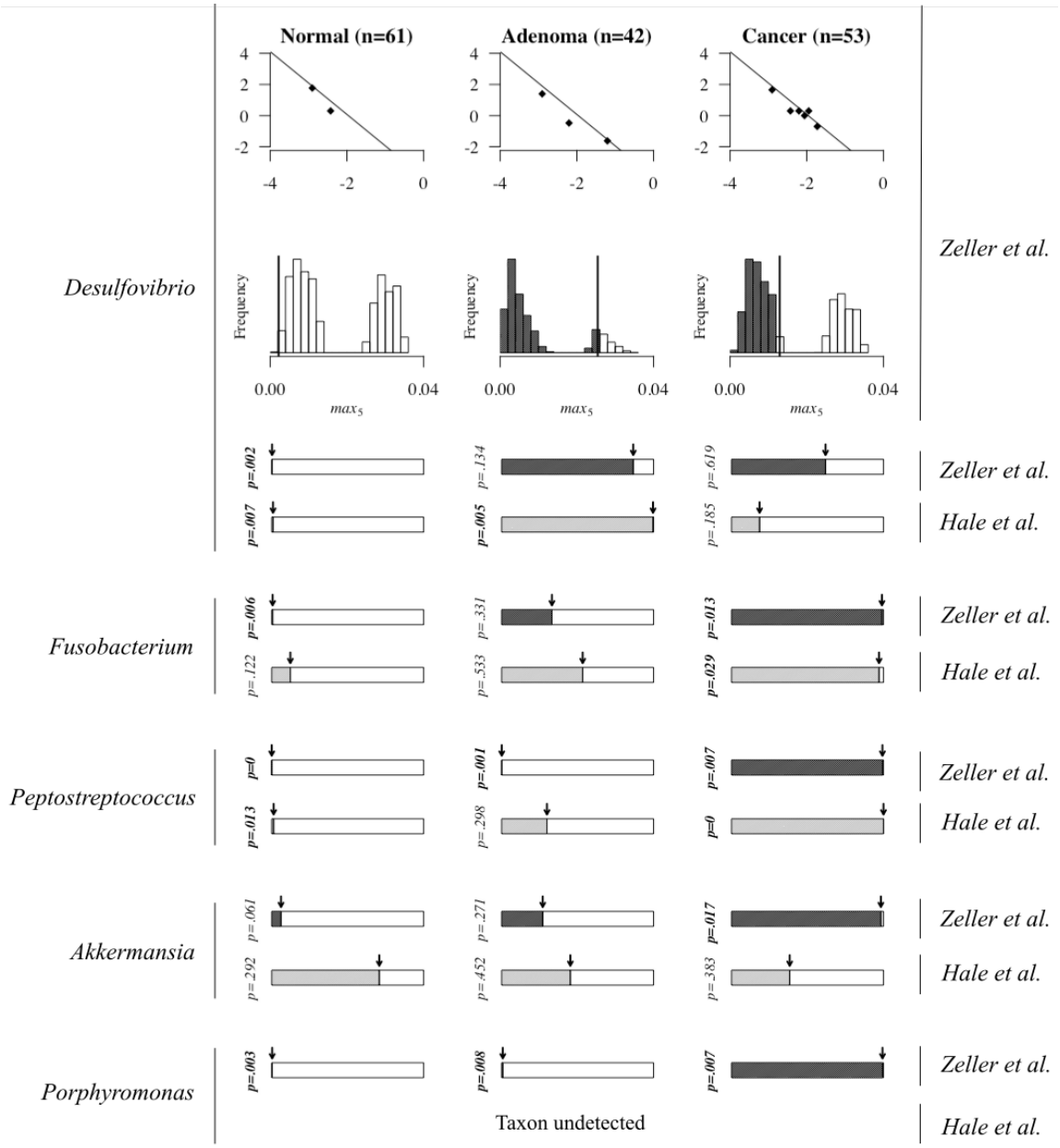

FIG. 5: Replication of results in separate, European cohort (Zeller et al.) and comparison to previous Hale et al. results. Log-log transformations and permutation distributions for *Desulfovibrio* using the max_5 statistic are shown as a representative example. One-sided p-values from the permutation test are visualized for both Zeller et al. and Hale et al. to compare. Note that results support consistent conclusions in both cohorts for *Desulfovibrio*, *Fusobacterium*, and *Peptostreptococcus*. *Porphyromonas* was undetected in Hale et al. Bold p-values $< .05$

sistent 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 an 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. Petrosino, 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. Iyadorai, J. Vadivelu, A.C. Roslani, E.C. Wick, E.F. Mongodin, 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. **6**, 320 (2012).

[11] G. Zeller, J. Tap, A.Y. Voigt, S. Sunagawa, J.R. Kultima, P.I. Costea, A. Amiot, J. Bohm, F. Brunetti, N. Habermann, et al., Mol. Syst. Biol. **10**, 766 (2014).

[12] Q. Feng, S. Liang, H. Jia, A. Stadlmayr, L. Tang, Z. Lan, D. Zhang, H. Xia, X. Xu, Z. Jie, et al., Nat. Commun. **6**, 6528 (2015).

[13] 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**, 8727 (2015).

[14] N.T. Baxter, M.T. IV Ruffin, M.A. Rogers, and P. D. Schloss, Genome Med. **8**, 37 (2016).

[15] J. Yu, Q. Feng, S.H. Wong, D. Zhang, Q. Liang, Y. Qin, L. Tang, H. Zhao, J. Stenvang, Y. Li, et al., Gut **66**, 70 (2017).

[16] M.S. Shah, T.Z. DeSantis, T. Weinmaier, P.J. McMurdie, J.L. Cope, A. Altrichter, J.M. Yamal, and E.B. Hollister, Gut **67**, 882 (2018).

[17] M. Castellarin, R.L. Warren, J.D. Freeman, L. Dreolini, M. Krzywinski, J. Strauss, R. Barnes, P. Watson, E. Allen-Vercoe, R.A. Moore, and R.A. Holt, Genome Res **22**, 299 (2012).

[18] A.D. Kostic, D. Gevers, C.S. Pedamallu, M. Michaud, F. Duke, A.M. Earl, A.I. Ojesina, J. Jung, A.J. Bass, J. Tabernero, et al., Genome Res. **22**, 292 (2012).

[19] J.P. Zackular, M.A. Rogers, M.T. IV Ruffin, and P.D. Schloss, Cancer Prev. Res. (Phila) **7**, 1112 (2014).

[20] B. Flemer, D.B. Lynch, J.M. Brown, I.B. Jeffery, F.J. Ryan, M.J. Claesson, M. O'Riordain, F. Shanahan, and P.W. O'Toole, Gut **66**, 633 (2017).

[21] V.L. Hale, P. Jeraldo, J. Chen, M. Mundy, J. Yao, S. Priya, G. Keeney, K. Lyke, J. Ridlon, B.A. White, et al., bioRxiv (2018).

[22] B.M. Bolker, M.E. Brooks, C.J. Clark, S.W. Geange, J.R. Poulsen, M.H. Stevens, and J.S. White, Trends Ecol Evol **24**, 127 (2009).

[23] D. Lambert, Technometrics **34**, 1 (1992).

[24] L. Xu, A.D. Paterson, W. Turpin, and W. Xu, PLoS One **10**(7), e0129606 (2015).

[25] C. Duvallet, S.M. Gibbons, T. Gurry, R.A. Irizarry, and E.J. Alm, Nat. Commun. **8**, 1784 (2017).

[26] M.E.J. Newman, Contemp. Phys. **46**, 323 (2005).

[27] I.J. Farkas, I. Derenyi, A.L. Barabasi, and T. Vicsek, Phys Rev E **64**, 026704 (2001).

[28] E. Vives, J. Ortin, L. Manosa, I.I. Rafols, R. Perez-Magrane, and A. Planes, Phys Rev Lett **72**(11), 1694 (1994).

[29] S. Redner, Eur Phys J B **4**, 131 (1998).

[30] J. Greenhough, P.C. Birch, S.C. Chapman, and G. Rowlands, Physica A **316**, 615 (2001).

[31] X. Gabaix, P. Gopikrishnan, V. Plerou, and H.E. Stanley, Nature **423**, 267 (2003).

[32] P. Andriani and B. McKelvey, J. Int. Bus. Stud. **38**, 1212 (2007).

[33] G. Mori, S. Rampelli, B.S. Orena, C. Rengucci, G. De Maio, G. Barbieri, A. Passardi, A. Casadei Gardini, G.L. Frassineti, S. Gaiarsa, et al., Sci. Rep. **8**, 10329 (2018).

[34] R.A. Fisher and L.H.C. Tippett, Math. Proc. Camb. Philos. Soc. **24**, 180 (1928).

[35] J.T. Leek, R.B. Scharpf, H.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K. Baggerly, and R.A. Irizarry, Nat Rev Genet **11**(10), 733 (2010).

[36] J. Chen, E. Ryu, M. Hathcock, K. Ballman, N. Chia, J.E. Olson, and H. Nelson, PeerJ **4**, e1514 (2016).

[37] R. Sinha, J. Chen, A. Amir, E. Vogtmann, J. Shi, K.S. Inman, R. Flores, J. Sampson, R. Knight, and N. Chia, Cancer Epidemiol. Biomarkers. Prev. **25**, 407 (2016).

[38] E. Loftfield, E. Vogtmann, J.N. Sampson, S.C. Moore, H. Nelson, R. Knight, N. Chia, and R. Sinha, Cancer Epidemiol Biomarkers Prev **25**(11), 1483 (2016).

[39] E. Vogtmann, J. Chen, A. Amir, J. Shi, C.C. Abnet, H. Nelson, R. Knight, N. Chia, and R. Sinha, Am. J. Epidemiol. **185**, 115 (2017).

[40] P. Jeraldo, N. Chia, and N. Goldenfeld, Environ. Microbiol. **13**, 3000 (2011).

[41] D.M. Gohl, P. Vangay, J. Garbe, A. MacLean, A. Hauge, A. Becker, T.J. Gould, J.B. Clayton, T.J. Johnson, R. Hunter, et al., Nat. Biotechnol. **34**, 942 (2016).

[42] R. Sinha, C.C. Abnet, O. White, R. Knight, and C. Huttenhower, Genome Biol. **16**, 276 (2015).

[43] Consortium Human Microbiome Project, Nature **486** (7402), 207 (2012).

[44] T.S.B. Schmidt, J. Raes, and P. Bork, Cell **172**(6), 1198 (2018).

[45] Taxa characterized include *Akkermansia, Bacteroides, Bilophila, Bifidobacterium, Clostridium, Coprobacillus, Coprococcus, Desulfovibrio, Epulopiscium, Fusobacterium, Lachnobacterium, Parvimonas, Peptostreptococcus, Prevotella, Pseudomonas, Ruminococcus, Streptococcus*, and *Sutterella*.

[46] Data obtained from Zeller et al. was provided as relative abundances and average read length was not given. However, the original report indicates that samples with fewer than 1000 reads were removed [11].

Appendix A: 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 (Fig. 6). The index of dispersion was used as a quantitative measure of the differences between the test statistics,

$$D = \sigma^2/\mu \tag{A1}$$

where σ^2 is the variance and μ 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 .014.

Appendix B: Parametric EVA Methods

The exploratory parametric EVA method first fits a power law to the microbial abundance data, as in Fig. 2. β_0 and β_1 from a linear fit to the log-log transformed data of bin interval I are exponentiated to derive a crude power law fit:

$$y = \beta_0 + \beta_1 x, \tag{B1}$$

$$f(x) = 10^{\beta_0} x^{\beta_1} \tag{B2}$$

A cutoff point, or critical value, is then determined to define the extreme values. A conditional inference tree is used to find the critical value, and the average of 1000 bootstrap samples is taken. We can now derive the total expected number of extreme values (E) from Eq. (B3), using the integral of the power law from the critical value to 1 (the maximum possible relative abundance):

$$\int_c^1 f(x)dx = \frac{10^{\beta_0}(1 - c^{\beta_1+1})}{\beta_1 + 1}, \tag{B3}$$

$$E = \frac{\int_c^1 f(x)dx}{I} \tag{B4}$$

The integral must be divided by bin interval I because counts are discrete. The number of observed extreme values in each group (normal, adenoma, and cancer) is then compared to expected using a binomial test, where x = observed counts, n = group sample size, and $p = E/$ total sample size.

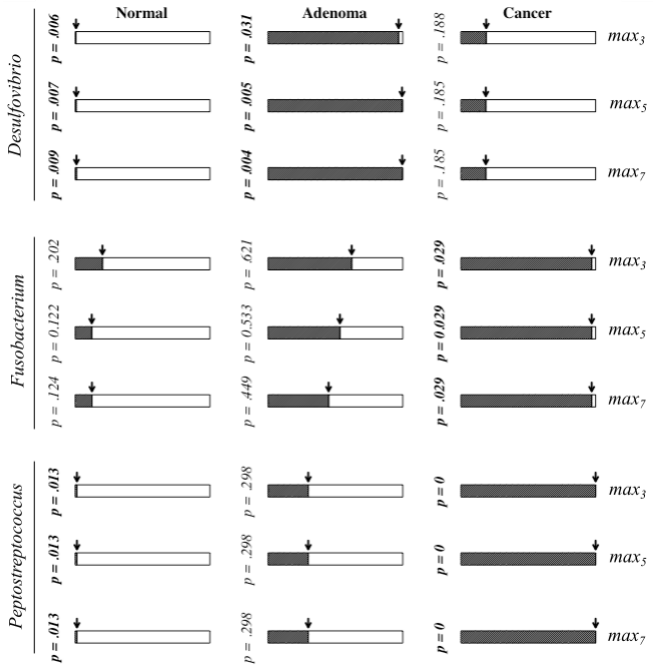

FIG. 6: Sensitivity analysis of permutation test statistic using average of maximum 3 (max_3), 5 (max_5), and 7 (max_7) values. Analyses were conducted using data from *Hale et al.* Bold p-values $< .05$